# Fully Polynomial-Time Approximation Scheme for a Problem of Finding a Subsequence

Alexander Kel'manov[1,2], Sergey Khamidullin[1], and Semyon Romanchenko[1,2]

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

**Abstract.** We consider a strongly NP-hard Euclidean problem of finding a subsequence in a finite sequence under the criterion of the minimum sum of squared distances from the elements of sought subsequence to its geometric center (centroid). It is assumed that the sought subsequence contains a given number of elements. In addition, sought subsequence has to satisfy the following condition: the difference between the indexes of each previous and next points is bounded with given lower and upper constants. We present an approximation algorithm for the problem and prove that it is a fully polynomial-time approximation scheme when the space dimension is bounded by a constant.

**Keywords:** Euclidean space, sequence, minimum sum of squared distances, NP-hardness, FPTAS.

## Introduction

In this paper we study one strongly NP-hard problem of searching a subsequence in a finite sequence of points from Euclidean space. Our aim is to provide an approximation algorithm for this problem.

This work was motivated by poor study of the problem. The problem is also interesting because of its importance for applications, in particular, for mathematical problems of time series analysis, approximation problems and also for applications dealing with data mining problems (see, for example, [1–4] and references therein).

The paper is organized as follows. In the next section the formal definition of the problem under study is given; an example of application (origin) of the problem is also presented. In Section 3, we provide a review of the previous results and announce the obtained algorithmic result. Basic definitions and statements that provide necessary elements to prove the properties of the proposed algorithm are presented in Section 4. Finally, in Section 5 we construct an approximation algorithm for solving the problem under study and prove that our algorithm is a fully polynomial-time approximation scheme (FPTAS) when the space dimension is fixed.



In: A. Kononov et al. (eds.): DOOR 2016, Vladivostok, Russia, published at http://ceur-ws.org

## 1 Problem formulation and its origin

Everywhere below $\mathbb{R}$ denotes the set of real numbers, $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^q$.

Problem under consideration has the following formulation (see [5], [6], [7]).

*Problem 1 (Finding a subsequence in a sequence). Given* a sequence $\mathcal{Y} = (y_1, \ldots, y_N)$ of points from $\mathbb{R}^q$ and positive integer numbers $T_{\min}$, $T_{\max}$ and $M > 1$. *Find* a subset $\mathcal{M} = \{n_1, \ldots, n_M\} \subseteq \mathcal{N} = \{1, \ldots, N\}$ of indexes of the sequence $\mathcal{Y}$ elements such that

$$F(\mathcal{M}) = \sum_{j \in \mathcal{M}} \|y_j - \overline{y}(\mathcal{M})\|^2 \to \min, \tag{1}$$

where $\overline{y}(\mathcal{M}) = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} y_i$ is a geometric center (centroid) of the subsequence $\{y_i \in \mathcal{Y} \,|\, i \in \mathcal{M}\}$ subject to constraints

$$1 \leq T_{\min} \leq n_m - n_{m-1} \leq T_{\max} \leq N, \ m = 2, \ldots, M, \tag{2}$$

on the elements of the tuple $(n_1, \ldots, n_M)$.

Problem 1 has the following interpretation (see [5], [6]). There is a time series containing $N$ measurements $y_1, \ldots, n_N$ of $q$ numerical characteristics of some objects. Each measurement result in the time series has an error, and no correspondence is known between the elements of the time series and the objects. Some of this objects have identical characteristics (or one can say that in the time series there are several measurements of one significant object). Other objects are distinguished and have different characteristics (or one can say that in the time series there are some measurements that are treated as "trash" which could be presented in this time series). The number of measurements for identical objects is known. In addition, it is known that the time interval between every two consequent results of measuring characteristics of the identical objects is bounded from above and below with some constants $T_{\max}$ and $T_{\min}$. The characteristics of identical objects in contrast to the characteristics of other objects have basic information value. It is required to find the subsequence of measures which corresponds to the identical objects using the criterion of minimum sum of squared distances and to estimate the characteristics of these objects (taking into account the measuring errors in the data).

## 2 Previous and obtained results

Problem 1 is among poorly studied discrete optimization problems. A special case of this problem when $T_{\min} = 1$ and $T_{\max} = N$ is equivalent [5] to the strongly NP-hard problem of searching points subset. In this case at the input there is a set of points instead of a sequence and time-related constraints (2) are absent.

First, let us recall the results obtained for the searching subset problem because it is a simple particular case of Problem 1.

In general, the searching subset problem is strongly NP-hard [8]. But in the case of fixed dimension $q$ of the space this problem could be solved [9] in $\mathcal{O}(N^{q+1})$ time.

Moreover, by this time for the searching subset problem the following algorithms have been presented. In [10] a 2-approximation polynomial algorithm is proposed, it's time complexity is $\mathcal{O}(qN^2)$. A polynomial-time approximation scheme (PTAS) is substantiated in [11]. This scheme allows to solve the searching subset problem for arbitrary relative error $\varepsilon$ in $\mathcal{O}(qN^{2/\varepsilon+1}(9/\varepsilon)^{3/\varepsilon})$ time. For the case of fixed dimension $q$ of the space and integer-valued points coordinates an exact pseudo-polynomial algorithm is constructed [12]. The running time of this algorithm is $\mathcal{O}(N(MD)^q)$, where $D$ is maximal absolute input point coordinate value. It is established [13], that the searching subset problem has no FPTAS, unless P=NP. In cited paper, an FPTAS was proposed for the case of fixed space dimension. This scheme allows to solve the problem with arbitrary relative error $\varepsilon$ in $\mathcal{O}(N^2(M/\varepsilon)^q)$ time.

By this time for the considered Problem 1 the following results were obtained. First, we should note that as far as the Problem 1 is generalization of strongly NP-hard subset problem, there are neither exact polynomial-time, nor pseudo-polynomial-time algorithms or FPTAS schemes, unless P=NP.

The case when $T_{\min}$ and $T_{\max}$ are parameters of Problem 1 is analyzed in [5]. In this work the authors showed that this problem is strongly NP-hard for any $T_{\min} < T_{\max}$. In the trivial case when $T_{\min} = T_{\max}$ this problem can be solved through a polynomial time.

In [6] a 2-approximation polynomial-time algorithm is proposed; the running time of the algorithm is $\mathcal{O}(N^2(MN+q))$. In the case of Problem 1 with integer components of the sequence elements and fixed dimension $q$ of the space in [7] an exact pseudo-polynomial-time algorithm is substantiated. This algorithm finds an optimal solution of Problem 1 in $\mathcal{O}(N^3(MD)^q)$ time.

Among the highest interest is the question of approximability of Problem 1. In particular, the question of FPTAS construction for the special case of Problem 1 (subclass of problem) has a great importance because in general case such a scheme does not exist. In the present work such scheme is presented for the case of fixed space dimension.

The main result of this work is an approximation algorithm which allows to find a $(1+\varepsilon)$–approximate solution for arbitrary relative error $\varepsilon > 0$ in $\mathcal{O}(N^2(M(T_{\max} - T_{\min} + 1) + q)(\sqrt{\frac{2q}{\varepsilon}} + 1)^q)$ time. If the dimension $q$ of the space is fixed then the time complexity of our algorithm is equal to $\mathcal{O}(MN^3(1/\varepsilon)^{q/2})$, and it implements an FPTAS.

## 3 Algorithm foundations

In order to substantiate our algorithm we'll need a few basic assertions and auxiliary problem with exact polynomial-time algorithm for this problem.

**Lemma 1.** *For an arbitrary point $x \in \mathbb{R}^q$ and a finite set $\mathcal{Z} \subset \mathbb{R}^q$ it is true that*

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

*where $\overline{z}$ is a centroid of set $\mathcal{Z}$.*

**Lemma 2.** *In assumptions of Lemma 1, if some point $u \in \mathbb{R}^q$ is closest (by distance) to the centroid $\overline{z}$ of set $\mathcal{Z}$ among all points from this set, then*

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

Both lemmas are well-known. Their proofs are presented in many publications (for example, in [12], [13]).

**Lemma 3.** *Let*

$$S(\mathcal{M}, x) = \sum_{n \in \mathcal{M}} \|y_n - x\|^2, \quad x \in \mathbb{R}^q, \ \mathcal{M} \subseteq \mathcal{N}, \tag{3}$$

*where $y_n \in \mathcal{Y}$, and $\mathcal{M} = \{n_1, \ldots, n_M\}$ satisfies the restrictions (2). Then for any fixed $\mathcal{M}$ the constrained minimum of $S(\mathcal{M}, x)$ over $x$ is reached at the point $x = \overline{y}(\mathcal{M})$ and is equal to $F(\mathcal{M})$.*

This assertion could be easily verified by differentiation of $S$ over $x$ and also follows from Lemma 1.

In addition to Lemma 3 for an arbitrary fixed point $x \in \mathbb{R}^q$ the restriction of the function $S(\mathcal{M}, x)$ to $\mathcal{M}$ we denote as $S^x(\mathcal{M})$ and argument of its minimum we denote as $\mathcal{M}^x$.

**Lemma 4.** *Let $\mathcal{M}^*$ be the optimal solution of Problem 1, and $\overline{y}(\mathcal{M}^*)$ be the centroid of the set $\{y_i | i \in \mathcal{M}^*\}$. Then for any point $x \in \mathbb{R}^q$ the following inequality holds*

$$F(\mathcal{M}^x) \leq F(\mathcal{M}^*) + M\|x - \overline{y}(\mathcal{M}^*)\|^2. \tag{4}$$

*Proof.* Let $\overline{y}(\mathcal{M}^x) = \frac{1}{|\mathcal{M}^x|} \sum_{i \in \mathcal{M}^x} y_i$ be a centroid of $\{y_i \,|\, i \in \mathcal{M}^x\}$. Then from definitions (1), (3) and Lemma 3 we obtain

$$F(\mathcal{M}^x) = \sum_{i \in \mathcal{M}^x} \|y_i - \overline{y}(\mathcal{M}^x)\|^2 \leq \sum_{i \in \mathcal{M}^x} \|y_i - x\|^2 = S^x(\mathcal{M}^x). \tag{5}$$

In addition, from the definition of the set $\mathcal{M}^x$ we have

$$S^x(\mathcal{M}^x) \leq S^x(\mathcal{M}^*). \tag{6}$$

Further, Lemma 1 applied up to the point $x$ and the set $\{y_i \,|\, i \in \mathcal{M}^*\}$ implies

$$S^x(\mathcal{M}^*) = \sum_{i \in \mathcal{M}^*} \|y_i - x\|^2 = \sum_{i \in \mathcal{M}^*} \|y_i - \overline{y}(\mathcal{M}^*)\|^2 + |\mathcal{M}^*| \cdot \|x - \overline{y}(\mathcal{M}^*)\|^2. \tag{7}$$

Finally, combining (5)–(7) we obtain

$$\begin{aligned} F(\mathcal{M}^x) \leq S^x(\mathcal{M}^x) &\leq S^x(\mathcal{M}^*) \\ &= \sum_{i \in \mathcal{M}^*} \|y_i - \overline{y}(\mathcal{M}^*)\|^2 + |\mathcal{M}^*| \cdot \|x - \overline{y}(\mathcal{M}^*)\|^2 \\ &= F(\mathcal{M}^*) + M\|x - \overline{y}(\mathcal{M}^*)\|^2. \end{aligned}$$

□

Lemma 4 shows that quality of feasible solution $\mathcal{M}^x$ obtained by some point $x \in \mathbb{R}^q$ could be estimated via distance from this point to (unknown) optimal centroid $\overline{y}(\mathcal{M}^*)$. The closer considered point $x$ up to the optimal centroid, the less absolute approximation error of an obtained solution.

**Lemma 5.** *Let $\mathcal{M}^*$ be the optimal solution of Problem 1, and let*

$$t = \arg \min_{y \in \{y_i \,|\, i \in \mathcal{M}^*\}} \|y - \overline{y}(\mathcal{M}^*)\|$$

*be the point of the optimal set $\{y_i | i \in \mathcal{M}^*\}$ closest to its centroid, then*

$$\|t - \overline{y}(\mathcal{M}^*)\|^2 \leq \frac{1}{M} F(\mathcal{M}^t), \tag{8}$$

*where $\mathcal{M}^t$ supplies minimum to $S^t(\mathcal{M})$ over $\mathcal{M} \subseteq \mathcal{N}$ under constraints* (2) *on elements of $\mathcal{M}$.*

*Proof.* From the definition of the point $t$ it follows that

$$\|t - \overline{y}(\mathcal{M}^*)\|^2 \leq \|y_i - \overline{y}(\mathcal{M}^*)\|^2$$

for any $i \in \mathcal{M}^*$. Summing up both sides of this inequality for all $i \in \mathcal{M}^*$ we obtain

$$M\|t - \overline{y}(\mathcal{M}^*)\|^2 \leq \sum_{i \in \mathcal{M}^*} \|y_i - \overline{y}(\mathcal{M}^*)\|^2. \tag{9}$$

From the fact that $\mathcal{M}^t$ is a feasible solution of Problem 1 and $\mathcal{M}^*$ is the optimal solution we have the following bound

$$F(\mathcal{M}^*) \leq F(\mathcal{M}^t). \tag{10}$$

Combining (9) and (10) we obtain final estimation

$$M\|t - \overline{y}(\mathcal{M}^*)\|^2 \leq \sum_{i \in \mathcal{M}^*} \|y_i - \overline{y}(\mathcal{M}^*)\|^2 = F(\mathcal{M}^*) \leq F(\mathcal{M}^t).$$

□

**Lemma 6.** *Assume that conditions of Lemma 5 are held. Then for $\mathcal{M}^x$ to be a $(1+\varepsilon)$–approximate solution of the Problem 1 for fixed $\varepsilon > 0$, it is enough to take such point $x$ that satisfies the inequality*

$$\|x - \overline{y}(\mathcal{M}^*)\|^2 \leq \frac{\varepsilon}{4M} F(\mathcal{M}^t). \tag{11}$$

*Proof.* Let $\overline{y}(\mathcal{M}^t) = \frac{1}{|\mathcal{M}^t|} \sum_{i \in \mathcal{M}^t} y_i$ be a centroid of the set $\{y_i \,|\, i \in \mathcal{M}^t\}$. From the fact that $\mathcal{M}^t = \arg\min_{\mathcal{M}} S^t(\mathcal{M})$ and definitions (1), (3) we have

$$F(\mathcal{M}^t) \leq S^t(\mathcal{M}^t). \tag{12}$$

In addition, from the optimality of the set $\mathcal{M}^t$ we have inequality

$$S^t(\mathcal{M}^t) \le S^t(\mathcal{M}^*). \tag{13}$$

Further, since the set $\{y_i \,|\, i \in \mathcal{M}^*\}$ and the point $t$ satisfy the conditions of Lemma 2, the following estimate holds

$$\sum_{i \in \mathcal{M}^*} \|y_i - t\|^2 \le 2 \sum_{i \in \mathcal{M}^*} \|y_i - \overline{y}(\mathcal{M}^*)\|^2$$

and, therefore,

$$S^t(\mathcal{M}^*) = \sum_{i \in \mathcal{M}^*} \|y_i - t\|^2 \le 2 \sum_{i \in \mathcal{M}^*} \|y_i - \overline{y}(\mathcal{M}^*)\|^2 = 2F(\mathcal{M}^*). \tag{14}$$

Combining (11)–(14) we obtain

$$\|x - \overline{y}(\mathcal{M}^*)\|^2 \le \frac{\varepsilon}{2M} F(\mathcal{M}^t) \le \frac{\varepsilon}{2M} S^t(\mathcal{M}^t) \le \frac{\varepsilon}{2M} S^t(\mathcal{M}^*) \le \frac{\varepsilon}{M} F(\mathcal{M}^*). \tag{15}$$

Finally, applying (15) up to the right side of inequality (4) we obtain inequality

$$F(\mathcal{M}^x) \le (1 + \varepsilon) F(\mathcal{M}^*),$$

that completes the proof of Lemma 6. □

Computational base of the proposed algorithm is an exact polynomial-time algorithm for solving the auxiliary problem. In this auxiliary problem for any point $x \in \mathbb{R}^q$ one needs to find such a set $\mathcal{M}^x$ that supplies minimum to $S^x(\mathcal{M})$ while elements from $\mathcal{M}$ are under constraints (2).

This auxiliary problem can be formulated as follows:

*Problem 2. Given* a sequence $\mathcal{Y} = (y_1, \ldots, y_N)$ of points from $\mathbb{R}^q$, point $x \in \mathbb{R}^q$, positive integer numbers $T_{\min}$, $T_{\max}$ and $M > 1$. *Find* a subset $\mathcal{M} = \{n_1, \ldots, n_M\} \subseteq \mathcal{N}$ of indexes of the sequence elements such that

$$S^x(\mathcal{M}) = \sum_{i \in \mathcal{M}} \|y_i - x\|^2 \to \min,$$

while elements of the tuple $(n_1, \ldots, n_M)$ satisfy the constraints (2).

A dynamic programming scheme is presented in the next lemma and its corollary. This scheme allows to find the optimal solution $\mathcal{M}^x$ of Problem 2. The presented scheme is based on results from [6], [14] and given here for completeness.

**Lemma 7.** *For any positive integer $M > 1$, such that $(M-1)T_{\min} \le N - 1$, and for arbitrary point $x \in \mathbb{R}^q$ the optimum $S^x_{\min} = \min_{\mathcal{M}} S^x(\mathcal{M})$ of Problem 2 could be found as*

$$S^x_{\min} = \min_{n \in \omega_M} S^x_M(n), \tag{16}$$

*where the values of the functions $S_M^x(n)$, $n \in \omega_M$, are calculated using the following recurrence formulas*

$$S_m^x(n) = \begin{cases} \|y_n - x\|^2, & \text{if } n \in \omega_1,\ m = 1; \\ \|y_n - x\|^2 + \max\limits_{j \in \gamma_{m-1}^-(n)} S_{m-1}^x(j), & \text{if } n \in \omega_m,\ m = 2, \ldots, M, \end{cases} \tag{17}$$

*where*

$$\omega_m = \big\{n \,|\, 1 + (m-1)T_{\min} \le n \le N - (M-m)T_{\min}\big\}, m = 1, \ldots, M,$$

$$\gamma_{m-1}^-(n) = \big\{j \mid \max\{1 + (m-2)T_{\min}, n - T_{\max}\} \le j \le n - T_{\min}\big\},$$
$$n \in \omega_m,\ m = 2, \ldots, M.$$

**Corollary 1.** *Elements $n_1^x, \ldots, n_M^x$ of the optimal tuple $\mathcal{M}^x$ could be found by the formulas:*

$$n_M^x = \arg \min_{n \in \omega_M} S_M^x(n), \tag{18}$$

$$n_{m-1}^x = \arg \min_{n \in \gamma_m^-(n_m^x)} S_m^x(n), \quad m = M, M-1, \ldots, 2. \tag{19}$$

Let us show the algorithm implementing above scheme in a step-by-step description.

*Algorithm* $\mathcal{A}_1$.

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

`Step 1`. Compute the values $\|y_n - x\|^2$ for each $n \in \mathcal{N}$.

`Step 2`. Using formulas (17), calculate the values $S_m^x(n)$ for each $n \in \omega_m$ while $m = 1, \ldots, M$.

`Step 3`. Find the minimum $S_{\min}^x$ of the objective function $S^x$ using (16) and the optimal tuple $\mathcal{M}^x = (n_1^x, \ldots, n_M^x)$ by formulas (18), (19).

*Output*: tuple $\mathcal{M}^x = (n_1^x, \ldots, n_M^x)$.

In [6], it was proved that the algorithm $\mathcal{A}_1$ finds an optimal solution of Problem 2 in $\mathcal{O}(N(M(T_{\max} - T_{\min} + 1) + q))$ time. In this expression, the value $T_{\max} - T_{\min} + 1$ is at most $N$. Therefore, the running time of the algorithm is estimated as $\mathcal{O}(N(MN + q))$.

## 4 Approximation algorithm

Proposed approximation algorithm for Problem 1 can be schematically described as follows. For all points from the input sequence we define a special box (cube with center in this point) such that at least one of these boxes contains unknown centroid of the optimal subsequence. By the given value of the relative error we construct multidimensional grid with uniform step by all coordinates. For all nodes of the grid we solve the auxiliary problem using dynamic programming scheme, and obtain a feasible solution of Problem 1. Finally, among all obtained feasible solutions we select one with minimal value of the objective function of Problem 1.

For an arbitrary point $z \in \mathbb{R}^q$ and positive numbers $h$ and $H$ we define a set of points

$$\mathcal{D}(z, h, H) = \{d \mid d = z + h(j_1, \ldots, j_q),\ j_i \in \mathbb{Z},\ |h \cdot j_i| \le H,\ i = 1, \ldots, q\}.$$

Note that points from this set are placed in the nodes of the uniform multidimensional rectangular grid with size $2H$ and step $h$ between nodes. The center of this grid is at the point $z$.

For the number of nodes of this grid the following bound holds

$$|\mathcal{D}(z,h,H)| \le (2\frac{H}{h}+1)^q.$$

Herewith, for any point $x \in \mathbb{R}^q$, such that $\|z-x\| \le H$, the distance up to the closest node of the grid $\mathcal{D}(z,h,H)$ does not exceed $\frac{h\sqrt{q}}{2}$.

Lemma 5 (right-hand side of inequality (8)), in fact, defines the size of the cube which guaranteed contains the unknown centroid of the optimal subsequence. So the size of the grid can be defined as follows:

$$H(y) = \sqrt{\frac{1}{M}F(\mathcal{M}^y)}, \quad y \in \mathcal{Y}. \tag{20}$$

Moreover, Lemma 6 defines the condition on the value of the grid spacing, wherein there is a node closed enough to the centroid of the optimal subsequence (in sense of guaranteed error $\varepsilon$). Therefore, define the grid spacing as follows:

$$h(y,\varepsilon) = \sqrt{\frac{2\varepsilon}{qM}F(\mathcal{M}^y)}, \quad y \in \mathcal{Y},\ \varepsilon > 0. \tag{21}$$

Let us present the algorithm for solving Problem 1.

*Algorithm* $\mathcal{A}$.

*Input*: sequence $\mathcal{Y}$ of points, numbers $T_{\min}$, $T_{\max}$, $M$ and $\varepsilon$.

For all points $y \in \mathcal{Y}$ execute the steps 1–5.

`Step 1`. Using Algorithm $\mathcal{A}_1$ find the optimal solution $\mathcal{M}^y$ of Problem 2 with $x = y$.

`Step 2`. Compute $F(\mathcal{M}^y)$, $h$ and $H$ by the formulas (1), (21) (20).

`Step 3`. If $F(\mathcal{M}^y) = 0$, then set $\mathcal{M}^y$ is declared as a result of the algorithm; Exit. Otherwise go to the next step.

`Step 4`. Build the grid $\mathcal{D}(y,h,H)$.

`Step 5`. For all points $d$ from the grid $\mathcal{D}(y,h,H)$ using Algorithm $\mathcal{A}_1$ find the optimal solution $\mathcal{M}^d$ of Problem 2 (with $x = d$) and compute the value $F(\mathcal{M}^d)$.

`Step 6`. In the set of solutions $\{\mathcal{M}^d \mid d \in \mathcal{D}(y,h,H),\ y \in \mathcal{Y}\}$ as a result select the tuple $\mathcal{M}^{d^*}$ supplying the minimal value for the objective function $F(\mathcal{M}^d)$.

*Output*: tuple $\mathcal{M}^{d^*}$.

**Theorem 1.** *For an arbitrary $\varepsilon > 0$ Algorithm $\mathcal{A}$ finds a $(1+\varepsilon)$–approximate solution of Problem 1 in $\mathcal{O}(N^2(M(T_{\max}-T_{\min}+1)+q)(\sqrt{\frac{2q}{\varepsilon}}+1)^q)$ time.*

*Proof.* Let $t = \arg\min_{y \in \{y_i \mid i \in \mathcal{M}^*\}} \|y - \overline{y}(\mathcal{M}^*)\|$ be the point from the set $\{y_i \mid i \in \mathcal{M}^*\}$, that is closest to the centroid of this set. If for this point on step 3 we have equality $F(\mathcal{M}^t) = 0$, then the set $\mathcal{M}^t$ is an optimal solution of Problem 1, because the value of the objective function $F$ is always nonnegative.

Now we consider the second case, when $F(\mathcal{M}^t) > 0$.

In accordance with Lemma 5, there is an inequality (8) for the point $t$. This inequality and definition (20) of the $H(\cdot)$ implies $\|t - \overline{y}(\mathcal{M}^*)\| \leq H$. In other words, centroid $\overline{y}(\mathcal{M}^*)$ of the optimal subsequence is inside area bounded by the grid $\mathcal{D}(t, h, H)$.

Take the point

$$d^* = \arg \min_{d \in \mathcal{D}(t,h,H)} \|d - \overline{y}(\mathcal{M}^*)\|$$

from the grid, that is closest to the optimal centroid. Since distance from $\overline{y}(\mathcal{M}^*)$ to the closest node $d^*$ from the grid $\mathcal{D}(t, h, H)$ does not exceed $\frac{h\sqrt{q}}{2}$, we obtain the estimate

$$\|d^* - \overline{y}(\mathcal{M}^*)\|^2 \leq \frac{h^2 q}{4} \frac{\varepsilon}{2M} F(\mathcal{M}^t).$$

Thus, the point $d^*$ satisfies the conditions of Lemma 6. Therefore, the set $\mathcal{M}^{d^*}$ is a $(1+\varepsilon)$–approximate solution of Problem 1.

Note, that either on step 3 optimal solution will be found or the point $d^*$ and the set $\mathcal{M}^{d^*}$ will be considered during operation of the algorithm in step 5. Therefore, at the end the algorithm is guaranteed to obtain at least $(1+\varepsilon)$–approximate solution.

Let us now estimate the time complexity of the algorithm. On step 1 to solve auxiliary problem $\mathcal{O}(N(M(T_{\max} - T_{\min} + 1) + q))$ operations is required. Step 2 requires $\mathcal{O}(qN)$ operations, and step 3 could be performed in $\mathcal{O}(1)$ time.

To generate the grid $\mathcal{D}(y, h, H)$ on step 4 it needs to perform $\mathcal{O}(q \cdot |\mathcal{D}(y, h, H)|)$ operations. The cost of constructing the sets $\mathcal{M}^d$ on step 5 and computing the values $F(\mathcal{M}^d)$ is equal to (as on step 1) $\mathcal{O}(N(M(T_{\max} - T_{\min} + 1) + q))$.

As a result, for each of $N$ points from the sequence $\mathcal{Y}$ performing steps 1–5 required $\mathcal{O}(N(M(T_{\max} - T_{\min} + 1) + q) \cdot |\mathcal{D}(y, h, H)|)$ operations. Finally, on step 6 to choose the best solution it needs to perform $\mathcal{O}(\sum_{y \in \mathcal{Y}} |\mathcal{D}(y, h, H)|)$ operations.

It remains to note that the size of the grid $\mathcal{D}(y, h, H)$ is bounded by

$$|\mathcal{D}(y, h, H)| \leq (2\frac{H}{h} + 1)^q \leq (\sqrt{\frac{2q}{\varepsilon}} + 1)^q.$$

Therefore, the total time complexity of the entire algorithm is $\mathcal{O}(N^2(M(T_{\max} - T_{\min} + 1) + q)(\sqrt{\frac{2q}{\varepsilon}} + 1)^q)$. □

Let us show that if the dimension $q$ of the space is fixed, then presented algorithm implements an FPTAS. Indeed, if $\varepsilon \in (0,\, 2q]$, then one of the factors in final complexity can be estimated as follows

$$\left(\sqrt{\frac{2q}{\varepsilon}} + 1\right)^q \leq 2^q \left(\sqrt{\frac{2q}{\varepsilon}}\right)^q = 2^{\frac{3q}{2}} q^{\frac{q}{2}} \left(\frac{1}{\varepsilon}\right)^{\frac{q}{2}} = \mathcal{O}\left(\left(\frac{1}{\varepsilon}\right)^{\frac{q}{2}}\right).$$

Therefore, as the value $T_{\max} - T_{\min} + 1$ does not exceed $N$, the time complexity of the algorithm is $\mathcal{O}(MN^3(1/\varepsilon)^{q/2})$. Thus, the proposed algorithm implements an FPTAS.

## 5 Conclusion

In this work we have constructed an approximation algorithm for one problem of finding a subsequence in the given sequence of points from the Euclidean space. The proposed algorithm implements a fully polynomial-time approximation scheme in the case when the dimension of the space is fixed (is not a parameter of the input).

**Acknowledgments.** This work was supported by the RFBR, projects 15-01-00462, 16-31-00186 and 16-07-00168.

## References

1. Fu, Tak-chung: A Review on Time Series Data Mining. Engineering Applications of Artificial Intelligence. 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 International Publishing Switzerland (2015)
3. T. Warren Liao: Clustering of Time Series Data — a Survey. Pattern Recognition. 38(11), 1857–1874 (2005)
4. Aggarwal C. C. Data Mining: The Textbook. Springer International Publishing (2015)
5. Kel'manov A.V., Pyatkin A.V.: On Complexity of Some Problems of Cluster Analysis of Vector Sequences. J. Appl. Indust. Math. 7(3) 363–369 (2013)
6. Kel'manov A.V., Romanchenko S.M., Khamidullin S.A.: Approximation algorithms for some intractable problems of choosing a vector subsequence. J. Appl. Indust. Math. 6(4) 443–450 (2012)
7. Kel'manov A.V., Romanchenko S.M., Khamidullin S.A.: Exact pseudopolynomial algorithms for some NP-hard problems of searching a vectors subsequence. Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki (in Russian). 53(1) 143–153 (2013)
8. Kel'manov A.V., Pyatkin A.V.: NP-Completeness of Some Problems of Choosing a Vector Subset. J. Appl. Indust. Math. 5(3) 352–357 (2011)
9. Aggarwal A., Imai H., Katoh N., Suri S.: Finding k points with minimum diameter and related problems. J. Algorithms. Vol. 12. P. 38–56 (1991)
10. Kel'manov A.V., Romanchenko S.M.: An Approximation Algorithm for Solving a Problem of Search for a Vector Subset. J. Appl. Indust. Math. 6(1), 90–96 (2012)
11. Shenmaier V.V.: An approximation scheme for a problem of search for a vector subset. J. Appl. Indust. Math. 6(3), 381–386 (2012)
12. Kel'manov A.V., Romanchenko S.M.: Pseudopolynomial Algorithms for Certain Computationally Hard Vector Subset and Cluster Analysis Problems. Automation and Remote Control. 73(2), 349–354 (2012)
13. Kel'manov A.V., Romanchenko S.M.: An FPTAS for a Vector Subset Search Problem. J. Appl. Indust. Math. 8(3), 329–336 (2014)
14. Kel'manov A.V., Khamidullin S.A.: Posterior Detection of a Given Number of Identical Subsequences in a Quasi-periodic Sequence. Comput. Math. Math. Phys. 41(5) 762-774 (2001)