Easy NP-hardness Proofs of Some Subset Choice Problems

Artem V. Pyatkin[1,2(✉)]

[1] Sobolev Institute of Mathematics, Koptyug Avenue, 4, Novosibirsk 630090, Russia
[2] Novosibirsk State University, Pirogova Street, 2, Novosibirsk 630090, Russia
artempyatkin@gmail.com

Abstract. We consider the following subset choice problems: given a family of Euclidean vectors, find a subset having the largest a) norm of the sum of its elements; b) square of the norm of the sum of its elements divided by the cardinality of the subset. The NP-hardness of these problems was proved in two papers about ten years ago by reduction of 3-SAT problem. However, that proofs were very tedious and hard to read. In the current paper much easier and natural proofs are presented.

Keywords: Euclidean space · Subset choice · Clustering · 2-partition · Strong np-hardness

1 Introduction

This paper deals with well-known vector subset choice problems that are induced by data analysis and pattern recognition problems. A typical problem in data analysis requires finding in a set of data a subset of the most similar elements where the similarity is defined according to some criterion. The cardinality of the sought subset could be known or unknown in advance. One of the possible criteria is minimum of the sum of squared deviations. This criterion arises, in particular, in a noise-proof data analysis where the aim is to detect informationally significant fragments in noisy datasets, to estimate them, and to classify them afterwards [8,12]. The problem of finding a subset of vectors with the longest sum has applications in the pattern recognition (finding a correct direction to a certain object) [25].

Although these problems are known to be NP-hard both in the case of known (given as a part of input) cardinality of a sought subset [3,8] and in the case of unknown one [14,15,22], the latter proofs are much more complicated and hard to read (see the discussion in the next section). In this paper we suggest much more easy and natural NP-hardness proofs for the case of unknown size of the

The research was supported by the program of fundamental scientific researches of the SB RAS, project 0314-2019-0014, by the Russian Foundation for Basic Research, project 19-01-00308, and by the Top-5-100 Program of the Ministry of Education and Science of the Russian Federation.

Y. Kochetov et al. (Eds.): MOTOR 2020, CCIS 1275, pp. 70–79, 2020.
https://doi.org/10.1007/978-3-030-58657-7_8

sought set. We believe that the new proofs can be helpful for analyzing related problems with the unknown cardinalities of the sought subset.

The paper is organized as follows. In the next section the mathematical formulation of the problems are given and the motivation of the research and some related results are discussed. In Sect. 3 the main results of the paper are presented. Section 4 concludes the paper.

2 Problem Formulation, Motivation and Related Results

The problem of noise-proof data analysis in noisy data sets [8,12,15] is as follows. Each record of the data is a vector representing a set of measured characteristics of an object transmitted via a noisy channel. The object can be either in an active or in a passive state. In the passive state all characteristics are 0, while in the active state all measured characteristics are stabile and at least one of them must be non-zero. The noise has a d-dimensional normal distribution with zero mean and an arbitrary dispersion. The goal is to determine the moments when the object was in the active state and to evaluate the measured characteristics.

As it was shown in [8,12,15], this problem can be reduced to the following optimization problem.

Problem 1. Given a set of vectors $\mathcal{Y} = \{y_1, \ldots, y_N\}$ in d-dimensional Euclidean space, find a non-empty subset $\mathcal{C} \subseteq \mathcal{Y}$ maximizing

$$h(\mathcal{C}) := \frac{\| \sum_{x \in \mathcal{C}} x \|^2}{|\mathcal{C}|}.$$

Everywhere in the paper the norm is Euclidean, unless otherwise stated. A version of Problem 1 with an additional restriction on the cardinality of the sought set $\mathcal{C}$ is referred to as

Problem 2. Given a set of vectors $\mathcal{Y} = \{y_1, \ldots, y_N\}$ in d-dimensional Euclidean space and a positive integer M, find a subset $\mathcal{C} \subseteq \mathcal{Y}$ of cardinality M maximizing $h(\mathcal{C})$.

The following two subset choice problems are very close in formulation to these ones.

Problem 3. Given a set of vectors $\mathcal{Y} = \{y_1, \ldots, y_N\}$ in d-dimensional Euclidean space, find a non-empty subset $\mathcal{C} \subseteq \mathcal{Y}$ maximizing $\| \sum_{x \in \mathcal{C}} x \|$.

Problem 4. Given a set of vectors $\mathcal{Y} = \{y_1, \ldots, y_N\}$ in d-dimensional Euclidean space and a positive integer M, find a subset $\mathcal{C} \subseteq \mathcal{Y}$ of cardinality M minimizing $\sum_{x \in \mathcal{C}} \|x - \overline{x}\|^2$ where $\overline{x} = (\sum_{x \in \mathcal{C}} x)/|\mathcal{C}|$ is the centroid of the set $\mathcal{C}$.

Note that the variant of Problem 3 with a given cardinality of the subset $\mathcal{C}$ is equivalent to Problem 2, while the variant of Problem 4 without the restriction on the cardinality of $\mathcal{C}$ is trivial (every subset of cadrinality 1 is an optimal solution).

Problem 3 has a following interpretation [25]. Each vector is a measurement result of a direction to some interesting object. Each measurement result has an additive error having a normal distribution, and there are some redundant vectors in the set (related to other objects or reflections). The goal is to delete the redundant vectors and find the correct direction. This can be done by finding a subset of vectors having the longest sum.

If the dimension of the space d is fixed then all Problems 1–4 are polynomially solvable. Namely, Problems 1 and 2 can be solved [9] in time $O(dN^{2d+2})$; Problem 3 is a particular case of Shaped Partition problem [11], which yields an $O(N^d)$ algorithm for it; a better algorithm of complexity $O(dN^{d-1}\log N)$ is presented in [25]. Problem 4 can be solved [1] in time $O(dN^{d+1})$. The universal algorithm solving Problems 1–4 in time $O(dN^{d+1})$ using Voronoi diagrams can be found in [24]. Note that this algorithm indeed can solve any vector subset choice problem satisfying one of the following two locality properties:

- For every input there is a point x^* such that the optimal solution consists of the set of M closest to x^* points of $\mathcal{Y}$.
- For every input there is a vector y^* such that the optimal solution consists of the set of M vectors of $\mathcal{Y}$ having minimum scalar products with y^*.

If the dimension of the space d is a part of input then all four problems mentioned above are NP-hard in a strong sense. Moreover, for Problems 2 and 3 an inapproximability bound $(16/17)^{1/p}$ was proved in [26] for an arbitrary norm l_p where $p \in [1, \infty)$.

There are a lot of approximation results for these problems. Let us mention randomized algorithms finding $(1+\varepsilon)$-approximate solution for Problems 2 and 3 of complexity $O(d^{3/2}N\log\log N/(2\varepsilon-\varepsilon^2)^{(d-1)/2})$ in [10] and of complexity $O(d^{O(1)}N(1+2/\varepsilon)^d)$ with probability $1-1/e$ in [26]. For Problem 4 a $(1+\varepsilon)$-approximation algorithm of complexity $O(N^2(M/\varepsilon)^d)$ was suggested in [19] and a PTAS of complexity $O(dN^{1+2/\varepsilon}(9/\varepsilon)^{3/\varepsilon})$ was constructed in [23]. For Problem 1 a $(1+\varepsilon)$-approximation algorithm of complexity $O(Nd(d+\log N)(\sqrt{(d-1)/\varepsilon}+1)^{d-1})$ can be found in [15].

The NP-hardness of Problem 2 (i. e. in case of known—given as a part of input—cardinality of a sought subset) was proved in [3,8]. The proof uses a natural reduction from the classical NP-hard Clique problem. In this reduction, each vector corresponds to a vertex of a graph and a subset $\mathcal{C}$ is optimal if and only if the corresponding subset of vertices induces a clique in the graph. This proof is so natural that the similar idea was used later, in particular, for proving NP-hardness of Problem 4 in [16], of Maximum Diversity problem in [5] and of 1-Mean and 1-Median 2-Clustering Problem in [18].

The NP-hardness of Problem 1 was proved in [14,15]. It uses quite complicated reduction of 3-SAT problem, where several vectors correspond to each clause and to each variable, and some irrational numbers (square roots) are used in their coordinates (and thus, additional arguments justifying the possibility of rational approximation become necessary). The NP-hardness of Problem 3 was proved in [22] also by reduction of 3-SAT; although there are no irrational numbers, the reduction still remains complicated and the proof is hard to follow.

These reductions are highly inconvenient and hard to generalize. So, many other vector choice or clustering problems with unknown cardinality of the sought set stay open (see, for example, [18]). In this paper we present an easy and natural NP-hardness proof for Problems 1 and 3 with almost the same reduction of Exact Cover by 3-Sets problem.

Let us mention some other problems that are related to Problems 1–4. Make use of the following well-known folklore identities (the proofs can be found, for instance, in [15,17]):

$$\sum_{y\in\mathcal{Y}}\|y\|^2-\frac{\|\sum_{x\in\mathcal{C}}x\|^2}{|\mathcal{C}|}=\sum_{y\in\mathcal{C}}\|y-\overline{y}\|^2+\sum_{y\in\mathcal{Y}\setminus\mathcal{C}}\|y\|^2$$
$$=\frac{\sum_{y\in\mathcal{C}}\sum_{z\in\mathcal{C}}\|y-z\|^2}{2|\mathcal{C}|}+\sum_{y\in\mathcal{Y}\setminus\mathcal{C}}\|y\|^2. \tag{1}$$

Since the sum of the squared norms of all vectors from $\mathcal{Y}$ does not depend on $\mathcal{C}$, Problems 1 and 2 are equivalent to minimization of the function

$$\sum_{y\in\mathcal{C}}\|y-\overline{y}\|^2+\sum_{y\in\mathcal{Y}\setminus\mathcal{C}}\|y\|^2,$$

that can be treated as a minimum sum of squares 2-clustering where the center of one cluster is known. This problem is very close to a classical MSSC (minimum sum of squares clustering) problem also known as k-means [2,6,20,21], but not equivalent to it. Note that in such equivalent formulations these problems admit polynomial 2-approximation algorithms of complexity $O(dN^2)$ both for known [4] and unknown [13] cardinality of the sought set (cluster with an unknown center). As far as we know, no polynomial approximation algorithm with a guaranteed exactness bound is known for Problem 1.

3 Main Results

In this section we present the new NP-hardness proofs for Problems 1 and 3.

3.1 NP-hardness of Problem 1

Let us rewrite Problem 1 in the equivalent (due to (1)) form of the decision problem.

Problem 5. Given a set of vectors $\mathcal{Y}=\{y_1,\dots,y_N\}$ in d-dimensional Euclidean space and a number $K>0$, is there a non-empty subset $\mathcal{C}\subseteq\mathcal{Y}$ such that

$$f(\mathcal{C}):=\frac{1}{2|\mathcal{C}|}\sum_{x\in\mathcal{C}}\sum_{y\in\mathcal{C}}\|x-y\|^2+\sum_{z\in\mathcal{Y}\setminus\mathcal{C}}\|z\|^2\le K?$$

We need the following well-known NP-hard [7] version of the Exact Cover by 3-Sets problem where each element lies in at most 3 subsets.

Problem 6 (X3C3). Given a family $E = \{e_1, \dots, e_m\}$ of 3-element subsets of the set $V = \{v_1, \dots, v_n\}$ where $n = 3q$ such that every $v \in V$ meets in at most 3 subsets from E, find out whether there exist a subfamily $E_0 = \{e_{i_1}, \dots, e_{i_q}\}$ covering the set V, i. e. such that $V = \cup_{j=1}^{q} e_{i_j}$.

The main result of this subsection is the following theorem.

Theorem 1. *Problem 1 is NP-hard in a strong sense.*

Proof. Consider an arbitrary instance of X3C3 problem and reduce it to an instance of Problem 5 in the following way. Put $N = m, d = 3n + 1$ and $K = 18a^2(m-1) + m - q$ where a is a positive integer such that $a^2 > m(m-q)/6$. Each vector $y_i \in \mathcal{Y}$ corresponds to a set $e_i \in E$. For every $i \in \{1, \dots, n\}$ refer to the coordinates $3i,\ 3i-1,\ 3i-2$ of a vector $y \in \mathcal{Y}$ as *i-th coordinate triple*. Denote by $y_i(j)$ the j-th coordinate of y_i. If $v_i \notin e_j$ then the i-th triple of the vector y_j contains zeroes: $y_j(3i-2) = y_j(3i-1) = y_j(3i) = 0$. Otherwise, let $k = |\{l < j \mid v_i \in e_l\}|$ be the number of subsets from E with lesser indices than j containing the element v_i. Since each v_i lies in at most 3 subsets from E, we have $k \in \{0, 1, 2\}$. Put

$$\begin{gathered} y_j(3i-2) = 2a,\ y_j(3i-1) = y_j(3i) = -a,\ \text{if } k = 0; \\ y_j(3i-1) = 2a,\ y_j(3i-2) = y_j(3i) = -a,\ \text{if } k = 1; \\ y_j(3i) = 2a,\ y_j(3i-2) = y_j(3i-1) = -a,\ \text{if } k = 2. \end{gathered}$$

Also, put $y_j(3n+1) = 1$ for all $j \in \{1, \dots, m\}$.

For example, if $E = \{(v_1, v_2, v_3), (v_1, v_3, v_4), (v_1, v_5, v_6), (v_2, v_3, v_5), (v_4, v_5, v_6)\}$ then the family $\mathcal{Y}$ contains the following five vectors of dimension 19:

$$y_1 = (2a, -a, -a \mid 2a, -a, -a \mid 2a, -a, -a \mid 0, 0, 0 \mid 0, 0, 0 \mid 0, 0, 0 \mid 1);$$

$$y_2 = (-a, 2a, -a \mid 0, 0, 0 \mid -a, 2a, -a \mid 2a, -a, -a \mid 0, 0, 0 \mid 0, 0, 0 \mid 1);$$

$$y_3 = (-a, -a, 2a \mid 0, 0, 0 \mid 0, 0, 0 \mid 0, 0, 0 \mid 2a, -a, -a \mid 2a, -a, -a \mid 1);$$

$$y_4 = (0, 0, 0 \mid -a, 2a, -a \mid -a, -a, 2a \mid 0, 0, 0 \mid -a, 2a, -a \mid 0, 0, 0 \mid 1);$$

$$y_5 = (0, 0, 0 \mid 0, 0, 0 \mid 0, 0, 0 \mid -a, 2a, -a \mid -a, -a, 2a \mid -a, 2a, -a \mid 1).$$

For the convenience, different coordinate triples are separated by the vertical lines.

Note that $\|y_i\|^2 = 18a^2 + 1$ for all i and also

$$\|y_i - y_j\|^2 = \begin{cases} 36a^2, & \text{if } e_i \cap e_j = \emptyset; \\ 42a^2, & \text{if } |e_i \cap e_j| = 1; \\ 48a^2, & \text{if } |e_i \cap e_j| = 2; \\ 54a^2, & \text{if } |e_i \cap e_j| = 3 \end{cases}$$

for every $i \neq j$.

Assume first that an exact cover E_0 exists. Put $\mathcal{C} = \{y_j \mid e_j \in E_0\}$. Then

$$f(\mathcal{C}) = \frac{q(q-1)36a^2}{2q} + (m-q)(18a^2+1) = 18a^2(m-1) + m - q = K,$$

as required.

Assume now that there is a subset $\mathcal{C}$ of size $t > 0$ such that $f(\mathcal{C}) \le K$. Note that each coordinate triple can be non-zero in at most 3 vectors from $\mathcal{C}$. For each $k \in \{0,1,2,3\}$ denote by a_k the number of coordinate triples that are non-zero in exactly k vectors from $\mathcal{C}$ and estimate the contributions of such triples into the first addend of $f(\mathcal{C})$. Note that $a_0 + a_1 + a_2 + a_3 = n = 3q$. Clearly, the contribution of a_0 zero triples is 0. If a triple is non-zero in one vector from $\mathcal{C}$ then it contributes

$$\frac{(t-1)(4a^2+a^2+a^2)}{t},$$

and the total contribution of such triples is

$$\frac{6a^2a_1(t-1)}{t}. \tag{2}$$

If a triple is non-zero in two vectors from $\mathcal{C}$ then it contributes

$$\frac{2(t-2)(4a^2+a^2+a^2)+(9a^2+9a^2)}{t};$$

so, the total contribution of such triples is

$$\frac{6a^2a_2(2t-1)}{t}. \tag{3}$$

Finally, the total contribution of triples that are non-zero in three vectors from $\mathcal{C}$ is

$$\frac{(3(t-3)6a^2+3\cdot 18a^2)a_3}{t} = 18a^2a_3. \tag{4}$$

Since $|e_j| = 3$ for all j, we have $a_1 + 2a_2 + 3a_3 = 3t$. Using (2)–(4), estimate the objctive function

$$f(\mathcal{C}) = \frac{6a^2}{t}((t-1)a_1 + (2t-1)a_2 + 3ta_3) + (m-t)(18a^2+1)$$

$$= \frac{6a^2}{t}(3t^2 - a_1 - a_2) + (m-t)(18a^2+1) = 18ma^2 + m - t - \frac{6a^2}{t}(a_1+a_2)$$

$$= K + 18a^2 - \frac{6a^2}{t}(a_1+a_2) + q - t.$$

If $t < q$ then $f(\mathcal{C}) > K$ since $a_1 + a_2 \le 3t$.

Assume now that $t \ge q$ and $a_2 + a_3 \ge 1$. Then $a_1 + a_2 = 3t - a_2 - 2a_3 \le 3t - 1$ and since $t \le m$ we obtain

$$f(\mathcal{C}) = K + 18a^2 - \frac{6a^2}{t}(a_1+a_2) + q - t$$

$$\geq K + 18a^2 - \frac{6a^2(3t-1)}{t} + q - t \geq K + \frac{6a^2}{m} + q - m > K$$

by the choice of a.

Therefore, $t \geq q$ and $a_2 = a_3 = 0$. But then $a_1 = 3t$ and $a_0 + a_1 = 3q$, i. e. $a_0 = 0$ and $t = q$. Hence, the set $E_0 = \{e_j \mid y_j \in \mathcal{C}\}$ induces an exact cover. □

3.2 NP-hardness of Problem 3

Since the norm is always non-negative, maximizing it is the same as maximizing its square, which is much more convenient. So, the decision version of Problem 3 is equivalent to the following

Problem 7. Given a set of vectors $\mathcal{Y} = \{y_1, \ldots, y_N\}$ in d-dimensional Euclidean space and a number K, is there a non-empty subset $\mathcal{C} \subseteq \mathcal{Y}$ such that

$$g(\mathcal{C}) := \| \sum_{x \in \mathcal{C}} x \|^2 \geq K?$$

In order to prove its NP-hardness we first need to show that X3C3 problem remains NP-complete for 3-uniform family of subsets (i. e. if each $v_i \in V$ lies in exactly 3 subsets from E). We refer to this variant of X3C3 problem as X3CE3 problem.

Proposition 1. *The* X3CE3 *problem is NP-complete.*

Proof. Consider an arbitrary instance of X3C3 problem. We may assume that each v_i lies in at least 2 subsets (if some v_i lies in a unique subset then this subset must always be in E_0 and the instance can be simplified). Denote by α and β the number of elements lying in 3 and 2 subsets from E respectively. Since $3\alpha + 2\beta = 3m$, there must be $\beta = 3\gamma$. Enumerate the elements of V so that $v_1, \ldots, v_{3\gamma}$ would lie in two subsets from E. Construct an instance of X3CE3 problem by adding to V a set of new elements $U = \{u_i \mid i = 1, \ldots, 3\gamma\}$ and by adding to E the subsets $\{v_{3i-2}, u_{3i-2}, u_{3i-1}\}, \{v_{3i-1}, u_{3i-2}, u_{3i}\}, \{v_{3i}, u_{3i-1}, u_{3i}\}$, and $\{u_{3i-2}, u_{3i-1}, u_{3i}\}$ for all $i = 1, \ldots, \gamma$. Clearly, no exact cover (a subfamily E_0) in the constructed instance can contain a subset that intersects both with U and V. Therefore, the constructed instance of X3CE3 problem has an exact cover if and only if the initial instance of X3C3 problem has one. □

Theorem 2. *Problem 3 is NP-hard in a strong sense.*

Proof. Consider an arbitrary instance of X3CE3 problem. Note that $m = n = 3q$. Reduce it to an instance of Problem 7 as follows. Put $N = n, d = 3n + 1$ and $K = 6a^2n + 4q^2$ where a is a positive integer such that $a^2 > (n^2 - 4q^2)/6$, and construct the set of vectors $\mathcal{Y}$ in exactly the same way as in proof of Theorem 1.

In an evident way, each $\mathcal{C} \subseteq \mathcal{Y}$ corresponds to a subfamily $E(\mathcal{C}) \subseteq E$. Put $u(\mathcal{C}) = \sum_{y \in \mathcal{C}} y$. Since $g(\mathcal{C}) = \|u(\mathcal{C})\|^2$, the contribution of the i-th coordinate triple into the objective function $g(\mathcal{C})$ is $6a^2$ if 1 or 2 vectors corresponding to subsets containing v_i lies in $E(\mathcal{C})$, and the contribution is 0 otherwise.

If there is an exact cover E_0 in X3CE3 problem then let $\mathcal{C}$ contain all $n-q = 2q$ vectors corresponding to the elements from $E \setminus E_0$. Since each element of V lies in exactly 2 subsets from $E \setminus E_0$, we have $g(\mathcal{C}) = 6a^2n + 4q^2 = K$.

Suppose now that there exists a subset $\mathcal{C} \subseteq \mathcal{Y}$ of cardinality $t > 0$ such that $g(\mathcal{C}) \geq K$. As in the proof of Theorem 1, for each $k \in \{0, 1, 2, 3\}$ denote by a_k the number of coordinate triples that are non-zero in exactly k vectors from $\mathcal{C}$. We have $a_0 + a_1 + a_2 + a_3 = n = 3q$ and $a_1 + 2a_2 + 3a_3 = 3t$. It follows form the arguments above that $g(\mathcal{C}) = 6a^2(a_1 + a_2) + t^2$.

If $t < 2q$ then $g(\mathcal{C}) < K$ since $a_1 + a_2 \leq n$.

If $t > 2q$ then $0 < 3t - 6q = a_3 - a_1 - 2a_0 \leq a_3$ and thus $a_3 \geq 1$ implying $a_1 + a_2 \leq n - 1$. Therefore,

$$g(\mathcal{C}) \leq 6a^2(n-1) + n^2 = K - 6a^2 + n^2 - 4q^2 < K$$

by the choice of a.

Hence, $t = 2q$ and $a_1 + a_2 = n$, which implies $a_0 = a_1 = a_3 = 0$ and $a_2 = 3q$. This means that each element $v_i \in V$ lies exactly in 2 subsets from $E(\mathcal{C})$. But then the subfamily $E_0 = E \setminus E(\mathcal{C})$ induces an exact cover in X3CE3 problem.□

4 Conclusions

In this paper we have presented two new NP-hardness proofs for the subset choice problems with unknown cardinalities of the sought subsets. Namely, the problems of finding a subset with the longest sum and a subset with the maximum squared norm of the sum normalized by the size of the subset were considered. These problems find their applications in the areas of data analysis and pattern recognition. Namely, the first problem can be used for finding a correct direction to a certain object, and the second one arises in problem of detection an informationally significant fragment in a noisy data.

The suggested new NP-hardness proofs use an easy and natural reduction from Ecact Cover by 3-Sets problem. We believe that new natural reductions could be helpful for proving NP-hardness of related problems with unknown cardinalities of the sought subsets.

Anknowledgement. The author is grateful to the unknown referees for their valuable comments.

References

1. Aggarwal, A., Imai, H., Katoh, N., Suri, S.: Finding k points with minimum diameter and related problems. J. Algorithms **12**(1), 38–56 (1991)
2. Aloise, D., Deshpande, A., Hansen, P., Popat, P.: NP-hardness of Euclidean sum-of-squares clustering. Mach. Learn. **75**(2), 245–248 (2009)
3. Baburin, A.E., Gimadi, E.K., Glebov, N.I., Pyatkin, A.V.: The problem of finding a subset of vectors with the maximum total weight. J. Appl. Ind. Math. **2**(1), 32–38 (2008). https://doi.org/10.1134/S1990478908010043

4. Dolgushev, A.V., Kel'manov, A.V.: An approximation algorithm for solving a problem of cluster analysis. J. Appl. Ind. Math. **5**(4), 551–558 (2011)
5. Eremeev, A.V., Kel'manov, A.V., Kovalyov, M.Y., Pyatkin, A.V.: Maximum diversity problem with squared euclidean distance. In: Khachay, M., Kochetov, Y., Pardalos, P. (eds.) MOTOR 2019. LNCS, vol. 11548, pp. 541–551. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-22629-9_38
6. Fisher, W.D.: On grouping for maximum homogeneity. J. Am. Stat. Assoc. **53**(284), 789–798 (1958)
7. Garey, M.R., Johnson, D.S.: Computers and Intractability. The Guide to the Theory of NP-Completeness. W. H. Freeman and Company, San Francisco (1979)
8. Gimadi, E.K., Kel'manov, A.V., Kel'manova, M.A., Khamidullin, S.A.: A posteriori detection of a quasi periodic fragment in numerical sequences with given number of recurrences. Sib. Zh. Ind. Mat. **9**(1), 55–74 (2006). (in Russian)
9. Gimadi, E.K., Pyatkin, A.V., Rykov, I.A.: On polynomial solvability of some problems of a vector subset choice in a Euclidean space of fixed dimension. J. Appl. Ind. Math. **4**(1), 48–53 (2010)
10. Gimadi, E., Rykov, I.A.: Efficient randomized algorithm for a vector subset problem. In: Kochetov, Y., Khachay, M., Beresnev, V., Nurminski, E., Pardalos, P. (eds.) DOOR 2016. LNCS, vol. 9869, pp. 148–158. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-44914-2_12
11. Hwang, F.K., Onn, S., Rothblum, U.G.: A polynomial time algorithm for shaped partition problems. SIAM J. Optim. **10**, 70–81 (1999)
12. 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**, 762–774 (2001)
13. Kel'manov, A.V., Khandeev, V.I.: A 2-approximation polynomial algorithm for a clustering problem. J. Appl. Ind. Math. **7**(4), 515–521 (2013). https://doi.org/10.1134/S1990478913040066
14. Kelmanov, A.V., Pyatkin, A.V.: On the complexity of a search for a subset of "similar" vectors. Doklady Math. **78**(1), 574–575 (2008)
15. Kel'manov, A.V., Pyatkin, A.V.: On a version of the problem of choosing a vector subset. J. Appl. Ind. Math. **3**(4), 447–455 (2009)
16. Kel'manov, A.V., Pyatkin, A.V.: NP-Completeness of some problems of choosing a vector subset. J. Appl. Ind. Math. **5**(3), 352–357 (2011)
17. Kel'manov, A.V., Pyatkin, A.V.: On the complexity of some quadratic euclidean 2-clustering problems. Comput. Math. Math. Phys. **56**(3), 491–497 (2016)
18. Kel'manov, A.V., Pyatkin, A.V., Khandeev, V.I.: NP-hardness of quadratic euclidean 1-mean and 1-median 2-clustering problem with constraints on the cluster sizes. Doklady Math. **100**(3), 545–548 (2019). https://doi.org/10.1134/S1064562419060127
19. Kelmanov, A.V., Romanchenko, S.M.: An FPTAS for a vector subset search problem. J. Appl. Ind. Math. **8**(3), 329–336 (2014). https://doi.org/10.1134/S1990478914030041
20. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of 5-th Berkeley Symposium on Mathematics, Statistics and Probability, vol. 1, pp. 281–297 (1967)
21. Mahajan, M., Nimbhorkar, P., Varadarajan, K.: The planar k-means problem is NP-hard. Theor. Comput. Sci. **442**, 13–21 (2012)
22. Pyatkin, A.V.: On complexity of a choice problem of the vector subset with the maximum sum length. J. Appl. Ind. Math. **4**(4), 549–552 (2010)

23. Shenmaier, V.V.: An approximation scheme for a problem of search for a vector subset. J. Appl. Ind. Math. **6**(3), 381–386 (2012)
24. Shenmaier, V.V.: Solving some vector subset problems by Voronoi diagrams. J. Appl. Ind. Math. **10**(4), 560–566 (2016). https://doi.org/10.1134/S199047891604013X
25. Shenmaier, V.V.: An exact algorithm for finding a vector subset with the longest sum. J. Appl. Ind. Math. **11**(4), 584–593 (2017). https://doi.org/10.1134/S1990478917040160
26. Shenmaier, V.V.: Complexity and approximation of finding the longest vector sum. Comput. Math. Math. Phys. **58**(6), 850–857 (2018). https://doi.org/10.1134/S0965542518060131