On the Complexity of Some Clustering Problems

A. V. Kel'manov,[*] A. V. Pyatkin [†]

[*] Sobolev Institute of Mathematics, Siberian Branch of the RAS, kelm@math.nsc.ru
[†] Durham University, Durham, UK; artempyatkin@gmail.com

In this work we study some partition problems of a vector set in the Euclidean space with the minimum sum of squares criterion. The goal of this paper is to provide the analysis of algorithmic complexity of these problems.

The problem of Euclidean vectors set partition into subsets (clusters) with the minimum sum of squares of the distances from cluster elements to cluster centers (the cluster center is defined as the average vector in the cluster) is well-known in the literature as MSSC (Minimum Sum-of-Squares Clustering) problem. The variety of typical problems of data analysis that arise in a wide range of applications can be reduced to the MSSC problem (e.g., see [1] and the references in this paper). This problem can be stated in the form of the properties verification as follows:

MSSC problem. *Input*: A set of vectors $\mathcal{Y} = \{y_1, \ldots, y_N\}$ from $\mathbb{R}^q$, a positive integer $J > 1$, and a positive real A. *Question*: Is there a partition of $\mathcal{Y}$ into nonempty subsets (clusters) $\mathcal{C}_1, \ldots, \mathcal{C}_J$ such that

$$\sum_{j=1}^{J} \sum_{y \in \mathcal{C}_j} \|y - \overline{y}(\mathcal{C}_j)\|^2 \leq A, \tag{1}$$

where $\overline{y}(\mathcal{C}_j) = \frac{1}{|\mathcal{C}_j|} \sum_{y \in \mathcal{C}_j} y$, $j = 1, \ldots, J$, is the center of the jth cluster?

Recall some known facts on the algorithmic complexity of MSSC. The one-dimensional variant of the problem is solvable in polynomial time [2]. The four possible cases of the multidimensional version of this problem are generated by combining the parameters of the space dimension and the number of clusters that can either be the part of the input or not. Regarding these cases we know the following.

If the dimension of the space and the number of clusters are not the parts of the input then the problem is solved exactly in polynomial time [3]. The NP-completeness of the three other cases of the problem was proved in [4]–[6].

Recently in [7], the NP-completeness (in the strong sense) of the following clustering problem was shown.

MSSC-Case problem. *Input*: A set of vectors $\mathcal{Y} = \{y_1, \ldots, y_N\}$ from $\mathbb{R}^q$, a positive integer $M > 1$, and a positive real A. *Question*: Is there a partition of $\mathcal{Y}$ into $J = N - M + 1$ nonempty clusters $\mathcal{C}_1, \ldots, \mathcal{C}_J$ such that one of the clusters has cardinality M and the inequality (1) holds?

Three following NP-complete problems of the vector subset choice are closely related to the MSSC-Case problem (see [7]).

VS-1 problem (Vector Subset 1). *Input*: A set of vectors $\mathcal{Y} = \{y_1, \ldots, y_N\}$ from $\mathbb{R}^q$, a positive integer $M > 1$, and a positive real B. *Question*: Is there a subset $\mathcal{C} \subseteq \mathcal{Y}$ such that

$$\frac{1}{|\mathcal{C}|} \Big\| \sum_{y \in \mathcal{C}} y \Big\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2 \geq B,$$

subject to the constraint $|\mathcal{C}| = M$?

VS-2 problem (Vector Subset 2) *Input*: A set of vectors $\mathcal{Y} = \{y_1, \ldots, y_N\}$ from $\mathbb{R}^q$, a positive integer $M > 1$, and a positive real A. *Question*: Is there a subset $\mathcal{C} \subseteq \mathcal{Y}$ such that

$$\sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 \leq A,$$

where $\overline{y}(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum_{y \in \mathcal{C}} y$, subject to the constraint $|\mathcal{C}| = M$?

VS-3 problem (Vector Subset 3) *Input*: A set of vectors $\mathcal{Y} = \{y_1, \dots, y_N\}$ from $\mathbb{R}^q$, a positive integer $M > 1$, and a positive real D. *Question*: Is there a subset $\mathcal{C} \subseteq \mathcal{Y}$ such that

$$\sum_{y\in\mathcal{C}} \sum_{z\in\mathcal{C}} \|y - z\|^2 \leq D,$$

subject to the constraint $|\mathcal{C}| = M$?

For the objective functions of the problems VS-1, VS-2 and VS-3 we have the following equations [7], [8]

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

In addition, if in the MSSC-Case problem, for example, the cardinality of mth cluster $\mathcal{C}_m$ is equal to M, then we have

$$\sum_{j=1}^{J} \sum_{y\in\mathcal{C}_j} \|y - \overline{y}(\mathcal{C}_j)\|^2 = \sum_{y\in\mathcal{C}_m} \|y - \overline{y}(\mathcal{C}_m)\|^2.$$

Thus, the problems MSSC-Case and VS-2 are equivalent.

The NP-completeness of the two following clustering problem was proved in [9]. Before formulating these problems, we explain their titles notation used below. The first four symbols MSSC there stand for Minimum Sum-of-Squares Clustering. The last symbol stands for the words Nonfixed or Fixed indicating whether the cardinalities of the clusters mentioned in the problem are the parts of the input or not respectively.

MSSC-N problem. *Input*: A set of vectors $\mathcal{Y} = \{y_1, \dots, y_N\}$ from $\mathbb{R}^q$, a positive integer $J > 1$, and a positive real A. *Question*: Is there a partition of $\mathcal{Y}$ into nonempty clusters $\mathcal{C}_1, \dots, \mathcal{C}_J$, and $\mathcal{B} = \mathcal{Y} \setminus (\cup_j \mathcal{C}_j)$ such that

$$\sum_{j=1}^{J} \sum_{y\in\mathcal{C}_j} \|y - \overline{y}(\mathcal{C}_j)\|^2 + \sum_{y\in\mathcal{B}} \|y\|^2 \leq A, \qquad (2)$$

where $\overline{y}(\mathcal{C}_j) = \frac{1}{|\mathcal{C}_j|}\sum_{y\in\mathcal{C}_j} y$, $j = 1, \dots, J$, is the center of the jth cluster?

MSSC-F problem. *Input*: A set of vectors $\mathcal{Y} = \{y_1, \dots, y_N\}$ from $\mathbb{R}^q$, positive integers $M_1, \dots, M_J$, and a positive real A. *Question*: Is there a partition of $\mathcal{Y}$ into nonempty clusters $\mathcal{C}_1, \dots, \mathcal{C}_J$, and $\mathcal{B} = \mathcal{Y} \setminus (\cup_j \mathcal{C}_j)$ such that the inequality (2) holds subject to the constraints $|\mathcal{C}_j| = M_j$, $j = 1, \dots, J$, on the clusters cardinalities?

The following two NP-complete problems the vector subsets choice are closely related to the MSSC-F and MSSC-F clustering problems (see [9]).

SSA-F problem. *Input*: A set of vectors $\mathcal{Y} = \{y_1, \dots, y_N\}$ from $\mathbb{R}^q$, positive integers $M_1, \dots, M_J$, and a positive real D. *Question*: Is there a family $\{\mathcal{C}_1, \dots, \mathcal{C}_J\}$ of disjoint subsets of $\mathcal{Y}$ such that

$$\sum_{j=1}^{J} \frac{1}{|\mathcal{C}_j|}\Big\|\sum_{y\in\mathcal{C}_j} y\Big\|^2 \geq D, \qquad (3)$$

subject to the constraints $|\mathcal{C}_j| = M_j$, $j = 1, \dots, J$, on the cardinalities of the subsets?

SSA-N problem. *Input*: A set of vectors $\mathcal{Y} = \{y_1, \dots, y_N\}$ from $\mathbb{R}^q$, a positive integer J, and a positive real D. *Question*: Is there a family $\{\mathcal{C}_1, \dots, \mathcal{C}_J\}$ of nonempty disjoint subsets of $\mathcal{Y}$ such that the inequality (3) holds?

For the objective functions of the problems MSSC-N, MSSC-F, SSA-N, and SSA-F we have the equation [9]

$$\sum_{j=1}^{J} \sum_{y\in\mathcal{C}_j} \|y - \overline{y}(\mathcal{C}_j)\|^2 + \sum_{y\in\mathcal{B}} \|y\|^2$$
$$= \sum_{y\in\mathcal{Y}} \|y\|^2 - \sum_{j=1}^{J} \frac{1}{|\mathcal{C}_j|}\Big\|\sum_{y\in\mathcal{C}_j} y\Big\|^2.$$

The first addend on the right side of this equation is a constant. Therefore, MSSC-N (MSSC-F) and SSA-N (SSA-F) is the pair of opposite (min-max) optimization problems.

In this paper we analyse some new clustering problems. Consider the following data structure represented by a set of Euclidean vectors. Let the vector sequence $x_n \in \mathbb{R}^q$, $n \in \mathcal{N}$, where $\mathcal{N} =$

$\{1, 2, \ldots, N\}$, has the property

$$x_n = \begin{cases} w_j, & n \in \mathcal{M}_j, \quad j = 1, \ldots, J, \\ v_k, & n \in \mathcal{N}_k, \quad k = 1, \ldots, K, \end{cases} \tag{4}$$

where $(\cup_j \mathcal{M}_j) \cup (\cup_k \mathcal{N}_k) = \mathcal{N}$, and $\mathcal{M}_1, \ldots, \mathcal{M}_J$, $\mathcal{N}_1, \ldots, \mathcal{N}_K$ are nonempty and mutually disjoint subsets.

Assume that a sequence

$$y_n = x_n + e_n, \qquad n \in \mathcal{N}, \tag{5}$$

is available for the analysis, where e_n is the noise (error) vector independent of x_n. Taking into account the fact that the elements of the sequence x_n depend on two families of subsets $\{\mathcal{M}_1, \ldots, \mathcal{M}_J\}$ and $\{\mathcal{N}_1, \ldots, \mathcal{N}_K\}$, and two vector sets $\{w_1, \ldots, w_J\}$ and $\{v_1, \ldots, v_K\}$, define

$$S(\mathcal{M}_1, \ldots, \mathcal{M}_K, \mathcal{N}_1, \ldots, \mathcal{N}_J, \\ w_1, \ldots, w_J, v_1, \ldots, v_K) = \sum_{n \in \mathcal{N}} \|y_n - x_n\|^2 \tag{6}$$

and consider the data analysis model in the form of the following optimization problem.

Given the sequence y_n, $n \in \mathcal{N}$, and vectors $v_1, \ldots, v_K$, find nonempty mutually disjoint subsets $\mathcal{M}_1, \ldots, \mathcal{M}_J$, $\mathcal{N}_1, \ldots, \mathcal{N}_K$ of $\mathcal{N}$ and vectors $w_1, \ldots, w_J$ minimizing $S(\cdot \,|\, v_1, \ldots, v_K)$ under the condition that the structure of the sequence is described by the formulas (4) and (5).

Rewriting the sum of squares on the right side of (6) and taking into account (4), we obtain

$$S(\cdot) = \sum_{j=1}^{J} \sum_{n \in \mathcal{M}_j} \|y_n - w_j\|^2 \\ + \sum_{k=1}^{K} \sum_{n \in \mathcal{N}_k} \|y_n - v_k\|^2. \tag{7}$$

The minimum of functional (7) over the unknown vectors $w_1, \ldots, w_J$ can be found analytically. It is easy to verify that for arbitrary nonempty mutually disjoint subsets $\mathcal{M}_1, \ldots, \mathcal{M}_J$, $\mathcal{N}_1, \ldots, \mathcal{N}_K$, the minimum is reached at the vectors $\overline{y}(\mathcal{M}_j) = \frac{1}{|\mathcal{M}_j|} \sum_{n \in \mathcal{M}_j} y_n$, $j = 1, \ldots, J$; this minimum equals

$$S_{\min}(\mathcal{M}_1, \ldots, \mathcal{M}_J, \mathcal{N}_1, \ldots, \mathcal{N}_K \,|\, v_1, \ldots, v_K) \\ = \sum_{j=1}^{J} \sum_{n \in \mathcal{M}_j} \|y_n - \overline{y}(\mathcal{M}_j)\|^2 \\ + \sum_{k=1}^{K} \sum_{n \in \mathcal{N}_k} \|y_n - v_k\|^2. \tag{8}$$

Consider possible variants of the optimization problems to which the problem formulated above can be reduced.

Let $\mathcal{Y} = \{y_n | \, n \in \mathcal{N}\}$, $\mathcal{C}_j = \{y_n | \, n \in \mathcal{M}_j\}$, $j = 1, \ldots, J$, and $\mathcal{B}_k = \{y_n | \, n \in \mathcal{N}_k\}$, $k = 1, \ldots, K$. Replacing in (8) the summation over the indices with the summation over the elements of these sets, we obtain four reduced optimization problems.

Before formulating these problems, we explain their titles notation used below. The first four symbols are MSSC. The next two symbols stand for the words Nonfixed and Fixed indicating whether the cardinalities of the clusters from two families appearing in the problem formulation are the parts of the input or not respectively; thus, we have four combinations FF, NF, FN, and NN. We state the reduced problems in the form of the properties verification.

MSSC-NN problem. *Input*: A set of vectors $\mathcal{Y} = \{y_1, \ldots, y_N\}$ and an alphabet of vectors $\{v_1, \ldots, v_K\}$ from $\mathbb{R}^q$, a positive integer $J > 1$, and a positive real A. *Question*: Is there a partition of $\mathcal{Y}$ into nonempty clusters $\mathcal{C}_1, \ldots, \mathcal{C}_J$, $\mathcal{B}_1, \ldots, \mathcal{B}_K$ such that

$$\sum_{j=1}^{J} \sum_{y \in \mathcal{C}_j} \|y - \overline{y}(\mathcal{C}_j)\|^2 + \sum_{k=1}^{K} \sum_{y \in \mathcal{B}_k} \|y - v_k\|^2 \le A, \tag{9}$$

where $\overline{y}(\mathcal{C}_j) = \frac{1}{|\mathcal{C}_j|} \sum_{y \in \mathcal{C}_j} y$, $j = 1, \ldots, J$, is the center of the jth cluster?

MSSC-FN problem. *Input*: A set of vectors $\mathcal{Y} = \{y_1, \ldots, y_N\}$ and an alphabet of vectors $\{v_1, \ldots, v_K\}$ from $\mathbb{R}^q$, positive integers $M_1, \ldots, M_J$, and a positive real A. *Question*: Is there a partition of $\mathcal{Y}$ into nonempty clusters $\mathcal{C}_1, \ldots, \mathcal{C}_J$, $\mathcal{B}_1, \ldots, \mathcal{B}_K$ such that the inequality

(9) holds subject to the constraints $|\mathcal{C}_j| = M_j$, $j = 1, \ldots, J$, on the clusters cardinalities?

MSSC-NF problem. *Input*: A set of vectors $\mathcal{Y} = \{y_1, \ldots, y_N\}$ and an alphabet of vectors $\{v_1, \ldots, v_K\}$ from $\mathbb{R}^q$, positive integers J, $N_1, \ldots, N_K$, and a positive real A. *Question*: Is there a partition of $\mathcal{Y}$ into nonempty clusters $\mathcal{C}_1, \ldots, \mathcal{C}_J$, $\mathcal{B}_1, \ldots, \mathcal{B}_K$ such that the inequality (9) holds subject to the constraints $|\mathcal{B}_k| = N_k$, $k = 1, \ldots, K$, on the clusters cardinalities?

MSSC-FF problem. *Input*: A set of vectors $\mathcal{Y} = \{y_1, \ldots, y_N\}$ and an alphabet of vectors $\{v_1, \ldots, v_K\}$ from $\mathbb{R}^q$, a positive integers $M_1, \ldots, M_J$, $N_1, \ldots, N_K$, and a positive real A. *Question*: Is there a partition of $\mathcal{Y}$ into nonempty clusters $\mathcal{C}_1, \ldots, \mathcal{C}_J$, $\mathcal{B}_1, \ldots, \mathcal{B}_K$ such that the inequality (9) holds subject to the constraints $|\mathcal{C}_j| = M_j$, $j = 1, \ldots, J$, and $|\mathcal{B}_k| = N_k$, $k = 1, \ldots, K$, on the clusters cardinalities?

The main result of the paper is the following

Theorem 1. *The problems MSSC-NN, MSSC-FN, MSSC-NF, and MSSC-FF are NP-complete.*

No efficient algorithms with guaranteed accuracy bounds for solving these problems in general case are known. A 2-approximation algorithm for MSSC-FF with $K = 1$, $J = 1$, and the alphabet of vectors containing only one zero vector was presented in [10]. The complexity of the algorithm is $\mathcal{O}(qN^2)$.

ACKNOWLEDGMENTS. This work was partially supported by EPSRC Grant EP/F064551/1, by the Russian Foundation for Basic Research (projects Nos. 09-01-00032 and 10-07-00195), the Program "Development of the Scientific Potential of Higher Education" of the Russian Federal Agency for Education (project No. 2.1.1/3235), the Siberian Division of the Russian Academy of Sciences (Integration project No. 44), and the Program No. 2 of the Presidium of the Russian Academy of Sciences (project No. 227).

References

[1] Anil K., Jain K. *Data Clustering: 50 Years Beyond k-Means* // Pattern Recognition Letters. 2010. Vol. 31. Pp. 651–666.

[2] Rao M. *Cluster Analysis and Mathematical Programming* // J. Am. Stat. Assoc. 1971. Vol. 66. P. 622–626.

[3] Inaba M., Katch N., Imai H. *Applications of Weighted Voronoi Diagrams and Randomization to Variance-Based Clustering* // Proc. Annual Symp. on Comput. Geom. 1994. P. 332–339.

[4] Aloise D., Deshpande A., Hansen P., Popat P. *NP-Hardness of Euclidean Sum-of-Squares Clustering* // Les Cahiers du GERAD, G-2008-33, 2008. 4 p.

[5] Mahajan M., Nimbhorkar P., Varadarajan K. *The Planar k-means Problem is NP-Hard* // Lecture Notes in Computer Science. 2009. Vol. 5431. P. 284–285.

[6] Dolgushev A. V., Kel'manov A. V. *On the Algorithmic Complexity of a Problem in Cluster Analysis* // Journal of Applied and Industial Mathematics. 2011. Vol. 5, No 2. P. 191–194.

[7] Kel'manov A. V., Pyatkin A. V. *NP-completness of Some Problems of Choosing Vector Subset* // Discrete Analysis and Operation Research. 2010. Vol. 17, No 5. Pp. 37–45 (in russian).

[8] Edwards A. W. F., Cavalli-Sforza L. L. *A Method for Cluster Analysis* // Biometrics. 1965. Vol. 21. P. 362–375.

[9] Kel'manov A. V., Pyatkin A. V. *Complexity of Certain Problems of Searching Subsets of Vectors and Cluster Analysis* // Computational Mathematics and Mathematical Physics. 2009. Vol. 49, No 11. P. 1966–1971.

[10] Dolgushev A. V., Kel'manov A. V. *An Approximation algorithm for One Cluster Analysis Problem* // Discrete Analysis and Operation Research. 2011. Vol. 18, No 2. Pp. 29–40 (in russian).