

ПОЛНОСТЬЮ ПОЛИНОМИАЛЬНАЯ  
АППРОКСИМАЦИОННАЯ СХЕМА ДЛЯ ОДНОЙ ЗАДАЧИ  
ДВУХКЛАСТЕРНОГО РАЗБИЕНИЯ  
ПОСЛЕДОВАТЕЛЬНОСТИ \*)

А.В. Кельманов<sup>1,2</sup>, С.А. Хамидуллин<sup>1</sup>, В.И. Хандеев<sup>1</sup>

<sup>1</sup>Институт математики им. С.Л. Соболева,  
пр. Коптюга, 4, 630090 Новосибирск, Россия

<sup>2</sup>Новосибирский государственный университет,  
ул. Пирогова, 2, 630090 Новосибирск, Россия

e-mail: kelm@math.nsc.ru, kham@math.nsc.ru, khandeev@math.nsc.ru

**Аннотация.** Рассматривается NP-трудная в сильном смысле задача разбиения конечной последовательности точек евклидова пространства на два кластера (подпоследовательности), имеющих заданные мощности, по критерию минимума суммы по обоим кластерам внутрикластерных сумм квадратов расстояний от элементов кластеров до их центров. Предполагается, что центр одного из искомым кластеров неизвестен и определяется как среднее значение по всем элементам, образующим этот кластер, а центр второго задан в начале координат. При этом разбиение подчинено условию: разность между номерами последующего и предыдущего элементов последовательности, входящих в первый кластер, ограничена сверху и снизу заданными константами. Обоснована полностью полиномиальная приближённая схема для случая задачи, в котором размерность пространства фиксирована. Библиогр. 27.

**Ключевые слова:** разбиение, последовательность, евклидово пространство, минимум суммы квадратов расстояний, NP-трудность, FPTAS.

**Введение**

Предметом исследования работы является NP-трудная в сильном смысле квадратичная задача разбиения конечной последовательности

---

\*) Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований (проекты 15-01-00462, 16-07-00168 и 16-31-00186-мол-а).

точек евклидова пространства на два кластера. Цель исследования — изучение вопросов аппроксимируемости этой задачи и обоснование полностью полиномиальной приближённой схемы (FPTAS) для её специального случая.

Исследование мотивировано слабой изученностью задачи и её актуальностью для многих естественнонаучных и технических приложений, в которых требуется классификация упорядоченных по времени данных численных экспериментов или результатов наблюдения за состояниями каких-либо материальных объектов. Ситуации, в которых требуется решение рассматриваемой задачи, характерны, в частности, для дистанционного зондирования, геофизики, биометрики, медицинской и технической диагностики, обработки речевых сигналов, электронной разведки, радиолокации, гидроакустики и др. (см., например, [6, 10, 12, 13, 21, 26] и цитированные там работы). Содержательная трактовка задачи приведена в разд. 1.

### 1. Формулировка задачи, известные и полученный результаты

Одной из наиболее известных (см., например, [20, 22, 24, 25, 27]) задач разбиения конечного множества точек евклидова пространства является задача MSSC (Minimum Sum-of-Squares Clustering), двухкластерный вариант которой имеет следующую формулировку.

**Задача 2-MSSC** (Minimum Sum-of-Squares 2-Clustering). *Дано множество  $\mathcal{Y} = \{y_1, \dots, y_N\}$  точек из  $\mathbb{R}^q$ . Найти разбиение множества  $\mathcal{Y}$  на два непустых кластера  $\mathcal{C}$  и  $\mathcal{Y} \setminus \mathcal{C}$  такое, что*

$$\sum_{y \in \mathcal{C}} \|y - \bar{y}(\mathcal{C})\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y - \bar{y}(\mathcal{Y} \setminus \mathcal{C})\|^2 \rightarrow \min,$$

где  $\bar{y}(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum_{y \in \mathcal{C}} y$  и  $\bar{y}(\mathcal{Y} \setminus \mathcal{C}) = \frac{1}{|\mathcal{Y} \setminus \mathcal{C}|} \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} y$  — геометрические центры (центроиды) кластеров.

Несмотря на полувековую известность задачи MSSC, её труднорешаемость установлена лишь несколько лет назад в [19]. С исследованием этой задачи и её приложениями связаны тысячи публикаций.

В последнее десятилетие активно изучалась (см. [4, 5, 7–9, 16–18]) близкая к задаче MSSC в постановочном плане NP-трудная в сильном смысле

**Задача 1** (Minimum Sum-of-Squares 2-Clustering with given center of one cluster and cluster cardinalities). *Дано множество  $\mathcal{Y} = \{y_1, \dots, y_N\}$*

точек из  $\mathbb{R}^q$  и натуральное число  $M$ . Найти разбиение множества  $\mathcal{Y}$  на два кластера  $\mathcal{C}$  и  $\mathcal{Y} \setminus \mathcal{C}$  такое, что

$$\sum_{y \in \mathcal{C}} \|y - \bar{y}(\mathcal{C})\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2 \rightarrow \min,$$

где  $\bar{y}(\mathcal{C})$  — центроид кластера  $\mathcal{C}$ , при ограничении  $|\mathcal{C}| = M$ .

В этой задаче так же, как и в задаче MSSC, требуется разбить конечное множество точек евклидова пространства на два кластера по критерию минимума суммы по обоим кластерам внутрикластерных сумм. При этом одна из внутрикластерных сумм — такая же, как в задаче MSSC, т. е. это сумма квадратов расстояний от элементов кластера до неизвестного центроида  $\bar{y}(\mathcal{C})$ , а другая — сумма квадратов расстояний от элементов кластера до заданного в произвольной точке желаемого центра. Без ограничения общности заданным центром может служить начало координат, поэтому в целевой функции задачи 1 вместо центроида  $\bar{y}(\mathcal{Y} \setminus \mathcal{C})$  фигурирует центр 0. Кроме того, предполагается, что мощности кластеров заданы на входе.

В настоящей работе рассматривается задача разбиения последовательности, которая обобщает задачу 1 разбиения множества. От задачи 1 рассматриваемую задачу отличает следующая особенность: входом задачи является не множество, а последовательность, при этом имеются ограничения на номера элементов подпоследовательностей, включаемых в кластеры.

Положим  $\mathcal{N} = \{1, \dots, N\}$ . Рассматриваемая задача имеет следующую формулировку (см. [10, 13]).

**Задача 2** (Minimum Sum-of-Squares 2-Clustering problem on sequence with given center of one cluster and cluster cardinalities). *Даны последовательность  $\mathcal{Y} = (y_1, \dots, y_N)$  точек из  $\mathbb{R}^q$ , натуральные числа  $T_{\min}$ ,  $T_{\max}$  и  $M > 1$ . Найти подмножество  $\mathcal{M} = \{n_1, \dots, n_M\} \subseteq \mathcal{N}$  номеров элементов последовательности  $\mathcal{Y}$  такое, что минимальна целевая функция*

$$F(\mathcal{M}) = \sum_{j \in \mathcal{M}} \|y_j - \bar{y}(\mathcal{M})\|^2 + \sum_{i \in \mathcal{N} \setminus \mathcal{M}} \|y_i\|^2, \quad (1)$$

где  $\bar{y}(\mathcal{M}) = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} y_i$ , при ограничениях

$$1 \leq T_{\min} \leq n_m - n_{m-1} \leq T_{\max} \leq N, \quad m = 2, \dots, M, \quad (2)$$

на элементы набора  $(n_1, \dots, n_M)$ .

В этой задаче центроид  $\bar{y}(\mathcal{M})$  подпоследовательности  $\{y_j \mid j \in \mathcal{M}\}$  неизвестен, а центр подпоследовательности  $\{y_i \mid i \in \mathcal{N} \setminus \mathcal{M}\}$  фиксирован в начале координат, как в задаче 1.

Задача 2 имеет следующую трактовку (см. [10, 13]). Имеется последовательность  $\mathcal{Y}$ , содержащая  $N$  упорядоченных по времени результатов  $y_1, \dots, y_N$  измерения набора  $y$  из  $q$  числовых характеристик некоторого объекта, который может находиться в двух состояниях — активном и пассивном. В пассивном состоянии все элементы набора равны нулю, а в активном — хотя бы одна из компонент набора не равна нулю. Соответствие элементов последовательности какому-либо состоянию объекта неизвестно. Натуральные числа  $T_{\min}$  и  $T_{\max}$  соответствуют минимальному и максимальному интервалам времени между двумя последовательными активными состояниями объекта. Номера из наборов  $\mathcal{M}$  и  $\mathcal{N} \setminus \mathcal{M}$  соответствуют моментам времени, в которые объект находился в активном и пассивном состояниях. Требуется разбить последовательность на два кластера, соответствующих активному и пассивному состояниям объекта, и оценить набор  $\bar{y}(\mathcal{M})$  характеристик этого объекта в активном состоянии.

Частный случай задачи 2, в котором  $T_{\min} = 1$  и  $T_{\max} = N$ , эквивалентен [10] NP-трудной в сильном смысле задаче 1, в которой ограничения (2) на номера элементов, включаемых в кластеры, отсутствуют, а входом является не последовательность, а множество.

Напомним сначала результаты, полученные для задачи 1, так как она является частным случаем задачи, рассматриваемой в настоящей работе.

Задача 1 полиномиально эквивалентна [4, 9] NP-трудной в сильном смысле задаче LVS (subset with the Longest Vector Sum) максимизации нормы суммы векторов подмножества заданной мощности [1, 23]. Из указанной полиномиальной эквивалентности и результатов [1, 23] о сложности задачи LVS следует, что задача 1 NP-трудна [4, 9] в сильном смысле. При этом в случае фиксированной размерности  $q$  пространства задача 1 полиномиально разрешима [17] за время  $\mathcal{O}(q^2 N^{2q})$ , что является следствием соответствующих результатов для задачи LVS [3].

Кроме того, к числу найденных к настоящему времени эффективных алгоритмических решений задачи 1 относятся следующие. В [4] предложен 2-приближённый полиномиальный алгоритм, трудоёмкость которого есть величина  $\mathcal{O}(qN^2)$ . Полиномиальная приближённая схема (PTAS) обоснована в [5]. Эта схема позволяет решать задачу 1 с произвольной относительной погрешностью  $\varepsilon$  за время  $\mathcal{O}(qN^{2/\varepsilon+1}(9/\varepsilon)^{3/\varepsilon})$ . В [16] предложен рандомизированный алгоритм, который в случае  $M \geq \beta N$

для некоторого  $\beta \in (0, 1)$  при заданных относительной погрешности  $\varepsilon$  и вероятности  $\gamma$  несрабатывания находит  $(1 + \varepsilon)$ -приближённое решение задачи за время  $\mathcal{O}(2^k q(k + N))$ , где  $k = \max(\lceil \frac{2}{\beta} \lceil \frac{2}{\gamma \varepsilon} \rceil \rceil, \lceil \frac{8}{\beta} \log \frac{2}{\gamma} \rceil)$ , линейное по  $N$  и  $q$  при фиксированных значениях  $\beta$ ,  $\gamma$  и  $\varepsilon$ . Там же найдены условия, при которых этот алгоритм асимптотически точен и имеет трудоёмкость  $\mathcal{O}(qN^2)$ . Для случая фиксированной размерности  $q$  пространства и целочисленных координат точек построены точные псевдополиномиальные алгоритмы, имеющие трудоёмкость  $\mathcal{O}(N(MD)^q)$  [17] и  $\mathcal{O}(NM(MD)^{q-1})$  [2], где  $D$  — максимальное абсолютное значение координат входных точек. В [18] установлено, что для задачи 1 не существует схемы FPTAS, если  $P \neq NP$ , и там же такая схема обоснована для специального случая, в котором размерность  $q$  пространства фиксирована. Эта схема позволяет решать задачу с произвольной относительной погрешностью  $\varepsilon$  за время  $\mathcal{O}(N^2(1/\varepsilon)^{q/2})$ .

Заметим сначала, что поскольку задача 2 — обобщение NP-трудной в сильном смысле задачи 1, для неё, как и для задачи 1, не существует ни точного полиномиального, ни точного псевдополиномиального алгоритмов, ни схемы FPTAS, если  $P \neq NP$ . К настоящему времени для рассматриваемой задачи 2 были получены следующие результаты.

В [10] анализировался вариант задачи 2, в котором  $T_{\min}$  и  $T_{\max}$  — параметры, и было установлено, что задача 2 NP-трудна в сильном смысле для любых  $T_{\min} < T_{\max}$ . В тривиальном случае, когда  $T_{\min} = T_{\max}$ , эта задача разрешима за полиномиальное время.

В [13] предложен 2-приближённый полиномиальный алгоритм, временная сложность которого есть величина  $\mathcal{O}(N^2(MN + q))$ . Для случая задачи, в котором компоненты векторов целочисленны, а размерность  $q$  пространства фиксирована, в [15] обоснован точный псевдополиномиальный алгоритм, находящий решение задачи за время  $\mathcal{O}(N^3(MD)^q)$ .

Представляет интерес выяснение вопроса аппроксимируемости задачи 2. В частности, актуален вопрос о построении схемы FPTAS для какого-либо специального случая (подкласса) задачи. В настоящей работе такая схема обоснована.

Основной результат настоящей работы — приближённый алгоритм, который при фиксированной размерности  $q$  пространства реализует схему FPTAS и при заданной относительной погрешности  $\varepsilon$  позволяет находить  $(1 + \varepsilon)$ -приближённое решение задачи 2 за время  $\mathcal{O}(MN^3(1/\varepsilon)^{q/2})$ .

## 2. Геометрические основы алгоритма

Для построения алгоритма нам потребуется несколько базовых утверждений. Часть этих утверждений приводится без доказательства со ссыл-

ками на публикации [4, 5, 11, 16], где эти доказательства представлены.

**Лемма 1** [11]. Для произвольной точки  $x \in \mathbb{R}^q$  и конечного множества  $\mathcal{Z} \subset \mathbb{R}^q$  имеет место равенство

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

где  $\bar{z}$  — центроид множества  $\mathcal{Z}$ .

**Лемма 2** [4]. Пусть  $\mathcal{Z}$  — непустое конечное множество точек из  $\mathbb{R}^q$ , а  $\bar{z}$  — центроид множества  $\mathcal{Z}$ . Тогда если точка  $x \in \mathbb{R}^q$  удовлетворяет условиям  $\|x - \bar{z}\| \leq \|z - \bar{z}\|$ ,  $z \in \mathcal{Z}$ , то имеет место неравенство

$$\sum_{z \in \mathcal{Z}} \|z - x\|^2 \leq 2 \sum_{z \in \mathcal{Z}} \|z - \bar{z}\|^2.$$

Справедливость следующей леммы следует из результатов [5, 16].

**Лемма 3.** Пусть

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

где элементы набора  $\mathcal{M} = \{n_1, \dots, n_M\}$  удовлетворяют ограничениям (2). Тогда справедливы следующие утверждения:

- (i) для любого фиксированного подмножества  $\mathcal{M} \subseteq \mathcal{N}$  минимум функции (3) по  $x$  достигается в точке  $x = \bar{y}(\mathcal{M})$  и равен  $F(\mathcal{M})$ ;
- (ii) для любой фиксированной точки  $x \in \mathbb{R}^q$  минимум функции

$$S^x(\mathcal{M}) = \sum_{n \in \mathcal{M}} \|y_n - x\|^2 + \sum_{n \in \mathcal{N} \setminus \mathcal{M}} \|y_n\|^2, \quad \mathcal{M} \subseteq \mathcal{N},$$

по всем наборам  $\mathcal{M}$  фиксированной размерности  $M$  достигается на наборе  $\mathcal{M}^x$  номеров элементов  $\{y_i \mid i \in \mathcal{M}^x\}$  последовательности  $\mathcal{Y}$ , для которых функция

$$G^x(\mathcal{M}) = \sum_{n \in \mathcal{M}} \langle y_n, x \rangle, \quad \mathcal{M} \subseteq \mathcal{N}, \quad (4)$$

максимальна.

**Лемма 4.** Пусть  $\mathcal{M}^*$  — оптимальное решение задачи 2,  $x \in \mathbb{R}^q$  — произвольная фиксированная точка, а  $\mathcal{M}^x$  — набор, доставляющий минимум функции  $S^x(\mathcal{M})$ ,  $\mathcal{M} \subseteq \mathcal{N}$ , при ограничениях (2) на элементы набора  $\mathcal{M}$ . Тогда справедлива оценка

$$F(\mathcal{M}^x) \leq F(\mathcal{M}^*) + M\|x - \bar{y}(\mathcal{M}^*)\|^2, \quad (5)$$

где  $\bar{y}(\mathcal{M}^*) = \frac{1}{|\mathcal{M}^*|} \sum_{i \in \mathcal{M}^*} y_i$  — центроид оптимального решения.

**Доказательство.** Пусть  $\bar{y}(\mathcal{M}^x) = \frac{1}{|\mathcal{M}^x|} \sum_{i \in \mathcal{M}^x} y_i$  — центроид множества  $\{y_i \mid i \in \mathcal{M}^x\}$ . Тогда из определений (1), (3) и утверждения (i) леммы 3 следует оценка

$$F(\mathcal{M}^x) = S^{\bar{y}(\mathcal{M}^x)}(\mathcal{M}^x) \leq S^x(\mathcal{M}^x), \quad (6)$$

а из (ii) — неравенство

$$S^x(\mathcal{M}^x) \leq S^x(\mathcal{M}^*). \quad (7)$$

Применяя лемму 1 к точке  $x$  и множеству  $\{y_i \mid i \in \mathcal{M}^*\}$ , получим

$$\sum_{i \in \mathcal{M}^*} \|y_i - x\|^2 = \sum_{i \in \mathcal{M}^*} \|y_i - \bar{y}(\mathcal{M}^*)\|^2 + |\mathcal{M}^*| \cdot \|x - \bar{y}(\mathcal{M}^*)\|^2. \quad (8)$$

Наконец, объединяя (6)–(8), имеем оценку

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

Лемма 4 доказана.

Лемма 4 показывает, что оптимальное решение задачи 2 может быть аппроксимировано условно-оптимальным решением  $\mathcal{M}^x$  для некоторой специально построенной точки  $x$ . При этом для абсолютной ошибки аппроксимации в соответствии с (5) справедлива оценка

$$F(\mathcal{M}^x) - F(\mathcal{M}^*) \leq M\|x - \bar{y}(\mathcal{M}^*)\|^2.$$

**Лемма 5.** Пусть выполнены условия леммы 4, и

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

— точка из множества  $\{y_i \mid i \in \mathcal{M}^*\}$ , ближайшая к центроиду этого множества. Тогда для того чтобы при фиксированном  $\varepsilon > 0$  множество  $\mathcal{M}^x$  было  $(1 + \varepsilon)$ -приближённым решением задачи 2, достаточно, чтобы точка  $x$  удовлетворяла неравенству

$$\|x - \bar{y}(\mathcal{M}^*)\|^2 \leq \frac{\varepsilon}{2M} F(\mathcal{M}^t), \quad (9)$$

где  $\mathcal{M}^t$  — набор, доставляющий минимум функции  $S^t(\mathcal{M})$ ,  $\mathcal{M} \subseteq \mathcal{N}$ , при ограничениях (2) на элементы набора  $\mathcal{M}$ .

Доказательство. Пусть  $\bar{y}(\mathcal{M}^t) = \frac{1}{|\mathcal{M}^t|} \sum_{i \in \mathcal{M}^t} y_i$  — центроид множества  $\{y_i \mid i \in \mathcal{M}^t\}$ . Поскольку  $\mathcal{M}^t = \arg \min_{\mathcal{M}} S^t(\mathcal{M})$ , из определений (1), (3) и утверждения (i) леммы 3 следует оценка

$$F(\mathcal{M}^t) = S^{\bar{y}(\mathcal{M}^t)}(\mathcal{M}^t) \leq S^t(\mathcal{M}^t), \quad (10)$$

а из (ii) — неравенство

$$S^t(\mathcal{M}^t) \leq S^t(\mathcal{M}^*). \quad (11)$$

Так как множество  $\{y_i \mid i \in \mathcal{M}^*\}$  и точка  $t$  удовлетворяют условиям леммы 2, имеет место неравенство

$$\sum_{i \in \mathcal{M}^*} \|y_i - t\|^2 \leq 2 \sum_{i \in \mathcal{M}^*} \|y_i - \bar{y}(\mathcal{M}^*)\|^2,$$

следовательно,

$$\begin{aligned} S^t(\mathcal{M}^*) &= \sum_{i \in \mathcal{M}^*} \|y_i - t\|^2 + \sum_{i \in \mathcal{N} \setminus \mathcal{M}^*} \|y_i\|^2 \\ &\leq 2 \sum_{i \in \mathcal{M}^*} \|y_i - \bar{y}(\mathcal{M}^*)\|^2 + \sum_{i \in \mathcal{N} \setminus \mathcal{M}^*} \|y_i\|^2 \\ &\leq 2 \sum_{i \in \mathcal{M}^*} \|y_i - \bar{y}(\mathcal{M}^*)\|^2 + 2 \sum_{i \in \mathcal{N} \setminus \mathcal{M}^*} \|y_i\|^2 = 2F(\mathcal{M}^*). \end{aligned} \quad (12)$$

Объединив (9)–(12), получим

$$\|x - \bar{y}(\mathcal{M}^*)\|^2 \leq \frac{\varepsilon}{2M} F(\mathcal{M}^t) \leq \frac{\varepsilon}{2M} S^t(\mathcal{M}^t) \leq \frac{\varepsilon}{2M} S^t(\mathcal{M}^*) \leq \frac{\varepsilon}{M} F(\mathcal{M}^*).$$



Наконец, применив последние соотношения к правой части неравенства (5), приходим к оценке

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

из которой следует справедливость утверждения леммы. Лемма 5 доказана.

Лемма 5 показывает, насколько точка  $x$  должна быть близка к оптимальному центроиду, чтобы условно-оптимальное решение  $\mathcal{M}^x$  гарантировало получение  $(1 + \varepsilon)$ -приближённого решения задачи 2.

**Лемма 6.** Пусть выполнены условия леммы 5. Тогда для точки  $t = \arg \min_{y \in \{y_i \mid i \in \mathcal{M}^*\}} \|y - \bar{y}(\mathcal{M}^*)\|$  справедлива оценка

$$\|t - \bar{y}(\mathcal{M}^*)\|^2 \leq \frac{1}{M} F(\mathcal{M}^t). \quad (13)$$

**ДОКАЗАТЕЛЬСТВО.** Из определения точки  $t$  следует, что для любого  $i \in \mathcal{M}^*$  справедливо неравенство

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

Просуммировав обе части этого неравенства по  $i \in \mathcal{M}^*$ , получим

$$M\|t - \bar{y}(\mathcal{M}^*)\|^2 \leq \sum_{i \in \mathcal{M}^*} \|y_i - \bar{y}(\mathcal{M}^*)\|^2. \quad (14)$$

Поскольку  $\mathcal{M}^t$  — допустимое решение задачи 2, а  $\mathcal{M}^*$  — её оптимальное решение, имеем

$$F(\mathcal{M}^*) \leq F(\mathcal{M}^t). \quad (15)$$

Объединяя (14) и (15), выводим оценку

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

которая устанавливает справедливость неравенства (13). Лемма 6 доказана.

Лемма 6 даёт оценку сверху на расстояние от оптимального центра до ближайшей к нему точки из входного множества.

### 3. Схема FPTAS

Суть предлагаемого алгоритмического решения — схемы FPTAS — состоит в следующем. Для каждой точки входной последовательности строится область (куб) так, что хотя бы одна из полученных областей гарантированно включала неизвестный центроид одного из искоемых кластеров. По заданной на входе желаемой относительной погрешности решения строится сетка (решётка), дискретизирующая куб с равномерным по всем координатам шагом. Для каждого узла решётки с помощью схемы динамического программирования решается задача максимизации вспомогательной целевой функции и строится допустимый набор номеров элементов последовательности, доставляющий максимум этой функции. Сформированный допустимый набор объявляется претендентом на решение. В качестве окончательного решения выбирается тот из допустимых наборов-претендентов, который доставляет наименьшее значение целевой функции.

За исключением вспомогательной задачи и алгоритма её решения, все шаги, реализующие сформулированный подход, имеют геометрическую базу в виде утверждений, приведённых в разд. 2. Сформулируем вспомогательную задачу и обоснуем алгоритм её решения.

Для произвольной фиксированной точки  $x \in \mathbb{R}^q$  положим

$$g^x(n) = \langle y_n, x \rangle, \quad n \in \mathcal{N}, \quad (16)$$

где  $y_n$  —  $n$ -й элемент входной последовательности  $\mathcal{Y}$ . Тогда согласно (4) имеем

$$G^x(\mathcal{M}) = \sum_{n \in \mathcal{M}} g^x(n), \quad \mathcal{M} \subseteq \mathcal{N}, \quad (17)$$

где элементы набора  $\mathcal{M} = \{n_1, \dots, n_M\}$  удовлетворяют ограничениям (2), кроме того, в соответствии с утверждением (ii) леммы 3 имеет место равенство

$$\mathcal{M}^x = \arg \min_{\mathcal{M}} S^x(\mathcal{M}) = \arg \max_{\mathcal{M}} G^x(\mathcal{M}).$$

Рассмотрим следующую вспомогательную задачу.

**Задача 3.** Даны последовательность  $\mathcal{Y} = (y_1, \dots, y_N)$  точек из  $\mathbb{R}^q$ , точка  $x \in \mathbb{R}^q$ , натуральные числа  $T_{\min}$ ,  $T_{\max}$  и  $M > 1$ . Найти подмножество  $\mathcal{M} = \{n_1, \dots, n_M\} \subseteq \mathcal{N}$  номеров элементов последовательности  $\mathcal{Y}$ , доставляющее максимум целевой функции (17) при ограничениях (2) на элементы набора  $(n_1, \dots, n_M)$ .

В следующей лемме и следствии к ней приведена схема динамического программирования, гарантирующая отыскание оптимального решения  $\mathcal{M}^x$  задачи 3. Схема опирается на результаты из [12, 13] и приводится здесь ради полноты изложения.

**Лемма 7.** Для любого натурального  $M > 1$ ,  $(M - 1)T_{\min} \leq N - 1$ , и произвольной точки  $x \in \mathbb{R}^q$  оптимальное значение  $G_{\max}^x = \max_{\mathcal{M}} G^x(\mathcal{M})$  целевой функции задачи 3 находится по формуле

$$G_{\max}^x = \max_{n \in \omega_M} G_M^x(n), \quad (18)$$

а значения функции  $G_M^x(n)$ ,  $n \in \omega_M$ , вычисляются по рекуррентным формулам

$$G_m^x(n) = \begin{cases} g^x(n) & \text{при } n \in \omega_1, \ m = 1, \\ g^x(n) + \max_{j \in \gamma_{m-1}^-(n)} G_{m-1}^x(j) & \text{при } n \in \omega_m, \ m = 2, \dots, M, \end{cases} \quad (19)$$

где множества  $\omega_m$  и  $\gamma_{m-1}^-(n)$  задаются следующими формулами:

$$\omega_m = \{n \mid 1 + (m - 1)T_{\min} \leq n \leq N - (M - m)T_{\min}\}, \quad m = 1, \dots, M,$$

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

**Следствие 1.** Элементы  $n_1^x, \dots, n_M^x$  оптимального набора  $\mathcal{M}^x$  находятся по следующим рекуррентным формулам:

$$n_M^x = \arg \max_{n \in \omega_M} G_M^x(n), \quad (20)$$

$$n_{m-1}^x = \arg \max_{n \in \gamma_m^-(n_m^x)} G_m^x(n), \quad m = M, M - 1, \dots, 2. \quad (21)$$

Запишем алгоритм, реализующий приведённую схему, в пошаговом виде.

АЛГОРИТМ  $\mathcal{A}_1$

ВХОД: множество  $\mathcal{U}$ , точка  $x$ , числа  $T_{\min}$ ,  $T_{\max}$  и  $M$ .

ШАГ 1. Вычислим значения  $g^x(n)$ ,  $n \in \mathcal{N}$ , по формуле (16).

ШАГ 2. Используя рекуррентные формулы (19), вычислим значения  $G_m^x(n)$  для каждого  $n \in \omega_m$  и  $m = 1, \dots, M$ .

ШАГ 3. Найдём значение  $G_{\max}^x$  максимума целевой функции  $G^x$  по формуле (18) и оптимальный набор  $\mathcal{M}^x = (n_1^x, \dots, n_M^x)$  по формулам (20) и (21).

Выход.

В [13] установлено, что алгоритм  $\mathcal{A}_1$  находит оптимальное решение задачи 3 за время  $\mathcal{O}(N(M(T_{\max} - T_{\min} + 1) + q))$ . В этом выражении значение  $T_{\max} - T_{\min} + 1$  не превосходит  $N$ , поэтому время работы алгоритма оценивается величиной  $\mathcal{O}(N(MN + q))$ .

Для произвольной точки  $z \in \mathbb{R}^q$  и положительных чисел  $h, H$  определим множество точек

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

— многомерную кубическую равномерную по каждой координате решётку размера  $2H$  с расстоянием  $h$  между узлами с центром в точке  $z$ . Для числа узлов этой сетки справедлива оценка

$$|\mathcal{D}(z, h, H)| \leq (2\lfloor H/h \rfloor + 1)^q \leq (2H/h + 1)^q.$$

При этом для любой точки  $x \in \mathbb{R}^q$  такой, что  $\|z - x\| \leq H$ , расстояние до ближайшего узла сетки  $\mathcal{D}(z, h, H)$ , очевидно, не превосходит  $\frac{h\sqrt{q}}{2}$ .

Заметим, что лемма 6 (правая часть неравенства (13)) фактически определяет размер решётки, которая гарантированно содержит неизвестный центроид оптимального решения задачи 2, если только  $t$  — ближайшая к этому центроиду точка. Поэтому для размера кубической решётки положим

$$H(y) = \sqrt{\frac{1}{M} F(\mathcal{M}^y)}, \quad y \in \mathcal{Y}. \quad (22)$$

Кроме того, лемма 5 устанавливает условие на размер шага решётки, при котором среди её узлов найдётся элемент, близкий (в смысле гарантированной погрешности  $\varepsilon$ ) к центроиду оптимального решения. Тем самым для шага решётки положим

$$h(y, \varepsilon) = \sqrt{\frac{2\varepsilon}{qM} F(\mathcal{M}^y)}, \quad y \in \mathcal{Y}, \varepsilon > 0. \quad (23)$$

Сформулируем следующий алгоритм решения задачи 2.

АЛГОРИТМ  $\mathcal{A}$

ВХОД: множество  $\mathcal{Y}$ , числа  $T_{\min}$ ,  $T_{\max}$ ,  $M$  и  $\varepsilon$ .

Для каждой точки  $y \in \mathcal{Y}$  выполним шаги 1–5.

ШАГ 1. С помощью алгоритма  $\mathcal{A}_1$  найдём оптимальное решение  $\mathcal{M}^y$  задачи 3 при  $x = y$ .

ШАГ 2. Вычислим  $F(\mathcal{M}^y)$ ,  $h$  и  $H$  по формулам (1), (23) и (22).

ШАГ 3. Если  $F(\mathcal{M}^y) = 0$ , то множество  $\mathcal{M}^y$  объявим результатом работы алгоритма; выход.

ШАГ 4. Построим решётку  $\mathcal{D}(y, h, H)$ .

ШАГ 5. Для каждой точки  $d$  решётки  $\mathcal{D}(y, h, H)$  с помощью алгоритма  $\mathcal{A}_1$  построим оптимальное решение  $\mathcal{M}^d$  задачи 3 (при  $x = d$ ) и вычислим значение  $F(\mathcal{M}^d)$ .

ШАГ 6. В семействе множеств  $\{\mathcal{M}^d \mid d \in \mathcal{D}(y, h, H), y \in \mathcal{Y}\}$  в качестве решения выберем то множество  $\mathcal{M}^d$ , для которого значение  $F(\mathcal{M}^d)$  минимально.

Выход.

**Теорема 1.** Для любого фиксированного  $\varepsilon > 0$  алгоритм  $\mathcal{A}$  находит  $(1 + \varepsilon)$ -приближённое решение задачи 2 за время

$$\mathcal{O}(N^2(M(T_{\max} - T_{\min} + 1) + q)(\sqrt{2q/\varepsilon} + 1)^q).$$

ДОКАЗАТЕЛЬСТВО. Пусть  $t = \arg \min_{y \in \{y_i \mid i \in \mathcal{M}^*\}} \|y - \bar{y}(\mathcal{M}^*)\|$  — точка из множества  $\{y_i \mid i \in \mathcal{M}^*\}$ , ближайшая к центроиду этого множества. Если для этой точки на шаге 3 алгоритма выполнено равенство  $F(\mathcal{M}^t) = 0$ , то в этом случае множество  $\mathcal{M}^t$  является оптимальным решением задачи 2, так как для любого множества  $\mathcal{M} \subseteq \mathcal{N}$  справедливо неравенство  $F(\mathcal{M}) \geq 0$ .

Рассмотрим случай, когда  $F(\mathcal{M}^t) > 0$ . По лемме 6 для точки  $t$  выполнено (13). Из этого неравенства и (22) следует, что  $\|t - \bar{y}(\mathcal{M}^*)\| \leq H$ . Другими словами, центроид  $\bar{y}(\mathcal{M}^*)$  оптимального множества лежит в области сетки  $\mathcal{D}(t, h, H)$ .

Положим  $d^* = \arg \min_{d \in \mathcal{D}(t, h, H)} \|d - \bar{y}(\mathcal{M}^*)\|$ . Поскольку расстояние от  $\bar{y}(\mathcal{M}^*)$  до ближайшего узла  $d^*$  сетки  $\mathcal{D}(t, h, H)$  не превосходит  $h\sqrt{q}/2$ , имеем оценку

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

поэтому точка  $d^*$  удовлетворяет условиям леммы 5, следовательно, множество  $\mathcal{M}^{d^*}$  является  $(1 + \varepsilon)$ -приближённым решением задачи 2.

Оценим временную сложность алгоритма. Шаг 1 — решение вспомогательной задачи — требует  $\mathcal{O}(N(M(T_{\max} - T_{\min} + 1) + q))$  операций [13].

На шаге 2 требуется  $\mathcal{O}(qN)$  операций, а шаг 3 выполняется за  $\mathcal{O}(1)$  операций. Для построения сетки  $\mathcal{D}(y, h, H)$  потребуется  $\mathcal{O}(q|\mathcal{D}(y, h, H)|)$  операций на шаге 4. Построение каждого из  $|\mathcal{D}(y, h, H)|$  множеств  $\mathcal{M}^d$  на шаге 5 и вычисление значений  $F(\mathcal{M}^d)$  выполняется, как и на шаге 1, за  $\mathcal{O}(N(M(T_{\max} - T_{\min} + 1) + q))$  операций. В итоге, для каждой из  $N$  точек  $y \in \mathcal{Y}$  выполнение шагов 1–5 потребует

$$\mathcal{O}(N(M(T_{\max} - T_{\min} + 1) + q)|\mathcal{D}(y, h, H)|)$$

операций. Наконец, на шаге 6 для выбора наименьшего элемента требуется  $\mathcal{O}(\sum_{y \in \mathcal{Y}} |\mathcal{D}(y, h, H)|)$  операций.

Остаётся заметить, что для мощности решётки  $\mathcal{D}(y, h, H)$  справедлива оценка

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

поэтому временная сложность алгоритма равна

$$\mathcal{O}(N^2(M(T_{\max} - T_{\min} + 1) + q)(\sqrt{2q/\varepsilon} + 1)^q).$$

Теорема 1 доказана.

Покажем, что в случае фиксированной размерности  $q$  пространства алгоритм  $\mathcal{A}$  реализует схему FPTAS. Действительно, если  $\varepsilon \in (0, 2q]$ , то

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

Следовательно, поскольку величина  $T_{\max} - T_{\min} + 1$  не превосходит  $N$ , при указанных условиях время работы алгоритма оценивается величиной  $\mathcal{O}(MN^3(1/\varepsilon)^{q/2})$ , которая ограничена полиномом как от размера входа задачи, так и от  $1/\varepsilon$ . Таким образом, предложенный алгоритм реализует схему FPTAS.

**Замечание.** С помощью алгоритма  $\mathcal{A}$  можно построить алгоритм решения задачи Minimum Sum-of-Squares 2-Clustering problem on sequence with given center of one cluster, в которой размеры подпоследовательностей являются оптимизируемыми переменными [14]. Действительно, для этого достаточно с помощью алгоритма  $\mathcal{A}$  найти решение задачи 2 для каждого допустимого размера  $M$  подпоследовательности с неизвестным центроидом, а затем среди  $\mathcal{O}(N)$  найденных решений выбрать наилучшее. Временная сложность такого алгоритма равна  $\mathcal{O}(N^3(M(T_{\max} - T_{\min} + 1) + q)(\sqrt{2q/\varepsilon} + 1)^q)$ . При фиксированной размерности пространства время работы этого алгоритма оценивается величиной  $\mathcal{O}(MN^4(1/\varepsilon)^{q/2})$  и он реализует схему FPTAS.

### Заключение

В работе обоснован приближённый алгоритм для одной из слабоизученных NP-трудных в сильном смысле задач разбиения конечной последовательности точек евклидова пространства на два кластера. Предложенный алгоритм реализует схему FPTAS в случае фиксированной размерности пространства.

Показано, что с помощью обоснованного алгоритма можно построить приближённый алгоритм для варианта задачи, в котором мощности кластеров не являются частью входа (неизвестны). В случае, когда размерность пространства фиксирована, этот алгоритм также реализует схему FPTAS. Однако для этого варианта задачи представляет интерес построение иного — менее трудоёмкого — алгоритма, реализующего схему FPTAS, для этого же случая задачи без перебора по всем допустимым размерам искоемых кластеров.

На наш взгляд, представленная в работе техника решения задачи будет полезной при построении эффективных приближённых алгоритмов с оценками точности для решения других (близких в постановочном плане) труднорешаемых задач, возникающих, в частности, в теории приближения, статистическом анализе временных рядов, анализе данных и распознавании образов, а также в естественнонаучных и технических приложениях.

Рассмотренная задача относится к числу слабоизученных в алгоритмическом плане. Поэтому продолжение исследования вопросов её аппроксимируемости, в частности, обоснование алгоритмов другого типа — асимптотически точных, рандомизированных, приближённых полиномиальных схем (PTAS) — для её решения представляется делом ближайшей перспективы.

### ЛИТЕРАТУРА

1. Бабурин А. Е., Гимади Э. Х., Глебов Н. И., Пяткин А. В. Задача отыскания подмножества векторов с максимальным суммарным весом // Дискрет. анализ и исслед. операций. Сер. 2. 2007. Т. 14, № 1. С. 32–42.
2. Гимади Э. Х., Глазков Ю. В., Рыков И. А. О двух задачах выбора подмножества векторов с целочисленными координатами в евклидовом пространстве с максимальной нормой суммы размерности // Дискрет. анализ и исслед. опер. 2008. Т. 15, № 4. С. 30–43.
3. Гимади Э. Х., Пяткин А. В., Рыков И. А. О полиномиальной разрешимости некоторых задач выбора подмножеств векторов в евклидовом пространстве фиксированной размерности // Дискрет. анализ и исслед. опер. 2008. Т. 15, № 6. С. 11–19.

4. Долгушев А. В., Кельманов А. В. Приближённый алгоритм решения одной задачи кластерного анализа // Дискрет. анализ и исслед. опер. 2011. Т. 18, № 2. С. 29–40.
5. Долгушев А. В., Кельманов А. В., Шенмайер В. В. Полиномиальная аппроксимационная схема для одной задачи разбиения конечного множества на два кластера // Тр. Ин-та математики и механики УрО РАН. 2015. Т. 21, № 3. С. 100–109.
6. Кельманов А. В. Проблема off-line обнаружения повторяющегося фрагмента в числовой последовательности // Тр. Ин-та математики и механики УрО РАН. 2008. Т. 14, № 2. С. 81–88.
7. Кельманов А. В. О сложности некоторых задач анализа данных // Журн. вычисл. математики и мат. физики. 2010. Т. 50, № 11. С. 2045–2051.
8. Кельманов А. В. О сложности некоторых задач кластерного анализа // Журн. вычисл. математики и мат. физики. 2011. Т. 51, № 11. С. 2106–2112.
9. Кельманов А. В., Пяткин А. В. О сложности некоторых задач поиска подмножеств векторов и кластерного анализа // Журн. вычисл. математики и мат. физики. 2009. Т. 49, № 11. С. 2059–2067.
10. Кельманов А. В., Пяткин А. В. О сложности некоторых задач кластерного анализа векторных последовательностей // Дискрет. анализ и исслед. операций. 2013. Т. 20, № 2. С. 47–57.
11. Кельманов А. В., Романченко С. М. FPTAS для одной задачи поиска подмножества векторов // Дискрет. анализ и исслед. операций. 2014. Т. 21, № 3. С. 41–52.
12. Кельманов А. В., Хамидуллин С. А. Апостериорное обнаружение заданного числа одинаковых подпоследовательностей в квазипериодической последовательности // Журн. вычисл. математики и мат. физики. 2001. Т. 41, № 5. С. 807–820.
13. Кельманов А. В., Хамидуллин С. А. Приближённый полиномиальный алгоритм для одной задачи разбиения последовательности // Дискрет. анализ и исслед. операций. 2014. Т. 21, № 1. С. 53–66.
14. Кельманов А. В., Хамидуллин С. А. Приближённый полиномиальный алгоритм для одной задачи бикластеризации последовательности // Журн. вычисл. математики и мат. физики. 2015. Т. 55, № 6. С. 1076–1085.
15. Кельманов А. В., Хамидуллин С. А., Хандеев В. И. Точный псевдополиномиальный алгоритм для одной задачи бикластеризации последовательности // Тез. докл. XV Всеросс. конф. «Математическое программирование и приложения» (Екатеринбург, 2–6 марта 2015 г.). Екатеринбург: Ин-т математики и механики УрО РАН, 2015. С. 139–140.
16. Кельманов А. В., Хандеев В. И. Рандомизированный алгоритм для одной задачи двухкластерного разбиения множества векторов // Журн. вычисл. математики и мат. физики. 2015. Т. 55, № 2. С. 335–344.
17. Кельманов А. В., Хандеев В. И. Точный псевдополиномиальный ал-



- горитм для одной задачи двухкластерного разбиения множества векторов // Дискрет. анализ и исслед. операций. 2015. Т. 22, № 3. С. 36–48.
18. **Кельманов А. В., Хандеев В. И.** Полностью полиномиальная аппроксимационная схема для специального случая одной квадратичной евклидовой задачи 2-кластеризации // Журн. вычисл. математики и мат. физики. 2016. Т. 56, № 2. С. 145–153.
  19. **Aloise D., Deshpande A., Hansen P., Popat P.** NP-hardness of Euclidean sum-of-squares clustering // Machine Learning. 2009. Vol. 75, No. 2. P. 245–248.
  20. **Bishop M. C.** Pattern recognition and machine learning. New York: Springer-Verl., 2006. 738 p.
  21. **Carter J. A., Agol E., et al.** Kepler-36: a pair of planets with neighboring orbits and dissimilar densities // Science. 2012. Vol. 337, No. 6094. P. 556–559.
  22. **Flach P.** Machine learning: the art and science of algorithms that make sense of data. New York: Cambridge Univ. Press, 2012. 396 p.
  23. **Gimadi E. Kh., Kel'manov A. V., Kel'manova M. A., Khamidullin S. A.** A posteriori detecting a quasiperiodic fragment in a numerical sequence // Pattern Recognit. Image Anal. 2008. Vol. 18, No. 1. P. 30–42.
  24. **Jain A. K.** Data clustering: 50 years beyond  $k$ -means // Pattern Recognit. Lett. 2010. Vol. 31, No. 8. P. 651–666.
  25. **James G., Witten D., Hastie T., Tibshirani R.** An introduction to statistical learning. New York: Springer-Verl., 2013. 426 p.
  26. **Kel'manov A. V., Jeon B.** A posteriori joint detection and discrimination of pulses in a quasiperiodic pulse train // IEEE Trans. Signal Process. 2004. Vol. 52, No. 3. P. 645–656.
  27. **Steger C., Ulrich M., Wiedemann C.** Machine vision algorithms and applications. Berlin: Wiley-VCH, 2007. 380 p.

Кельманов Александр Васильевич,  
Хамидуллин Сергей Асгадуллович,  
Хандеев Владимир Ильич

Статья поступила  
15 сентября 2015 г.  
Исправленный вариант —  
12 января 2016 г.

FULLY POLYNOMIAL-TIME APPROXIMATION SCHEME  
FOR A SEQUENCE 2-CLUSTERING PROBLEMA. V. Kel'manov<sup>1,2</sup>, S. A. Khamidullin<sup>1</sup>, V. I. Khandeev<sup>1</sup><sup>1</sup>Sobolev Institute of Mathematics,

4 Koptug Ave., 630090 Novosibirsk, Russia

<sup>2</sup>Novosibirsk State University,

2 Pirogova St., 630090 Novosibirsk, Russia

e-mail: kelm@math.nsc.ru, kham@math.nsc.ru, khandeev@math.nsc.ru

**Abstract.** We consider a strongly NP-hard problem of partitioning a finite sequence of points in Euclidean space into two clusters minimizing the sum over both clusters of intra-cluster sum of squared distances from the clusters elements to their centers. The sizes of the clusters are fixed. The centroid of the first cluster is defined as the mean value of all vectors in the cluster, and the center of the second one is given in advance and is equal to 0. Additionally, the partition must satisfy the restriction that for all vectors in the first cluster the difference between the indices of two consequent points from this cluster is bounded from below and above by some given constants. We present a fully polynomial-time approximation scheme for the case of fixed space dimension. Bibliogr. 27.

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

## REFERENCES

1. A. E. Baburin, E. Kh. Gimadi, N. I. Glebov, and A. V. Pyatkin, The problem of finding a subset of vectors with the maximum total weight, *Diskretn. Anal. Issled. Oper., Ser. 2*, **14**, No. 1, 32–42, 2007. Translated in *J. Appl. Ind. Math.*, **2**, No. 1, 32–38, 2008.
2. E. Kh. Gimadi, Yu. V. Glazkov, and I. A. Rykov, On two problems of choosing some subset of vectors with integer coordinates that has maximum norm of the sum of elements in Euclidean space, *Diskretn. Anal. Issled. Oper.*, **15**, No. 4, 30–43, 2008. Translated in *J. Appl. Ind. Math.*, **3**, No. 3, 343–352, 2009.
3. E. Kh. Gimadi, A. V. Pyatkin, and I. A. Rykov, On polynomial solvability of some problems of a vector subset choice in a Euclidean space of fixed dimension, *Diskretn. Anal. Issled. Oper.*, **15**, No. 6, 11–19, 2008. Translated in *J. Appl. Ind. Math.*, **4**, No. 1, 48–53, 2010.

4. **A. V. Dolgushev** and **A. V. Kel'manov**, An approximation algorithm for solving a problem of cluster analysis, *Diskretn. Anal. Issled. Oper.*, **18**, No. 2, 29–40, 2011. Translated in *J. Appl. Ind. Math.*, **5**, No. 4, 551–558, 2011.
5. **A. V. Dolgushev**, **A. V. Kel'manov**, and **V. V. Shenmaier**, Polynomial-time approximation scheme for a problem of partitioning a finite set into two clusters, *Tr. Inst. Mat. Mekh.*, **21**, No. 3, 100–109, 2015.
6. **A. V. Kel'manov**, Off-line detection of a quasi-periodically recurring fragment in a numerical sequence, *Tr. Inst. Mat. Mekh.*, **14**, No. 2, 81–88, 2008. Translated in *Proc. Steklov Inst. Math.*, **263**, Suppl. 2, S84–S92, 2008.
7. **A. V. Kel'manov**, On the complexity of some data analysis problems, *Zh. Vychisl. Mat. Mat. Fiz.*, **50**, No. 11, 2045–2051, 2010. Translated in *Comput. Math. Math. Phys.*, **50**, No. 11, 1941–1947, 2010.
8. **A. V. Kel'manov**, On the complexity of some cluster analysis problems, *Zh. Vychisl. Mat. Mat. Fiz.*, **51**, No. 11, 2106–2112, 2011. Translated in *Comput. Math. Math. Phys.*, **51**, No. 11, 1983–1988, 2011.
9. **A. V. Kel'manov** and **A. V. Pyatkin**, Complexity of certain problems of searching for subsets of vectors and cluster analysis, *Zh. Vychisl. Mat. Mat. Fiz.*, **49**, No. 11, 2059–2067, 2009. Translated in *Comput. Math. Math. Phys.*, **49**, No. 11, 1966–1971, 2009.
10. **A. V. Kel'manov** and **A. V. Pyatkin**, On complexity of some problems of cluster analysis of vector sequences, *Diskretn. Anal. Issled. Oper.*, **20**, No. 2, 47–57, 2013. Translated in *J. Appl. Ind. Math.*, **7**, No. 3, 363–369, 2013.
11. **A. V. Kel'manov** and **S. M. Romanchenko**, An FPTAS for a vector subset search problem, *Diskretn. Anal. Issled. Oper.*, **21**, No. 3, 41–52, 2014. Translated in *J. Appl. Ind. Math.*, **8**, No. 3, 329–336, 2014.
12. **A. V. Kel'manov** and **S. A. Khamidullin**, Posterior detection of a given number of identical subsequences in a quasi-periodic sequence, *Zh. Vychisl. Mat. Mat. Fiz.*, **41**, No. 5, 807–820, 2001. Translated in *Comput. Math. Math. Phys.*, **41**, No. 5, 762–774, 2001.
13. **A. V. Kel'manov** and **S. A. Khamidullin**, An approximating polynomial algorithm for a sequence partitioning problem, *Diskretn. Anal. Issled. Oper.*, **21**, No. 1, 53–66, 2014. Translated in *J. Appl. Ind. Math.*, **8**, No. 2, 236–244, 2014.
14. **A. V. Kel'manov** and **S. A. Khamidullin**, An approximation polynomial-time algorithm for a sequence bi-clustering problem, *Zh. Vychisl. Mat. Mat. Fiz.*, **55**, No. 6, 1076–1085, 2015. Translated in *Comput. Math. Math. Phys.*, **55**, No. 6, 1068–1076, 2015.
15. **A. V. Kel'manov**, **S. A. Khamidullin**, and **V. I. Khandeev**, An exact pseudopolynomial algorithm for a sequence bi-clustering problem, in *Tezisy dokladov XV Vserossiiskoy konferentsii "Matematicheskoe programmirovaniye i prilozheniya"* (Abstr. XV All-Russ. Conf. "Mathematical Programming and Applications"), *Ekaterinburg, Russia, Mar. 2–6, 2015*, pp. 139–140, Inst. Mat. Mekh. UrO RAN, Ekaterinburg, 2015.

16. **A. V. Kel'manov** and **V. I. Khandeev**, A randomized algorithm for two-cluster partition of a set of vectors, *Zh. Vychisl. Mat. Mat. Fiz.*, **55**, No. 2, 335–344, 2015. Translated in *Comput. Math. Math. Phys.*, **55**, No. 2, 330–339, 2015.
17. **A. V. Kel'manov** and **V. I. Khandeev**, An exact pseudopolynomial algorithm for a problem of the two-cluster partitioning of a set of vectors, *Diskretn. Anal. Issled. Oper.*, **22**, No. 3, 36–48, 2015. Translated in *J. Appl. Ind. Math.*, **9**, No. 4, 497–502, 2015.
18. **A. V. Kel'manov** and **V. I. Khandeev**, Fully polynomial-time approximation scheme for special case of a quadratic Euclidean 2-clustering problem, *Zh. Vychisl. Mat. Mat. Fiz.*, **56**, No. 2, 145–153, 2016.
19. **D. Aloise**, **A. Deshpande**, **P. Hansen**, and **P. Popat**, NP-hardness of Euclidean sum-of-squares clustering, *Mach. Learn.*, **75**, No. 2, 245–248, 2009.
20. **C. M. Bishop**, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.
21. **J. A. Carter** [et al.], Kepler-36: A pair of planets with neighboring orbits and dissimilar densities, *Science*, **337**, No. 6094, 556–559, 2012.
22. **P. Flach**, *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*, Cambridge Univ. Press, New York, 2012.
23. **E. Kh. Gimadi**, **A. V. Kel'manov**, **M. A. Kel'manova**, and **S. A. Khamidullin**, A posteriori detecting a quasiperiodic fragment in a numerical sequence, *Pattern Recognit. Image Anal.*, **18**, No. 1, 30–42, 2008.
24. **A. K. Jain**, Data clustering: 50 years beyond K-means, *Pattern Recognit. Lett.*, **31**, No. 8, 651–666, 2010.
25. **G. James**, **D. Witten**, **T. Hastie**, and **D. Tibshirani**, *An Introduction to Statistical Learning: With Applications in R*, Springer, New York, 2013.
26. **A. V. Kel'manov** and **B. Jeon**, A posteriori joint detection and discrimination of pulses in a quasiperiodic pulse train, *IEEE Trans. Signal Process.*, **52**, No. 3, 645–656, 2004.
27. **C. Steger**, **M. Ulrich**, and **C. Wiedemann**, *Machine Vision Algorithms and Applications*, Wiley-VCH, Berlin, 2007.

Alexander V. Kel'manov,  
Sergey A. Khamidullin,  
Vladimir I. Khandeev

Received  
15 September 2015  
Revised  
12 January 2016