Efficient Approximation Algorithms for Some NP-hard Problems of Partitioning a Set and a Sequence

Alexander Kel'manov

Sobolev Institute of Mathematics
Siberian Branch of the Russian Academy of Sciences,
Novosibirsk State University,
Novosibirsk, Russia

XIIIth International Conference
«Problems of complex systems' optimization»
Novosibirsk, September 18-23, 2017

Outline

1. Introduction (subject, goal and motivation of the investigation)
2. Some quadratic Euclidean clustering problems:
   1. Problem complexity and algorithms with performance garantees
   2. Successful techniques for these problems
3. Conclusion, open problems

Introduction

The subject of investigation

is some quadratic Euclidean clustering problems.

The goal

is short review of some new results on the complexity of these problems, and on efficient algorithms with performance guarantees for their solutions.

The research is motivated by

poor research record on the problems and their relevance to many applications, in particular, to
(1) Geometric, approximation and statistical problems;

(2) Data clustering, Data mining, Machine learning, Big data;

(3) Applied problems in technical and medical diagnostics, remote monitoring, biometrics, bioinformatics, econometrics, criminology, processing of experimental data, processing and recognition of signals, etc.

List of considered clustering problems

1. Subset of points with the largest cardinality under a constraint on the total quadratic variation
(Подмножество точек наибольшей мощности при ограничении на суммарный квадратичный разброс)
(**Ageev, Kel'manov, Pyatkin, Khamidullin, Shenmaier**).

2. Finding a family of disjoint subsets
(Поиск семейства непересекающихся подмножеств)
(**Galashov, Kel'manov**).

3. Cardinality-weighted variance-based 2-clustering with given center
(Мощностно-взвешенная 2-кластеризация при заданном центре одного кластера)
(**Kel'manov, Motkova**).

4. Weighted variance-based 2-clustering with given center
(Взвешенная 2-кластеризация при заданном центре одного кластера)
(**Kel'manov, Motkova, Shenmaier**).

List of of considered clustering problems

5. Finding a subsequence in a sequence
(Поиск подпоследовательности заданного размера)
(**Kel'manov, Romanchenko, Khamidullin**)

6. Minimum sum-of-squares 2**-clustering problem on sequence with given center of one cluster**
(2-кластеризация последовательности при заданном центре одного кластера)
(**Kel'manov, Khamidullin, Khandeev**)

7. Partitioning a sequence into clusters with given center of one cluster and cluster cardinalities
(Кластеризация последовательности при при заданных центре одного кластера и мощностях кластеров)
(**Kel'manov, Mikhailova, Khamidullin, Khandeev**)

8. Partitioning a sequence into clusters
(Кластеризация последовательности при заданном центре одного кластера)
(**Kel'manov, Mikhailova, Khamidullin, Khandeev**)

One of the well-known (Fisher, 1958) data analysis problems is the MSSC (Minimum Sum-of-Squares Clustering) problem which is strongly NP-hard (Aloise D., Deshpande A., Hansen P., Popat P., 2009) and has the following formulation.

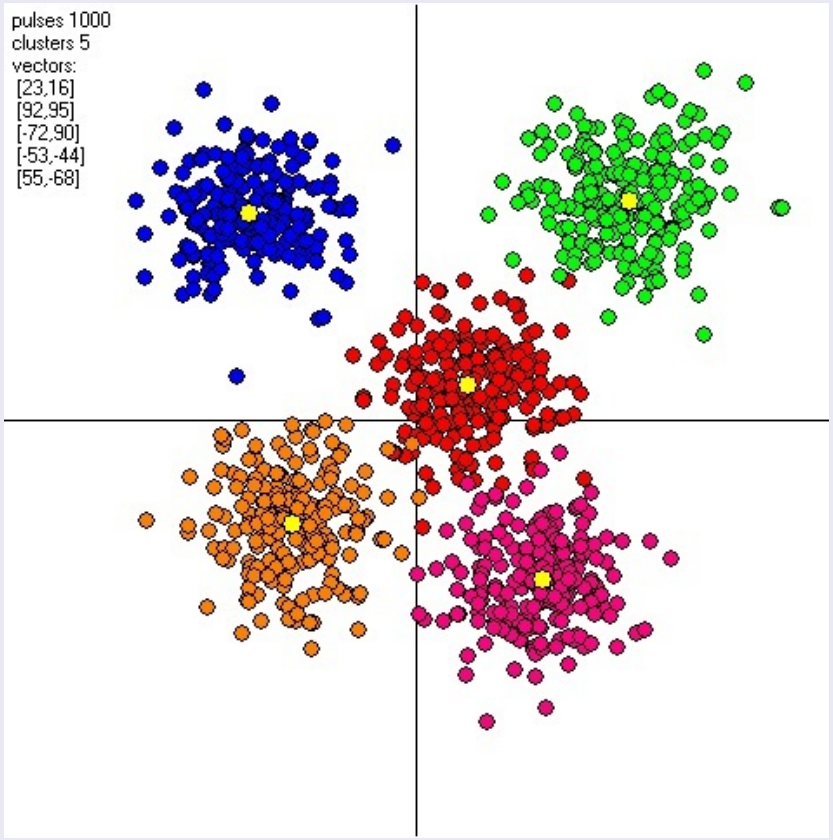
MSSC

Given a set $\mathcal{Y} = \{y_1, \ldots, y_N\}$ of points from $\mathbb{R}^q$ and positive integer $J > 1$.
Find a partition of $\mathcal{Y}$ into non-empty clusters $\mathcal{C}_1, \ldots, \mathcal{C}_J$ such that

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

where $\overline{y}(\mathcal{C}_j) = \frac{1}{|\mathcal{C}_j|} \sum_{y \in \mathcal{C}_j} y$ is the centroid (geometrical center) of $\mathcal{C}_j$.

MSSC problem. Two-dimensional example



1. Subset of points with the largest cardinality under a constraint on the total quadratic variation

Problem 1 (Subset of points with the largest cardinality under a constraint on the total quadratic variation)

Given a set $\mathcal{Y} = \{y_1, \ldots, y_N\}$ of points from $\mathbb{R}^q$ and number $\alpha \in (0, 1)$.
Find a subset $\mathcal{C} \subset \mathcal{Y}$ with the largest cardinality such that

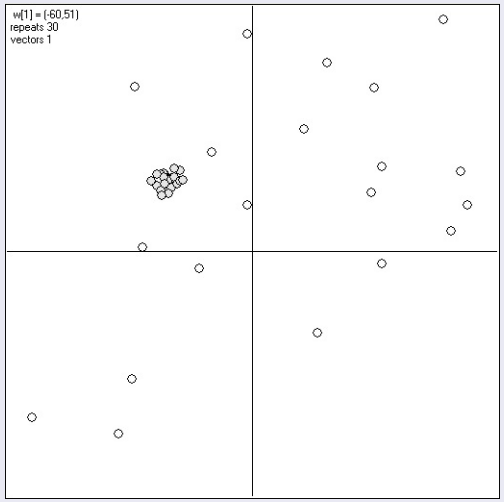
$$\sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 \leq \alpha \sum_{y \in \mathcal{Y}} \|y - \overline{y}(\mathcal{Y})\|^2$$

where $\overline{y}(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum_{y \in \mathcal{C}} y$ is the centroid (the geometrical center) of the subset $\mathcal{C}$, and $\overline{y}(\mathcal{Y}) = \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} y$ is the centroid of the input set.

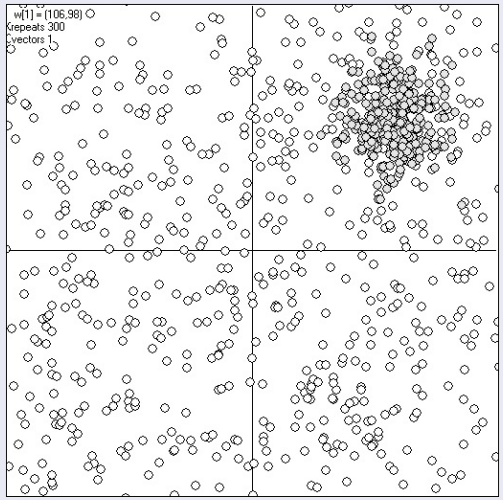
1. Subset of points with the largest cardinality under a constraint on the total quadratic variation

Two-dimensional examples of the input sets

Example 1



Example 2



The problem has a simple interpretation, namely, searching for the largest by cardinality subset $\mathcal{C}$ of points, whose total quadratic variation from the unknown centroid $\overline{y}(\mathcal{C})$ doesn't exceed the total quadratic variation of the input set $\mathcal{Y}$ from its centroid $\overline{y}(\mathcal{Y})$ multiplied by α.

1. Subset of points with the largest cardinality under a constraint on the total quadratic variation

Currently, there are no available algorithmic results for Problem 1. Some results are known for the closest problem related to Problem 1, that is for

M-Variance problem

Given: a set $\mathcal{Y} = \{y_1, \ldots, y_N\}$ of points from $\mathbb{R}^q$ and positive integer number $M > 1$.
Find: a subset $\mathcal{C} \subset \mathcal{Y}$ of cardinality M such that

$$\sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 \longrightarrow \min$$

Known results for M-Variance problem

1. The strong NP-hardness of the problem (Kel'manov and Pyatkin, 2010).

2. An exact algorithm, $\mathcal{O}(qN^{q+1})$, Aggarwal, Imai, Katoh, Suri (1991), Shenmaier, 2016.

1. Subset of points with the largest cardinality under a constraint on the total quadratic variation

Known results for M-Variance problem

3. A 2-approximation polynomial-time algorithm, $\mathcal{O}(qN^2)$, Kel'manov, Romanchenko (2012).

4. An exact algorithm for the integer-valued variant of the data input. In the case of fixed space dimension the algorithm has $\mathcal{O}(N(MB)^q)$ running time, where B is the maximum absolute value of the coordinates of the input points, Kel'manov, Romanchenko (2012).

5. PTAS of complexity $\mathcal{O}(qN^{2/\varepsilon+1}(9/\varepsilon)^{3/\varepsilon})$, where ε is a guaranteed relative error, Shenmaier (2012).

6. $(1+\varepsilon)$-Approximation algorithm, which implements an FPTAS with $\mathcal{O}(N^2(M/\varepsilon)^q)$-time complexity, in the case of fixed space dimension, Kel'manov, Romanchenko (2014).

1. Subset of points with the largest cardinality under a constraint on the total quadratic variation

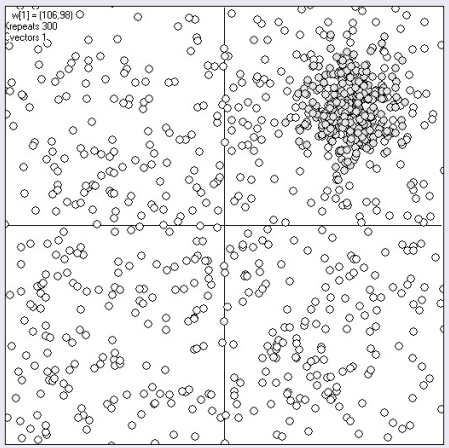
New results (Ageev, Kel'manov, Pyatkin, Khamidullin, Shenmaier, 2017)

1. Problem 1 is strongly NP-hard.

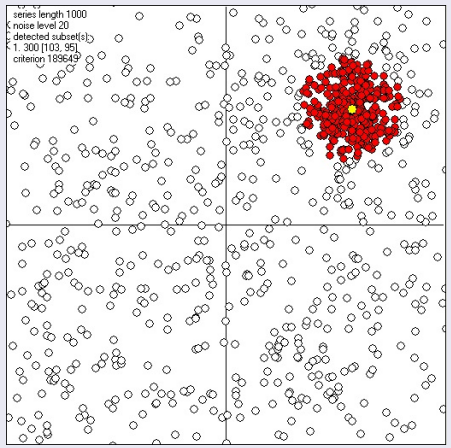
2. 1/2-approximation polynomial-time algorithm with running time

$$\mathcal{O}(N^2(q + \log N)).$$

example of an input set



found subset



1. Subset of points with the largest cardinality under a constraint on the total quadratic variation

The idea of approximation algorithm

Input: set $\mathcal{Y}$ and number $\alpha \in (0,1)$.
Step 1. Compute the value $A = \alpha \sum_{y \in \mathcal{Y}} \|y - \overline{y}(\mathcal{Y})\|^2$.

For each point $y \in \mathcal{Y}$ perform steps 2 and 3.
Step 2. Compute the distances from the point y to all points in $\mathcal{Y}$ and sort the set $\mathcal{Y}$ in the nondecreasing order according to these distances. Denote this sequence by $y_1, \ldots, y_N$.

Step 3. Find the subsequence $y_1, \ldots y_M$ of maximum length such that

$$\sum_{i=1}^{M} \|y - y_i\|^2 \leq A$$

Define the subset $\mathcal{C}^y = \{y_1 \ldots, y_M\}$.

Step 4. In the family $\{\mathcal{C}^y \,|\, y \in \mathcal{Y}\}$ of admissible subsets constructed in step 3 choose as the output $\mathcal{C}_A$ any subset $\mathcal{C}^y$ of the largest cardinality.

Output: subset $\mathcal{C}_A$.

2. Finding a family of disjoint subsets

In MSSC problem the cardinalities of the required clusters are unknown and we have to find a partition of the set. But in the considered problem we have to find a family of subsets which union might not cover input set and the cardinalities of the required clusters are given.

Problem 2 (Finding a family of disjoint subsets)

Given a set $\mathcal{Y} = \{y_1, \ldots, y_N\}$ of points from $\mathbb{R}^q$ and some positive integers $M_1, \ldots, M_J$.
Find a family $\{\mathcal{C}_1, \ldots, \mathcal{C}_J\}$ of disjoint subsets of $\mathcal{Y}$ such that

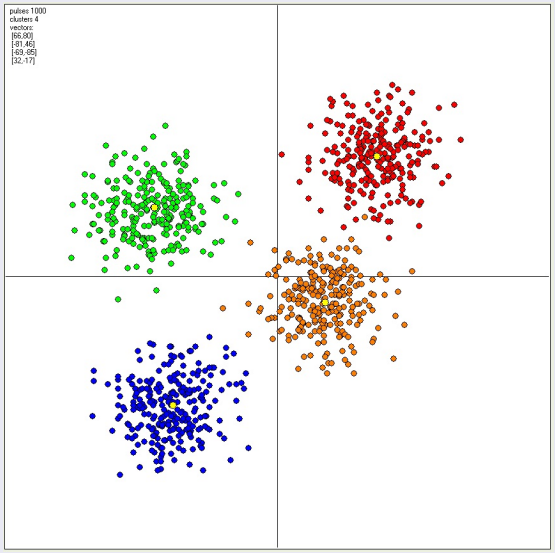
$$\sum_{j=1}^{J} \sum_{y \in \mathcal{C}_j} \|y - \overline{y}(\mathcal{C}_j)\|^2 \to \min,$$

where $\overline{y}(\mathcal{C}_j) = \frac{1}{|\mathcal{C}_j|} \sum\limits_{y \in \mathcal{C}_j} y$ is the centroid (geometrical center) of the subset $\mathcal{C}_j$, under constraints $|\mathcal{C}_j| = M_j, j = 1, \ldots, J$, on the cardinalities of the required subsets.

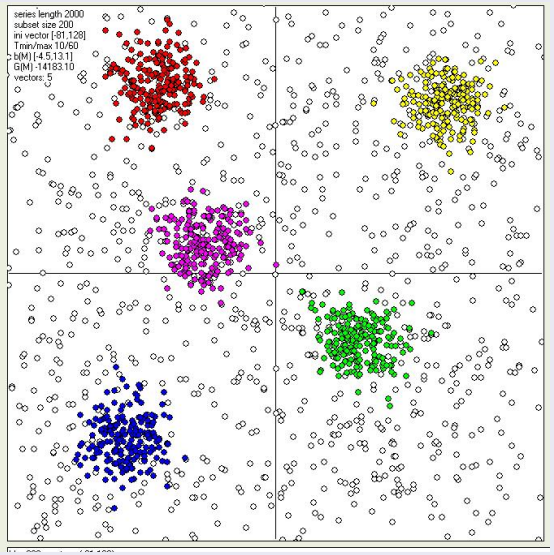
2. Finding a family of disjoint subsets

Two-dimensional examples

Minimum Sum-of-Squares Clustering



Searching a family of disjoint subsets



Known results

1. The strong NP-hardness of the problem is implied from the results of Kel'manov and Pyatkin (it was proved that the special case of the problem when $J = 1$ is strongly NP-hard, 2011).

2. A 2-approximation algorithm, $\mathcal{O}(N^2(N^{J+1} + q))$,
Galashov, Kel'manov, 2014.

Some results were obtained for the **case** of the problem when $J = 1$.

3. An exact algorithm, $\mathcal{O}(qN^{q+1})$,
Aggarwal, Imai, Katoh, Suri (1991), Shenmaier, 2016.

4. A 2-approximation polynomial-time algorithm, $\mathcal{O}(qN^2)$,
Kel'manov, Romanchenko (2012).

Known results

5. An exact algorithm for the integer-valued variant of the data input. In the case of fixed space dimension the algorithm has $\mathcal{O}(N(MB)^q)$ running time, where B is the maximum absolute value of the coordinates of the input points,
Kel'manov, Romanchenko (2012).

6. PTAS of complexity $\mathcal{O}(qN^{2/\varepsilon+1}(9/\varepsilon)^{3/\varepsilon})$, where ε is a guaranteed relative error,
Shenmaier (2012).

7. $(1+\varepsilon)$-Approximation algorithm, which implements an FPTAS with $\mathcal{O}(N^2(M/\varepsilon)^q)$-time complexity in the case of fixed space dimension,
Kel'manov, Romanchenko (2014).

New result (Galashov, Kel'manov, 2016)

An **exact algorithm** for the case of integer components of the input points with

$$\mathcal{O}(N(N^2 + qJ)(2MB + 1)^{qJ} + (J - 1)\lg N)$$

running time, where B is the maximum absolute value of the coordinates of the input points and M is the least common multiple for the numbers $M_1, \ldots, M_J$.

In the case of the fixed dimension q of the space and of the fixed number J of required subsets, the proposed algorithm is pseudopolynomial and its time complexity is bounded by

$$\mathcal{O}(N^3(MB)^{qJ}).$$

The idea of algorithm which implements a grid approach

1. Find the least common multiple M for the numbers $M_1, \ldots, M_J$ and the maximum absolute value B of the coordinates of the input points. Construct the multi-dimensional grid $\mathcal{D}$ using formula

$$\mathcal{D} = \{z \in \mathbb{R}^q | \ (z)^k = \frac{1}{M}(v)^k, (v)^k \in \mathbb{Z}, |(v)^k| \leq B, k = 1, \ldots, q\},$$

2. For every tuple $d = (d_1, \ldots, d_J) \in \mathcal{D}^J$ of J points from the grid find and save the exact solution $\{\mathcal{B}_1(d), \ldots, \mathcal{B}_J(d)\}$ of the auxiliary problem

$$G^d(\mathcal{B}_1, \ldots, \mathcal{B}_J) = \sum_{j=1}^{J} \sum_{y \in \mathcal{B}_j} \|y - d_j\|^2 \to \min .$$

Save the value of G^d.

3. Take as a solution the family $\{\mathcal{C}_1^A, \ldots, \mathcal{C}_J^A\}$ of the subsets with minimum value of the function G^d, $d \in \mathcal{D}^J$.

3. Cardinality-weighted variance-based 2-clustering with given center

Problem 3 (Cardinality-weighted variance-based 2-clustering with given center)

Given a set $\mathcal{Y} = \{y_1, \ldots, y_N\}$ of points from $\mathbb{R}^q$ and a positive integer M.

Find a partition of $\mathcal{Y}$ into two non-empty clusters $\mathcal{C}$ and $\mathcal{Y} \setminus \mathcal{C}$ such that

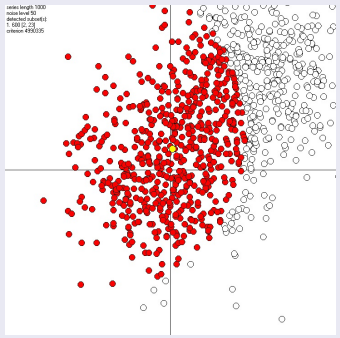
$$F(\mathcal{C}) = |\mathcal{C}| \sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 + |\mathcal{Y} \setminus \mathcal{C}| \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2 \longrightarrow \min,$$

where $\overline{y}(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum_{y \in \mathcal{C}} y$ is the geometric center (centroid) of $\mathcal{C}$, subject to constraint $|\mathcal{C}| = M$.

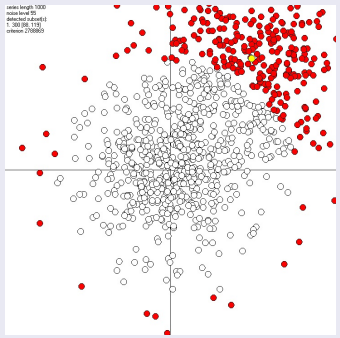
3. Cardinality-weighted variance-based 2-clustering with given center

Two-dimensional examples

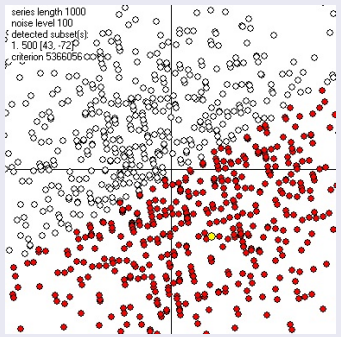
Example 1, $N = 1000, M = 600$



Example 2, $N = 1000, M = 300$



Example 3, $N = 1000, M = 500$



3. Cardinality-weighted variance-based 2-clustering with given center

Known results

1. The strong NP-hardness of the problem was proved in 2015, Kel'manov, Pyatkin.

2. An exact algorithm for the case of integer components of the input points was presented. If the dimension q of the space is bounded by a constant, then this algorithm has a pseudopolynomial $\mathcal{O}(N(MB)^q)$-time complexity, where B is the maximum absolute value of the coordinates of the input points; Kel'manov, Motkova (2015).

New result (Kel'manov, Motkova, 2016)

An approximation algorithm that allows to find a $(1+\varepsilon)$-approximate solution in $\mathcal{O}(qN^2(\sqrt{\frac{2q}{\varepsilon}}+2)^q)$ time for a given relative error ε. If the space dimension q is bounded by a constant this algorithm implements a fully polynomial-time approximation scheme with $\mathcal{O}\left(N^2\left(\frac{1}{\varepsilon}\right)^{q/2}\right)$-time complexity.

3. Cardinality-weighted variance-based 2-clustering with given center

The main idea of algorithm which implements an adaptive-grid-approach

For each point $y \in \mathcal{Y}$ steps 1-2 are executed:

1. Construct the cubic grid centered at the point y with node spacing h and edge cube size $2H + h$:

$$\mathcal{D}(y, h, H)$$
$$= \{d \in \mathbb{R}^q | \, d = y + h \cdot (i_1, \ldots, i_q), i_k \in \mathbb{Z}, |hi_k| \leq H, k \in \{1, \ldots, q\}\},$$

where

$$H = H(y) = \frac{1}{M}\sqrt{F(\mathcal{B}^y)}, \quad h = h(y, \varepsilon) = \frac{1}{M}\sqrt{\frac{2\varepsilon}{q}F(\mathcal{B}^y)},$$

where $\mathcal{B}^y$ is a subset of $\mathcal{Y}$ with M smallest values of the function

$$g^y(z) = (2M - N)\|z\|^2 - 2M\langle z, y\rangle, \quad z \in \mathcal{Y}.$$

3. Cardinality-weighted variance-based 2-clustering with given center

The main idea of algorithm which implements an adaptive-grid-approach

2. For each node x of the grid $\mathcal{D}(y, h, H + h/2)$ find a subset $\mathcal{B}^x \subseteq \mathcal{Y}$ with M smallest values of

$$g^x(y) = (2M - N)\|y\|^2 - 2M\langle y, x\rangle, \quad y \in \mathcal{Y}.$$

Compute $F(\mathcal{B}^x)$ and remember this value and the subset $\mathcal{B}^x$.

3. In the family $\{\mathcal{B}^x \,|\, x \in \mathcal{D}(y, h, H + h/2), y \in \mathcal{Y}\}$ choose as a solution $\mathcal{C}_{\mathcal{A}}$ the set $\mathcal{B}^x$ for which the value of $F(\mathcal{B}^x)$ is the smallest.

Problem 4 (Weighted variance-based 2-clustering with given center)

Given: an N-element set $\mathcal{Y}$ of points from $\mathbb{R}^q$, a positive integer $M \leq N$, and real numbers (weights) $\omega_1 > 0$ and $\omega_2 \geq 0$.
Find: a partition of $\mathcal{Y}$ into two clusters $\mathcal{C}$ and $\mathcal{Y}\backslash\mathcal{C}$ minimizing the value of

$$w_1 \sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 + w_2 \sum_{y \in \mathcal{Y}\backslash\mathcal{C}} \|y\|^2$$

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

4. Weighted variance-based 2-clustering with given center

Existing results for the case $w_1 = 1$ and $w_2 = 0$ (M-variance problem)

Aggarval, Imai, Katoh, Suri (exact algoritm, 1991),

Kel'manov, Pyatkin (strong NP-hardness, 2010),

Kel'manov, Romanchenko:
- 2-approximation polynomial-time algorithm, 2011,
- exact pseudopolynomial-time algorithm for the special case, 2012,
- FPTAS for the special case, 2014,

Shenmaier:
- PTAS, 2012,
- exact algorithm, 2016.

4. Weighted variance-based 2-clustering with given center

Existing results for the case $w_1 = w_2 = const$

Gimadi, Kel'manov, Pyatkin et al. (strong NP-hardness, 2006–2008),

Gimadi, Pyatkin, Rykov, Shenmaier (exact algorithms, 2007–2016),

Dolgushev, Kel'manov (2-approximation polynomial-time algorithm, 2011),

Baburin, Gimadi, Glebov, Glaskov, Pyatkin, Rykov, Kel'manov, Khandeev (exact pseudopolynomial-time algorithms for the special case,

2007-2015),

Kel'manov, Khandeev (randomized algorithm, 2015),

Dolgushev, Kel'manov, Shenmaier (PTAS, 2015).

4. Weighted variance-based 2-clustering with given center

Existing results for the case $w_1 = |\mathcal{C}|$, $w_2 = |\mathcal{Y} \setminus \mathcal{C}|$

Kel'manov, Pyatkin, 2015 (strong NP-hardness),

Kel'manov, Motkova:

- exact algorithm for the special case, 2015,
- 2-approximation polynomial-time algorithm, 2016 (accepted),
- FPTAS for the special case, 2016.

New results (Kel'manov, Motkova, Shenmaier, 2017)

1. An approximation algorithm for any $w_1 > 0$ and $w_2 \geq 0$ that allows to find a $(1+\varepsilon)$-approximate solution in $\mathcal{O}(qN^2(\sqrt{\frac{2q}{\varepsilon}}+2)^q)$ time for a given relative error ε. If the space dimension q is bounded by a constant this algorithm implements a fully polynomial-time approximation scheme with $\mathcal{O}\left(N^2\left(\frac{1}{\varepsilon}\right)^{q/2}\right)$-time complexity.

2. We propose the modification of this algorithm with improved time complexity $\mathcal{O}\left(\sqrt{q}N^2\left(\frac{\pi e}{2}\right)^{q/2}\left(\sqrt{\frac{2}{\varepsilon}}+2\right)^q\right)$. The algorithm implements an FPTAS for the fixed space dimension.
If the space dimension is $\mathcal{O}(\log N)$ the algorithm remains polynomial and implements a PTAS.

Algorithm $\mathcal{A}_1$ (the main idea)

For each point $y \in \mathcal{Y}$ steps 1-2 are executed:

1. Construct the cubic grid centered at the point y with node spacing h and edge cube size $2H + h$:

$$\mathcal{D} = \mathcal{D}(y, h, H)$$
$$= \{d \in \mathbb{R}^q | \, d = y + h \cdot (i_1, \ldots, i_q), i_k \in \mathbb{Z}, |hi_k| \leq H, k \in \{1, \ldots, q\}\},$$

where

$$H = H(y) = \sqrt{\frac{1}{Mw_1} F(\mathcal{B}^y)}, \quad h = h(y, \varepsilon) = \sqrt{\frac{2\varepsilon}{qMw_1} F(\mathcal{B}^y)}, \qquad (1)$$

where $\mathcal{B}^y$ is a subset of $\mathcal{Y}$ with M smallest values of the function

$$g^y(z) = (w_1 - w_2)\|z\|^2 - 2w_1 \langle z, y \rangle, \quad z \in \mathcal{Y}.$$

Algorithm $\mathcal{A}_1$ (the main idea)

2. For each node x of the grid $\mathcal{D}(y, h, H)$ find a subset $\mathcal{B}^x \subseteq \mathcal{Y}$ with M smallest values of

$$g^x(y) = (w_1 - w_2)\|y\|^2 - 2w_1 \langle y, x \rangle, \quad y \in \mathcal{Y}.$$

Compute $F(\mathcal{B}^x)$ and remember this value and the subset $\mathcal{B}^x$.

3. In the family $\{\mathcal{B}^x \,|\, x \in \mathcal{D}(y, h, H + h/2), y \in \mathcal{Y}\}$ choose as a solution $\mathcal{C}_{\mathcal{A}}$ the set $\mathcal{B}^x$ for which the value of $F(\mathcal{B}^x)$ is the smallest.

Improved Algorithm $\mathcal{A}_2$ (the main idea)

1. For each $y \in \mathcal{Y}$ let

$$R = H + \frac{h\sqrt{q}}{2},$$

where $H = H(y)$, $h = h(y, \varepsilon)$ are the grid parameters defined by (1). Construct the reduced spherical lattice

$$\mathcal{D}_R(y, h, H + h/2) = \mathcal{D}(y, h, H + h/2) \cap B(y, R),$$

where

$$B(y, R) = \{x \in \mathbb{R}^q | \, \|x - y\| \leq R\}$$

is the ball of radius R and center y.

2. Instead of grids $\mathcal{D}(y, h, H + h/2)$ we use grids $\mathcal{D}_R(y, h, H + h/2)$.

Problem 5 (Finding a subsequence in a sequence)

Given a sequence $\mathcal{Y} = (y_1, \ldots, y_N)$ of points from $\mathbb{R}^q$, and some positive integer numbers $T_{\min}$, $T_{\max}$ and M.
Find a tuple $\mathcal{M} = (n_1, \ldots, n_M)$, where
$n_m \in \mathcal{N} = \{1, \ldots, N\}, m = 1, \ldots, N$, such that

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

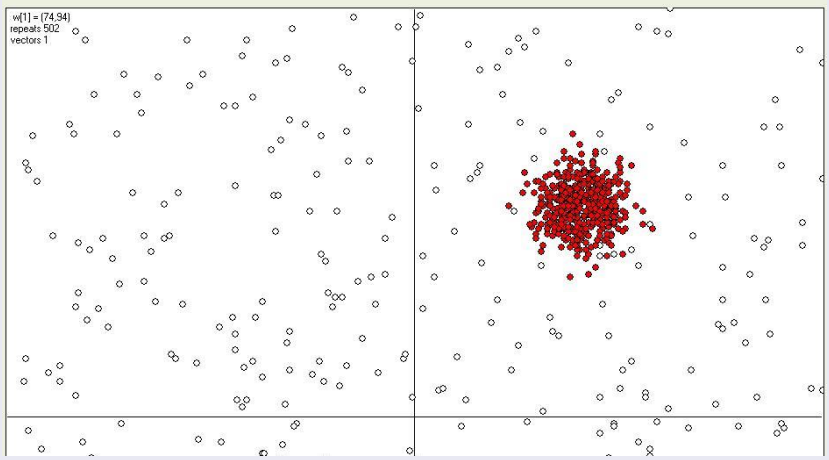
where $\overline{y}(\mathcal{M})$ is the centroid of $\{y_j | j \in \mathcal{M}\}$, under constraints

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

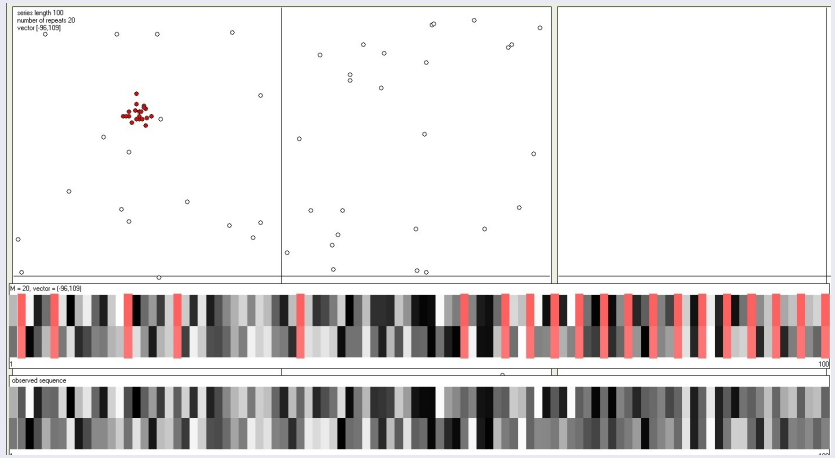
on the elements of $\mathcal{M}$.

5. Finding a subsequence in a sequence

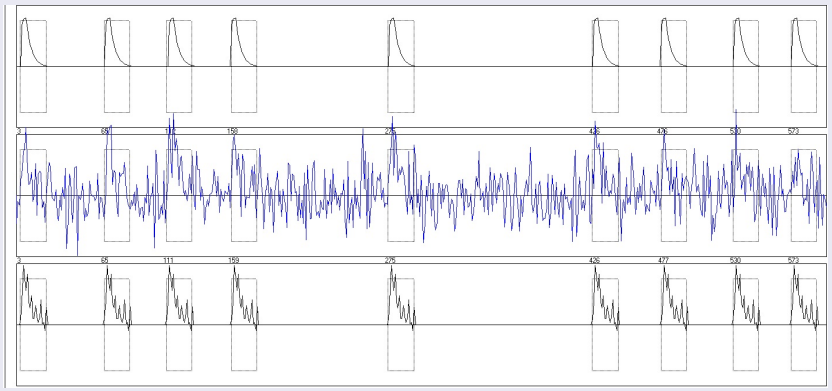
Finding a subset, $q = 2$



Finding a subsequence, $q = 2$



Finding a subsequence of repeating fragments (fragment size $q = 20$) in a one-dimensional sequence



5. Finding a subsequence in a sequence

Known results

1. The strong NP-hardness of the problem is implied from the results of Kel'manov, Pyatkin (2010).

2. A 2-approximation polynomial algorithm having $\mathcal{O}(N^2(N+q))$ running time was proposed in 2012 (Kel'manov, Romanchenko, Khamidullin).

3. For the case of fixed space dimension and integer input points coordinates, an exact pseudopolynomial algorithm with $\mathcal{O}(N^3(MD)^q)$-time complexity where D is the maximum absolute coordinate value of the points was presented in 2013 (Kel'manov, Romanchenko, Khamidullin).

Currently, there are no other algorithmic results for the problem.

New result (Kel'manov, Romanchenko, Khamidullin, 2016)

An FPTAS for the case of fixed space dimension with $\mathcal{O}(MN^3(1/\varepsilon)^{q/2})$-time complexity for an arbitrary relative error ε.

The main idea of algorithm which implements an adaptive-grid-approach

For each point $y \in \mathcal{Y}$ steps 1-2 are executed:

1. Construct the cubic grid centered at the point y with node spacing h and edge cube size $2H + h$

$$\mathcal{D}(y, h, H)$$
$$= \{d \in \mathbb{R}^q | \, d = y + h \cdot (i_1, \ldots, i_q), i_k \in \mathbb{Z}, |hi_k| \leq H, k \in \{1, \ldots, q\}\},$$

where

$$H = H(y) = \sqrt{\frac{1}{M} F(\mathcal{M}^y)}, \quad h = h(y, \varepsilon) = \sqrt{\frac{2\varepsilon}{qM} F(\mathcal{M}^y)},$$

where $\mathcal{M}^y$ is the optimal solution (tuple of indices) of the auxiliary problem

$$\sum_{i \in \mathcal{M}} \|y_i - y\|^2 \to \min_{\mathcal{M} \subseteq \mathcal{N}}.$$

The main idea of algorithm which implements an adaptive-grid-approach

2. For each node d of the grid $\mathcal{D}(y, h, H + h/2)$ find the optimal solution (tuple of indices) $\mathcal{M}^d$ of the auxiliary problem

$$\sum_{i \in \mathcal{M}} \|y_i - d\|^2 \to \min_{\mathcal{M} \subseteq \mathcal{N}}.$$

Compute $F(\mathcal{M}^d)$ and remember this value and the subset $\mathcal{M}^d$.

3. In the family $\{\mathcal{M}^d \,|\, d \in \mathcal{D}(y, h, H + h/2), y \in \mathcal{Y}\}$ choose as a solution $\mathcal{M}_{\mathcal{A}}$ the set $\mathcal{M}^d$ for which the value of $F(\mathcal{M}^d)$ is the smallest.

6. Minimum sum-of-squares 2-clustering problem on sequence with given center of one cluster

Problem 6 (Minimum sum-of-squares 2-clustering problem on sequence with given center of one cluster)

Given: a sequence $\mathcal{Y} = (y_1, \ldots, y_N)$ of points from $\mathbb{R}^q$, and some positive integer numbers $T_{\min}$, $T_{\max}$, and M.
Find: a tuple $\mathcal{M} = (n_1, \ldots, n_M)$, where $n_m \in \mathcal{N} = \{1, \ldots, N\}, m = 1, \ldots, N$, such that

$$\sum_{j \in \mathcal{M}} \|y_j - \overline{y}(\mathcal{M})\|^2 + \sum_{j \in \mathcal{N} \setminus \mathcal{M}} \|y_j\|^2 \to \min,$$

where $\overline{y}(\mathcal{M}) = \frac{1}{M} \sum_{i \in \mathcal{M}} y_i$, under constraints

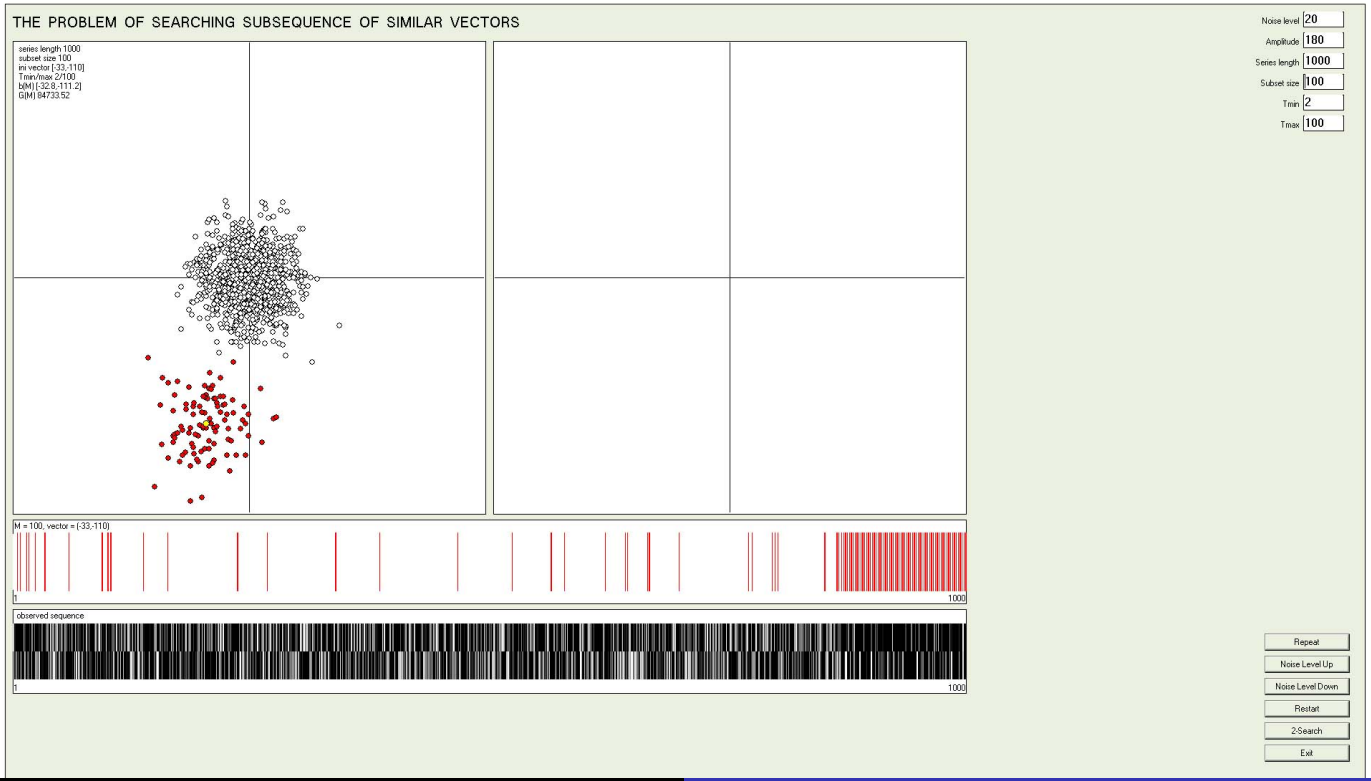
$$T_{\min} \leq n_m - n_{m-1} \leq T_{\max} \leq N, \ m = 2, \ldots, M,$$

on the elements of $\mathcal{M}$.

6. Minimum sum-of-squares 2-clustering problem on sequence with given center of one cluster

2-dimensional example

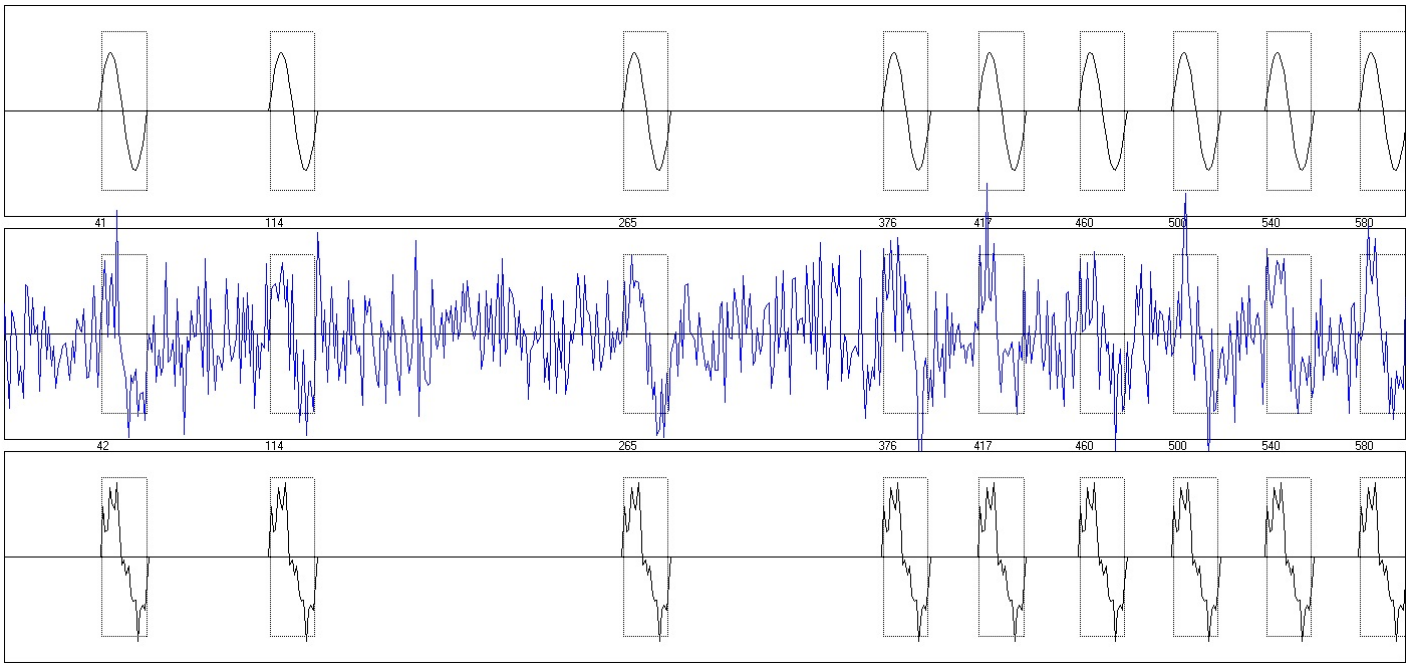
1000 results of the measurements of a tuple of numerical characteristics of some object.
100 measurements correspond to the active state; 900 measurements correspond to the passive state.



6. Minimum sum-of-squares 2-clustering problem on sequence with given center of one cluster

Example. Partitioning a one-dimensional sequence: finding a repeated fragment (fragment size $q = 20$)

Subsequence of 600 results of the measurements of one numerical characteristics of some object (9 times the object was in a passive state)



6. Minimum sum-of-squares 2-clustering problem on sequence with given center of one cluster

Known results

1. The problem is strongly NP-hard. Therefore, the problem admits neither exact polynomial, nor exact pseudopolynomial algorithm, P≠NP (Kel'manov, Pyatkin, 2013).

2. If $T_{\min}$, $T_{\max}$ — are the parameters, then the problem:
(1) strongly NP-hard for any $T_{\min} < T_{\max}$;
(2) solvable in polynomial time when $T_{\min} = T_{\max}$.
(Kel'manov, Pyatkin, 2013)

3. A 2-approximation polynomial-time algorithm running in $\mathcal{O}(N^2(MN + q))$ time was presented
(Kel'manov, Khamidullin, 2013).

6. Minimum sum-of-squares 2-clustering problem on sequence with given center of one cluster

Known results

4. For the case of integer inputs and fixed space dimension q an exact pseudopolynomial algorithm was constructed. The time complexity of this algorithm is $\mathcal{O}(MN^2(MD)^q)$, where D is the maximum absolute in any coordinate of the input points
(Kel'manov, Khamidullin, Khandeev, 2015).

5. For the case of fixed space dimension a fully polynomial-time approximation scheme (FPTAS) was proposed which, given a relative error ε, finds a $(1+\varepsilon)$-approximate solution of the problem in $\mathcal{O}(MN^3(1/\varepsilon)^{q/2})$ time
(Kel'manov, Khamidullin, Khandeev, 2016).

6. Minimum sum-of-squares 2-clustering problem on sequence with given center of one cluster

New results (Kel'manov, Khamidullin, Khandeev, 2017)

1. A **randomized** algorithm is proposed which, given an $\varepsilon > 0$ and fixed $\gamma \in (0, 1)$, for an established parameter value allows to find a $(1 + \varepsilon)$-approximate solution of the problem with a probability of at least $1 - \gamma$ in $\mathcal{O}(qMN^2)$ time.

2. The conditions are established under which the algorithm is **asymptotically exact** and its time complexity is $\mathcal{O}(qMN^3)$.

6. Minimum sum-of-squares 2-clustering problem on sequence with given center of one cluster

Algorithm $\mathcal{A}$ (the approach to Problem 6)

Input: sequence $\mathcal{Y}$, positive integers $T_{\min}$, $T_{\max}$, M, and positive integer parameter k.

Step 1. Generate a multiset $\mathcal{T}$ by independently and randomly choosing k elements one after another (with replacement) from $\mathcal{Y}$.

Step 2. For each nonempty $\mathcal{H} \subseteq \mathcal{T}$, compute the centroid $\overline{y}(\mathcal{H})$ and find the optimal solution $\mathcal{M}^{\overline{y}(\mathcal{H})}$ of the problem

$$\sum_{n \in \mathcal{M}} (2\langle y_n, \overline{y}(\mathcal{H})\rangle - \|\overline{y}(\mathcal{H})\|^2) \to \max_{\mathcal{M}}$$

under constraints $T_{\min} \le n_m - n_{m-1} \le T_{\max} \le N,\ m = 2, \ldots, M$, on the elements of $\mathcal{M} = (n_1, \ldots, n_M)$.

Step 3. In the family of tuples found at Step 2 choose $\mathcal{M}^{\overline{y}(\mathcal{H})}$ minimizing the value $F(\mathcal{M}^{\overline{y}(\mathcal{H})})$ as a solution $\mathcal{M}_{\mathcal{A}}$ of the problem. If there are several optimal values, then choose any of them.

Output: the tuple $\mathcal{M}_A$.

Problem 7 (Partitioning a sequence into clusters with restrictions on their cardinalities)

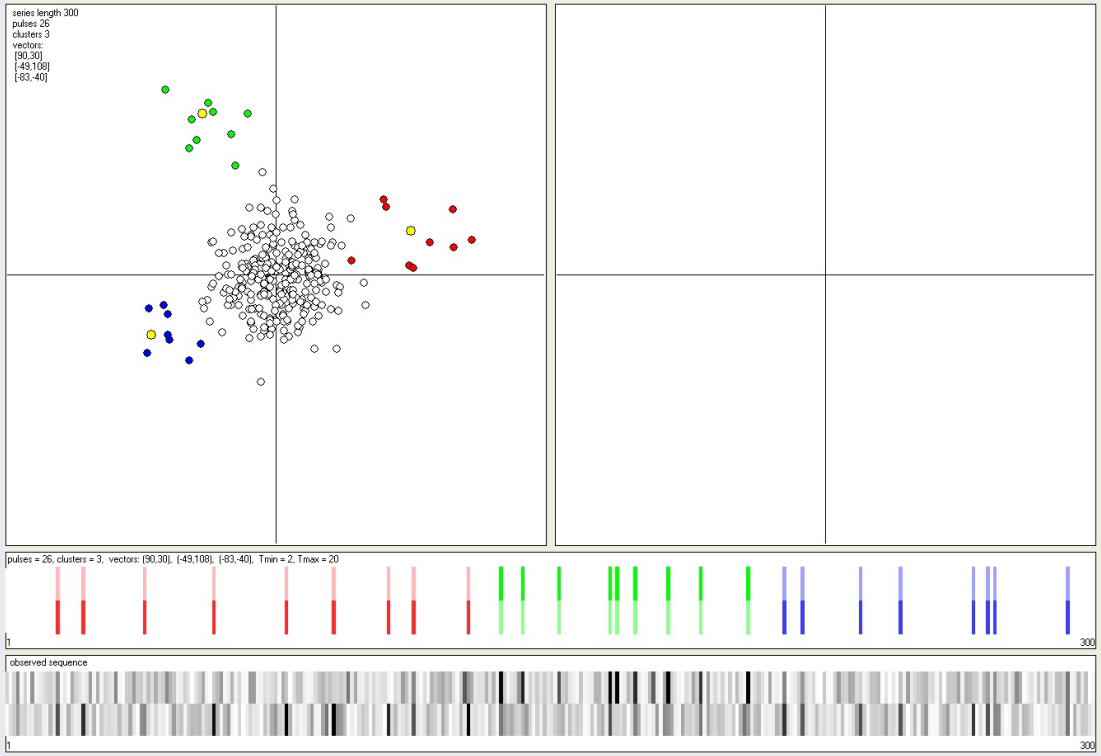
Given a sequence $\mathcal{Y} = (y_1, \ldots, y_N)$ of points from $\mathbb{R}^q$ and some positive integers $T_{\min}$, $T_{\max}$, L, and M.
Find nonempty disjoint subsets $\mathcal{M}_1, \ldots, \mathcal{M}_L$ of $\mathcal{N} = \{1, \ldots, N\}$ such that

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

where $\mathcal{M} = \bigcup_{l=1}^{L} \mathcal{M}_l$, and $\overline{y}(\mathcal{M}_l)$ is the centroid of subset $\{y_j | j \in \mathcal{M}_l\}$, under the following constraints:
(i) the cardinality of $\mathcal{M}$ is equal to M,
(ii) concatenation of elements of subsets $\mathcal{M}_1, \ldots, \mathcal{M}_L$ is an increasing sequence, provided that the elements of each subset are in ascending order,
(iii) the following inequalities for the elements of $\mathcal{M} = \{n_1, \ldots, n_M\}$ are satisfied: $T_{\min} \leq n_m - n_{m-1} \leq T_{\max} \leq N$, $m = 2, \ldots, M$.

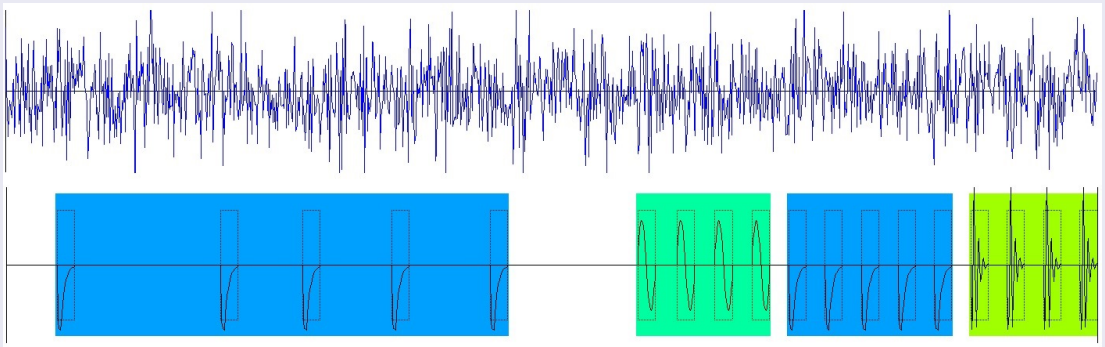
7. Partitioning a sequence into clusters with restrictions on their cardinalities

2-dimensional example



7. Partitioning a sequence into clusters with restrictions on their cardinalities

Example. Partitioning a one-dimensional sequence into clusters containing a repeated fragment (fragment size $q = 20$)



7. Partitioning a sequence into clusters with restrictions on their cardinalities

Known results

1. The strong NP-hardness of the problem is implied from the results of Kel'manov, Pyatkin (2011, 2013).

At present, for Problem 4, except for its **particular case** when $L = 1$, there are no efficient algorithms with guaranteed accuracy. For the mentioned case of Problem 4 the following results were obtained.

2. A 2-approximation polynomial-time algorithm having $\mathcal{O}(N^2(MN + q))$ running time was proposed in 2014 (Kel'manov, Khamidullin).

3. For the case of fixed space dimension and integer input points coordinates, an exact pseudopolynomial algorithm with $\mathcal{O}(MN^2(MD)^q)$-time complexity where D is the maximum absolute coordinate value of the points was presented in 2015 (Kel'manov, Khamidullin, Khandeev).

7. Partitioning a sequence into clusters with restrictions on their cardinalities

Known results

4. A randomized algorithm was proposed. For an established parameter value, the algorithm finds an approximate solution of the problem in $\mathcal{O}(qMN^2)$-time for given values of the relative error and failure probability. The conditions are established under which the algorithm is asymptotically exact and runs in $\mathcal{O}(qMN^3)$-time (Kel'manov, Khamidullin, Khandeev, 2015).

Currently, there are no other algorithmic results for Problem 4.

New result (Kel'manov, Mikhailova, Khamidullin, Khandeev, 2016)

An algorithm that allows to find a 2-approximate solution of Problem 4 in $\mathcal{O}(LN^{L+1}(MN+q))$ time, which is polynomial if L is fixed (bounded by some constant).

7. Partitioning a sequence into clusters with restrictions on their cardinalities

The main idea of algorithm

For each point $y \in \mathcal{Y}$ steps 1-2 are executed:

1. For every tuple $x = (x_1, \ldots, x_L) \in \mathcal{Y}^L$ of elements of the sequence $\mathcal{Y}$, using **special dynamic programming scheme**, find the optimal solution $\{\mathcal{M}_1^x, \ldots, \mathcal{M}_L^x\}$ of the auxiliary problem

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

2. Find a tuple $x(A) = \arg\max_{x \in \mathcal{Y}^L} G^x(\mathcal{M}_1^x, \ldots, \mathcal{M}_L^x)$ and a family $\{\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A\} = \{\mathcal{M}_1^{x(A)}, \ldots, \mathcal{M}_L^{x(A)}\}$.

If the optimum is taken by several tuples, we choose any of them.

Problem 8 (Partitioning a sequence into clusters)

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

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

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

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

(the cardinality of $\mathcal{M}$ assumed to be unknown)

8. Partitioning a sequence into clusters

Known results

1. The strong NP-hardness of the problem is implied from the results of Kel'manov, Pyatkin (2011, 2013).

At present, for Problem 5, except for its **particular case when** $L = 1$, there are no efficient algorithms with guaranteed accuracy. For the mentioned case of Problem 5 the following results were obtained.

2. A 2-approximation polynomial-time algorithm having $\mathcal{O}(N^2(N+q))$ running time was proposed in 2015 (Kel'manov, Khamidullin).

Currently, there are no other algorithmic results for Problem 5.

New result (Kel'manov, Mikhailova, Khamidullin, Khandeev, 2016)

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

The main idea of algorithm

For each point $y \in \mathcal{Y}$ steps 1-2 are executed:

1. For every tuple $x = (x_1, \ldots, x_L) \in \mathcal{Y}^L$ of elements of the sequence $\mathcal{Y}$, using **special dynamic programming scheme**, find the optimal solution $\{\mathcal{M}_1^x, \ldots, \mathcal{M}_L^x\}$ of the auxiliary problem

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

2. Find a tuple $x(A) = \arg\max_{x \in \mathcal{Y}^L} G^x(\mathcal{M}_1^x, \ldots, \mathcal{M}_L^x)$ and a family $\{\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A\} = \{\mathcal{M}_1^{x(A)}, \ldots, \mathcal{M}_L^{x(A)}\}$.

If the optimum is taken by several tuples, we choose any of them.

Thank you for your attention!