Odjel za matematiku
Sveučilište Josipa Jurja Strossmayera u Osijeku

Department of Mathematics
J. J. Strossmayer University of Osijek

Serije: Rukopisi u pripremi
Series: Technical Reports

One-dimensional center-based l_1-clustering method

Kristian Sabo, Rudolf Scitovski, Ivan Vazler

One-dimensional center-based l_1-clustering method[1]

Kristian Sabo
Department of Mathematics, University of Osijek
Trg Lj. Gaja 6, HR – 31 000 Osijek, Croatia
e-mail: ksabo@mathos.hr

Rudolf Scitovski[2]
Department of Mathematics, University of Osijek
Trg Lj. Gaja 6, HR – 31 000 Osijek, Croatia
e-mail: scitowsk@mathos.hr

Ivan Vazler
Department of Mathematics, University of Osijek
Trg Lj. Gaja 6, HR – 31 000 Osijek, Croatia
e-mail: ivazler@mathos.hr

Abstract. Motivated by the method for solving center-based Least Squares – clustering problem [21, 39] we construct a very efficient iterative process for solving a one-dimensional center-based l_1 – clustering problem, on the basis of which it is possible to determine the optimal partition. We analyze the basic properties and convergence of our iterative process, which converges to a stationary point of the corresponding objective function for each choice of the initial approximation. Given is also a corresponding algorithm, which in only few steps gives a stationary point and the corresponding partition. The method is illustrated and visualized on the example of looking for an optimal partition with two clusters, where we check all stationary points of the corresponding minimizing functional. Also, the method is tested on the basis of large numbers of data points and clusters and compared with the method for solving the center-based Least Squares – clustering problem described in [21, 39].

Key words: clustering, data mining, optimization, weighted median problem

MSC2010: 62H30, 68T10, 90C26, 90C27, 91C20, 47N10

1 Introduction

Clustering or grouping a data set into conceptually meaningful clusters is a well-studied problem in recent literature, and it has practical importance in a wide variety of applications such as biology, classification of the plough-lands according to fertility, classification of insects into groups, ranking of municipalities for financial support, pattern recognition, information retrieval, text classification, machine learning, business, facility location problem, medicine, understanding the Earth's climate, psychology, and other social sciences [5, 10, 16, 28, 34, 38].

[1] This work is supported by the Ministry of Science, Education and Sports, Republic of Croatia, through research grant 235-2352818-1034.

[2] Corresponding author: Rudolf Scitovski, e-mail: scitowsk@mathos.hr, telephone number: ++385-224-800, fax number: ++385-224-801

Classification and ranking of objects are becoming more and more interesting topics for researchers, decision makers, state administrations, etc.

One of the most popular clustering algorithm is k-means. A classical version of k-means uses the squared Euclidean distance. However, this distance measure is often inappropriate [11]. Various other distance-like functions can be found in literature, like e.g. Bregman distance [1, 21, 24, 39]. The k-means algorithm generally faces a nonconvex and nonsmooth optimization problem. Therefore, a well-known disadvantage of this algorithm lies in its strong dependency on the choice of the initial partition. A probabilistic approach to data clustering, which is based on the Weiszfeld method for solving the Fermat–Weber location problem is described in [2, 18].

A partition of the set $\mathcal{A} = \{a_i \in \mathbb{R}^n : i = 1, \dots, m\} \subset \mathbb{R}^n$ into k disjoint subsets $\pi_1, \dots, \pi_k$, $1 \leq k \leq m$, such that

$$\bigcup_{i=1}^{k} \pi_i = \mathcal{A}, \qquad \pi_i \cap \pi_j = \emptyset, \quad i \neq j, \qquad |\pi_j| \geq 1, \quad j = 1, \dots, k, \tag{1}$$

will be denoted by $\Pi(\mathcal{A}) = \{\pi_1, \dots, \pi_k\}$, and the elements $\pi_1, \dots, \pi_k$ of such partition are called *clusters in* $\mathbb{R}^n$.

If $d \colon \mathbb{R}^n \times \mathbb{R}^n \to [0, +\infty\rangle$ is some distance-like function (see e.g. [21, 39]), then, by applying the *minimal distance condition* (see e.g. [21, 37]), with each cluster $\pi_j \in \Pi$ we can associate its center c_j, defined by

$$c_j = c(\pi_j) := \operatorname*{argmin}_{x \in \mathcal{C}_j} \sum_{a_i \in \pi_j} d(x, a_i), \tag{2}$$

where $\mathcal{C}_j = \operatorname{conv}(\pi_j)$.

If we define an objective function $\mathcal{F} \colon \mathcal{P}(\mathcal{A}, k) \to [0, +\infty\rangle$ on the set of all partitions $\mathcal{P}(\mathcal{A}, k)$ of the set $\mathcal{A}$ containing k clusters by

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a_i \in \pi_j} d(c_j, a_i), \tag{3}$$

then we define an optimal partition $\Pi^\star$, such that

$$\mathcal{F}(\Pi^\star) = \min_{\Pi \in \mathcal{P}(\mathcal{A}, k)} \mathcal{F}(\Pi).$$

Conversely, for a given set of centers $c_1, \dots, c_k \in \mathbb{R}^n$ applying the minimal distance condition we can define the partition $\Pi = \{\pi_1, \dots, \pi_k\}$ of the set $\mathcal{A}$ in the following way:

$$\pi_j = \{a \in \mathcal{A} : d(c_j, a) \leq d(c_s, a), \forall s = 1, \dots, k\}, \qquad j = 1, \dots, k, \tag{4}$$

where one has to take care that every element of the set $\mathcal{A}$ occurs in one and only one cluster. Therefore the problem of finding an optimal partition of the set $\mathcal{A}$ can be reduced to the following optimization problem

$$\min_{c_1, \dots, c_k \in \mathbb{R}^n} F(c_1, \dots, c_k), \qquad F(c_1, \dots, c_k) = \sum_{i=1}^{m} \min_{j=1,\dots,k} d(c_j, a_i), \tag{5}$$

where $F\colon \mathbb{R}^{kn} \to \mathbb{R}_+$, and $\mathbb{R}_+$ is the set of all vectors in $\mathbb{R}^n$ with nonnegative components. In general, this functional is not differentiable and it may have several local minima. Optimization problem (5) can also be found in literature as a k-median problem and it is most frequently solved by various metaheuristic methods [12] or by applying integer programming [27, 32, 36]. An overview of papers in this area up to the year 2006 can be found in [31].

Kogan [21] and Teboulle [39] considered problem (5) in the Least Squares sense for the distance-like function $d(x,y) = \|x-y\|_2^2$, $x, y \in \mathbb{R}^n$. Generally, since the function $f\colon \mathbb{R}^k \to \mathbb{R}$, $f(z) = \max\limits_{j=1,\dots,k} z_j$ can be approximated by a smooth function $f_\epsilon(z) = \epsilon \ln \sum\limits_{j=1}^{k} \exp\left(\frac{z_j}{\epsilon}\right)$, instead of solving problem (5), the following optimization problem is considered:

$$\min_{c_1,\dots,c_k \in \mathbb{R}} F_\epsilon(c_1,\dots,c_k), \qquad F_\epsilon(c_1,\dots,c_k) = -\epsilon \sum_{i=1}^{m} \ln \sum_{j=1}^{k} e^{-\frac{\|c_j - a_i\|_2^2}{\epsilon}}, \tag{6}$$

and a simple iterative procedure is proposed for finding a stationary point of the differentiable functional (6) as a sequence of a corresponding weighted arithmetic mean of the data. Motivated by this, in our paper we construct an efficient iterative process for solving a one-dimensional center-based l_1 – clustering problem.

This paper is organized as follows. In Section 2, a one-dimensional clustering problem is described and the main properties of the minimizing functional are given. Section 3 gives the method for finding centers of clusters as stationary points of the minimizing functional, analyzes convergence of the iterative process and gives an appropriate algorithm. An illustrative example as well as testing and a comparison of the proposed method on a larger number of data that should be grouped into a larger number of clusters are given in Section 4. Measurements of the CPU time indicate high efficiency of the proposed method.

2 One-dimensional l_1-clustering

Next, we consider a one-dimensional ($n = 1$) clustering problem, which also has many applications [18, 21]. An even more motivating reason for investigating such a special clustering problem is a possibility of solving large and high-dimensional data clustering problems by reducing them to one-dimensional ones (see e.g. *Principal direction divisive partitioning* in [3, 22, 26]).

The set $\mathcal{A} = \{a_i \in \mathbb{R} : i = 1,\dots,m\} \subset I \subset \mathbb{R}$, $I = [\alpha, \beta]$, has to be divided into k disjoint subsets $\pi_1,\dots,\pi_k$, $1 \le k \le m$, satisfying (1). In the present paper, we consider a one-dimensional clustering problem using the Least Absolute Deviations (LAD) – optimality criterion [33], by applying l_1-distance function $d(x,y) = |x-y|$. The LAD principle is not sensitive to the presence of outliers among the data, and the problem of finding an optimal partition of the set $\mathcal{A}$ according to (5) reduces to the following nonconvex and nonsmooth optimization problem

$$\min_{c_1,\dots,c_k \in I} \Phi(c_1,\dots,c_k), \qquad \Phi(c_1,\dots,c_k) = \sum_{i=1}^{m} \min_{j=1,\dots,k} |c_j - a_i|, \tag{7}$$

where $\Phi : I^k \to \mathbb{R}_+$ is a continuous function. Similarly to [21, 39], instead of solving problem (7), we can solve the following optimization problem

$$\min_{c_1,\dots,c_k\in I}\Phi_\epsilon(c_1,\dots,c_k),\qquad \Phi_\epsilon(c_1,\dots,c_k)=-\epsilon\sum_{i=1}^m\ln\sum_{j=1}^k\exp\left(-\frac{|c_j-a_i|}{\epsilon}\right), \tag{8}$$

where $\Phi_\epsilon\colon I^k\to\mathbb{R}_+$. This is an optimization problem for the continuous objective function, which is further neither convex nor differentiable. Thereby the objective function can have a great number of independent variables (the number of clusters in the partition multiplied by the dimension of data points $(k\cdot n)$).

2.1 Properties of the functional Φ_ϵ

In this subsection we are going to analyze some properties of the functional Φ_ϵ. To simplify the notation we denote by $\theta:=(c_1,\dots,c_k)$. The next theorem relates the function Φ given by (7) and Φ_ϵ given by (8) (see also [39]).

Theorem 1. *Let $\mathcal{A}=\{a_i\in\mathbb{R}:i=1,\dots,m\}\subset I\subset\mathbb{R}$, $I=[\alpha,\beta]$, be a given set, and let Φ_ϵ, $\epsilon>0$, be a functional given by* (8).

Then for all $\theta=(c_1,\dots,c_k)\in I^k$, the following inequalities hold

$$0<\epsilon\, m\ln\left(1+(k-1)e^{-\frac{1}{\epsilon}(\beta-\alpha)}\right)\le\Phi(\theta)-\Phi_\epsilon(\theta)\le\epsilon\, m\ln k. \tag{9}$$

Proof. Denote by $\Delta_{ij}=\frac{|c_j-a_i|}{\epsilon}$, $i=1,\dots,m$, $j=1,\dots,k$. Without loss of generality, we may assume that $\Delta_{i1}\le\cdots\le\Delta_{ik}$ for all $i\in\{1,\dots,m\}$. By definition of Φ and Φ_ϵ, we find

$$\begin{aligned}\Phi(\theta)-\Phi_\epsilon(\theta)&=\epsilon\sum_{i=1}^m\min_{j=1,\dots,k}\frac{|c_j-a_i|}{\epsilon}+\epsilon\sum_{i=1}^m\ln\sum_{j=1}^k e^{-\frac{|c_j-a_i|}{\epsilon}}\\&=\epsilon\sum_{i=1}^m\ln\left(e^{\Delta_{i1}}\sum_{j=1}^k e^{-\Delta_{ij}}\right)=\epsilon\sum_{i=1}^m\ln\sum_{j=1}^k e^{-(\Delta_{ij}-\Delta_{i1})}.\end{aligned} \tag{10}$$

Furthermore, $0\le\Delta_{ij}-\Delta_{i1}=\frac{1}{\epsilon}\Big||c_j-a_i|-|c_1-a_i|\Big|\le\frac{1}{\epsilon}|c_j-c_1|\le\frac{1}{\epsilon}(\beta-\alpha)$, which implies

$$\sum_{j=1}^k e^{-(\Delta_{ij}-\Delta_{i1})}=1+\sum_{j=2}^k e^{-(\Delta_{ij}-\Delta_{i1})}\ge1+(k-1)e^{-\frac{1}{\epsilon}(\beta-\alpha)}>1,$$

hence the left-hand side of inequality (9) follows from (10). Finally, the right-hand side of inequality (9) follows from (10) since $\sum_{j=1}^k e^{-(\Delta_{ij}-\Delta_{i1})}<k$. □

The functional Φ_ϵ is continuous, and according to Theorem 1, it is bounded below,

$$\Phi_\epsilon(\theta)\ge\Phi(\theta)-\epsilon\, m\ln k\ge-\epsilon\, m\ln k.$$

Therefore, since $I^k\subset\mathbb{R}^k$ is compact, Φ_ϵ attains its global minimum.

The next lemma shows that the functional Φ_ϵ, in addition to being continuous, satisfies the Lipschitz property.

Lemma 1. *For all $\theta_1, \theta_2 \in I^k$ there holds*

$$|\Phi_\epsilon(\theta_2) - \Phi_\epsilon(\theta_1)| \leq m \, \|\theta_2 - \theta_1\|_\infty. \tag{11}$$

Proof. Let $\theta_1 = (c_1, \dots, c_k), \theta_2 = (d_1, \dots, d_k) \in I^k$. Then

$$\begin{aligned}
\Phi_\epsilon(\theta_2) - \Phi_\epsilon(\theta_1) &= \epsilon \sum_{i=1}^m \ln \sum_{s=1}^k \frac{e^{-\frac{|c_s - a_i|}{\epsilon}}}{\sum_{j=1}^k e^{-\frac{|d_j - a_i|}{\epsilon}}} \\
&= \epsilon \sum_{i=1}^m \ln \sum_{s=1}^k \frac{e^{-\frac{|d_s - a_i|}{\epsilon}}}{\sum_{j=1}^k e^{-\frac{|d_j - a_i|}{\epsilon}}} e^{\frac{|d_s - a_i| - |c_s - a_i|}{\epsilon}} \\
&\leq \epsilon \sum_{i=1}^m \ln \sum_{s=1}^k \frac{e^{-\frac{|d_s - a_i|}{\epsilon}}}{\sum_{j=1}^k e^{-\frac{|d_j - a_i|}{\epsilon}}} e^{\frac{|d_s - c_s|}{\epsilon}} \\
&\leq \epsilon \sum_{i=1}^m \ln \exp \left(\max_{s=1,\dots,k} \frac{|d_s - c_s|}{\epsilon} \right) \\
&= m \max_{s=1,\dots,k} |d_s - c_s| = m \, \|\theta_2 - \theta_1\|_\infty
\end{aligned}$$

Similarly, one can show that $\Phi_\epsilon(\theta_1) - \Phi_\epsilon(\theta_2) \leq m \|\theta_2 - \theta_1\|_\infty$, and therefore follows (11). □

In what follows we will need the next lemma [4, 21].

Lemma 2. *Let $\psi \colon \mathbb{R}^k \to \mathbb{R}$ be defined by $\psi(x) = \ln \sum\limits_{s=1}^k e^{-x_s}$. Then*

(*i*) *ψ is a convex function differentiable of class $C^\infty(\mathbb{R}^k)$.*

(*ii*) *For every pair of points $x, y \in \mathbb{R}^k$ the following holds:*

$$\psi(y) - \psi(x) \leq \sum_{s=1}^k (x_s - y_s) \mu_s, \qquad \textit{where} \quad \mu_s = e^{-y_s} \left(\sum_{j=1}^k e^{-y_j} \right)^{-1}. \tag{12}$$

Proof. (*i*) Recall the Hölder inequality for $a, b \in \mathbb{R}^k$:

$$\sum_{i=1}^k |a_i b_i| \leq \left(\sum_{i=1}^k |a_i|^p \right)^{1/p} \left(\sum_{i=1}^k |b_i|^q \right)^{1/q}, \qquad p, q \in \langle 0, +\infty \rangle, \quad \frac{1}{p} + \frac{1}{q} = 1.$$

Given are two points $x = (x_1, \dots, x_k)$, $y = (y_1, \dots, y_k) \in \mathbb{R}^k$, and substituting $\alpha = \frac{1}{p}$, $\beta = \frac{1}{q}$, the Hölder inequality for $a = (e^{-\alpha x_1}, \dots, e^{-\alpha x_k})$ and $b = (e^{-\beta y_1}, \dots, e^{-\beta y_k})$ gives

$$\sum_{i=1}^k e^{-\alpha x_i - \beta y_i} \leq \left(\sum_{i=1}^k e^{-x_i} \right)^\alpha \left(\sum_{i=1}^k e^{-y_i} \right)^\beta,$$

and taking the logarithm we obtain

$$\psi(\alpha x + \beta y) \leq \alpha \psi(x) + \beta \psi(y), \qquad \alpha + \beta = 1.$$

(ii) Since ψ is a convex function, (ii) follows from the gradient inequality

$$\psi(x) - \psi(y) \geq (x - y)\nabla\psi(y).$$

Namely, since $\frac{\partial\psi(y)}{\partial y_s} = \frac{-\exp(-y_s)}{\sum_{j=1}^{k}\exp(-y_j)}$, $s = 1, \dots, k$, the gradient inequality readily implies inequality (12). □

3 A method for finding stationary points of the functional Φ_ϵ

Assuming that $\theta^{(t)} = (c_1^{(t)}, \dots, c_k^{(t)}) \in I^k$ is known, we are going to look for the next iteration $\theta^{(t+1)} = (c_1^{(t+1)}, \dots, c_k^{(t+1)})$, where $c_s^{(t+1)}$ is the weighted median of set $\mathcal{A}$ [35, 40][3] with appropriate weights, i.e.

$$c_s^{(t+1)} = \operatorname{med}\left(w^{(s)}(\theta^{(t)}), \mathcal{A}\right), \qquad s = 1, \dots, k, \tag{13}$$

where $w^{(s)}(\theta^{(t)}) = \left(w_1^{(s)}(\theta^{(t)}), \dots, w_m^{(s)}(\theta^{(t)})\right)$, and

$$w_i^{(s)}(\theta^{(t)}) = \frac{\exp\left(-\frac{1}{\epsilon}|c_s^{(t)} - a_i|\right)}{\sum\limits_{j=1}^{k}\exp\left(-\frac{1}{\epsilon}|c_j^{(t)} - a_i|\right)}, \quad i = 1, \dots, m. \tag{14}$$

Therefore, we can assume that each component $c_s^{(t+1)}$ of the next approximation $\theta^{(t+1)}$ is obtained as a solution of a weighted median problems [35, 40]

$$c_s^{(t+1)} = \operatorname*{argmin}_{\zeta\in\mathbb{R}} g_s(\zeta; \theta^{(t)}), \qquad s = 1, \dots, k, \tag{15}$$

where

$$g_s \colon \mathbb{R} \to \mathbb{R}_+, \quad g_s(\zeta; \theta^{(t)}) = \sum_{i=1}^{m} w_i^{(s)}(\theta^{(t)})|\zeta - a_i|.$$

Note that g_s are continuous, but nondifferentiable convex functions.

Let $g(\,\cdot\,; \theta^{(t)}) \colon \mathbb{R}^k \to \mathbb{R}_+$ be a convex function defined by

$$g(\theta; \theta^{(t)}) = \sum_{s=1}^{k} g_s(c_s; \theta^{(t)}), \qquad \theta = (c_1, \dots, c_k). \tag{16}$$

Because of convexity of the function g, there exists

$$\theta^{(t+1)} = (c_1^{(t+1)}, \dots, c_k^{(t+1)}) = \operatorname*{argmin}_{\theta\in I^k} g(\theta; \theta^{(t)}), \tag{17}$$

where

$$c_s^{(t+1)} = \operatorname*{argmin}_{c_s\in\mathbb{R}} g_s(c_s; \theta^{(t)}) = \operatorname{med}\left(w^{(s)}(\theta^{(t)}), \mathcal{A}\right), \quad s = 1, \dots, k. \tag{18}$$

In that way we defined the iterative process associating a k-tuple $\theta^{(t)}$ with a k-tuple $\theta^{(t+1)}$.

[3] Generally, a median of the data can be some real number a_ν or any number from segment $[a_{\nu-1}, a_\nu]$. In that second case, under the term median of the data we imply the right edge a_ν of the interval.

Remark 1. Because of the weighted median of data properties [40], we can suppose that $\theta^{(t)} \in \mathcal{A}^k$, i.e. $c_s^{(t)} \in \mathcal{A}$ for all $s = 1, \dots, k$. This means that the iterative process (17)-(18) can be defined in such a way that it searches for the minimum of Φ_ϵ among points of the set $\mathcal{A}^k$.

Also, since the function Φ given by (7) is a piecewise linear function, it can always be expected that its global minimum is attained at some point from $\mathcal{A}^k$.

Furthermore, since Φ and Φ_ϵ are symmetric functions, if $\theta^\star = (c_1^\star, \dots, c_k^\star)$ minimizes the functional Φ_ϵ, and $\tilde{\theta}$ is an arbitrary permutation of $\theta^\star$, then also $\tilde{\theta}$ minimizes Φ_ϵ. This means that the iterative process could be restricted to the set

$$\mathcal{C} = \{(c_1, \dots, c_k) \in I^k : \alpha \le c_1 \le c_2 \le \cdots \le c_k \le \beta\}.$$

Note also that the iterative procedure (17)-(18) can be constructed as a Gauss-Seidel iterative procedure, and in this way it accelerates the process even more.

3.1 Convergence of the iterative process

The following proposition can be checked easily.

Proposition 1.

(i) *For every* $i = 1, \dots, m$ *and an arbitrary* $\theta \in \mathbb{R}^k$, *the sequence of weights* $w_i^{(s)}(\theta)$, $s = 1, \dots, k$, *satisfies* $0 < w_i^{(s)}(\theta) < 1$.

(ii) *For an arbitrary* $\theta^{(0)} \in I^k$, *the sequence* $\left(\theta^{(t)}\right)$, *defined by the iterative process* (17)-(18), *remains in* I^k, *and hence it is bounded.*

Proposition 2. *Let* $\theta^{(0)} \in I^k$ *be an arbitrary point, let the sequence* $\left(\theta^{(t)}\right)$ *be given by the iterative process* (17)-(18), *and let* $\Phi_\epsilon \colon I^k \to \mathbb{R}$ *be the functional given by* (8).

If $\theta^{(t+1)} \neq \theta^{(t)}$, *then* $\Phi_\epsilon(\theta^{(t+1)}) < \Phi_\epsilon(\theta^{(t)})$.

Proof. The function $\theta \mapsto g(\theta; \theta^{(t)})$ defined by (16) is a convex function, and its minimizer is $\theta^{(t+1)}$. By our assumption $\theta^{(t+1)} \neq \theta^{(t)}$, and therefore

$$g(\theta^{(t+1)}; \theta^{(t)}) \le g(\theta^{(t)}; \theta^{(t)}). \tag{19}$$

On the other hand, by Lemma 2, the function $\psi \colon \mathbb{R}^k \to \mathbb{R}$ given by $\psi(x) = \ln \sum\limits_{s=1}^k e^{-x_s}$, is a convex function which satisfies

$$\psi(y) - \psi(x) \le \sum_{s=1}^k (x_s - y_s)\mu_s, \qquad \mu_s = e^{-y_s} \left(\sum_{j=1}^k e^{-y_j} \right)^{-1}$$

for all $x, y \in \mathbb{R}^k$. In particular, for points x, y with components $x_s = \frac{1}{\epsilon}|c_s^{(t+1)} - a_i|$, $y_s = \frac{1}{\epsilon}|c_s^{(t)} - a_i|$, one gets

$$\ln \sum_{s=1}^k e^{-\frac{1}{\epsilon}|c_s^{(t)} - a_i|} - \ln \sum_{s=1}^k e^{-\frac{1}{\epsilon}|c_s^{(t+1)} - a_i|} \le \frac{1}{\epsilon} \sum_{s=1}^k \left(|c_s^{(t+1)} - a_i| - |c_s^{(t)} - a_i| \right) \mu_s, \tag{20}$$

where, cf. (14),

$$\mu_s = \frac{\exp(-\frac{1}{\epsilon}|c_s^{(t)} - a_i|)}{\sum_{j=1}^k \exp(-\frac{1}{\epsilon}|c_j^{(t)} - a_i|)} = w_i^{(s)}(\theta^{(t)}).$$

Adding up (20) for $i = 1, \dots, m$ and multiplying by ϵ, using (19) we obtain

$$\begin{aligned}\Phi_\epsilon(\theta^{(t+1)}) - \Phi_\epsilon(\theta^{(t)}) &\le \sum_{i=1}^m \sum_{s=1}^k \left(|c_s^{(t+1)} - a_i| - |c_s^{(t)} - a_i|\right) w_i^{(s)}(\theta^{(t)}) \\ &= g(\theta^{(t+1)}; \theta^{(t)}) - g(\theta^{(t)}; \theta^{(t)}) \le 0.\end{aligned}$$ □

Furthermore, note that the functional $\Phi_\epsilon \colon I^k \to \mathbb{R}_+$ being Lipschitz (Lemma 1) is obviously locally Lipschitz, hence by [6, 7, 29] we have a well defined Clarke's generalized subdifferential

$$\partial\Phi_\epsilon(\theta) = \operatorname{conv}\left\{\lim_{i\to\infty} (\nabla\Phi_\epsilon(\theta_i))^T : (\theta_i) \text{ sequence in } S \text{ such that } \lim_{i\to\infty} \theta_i = \theta\right\},$$

where S is a set of all points in I^k at which $\nabla\Phi_\epsilon$ exists and it is bounded. In addition, if a locally Lipschitz functional $\Phi_\epsilon \colon I^k \to \mathbb{R}_+$ attains its local minimum in $\theta^\star \in I^k$, then $0 \in \partial\Phi_\epsilon(\theta^*)$. Conversely, every point $\hat\theta \in I^k$ for which $0 \in \partial\Phi_\epsilon(\hat\theta)$ will be called a *stationary point* of the functional Φ_ϵ, where

$$\partial\Phi_\epsilon(\theta) = \{(u_1, \dots, u_k) \in \mathbb{R}^k \colon u_s = \sum_{i=1}^m w_i^{(s)}(\theta)\sigma_\lambda(c_s, a_i),\ \lambda \in [-1, 1]\}, \tag{21}$$

$$\sigma_\zeta(c, a) = \begin{cases} \operatorname{sign}(c - a), & c \ne a \\ \zeta, & c = a. \end{cases} \tag{22}$$

Theorem 2. *Let $\theta^{(0)} \in I^k$ be an arbitrary point, let the sequence $\left(\theta^{(t)}\right)$ be defined by the iterative process* (17)-(18)*, and let $\Phi_\epsilon \colon I^k \to \mathbb{R}$ be the functional given by* (8)*. Then*

- (*i*) *The sequence $\left(\theta^{(t)}\right)$ has an accumulation point.*
- (*ii*) *The sequence $\left(\Phi_\epsilon^{(t)}\right)$, where $\Phi_\epsilon^{(t)} := \Phi_\epsilon(\theta^{(t)})$, converges.*
- (*iii*) *Every accumulation point $\hat\theta$ of sequence $\left(\theta^{(t)}\right)$ is a stationary point of the functional Φ_ϵ, and it is obtained by the iterative process* (17)-(18) *in finitely many steps, i.e. there exists a $\mu \in \mathbb{N}$, such that $\theta^{(\mu+1)} = \theta^{(\mu)} = \hat\theta$.*
- (*iv*) *If $\hat\theta_1$ and $\hat\theta_2$ are two accumulation points of the sequence $\left(\theta^{(t)}\right)$, then $\Phi_\epsilon(\hat\theta_1) = \Phi_\epsilon(\hat\theta_2)$.*

Proof. (*i*) By Proposition 1, the sequence $\left(\theta^{(t)}\right)$ is bounded, and therefore it has an accumulation point.

(*ii*) By Proposition 2, the sequence $\left(\Phi_\epsilon^{(t)}\right)$ is monotonously decreasing, and by Theorem 1, the functional Φ_ϵ is bounded below. Therefore, there exists a $\Phi_\epsilon^\star$, such that $\Phi_\epsilon^\star = \lim_{t\to\infty} \Phi_\epsilon^{(t)}$.

(iii) Since the sequence $\left(\Phi_\epsilon^{(t)}\right)$ converges and $\theta^{(t)}$ belongs to a finite set $\mathcal{A}^k$, there exists a $\mu \in \mathbb{N}$ such that $\Phi_\epsilon(\theta^{(\mu+1)}) = \Phi_\epsilon(\theta^{(\mu)})$. According to Proposition 2, we have

$$\theta^{(\mu+1)} = \theta^{(\mu)} = \hat{\theta}. \tag{23}$$

Because $\theta^{(\mu+1)} = \operatorname*{argmin}_{\theta \in I^k} g(\theta; \theta^{(\mu)})$, we conclude that $0 \in \partial g(\theta^{(\mu+1)}; \theta^{(\mu)})$, where $\partial g(\theta; \theta^{(t)})$ is a Clarke's generalized subdifferential of the function g at the point $\theta = (c_1, \ldots, c_k)$,

$$\partial g(\theta; \theta^{(t)}) = \left\{ (u_1, \ldots, u_k) \in \mathbb{R}^k \colon u_s = \sum_{i=1}^m w_i^{(s)}(\theta^{(t)}) \sigma_\lambda(c_s, a_i),\ \lambda \in [-1, 1] \right\}, \tag{24}$$

where the function σ_ζ is given by (22). From (23) it follows that

$$\begin{aligned} 0 \in \partial g(\theta^{(\mu+1)}; \theta^{(\mu)}) &= \partial g(\theta^{(\mu)}; \theta^{(\mu)}) \\ &= \left\{ (u_1, \ldots, u_k) \in \mathbb{R}^k \colon u_s = \sum_{i=1}^m w_i^{(s)}(\theta^{(\mu)}) \sigma_\lambda(c_s^{(\mu)}, a_i),\ \lambda \in [-1, 1] \right\}, \end{aligned}$$

which coincides with the Clarke's generalized subdifferential $\partial \Phi_\epsilon(\theta^{(\mu)})$ of the functional Φ_ϵ given by (21), at the point $\theta^{(\mu)}$. Therefore, $\theta^{(\mu)} = \hat{\theta}$ is a stationary point of the functional Φ_ϵ.

(iv) Let $\left(\theta_1^{(t)}\right)$ and $\left(\theta_2^{(t)}\right)$ be two subsequences of the sequence $\left(\theta^{(t)}\right)$, such that $\hat{\theta}_1 = \lim_{t \to \infty} \theta_1^{(t)}$ and $\hat{\theta}_2 = \lim_{t \to \infty} \theta_2^{(t)}$. Since the sequence $\left(\Phi_\epsilon^{(t)}\right)$ converges, we have

$$\Phi_\epsilon(\hat{\theta}_1) = \lim_{t \to \infty} \Phi_\epsilon(\theta_1^{(t)}) = \lim_{t \to \infty} \Phi_\epsilon(\theta_2^{(t)}) = \Phi_\epsilon(\hat{\theta}_2). \qquad \square$$

In the next corollary we discuss a special choice of the initial approximation.

Corollary 1. *If* $\theta^{(0)} = (c_1^{(0)}, \ldots, c_k^{(0)})$ *is an initial approximation such that* $c_1^{(0)} = \ldots = c_k^{(0)}$, *then the first step of the iterative process* (17)-(18) *will give the stationary point*

$$\hat{\theta} = (\operatorname{med}(\mathcal{A}), \ldots, \operatorname{med}(\mathcal{A})),$$

where $\operatorname{med}(\mathcal{A})$ *is the ordinary median of the set* $\mathcal{A} = \{a_1, \ldots, a_m\}$.

Proof. If $c_1^{(0)} = \ldots = c_k^{(0)}$, then $w^{(s)}(\theta^{(0)}) = \left(\frac{1}{k}, \ldots, \frac{1}{k}\right) \in \mathbb{R}^m$, $s = 1, \ldots, k$, and therefore

$$c_s^{(1)} = \operatorname{med}\left(w^{(s)}(\theta^{(0)}), \mathcal{A}\right) = \operatorname{med}(\mathcal{A}), \quad s = 1, \ldots, k.$$

Consequently, for every $t = 0, 1, 2, \ldots$ we have $w^{(s)}(\theta^{(t)}) = \left(\frac{1}{k}, \ldots, \frac{1}{k}\right) \in \mathbb{R}^m$, and

$$c_s^{(t+1)} = \operatorname{med}\left(w^{(s)}(\theta^{(t)}), \mathcal{A}\right) = \operatorname{med}(\mathcal{A}), \quad s = 1, \ldots, k. \qquad \square$$

Note that because of $\Phi(c_1, \ldots, c_k) \leq \Phi(c_s, \ldots, c_s)$, for all $s = 1, \ldots, k$, it is not real to expect that the choice of an initial approximation as in Corollary 1 would give a global minimizer of the functional Φ_ϵ.

3.2 One-dimensional l_1-clustering algorithm

Theorem 2 and Corollary 1 show that given an initial approximation $\theta^{(0)} \in I^k$, the iterative process (17)-(18) always converges to some stationary point, which needs not be unique. Therefore, special attention should be paid to the choice of the initial approximation of centers $\theta^{(0)} = (c_1^{(0)}, \ldots, c_k^{(0)})$. It is shown that a very good initial approximation $\theta^{(0)}$ can be obtained in the following way. Sorted data $a_1 \leq a_2 \leq \cdots \leq a_m$ should be divided into k approximately equal subsets and for each of them the median should be calculated. In most of the cases, with such initial approximation our algorithm gives the best partition in only few steps. However, a question remains open as to how to test whether a stationary point obtained by the iterative process (17)-(18) is a global minimizer of the functional Φ_ϵ [9, 14, 15, 17].

In addition, Theorem 2 (*iii*) gives a criterion for terminating the iterative process (17)-(18). Next, we give the following algorithm.

Algorithm 1. (One-dimensional l_1-clustering)

Step 1: Input $m \geq 1$, $1 \leq k \leq m$, $\epsilon > 0$, $\mathcal{A} = \{a_i \in \mathbb{R} : i = 1, \ldots, m\}$, and choose an initial approximation of centers $\theta^{(0)} = (c_1^{(0)}, \ldots, c_k^{(0)})$;

Step 2: For all $s = 1, \ldots, k$ define an m-tuple $w^{(s)}$ with components

$$w_i^{(s)} = \frac{\exp\left(-\frac{1}{\epsilon}|c_s^{(0)} - a_i|\right)}{\sum\limits_{j=1}^{k} \exp\left(-\frac{1}{\epsilon}|c_j^{(0)} - a_i|\right)}, \quad i = 1, \ldots, m;$$

Step 3: For all $s = 1, \ldots, k$ solve the weighted median problem[4]

$$g_s(\zeta) = \sum_{i=1}^{m} w_i^{(s)} |\zeta - a_i| \to \min_{\zeta},$$

and set $\theta^{(1)} = (c_1^{(1)}, \ldots, c_k^{(1)})$, where $c_s^{(1)} = \operatorname{argmin} g_s(\zeta)$;

Step 4: If $\theta^{(1)} = \theta^{(0)}$, Go To Step 5; Else set $\theta^{(0)} = \theta^{(1)}$ and go to Step 2;

Step 5: According to the minimal distance principle, define a partition $\Pi = \{\pi_1, \ldots, \pi_k\}$ with centers $c_1^{(1)}, \ldots, c_k^{(1)}$:

$$\pi_1 = \{a_i \in \mathcal{A} : |a_i - c_1^{(1)}| \leq |a_i - c_l^{(1)}|,\ l = 1, \ldots, k\},$$

$$\pi_j = \{a_i \in \mathcal{A} \setminus \bigcup_{s=1}^{j-1} \pi_s : |a_i - c_j^{(1)}| \leq |a_i - c_l^{(1)}|,\ \forall\, l = 1, \ldots, k\}, \quad j = 2, \ldots, k.$$

Remark 2. Let us mention one possibility for the choice of the smoothing parameter $\epsilon > 0$. If we want a relative deviation $\frac{\Phi(\theta^{(0)}) - \Phi_\epsilon(\theta^{(0)})}{\Phi(\theta^{(0)})}$ between the function Φ and Φ_ϵ in the initial approximation $\theta^{(0)}$ to be less than the number $\delta > 0$ set in advance, then by using Theorem 1 we obtain

$$\epsilon \leq \delta \tfrac{\Phi(\theta^{(0)})}{m \ln k}. \tag{25}$$

[4] *Mathematica*-code for solving a weighted median problem is available at: `http://www.mathos.hr/seminar/Software.html`

4 Numerical examples

In order to visualize and analyze the problem and the proposed method from Section 3 further, we consider a simple example where the data set $\mathcal{A}$ consists of grade point averages (GPA) of successful second year students majoring in mathematics at the Department of Mathematics, University of Osijek (see Table 1).

Student	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}
GPA	2.2	2.35	2.5	2.64	2.85	3.	3.25	3.35	3.4	3.54
Student	s_{11}	s_{12}	s_{13}	s_{14}	s_{15}	s_{16}	s_{17}	s_{18}	s_{19}	s_{20}
GPA	3.54	3.7	3.72	3.72	3.8	3.85	3.95	4.05	4.15	4.2
Student	s_{21}	s_{22}	s_{23}	s_{24}	s_{25}	s_{26}	s_{27}	s_{28}	s_{29}	s_{30}
GPA	4.2	4.3	4.41	4.41	4.54	4.6	4.6	4.65	4.84	5.

Table 1: Students' GPAs

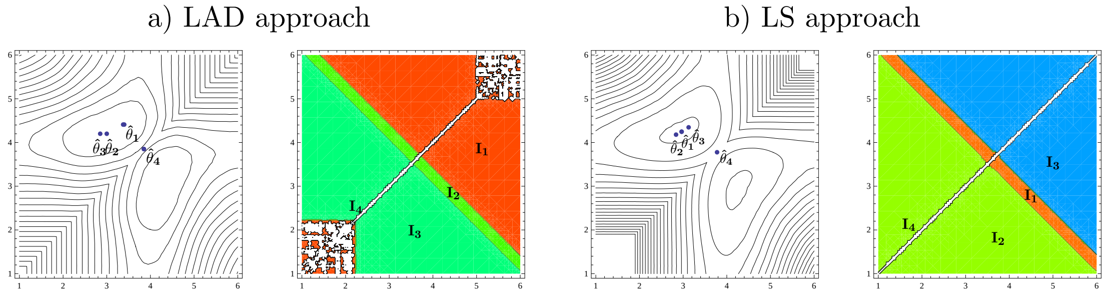
Example 1. *We are going to split the set $\mathcal{A}$ of GPAs of successful students shown in Table 1 into two clusters with objective function $\Phi_\epsilon\colon I^2 \to \mathbb{R}_+$, $I = [2.2, 5]$ given by* (8) *for $\epsilon = 0.005$.*

Looking at the `ContourPlot` of the objective function Φ_ϵ (see Fig. 1), one immediately notices the symmetry property of the functional Φ_ϵ: $\Phi_\epsilon(c_1, c_2) = \Phi_\epsilon(c_2, c_1)$.

Application of our Algorithm 1 to various choices of initial approximations $\theta^{(0)} \in I^2$ results in four different stationary points $\hat{\theta}_i$, $i = 1, \dots, 4$ (see Fig. 1), at which the functional Φ_ϵ assumes different values shown in Table 2.

i	$\theta_i^{(0)}$	$\hat{\theta}_i$	$\Phi_\epsilon(\hat{\theta}_i)$
1	$\theta^{(0)} \in I_1$	$\{(\vartheta_1, 4.41) \in \mathbb{R}^2 : \vartheta_1 \in [a_8, a_9] = [3.35, 3.4]\}$	10.51
2	$\theta^{(0)} \in I_2$	$(3.00, 4.20)$	10.75
3	$\theta^{(0)} \in I_3$	$(2.85, 4.20)$	10.84
4	$\theta^{(0)} \in I_4$	$\{(\vartheta_1, \vartheta_2) \in \mathbb{R}^2 : \vartheta_i \in [a_{15}, a_{16}] = [3.8, 3.85]\}$	18.086

Table 2: Stationary points of the functional Φ_ϵ


a) LAD approach

b) LS approach

Figure 1: `ContourPlot` with stationary points of the objective function and the appropriate area of choice of initial approximations

Fig. 1a shows four different regions I_1, I_2, I_3, and I_4 for choosing initial approximations, starting from which Algorithm 1 yields the corresponding stationary points $\hat{\theta}_1, \dots, \hat{\theta}_4$, with different values of objective function Φ_ϵ. Note that region I_4 is located around the bisector of the first quadrant and as shown in Corollary 1, the iterative process always terminates at the ordinary median of data $a_1, \dots, a_m$, and in this way the global minimizer of the functional Φ_ϵ is not attained. It is natural that in order to choose the initial approximation in the data region $\theta^{(0)} \in \operatorname{conv}(\mathcal{A}) \times \operatorname{conv}(\mathcal{A})$. As the calculations show, the functional Φ_ϵ has four stationary points, and it attains its global minimum $\Phi_\epsilon(\theta^\star)$ at any $\theta^\star = (\theta_1^\star, \theta_2^\star)$, $\theta_1^\star \in [a_8, a_9] = [3.35, 3.4]$, $\theta_2^\star = a_{23} = 4.41$, which is obtained by our Algorithm 1, where the initial approximation has to be chosen in region I_1. As proposed at the beginning of Section 3.2, if the data are divided into two equal parts in which the median is calculated, we will obtain a very good initial approximation $\theta^{(0)} = (3.5, 4.41)$, which belongs to region I_1. Moreover, it turns out that our Algorithm 1 is very efficient, and it usually terminates after only a few iterations.

For the purpose of comparison, as already mentioned, the same problem in [21] is solved in the Least Squares sense by minimizing the functional (6) with $d(x, y) = (x - y)^2$. In this case the objective function is differentiable, and the corresponding iterative process again yields four different stationary points $\hat{\theta}_1, \dots, \hat{\theta}_4$ (see Fig. 1b), but the initial approximation region which yields the global minimum is a narrow strip, denoted by I_1. In this case the iterative process terminates when the distance $\|\theta^{(t+1)} - \theta^{(t)}\|_2$ becomes smaller than some prescribed $\eta > 0$. From Fig. 1 it can be seen that the red area of choice of good initial approximations for which the LAD algorithm converges to the global minimum of the objective function is significantly larger than the analogous area for the LS algorithm. Other numerical examples also point to this property. This means that the probability of a random choice of a good initial approximation is significantly bigger in the case of the LAD algorithm.

In the next example we will test the proposed method on a greater number of data that should be grouped into a greater number of clusters. We will thereby compare calculation performances of the algorithm described in [21, 39] for solving the optimization problem (6) (LS algorithm) and the proposed algorithm for solving the optimization problem (8) (LAD algorithm).

Motivation for this example originates from the problem of determining spatial clusters of accidents along a continuous highway "New Jersey Turnpike" using different objectives [18]. Identifying such spatial clusters of accidents according to different objectives can provide useful insights into various operational and safety issues. On the basis of the location of these accidents we determine optimal clusters of accidents.

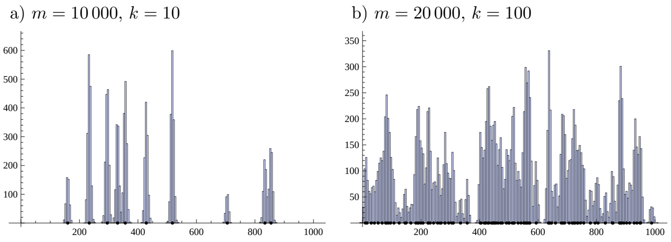


Figure 2: Centers and distribution of data points

Example 2. *Instead of historical crash data-sets we simulate random locations of accidents in the following way. In interval $I = [0, 1000]$ we choose k centers $c_1, \dots, c_k \in I$ at random (locations k of most frequently occurring accidents). The data set $\mathcal{A}$ containing m randomly chosen real numbers from the interval I (locations m of observed accidents) is generated in the following way:*

(i) let $i_1, \dots, i_k$ be randomly generated integers such that $\sum_{s=1}^{k} i_s = m$;

(ii) in the neighbourhood of the center c_s we generate a set A_s, which consists of i_s random real numbers from $\mathcal{N}(c_s, 5)$ (variance $\sigma^2 = 5$ is relatively very small, which enables simulation of accidents in the immediate neighbourhood of centers $c_1, \dots, c_k$);

(iii) $\mathcal{A} = \bigcup_{s=1}^{k} A_s$.

Fig. 2 shows centers and distribution of data points for two different choices of pairs (m, k).

First note that for each fixed $i \in \{1, \dots, m\}$, there exists $r \in \{1, \dots, k\}$, such that $|c_r^{(t)} - a_i| \le \min\limits_{s \ne r} |c_s^{(t)} - a_i|$, and therefore the following holds

$$w_i^{(s)}(\theta^{(t)}) = \begin{cases} \dfrac{1}{1 + \sum\limits_{j=1, j\ne r}^{k} \exp\Big(-\frac{1}{\epsilon}(|c_j^{(t)} - a_i| - |c_r^{(t)} - a_i|)\Big)} & \text{if } s = r, \\ \dfrac{\exp\Big(-\frac{1}{\epsilon}(|c_s^{(t)} - a_i| - |c_r^{(t)} - a_i|)\Big)}{1 + \sum\limits_{j=1, j\ne r}^{k} \exp\Big(-\frac{1}{\epsilon}(|c_j^{(t)} - a_i| - |c_r^{(t)} - a_i|)\Big)} & \text{if } s \ne r, \end{cases}$$

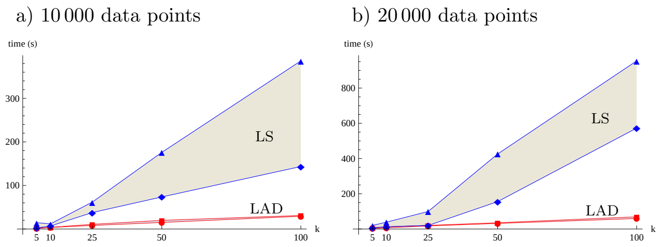
from where

$$\lim_{\epsilon \to 0^+} w_i^{(s)}(\theta^{(t)}) = \begin{cases} 1 & \text{if } s = r, \\ 0 & \text{if } s \ne r. \end{cases} \tag{26}$$

This means that $\epsilon > 0$ can always be chosen such that the data a_i belongs to the closest center. This means that in Step 3 of Algorithm 1 instead of solving the weighted median problem for all data points, for every $s = 1, \dots, k$ we can calculate the ordinary median only of the data closest to the center $c_s^{(0)}$.

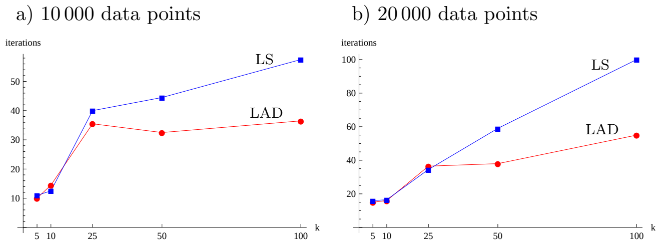
By using previous considerations, we are going to split the set $\mathcal{A}$ into k clusters with objective function Φ_ϵ given by (8) (LAD approach) and with objective function F_ϵ given by (6) (LS approach) for $\epsilon = 0.005$. The experiment will be conducted by taking $m \in \{1000, 5000, 10000, 20000\}$ and $k \in \{5, 10, 25, 50, 100\}$ (see [25]). Since the LS and LAD algorithm converges to some local minima of the corresponding objective function, similarly to [25] each algorithm will start running 10 times with some different random initializations on each combinations of m and k.

Fig. 3 and Fig. 4 show movement of the `CPU` times in seconds for each running depending on the number of centers and movement of the number of iterations for the LAD and LS algorithm on a Pentium `M` processor with `1.4 GHz`, respectively. The grey area in Fig. 3 shows a range of the `CPU` time required for the execution of the LS algorithm for various choices of initial approximations of centers. It can be noticed that, in contrast to the LAD algorithm, the `CPU` time required for the execution of the LS algorithm depends heavily on the selected initial approximation. As can be seen in Fig. 4, an increase in the number of clusters also causes the number of necessary iterations of the LS algorithm to increase in relation to the LAD algorithm. We will also try to estimate the error during reconstruction of centers


a) 10 000 data points

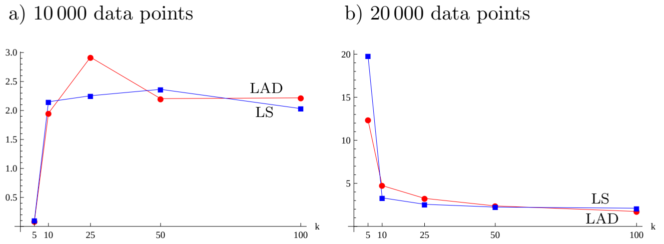
b) 20 000 data points

Figure 3: CPU time (in seconds) necessary for the execution of the LAD and LS algorithm


a) 10 000 data points

b) 20 000 data points

Figure 4: Number of iterations necessary for the execution of the LAD and LS algorithm


a) 10 000 data points

b) 20 000 data points

Figure 5: Reconstruction error of centers for LAD and LS algorithm

by applying these two algorithms. Let $\theta = (c_1, \ldots, c_k)$ be a vector whose components are original centers, and let $\theta^\star = (c_1^\star, \ldots, c_k^\star)$ be a vector whose components are estimated centers. One possibility for error assessment is the Hausdorff distance between these two vectors [41], which is given by the following formula

$$d(\theta, \theta^\star) = \max\left\{ \max_{i=1,\ldots,k} \left(\min_{j=1,\ldots,k} |c_i - c_j^\star| \right), \max_{j=1,\ldots,k} \left(\min_{i=1,\ldots,k} |c_i - c_j^\star| \right) \right\}. \tag{27}$$

In this way we would point out the maximum error that might occur. We believe that, for the purpose of comparing the LAD and the LS algorithm, a better indicator is average distance of each center $c_i^\star$ to the closest of centers $c_1, \ldots, c_k$ (see Fig. 5). A significant difference between the LAD and the LS algorithm with respect to the reconstruction quality of centers is not indicated.

5 Conclusions

In this paper we consider a one-dimensional data clustering problem in case outliers are to be expected among the data. Usage of the Least Absolute Deviations-optimality criterion is proposed for solving this problem. By knowing a good initial approximation, the proposed method can provide acceptable solutions in only a few steps. In case we do not have a good initial approximation, what is usually recommended [25] are multi-run algorithms with various random initializations, as done in Example 2. It is shown that the LS and the LAD algorithm give approximately equally good reconstructions of centers. On the other hand, numerous numerical experiments show a series of advantages of the LAD approach, such as:

(*i*) the probability of a random choice of a good initial approximation is significantly larger in the case of the LAD algorithm (see Example 1);

(*ii*) increasing the number of clusters causes an increase in the number of necessary iterations of the LS algorithm in relation to the LAD algorithm;

(*iii*) in contrast to the LAD algorithm, the CPU time required for the execution of the LS algorithm depends heavily on the initial approximation;

(*iv*) generally, the LAD approach ignores outliers among the data [8, 33], while the LS approach stresses them.

Solving the global optimization problem is a very common issue in recent literature. An overview of papers published lately in this field can be found in [14]. One approach to the solution of this problem by applying interval analysis can be found in [17]. Since our functional Φ_ϵ given by (8) satisfies a Lipschitz condition, global optimization methods for Lipschitz functions [15, 30] are especially interesting, among which the most popular is the `DI`(viding)`RECT`(angles) algorithm [13, 20].

References

[1] A. Banerjee, S. Merugu, I. S. Dhillon, J. Ghosh, *Clustering with Bregman divergences*, Journal of Machine Learning Research **6**(2005), 1705–1749

[2] A. Ben-Israel, C. Iyigun, *Probabilistic D-clustering*, Journal of Classification **25**(2007) `DOI: 10.1007/s00357-007-0021-y`

[3] D. L. Boley, *Principal direction divisive partitioning*, Data Mining and Knowledge Discovery 2(1998), 325–344

[4] D. L. Boyd, L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, 2004.

[5] W. A. Chaovalitwongse, S. Butenko, P. M. Pardalos, *Clustering Challenges in Biological Networks*, World Scientific, 2009.

[6] F. H. Clarke, *Generalized gradients an applications*, Transactions of the America Mathematica Society, **205**(1975), 247–262

[7] F. H. Clarke, *Optimization and Nonsmooth Analysis*, SIAM, Philadelphia, 1990.

[8] R. Cupec, R. Grbić, K. Sabo, R. Scitovski, *Three points method for searching the best least absolute deviations plane*, Applied Mathematics and Computation, **215**(2009), 983–994

[9] E. Demidenko, *Criteria for unconstrained global optimization*, J. Optim. Theory Appl. **136**(2008), 375–395

[10] I. S. Dhillon, S. Mallela, R. Kumar, *A divisive information-theoretic feature clustering algorithm for text classification*, Journal of Machine Learning Research **3**(2003), 1265–1287.

[11] I. S. Dhillon, Y. Guan, B. Kulis, *Kernel k-means, spectral clustering and normalized cuts*, Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), August 22–25, 2004, Seattle, Washington, USA, 551–556, 2004

[12] E. Domínguez, J. Muñoz, *Applying bio-inspired techniques to the p-median problem*, IWANN 2005; Computational Intelligence Bioinspired Syst., 8th Int. Workshop Artificial Neural Networks, Lecture Notes in Computer Science, Springer-Verlag, Berlin, 2005, 67 – 74

[13] D. E. Finkel, C. T. Kelley, *Additive scaling and the* `DIRECT` *algorithm*, J. Glob. Optim. **36**(2006), 597-–608

[14] C. A. Floudas, C. E. Gounaris, *A review of recent advances in global optimization*, J. Glob. Optim. **45**(2009), 3—38

[15] M. Gaviano, D. Lera, *A global minimization algorithm for Lipschitz functions*, Optimization Letters **2**(2008) 1—13

[16] G. Gan, C Ma, J. Wu, *Data Clustering: Theory, Algorithms, and Applications*, SIAM, Philadelphia, 2007.

[17] E. R. Hansen, G. W. Walster, *Global Optimization Using Interval Analysis.* Marcel Dekker, New York, Second Edition, Revised and Expanded, 2004.

[18] C. Iyigun, *Probabilistic Distance Clustering*, Dissertation, Graduate School – New Brunswick, Rutgers, 2007.

[19] C. Iyigun, A. Ben-Israel, *A generalized Weiszfeld method for the multi-facility location problem*, Operations Research Letters **38**(2010), 207–214

[20] D. R. Jones, C. D. Perttunen, B. E. Stuckman, *Lipschitzian optimization without the Lipschitz constant*, JOTA **79**(1993), 157-181

[21] J. Kogan, *Introduction to Clustering Large and High-Dimensional Data*, Cambridge University Press, 2007.

[22] J. Kogan, C. Nicholas, M. Wiacek, *Hybrid Clustering of large high dimensional data*, In M. Castellanos and M. W. Berry (Eds.), Proceedings of the Workshop on Text Mining, SIAM, 2007.

[23] J. Kogan, M. Teboulle, *Scaling clustering algorithms with Bregman distances.* In: M. W. Berry and M. Castellanos (Eds.), Proceedings of the Workshop on Text Mining at the Sixth SIAM International Conference on Data Mining, 2006.

[24] J. Kogan, C. Nicholas, M. Wiacek, *Hybrid clustering with divergences.* In: M. W. Berry and M. Castellanos (Eds.), Survey of Text Mining: Clustering, Classification, and Retrieval, Second Edition, Springer, 2007.

[25] F. Leisch, *A toolbox for K-centroids cluster analysis*, Computational Statistics & Data Analysis **51**(2006), 526 – 544

[26] D. Littau, D. L. Boley, *Clustering very large data sets with PDDP*. In J. Kogan, C. Nicholas, M. Teboulle (eds), *Grouping Multidimensional Data: Recent Advances in Clustering*, 99–126, Springer-Verlag, New York, 2006.

[27] S. Pan, J.,S. Chen, *Two unconstrained optimization approaches for the Euclidean k-centrum location problem*, Applied Mathematics and Computation **189**(2007), 1368–1383

[28] P. M. Pardalos, P. Hansen, *Data Mining and Mathematical Programming*, American Mathematical Society, Providence, 2008.

[29] J. Petrić, S. Zlobec, *Nonlinear Programming* (in Croatian), Naučna knjiga, Beograd, 1989.

[30] S. A. Piyavskij, *An algorithm for finding the absolute extrernum of a function*, Journal Computational Mathematics and Mathematical Physics **12**(1972), 888-896. (in Russian)

[31] J. Reese, *Solution methods for the p-median problem: an annotated bibliography*, Published online in Wiley InterScience, Wiley, 2006.

[32] A. M. Rodrígues-Chia, I. Espejo, Z. Drezner, *On solving the planar k-centrum problem with Euclidean distances*, European Journal of Operational Research **207**(2010), 1169-1186

[33] K. Sabo, R. Scitovski, I. Vazler, *Searching for a best LAD-solution of an overdetermined system of linear equations motivated by searching for a best LAD-hyperplane on the basis of given data*, J. Optim. Theory Appl. **149**(2011), 293–314

[34] K. Sabo, R. Scitovski, I. Vazler, M. Zekić-Sušac, *Mathematical models of natural gas consumption*, Energy Conversion and Management **52**(2011), 1721–1727

[35] K. Sabo, R. Scitovski, *The best least absolute deviations line – properties and two efficient methods*, ANZIAM Journal **50**(2008), 185–198

[36] A. Schöbel, D. Scholz, *The big cube small cube solution method for multidimensional facility location problems*, Computers & Operations Research **37**(2010), 115–122

[37] H. Späth, *Cluster-Formation und Analyse*, R. Oldenburg Verlag, München, 1983.

[38] Z. Su, J. Kogan, C. Nicholas, *Constrained clustering with k-means type algorithms*, In M.W. Berry, J. Kogan (Eds.), *Text Mining Applications and Theory*, 81–103, Willey, Chichester, 2010.

[39] M. Teboulle, *A unified continuous optimization framework for center-based clustering methods*, Journal of Machine Learning Research **8**(2007), 65–102

[40] I. Vazler, K. Sabo, R. Scitovski, *Weighted median of the data in solving least absolute deviations problems*, Communications in Statistics - Theory and Methods, to appear in 2011

[41] E. P. Vivek, N. Sudha, *Robust Hausdorff distance measure for face recognition*, Pattern Recognition **40**(2007), 431—442