A Novel Class of Globally Convergent Algorithms For Clustering Problems

Sergey Voldman

Raymond and Beverly Sackler Faculty of Exact Sciences

Tel-Aviv University

06.04.2016

Under the supervision of

Prof. Marc Teboulle (Tel Aviv University)
and Prof. Shoham Sabach (Technion)

Goal and Outline

Develop and analyze two center-based clustering algorithms each with different distance-like function.

Outline

- Introduction to the clustering problem.
- Introduction to the convergence methodology.
- Clustering with the squared Euclidean norm: KPALM algorithm and its analysis.
- Clustering with the Euclidean norm: ε-KPALM algorithm and its analysis.
- Numerical results of the proposed algorithms.

The Clustering Problem

- Clustering is fundamental in fields such as machine learning, data mining, etc.
- The clustering problem focused a lot of research and there are many algorithms tackling it, such as k-means, Expectation-Maximization and others.
- It has been shown that the clustering problem is NP-hard.
- Let $\mathcal{A} = \{a^1, a^2, \ldots, a^m\} \subset \mathbb{R}^n$ set of points, and $1 < k < m$ a given number of clusters.
- The goal is to partition the data $\mathcal{A}$ into k subsets $\{\mathcal{C}^1, \mathcal{C}^2, \ldots, \mathcal{C}^k\}$ called clusters.
- Each cluster $\mathcal{C}^l$ is represented by its center $x^l \in \mathbb{R}^n$.
- The clustering problem is given by

$$(P_0) \qquad \min_{x \in \mathbb{R}^{nk}} \left\{ F(x) := \sum_{i=1}^{m} \min_{1 \leq l \leq k} d(x^l, a^i) \right\},$$

with $d(\cdot, \cdot)$ being a distance-like function, such as the squared Euclidean norm.

Problem Reformulation

- Using the fact that

$$\min_{1 \le l \le k} u_l = \min \{ \langle u, v \rangle : v \in \Delta \},$$

where Δ is the simplex in $\mathbb{R}^k$, problem (P_0) can be transformed into

$$(P_1) \qquad \min_{x \in \mathbb{R}^{nk}} \left\{ \sum_{i=1}^{m} \min_{w^i \in \Delta} \langle w^i, d^i(x) \rangle \right\},$$

with $d^i(x) = (d(x^1, a^i), d(x^2, a^i), \ldots, d(x^k, a^i)) \in \mathbb{R}^k, \quad i = 1, 2, \ldots, m.$

- Replacing the constrain $w^i \in \Delta$ by adding the indicator function $\delta_\Delta(\cdot)$ results in

$$(P_2) \qquad \min_{x \in \mathbb{R}^{nk}, w \in \mathbb{R}^{km}} \left\{ \sum_{i=1}^{m} \left(\langle w^i, d^i(x) \rangle + \delta_\Delta(w^i) \right) \right\},$$

where $w = (w^1, w^2, \ldots, w^m) \in \mathbb{R}^{km}$.

- The final version is given by

$$(P) \qquad min \left\{ \Psi(z) := H(w, x) + G(w) \mid z := (w, x) \in \mathbb{R}^{km} \times \mathbb{R}^{nk} \right\},$$

with $H(w, x) = \sum_{i=1}^{m} H^i(w, x) = \sum_{i=1}^{m} \langle w^i, d^i(x) \rangle$ and $G(w) = \sum_{i=1}^{m} G^i(w^i) = \sum_{i=1}^{m} \delta_\Delta(w^i)$.

Convergence Methodology: Goal and Definitions

Given:

Let $\Psi : \mathbb{R}^N \to \overline{\mathbb{R}}$ be a proper, lsc and bounded from below function.

$$(P) \qquad \inf \left\{ \Psi(z) : \ z \in \mathbb{R}^N \right\}.$$

Suppose $\mathcal{A}$ is a generic algorithm which generates a sequence $\{z^k\}_{k \in \mathbb{N}}$ via:

$$z^0 \in \mathbb{R}^N, z^{k+1} \in \mathcal{A}\left(z^k\right), \ k = 0, 1, \ldots .$$

Goal: To prove that whole $\{z^k\}_{k \in \mathbb{N}}$ converges to a critical point of Ψ.

Definition (Limiting Subdifferential $\partial\sigma(x)$)

Let $\sigma : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lower semicontinuous function.
The (Limiting) Subdifferential $\partial\sigma(x)$ is defined via:

$$\begin{aligned} u^* \in \partial\sigma(x) \quad & \text{iff} \quad (x_k, u_k^*) \to (x, u^*) \text{ s.t. } \sigma(x_k) \to \sigma(x) \text{ and} \\ \sigma(u) \quad & \geq \quad \sigma(x_k) + \langle u_k^*, u - x_k \rangle + o(\|u - x_k\|) \end{aligned}$$

- $x \in \mathbb{R}^d$ is a critical point of σ if $\partial\sigma(x) \ni 0$.
- The set of critical points of $\sigma \equiv \text{crit}\,\sigma$.
- $r \in \mathbb{R}$ is a critical value if $\exists x \in \text{crit}\,\sigma : \sigma(x) = r$.

KL property

Denote the following class of concave functions

$$\Phi_\eta = \left\{ \varphi \in C\left([0,\eta), \mathbb{R}_+\right) : \ \varphi \in C^1\left((0,\eta)\right), \ \varphi' > 0, \ \varphi(0) = 0 \right\}.$$

Definition (Kurdyka-Łojasiewicz property)

Let $\sigma : \mathbb{R}^d \to (-\infty, +\infty]$ be proper and lower semicontinuous.

(i) σ admits the KL property at $\overline{u} \in dom\, \partial\sigma := \{u \in \mathbb{R}^d : \ \partial\sigma \neq \emptyset\}$ if there exist $\eta \in (0, +\infty]$, a neighborhood U of $\overline{u}$ and a function $\varphi \in \Phi_\eta$, such that for all

$$u \in U \cap \left\{ x \in \mathbb{R}^d : \ \sigma(\overline{u}) < \sigma(x) < \sigma(\overline{u}) + \eta \right\},$$

the following inequality holds

$$\varphi'(\sigma(u) - \sigma(\overline{u}))\, \mathrm{dist}(0, \partial\sigma(u)) \geq 1,$$

where $\mathrm{dist}(x, S) := \inf\{\|y - x\| : \ y \in S\}$ denotes the distance from $x \in \mathbb{R}^d$ to $S \subset \mathbb{R}^d$.

(ii) If σ satisfy the KL property at each point of *dom* σ then σ is called a *KL function*.

Semi-Algebraic Functions

Theorem (Bolte-Daniilidis-Lewis (2006))

Let $\sigma : \mathbb{R}^d \to \overline{\mathbb{R}}$ *be a proper and lsc function. If* σ *is semi-algebraic then it satisfy the KL property at any point of* $\operatorname{dom} \sigma$*.*

Definition

(i) A subset S of $\mathbb{R}^n$ is a real semi-algebraic set if there exists a finite number of real polynomial functions $g_{ij}, h_{ij} : \mathbb{R}^n \to \mathbb{R}$ such that

$$S = \bigcup_{j=1}^{p} \bigcap_{i=1}^{q} \left\{ u \in \mathbb{R}^n : \; g_{ij}(u) = 0 \text{ and } h_{ij}(u) < 0 \right\}.$$

(ii) A function $\sigma : \mathbb{R}^n \to \overline{\mathbb{R}}$ is called semi-algebraic if its graph

$$\left\{ (u, t) \in \mathbb{R}^{n+1} : \; \sigma(u) = t \right\}$$

is a semi-algebraic subset of $\mathbb{R}^{n+1}$.

The Wealth of Semi-Algebraic Functions

- Real polynomial functions.
- Indicator functions of semi-algebraic sets.
- Finite sums and product of semi-algebraic functions.
- Composition of semi-algebraic functions.
- Sup/Inf type function, *e.g.*, $\sup \{g(u, v) : v \in C\}$ is semi-algebraic when g is a semi-algebraic function and C a semi-algebraic set.
- The function $x \to \operatorname{dist}(x, S)^2$ is semi-algebraic whenever S is a nonempty semi-algebraic subset of $\mathbb{R}^n$.
- $\|\cdot\|_0$ (counts the non-zero values) is semi-algebraic.
- $\|\cdot\|_p$ is semi-algebraic whenever $p > 0$ is rational.

In particular, for distance-like functions $d(x, y) = \|x - y\|^2$ and $d(x, y) = \|x - y\|$ the resulting clustering function defined in (P) is semi-algebraic.

Gradient-Like Descent Sequence

Definition

Let $\sigma : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lower semicontinuous function. A sequence $\{z^k\}_{k\in\mathbb{N}}$ is called a gradient-like descent sequence for σ if for all $k \in \mathbb{N}$ the following two conditions hold:

(C1) Sufficient decrease property: There exists a positive scalar ρ_1 such that

$$\rho_1 \left\|z^{k+1} - z^k\right\|^2 \leq \sigma\left(z^k\right) - \sigma\left(z^{k+1}\right).$$

(C2) A subgradient lower bound for the iterates gap:

- $\{z^k\}_{k\in\mathbb{N}}$ is bounded.
- There exists a positive scalar ρ_2 such that

$$\left\|w^{k+1}\right\| \leq \rho_2 \left\|z^{k+1} - z^k\right\|, \; w^{k+1} \in \partial\sigma\left(z^{k+1}\right).$$

Theorem (Bolte-Shabach-Teboulle (2014))

Let $\sigma : \mathbb{R}^d \to (-\infty, \infty]$ be a proper, lower semicontinuous and semi-algebraic function with $\inf \sigma > -\infty$, and assume that $\{z^k\}_{k\in\mathbb{N}}$ is a gradient-like descent sequence for σ. If $\omega\left(z^0\right) \subset crit(\sigma)$ then the sequence $\{z^k\}_{k\in\mathbb{N}}$ convergences to a critical point z^* of σ.

The Optimization Model

$$(M) \qquad \text{minimize}_{x,y} \Psi(x,y) := f(x) + g(y) + H(x,y)$$

Assumption

(i) $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ *and* $g : \mathbb{R}^m \to \overline{\mathbb{R}}$ *are proper and lsc functions.*

(ii) $H : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ *is a* C^1 *function.*

(iii) *Partial gradients of H are Lipshitz continuous:* $H(\cdot, y) \in C^{1,1}_{L(y)}$ *and likewise* $H(x, \cdot) \in C^{1,1}_{L}(x)$.

- **NO convexity** will be assumed in the objective or/and the constraints (built-in through f and g extended valued).

The Algorithm: Proximal Altertnating Linearization Minimization (PALM)

PALM algorithm is nothing but alternating the classical proximal gradient over the two blocks (x, y). This leads towards the following approximations:

$$\widehat{\Psi}(x, y^k) = \left\langle x - x^k, \nabla_x H(x^k, y^k) \right\rangle + \frac{c_k}{2} \left\| x - x^k \right\|^2 + f(x),$$

$$\widetilde{\Psi}(x^{k+1}, y) = \left\langle y - y^k, \nabla_y H(x^{k+1}, y^k) \right\rangle + \frac{d_k}{2} \left\| y - y^k \right\|^2 + g(y).$$

1. Initialization: start with any $\left(x^0, y^0\right) \in \mathbb{R}^n \times \mathbb{R}^m$.
2. For each $k = 0, 1, \ldots$ generate a sequence $\left\{\left(x^k, y^k\right)\right\}_{k \in \mathbb{N}}$:
   - 2.1. Take $\gamma_1 > 1$, set $c_k = \gamma_1 L_1\left(y^k\right)$ and compute

     $$x^{k+1} \in \operatorname{argmin}\left\{\widehat{\Psi}(x, y^k) : \ x \in \mathbb{R}^n\right\} = \operatorname{prox}_{c_k}^{f}\left(x^k - {c_k}^{-1} \nabla_x H\left(x^k, y^k\right)\right).$$

   - 2.2. Take $\gamma_2 > 1$, set $d_k = \gamma_2 L_2\left(x^{k+1}\right)$ and compute

     $$y^{k+1} \in \operatorname{argmin}\left\{\widetilde{\Psi}(x^{k+1}, y) : \ y \in \mathbb{R}^m\right\} = \operatorname{prox}_{d_k}^{g}\left(y^k - {d_k}^{-1} \nabla_y H\left(x^{k+1}, y^k\right)\right).$$

Main computational step: Computing prox of a nonconvex function.

KPALM Algorithm For The Squared Euclidean Norm

We devise a PALM-like algorithm, exploiting the specific structure of H, namely

- The function $w \mapsto H(w, x)$, for fixed x, is linear and therefore there is no need to linearize it as suggested in PALM.

$$w^i(t+1) = \arg\min_{w^i \in \Delta} \left\{ \langle w^i, d^i(x(t)) \rangle + \frac{\alpha_i(t)}{2} \|w^i - w^i(t)\|^2 \right\}.$$

- The function $x \mapsto H(w, x)$, for fixed w, is quadratic and convex. Hence, there is no need to add a proximal term as suggested in PALM.

$$x(t+1) = \operatorname{argmin} \left\{ H(w(t+1), x) \mid x \in \mathbb{R}^{nk} \right\}.$$

(1) Initialization: $z(0) = (w(0), x(0)) \in \Delta^m \times \mathbb{R}^{nk}$.

(2) General step ($t = 0, 1, \ldots$):

(2.1) Cluster assignment: choose certain $\alpha_i(t) > 0$, $i = 1, 2, \ldots, m$, and compute

$$w^i(t+1) = P_\Delta \left(w^i(t) - \frac{d^i(x(t))}{\alpha_i(t)} \right).$$

(2.2) Centers update: for each $l = 1, 2, \ldots, k$ compute

$$x^l(t+1) = \frac{\sum_{i=1}^m w_l^i(t+1) a^i}{\sum_{i=1}^m w_l^i(t+1)}.$$

KPALM Analysis

In order to prove the convergence of the sequence that is generated by KPALM, $\{z(t) := (w(t), x(t))\}_{n\in\mathbb{N}}$, to a critical point, we need to show the properties required by PALM theory.

- Boundedness: $w^i(t) \in \Delta$ and

$$x^l(t) = \frac{\sum_{i=1}^m w_l^i(t) a^i}{\sum_{i=1}^m w_l^i(t)} = \sum_{i=1}^m \left(\frac{w_l^i(t)}{\sum_{j=1}^m w_l^j(t)} \right) a^i \in Conv(\mathcal{A}).$$

- KL Property: H is a weighted sum of squared Euclidean norms hence semi-algebraic, and Δ is semi-algebraic set, thus $\delta_\Delta(\cdot)$ is semi-algebraic, and in turn Ψ since it is sum of these functions.

Assumption

(i) *The chosen sequences of parameters* $\{\alpha_i(t)\}_{t\in\mathbb{N}}$, $1 \leq i \leq m$, *are bounded:* $0 < \underline{\alpha_i} \leq \alpha_i(t) \leq \overline{\alpha_i} < \infty, \quad \forall\, t \in \mathbb{N}$.

(ii) *For all* $t \in \mathbb{N}$ *there exists* $\underline{\beta} > 0$ *such that* $2 \min_{1\leq l\leq k} \sum_{i=1}^m w_l^i(t) := \beta(w(t)) \geq \underline{\beta}$.

Denote $\underline{\alpha} = \min_{1\leq i\leq m} \underline{\alpha_i}$ and $\overline{\alpha} = \max_{1\leq i\leq m} \overline{\alpha_i}$.

Sufficient Decrease Proof

Since $x \mapsto H(w, x) = \sum_{l=1}^{k} \sum_{i=1}^{m} w_l^i \left\| x^l - a^i \right\|^2$ is C^2, and it Hessian is given by

$$\nabla_{x^j} \nabla_{x^l} H(w, x) = \begin{cases} 0 & \text{if } j \neq l, \quad 1 \leq j, l \leq k, \\ 2 \sum_{i=1}^{m} w_l^i & \text{if } j = l, \quad 1 \leq j, l \leq k, \end{cases}$$

then it is strongly convex with parameter $\beta(w)$, whenever $\beta(w) = 2 \min_{1 \leq l \leq k} \sum_{i=1}^{m} w_l^i > 0$.

Assumption 2(ii) ensures that $x \mapsto H(w(t), x)$ is strongly convex with parameter $\beta(w(t))$, hence

$$\begin{aligned} & H(w(t+1), x(t)) - H(w(t+1), x(t+1)) \geq \\ & \qquad \geq \langle \nabla_x H(w(t+1), x(t+1)), x(t) - x(t+1) \rangle + \frac{\beta(w(t))}{2} \|x(t) - x(t+1)\|^2 \\ & \qquad = \frac{\beta(w(t))}{2} \|x(t+1) - x(t)\|^2 \\ & \qquad \geq \frac{\beta}{2} \|x(t+1) - x(t)\|^2, \end{aligned}$$

where the equality follows from $\nabla_x H(w(t+1), x(t+1)) = 0$.

Sufficient Decrease Proof-Contd.

From the w update step we derive

$$\begin{aligned} H^i(w(t+1),x(t)) + \frac{\alpha_i(t)}{2}\|w^i(t+1)-w^i(t)\|^2 &= \\ &= \langle w^i(t+1), d^i(x(t))\rangle + \frac{\alpha_i(t)}{2}\|w^i(t+1)-w^i(t)\|^2 \\ &\leq \langle w^i(t), d^i(x(t))\rangle + \frac{\alpha_i(t)}{2}\|w^i(t)-w^i(t)\|^2 \\ &= \langle w^i(t), d^i(x(t))\rangle = H^i(w(t),x(t)). \end{aligned}$$

Summing the last inequality over all $1 \leq i \leq m$ yields

$$\frac{\underline{\alpha}}{2}\|w(t+1)-w(t)\|^2 \leq H(w(t),x(t)) - H(w(t+1),x(t))$$

Set $\rho_1 = \frac{1}{2}\min\left\{\underline{\alpha}, \underline{\beta}\right\}$, by combining the sufficient decrease in x and w variables we get

$$\begin{aligned} &\rho_1\|z(t+1)-z(t)\|^2 = \rho_1\left(\|w(t+1)-w(t)\|^2 + \|x(t+1)-x(t)\|^2\right) \leq \\ &\quad\leq [H(w(t),x(t)) - H(w(t+1),x(t))] + [H(w(t+1),x(t)) - H(w(t+1),x(t+1))] \\ &\quad= H(z(t)) - H(z(t+1)) = \Psi(z(t)) - \Psi(z(t+1)). \end{aligned}$$

Subgradient Lower Bound The Iterates Gap Proof

$$H(w,x) = \sum_{i=1}^{m} H^i(w,x) = \sum_{i=1}^{m} \langle w^i, d^i(x) \rangle, \quad G(w) = \sum_{i=1}^{m} G^i(w^i) = \sum_{i=1}^{m} \delta_\Delta(w^i)$$

$$w^i(t+1) = \arg\min_{w^i \in \Delta} \left\{ \langle w^i, d^i(x(t)) \rangle + \frac{\alpha_i(t)}{2} \|w^i - w^i(t)\|^2 \right\} \tag{1}$$

$$x(t+1) = \operatorname{argmin} \left\{ H(w(t+1), x) \mid x \in \mathbb{R}^{nk} \right\} \tag{2}$$

$$\partial \Psi = \nabla H + \partial G = \left(\left(\nabla_{w^i} H^i + \partial_{w^i} \delta_\Delta \right)_{i=1,2,\ldots,m}, \nabla_x H \right).$$

Evaluating the last relation at $z(t+1)$ and using (2)

$$\begin{aligned} \partial \Psi(z(t+1)) &= \left(\left(d^i(x(t+1)) + \partial_{w^i} \delta_\Delta(w^i(t+1)) \right)_{i=1,2,\ldots,m}, \nabla_x H(w(t+1), x(t+1)) \right) \\ &= \left(\left(d^i(x(t+1)) + \partial_{w^i} \delta_\Delta(w^i(t+1)) \right)_{i=1,2,\ldots,m}, \mathbf{0} \right). \end{aligned}$$

The optimality condition of $w^i(t+1)$ (see (1)), implies that there exists $u^i(t+1) \in \partial \delta_\Delta(w^i(t+1))$ such that

$$d^i(x(t)) + \alpha_i(t) \left(w^i(t+1) - w^i(t) \right) + u^i(t+1) = \mathbf{0}.$$

Subgradient Lower Bound The Iterates Gap Proof-Contd.

Setting $\gamma(t+1) := \left(\left(d^i(x(t+1)) + u^i(t+1)\right)_{i=1,2,\ldots,m}, \mathbf{0}\right) \in \partial\Psi(z(t+1))$.

$$\begin{aligned}
\|\gamma(t+1)\| &\leq \sum_{i=1}^{m} \left\| d^i(x(t+1)) - d^i(x(t)) - \alpha_i(t)\left(w^i(t+1) - w^i(t)\right) \right\| \\
&\leq \sum_{i=1}^{m} \left\| d^i(x(t+1)) - d^i(x(t)) \right\| + \sum_{i=1}^{m} \alpha_i(t) \left\| w^i(t+1) - w^i(t) \right\| \\
&\leq \sum_{i=1}^{m} 4M \left\| x(t+1) - x(t) \right\| + m\overline{\alpha} \left\| z(t+1) - z(t) \right\| \\
&\leq m\left(4M + \overline{\alpha}\right) \left\| z(t+1) - z(t) \right\|,
\end{aligned}$$

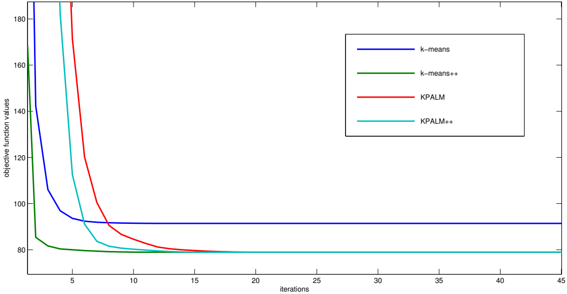
where the third inequality follows from the inequality

$$\|d^i(x(t+1) - d^i(x(t))\| \leq 4M\|x(t+1) - x(t)\|, \quad \forall\, i = 1, 2, \ldots, m,\ t \in \mathbb{N},$$

with $M = \max\limits_{1 \leq i \leq m} \|a^i\|$ and the result follows with $\rho_2 = m\left(4M + \overline{\alpha}\right)$.

Numerical Results: The Squared Norm

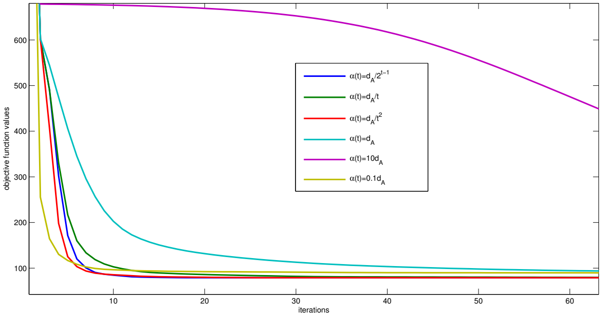
- Initialization issues: randomly picking *k* data points as starting centers vs. k-means++.
- Comparison of the objective function values performed on the Iris dataset for squared norm algorithms



- In the squared Euclidean setting, KPALM achieves lower objective function values than k-means. When using a more sophisticated initialization step, such as the one in k-means++, then k-means++ and KPALM++ achieve similar objective function values.
- k-means needs less number of iterations then KPALM to reach a certain precision.

Numerical Results: The Squared Norm

- Comparison of the objective function values performed on the Iris dataset for KPALM algorithm with different α parameter updates.



- It is preferable to use dynamic update of $\alpha(t)$ parameter to achieve a faster convergence, both in KPALM and ϵ-KPALM. Example for suitable choices can be $\alpha(t) = \mathrm{diam}(\mathcal{A})/t$ and $\alpha(t) = \mathrm{diam}(\mathcal{A})/2^t$.

Numerical Results: Squared Norm Algorithms vs. ε-KPALM

- Generated two synthetic datasets
- When the convex hulls of the desired clusters are mutually exclusive, algorithms which solve the clustering problem with the squared Euclidean distance are preferable to ε-KPALM.
- In datasets with outliers, the clustering obtained with ε-KPALM is more similar to the desired clustering, in terms of clustering metrics, than the clusterings obtained via the squared Euclidean algorithms. Therefore, as expected, for data with outliers, the choice of a norm instead of the squared norm is a more natural choice, and the ε-KPALM algorithm appears to be a promising algorithm to handle such data.

Summary of Numerical Results

- In the squared Euclidean setting, KPALM achieves lower objective function values than k-means. When using a more sophisticated initialization step, such as the one in k-means++, then k-means++ and KPALM++ achieve similar objective function values.
- k-means needs less number of iterations then KPALM to reach a certain precision.
- It is preferable to use dynamic update of $\alpha(t)$ parameter to achieve a faster convergence, both in KPALM and ϵ-KPALM. Example for suitable choices can be $\alpha(t) = \mathrm{diam}(\mathcal{A})/t$ and $\alpha(t) = \mathrm{diam}(\mathcal{A})/2^t$.
- When the convex hulls of the desired clusters are mutually exclusive, algorithms which solve the clustering problem with the squared Euclidean distance are preferable to ε-KPALM.
- In datasets with outliers, the clustering obtained with ε-KPALM is more similar to the desired clustering, in terms of clustering metrics, than the clusterings obtained via the squared Euclidean algorithms. Therefore, as expected, for data with outliers, the choice of a norm instead of the squared norm is a more natural choice, and the ε-KPALM algorithm appears to be a promising algorithm to handle such data.