A PROBABILISTIC ℓ_1 METHOD FOR CLUSTERING HIGH DIMENSIONAL DATA

TSVETAN ASAMOV AND ADI BEN–ISRAEL

Abstract. In general, the clustering problem is NP–hard, and global optimality cannot be established for non–trivial instances. For high–dimensional data, distance–based methods for clustering or classification face an additional difficulty, the unreliability of distances in very high–dimensional spaces. We propose a probabilistic, distance–based, iterative method for clustering data in very high–dimensional space, using the ℓ_1–metric that is less sensitive to high dimensionality than the Euclidean distance. For K clusters in $\mathbb{R}^n$, the problem decomposes to K problems coupled by probabilities, and an iteration reduces to finding Kn weighted medians of points on a line. The complexity of the algorithm is linear in the dimension of the data space, and its performance was observed to improve significantly as the dimension increases.

1. Introduction

The emergence and growing applications of big data have underscored the need for efficient algorithms based on optimality principles, and scalable methods that can provide valuable insights at a reasonable computational cost.

In particular, problems with high–dimensional data have arisen in several scientific and technical areas (such as genetics [19], medical imaging [29] and spatial databases [21], etc.) These problems pose a special challenge because of the unreliability of distances in very high dimensions. In such problems it is often advantageous to use the ℓ_1–metric which is less sensitive to the "curse of dimensionality" than the Euclidean distance.

We propose a new probabilistic distance–based method for clustering data in very high–dimensional spaces. The method uses the ℓ_1–distance, and computes the cluster centers using weighted medians of the given data points. Our algorithm resembles well–known techniques such as fuzzy clustering [9] and K–means, and inverse distance interpolation [26].

The cluster membership probabilities are derived from necessary optimality conditions for an approximate problem, and decompose a clustering problem with K clusters in $\mathbb{R}^n$ into Kn one–dimensional problems, which can be solved separately. The algorithm features a straightforward implementation and a polynomial running time, in particular, its complexity is linear in the dimension n. In numerical experiments it outperformed several commonly used methods, with better results for higher dimensions.

While the cluster membership probabilities simplify our notation, and link our results to the theory of subjective probability, these probabilities are not needed by themselves, since they are given in terms of distances, that have to be computed at each iteration.

1.1. Notation.

We use the abbreviation $\overline{1,K} := \{1, 2, \ldots, K\}$ for the indicated index set. The j_{th} component of a vector $\mathbf{x}_i \in \mathbb{R}^n$ is denoted $\mathbf{x}_i[j]$. The ℓ_p–**norm** of a vector $\mathbf{x} = (\mathbf{x}[j]) \in \mathbb{R}^n$ is

$$\|\mathbf{x}\|_p := (\sum_{j=1}^{n} \left|\mathbf{x}[j]\right|^p)^{1/p}$$

Date: April 13, 2016.
2010 *Mathematics Subject Classification.* Primary 62H30, 90B85; Secondary 90C59.
Key words and phrases. Clustering, ℓ_1–norm, high–dimensional data, continuous location.

and the associated ℓ_p–**distance** between two vectors $\mathbf{x}$ and $\mathbf{y}$ is $d_p(\mathbf{x},\mathbf{y}) := \|\mathbf{x}-\mathbf{y}\|_p$, in particular, the Euclidean distance with $p=2$, and the ℓ_1–distance,

$$d_1(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_1 = \sum_{j=1}^{n} \big|\mathbf{x}[j]-\mathbf{y}[j]\big|. \tag{1}$$

1.2. **The clustering problem.** Given

- a set $\mathbf{X} = \{\mathbf{x}_i : i \in \overline{1,N}\} \subset \mathbb{R}^n$ of N points in $\mathbb{R}^n$,
- their **weights** $W = \{w_i > 0 : i \in \overline{1,N}\}$, and
- an integer $1 \le K \le N$,

partition $\mathbf{X}$ into K **clusters** $\{\mathbf{X}_k : k \in \overline{1,K}\}$, defined as disjoint sets where the points in each cluster are **similar** (in some sense), and points in different clusters are dissimilar. If by **similar** is meant **close** in some metric $d(\mathbf{x},\mathbf{y})$, we have a **metric** (or **distance based**) **clustering problem**, in particular ℓ_1–**clustering** if the ℓ_1–distance is used, **Euclidean clustering** for the ℓ_2–distance, etc.

1.3. **Centers.** In metric clustering each cluster has a representative point, or **center**, and distances to clusters are defined as the distances to their centers. The center $\mathbf{c}_k$ of cluster $\mathbf{X}_k$ is a point $\mathbf{c}$ that minimizes the sum of weighted distances to all points of the cluster,

$$\mathbf{c}_k := \arg\min \Big\{ \sum_{\mathbf{x}_i \in \mathbf{X}_k} w_i\, d(\mathbf{x}_i,\mathbf{c}) \Big\}. \tag{2}$$

The metric clustering problem is formulated as follows: Given $\mathbf{X}, W$ and K as above, find centers $\{\mathbf{c}_k : k \in \overline{1,K}\} \subset \mathbb{R}^n$ so as to minimize

$$\min_{\mathbf{c}_1,\cdots,\mathbf{c}_K} \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in \mathbf{X}_k} w_i\, d(\mathbf{x}_i,\mathbf{c}_k), \tag{$\mathbf{L}.K$}$$

where $\mathbf{X}_k$ is the cluster of points in $\mathbf{X}$ assigned to the center $\mathbf{c}_k$.

1.4. **Location problems.** Metric clustering problems often arise in location analysis, where $\mathbf{X}$ is the set of the locations of customers, W is the set of their weights (or demands), and it is required to locate K facilities $\{\mathbf{c}_k\}$ to serve the customers optimally in the sense of total weighted-distances traveled. The problem ($\mathbf{L}.K$) is then called a **multi–facility location problem**, or a **location–allocation problem** because it is required to locate the centers, and to assign or allocate the points to them.

Problem ($\mathbf{L}.K$) is trivial for $K = N$ (every point is its own center) and reduces for $K = 1$ to the **single facility location problem**: find the location of a **center** $\mathbf{c} \in \mathbb{R}^n$ so as to minimize the sum of weighted distances,

$$\min_{\mathbf{c}\,\in\,\mathbb{R}^n} \sum_{i=1}^{N} w_i\, d(\mathbf{x}_i,\mathbf{c}). \tag{$\mathbf{L}.1$}$$

For $1 < K < N$, the problem ($\mathbf{L}.K$) is NP-hard in general [24], while the planar case can be solved polynomially in N, [13].

1.5. **Probabilistic approximation.** ($\mathbf{L}.K$) can be approximated by a continuous problem

$$\min_{\mathbf{c}_1,\cdots,\mathbf{c}_K} \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in \mathbf{X}} w_i\, p_k(\mathbf{x}_i) d(\mathbf{x}_i,\mathbf{c}_k), \tag{$\mathbf{P}.K$}$$

where rigid assignments $\mathbf{x}_i \in \mathbf{X}_k$ are replaced by probabilistic (soft) assignments, expressed by probabilities $p_k(\mathbf{x}_i)$ that a point $\mathbf{x}_i$ belongs to the cluster $\mathbf{X}_k$.

The **cluster membership probabilities** $p_k(\mathbf{x}_i)$ of any point $\mathbf{x}_i$ and are assumed to depend on the distances $d(\mathbf{x}_i, \mathbf{c}_k)$ as follows

$$\boxed{\text{membership in a cluster is more likely the closer is its center}} \qquad (\mathbf{A})$$

Given these probabilities, the problem (P.K) can be decomposed into K single facility location problems,

$$\min_{\mathbf{c}_k} \sum_{\mathbf{x}_i \in \mathbf{X}} p_k(\mathbf{x}_i) w_i \, d(\mathbf{x}_i, \mathbf{c}_k), \;\; k \in \overline{1,K}. \qquad (\text{P}.k)$$

The solutions $\mathbf{c}_k$ of the K problems (P.k), are then used to calculate the new distances $d(\mathbf{x}_i, \mathbf{c}_k)$ for all $i \in \overline{1,N}$, $k \in \overline{1,K}$, and from them, new probabilities $\{p_k(\mathbf{x}_i)\}$, etc.

1.6. **The case for the ℓ_1 norm.** In high dimensions, distances between points become unreliable [7], and this in particular "makes a proximity query meaningless and unstable because there is poor discrimination between the nearest and furthest neighbor" [1]. For very large n, the Euclidean distance between random points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$d_2(\mathbf{x}, \mathbf{y}) = (\sum_{j=1}^{n} \big|\mathbf{x}[j] - \mathbf{y}[j]\big|^2)^{1/2} = (\|\mathbf{x}\|_2^2 - 2 \sum_{j=1}^{n} \mathbf{x}[j]\, \mathbf{y}[j] + \|\mathbf{y}\|_2^2)^{1/2} \qquad (3)$$

is approximately $d_2(\mathbf{x}, \mathbf{y}) \approx (\|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2)^{1/2}$ because the cross products $\mathbf{x}[j]\, \mathbf{y}[j]$ in (3) tend to cancel. In particular, if $\mathbf{x}, \mathbf{y}$ are random points on the unit sphere in $\mathbb{R}^n$ then $d_2(\mathbf{x}, \mathbf{y}) \approx \sqrt{2}$ for very large n. This "curse of high dimensionality" limits the applicability of distance based methods in high dimension.

The ℓ_1–distance is less sensitive to high dimensionality, and has been shown to "provide the best discrimination in high–dimensional data spaces", [1]. We use it throughout this paper.

The plan of the paper. The ℓ_1–metric clustering problem is solved in § 2 for one center. A probabilistic approximation of (**L**.K) is discussed in § 3, the probabilities studied in §§ 4–5. The centers of the approximate problem are computed in § 6. Our main result, Algorithm PCM(ℓ_1) of § 8, uses the power probabilities of § 7, and has running time that is linear in the dimension of the space, see Corollary 1. Theorem 1, a monotonicity property of Algorithm PCM(ℓ_1), is proved in § 9. Section 10 lists conclusions. Appendix A shows relations to previous work, and Appendix B reports some numerical results.

2. The single facility location problem with the ℓ_1–norm

For the ℓ_1–distance (1) the problem (**L**.1) becomes

$$\min_{\mathbf{c} \in \mathbb{R}^n} \sum_{i=1}^{N} w_i \, d_1(\mathbf{x}_i, \mathbf{c}), \quad \text{or} \quad \min_{\mathbf{c} \in \mathbb{R}^n} \sum_{i=1}^{N} w_i \sum_{j=1}^{n} \big|\mathbf{x}_i[j] - \mathbf{c}[j]\big|, \qquad (4)$$

in the variable $\mathbf{c} \in \mathbb{R}^n$, which can be solved separately for each component $\mathbf{c}[j]$, giving the n problems

$$\min_{\mathbf{c}[j] \in \mathbb{R}} \sum_{i=1}^{N} w_i \, \big|\mathbf{x}_i[j] - \mathbf{c}[j]\big|, \;\; j \in \overline{1,n}. \qquad (5)$$

Definition 1. Let $\mathbf{X} = \{x_1, \cdots, x_N\} \in \mathbb{R}$ be an ordered set of points

$$x_1 \le x_2 \le \cdots \le x_N$$

and let $\mathbf{W} = \{w_1, \cdots, w_N\}$ be a corresponding set of positive weights. A point x is a **weighted median** (or **W–median**) of $\mathbf{X}$ if there exist $\alpha, \beta \ge 0$ such that

$$\sum \{w_i : \, x_i < x\} + \alpha = \sum \{w_i : \, x_i > x\} + \beta \qquad (6)$$

where $\alpha + \beta$ is the weight of x if $x \in \mathbf{X}$, and $\alpha = \beta = 0$ if $x \notin \mathbf{X}$.

The weighted median always exists, but is not necessarily unique.

Lemma 1. For $\mathbf{X}$, $\mathbf{W}$ as above, define

$$\theta_k := \frac{\sum_{i=1}^{k} w_i}{\sum_{i=1}^{N} w_i}, \quad k \in \overline{1,N}, \tag{7}$$

and let k^* be the smallest k with $\theta_k \geq \frac{1}{2}$. If

$$\theta_{k^*} > \tfrac{1}{2} \tag{8}$$

then x_{k^*} is the unique weighted median, with

$$\alpha = \tfrac{1}{2}\left(w_{k^*} + \sum_{k>k^*} w_k - \sum_{k<k^*} w_k\right),\ \beta = w_{k^*} - \alpha. \tag{9}$$

Otherwise, if

$$\theta_{k^*} = \tfrac{1}{2}, \tag{10}$$

then any point in the open interval (x_{k^*}, x_{k^*+1}) is a weighted median with $\alpha = \beta = 0$.

Proof. The statement holds since the sequence (7) is increasing from $\theta_1 = (w_1/\sum_{k=1}^{N} w_k)$ to $\theta_N = 1$. □

Note: In case (10),

$$\sum\{w_k : x_k \leq x_{k^*}\} = \sum\{w_k : x_k \geq x_{k^*+1}\},$$

we can take the median as the midpoint of x_{k^*} and x_{k^*+1}, in order to conform with the classical definition of the median (for an even number of points of equal weight) .

Lemma 2. Given $\mathbf{X}$ and $\mathbf{W}$ as in Definition 1, the set of minimizers c of

$$\sum_{i=1}^{N} w_i \left|x_i - c\right|$$

is the set of $\mathbf{W}$–medians of $\mathbf{X}$.

Proof. The result is well known if all weights are 1. If the weights are integers, consider a point x_i with weight w_i as w_i coinciding points of weight 1 and the result follows. Same if the weights are rational. Finally, if the weights are real, consider their rational approximations. □

3. Probabilistic approximation of (L.K)

We relax the assignment problem in (**L**.K) of § 1.2 by using a **continuous approximation** as follows,

$$\min \sum_{k=1}^{K} \sum_{i=1}^{N} w_i\, p_k(\mathbf{x}_i)\, d(\mathbf{x}_i, \mathbf{c}_k) \tag{**P**.K}$$

with two sets of variables,

the **centers** $\{\mathbf{c}_k\}$, and

the **cluster membership probabilities** $\{p_k(\mathbf{x}_i)\}$,

$$p_k(\mathbf{x}_i) := \text{Prob}\,\{\mathbf{x}_i \in \mathbf{X}_k\},\ i \in \overline{1,N},\ k \in \overline{1,K}, \tag{11}$$

Because the probabilities $\{p_k(\mathbf{x}_i)\}$ add to 1 for each $i \in \overline{1,N}$, the objective function of $(\mathbf{P}.K)$ is an upper bound on the optimal value of $(\mathbf{L}.K)$,

$$\sum_{k=1}^{K}\sum_{i=1}^{N} w_i\, p_k(\mathbf{x}_i)\, d(\mathbf{x}_i, \mathbf{c}_k) \geq \min\,(\mathbf{L}.K), \tag{12}$$

and therefore so is the optimal value of $(\mathbf{P}.K)$,

$$\min\,(\mathbf{P}.K) \geq \min\,(\mathbf{L}.K). \tag{13}$$

4. Axioms for probabilistic distance clustering

In this section, $d_k(\mathbf{x})$ stands for $d_k(\mathbf{x}, \mathbf{c}_k)$, the distance of $\mathbf{x}$ to the center $\mathbf{c}_k$ of the k_{th}–cluster, $k \in \overline{1,K}$. To simplify notation, the point $\mathbf{x}$ is assumed to have weight $w = 1$.

The **cluster membership probabilities** $\{p_k(\mathbf{x}) : k \in \overline{1,K}\}$ of a point $\mathbf{x}$ depend only on the **distances** $\{d_k(\mathbf{x}) :\ k \in \overline{1,K}\}$,

$$\mathbf{p}(\mathbf{x}) = \mathbf{f}(\mathbf{d}(\mathbf{x})) \tag{14}$$

where $\mathbf{p}(\mathbf{x}) \in \mathbb{R}^K$ is the vector of probabilities $(p_k(\mathbf{x}))$, and $\mathbf{d}(\mathbf{x})$ is the vector of distances $(d_k(\mathbf{x}))$. Natural assumptions for the relation (14) include

$$d_i(\mathbf{x}) < d_j(\mathbf{x}) \implies p_i(\mathbf{x}) > p_j(\mathbf{x}),\ \text{for all } i, j \in \overline{1,K} \tag{15a}$$

$$\mathbf{f}(\lambda\, \mathbf{d}(\mathbf{x})) = \mathbf{f}(\mathbf{d}(\mathbf{x})),\ \text{for any } \lambda > 0 \tag{15b}$$

$$Q\, \mathbf{p}(\mathbf{x}) = \mathbf{f}(Q\, \mathbf{d}(\mathbf{x})),\ \text{for any permutation matrices } Q \tag{15c}$$

Condition (15a) states that membership in a cluster is more probable the closer it is, which is Assumption (A) of § 1.5. The meaning of (15b) is that the probabilities $p_k(\mathbf{x})$ do not depend on the scale of measurement, i.e., $\mathbf{f}$ is homogeneous of degree 0. It follows that the probabilities $p_k(\mathbf{x})$ depend only on the ratios of the distances $\{d_k(\mathbf{x}) :\ k \in \overline{1,K}\}$.

The symmetry of $\mathbf{f}$, expressed by (15c), guarantees for each $k \in \overline{1,K}$, that the probability $p_k(\mathbf{x})$ does not depend on the numbering of the other clusters.

Assuming continuity of $\mathbf{f}$ it follows from (15a) that

$$d_i(\mathbf{x}) = d_j(\mathbf{x}) \implies p_i(\mathbf{x}) = p_j(\mathbf{x}),$$

for any $i, j \in \overline{1,K}$.

For any nonempty subset $\mathcal{S} \subset \overline{1,K}$, let

$$p_{\mathcal{S}}(\mathbf{x}) = \sum_{s \in \mathcal{S}} p_s(\mathbf{x}),$$

the probability that $\mathbf{x}$ belongs to one of the clusters $\{\mathcal{C}_s :\ s \in \mathcal{S}\}$, and let $p_k(\mathbf{x}|\mathcal{S})$ denote the **conditional probability** that $\mathbf{x}$ belongs to the cluster $\mathcal{C}_k$, given that it belongs to one of the clusters $\{\mathcal{C}_s :\ s \in \mathcal{S}\}$.

Since the probabilities $p_k(\mathbf{x})$ depend only on the ratios of the distances $\{d_k(\mathbf{x}) :\ k \in \overline{1,K}\}$, and these ratios are unchanged in subsets $\mathcal{S}$ of the index set $\overline{1,K}$, it follows that for all $k \in \overline{1,K},\ \emptyset \neq \mathcal{S} \subset \overline{1,K}$,

$$p_k(\mathbf{x}) = p_k(\mathbf{x}|\mathcal{S})\, p_{\mathcal{S}}(\mathbf{x}) \tag{16}$$

which is the **choice axiom** of Luce, [22, Axiom 1], and therefore, [30],

$$p_k(\mathbf{x}|\mathcal{S}) = \frac{v_k(\mathbf{x})}{\sum\limits_{s \in \mathcal{S}} v_s(\mathbf{x})} \tag{17}$$

where $v_k(\mathbf{x})$ is a scale function, in particular,

$$p_k(\mathbf{x}) = \frac{v_k(\mathbf{x})}{\sum\limits_{s\in\overline{1,K}} v_s(\mathbf{x})} \,. \tag{18}$$

Assuming $v_k(\mathbf{x}) \neq 0$ for all k, it follows that

$$p_k(\mathbf{x}) v_k(\mathbf{x})^{-1} = \frac{1}{\sum\limits_{s\in\overline{1,K}} v_s(\mathbf{x})}, \tag{19}$$

where the right hand side is a function of $\mathbf{x}$, and does not depend on k.

Property (15a) implies that the function $v_k(\cdot)$ is a monotone decreasing function of $d_k(\mathbf{x})$.

5. Cluster membership probabilities as functions of distance

Given K centers $\{\mathbf{c}_k\}$, and a point $\mathbf{x}$ with weight w and distances $\{d(\mathbf{x},\mathbf{c}_k) : k \in \overline{1,K}\}$ from these centers, a simple choice for the function $v_k(\mathbf{x})$ in (17) is

$$v_k(\mathbf{x}) = \frac{1}{w\, d_k(\mathbf{x})} \,, \tag{20}$$

for which (19) gives[1],

$$w\, p_k(\mathbf{x})\, d(\mathbf{x},\mathbf{c}_k) = D(\mathbf{x}),\ k \in \overline{1,K}, \tag{21}$$

where the function $D(\mathbf{x})$, called the **joint distance function** (JDF) at $\mathbf{x}$, does not depend on k.

For a given point $\mathbf{x}$ and given centers $\{\mathbf{c}_k\}$, equations (21) are optimality conditions for the extremum problem

$$\min \left\{ w \sum_{k=1}^{K} p_k^2\, d(\mathbf{x},\mathbf{c}_k) : \sum_{k=1}^{K} p_k = 1,\ p_k \geq 0,\ k \in \overline{1,K} \right\} \tag{22}$$

in the probabilities $\{p_k := p_k(\mathbf{x})\}$. The squares of probabilities in the objective of (22) serve to smooth the underlying non–smooth problem, see the seminal paper by Teboulle [27]. Indeed, (21) follows by differentiating the Lagrangian

$$L(\mathbf{p},\lambda) = w \sum_{k=1}^{K} p_k^2\, d(\mathbf{x},\mathbf{c}_k) + \lambda \left(\sum_{k=1}^{K} p_k - 1 \right), \tag{23}$$

with respect to p_k and equating the derivative to zero.

Since probabilities add to one we get from (21),

$$p_k(\mathbf{x}) = \frac{\prod\limits_{j\neq k} d(\mathbf{x},\mathbf{c}_j)}{\sum\limits_{\ell=1}^{K} \prod\limits_{m\neq\ell} d(\mathbf{x},\mathbf{c}_m)},\ k \in \overline{1,K}, \tag{24}$$

and the JDF at $\mathbf{x}$,

$$D(\mathbf{x}) = w\, \frac{\prod\limits_{j=1}^{K} d(\mathbf{x},\mathbf{c}_j)}{\sum\limits_{\ell=1}^{K} \prod\limits_{m\neq\ell} d(\mathbf{x},\mathbf{c}_m)} \,, \tag{25}$$

which is (up to a constant) the **harmonic mean** of the distances $\{d(\mathbf{x},\mathbf{c}_k) : k \in \overline{1,K}\}$, see also (A-4) below.

[1]There are other ways to model Assumption (**A**), e.g. [5], but the simple model (21) works well enough for our purposes.

Note that the probabilities $\{p_k(\mathbf{x}) : k \in \overline{1,K}\}$ are determined by the centers $\{\mathbf{c}_k : k \in \overline{1,K}\}$ alone, while the function $D(\mathbf{x})$ depends also on the weight w. For example, in case $K = 2$,

$$p_1(\mathbf{x}) = \frac{d(\mathbf{x},\mathbf{c}_2)}{d(\mathbf{x},\mathbf{c}_1)+d(\mathbf{x},\mathbf{c}_2)},\ p_2(\mathbf{x}) = \frac{d(\mathbf{x},\mathbf{c}_1)}{d(\mathbf{x},\mathbf{c}_1)+d(\mathbf{x},\mathbf{c}_2)}, \tag{26a}$$

$$D(\mathbf{x}) = w\,\frac{d(\mathbf{x},\mathbf{c}_1)\,d(\mathbf{x},\mathbf{c}_2)}{d(\mathbf{x},\mathbf{c}_1)+d(\mathbf{x},\mathbf{c}_2)}. \tag{26b}$$

6. Computation of centers

We use the ℓ_1–distance (1) throughout. The objective function of $(\mathbf{P}.K)$ is a separable function of the cluster centers,

$$f(\mathbf{c}_1,\ldots,\mathbf{c}_K) := \sum_{k=1}^{K} f_k(\mathbf{c}_k), \tag{27a}$$

$$\text{where}\quad f_k(\mathbf{c}) := \sum_{i=1}^{N} w_i\,p_k(\mathbf{x}_i)\,d_1(\mathbf{x}_i,\mathbf{c}),\quad k \in \overline{1,K}. \tag{27b}$$

The centers problem thus separates into K problems of type (4),

$$\min_{\mathbf{c}_k\,\in\,\mathbb{R}^n} \sum_{i=1}^{N} w_i\,p_k(\mathbf{x}_i) \sum_{j=1}^{n} \big|\mathbf{x}_i[j]-\mathbf{c}_k[j]\big|,\ k \in \overline{1,K}, \tag{28}$$

coupled by the probabilities $\{p_k(\mathbf{x}_i)\}$. Each of these problems separates into n problems of type (5) for the components $\mathbf{c}_k[j]$,

$$\min_{c_k[j]\,\in\,\mathbb{R}} \sum_{i=1}^{N} w_i\,p_k(\mathbf{x}_i)\,\big|\mathbf{x}_i[j]-\mathbf{c}_k[j]\big|,\ k \in \overline{1,K},\ j \in \overline{1,n}, \tag{29}$$

whose solution, by Lemma 2, is a weighted median of the points $\{\mathbf{x}_i[j]\}$ with weights $\{w_i\,p_k(\mathbf{x}_i)\}$.

7. Power probabilities

The cluster membership probabilities $\{p_k(\mathbf{x}) : k \in \overline{1,K}\}$ of a point $\mathbf{x}$ serve to relax the rigid assignment of $\mathbf{x}$ to any of the clusters, but eventually it may be necessary to produce such an assignment. One way to achieve this is to raise the membership probabilities $p_k(\mathbf{x})$ of (24) to a power $\nu \geq 1$, and normalize, obtaining the **power probabilities**

$$p_k^{(\nu)}(\mathbf{x}) := \frac{p_k^{\nu}(\mathbf{x})}{\sum\limits_{j=1}^{K} p_j^{\nu}(\mathbf{x})}, \tag{30}$$

which, by (24), can also be expressed in terms of the distances $d(\mathbf{x},\mathbf{c}_k)$,

$$p_k^{(\nu)}(\mathbf{x}) := \frac{\prod\limits_{j\neq k} d(\mathbf{x},\mathbf{c}_j)^{\nu}}{\sum\limits_{\ell=1}^{K}\prod\limits_{m\neq\ell} d(\mathbf{x},\mathbf{c}_m)^{\nu}},\ k \in \overline{1,K}. \tag{31}$$

As the exponent ν increases the power probabilities $p_k^{(\nu)}(\mathbf{x})$ tend to hard assignments: If M is the index set of maximal probabilities, and M has $\#M$ elements, then,

$$\lim_{\nu\to\infty} p_k^{(\nu)}(\mathbf{x}) = \begin{cases} \frac{1}{\#M}, & \text{if } k \in M; \\ 0, & \text{otherwise,} \end{cases} \tag{32}$$

and the limit is a hard assignment if $\#M = 1$, i.e. if the maximal probability is unique.

Numerical experience suggests an increase of ν at each iteration, see, e.g., (33) below.

8. Algorithm PCM(ℓ_1): Probabilistic Clustering Method with ℓ_1 distances

The problem $(\mathbf{P}.K)$ is solved iteratively, using the following updates in succession.

Probabilities computation: Given K centers $\{\mathbf{c}_k\}$, the assignments probabilities $\{p_k^{(\nu)}(\mathbf{x}_i)\}$ are calculated using (31). The exponent ν is updated at each iteration, say by a constant increment $\Delta \geq 0$,

$$\nu := \nu + \Delta \tag{33}$$

starting with an initial ν_0.

Centers computation: Given the assignment probabilities $\{p_k^{(\nu)}(\mathbf{x}_i)\}$, the problem $(\mathbf{P}.K)$ separates into Kn problems of type (29),

$$\min_{\mathbf{c}_k[j]\in\mathbb{R}} \sum_{i=1}^{N} w_i\, p_k^{(\nu)}(\mathbf{x}_i)\, \big|\mathbf{x}_i[j] - \mathbf{c}_k[j]\big|,\ k \in \overline{1,K},\ j \in \overline{1,n}, \tag{34}$$

one for each component $\mathbf{c}_k[j]$ of each center $\mathbf{c}_k$, that are solved by Lemma 2.

These results are presented in an algorithm form as follows.

Algorithm PCM(ℓ_1) : An algorithm for the ℓ_1 clustering problem

---

Data: $\mathbf{X} = \{\mathbf{x}_i : i \in \overline{1,N}\}$ data points, $\{w_i : i \in \overline{1,N}\}$ their weights,
K the number of clusters,
$\epsilon > 0$ (stopping criterion),
$\nu_0 \geq 1$ (initial value of the exponent ν), $\Delta > 0$ (the increment in (33).)

Initialization: K arbitrary centers $\{\mathbf{c}_k : k \in \overline{1,K}\}$, $\nu := \nu_0$.

Iteration:

Step 1 **compute** distances $\{d_1(\mathbf{x}, \mathbf{c}_k) : k \in \overline{1,K}\}$ for all $\mathbf{x} \in \mathbf{X}$

Step 2 **compute** the assignments $\{p_k^{(\nu)}(\mathbf{x}) : \mathbf{x} \in \mathbf{X},\ k \in \overline{1,K}\}$ (using (31))

Step 3 **compute** the new centers $\{\mathbf{c}_{k+} : k \in \overline{1,K}\}$ (applying Lemma 2 to (34))

Step 4 **if** $\sum_{k=1}^{K} d_1(\mathbf{c}_{k+}, \mathbf{c}_k) < \epsilon$ **stop**
else $\nu := \nu + \Delta$, **return** to step 1

Corollary 1. The running time of Algorithm PCM(ℓ_1) is

$$O(NK(K^2 + n)I), \tag{35}$$

where n is the dimension of the space, N the number of points, K the number of clusters, and I is the number of iterations.

Proof. The number of operations in an iteration is calculated as follows:

Step 1: $O(nNK)$, since computing the ℓ_1 distance between two n-dimensional vectors takes $O(n)$ time, and there are $N\,K$ distances between all points and all centers.

Step 2: $O(NK^3)$, there are NK assignments, each taking $O(K^2)$.

Step 3: $O(nNK)$, computing the weighted median of N points in $\mathbb{R}$ takes $O(N)$ time, and $K\,n$ such medians are computed.

Step 4: $O(nK)$, since there are K cluster centers of dimension n.

The corollary is proved by combining the above results. □

Remark 1.
(a) The result (35) shows that Algorithm PCM(ℓ_1) is linear in n, which in high–dimensional data is much greater than N and K.
(b) The first few iterations of the algorithm come close to the final centers, and thereafter the iterations are slow, making the stopping rule in Step 4 ineffective. A better stopping rule is a bound on the number of iterations I, which can then be taken as a constant in (35).
(c) Algorithm PCM(ℓ_1) can be modified to account for very unequal cluster sizes, as in [14]. This modification did not significantly improve the performance of the algorithm in our experiments.
(d) The centers here are computed from scratch at each iteration using the current probabilities, unlike the Weiszfeld method [28] or its generalizations, [17]–[18], where the centers are updated at each iteration.

9. Monotonicity

The centers computed iteratively by Algorithm PCM(ℓ_1) are confined to the convex hull of $\mathbf{X}$, a compact set, and therefore a subsequence converges to an optimal solution of the approximate problem ($\mathbf{P}.K$), that in general is not an optimal solution of the original problem ($\mathbf{L}.K$).

The JDF of the data set $\mathbf{X}$ is defined as the sum of the JDF's of its points,

$$D(\mathbf{X}) := \sum_{\mathbf{x}\in\mathbf{X}} D(\mathbf{x}). \tag{36}$$

We prove next a monotonicity result for $D(\mathbf{X})$.

Theorem 1. The function $D(\mathbf{X})$ decrease along any sequence of iterates of centers.

Proof. The function $D(\mathbf{X})$ can be written as

$$\begin{aligned} D(\mathbf{X}) &:= \sum_{\mathbf{x}\in\mathbf{X}} \left(\sum_{k=1}^{K} p_k(\mathbf{x}) \right) D(\mathbf{x}), \text{ since the probabilities add to 1,} \\ &= \sum_{\mathbf{x}\in\mathbf{X}} \sum_{k=1}^{K} w(\mathbf{x})\, p_k(\mathbf{x})^2\, d_1(\mathbf{x},\mathbf{c}_k), \text{ by (21).} \end{aligned} \tag{37}$$

The proof is completed by noting that, for each $\mathbf{x}$, the probabilities $\{p_k(\mathbf{x}) : k \in \overline{1,K}\}$ are chosen as to minimize the function

$$\sum_{k=1}^{K} w(\mathbf{x})\, p_k(\mathbf{x})^2\, d_1(\mathbf{x},\mathbf{c}_k) \tag{38}$$

for the given centers, see (22), and the centers $\{\mathbf{c}_k : k \in \overline{1,K}\}$ minimize the function (38) for the given probabilities. □

Remark 2. The function $D(\mathbf{X})$ also decreases if the exponent ν is increased in (30), for then shorter distances are becoming more probable in (37).

10. Conclusions

In summary, our approach has the following advantages.

(1) In numerical experiments, see Appendix B, Algorithm PCM(ℓ_1) outperformed the fuzzy clustering ℓ_1–method, the K–means ℓ_1 method, and the generalized Weiszfeld method [17].

(2) The solutions of (22) are less sensitive to outliers than the solutions of (A-5), which uses squares of distances.
(3) The probabilistic principle (A-8) allows using other monotonic functions, in particular the exponential function $\phi(d) = e^d$, that gives sharper results, and requires only that every distance $d(\mathbf{x}, \mathbf{c})$ be replaced by $\exp\{d(\mathbf{x}, \mathbf{c})\}$, [5].
(4) The JDF (36) of the data set, provides a guide to the "right" number of clusters for the given data, [6].

References

[1] C.C. Aggarwal, A. Hinneburg and D.A. Keim, On the surprising behavior of distance metrics in high dimensional spaces, *Lecture Notes in Mathematics* **1748**(2000), 420–434, Springer–Verlag.

[2] A. Andoni and P. Indyk, Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions, *Proceedings of the 47th Annual IEEE Symposium on the Foundations of Computer Science*, 2006.

[3] M. Arav, Contour approximation of data and the harmonic mean, *J. Math. Inequalities* **2**(2008), 161–167.

[4] A. Beck and S. Sabach, Weiszfeld's Method: Old and New Results, *J. Optimiz. Th. Appl.* **164**(2015), 1–40.

[5] A. Ben–Israel and C. Iyigun, Probabilistic distance clustering, *J. Classification* **25**(2008), 5–26.

[6] A. Ben–Israel and C. Iyigun, Clustering, Classification and Contour Approximation of Data, pp. 75–100 in *Biomedical Mathematics: Promising Directions in Imaging, Therapy Planning and Inverse Problems*, Y. Censor, Ming Jiang and Ge Wang (Editors), Medical Physics Publishing, Madison, Wisconsin, 2010, ISBN 978-1-930524-48-4.

[7] K. Beyer, J. Goldstein, R. Ramakrishnan and U. Shaft, When is nearest neighbors meaningful?, *Int. Conf. Database Theory (ICDT) Conference Proceedings*, 1999, 217– 235.

[8] J.C. Bezdek, *Fuzzy mathematics in pattern classification*, Doctoral Dissertation, Cornell University, Ithaca, 1973.

[9] J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum, New York, 1981, ISBN 0-306-40671-3.

[10] J.C. Bezdek and S.K. Pal, (editors), *Fuzzy Models for Pattern Recognition: Methods that Search for Structure in Data*, IEEE Press, New York, 1992

[11] B. Chazelle, Finding a good neighbor, near and fast, *Comm. ACM* **51**(2008), 115.

[12] K. R. Dixon and J. A. Chapman, Harmonic mean measure of animal activity areas, *Ecology* **61**(1980), 1040–1044

[13] Z. Drezner, The planar two–center and two–median problems, *Transportation Science* **18**(1984), 351–361.

[14] C. Iyigun and A. Ben–Israel, Probabilistic distance clustering adjusted for cluster size, *Probability in Engineering and Informational Sciences* **22**(2008), 1–19.

[15] C. Iyigun and A. Ben–Israel, Contour approximation of data: A duality theory, *Lin. Algeb. and Appl.* **430**(2009), 2771–2780.

[16] C. Iyigun and A. Ben–Israel, Semi–supervised probabilistic distance clustering and the uncertainty of classification, pp. 3–20 in *Advances in Data Analysis, Data Handling and Business Intelligence*, A. Fink, B. Lausen, W. Seidel and A. Ultsch (Editors), Studies in Classification, Data Analysis and Knowledge Organization, Springer 2010, ISBN 978-3-642-01043-9.

[17] C. Iyigun and A. Ben–Israel, A generalized Weiszfeld method for the multi–facility location problem, *O.R. Letters* **38**(2010), 207–214.

[18] C. Iyigun and A. Ben–Israel, Contributions to the multi–facility location problem, (to appear)

[19] K. Kailing, H. Kriegel and P. Kröger, Density-connected subspace clustering for high-dimensional data, *In Proc. 4th SIAM Int. Conf. on Data Mining* (2004), 246–257

[20] F. Klawonn, What Can Fuzzy Cluster Analysis Contribute to Clustering of High-Dimensional Data?, pp. 1–14 in *Fuzzy Logic and Applications*, F. Masulli, G.Pasi and R. Yager, (Editors), Lecture Notes in Artificial Intelligence, Springer 2013, ISBN 978-3-319-03199-6.

[21] E. Kolatch, Clustering algorithms for spatial databases: A survey, *http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.28.1145&rep=rep1&type=pdf*

[22] R.D. Luce, *Individual Choice Behavior: A Theoretical Analysis*, Wiley, New York, 1959, ISBN 0-486-44136-9.

[23] MATLAB version 7.14.0.739. Natick, Massachusetts: The MathWorks Inc., 2012.
[24] N. Megiddo and K.J. Supowit, On the complexity of some common geometric location problems, *SIAM Journal on Computing* **13**(1984), 182–196.
[25] R.W. Stanforth, E. Kolossov and B. Mirkin, A measure of domain of applicability for QSAR modelling based on intelligent K–means clustering, *QSAR Comb. Sci.* **26** (2007), 837–844.
[26] D.S. Shepard, A two-dimensional interpolation function for irregularly spaced data, *Proceedings of 23rd National Conference*, Association for Computing Machinery. Princeton, NJ: Brandon/Systems Press, 1968, pp. 517–524.
[27] M. Teboulle, A unified continuous optimization framework for center–based clustering methods, *J. Machine Learning* **8**(2007), 65–102.
[28] E. Weiszfeld, Sur le point par lequel la somme des distances de n points donnés est minimum, *Tohoku Math. J.* **43** (1937), 355–386.
[29] D.H. Ye, K.M. Pohl, H. Litt and C. Davatzikos, Groupwise morphometric analysis based on high dimensional clustering, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2010), 47–54
[30] J.I. Yellott, Jr., Luce's choice axiom. In N.J. Smelser and P.B. Baltes, editors. *International Encyclopedia of the Social & Behavioral Sciences*, pp. 9094–9097. ISBN 0-08-043076-7, 2001.
[31] B. Zhang, M. Hsu, and U. Dayal, k–Harmonic Means A Spatial Clustering Algorithm with Boosting, *Temporal, Spatial, and SpatioTemporal Data Mining*, pp. 31–45, 2000.
[32] B. Zhang, M. Hsu, and U. Dayal, Harmonic Average Based Clustering Method and System, US Patent 6,584,433, 2000.

Appendix A: Relation to previous work

Our work brings together ideas from four different areas: inverse distance weighted interpolation, fuzzy clustering, subjective probability, and optimality principles.

1. Inverse distance weighted (or **IDW**) **interpolation** was introduced in 1965 by Donald Shepard, who published his results [26] in 1968. Shepard, then an undergraduate at Harvard, worked on the following problem:

A function $u : \mathbb{R}^n \to \mathbb{R}$ is evaluated at K given points $\{\mathbf{x}_k : k \in \overline{1,K}\}$ in $\mathbb{R}^n$, giving the values $\{u_k : k \in \overline{1,K}\}$, respectively. These values are the only information about the function. It is required to estimate u at any point $\mathbf{x}$.

Examples of such functions include rainfall in meteorology, and altitude in topography. The point $\mathbf{x}$ cannot be too far from the data points, and ideally lies in their convex hull.

Shepard estimated the value $u(\mathbf{x})$ as a convex combination of the given values u_k,

$$u(\mathbf{x}) = \sum_{k=1}^{K} \lambda_k(\mathbf{x})\, u_k \tag{A-1}$$

where the weights $\lambda_k(\mathbf{x})$ are inversely ptoportional to the distances $d(\mathbf{x}, \mathbf{x}_k)$ between $\mathbf{x}$ and $\mathbf{x}_k$, say

$$u(\mathbf{x}) = \sum_{k=1}^{K} \left(\frac{\dfrac{1}{d(\mathbf{x}, \mathbf{x}_k)}}{\displaystyle\sum_{j=1}^{K} \frac{1}{d(\mathbf{x}, \mathbf{x}_j)}} \right) u_k \tag{A-2}$$

giving the weights

$$\lambda_k(\mathbf{x}) = \frac{\prod\limits_{j\neq k} d(\mathbf{x},\mathbf{x}_j)}{\sum\limits_{\ell=1}^{K}\prod\limits_{m\neq\ell} d(\mathbf{x},\mathbf{x}_m)} \tag{A-3}$$

that are identical with the probabilities (24), if the data points are identified with the centers. IDW interpolation is used widely in spatial data analysis, geology, geography, ecology and related areas.

Interpolating the K distances $d(\mathbf{x},\mathbf{x}_k)$, i.e. taking $u_k = d(\mathbf{x},\mathbf{x}_k)$ in (A-2), gives

$$K\,\frac{\prod\limits_{j=1}^{K} d(\mathbf{x},\mathbf{x}_j)}{\sum\limits_{\ell=1}^{K}\prod\limits_{m\neq\ell} d(\mathbf{x},\mathbf{x}_m)} \tag{A-4}$$

the harmonic mean of the distances $\{d(\mathbf{x},\mathbf{x}_k) : k \in \overline{1,K}\}$, which is the JDF in (25) multiplied by a scalar.

The harmonic mean pops up in several areas of spatial data analysis. In 1980 Dixon and Chapman [12] posited that the *home–range* of a species is a contour of the harmonic mean of the areas it frequents, and this has since been confirmed for hundreds of species. The importance of the harmonic mean in clustering was established by Teboulle [27], Stanforth, Kolossov and Mirkin [25], Zhang, Hsu, and Dayal [31]–[32], Ben–Israel and Iyigun [5] and others. Arav [3] showed the harmonic mean of distances to satisfy a system of reasonable axioms for contour approximation of data.

2. Fuzzy clustering introduced by J.C. Bezdek in 1973, [8], is a relaxation of the original problem, replacing the hard assignments of points to clusters by soft, or fuzzy, assignments of points simultaneously to all clusters, the strength of association of $\mathbf{x}_i$ with the k_{th} cluster is measured by $w_{ik} \in [0,1]$.

In the fuzzy c–means (FCM) method [9] the centers $\{\mathbf{c}_k\}$ are computed by

$$\min \sum_{i=1}^{N}\sum_{k=1}^{K} w_{ik}^m \,\|\mathbf{x}_i - \mathbf{c}_k\|_2^2, \tag{A-5}$$

where the weights x_{ik} are updated as[2]

$$w_{ik} = \frac{1}{\sum\limits_{j=1}^{K}\left(\dfrac{\|\mathbf{x}_i - \mathbf{c}_k\|_2}{\|\mathbf{x}_i - \mathbf{c}_j\|_2}\right)^{2/m-1}}, \tag{A-6}$$

and the centers are then calculated as convex combinations of the data points,

$$\mathbf{c}_k = \sum_{i=1}^{N}\left(\frac{w_{ik}^m}{\sum\limits_{j=1}^{N} w_{jk}^m}\right)\mathbf{x}_i.\ k \in \overline{1,K}. \tag{A-7}$$

[2]The weights (A-6) are optimal for the problem (A-5) if they are probabilities, i.e. if they are required to add to 1 for every point $\mathbf{x}_i$.

The constant $m \geq 1$ (the "fuzzifier") controls he fuzziness of the assignments, which become hard assignments in the limit as $m \downarrow 1$. For $m = 1$, FCM is the classical K–means method. If $m = 2$ then the weights w_{ik} are inversely proportional to the square distance $\|\mathbf{x}_i - \mathbf{c}_k\|_2^2$, analogously to (21).

Fuzzy clustering is one of the best known, and most widely used, clustering methods. However, it may need some modification if the data in question is very high–dimensional, see, e.g. [20].

3. Subjective probability. There is some arbitrariness in the choice of the model and the fuzzifier m in (A-5)–(A-6). In contrast, the probabilities (24) can be justified axiomatically. Using ideas and classical results from subjective probability ([22], [30]) it is shown in § 4 that the cluster membership probabilities $p_k(\mathbf{x})$, and distances $d_k(\mathbf{x})$, satisfy an inverse relationship, such as,

$$p_k(\mathbf{x})\, \phi(d(\mathbf{x}, \mathbf{c}_k)) = f(\mathbf{x}),\ k \in \overline{1,K}, \tag{A-8}$$

where $\phi(\cdot)$ is non–decreasing, and $f(\mathbf{x})$ does not depend on k. In particular, the choice $\phi(d) = d$ gives (21), which works well in practice.

4. Optimality principle. Equation (A-8) is a necessary optimality condition for the problem

$$\min \left\{ \sum_{k=1}^{K} p^2\, \phi(d(\mathbf{x}, \mathbf{c}_k)) : \sum_{k=1}^{K} p_k = 1,\ p_k \geq 0, k \in \overline{1,K} \right\}, \tag{A-9}$$

that reduces to (22) for the choice $\phi(d) = d$. This shows the probabilities $\{p_k(\mathbf{x})\}$ of (24) to be optimal, for the model chosen.

Remark 3. Minimizing a function of squares of probabilities seems unnatural, so a physical analogy may help. Consider an electric circuit with K resistances $\{R_k\}$ in parallel. A current I through the circuit splits into K currents, with current I_k through the resistance R_k. These currents solve an optimization problem (the **Kelvin principle**)

$$\min_{I_1, \cdots, I_K} \left\{ \sum_{k=1}^{K} I_k^2\, R_k : \sum_{k=1}^{K} I_k = I \right\} \tag{A-10}$$

that is analogous to (22). The optimality condition for (A-10) is **Ohm's law**,

$$I_k\, R_k = \text{constant}$$

a statement that potential is well defined, and an analog of (21). The equivalent resistance of the circuit, i.e. the resistance R such that $I^2\, R$ is equal to the minimal value in (A-10), is then the JDF (25) with R_j instead of $d(\mathbf{x}, \mathbf{c}_j)$ and $w = 1$.

Appendix B: Numerical Examples

In the following examples we use synthetic data to be clustered into $K = 2$ clusters. The data consists of two randomly generated clusters, $\mathbf{X}_1$ with N_1 points, and $\mathbf{X}_2$ with N_2 points.

The data points $\mathbf{x} = (x_1, \cdots, x_n) \in \mathbb{R}^n$ of cluster $\mathbf{X}_k$ are such that all of their components $x_i, 1 \leq i \leq n$ are generated by sampling from a distribution F_k with mean μ_k, $k = 1, 2$. In cluster $\mathbf{X}_1$ we take $\mu_1 = 1$, and in cluster $\mathbf{X}_2$, $\mu_2 = -1$.

We ran Algorithm PCM(ℓ_1), with the parameters $\nu_0 = 1$, $\Delta = 0.1$, and compared its performance with that of the fuzzy clustering method [9] with the ℓ_1 norm, as well as the generalized Weiszfeld algorithm

σ	Method	$n = 10^4$	$n = 5 \cdot 10^4$	$n = 10^5$	$n = 5 \cdot 10^5$	$n = 10^6$
$\sigma = 8$	PCM (ℓ_1)	0.0	0.0	0.0	0.0	0.0
	FCM (ℓ_1)	27.1	38.6	24.4	40.9	40.1
	K-means (ℓ_1)	28.9	26.8	12.7	22.4	22.9
	Gen. Weiszfeld	48.5	48.8	48.0	48.2	47.9
$\sigma = 16$	PCM (ℓ_1)	4.3	0.0	0.0	4.7	0.0
	FCM (ℓ_1)	41.0	42.1	44.5	43.9	39.5
	K-means (ℓ_1)	41.8	35.2	23.7	23.5	23.6
	Gen. Weiszfeld	48.0	47.0	48.4	48.6	48.0
$\sigma = 24$	PCM (ℓ_1)	42.6	8.8	0.8	4.8	0.0
	FCM (ℓ_1)	46.4	45.9	47.5	39.5	45.1
	K-means (ℓ_1)	45.5	42.6	35.6	28.0	24.5
	Gen. Weiszfeld	47.9	47.8	47.1	48.0	48.2
$\sigma = 32$	PCM (ℓ_1)	46.0	42.2	13.4	13.6	0.0
	FCM (ℓ_1)	47.4	46.0	44.8	46.0	46.0
	K-means (ℓ_1)	46.4	45.7	40.3	36.0	30.7
	Gen. Weiszfeld	48.2	48.9	48.5	48.9	47.8

TABLE 1. Percentages of misclassified data in Example 1

of [18] (that uses Euclidean distances), and the ℓ_1–K-Means algorithm [23]. For each method we used a stopping rule of at most 100 iterations (for Algorithm PCM(ℓ_1) this replaces Step 4). For each experiment we record the average percentage of misclassification (a misclassification occurs when a point in $\mathbf{X}_1$ is declared to be in $\mathbf{X}_2$, or vice versa) from 10 independent problems.

In examples 1,2,3 we take F_k as the normal distribution $N(\mu_k, \sigma)$.

Example 1. In this example the clusters are of equal size, $N_1 = N_2 = 100$. Table 1 gives the percentages of misclassification under the five methods tested, for different values of σ and dimension n.

Example 2. We use $N_1 = 200$ and $N_2 = 100$. Table 2 gives the percentages of misclassifications for different values of σ and dimension n.

Example 3. In this case $N_1 = 1000$, $N_2 = 10$. The percentages of misclassification are included in Table 3.

In addition to experiments with normal data, we also consider instances with uniform data in Examples 4 and 5. In this case F_k is a uniform distribution with mean μ_k and support length $|\text{supp}(F_k)|$.

Example 4. We use $N_1 = 100$, $N_2 = 100$. The results are shown in Table 4.

Example 5. In this instance $N_1 = 200$, $N_2 = 100$. The results are shown in Table 5.

σ	Method	$n = 10^4$	$n = 5 \cdot 10^4$	$n = 10^5$	$n = 5 \cdot 10^5$	$n = 10^6$
$\sigma = 8$	PCM (ℓ_1)	0.0	0.0	0.0	0.0	0.0
	FCM (ℓ_1)	11.9	19.1	25.2	30.1	22.6
	K-means (ℓ_1)	20.8	25.9	18.4	31.4	13.6
	Gen. Weiszfeld	37.8	37.9	37.2	36.7	36.4
$\sigma = 16$	PCM (ℓ_1)	10.4	0.0	0.0	0.0	0.0
	FCM (ℓ_1)	37.7	35.6	35.0	36.2	39.4
	K-means (ℓ_1)	35.8	32.0	23.6	31.7	14.1
	Gen. Weiszfeld	38.0	37.7	35.8	36.6	37.8
$\sigma = 24$	PCM (ℓ_1)	44.1	5.9	1.2	0.0	0.0
	FCM (ℓ_1)	41.3	37.7	38.9	36.7	34.6
	K-means (ℓ_1)	40.3	39.9	32.7	33.3	15.5
	Gen. Weiszfeld	36.8	37.7	36.7	36.9	37.2
$\sigma = 32$	PCM (ℓ_1)	47.2	38.7	18.5	0.0	0.0
	FCM (ℓ_1)	42.3	38.8	37.0	39.7	38.9
	K-means (ℓ_1)	41.5	42.9	37.2	36.8	22.6
	Gen. Weiszfeld	36.7	36.9	36.0	36.5	37.4

TABLE 2. Percentages of misclassified data in Example 2

σ	Method	$n = 10^3$	$n = 5 \cdot 10^3$	$n = 10^4$	$n = 5 \cdot 10^4$	$n = 10^5$
$\sigma = 0.4$	PCM (ℓ_1)	46.4	41.1	24.1	5.1	0.9
	FCM (ℓ_1)	13.4	0.5	0.0	19.5	32.0
	K-means (ℓ_1)	37.4	31.6	27.1	36.5	32.6
	Gen. Weiszfeld	35.4	36.7	32.6	33.7	38.7
$\sigma = 0.8$	PCM (ℓ_1)	47.4	31.4	23.4	5.4	1.8
	FCM (ℓ_1)	29.3	9.5	13.0	27.3	37.2
	K-means (ℓ_1)	37.5	32.0	27.1	36.4	32.6
	Gen. Weiszfeld	30.3	31.6	25.9	27.9	34.5
$\sigma = 1.2$	PCM (ℓ_1)	47.3	33.9	26.2	7.7	1.6
	FCM (ℓ_1)	36.4	20.8	23.2	31.1	23.9
	K-means (ℓ_1)	38.4	32. 2	28.3	36.4	32.6
	Gen. Weiszfeld	22.1	23.8	26.8	21.6	25.5
$\sigma = 1.6$	PCM (ℓ_1)	47.8	35.4	27.9	9.8	3.6
	FCM (ℓ_1)	41.1	27.8	30.0	27.6	24.2
	K-means (ℓ_1)	37.6	32.3	28.3	36.4	33.4
	Gen. Weiszfeld	23.1	23.2	21.1	25.4	31.6

TABLE 3. Percentages of misclassified data in Example 3

In all examples Algorithm PCM(ℓ_1) was unsurpassed and was the clear winner in Examples 1, 2, 4 and 5.

(Tsvetan Asamov) DEPARTMENT OF OPERATIONS RESEARCH AND FINANCIAL ENGINEERING, PRINCETON UNIVERSITY, 98 CHARLTON STREET, PRINCETON, NJ 08540, USA

E-mail address: tasamov@princeton.edu

(Adi Ben–Israel) RUTGERS BUSINESS SCHOOL, RUTGERS UNIVERSITY, 100 ROCKAFELLER ROAD, PISCATAWAY, NJ 08854, USA

E-mail address: adi.benisrael@gmail.com

$\lvert\mathrm{supp}(F)\rvert$	Method	$n = 10^4$	$n = 5 \cdot 10^4$	$n = 10^5$	$n = 5 \cdot 10^5$	$n = 10^6$
$\lvert\mathrm{supp}(F)\rvert = 8$	PCM (ℓ_1)	0.0	0.0	0.0	0.0	0.0
	FCM (ℓ_1)	0.0	0.1	0.3	2.7	0.1
	K-means (ℓ_1)	5.0	5.0	4.8	0.0	0.0
	Gen. Weiszfeld	0.0	0.0	0.0	0.0	0.0
$\lvert\mathrm{supp}(F)\rvert = 16$	PCM (ℓ_1)	0.0	0.0	0.0	0.0	0.0
	FCM (ℓ_1)	8.9	29.1	26.6	21.9	25.6
	K-means (ℓ_1)	23.8	25.9	18.0	23.4	17.8
	Gen. Weiszfeld	47.0	49.2	46.8	46.2	47.4
$\lvert\mathrm{supp}(F)\rvert = 24$	PCM (ℓ_1)	0.0	0.0	0.0	0.0	0.0
	FCM (ℓ_1)	23.6	39.1	20.1	27.8	25.4
	K-means (ℓ_1)	32.0	27.2	18.7	23.2	18.1
	Gen. Weiszfeld	47.1	47.4	48.0	47.4	47.3
$\lvert\mathrm{supp}(F)\rvert = 32$	PCM (ℓ_1)	0.3	0.0	0.0	0.0	0.0
	FCM (ℓ_1)	28.8	39.9	36.6	42.6	38.8
	K-means (ℓ_1)	35.7	27.5	19.3	23.4	18.6
	Gen. Weiszfeld	48.1	48.0	47.9	47.8	47.9

TABLE 4. Percentages of misclassified data in Example 4

$\lvert\mathrm{supp}(F)\rvert$	Method	$n = 10^4$	$n = 5 \cdot 10^4$	$n = 10^5$	$n = 5 \cdot 10^5$	$n = 10^6$
$\lvert\mathrm{supp}(F)\rvert = 8$	PCM (ℓ_1)	0.0	0.0	0.0	0.0	0.0
	FCM (ℓ_1)	0.0	10.0	7.2	0.4	0.4
	K-means (ℓ_1)	4.9	13.5	0.0	4.9	14.4
	Gen. Weiszfeld	0.0	0.0	13.1	0.0	0.0
$\lvert\mathrm{supp}(F)\rvert = 16$	PCM (ℓ_1)	0.0	0.0	0.0	0.0	0.0
	FCM (ℓ_1)	30.8	28.0	28.3	14.8	18.3
	K-means (ℓ_1)	22.2	31.8	18.4	17.6	32.0
	Gen. Weiszfeld	39.2	36.6	35.7	36.7	36.3
$\lvert\mathrm{supp}(F)\rvert = 24$	PCM (ℓ_1)	0.0	0.0	0.0	0.0	0.0
	FCM (ℓ_1)	21.0	26.1	30.3	27.5	37.6
	K-means (ℓ_1)	32.3	36.3	22.6	18.1	35.1
	Gen. Weiszfeld	37.4	38.4	37.6	36.5	37.8
$\lvert\mathrm{supp}(F)\rvert = 32$	PCM (ℓ_1)	1.5	0.0	0.0	0.0	0.0
	FCM (ℓ_1)	38.0	35.0	36.5	38.5	33.5
	K-means (ℓ_1)	35.1	36.6	23.1	18.6	35.4
	Gen. Weiszfeld	39.7	36.0	37.5	40.0	38.0

TABLE 5. Percentages of misclassified data in Example 5