Journal of Classification 25:5-26 (2008)
DOI: 10.1007/s00357-008-9002-z

Probabilistic D-Clustering

Adi Ben-Israel

Rutgers University, New Jersey

Cem Iyigun

Rutgers University, New Jersey

Abstract: We present a new iterative method for probabilistic clustering of data. Given clusters, their centers and the distances of data points from these centers, the probability of cluster membership at any point is assumed inversely proportional to the distance from (the center of) the cluster in question. This assumption is our working principle.

The method is a generalization, to several centers, of the Weiszfeld method for solving the Fermat–Weber location problem. At each iteration, the distances (Euclidean, Mahalanobis, etc.) from the cluster centers are computed for all data points, and the centers are updated as convex combinations of these points, with weights determined by the above principle. Computations stop when the centers stop moving.

Progress is monitored by the joint distance function, a measure of distance from all cluster centers, that evolves during the iterations, and captures the data in its low contours.

The method is simple, fast (requiring a small number of cheap iterations) and insensitive to outliers.

Keywords: Clustering; Probabilistic clustering; Mahalanobis distance; Harmonic mean; Joint distance function; Weiszfeld method; Similarity matrix.

Authors' Address: RUTCOR–Rutgers Center for Operations Research, Rutgers University, 640 Bartholomew Rd., Piscataway, NJ 08854-8003, USA, e-mail: adi.benisrael @gmail. com, iyigun@rutcor.rutgers.edu

Published online 18 June 2008

1. Introduction

A cluster is a set of data points that are similar, in some sense, and clustering is a process of partitioning a data set into disjoint clusters.

Clustering is a basic tool in statistics and machine learning, and has been applied in pattern recognition, medical diagnostics, data mining, biology, finance and other areas.

We take data points to be vectors $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$, and interpret "similar" as "close", in terms of a distance function $d(\mathbf{x}, \mathbf{y})$ in $\mathbb{R}^n$, such as

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|, \ \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \tag{1}$$

where the norm $\|\cdot\|$ is elliptic, defined for $\mathbf{u} = (u_i)$ by

$$\|\mathbf{u}\| = \langle \mathbf{u}, Q\mathbf{u} \rangle^{1/2}, \tag{2}$$

with $\langle \cdot, \cdot \rangle$ the standard inner product, and Q a positive definite matrix. In particular, $Q = I$ gives the Euclidean norm,

$$\|\mathbf{u}\| = \langle \mathbf{u}, \mathbf{u} \rangle^{1/2}, \tag{3}$$

and the Mahalanobis distance corresponds to $Q = \Sigma^{-1}$, where Σ is the covariance matrix of the data involved.

Example 1. *A data set in $\mathbb{R}^2$ with $N = 200$ data points is shown in Figure 1. The data was simulated, from normal distributions $N(\boldsymbol{\mu_i}, \Sigma_i)$, with:*

$$\boldsymbol{\mu}_1 = (0, 0), \ \Sigma_1 = \begin{pmatrix} 0.1 & 0 \\ 0 & 1 \end{pmatrix}, \ \textit{(100 points)}\ ,$$

$$\boldsymbol{\mu}_2 = (3, 0), \ \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 0.1 \end{pmatrix}, \ \textit{(100 points)}\ .$$

This data will serve to illustrate Examples 2–5 below.

The clustering problem is, given a dataset $\mathcal{D}$ consisting of N data points

$$\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \subset \mathbb{R}^n,$$

and an integer K, $1 < K < N$, to partition $\mathcal{D}$ into K clusters $\mathcal{C}_1, \ldots, \mathcal{C}_K$.

Data points are assigned to clusters using a clustering criterion. In distance clustering, abbreviated d–clustering, the clustering criterion is metric: With each cluster $\mathcal{C}_k$ we associate a center $\mathbf{c}_k$, for example its centroid, and each data point is assigned to the cluster to whose center it is the nearest. After each such assignment, the cluster centers may change, resulting in

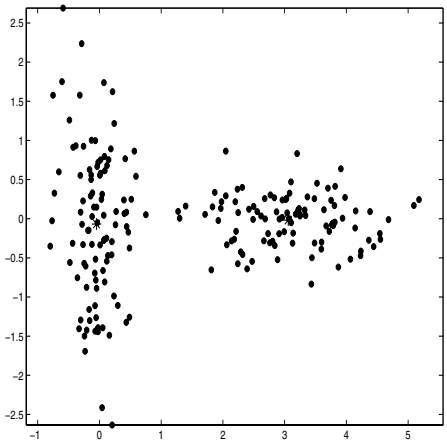

Figure 1. A data set in $\mathbb{R}^2$

re–assignments. Such an algorithm will therefore iterate between updating the centers and re–assignments.

A commonly used clustering criterion is the sum–of–squares of Euclidean distances,

$$\sum_{k=1}^{K} \sum_{\mathbf{x}_i \in \mathcal{C}_k} \| \mathbf{x}_i - \mathbf{c}_k \|^2, \tag{4}$$

to be minimized by the sought clusters $\mathcal{C}_1, \ldots, \mathcal{C}_K$. The well known k–means clustering algorithm (Hartigan 1975) uses this criterion.

In probabilistic clustering the assignment of points to clusters is "soft", in the sense that the membership of a data point $\mathbf{x}$ in a cluster $\mathcal{C}_k$ is given as a probability, denoted by $p_k(\mathbf{x})$. These are subjective probabilities, indicating strength of belief in the event in question.

Let a distance function

$$d_k(\,\cdot\,,\,\cdot\,) \tag{5}$$

be defined for each cluster $\mathcal{C}_k$. These distance functions are, in general, different from one cluster to another. For each data point $\mathbf{x} \in \mathcal{D}$, we then compute:

• the distance $d_k(\mathbf{x}, \mathbf{c}_k)$, also denoted by $d_k(\mathbf{x})$ (since d_k is used only for distances from $\mathbf{c}_k$), or just d_k if $\mathbf{x}$ is understood, and

• a probability that $\mathbf{x}$ is a member of $\mathcal{C}_k$, denoted by $p_k(\mathbf{x})$, or just p_k.

Various relations between probabilities and distances can be assumed, resulting in different ways of clustering the data. In our experience, the following assumption has proved useful: For any point $\mathbf{x}$, and all $k = 1, \cdots, K$

$$p_k(\mathbf{x})\, d_k(\mathbf{x}) = \text{constant, depending on } \mathbf{x}\ .$$

This model is our working principle in what follows, and the basis of the probabilistic d–clustering approach of Section 2.

The above principle owes its versatility to the different ways of choosing the distances $d_k(\cdot)$. It is also natural to consider increasing functions of such distances, and one useful choice is

$$p_k(\mathbf{x})e^{d_k(\mathbf{x})} = \text{constant, depending on } \mathbf{x} \ ,$$

giving the probabilistic exponential d–clustering approach of Section 3.

The probabilistic d–clustering algorithm is presented in Section 4. It is a generalization, to several centers, of the Weizsfeld method for solving the Fermat–Weber location problem, see Section 2.5, and convergence follows as in Kuhn (1973). The updates of the centers use an extremal principle, described in Section 2.3. The progress of the algorithm is monitored by the joint distance function, a distance function that captures the data in its low contours, see Section 2.2. The centers updated by the algorithm are stationary points of the joint distance function.

The paper concludes with a small example, Section 5, analyzing the liberal–conservative divide of the U.S. Supreme Court.

For other approaches to probabilistic clustering see the surveys in Höppner, Klawonn, Kruse, and Runkler (1999), Tan, Steinbach, and Kumar (2006), and the seminal article by Teboulle (2007) unifying clustering methods in the framework of modern optimization theory.

2. Probabilistic D–Clustering

There are several ways to model the relationship between distances and probabilities. The simplest model, and our working principle (or axiom), is the following:

Principle 1. *For each* $\mathbf{x} \in \mathcal{D}$, *and each cluster* $\mathcal{C}_k$,

$$p_k(\mathbf{x})\, d_k(\mathbf{x}) = \textit{constant, depending on } \mathbf{x} \ . \tag{6}$$

Cluster membership is thus more probable the closer the data point is to the cluster center. Note that the constant in (6) is independent of the cluster k.

2.1 Probabilities

From Principle 1, and the fact that probabilities add to one, we get

Theorem 1. *Let the cluster centers* $\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K\}$ *be given, let* $\mathbf{x}$ *be a data point, and let* $\{d_k(\mathbf{x}) : k = 1, \ldots, K\}$ *be its distances from the given centers. Then the membership probabilities of* $\mathbf{x}$ *are*

$$p_k(\mathbf{x}) = \frac{\prod\limits_{j\neq k} d_j(\mathbf{x})}{\sum\limits_{t=1}^{K} \prod\limits_{j\neq t} d_j(\mathbf{x})} \ , \ k = 1, \ldots, K. \tag{7}$$

Proof. Using (6) we write for t, k

$$p_t(\mathbf{x}) = \left(\frac{p_k(\mathbf{x}) d_k(\mathbf{x})}{d_t(\mathbf{x})} \right) \ .$$

Since $\sum\limits_{t=1}^{K} p_t(\mathbf{x}) = 1$,

$$p_k(\mathbf{x}) \sum_{t=1}^{K} \left(\frac{d_k(\mathbf{x})}{d_t(\mathbf{x})} \right) = 1.$$

$$\therefore \ p_k(\mathbf{x}) = \frac{1}{\sum\limits_{t=1}^{K} \left(\frac{d_k(\mathbf{x})}{d_t(\mathbf{x})} \right)} = \frac{\prod\limits_{j\neq k} d_j(\mathbf{x})}{\sum\limits_{t=1}^{K} \prod\limits_{j\neq t} d_j(\mathbf{x})} \ .$$

■

In particular, for $K = 2$,

$$p_1(\mathbf{x}) = \frac{d_2(\mathbf{x})}{d_1(\mathbf{x}) + d_2(\mathbf{x})} \ , \ p_2(\mathbf{x}) = \frac{d_1(\mathbf{x})}{d_1(\mathbf{x}) + d_2(\mathbf{x})} \ , \tag{8}$$

and for $K = 3$,

$$p_1(\mathbf{x}) = \frac{d_2(\mathbf{x}) d_3(\mathbf{x})}{d_1(\mathbf{x}) d_2(\mathbf{x}) + d_1(\mathbf{x}) d_3(\mathbf{x}) + d_2(\mathbf{x}) d_3(\mathbf{x})} \ , \text{ etc.} \tag{9}$$

Note: See Heiser (2004) for related work in a different context. In particular, our equation (8) is closely related to equation (5) in Heiser.

2.2 The Joint Distance Function

We denote the constant in (6) by $D(\mathbf{x})$, a function of $\mathbf{x}$. Then

$$p_k(\mathbf{x}) = \frac{D(\mathbf{x})}{d_k(\mathbf{x})} \ , \ k = 1, \ldots, K.$$

Since the probabilities add to one we get,

$$D(\mathbf{x}) = \frac{\prod_{k=1}^{K} d_k(\mathbf{x}, \mathbf{c}_k)}{\sum_{t=1}^{K} \prod_{j \neq t} d_j(\mathbf{x}, \mathbf{c}_j)} \,. \tag{10}$$

The function $D(\mathbf{x})$, called the joint distance function (abbreviated JDF) of $\mathbf{x}$, has the dimension of distance, and measures the distance of $\mathbf{x}$ from all cluster centers. Here are special cases of (10), for $K = 2$,

$$D(\mathbf{x}) = \frac{d_1(\mathbf{x})\, d_2(\mathbf{x})}{d_1(\mathbf{x}) + d_2(\mathbf{x})} \,, \tag{11}$$

and for $K = 3$,

$$D(\mathbf{x}) = \frac{d_1(\mathbf{x})\, d_2(\mathbf{x})\, d_3(\mathbf{x})}{d_1(\mathbf{x})\, d_2(\mathbf{x}) + d_1(\mathbf{x})\, d_3(\mathbf{x}) + d_2(\mathbf{x})\, d_3(\mathbf{x})} \,. \tag{12}$$

The JDF of the whole data set $\mathcal{D}$ is the sum of (10) over all points, and is a function of the K cluster centers, say,

$$F(\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_K) = \sum_{i=1}^{N} \frac{\prod_{k=1}^{K} d_k(\mathbf{x}_i, \mathbf{c}_k)}{\sum_{t=1}^{K} \prod_{j \neq t} d_j(\mathbf{x}_i, \mathbf{c}_j)} \,. \tag{13}$$

Example 2. *Figure 2 shows level sets of the JDF (11), with Mahalanobis distances*

$$d_k(\mathbf{x}, \mathbf{c}_k) = \sqrt{(\mathbf{x} - \mathbf{c}_k)^T \Sigma_k^{-1} (\mathbf{x} - \mathbf{c}_k)} \,, \tag{14}$$

$\mathbf{c}_1 = \boldsymbol{\mu}_1,\ \mathbf{c}_2 = \boldsymbol{\mu}_2$, *and* $\Sigma_1,\ \Sigma_2$ *as in Example 1.*

Notes:

(a) The JDF $D(\mathbf{x})$ of (10) is a measure of the classifiability of the point $\mathbf{x}$ in question. It is zero if and only if $\mathbf{x}$ coincides with one of the cluster centers, in which case $\mathbf{x}$ belongs to that cluster with probability 1. If all the distances $d_k(\mathbf{x}, \mathbf{c}_k)$ are equal, say equal to d, then $D(\mathbf{x}) = d/k$ and all $p_k(\mathbf{x}) = 1/K$, showing indifference between the clusters. As the distances $d_k(\mathbf{x})$ increase, so does $D(\mathbf{x})$, indicating greater uncertainty about the cluster where $\mathbf{x}$ belongs.

(b) The JDF (10) is, up to a constant, the harmonic mean of the distances involved, see Arav (2008) for an elucidation of the role of the harmonic mean in contour approximation of data. A related concept in ecology is the home range, shown in Dixon and Chapman (1980) to be the harmonic mean of the area moments in question.

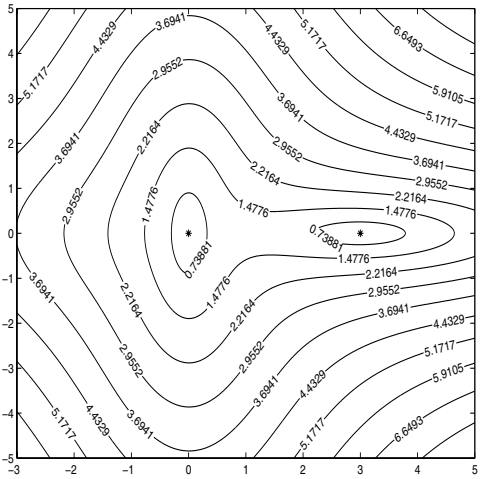

Figure 2. Level sets of a joint distance function

2.3 An Extremal Principle

For simplicity consider the case of two clusters (the results are easily extended to the general case.)

Let $\mathbf{x}$ be a given data point with distances $d_1(\mathbf{x}),\ d_2(\mathbf{x})$ to the cluster centers. Then the probabilities in (8) are the optimal solutions p_1, p_2 of the extremal problem

$$\begin{array}{ll} \text{Minimize} & d_1(\mathbf{x})\, p_1^2 + d_2(\mathbf{x})\, p_2^2 \\ \text{subject to} & p_1 + p_2 = 1 \\ & p_1,\ p_2 \geq 0 \end{array} \tag{15}$$

Indeed, the Lagrangian of this problem is

$$L(p_1, p_2, \lambda) = d_1(\mathbf{x})\, p_1^2 + d_2(\mathbf{x})\, p_2^2 - \lambda(p_1 + p_2 - 1) \tag{16}$$

and setting the partial derivatives (with respect to $p_1,\ p_2$) equal to zero gives the principle (6),

$$p_1\, d_1(\mathbf{x}) = p_2\, d_2(\mathbf{x})\ .$$

Substituting the probabilities (8) in the Lagrangian (16) we get the optimal value of (15),

$$L^*(p_1(\mathbf{x}), p_2(\mathbf{x}), \lambda) = \frac{d_1(\mathbf{x})\, d_2(\mathbf{x})}{d_1(\mathbf{x}) + d_2(\mathbf{x})}\ , \tag{17}$$

which is the JDF (11) again.

The extremal problem for a data set $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \subset \mathbb{R}^n$ is, accordingly,

$$\begin{aligned} \text{Minimize} \quad & \sum_{i=1}^{N} \left(d_1(\mathbf{x}_i)\, p_1(\mathbf{x}_i)^2 + d_2(\mathbf{x}_i)\, p_2(\mathbf{x}_i)^2 \right) \\ \text{subject to} \quad & p_1(\mathbf{x}_i) + p_2(\mathbf{x}_i) = 1, \\ & p_1(\mathbf{x}_i),\, p_2(\mathbf{x}_i) \geq 0,\ i = 1, \ldots, N. \end{aligned} \tag{18}$$

This problem separates into N problems like (15), and its optimal value is

$$\sum_{i=1}^{N} \frac{d_1(\mathbf{x}_i)\, d_2(\mathbf{x}_i)}{d_1(\mathbf{x}_i) + d_2(\mathbf{x}_i)} \tag{19}$$

the JDF (13) of the data set, with $K = 2$.

Note: The explanation for the strange appearance of "probabilities squared" above, is that (15) is a smoothed version of the "real" clustering problem, namely,

$$\min \{d_1, d_2\},$$

which is nonsmooth, see Teboulle (2007) for a unified development of smoothed clustering methods.

2.4 Centers

We write (18) as a function of the cluster centers $\mathbf{c}_1, \mathbf{c}_2$,

$$f(\mathbf{c}_1, \mathbf{c}_2) = \sum_{i=1}^{N} \left(d_1(\mathbf{x}_i, \mathbf{c}_1)\, p_1(\mathbf{x}_i)^2 + d_2(\mathbf{x}_i, \mathbf{c}_2)\, p_2(\mathbf{x}_i)^2 \right) . \tag{20}$$

If a point $\mathbf{x}_i$ coincides with a center, say $\mathbf{x}_i = \mathbf{c}_1$, then $d_1(\mathbf{x}_i) = 0$, $p_1(\mathbf{x}_i) = 1$ and $p_2(\mathbf{x}_i) = 0$. This point contributes zero to the summation.

For the special case of Euclidean distances, the minimizers of (20) assume a simple form as convex combinations of the data points.

Theorem 2. *Let the distance functions d_1, d_2 in (20) be Euclidean,*

$$d_k(\mathbf{x}, \mathbf{c}_k) = \|\mathbf{x} - \mathbf{c}_k\| \ , \ k = 1, 2 \ , \tag{21}$$

so that

$$f(\mathbf{c}_1, \mathbf{c}_2) = \sum_{i=1,\ldots,N} \left(\|\mathbf{x}_i - \mathbf{c}_1\|\, p_1(\mathbf{x}_i)^2 + \|\mathbf{x}_i - \mathbf{c}_2\|\, p_2(\mathbf{x}_i)^2 \right) , \tag{22}$$

and let the probabilities be given for $i = 1, \ldots, N$. *We make the following assumption about the minimizers* $\mathbf{c}_1, \mathbf{c}_2$ of (22):

$$\mathbf{c}_1, \mathbf{c}_2 \textit{ do not coincide with any of the points } \mathbf{x}_i,\ i = 1, \ldots, N. \tag{23}$$

Then the minimizers $\mathbf{c}_1, \mathbf{c}_2$ *are given by*

$$\mathbf{c}_k = \sum_{i=1,\ldots,N} \left(\frac{u_k(\mathbf{x}_i)}{\sum\limits_{j=1,\ldots,N} u_k(\mathbf{x}_j)} \right) \mathbf{x}_i \ , \tag{24}$$

where

$$u_k(\mathbf{x}_i) = \frac{p_k(\mathbf{x}_i)^2}{d_k(\mathbf{x}_i, \mathbf{c}_k)} \ , \tag{25}$$

for $k = 1, 2$, *or equivalently, using* (8),

$$\begin{aligned} u_1(\mathbf{x}_i) &= \frac{d_2(\mathbf{x}_i, \mathbf{c}_2)^2}{d_1(\mathbf{x}_i, \mathbf{c}_1)\left(d_1(\mathbf{x}_i, \mathbf{c}_1) + d_2(\mathbf{x}_i, \mathbf{c}_2)\right)^2} \ , \\ u_2(\mathbf{x}_i) &= \frac{d_1(\mathbf{x}_i, \mathbf{c}_1)^2}{d_2(\mathbf{x}_i, \mathbf{c}_2)\left(d_1(\mathbf{x}_i, \mathbf{c}_1) + d_2(\mathbf{x}_i, \mathbf{c}_2)\right)^2} \ . \end{aligned} \tag{26}$$

Proof. The gradient of $d(\mathbf{x}, \mathbf{c}) = \|\mathbf{x} - \mathbf{c}\|$ with respect to $\mathbf{c}$ is, for $\mathbf{x} \neq \mathbf{c}$,

$$\nabla_{\mathbf{c}} \|\mathbf{x} - \mathbf{c}\| = -\frac{\mathbf{x} - \mathbf{c}}{\|\mathbf{x} - \mathbf{c}\|} = -\frac{\mathbf{x} - \mathbf{c}}{d(\mathbf{x}, \mathbf{c})} \ . \tag{27}$$

By Assumption (23), the gradient of (22) with respect to $\mathbf{c}_k$ is

$$\begin{aligned} \nabla_{\mathbf{c}_k} f(\mathbf{c}_1, \mathbf{c}_2) &= -\sum_{i=1,\ldots,N} \frac{\mathbf{x}_i - \mathbf{c}_k}{\|\mathbf{x}_i - \mathbf{c}_k\|} p_k(\mathbf{x}_i)^2 \\ &= -\sum_{i=1,\ldots,N} \frac{\mathbf{x}_i - \mathbf{c}_k}{d_k(\mathbf{x}_i, \mathbf{c}_k)} p_k(\mathbf{x}_i)^2 \ , k = 1, 2 \ . \end{aligned} \tag{28}$$

Setting the gradient equal to zero, and summing like terms, we get

$$\sum_{i=1,\ldots,N} \left(\frac{p_k(\mathbf{x}_i)^2}{d_k(\mathbf{x}_i, \mathbf{c}_k)} \right) \mathbf{x}_i = \left(\sum_{i=1,\ldots,N} \frac{p_k(\mathbf{x}_i)^2}{d_k(\mathbf{x}_i, \mathbf{c}_k)} \right) \mathbf{c}_k \ , \tag{29}$$

proving (24)–(26).

■

The same formulas for the centers $\mathbf{c}_1, \mathbf{c}_2$ hold if the norm used in (22) is elliptic.

Corollary 1. *Let the distance functions* d_1, d_2 *in* (20) *be elliptic,*

$$d_k(\mathbf{x}, \mathbf{c}_k) = \langle (\mathbf{x} - \mathbf{c}_k), Q_k(\mathbf{x} - \mathbf{c}_k) \rangle^{1/2} , \tag{30}$$

with positive–definite matrices Q_k. *Then the minimizers* $\mathbf{c}_1, \mathbf{c}_2$ *of* (20) *are given by* (24)–(26).

Proof. The gradient of $d(\mathbf{x}, \mathbf{c}) = \langle (\mathbf{x} - \mathbf{c}), Q(\mathbf{x} - \mathbf{c}) \rangle^{1/2}$ with respect to $\mathbf{c}$ is, for $\mathbf{x} \neq \mathbf{c}$,

$$\nabla_{\mathbf{c}}\, d(\mathbf{x}, \mathbf{c}) = -\frac{Q(\mathbf{x} - \mathbf{c})}{d(\mathbf{x}, \mathbf{c})} .$$

Therefore the analog of (28) is

$$\nabla_{\mathbf{c}_k} f(\mathbf{c}_1, \mathbf{c}_2) = -Q_k \sum_{i=1,\dots,N} \frac{\mathbf{x}_i - \mathbf{c}_k}{d_k(\mathbf{x}_i, \mathbf{c}_k)}\, p_k(\mathbf{x}_i)^2 , \tag{31}$$

and since Q_k is nonsingular, it can be "cancelled" when we set the gradient equal to zero. The rest of the proof is as in Theorem 2.
■

Corollary 1 applies, in particular, to the Mahalanobis distance (14)

$$d_k(\mathbf{x}, \mathbf{c}_k) = \sqrt{(\mathbf{x} - \mathbf{c}_k)^T \Sigma_k^{-1} (\mathbf{x} - \mathbf{c}_k)} ,$$

where Σ_k is the covariance matrix of the cluster $\mathcal{C}_k$.

The formulas (24)–(25) are also valid in the general case of K clusters, where the analog of (20) is

$$f(\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_K) = \sum_{i=1,\dots,N} \sum_{k=1}^{K} d_k(\mathbf{x}_i, \mathbf{c}_k)\, p_k(\mathbf{x}_i)^2 . \tag{32}$$

Corollary 2. *Let the distance functions* d_k *in* (32) *be elliptic, as in* (30), *and let the probabilities* $p_k(\mathbf{x}_i)$ *be given. Then the minimizers* $\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_K$ *of* (32) *are given by* (24)–(25) for $k = 1, 2, \cdots, K$.

Proof. The proof of Corollary 1 holds in the general case, since the minimizers are calculated separately.
■

2.5 The Weiszfeld Method

In the case of one cluster (where the probabilities are all 1 and therefore of no interest) the center formulas (24)–(25) reduce to

$$\mathbf{c} = \sum_{i=1,\dots,N} \left(\frac{1/d(\mathbf{x}_i, \mathbf{c})}{\sum\limits_{j=1,\dots,N} 1/d(\mathbf{x}_j, \mathbf{c})} \right) \mathbf{x}_i , \tag{33}$$

giving the minimizer of $f(\mathbf{c}) = \sum_{i=1}^{N} d(\mathbf{x}_i, \mathbf{c})$. Formula (33) can be used iteratively to update the center $\mathbf{c}$ (on the left) as a convex combination of the points $\mathbf{x}_i$ with weights depending on the current center. This iteration is the Weiszfeld method (Weiszfeld 1937) for solving the Fermat–Weber location problem, see Weiszfeld (1937), and Love, Morris, and Wesolowsky (1988). Convergence of Weiszfeld's method was established in Kuhn (1973) by modifying the gradient $\nabla f(\mathbf{c})$ so that it is always defined, see Ostresh (1978) for further details. However, the modification is not carried out in practice since, as shown by Kuhn, the set of initial points $\mathbf{c}$ for which it ever becomes necessary is denumerable.

In what follows we use the formulas (24)–(25) iteratively to update the centers. Convergence can be proved by adapting the arguments of Kuhn (1973), but as there it requires no special steps in practice.

2.6 The Centers and the Joint Distance Function

The centers given by (24)–(25) are related to the JDF (13) of the data set. Consider first the case of $K = 2$ clusters, where (13) reduces to

$$F(\mathbf{c}_1, \mathbf{c}_2) = \sum_{i=1}^{N} \frac{d_1(\mathbf{x}_i, \mathbf{c}_1)\, d_2(\mathbf{x}_i, \mathbf{c}_2)}{d_1(\mathbf{x}_i, \mathbf{c}_1) + d_2(\mathbf{x}_i, \mathbf{c}_2)} \,. \tag{34}$$

The points $\mathbf{c}_k$ where $\nabla_{\mathbf{c}_k} F(\mathbf{c}_1, \mathbf{c}_2) = \mathbf{0}, k = 1, 2$, are called stationary points of (34).

Theorem 3. *Let the distances d_1, d_2 in* (34) *be elliptic, as in* (30). *Then the stationary points of $F(\mathbf{c}_1, \mathbf{c}_2)$ are given by* (24)–(26).

Proof. Let the distances d_k be Euclidean, $d_k(\mathbf{x}) = \|\mathbf{x} - \mathbf{c}_k\|$. It is enough to prove the theorem for one center, say $\mathbf{c}_1$. Using (27) we derive

$$\nabla_{\mathbf{c}_1} F(\mathbf{c}_1, \mathbf{c}_2) =$$

$$\sum_{i=1}^{N} \frac{(d_1(\mathbf{x}_i) + d_2(\mathbf{x}_i))\, d_2(\mathbf{x}_i) \left(-\dfrac{\mathbf{x}_i - \mathbf{c}_1}{d_1(\mathbf{x}_i)}\right) + d_1(\mathbf{x}_i)\, d_2(\mathbf{x}_i) \left(\dfrac{\mathbf{x}_i - \mathbf{c}_1}{d_1(\mathbf{x}_i)}\right)}{(d_1(\mathbf{x}_i) + d_2(\mathbf{x}_i))^2}$$

$$= \sum_{i=1}^{N} \frac{-d_2(\mathbf{x}_i)^2\, (\mathbf{x}_i - \mathbf{c}_1)}{d_1(\mathbf{x}_i)\, (d_1(\mathbf{x}_i) + d_2(\mathbf{x}_i))^2} \,. \tag{35}$$

Setting (35) equal to zero, and summing like terms, we get

$$\left(\sum_{j=1}^{N} \frac{d_2(\mathbf{x}_j)^2}{d_1(\mathbf{x}_j)\,(d_1(\mathbf{x}_j)+d_2(\mathbf{x}_j))^2}\right)\mathbf{c}_1 = \sum_{i=1}^{N}\left(\frac{d_2(\mathbf{x}_i)^2}{d_1(\mathbf{x}_i)\,(d_1(\mathbf{x}_i)+d_2(\mathbf{x}_i))^2}\right)\mathbf{x}_i\ ,$$

duplicating (24)–(26). If the distances are elliptic, as in (30), then the analog of (35) is,

$$\nabla_{\mathbf{c}_1} F(\mathbf{c}_1,\mathbf{c}_2) = \sum_{i=1}^{N} \frac{-d_2(\mathbf{x}_i)^2\, Q_1\,(\mathbf{x}_i-\mathbf{c}_1)}{d_1(\mathbf{x}_i)\,(d_1(\mathbf{x}_i)+d_2(\mathbf{x}_i))^2}$$

and since Q_1 is nonsingular, it can be "cancelled" when the gradient is set equal to zero.

■

In the above proof the stationary points $\mathbf{c}_1, \mathbf{c}_2$ are calculated separately, and the calculation does not depend on there being 2 clusters. We thus have:

Corollary 3. *Consider a data set with K clusters, and elliptic distances d_k. Then the stationary points of the JDF* (13) *are the centers $\mathbf{c}_k$ given by* (24)–(25).

Note: The JDF (10) is zero exactly at the K centers $\{\mathbf{c}_k\}$, and is positive elsewhere. These centers are therefore the global minimizers of (10). However, the function (10) is not convex, not even quasi–convex, and may have other stationary points, that are necessarily saddle points.

2.7 Why d and Not d^2?

The extremal principle (18), which is the basis of our work, is linear in the distances d_k,

$$\text{Minimize} \quad \sum_k d_k\, p_k^2.$$

We refer to this as the d–model.

In clustering, and statistics in general, it is customary to use the distances squared in the objective function,

$$\text{Minimize} \quad \sum_k d_k^2.$$

We call this the d^2–model.

The d^2–model has a long tradition, dating back to Gauss, and is endowed with a rich statistical theory. There are geometrical advantages (Pythagoras Theorem), as well as analytical (linear derivatives).

The d–model is suggested by the analogy between clustering and location problems, where sums of distances (not distances squared) are minimized. Our center formulas (24)–(25) are thus generalizations of the Weiszfeld Method to several facilities, see Section 2.5.

An advantage of the d–model is its robustness. Indeed the formula (25), which does not follow from the d^2–model, guarantees that outliers will not affect the center locations.

2.8 Other Principles

There are alternative ways of modelling the relations between distances and probabilities. For example:

Principle 2. *For each $\mathbf{x} \in \mathcal{D}$, and each cluster $\mathcal{C}_k$, the probability $p_k = p_k(\mathbf{x})$ and distance $d_k = d_k(\mathbf{x})$ are related by*

$$p_k^{\alpha}\, d_k^{\beta} = \textit{constant, depending on } \mathbf{x}\ . \tag{36}$$

where the exponents α, β are positive.

For the case of 2 clusters we get, by analogy with (8) and (18) respectively, the probabilities

$$p_1(\mathbf{x}) = \frac{d_2(\mathbf{x})^{\beta/\alpha}}{d_1(\mathbf{x})^{\beta/\alpha} + d_2(\mathbf{x})^{\beta/\alpha}}\ ,\ p_2(\mathbf{x}) = \frac{d_1(\mathbf{x})^{\beta/\alpha}}{d_1(\mathbf{x})^{\beta/\alpha} + d_2(\mathbf{x})^{\beta/\alpha}}\ , \tag{37}$$

and an extremal principle,

$$\begin{aligned} \text{Minimize} \quad & \sum_{i=1}^{N} \left(d_1(\mathbf{x}_i)^{\beta}\, p_1(i)^{\alpha+1} + d_2(\mathbf{x}_i)^{\beta}\, p_2(i)^{\alpha+1} \right) \\ \text{subject to} \quad & p_1(i) + p_2(i) = 1 \\ & p_1(i),\, p_2(i) \geq 0 \end{aligned} \tag{38}$$

where $p_1(i), p_2(i)$ are the cluster probabilities at $\mathbf{x}_i$.

The Fuzzy Clustering Method (Bezdek 1973, Bezdek 1981), which is an extension of k-means method, uses $\beta = 2$ and allows different choices of α. For $\alpha = 2$, it gives the same probabilities as (7), however the center updates are different than (24)–(25).

3. Probabilistic Exponential D–Clustering

Any increasing function of the distance can be used in Principle 1. The following model, with probabilities decaying exponentially as distances increase, has proved useful in our experience.

Principle 3. *For each* $\mathbf{x} \in \mathcal{D}$, *and each cluster* $\mathcal{C}_k$, *the probability* $p_k(\mathbf{x})$ *and distance* $d_k(\mathbf{x})$ *are related by*

$$p_k(\mathbf{x})\, e^{d_k(\mathbf{x})} = E(\mathbf{x}),\ \textit{a constant depending on } \mathbf{x}\ . \tag{39}$$

Most results of Section 2 hold also for Principle 3, with the distance $d_k(\mathbf{x})$ replaced by $e^{d_k(\mathbf{x})}$. Thus the analog of the probabilities (8) is

$$p_1(\mathbf{x}) = \frac{e^{d_2(\mathbf{x})}}{e^{d_1(\mathbf{x})} + e^{d_2(\mathbf{x})}}\ ,\ p_2(\mathbf{x}) = \frac{e^{d_1(\mathbf{x})}}{e^{d_1(\mathbf{x})} + e^{d_2(\mathbf{x})}}\ , \tag{40}$$

or equivalently

$$p_1(\mathbf{x}) = \frac{e^{-d_1(\mathbf{x})}}{e^{-d_1(\mathbf{x})} + e^{-d_2(\mathbf{x})}}\ ,\ p_2(\mathbf{x}) = \frac{e^{-d_2(\mathbf{x})}}{e^{-d_1(\mathbf{x})} + e^{-d_2(\mathbf{x})}}\ . \tag{41}$$

Similarly, since the probabilities add to 1, the constant in (39) is

$$E(\mathbf{x}) = \frac{e^{d_1(\mathbf{x})+d_2(\mathbf{x})}}{e^{d_1(\mathbf{x})} + e^{d_2(\mathbf{x})}}\ , \tag{42}$$

called the exponential JDF.

3.1 An Extremal Principle

The probabilities (40) are the optimal solutions of the problem

$$\min_{p_1,p_2} \quad \left\{ e^{d_1} p_1^2 + e^{d_2} p_2^2 \ :\ p_1 + p_2 = 1\ ,\ p_1,\, p_2 \geq 0 \right\}\ , \tag{43}$$

whose optimal value, obtained by substituting the probabilities (40), is again the exponential JDF (42).

The extremal problem for a data set $\mathcal{D} = \{\mathbf{x}_1,\, \mathbf{x}_2,\, \ldots,\, \mathbf{x}_N\} \subset \mathbb{R}^n$, partitioned into 2 clusters, is the following analog of (18)

$$\begin{aligned} \text{Minimize} \quad & \sum_{i=1}^{N} \left(e^{d_1(\mathbf{x}_i)}\, p_1(i)^2 + e^{d_2(\mathbf{x}_i)}\, p_2(i)^2 \right) \\ \text{subject to} \quad & p_1(i) + p_2(i) = 1 \\ & p_1(i),\, p_2(i) \geq 0 \end{aligned} \tag{44}$$

where $p_1(i), p_2(i)$ are the cluster probabilities at $\mathbf{x}_i$. The problem separates into N problems like (43), and its optimal value is

$$\sum_{i=1}^{N} \frac{e^{d_1(\mathbf{x}_i)+d_2(\mathbf{x}_i)}}{e^{d_1(\mathbf{x}_i)} + e^{d_2(\mathbf{x}_i)}} \ , \tag{45}$$

the exponential JDF of the whole data set.

Alternatively, (39) follows from the "smoothed" extremal principle

$$\min_{p_1,p_2} \quad \left\{ \sum_{k=1}^{2} p_k \, d_k + \sum_{k=1}^{2} p_k \, \log p_k \ : \ p_1 + p_2 = 1 \ , \ p_1, \, p_2 \geq 0 \right\} \ , \tag{46}$$

obtained by adding an entropy term to $\sum p_k \, d_k$. Indeed the Lagrangian of (46) is

$$L(p_1, p_2, \lambda) = \sum_{k=1}^{2} p_k \, d_k + \sum_{k=1}^{2} p_k \, \log p_k - \lambda \, (p_1 + p_2 - 1) \ .$$

Differentiation with respect to p_k, and equating to 0, gives

$$d_k + 1 + \log p_k - \lambda = 0$$

which is (39).

3.2 Centers

We write (44) as a function of the cluster centers $\mathbf{c}_1, \mathbf{c}_2$,

$$f(\mathbf{c}_1, \mathbf{c}_2) = \sum_{i=1}^{N} \left(e^{d_1(\mathbf{x}_i, \mathbf{c}_1)} \, p_1(\mathbf{x}_i)^2 + e^{d_2(\mathbf{x}_i, \mathbf{c}_2)} \, p_2(\mathbf{x}_i)^2 \right) \tag{47}$$

and for elliptic distances we can verify, as in Theorem 2, that the minimizers of (47) are given by,

$$\mathbf{c}_k = \sum_{i=1}^{N} \left(\frac{u_k(\mathbf{x}_i)}{\sum\limits_{j=1}^{N} u_k(\mathbf{x}_j)} \right) \mathbf{x}_i \ , \tag{48}$$

where (compare with (25)),

$$u_k(\mathbf{x}_i) = \frac{p_k(\mathbf{x}_i)^2 \, e^{d_k(\mathbf{x}_i)}}{d_k(\mathbf{x}_i)} \ , \tag{49}$$

or equivalently,

$$u_1(\mathbf{x}_i) = \frac{e^{-d_1(\mathbf{x}_i)}/d_1(\mathbf{x}_i)}{(e^{-d_1(\mathbf{x}_i)} + e^{-d_2(\mathbf{x}_i)})^2}\ , \quad u_2(\mathbf{x}_i) = \frac{e^{-d_2(\mathbf{x}_i)}/d_2(\mathbf{x}_i)}{(e^{-d_1(\mathbf{x}_i)} + e^{-d_2(\mathbf{x}_i)})^2}\ . \tag{50}$$

As in Theorem 3, these minimizers are the stationary points of the JDF, given here as

$$F(\mathbf{c}_1, \mathbf{c}_2) = \sum_{i=1}^{N} \frac{e^{d_1(\mathbf{x}_i,\mathbf{c}_1)+d_2(\mathbf{x}_i,\mathbf{c}_2)}}{e^{d_1(\mathbf{x}_i,\mathbf{c}_1)} + e^{d_2(\mathbf{x}_i,\mathbf{c}_2)}}\ . \tag{51}$$

Finally we can verify, as in Corollary 2, that the results hold in the general case of K clusters.

4. A Probabilistic D–clustering Algorithm

The ideas of Sections 2–3 are implemented in the following algorithm for unsupervised clustering of data. A schematic description, presented – for simplicity – for the case of 2 clusters, follows.

Initialization:	given data $\mathcal{D}$, any two points $\mathbf{c}_1, \mathbf{c}_2$, and $\epsilon > 0$
Iteration:	
Step 1	compute distances $d_1(\mathbf{x})$, $d_2(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{D}$
Step 2	update the centers $\mathbf{c}_1^+, \mathbf{c}_2^+$
Step 3	if $\|\mathbf{c}_1^+ - \mathbf{c}_1\| + \|\mathbf{c}_2^+ - \mathbf{c}_2\| < \epsilon$ stop
	return to step 1

The algorithm iterates between the cluster centers, (24) or (48), and the distances of the data points to these centers. The cluster probabilities, (8) or (40), are not used explicitly.

Notes:

(a) The distance used in Step 1 can be Euclidean or elliptic (the formulas (24)–(26), and (48)–(50), are valid in both cases.)

(b) In Step 2, the centers are updated by (24)–(26) if Principle 1 is used, and by (48)–(50) for Principle 3.

(c) In particular, if the Mahalanobis distance (14)

$$d(\mathbf{x}, \mathbf{c}_k) = \sqrt{(\mathbf{x} - \mathbf{c}_k)^T \Sigma_k^{-1} (\mathbf{x} - \mathbf{c}_k)}$$

is used, the covariance matrix Σ_k of the k^{th}–cluster, can be estimated at each iteration by

$$\Sigma_k = \frac{\sum_{i=1}^{N} u_k(\mathbf{x}_i)(\mathbf{x}_i - \mathbf{c}_k)(\mathbf{x}_i - \mathbf{c}_k)^T}{\sum_{i=1}^{N} u_k(\mathbf{x}_i)} \tag{52}$$

with $u_k(\mathbf{x}_i)$ given by (26) or (50).

(d) The computations stop (in Step 3) when the centers stop moving, at which point the cluster membership probabilities may be computed by (8) or (40). These probabilities are not needed in the algorithm, but may be used for classifying the data points, after the cluster centers have been computed.

(e) Using the arguments of Kuhn (1973) it can be shown that the objective function (32) decreases at each iteration, and the Algorithm converges.

(f) The cluster centers and distance functions change at each iteration, and so does the function (13) itself, which decreases at each iteration. The JDF may have stationary points that are not minimizers, however such points are necessarily saddle points, and will be missed by the Algorithm with probability 1.

Example 3. *We apply the algorithm, using d–clustering as in Section 2 and Mahalanobis distance, to the data of Example 1. Figure 3 shows the evolution of the joint distance function, represented by its level sets. The initial function, shown in the top-left pane, corresponds to the (arbitrarily chosen) initial centers and initial covariances* $\Sigma_1 = \Sigma_2 = I$. *The covariances are updated at each iteration using (52), and by iteration 8 the function is already very close to its final form, shown in the bottom-right pane. For a tolerance of* $\epsilon = 0.01$ *the algorithm terminated in 12 iterations.*

Example 4. *In Figure 4 we illustrate the movement of the cluster centers for different initial centers. The centers at each run are shown with the final level sets of the joint distance function found in Example 3.*

The algorithm gives the correct cluster centers, for all initial starts. In particular, the two initial centers may be arbitrarily close, as shown in the top-left pane of Figure 4.

Example 5. *The class membership probabilities (8) were then computed using the centers determined by the algorithm. The level sets of the probability* $p_1(\mathbf{x})$ *are shown in Figure 5. The curve* $p_1(\mathbf{x}) = 0.5$, *the thick curve shown in the left pane of Figure 5, may serve as the clustering rule. Alternatively, the 2 clusters can be defined as*

$$\mathcal{C}_1 = \{\mathbf{x} : \, p_1(\mathbf{x}) \geq 0.6\}, \, \mathcal{C}_2 = \{\mathbf{x} : \, p_1(\mathbf{x}) \leq 0.4\} \, ,$$

with points $\{\mathbf{x} : \, 0.4 < p_1(\mathbf{x}) < 0.6\}$ *left unclassified, see the right pane of Figure 5.*

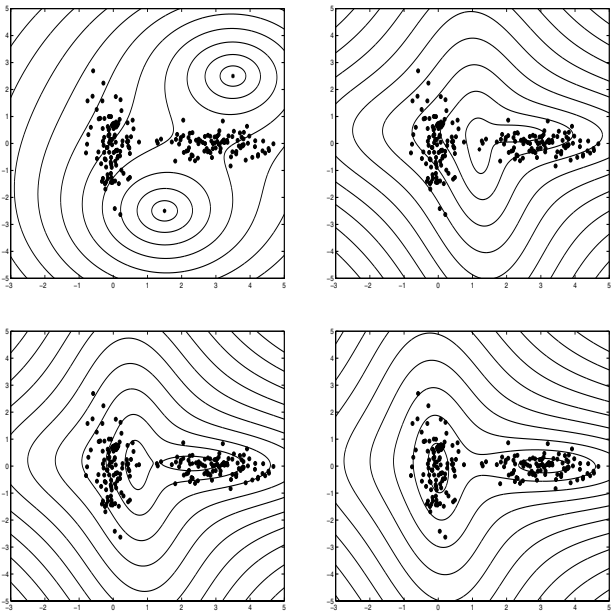

Figure 3. The level sets of the evolving joint distance function at iteration 0 (top left), iteration 1 (top right), iteration 2 (bottom left) and iteration 12 (bottom right)

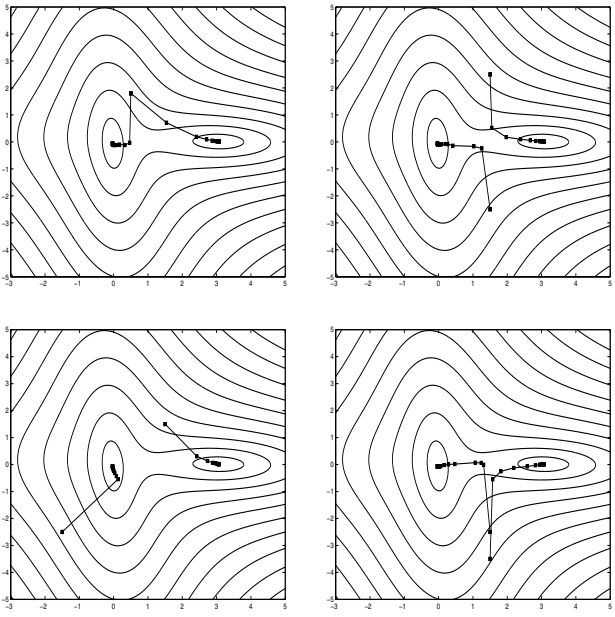

Figure 4. Movements of the cluster centers for different starts. The top–right pane shows the centers corresponding to Fig. 3. The top–left pane shows very close initial centers.

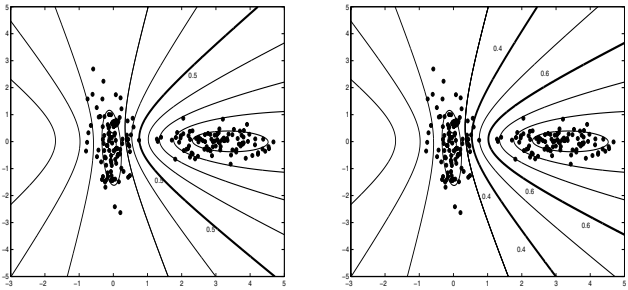

Figure 5. The level sets of the probabilities $p_1(\mathbf{x})$ and two clustering rules.

5. The Liberal-Conservative Divide of the Rehnquist Court

In many applications the data is given as similarity matrices. A small example of this type is considered next.

The Rehnquist Supreme Court was analyzed by Hubert and Steinley in 2005, where the justices were ranked as follows, from most liberal to most conservative.

Liberals	Conservatives
1. John Paul Stevens (St)	5. Sandra Day O'Connor (Oc)
2. Stephen G.Breyer (Br)	6. Anthony M. Kennendy (Ke)
3. Ruth Bader Ginsberg (Gi)	7. William H. Rehnquist (Re)
4. David Souter (So)	8. Antonin Scalia (Sc)
	9. Clarence Thomas (Th)

The data used in the analysis is a 9×9 similarity matrix, giving the percentages of non-unanimous cases in which justices *agreed*, see Table 1 (a mirror image of Table 1 in Hubert and Steinley (2005), listing the disagreements.)

Hubert and Steinley used two methods, unidimensional scaling (mapping the data from $\mathbb{R}^9$ to $\mathbb{R}$), and hierarchical classification.

We applied our method to the Rehnquist Court, with Justices represented by points $\mathbf{x}$ in $\mathbb{R}^9$ (the columns of Table 1), using the Euclidean distance in $\mathbb{R}^9$. Our results are given in Table 2, listing the clusters and their membership probabilities.

The membership probability of a Justice in a cluster is, by (6), proportional to the proximity to the cluster center, and is thus a measure of the agreement of the Justice with others in the cluster.

Since not all non–unanimous cases were equally important, or equally revealing of ideology, we should not read into these probabilities more than is supported by the data. For example, Justice Kennedy (probability 0.7540) is not "more conservative" than Justice Scalia (probability 0.7173), but perhaps "more conformist" with the "conservative center".

Table 1. Similarities among the nine Supreme Court justices

	St	Br	Gi	So	Oc	Ke	Re	Sc	Th
1 St	1.00	.62	.66	.63	.33	.36	.25	.14	.15
2 Br	.62	1.00	.72	.71	.55	.47	.43	.25	.24
3 Gi	.66	.72	1.00	.78	.47	.49	.43	.28	.26
4 So	.63	.71	.78	1.00	.55	.50	.44	.31	.29
5 Oc	.33	.55	.47	.55	1.00	.67	.71	.54	.54
6 Ke	.36	.47	.49	.50	.67	1.00	.77	.58	.59
7 Re	.25	.43	.43	.44	.71	.77	1.00	.66	.68
8 Sc	.14	.25	.28	.31	.54	.58	.66	1.00	.79
9 Th	.15	.24	.26	.29	.54	.59	.68	.79	1.00

Table 2. The liberal–conservative divide of the Rehnquist Court

Cluster	Justice	Membership Probability
Liberal	Ruth Bader Ginsburg	0.8685
	David Souter	0.8390
	Stephen Breyer	0.7922
	John Paul Stevens	0.7144
Conservative	William Rehnquist	0.8966
	Anthony Kennedy	0.7540
	Clarence Thomas	0.7220
	Antonin Scalia	0.7173
	Sandra Day O'Connor	0.6740

Similarly, Justice Stevens, ranked "most liberal" in Hubert and Steinley (2005), is in our analysis the "least conformist" in the liberal cluster.

Overall, the liberal cluster is tighter, and more conformist, than the conservative cluster.

6. Related work

There are applications where the cluster sizes (ignored here) need to be estimated. An important example is parameter estimation in mixtures of distributions. The above method, adjusted for cluster sizes, is applicable, and in particular presents a viable alternative to the EM method, see Iyigun and Ben-Israel (2008a) and Iyigun and Ben-Israel (2008).

As noted at the end of Section 2.4, our method allows an extension of the classical Weiszfeld method to several facilities. This is the subject of Iyigun and Ben-Israel (2008b), giving the solution of multi-facility location problems, including the capacitated case (which corresponds to given cluster sizes.)

A simple and practical criterion for clustering validity, determining the "right" number of clusters that fit a given data, is given in Iyigun and

Ben-Israel (2008c). This criterion is based on the monotonicity of the JDF (13) as a function of the number of clusters.

Semi–supervised clustering is a framework for reconciling supervised learning, using any prior information ("labels") on the data, with unsupervised clustering, based on the intrinsic properties and geometry of the data set. A new method for semi-supervised clustering, combining probabilistic distance clustering for the unlabelled data points and a least squares criterion for the labelled ones, is given in Iyigun and Ben-Israel (2008d).

7. Conclusions

The probabilistic distance clustering algorithm presented here is simple, fast (requiring a small number of cheap iterations), robust (insensitive to outliers), and gives a high percentage of correct classifications.

It was tried on hundreds of problems with both simulated and real data sets. In simulated examples, where the answers are known, the algorithm, starting at random initial centers, always converged – in our experience – to the true cluster centers.

Results of our numerical experiments, and comparisons with other distance–based clustering algorithms, will be reported elsewhere.

References

ARAV, M. (2008), "Contour Approximation of Data and the Harmonic Mean", *Mathematical Inequalities & Applications*, to appear in Volume 11, 2008.

BEZDEK, J.C. (1973), "Fuzzy Mathematics in Pattern Classification", Ph.D. Thesis (Applied Mathematics), Cornell University, Ithaca, New York.

BEZDEK, J.C. (1981), *Pattern Recognition with Fuzzy Objective Function Algorithms*, New York: Plenum.

DIXON, K.R., and CHAPMAN J.A. (1980), "Harmonic Mean Measure of Animal Activity Areas", *Ecology 61*, 1040–1044.

HARTIGAN, J. (1975), *Clustering Algorithms*, New York: John Wiley & Sons, Inc.

HEISER, W.J. (2004), "Geometric Representation of Association Between Categories", *Psychometrika 69*, 513–545.

HÖPPNER, F., KLAWONN, F., KRUSE, R., and RUNKLER, T. (1999), *Fuzzy Cluster Analysis*, Chichester: John Wiley & Sons, Inc.

HUBERT, L., and STEINLEY, D. (2005), "Agreement Among Supreme Court Justices: Categorical vs. Continuous Representation", *SIAM News, 38(7).*

IYIGUN, C., and BEN-ISRAEL, A. (2008), "Probabilistic Distance Clustering, Theory and Applications", in *Clustering Challenges in Biological Networks*, eds. W. Chaovalitwongse and P.M. Pardalos, World Scientific, to appear.

IYIGUN, C., and BEN-ISRAEL, A. (2008a), "Probabilistic Distance Clustering Adjusted for Cluster Size", *Probability in the Engineering and Informational Sciences*, to appear.

IYIGUN, C., and BEN-ISRAEL, A. (2008b), "A Generalized Weiszfeld Method for Multi-facility Location Problems", to appear.

IYIGUN, C., and BEN-ISRAEL, A. (2008c), "A New Criterion for Clustering Validity via Contour Approximation of Data", to appear.

IYIGUN, C., and BEN-ISRAEL, A. (2008d), "Probabilistic Semi-Supervised Clustering", to appear.

JAIN, A.K., and DUBES, R.C. (1988), *Algorithms for Clustering Data*, New Jersey: Prentice Hall.

KUHN, H.W. (1973), "A Note on Fermat's Problem", *Mathematical Programming 4*, 98–107.

LOVE, R., MORRIS, J., and WESOLOWSKY, G. (1988), *Facilities Location: Models and Methods*, Amsterdam: North-Holland.

OSTRESH Jr., L.M. (1978), "On the Convergence of a Class of Iterative Methods for Solving the Weber Location Problem", *Operations Research 26*, 597–609.

TAN, P., STEINBACH, M., and KUMAR, V. (2006), *Introduction to Data Mining*, Boston: Addison Wesley.

TEBOULLE, M. (2007), "A Unified Continuous Optimization Framework for Center-Based Clustering Methods", *Journal of Machine Learning 8*, 65–102.

WEISZFELD, E. (1937), "Sur le point par lequel la somme des distances de n points donnés est minimum", *Tohoku Mathematical Journal 43*, 355–386.