CONTOUR APPROXIMATION OF DATA: A DUALITY THEORY

CEM IYIGUN AND ADI BEN-ISRAEL

Abstract. Given a dataset $\mathcal{D}$ partitioned in clusters, the joint distance function (JDF) $J(\mathbf{x})$ at any point $\mathbf{x}$ is the harmonic mean of the distances of $\mathbf{x}$ from the cluster centers. The JDF is a continuous function capturing the data points in its lower level sets (a property called contour approximation), and is a useful concept in probabilistic clustering and data analysis.

The JDF of the whole dataset, $J(\mathcal{D}) := \sum\{J(\mathbf{x}) : \mathbf{x} \in \mathcal{D}\}$, is a measure of the classifiability of $\mathcal{D}$, and can be used to determine the "right" number of clusters for $\mathcal{D}$.

A duality theory for the JDF of a dataset is given, in analogy with Kuhn's geometric duality theory for the Fermat–Weber location problem.

1. Introduction

We use the abbreviation

$$\overline{1,K} := \{1, 2, \ldots, K\} \tag{1}$$

for the indicated index set.

In $\mathbb{R}^n$ we denote the standard inner product by $\mathbf{x}\cdot\mathbf{y}$, and for a positive definite matrix Q define the **elliptic norm**,

$$\|\mathbf{u}\|_Q = (\mathbf{u}\cdot Q\mathbf{u})^{1/2} . \tag{2}$$

The **Euclidean norm**

$$\|\mathbf{u}\| = (\mathbf{u}\cdot\mathbf{u})^{1/2} \tag{3}$$

corresponds to $Q = I$, in which case we omit the subscript of the norm. We note the relation,

$$\|\mathbf{u}\|_Q = \|Q^{1/2}\mathbf{u}\| \ , \ \forall\ \mathbf{u} \in \mathbb{R}^n \ . \tag{4}$$

We take data points $\mathbf{x} = (x_1, \ldots, x_n)$ as vectors in $\mathbb{R}^n$, and consider a dataset consisting of N points, $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \mathbb{R}^n$. Let $\mathcal{D}$ be partitioned into K **clusters**

$$\mathcal{D} = \mathcal{C}_1 \cup \mathcal{C}_2 \cup \cdots \cup \mathcal{C}_K, \quad \mathcal{C}_i \cap \mathcal{C}_j = \emptyset \quad \text{if } i \neq j \ ,$$

and let the k^{th} cluster have a **center** $\mathbf{c}_k$ and be associated with a **distance function** $d_k(\mathbf{x}, \mathbf{c}_k)$, defined by

$$d_k(\mathbf{x}, \mathbf{y}) := \|\mathbf{x} - \mathbf{y}\|_{Q_k} \ , \tag{5}$$

where the positive definite matrix Q_k depends on the cluster, modelling its geometry. In particular, the **Mahalanobis distance** corresponds to $Q_k = \Sigma_k^{-1}$, where Σ_k is the covariance matrix of the data involved.

The distances are always measured from the cluster centers, so we can abbreviate $d_k(\mathbf{x}, \mathbf{c}_k)$ by $d_k(\mathbf{x})$.

Date: December 12, 2007.

Key words and phrases. Clustering, probabilistic clustering, duality, Mahalanobis distance, harmonic mean, joint distance function, Weiszfeld method.

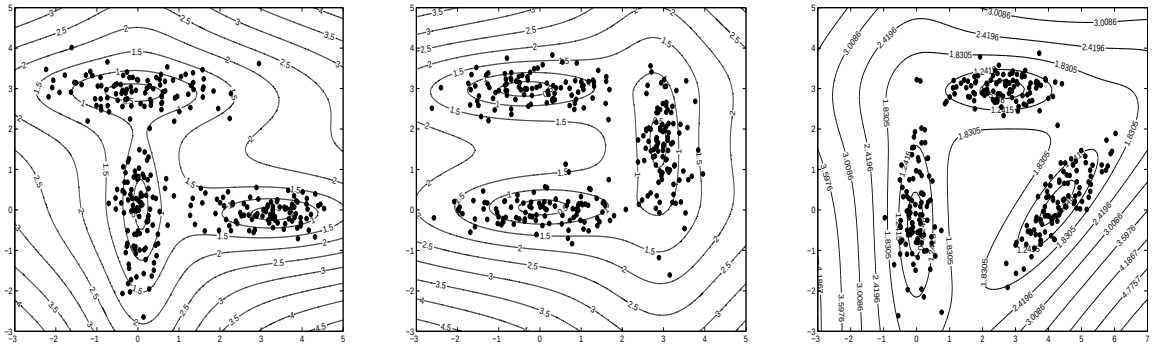

FIGURE 1. Contour approximation of data by the joint distance function

In probabilistic distance clustering, [2], we associate with each point $\mathbf{x}$ the cluster membership probabilities,

$$p_k(\mathbf{x}) = \text{Prob}\{\mathbf{x} \in \mathcal{C}_k\} \ , \ k \in \overline{1,K} \ , \tag{6}$$

that depend on the distances $d_k(\mathbf{x})$ from the clusters' centers. A simple relationship between probabilities and distances is

$$p_k(\mathbf{x})\, d_k(\mathbf{x}) = J(\mathbf{x}) \ , \ k \in \overline{1,K} \ , \tag{7}$$

where $J(\mathbf{x})$ is a function of $\mathbf{x}$, but does not depend on k. The meaning of (7) is that cluster membership is more probable the closer is the cluster center. In what follows we use (7) as our working principle.

Since the probabilities (6) add to one, (7) gives

$$J(\mathbf{x}) = \frac{\prod_{k=1}^{K} d_k(\mathbf{x})}{\sum_{k=1}^{K} \prod_{j \neq k} d_j(\mathbf{x})} \ , \tag{8}$$

which is (up to a constant) the harmonic mean of the K distances $d_k(\mathbf{x})$. We call $J(\mathbf{x})$ the **joint distance function** (or JDF for short) at $\mathbf{x}$. It has the dimension of distance, and is an indicator of the **classifiability** of the point $\mathbf{x}$, with $\mathbf{x}$ easier to classify the smaller is $J(\mathbf{x})$. In particular, $J(\mathbf{x}) = 0$ if and only if $\mathbf{x}$ coincides with one of the centers $\mathbf{c}_k$, in which case $p_k(\mathbf{x}) = 1$.

Since $J(\mathbf{x}) = \left(\sum_k p_k(\mathbf{x})\right) J(\mathbf{x}) = \sum_k p_k(\mathbf{x}) \left(p_k(\mathbf{x}) d_k(\mathbf{x})\right)$ it follows that

$$J(\mathbf{x}) = \sum_{k=1}^{K} p_k(\mathbf{x})^2\, d_k(\mathbf{x}) \ , \tag{9}$$

an alternative expression of the JDF.

The JDF $J(\mathbf{x})$ has an important approximation property: it captures the data points $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ in its lower level sets. This property, called **contour approximation**, was studied in [1] where the significance of the harmonic mean was elucidated. See Figure 1 for an illustration of contour approximation for datasets with 3 clusters in $\mathbb{R}^2$.

The JDF of the whole data set $\mathcal{D}$ is defined as the sum over all data points,

$$\begin{aligned} J(\mathcal{D}) &= \sum_{i=1}^{N} D(\mathbf{x}_i) \\ &= \sum_{k=1}^{K} \sum_{i=1}^{N} p_k(\mathbf{x}_i)^2 \, d_k(\mathbf{x}_i) \ , \ \text{by (9)} \ , \end{aligned} \tag{10}$$

measuring the uncertainty of classifying the dataset $\mathcal{D}$. This suggests the following formulation of clustering as a minimization problem,

$$\begin{aligned} \min & \sum_{k=1}^{K} \sum_{i=1}^{N} p_k(\mathbf{x}_i)^2 \, d_k(\mathbf{x}_i, \mathbf{c}_k) \\ \text{s.t.} & \sum_{k=1}^{K} p_k(\mathbf{x}_i) = 1 \ , \ i \in \overline{1,N} \ , \\ & p_k(\mathbf{x}_i) \geq 0 \ , \ k \in \overline{1,K} \ , i \in \overline{1,N} \ , \end{aligned} \tag{P}$$

called the **primal problem**.

The duality theory given here is an adaptation of the geometric duality developed by Kuhn [4] for the Fermat–Weber location problem. Section 2 is a schematic description of an iterative solution of (P). The problem encountered when one of the centers coincides with a data point is addressed in Section 3, where a modified gradient is constructed, and applied in Theorem 1 to characterize optimality. The dual problem (D) introduced in Section 4 is shown to satisfy weak duality, Theorem 2. Theorems 3–4 of Section 5 show that the dual pair {(P),(D)} have no duality gap.

2. Probabilistic distance clustering

In problem (P) there are two sets of variables, the **centers** $\{\mathbf{c}_1, \ldots, \mathbf{c}_K\}$ and the **probabilities** $\{p_k(\mathbf{x}_i) : k \in \overline{1,K} \ , i \in \overline{1,N}\}$. A natural approach is to fix one set of variables, and minimize (P) with respect to the other set, then fix the other set, etc. This is the idea of the **probabilistic distance clustering** method of [2] that computes the centers iteratively, using the following two steps:

Probabilities update. The centers $\{\mathbf{c}_1, \ldots, \mathbf{c}_K\}$ are given, and the distances $d_k(\mathbf{x}_i)$ are computed for all centers $\mathbf{c}_k$ and data points $\mathbf{x}_i$.

The minimizing probabilities are then

$$p_k(\mathbf{x}_i) = \frac{\prod\limits_{j \neq k} d_j(\mathbf{x}_i)}{\sum\limits_{m=1}^{K} \prod\limits_{j \neq m} d_j(\mathbf{x}_i)} \ , \ k \in \overline{1,K} \ , \tag{11}$$

called the probabilities **corresponding** to the centers $\{\mathbf{c}_1, \ldots, \mathbf{c}_K\}$.

Centers update. The probabilities $p_k(\mathbf{x}_i)$ and distances $d_k(\mathbf{x}_i)$ are given for all $k \in \overline{1,K}$, $i \in \overline{1,N}$. The centers $\{\mathbf{c}_1, \ldots, \mathbf{c}_K\}$ are then convex combinations of the data points,

$$\mathbf{c}_k = \sum_{i=1}^{N} \lambda_k(\mathbf{x}_i) \, \mathbf{x}_i \ , \tag{12}$$

where the weights $\lambda_k(\mathbf{x}_i)$ are given by

$$\lambda_k(\mathbf{x}_i) = \frac{p_k(\mathbf{x}_i)^2/d_k(\mathbf{x}_i)}{\sum_{m=1}^{K} p_m(\mathbf{x}_i)^2/d_m(\mathbf{x}_i)} \ , \ k \in \overline{1,K} \ , \ i \in \overline{1,N} \ . \tag{13}$$

Substituting (11) in (13) shows that the centers update can be done in terms of the distances $\{d_k(\mathbf{x}_i)\}$ alone, and that the probabilities $\{p_k(\mathbf{x}_i)\}$ are not explicitly needed.

The centers update (12)–(13) is a generalization, to several centers, of the well–known **Weiszfeld method** [9] for solving the Fermat–Weber location problem.

The method [2] also provides estimates of the matrices $\{Q_1, \ldots, Q_K\}$ modelling the geometry of the clusters, that are needed for the distances d_k.

3. The Modified Gradient

Fixing the probabilities $p_k(\mathbf{x}_i)$ in (P), the objective function is a function of the cluster centers, $\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K\}$,

$$f(\mathbf{c}_1, \ldots, \mathbf{c}_K) = \sum_{k=1}^{K} \sum_{i=1}^{N} p_k(\mathbf{x}_i)^2 \, d_k(\mathbf{x}_i, \mathbf{c}_k) \ , \tag{14}$$

where $d_k(\mathbf{x}_i, \mathbf{c}_k) = \|\mathbf{x}_i - \mathbf{c}_k\|_{Q_k}$, an elliptic distance defined by some positive definite matrix Q_k, associated with the k^{th} cluster.

The gradient of (14) with respect to $\mathbf{c}_k$, at a variable point $\mathbf{c}$,

$$\nabla_{\mathbf{c}_k} f(\mathbf{c}) = -Q_k \sum_{i=1}^{N} \frac{p_k(\mathbf{x}_i)^2}{d_k(\mathbf{x}_i, \mathbf{c}_k)} (\mathbf{x}_i - \mathbf{c}) \ , \tag{15}$$

is undefined (0/0) if $\mathbf{c}$ coincides with any of the data points $\mathbf{x}_i$. In this case we modify the gradient, following [4]–[5], and denote the modified gradient by $-\mathbf{R}_k$.

First copy (15), with a change of sign,

$$\mathbf{R}_k(\mathbf{c}) := Q_k \sum_{i=1}^{N} \frac{p_k(\mathbf{x}_i)^2}{d_k(\mathbf{x}_i, \mathbf{c}_k)} (\mathbf{x}_i - \mathbf{c}) \ , \quad \text{if } \mathbf{c} \neq \mathbf{x}_j, \ j \in \overline{1,N}, \tag{16a}$$

and otherwise define,

$$\mathbf{R}_k(\mathbf{x}_j) := \max\left(\|Q_k^{-1/2} \mathbf{R}_k^j\| - p_k(\mathbf{x}_j)^2, \, 0 \right) \frac{\mathbf{R}_k^j}{\|\mathbf{R}_k^j\|} \ , \ j \in \overline{1,N} \ , \tag{16b}$$

where

$$\mathbf{R}_k^j = Q_k \sum_{i \neq j} \frac{p_k(\mathbf{x}_i)^2}{d_k(\mathbf{x}_i, \mathbf{x}_j)} (\mathbf{x}_i - \mathbf{x}_j) \ . \tag{16c}$$

In (16b) the length of $Q_k^{-1/2} \, \mathbf{R}_k^j$ is compared with $p_k(\mathbf{x}_j)^2$: if $\|Q_k^{-1/2} \, \mathbf{R}_k^j\| < p_k(\mathbf{x}_j)^2$ then $\mathbf{R}_k(\mathbf{x}_j) = \mathbf{0}$; otherwise, $\mathbf{R}_k(\mathbf{x}_j)$ is a vector with magnitude $\|Q_k^{-1/2} \, \mathbf{R}_k^j\| - p_k(\mathbf{x}_j)^2$ and direction $\mathbf{R}_k^j$.

Next, a characterization of optimality in terms of the modified gradient.

Theorem 1. Given the data $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, let $\{\mathbf{c}_1, \ldots, \mathbf{c}_K\}$ be any K points, and let $\{p_k(\mathbf{x}_i) :\ k \in \overline{1,K},\ i \in \overline{1,N}\}$ be the probabilities corresponding by (11). Then the points $\{\mathbf{c}_1, \ldots, \mathbf{c}_K\}$ minimize the function f of (14) if and only if $\mathbf{R}_k(\mathbf{c}_k) = \mathbf{0}$, for $k \in \overline{1,K}$.

Proof. If $\mathbf{c}_k$ is not one of the data points, then $\mathbf{R}_k(\mathbf{c}_k)$ is the gradient (15) at $\mathbf{c}_k$, and by the convexity of f, $\mathbf{R}_k(\mathbf{c}_k) = \mathbf{0}$ is both necessary and sufficient for a minimum.

If an optimal $\mathbf{c}_k$ coincides with a data point $\mathbf{x}_j$, then $\mathbf{x}_j$ belongs with certainty to the k^{th} cluster and, by (11),

$$p_k(\mathbf{x}_j) = 1\ ,\ p_k(\mathbf{x}_i) = 0 \text{ for all } i \neq j\ . \tag{17}$$

If $\mathbf{R}_k(\mathbf{x}_j) \neq \mathbf{0}$ then

$$\begin{aligned} p_k(\mathbf{x}_j)^2 &< \|Q_k^{-1/2}\mathbf{R}_k^j\|\ ,\ \text{by (16b)}\ , \\ &= \|Q_k^{1/2}\sum_{i\neq j}\frac{p_k(\mathbf{x}_i)^2}{d_k(\mathbf{x}_i,\mathbf{x}_j)}(\mathbf{x}_i - \mathbf{x}_j)\|\ ,\ \text{by (16c)}\ , \\ &\leq \sum_{i\neq j} p_k(\mathbf{x}_i)^2\ ,\ \text{since } \|Q_k^{1/2}(\mathbf{x}_i - \mathbf{x}_j)\| = d_k(\mathbf{x}_i,\mathbf{x}_j)\ , \end{aligned}$$

contradicting (17). This proves that $\mathbf{R}_k(\mathbf{c}_k) = \mathbf{0}$ is necessary for optimality.

To prove sufficiency, consider the change from $\mathbf{x}_j$ to $\mathbf{x}_j + t\,\mathbf{z}$ where $\|\mathbf{z}\|_{Q_k} = 1$. Then,

$$\tfrac{d}{dt} f(\mathbf{c}_1, \ldots, \mathbf{c}_{k-1}, \mathbf{x}_j + t\,\mathbf{z}, \mathbf{c}_{k+1} \ldots, \mathbf{c}_K)\big|_{t=0} = p_k(\mathbf{x}_j)^2 - \mathbf{R}_k^j \cdot \mathbf{z}\ . \tag{18}$$

The greatest decrease of (18) is when $\mathbf{z}$ is along $\mathbf{R}_k^j$, i.e., when

$$\mathbf{z} = \frac{\mathbf{R}_k^j}{\|\mathbf{R}_k^j\|_{Q_k}}\ ,\ \text{since } \|\mathbf{z}\|_{Q_k} = 1.$$

Therefore $\mathbf{x}_j$ is a local minimum if and only if,

$$p_k(\mathbf{x}_j)^2 - \frac{\mathbf{R}_k^j \cdot \mathbf{R}_k^j}{\|\mathbf{R}_k^j\|_{Q_k}} \geq 0\ , \quad \text{or} \quad \frac{(Q_k^{1/2}\mathbf{R}_k^j)\cdot(Q_k^{-1/2}\mathbf{R}_k^j)}{\|Q_k^{1/2}\mathbf{R}_k^j\|} \leq p_k(\mathbf{x}_j)^2\ ,$$

which by the Cauchy–Schwartz inequality follows from

$$\|Q_k^{-1/2}\mathbf{R}_k^j\| = \frac{\|Q_k^{1/2}\mathbf{R}_k^j\|\|Q_k^{-1/2}\mathbf{R}_k^j\|}{\|Q_k^{1/2}\mathbf{R}_k^j\|} \leq p_k(\mathbf{x}_j)^2$$

proving by (16b) that $\mathbf{R}_k(\mathbf{c}_k) = \mathbf{0}$ is also sufficient for optimality. □

4. The Dual Problem

In problem (P) let the probabilities $p_k(\mathbf{x}_i)$ be fixed. We abbreviate $p_k(\mathbf{x}_i)$ by p_{ki}, for $k \in \overline{1,K}$, $i \in \overline{1,N}$.

A dual problem (D) for (P) is now given. It uses the data

$$\mathcal{S} := \{\mathbf{x}_i :\ i \in \overline{1,N}\} \bigcup \{p_{ki} :\ k \in \overline{1,K},\ i \in \overline{1,N}\} \bigcup \{Q_k :\ k \in \overline{1,K}\} \tag{19}$$

consisting of the original data points $\{\mathbf{x}_i\}$, and (computed or assumed) values for the probabilities $\{p_{ki}\}$ and the matrices $\{Q_k\}$ used in the elliptic distances. The dual variables are $K\,N$ vectors $\{\mathbf{u}_{ki} :\ k \in \overline{1,K},\ i \in \overline{1,N}\}$, one for each pair {cluster, data point}. We denote by $\mathbf{U}$ the set of dual variables.

The **dual problem** is:

$$\max \quad g(\mathbf{U}) = \sum_{k=1}^{K} \sum_{i=1}^{N} \mathbf{u}_{ki} \cdot \mathbf{x}_i \tag{D}$$

$$\text{s.t.} \quad \sum_{i=1}^{N} \mathbf{u}_{ki} = \mathbf{0} \ , \ k \in \overline{1,K} \ , \tag{20}$$

$$\|Q_k^{-1/2} \mathbf{u}_{ki}\| \le p_{ki}^2 \ , \ i \in \overline{1,N}, \ k \in \overline{1,K} \ . \tag{21}$$

Variables $\mathbf{U} = \{\mathbf{u}_{ki}\}$ satisfying (20)–(21) are called **feasible**.

Theorem 2. Let the data $\mathcal{S}$ in (19) be given. Then for any set of centers $\{\mathbf{c}_1, \dots, \mathbf{c}_K\}$, and any set of feasible dual variable $\mathbf{U} = \{\mathbf{u}_{ki}\}$,

$$g(\mathbf{U}) \ \le \ f(\mathbf{c}_1, \dots, \mathbf{c}_K). \tag{22}$$

Proof. The objective function $g(\mathbf{U})$ can be written, using (20), as

$$g(\mathbf{U}) = \sum_{k=1}^{K} \left(\sum_{i=1}^{N} \mathbf{u}_{ki} \cdot \mathbf{x}_i - \left(\sum_{i=1}^{N} \mathbf{u}_{ki} \right) \cdot \mathbf{c}_k \right) = \sum_{k=1}^{K} \sum_{i=1}^{N} \mathbf{u}_{ki} \cdot (\mathbf{x}_i - \mathbf{c}_k). \tag{23}$$

Setting

$$\mathbf{u}_{ki} \cdot (\mathbf{x}_i - \mathbf{c}_k) = \left(Q_k^{-1/2} \mathbf{u}_{ki} \right) \cdot \left(Q_k^{1/2} (\mathbf{x}_i - \mathbf{c}_k) \right) ,$$

we get from the Cauchy–Schwartz inequality,

$$\begin{aligned} |\mathbf{u}_{ki} \cdot (\mathbf{x}_i - \mathbf{c}_k)| &= \left| Q_k^{-1/2} \mathbf{u}_{ki} \cdot Q_k^{1/2} (\mathbf{x}_i - \mathbf{c}_k) \right| \\ &\le \|Q_k^{-1/2} \mathbf{u}_{ki}\| \|\mathbf{x}_i - \mathbf{c}_k\|_{Q_k}. \end{aligned}$$

$$\therefore \ g(\mathbf{U}) = \sum_{k=1}^{K} \sum_{i=1}^{N} \mathbf{u}_{ki} \cdot (\mathbf{x}_i - \mathbf{c}_k) \le \sum_{k=1}^{K} \sum_{i=1}^{N} \|Q_k^{-1/2} \mathbf{u}_{ki}\| \|\mathbf{x}_i - \mathbf{c}_k\|_{Q_k} \tag{24a}$$

$$\le \sum_{k=1}^{K} \sum_{i=1}^{N} p_{ki}^2 \, d_k(\mathbf{x}_i, \, \mathbf{c}_k) \ , \ \text{by (21)}, \tag{24b}$$

$$= f(\mathbf{c}_1, \dots, \mathbf{c}_K) \ .$$

□

5. Strong Duality

Theorem 2 is a **weak duality** theorem in the sense that any feasible solution $\mathbf{U}$ of (D) gives a lower bound for the optimal value of (P), and conversely, any set of centers $\{\mathbf{c}_k\}$ for (P) gives an upper bound on the optimal value of (D). The next two theorems show that there is no duality gap between (P) and (D).

Theorem 3. Given the data $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, and an optimal solution

$$\{\mathbf{c}_1, \dots, \mathbf{c}_K\}, \ \{p_{ki} : i \in \overline{1,N}, \ k \in \overline{1,K}\} \ ,$$

of the primal problem (P), there exist feasible dual variables $\mathbf{U}$ such that

$$g(\mathbf{U}) = f(\mathbf{c}_1, \dots, \mathbf{c}_K). \tag{25}$$

Proof. We distinguish two cases.

Case 1. None of the centers $\{\mathbf{c}_k\}$ coincides with any of the data points $\{\mathbf{x}_i\}$.

For $k \in \overline{1,K}$ and $i \in \overline{1,N}$ define

$$\mathbf{u}_{ki} := \frac{p_{ki}^2}{d_k(\mathbf{x}_i,\, \mathbf{c}_k)} Q_k \,(\mathbf{x}_i - \mathbf{c}_k) \;. \tag{26}$$

Then from (16a) and Theorem 1 it follows that,

$$\sum_{i=1}^{N} \mathbf{u}_{ki} = \mathbf{R}_k(\mathbf{c}_k) = \mathbf{0}, \text{ verifying (20) .}$$

Rewriting (26) as

$$Q_k^{-1/2}\, \mathbf{u}_{ki} = \frac{p_{ki}^2}{d_k(\mathbf{x}_i,\, \mathbf{c}_k)}\, Q_k^{1/2}\, (\mathbf{x}_i - \mathbf{c}_k),$$

we get, for all k, i,

$$\|Q_k^{-1/2}\mathbf{u}_{ki}\| = \frac{p_{ki}^2}{d_k(\mathbf{x}_i,\, \mathbf{c}_k)} \|\mathbf{x}_i - \mathbf{c}_k\|_{Q_k} = p_{ki}^2 \;, \tag{27}$$

proving (21) as equations. Therefore the $\{\mathbf{u}_{ki}\}$ defined by (26) are feasible.

From (26) and (4)–(5) it follows that

$$\mathbf{u}_{ki} \cdot (\mathbf{x}_i - \mathbf{c}_k) = p_{ki}^2\, d_k(\mathbf{x}_i, \mathbf{c}_k) \tag{28}$$

and (25) follows then from (23).

Case 2. A center coincides with one of the data points, say

$$\mathbf{c}_k = \mathbf{x}_j \;, \tag{29}$$

for some $k \in \overline{1,K}$, $j \in \overline{1,N}$, and define

$$\mathbf{u}_{ki} := \frac{p_{ki}^2}{d_k(\mathbf{x}_i,\, \mathbf{x}_j)} Q_k \,(\mathbf{x}_i - \mathbf{x}_j) \;, \quad \text{for } i \neq j \;, \tag{30a}$$

$$\mathbf{u}_{kj} := -\sum_{i \neq j} \mathbf{u}_{ki} \;. \tag{30b}$$

Then $\sum_i \mathbf{u}_{ki} = \mathbf{0}$ by definition, and $\|Q_k^{-1/2}\, \mathbf{u}_{ki}\| = p_{ki}^2$ for all $i \neq j$, as in (27). Next,

$$\mathbf{u}_{kj} = -\mathbf{R}_k^j \text{ by (16c), and therefore by (16b),}$$

$$\mathbf{R}_k(\mathbf{x}_j) = \mathbf{0} \quad \text{implies} \quad p_{kj}^2 \geq \|Q^{-1/2}\mathbf{R}_k^j\| = \|Q^{-1/2}\mathbf{u}_{kj}\| \;,$$

proving that the variables $\mathbf{U}$ defined by (30a)–(30b) are feasible.

Finally we prove (25). As in Case 1 we have the equality (28) for all $i \neq j$. Consider now the inequalities in (24a)–(24b) corresponding to $i = j$,

$$\mathbf{u}_{kj} \cdot (\mathbf{x}_j - \mathbf{c}_k) \;\leq\; \|Q_k^{-1/2}\, \mathbf{u}_{kj}\| \|\mathbf{x}_j - \mathbf{c}_k\|_{Q_k} \;\leq\; p_{kj}^2\, d_k(\mathbf{x}_j,\, \mathbf{c}_k)$$

These become trivial equalities, since by (29) all three terms are zero. Similarly,

$$0 = \mathbf{u}_{mj} \cdot (\mathbf{x}_j - \mathbf{c}_m) \;\leq\; \|Q_m^{-1/2}\, \mathbf{u}_{mj}\| \|\mathbf{x}_j - \mathbf{c}_m\|_{Q_m} \;\leq\; p_{mj}^2\, d_m(\mathbf{x}_j,\, \mathbf{c}_m) = 0$$

for any other cluster $m \in \overline{1,K}$, $m \neq k$, since $p_{mj} = 0$ (by (17), which follows from (29)), and therefore $\mathbf{u}_{mj} = \mathbf{0}$.

The two inequalities in (24a)–(24b) therefore hold as equalities, proving (25). □

The next theorem is the converse of Theorem 3.

Theorem 4. Let $\mathbf{U}$ be an optimal solution of the dual problem (D). Then there exist $\{\mathbf{c}_1, \ldots, \mathbf{c}_K\}$ such that

$$g(\mathbf{U}) = f(\mathbf{c}_1, \ldots, \mathbf{c}_K) \; . \tag{25}$$

Proof. Writing the objective function $g(\mathbf{U})$ as in (23), the Lagrangian of (D) is

$$\sum_{k=1}^{K} \sum_{i=1}^{N} \mathbf{u}_{ki} \cdot (\mathbf{x}_i - \mathbf{c}_k) - \sum_{k=1}^{K} \sum_{i=1}^{N} t_{ki} \left(\|Q_k^{-1/2} \mathbf{u}_{ki}\| - p_{ki}^2 \right)$$

with Lagrange multipliers $\{t_{ki}\}$. The Karush–Kuhn–Tucker necessary conditions for optimality are

$$(\mathbf{x}_i - \mathbf{c}_k) - t_{ki} \, Q_k^{-1} \frac{\mathbf{u}_{ki}}{\|Q_k^{-1/2} \mathbf{u}_{ki}\|} = \mathbf{0} \; , \tag{31a}$$

$$\sum_{i=1}^{N} \mathbf{u}_{ki} = \mathbf{0} \; , \tag{31b}$$

$$\|Q_k^{-1/2} \mathbf{u}_{ki}\| \le p_{ki}^2 \; , \tag{31c}$$

$$t_{ki} \ge 0 \; , \tag{31d}$$

$$t_{ki} (\|Q_k^{-1/2} \mathbf{u}_{ki}\| - p_{ki}^2) = 0 \; , \tag{31e}$$

for all $k \in \overline{1,K}$, $i \in \overline{1,N}$. Again we distinguish two cases.

Case 1. All $t_{ki} > 0$. Then (27) follows from (31e) for all k, i, and from (31a),

$$Q_k^{1/2} (\mathbf{x}_i - \mathbf{c}_k) = \; t_{ki} \, Q_k^{-1/2} \frac{\mathbf{u}_{ki}}{\|Q_k^{-1/2} \mathbf{u}_{ki}\|} \; .$$

Taking norms on both sides gives

$$d_k(\mathbf{x}_i, \, \mathbf{c}_k) = t_{ki} \; , \tag{32}$$

and by substituting (27) and (32) in (31a),

$$\mathbf{u}_{ki} := \frac{p_{ki}^2}{d_k(\mathbf{x}_i, \, \mathbf{c}_k)} Q_k \left(\mathbf{x}_i - \mathbf{c}_k \right) \; ,$$

and the equality (25) follows as in the proof of Theorem 3, Case 1.

Case 2. Some Lagrange multipliers are zero, say $t_{kj} = 0$. Then by (31a), $\mathbf{c}_k = \mathbf{x}_j$, and $t_{ki} > 0$ for $i \ne j$, by (31a) and (31d), and therefore, by (31e),

$$\|Q_k^{-1/2} \mathbf{u}_{ki}\| = \; p_{ki}^2 \; , \quad \text{for all} \quad i \ne j \; .$$

From $\mathbf{c}_k = \mathbf{x}_j$ and (31a) it follows that

$$Q_k^{1/2} (\mathbf{x}_i - \mathbf{x}_j) = \; t_{ki} \, Q_k^{-1/2} \frac{\mathbf{u}_{ki}}{\|Q_k^{-1/2} \mathbf{u}_{ki}\|}$$

and by taking norms,

$$d_k(\mathbf{x}_i, \, \mathbf{x}_j) = t_{ki} \; , \quad \text{for all} \quad i \ne j \; .$$

Substituting t_{ki} and $\|Q_k^{-1/2} \mathbf{u}_{ki}\|$ in (31a) gives,

$$\mathbf{u}_{ki} := \frac{p_{ki}^2}{d_k(\mathbf{x}_i, \, \mathbf{x}_j)} Q_k \left(\mathbf{x}_i - \mathbf{x}_j \right) \; , \text{ for } i \ne j \; ,$$

and from (31b),

$$\mathbf{u}_{kj} := -\sum_{i \neq j} \mathbf{u}_{ki} \ .$$

These are the same as (30a)–(30b), and equality in (25) follows as in the proof of Theorem 3, Case 2. □

Notes.

(a) The probabilistic distance clustering method of [2] was adjusted in [3] to account for the cluster sizes. The duality theory given here can be adapted to problems where the cluster sizes are constrained, or are variables to be estimated.

(b) The seminal paper [7] gives a unified optimization framework for clustering problems, and a duality theory more general than the one attempted here.

(c) The practical applicability of Theorems 2–4 above is limited by the fact that the matrices $\{Q_1, \ldots, Q_K\}$ modelling the geometry of the clusters are not known a priori, and may require a solution of the primal problem (P). However, useful bounds on the optimal value of (P) can be found by taking Euclidean distances, i.e., approximating the matrices $\{Q_k\}$ by the identity matrix.

(d) For other results on duality in multi–facility location problems see [6], [8] and their references.

References

[1] M. Arav, Contour approximation of data and the harmonic mean, *Mathematical Inequalities & Applications* (to appear)

[2] A. Ben–Israel and C. Iyigun, Probabilistic distance clustering, *J. Classification* (to appear)

[3] C. Iyigun and A. Ben–Israel, Probabilistic distance clustering adjusted for cluster size, *Probability Engrg. Info. Sci.* (to appear)

[4] H. W. Kuhn, On a pair of dual nonlinear programs, in J. Abadie (ed.), *Methods of Nonlinear Programming*, Amsterdam, North-Holland, (1967), 38–54

[5] H. W. Kuhn, A note on Fermat's problem, *Mathematical Programming* **4**(1973), 98–107

[6] R. F. Love and H. Juel, Properties and solution methods for large location–allocation problems, *The Journal of the Operational Research Society* **33**(1982), 443–452

[7] M. Teboulle, A unified continuous optimization framework for center–based clustering methods, *J. Machine Learning* **8**(2007), 65–102

[8] H. Üster and R. F. Love, Duality in constrained multi–facility location models, *Naval Research Logistics* **49**(2002), 410–421

[9] E. Weiszfeld, Sur le point par lequel la somme des distances de n points donnés est minimum, *Tohoku Math. J.* **43** (1937), 355–386

Cem Iyigun
E-mail address: iyigun@rutcor.rutgers.edu

Adi Ben-Israel
E-mail address: adi.benisrael@gmail.com

RUTCOR–Rutgers Center for Operations Research, Rutgers University, 640 Bartholomew Rd., Piscataway, NJ 08854-8003, USA