DIMACS Technical Report 2006-20
December 2006

Contour approximation of data and the harmonic mean

by

Marina Arav
Department of Mathematics and Statistics
Georgia State University
30 Pryor St., Atlanta, GA 30303-3083
email: matmxa@langate.gsu.edu

ABSTRACT

A contour approximation of data is a function capturing the data points in its lower level sets. Desirable properties of contour approximation are posited, and shown to be satisfied uniquely (up to a multiplicative constant) by the weighted harmonic mean of distances to the cluster centers. This harmonic mean is the joint distance function used in probabilistic clustering, expressing the uncertainty of classification.

Keywords: Harmonic mean, quasi-linear mean, clustering, contour approximation of data, distance functions.

1 Introduction

Consider a **data set** $\mathcal{D}$ consisting of N data points $\{\mathbf{x}_i : i \in \overline{1,N}\} \subset \mathbb{R}^n$. We assume that $\mathcal{D}$ is partitioned into K clusters, $1 < K < N$, each consisting of points close to each other. We assume the clusters given (in practice they need be computed.)

For example, the data set in Figs. 1(a) consists of 200 points, in two equal clusters each sampled from a bi–normal distribution. The data set in Fig. 2(a), has 1100 data points, in two unequal clusters. The large cluster, with 1000 points, is simulated from a spherical distribution. The small cluster is a sample from a bi–normal distribution.

With each cluster $\mathcal{C}_k, k \in \overline{1,K}$, we associate a **center** $\mathbf{c}_k$, and a **distance function** $d_k(\cdot\,,\cdot)$. The distances used include

$$d_k\,(\mathbf{x},\mathbf{y}) = <\mathbf{x}-\mathbf{y},\Sigma_k^{-1}(\mathbf{x}-\mathbf{y})>^{1/2},\ \text{(Mahalonobis distance)}, \tag{1}$$

with Σ_k = the covariance matrix of $\mathcal{C}_k$ (assumed positive definite), in particular,

$$d_k\,(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|,\ \text{(Euclidean distance)}. \tag{2}$$

The **cluster sizes** q_k are assumed known.

The purpose of this paper is to study a new kind of data approximation suggested by the **joint distance function** (JDF) used in probabilistic clustering, [5, 8]. The JDF, denoted $D(\mathbf{x})$, is (up to a scalar) the weighted harmonic mean of the distances $d_k(\mathbf{x},\mathbf{c}_k)$ from all the cluster centers $\mathbf{c}_k$,

$$D(\mathbf{x}) = \frac{\displaystyle\prod_{j=1}^{K} \frac{d_j(\mathbf{x},\mathbf{c}_j)}{q_j}}{\displaystyle\sum_{i=1}^{K}\prod_{j\neq i}\frac{d_j(\mathbf{x},\mathbf{c}_j)}{q_j}}\ , \tag{3}$$

in particular,

$$D(\mathbf{x}) = \frac{d_1(\mathbf{x},\mathbf{c}_1)\,d_2(\mathbf{x},\mathbf{c}_2)}{q_2\,d_1(\mathbf{x},\mathbf{c}_1)+q_1\,d_2(\mathbf{x},\mathbf{c}_2)}\ ,\ \text{for } K=2\ . \tag{4}$$

The JDF approximates the data in a sense illustrated in Fig. 1(b) and Fig. 2(b). This suggests the following definition.

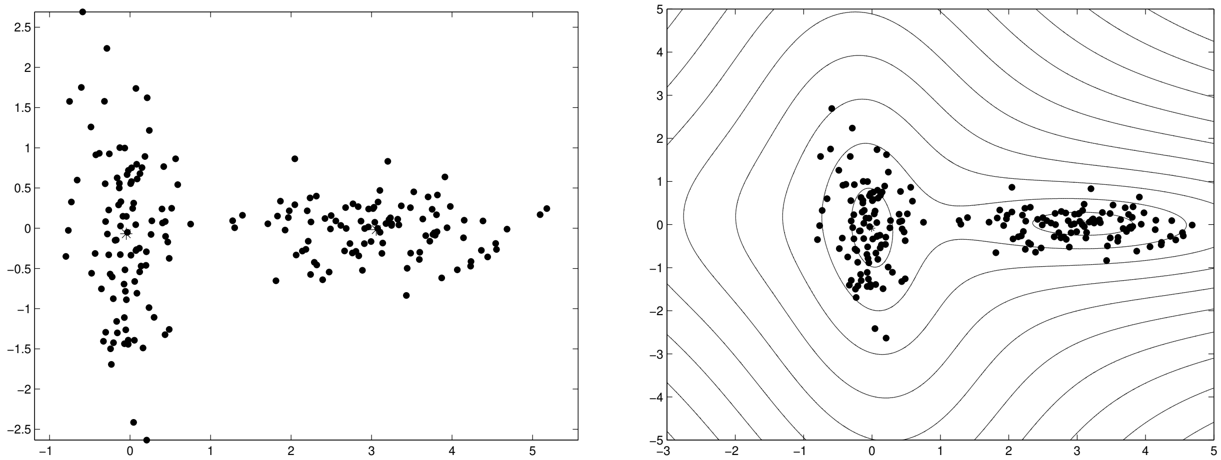
Definition 1. A function $D : \mathbb{R}^{2K} \to \mathbb{R}_+$, of the distances $d_k(\mathbf{x},\mathbf{c}_k)$ and cluster sizes q_k, say

$$D(\mathbf{x}) = F(d_1(\mathbf{x},\mathbf{c}_1),\cdots,d_K(\mathbf{x},\mathbf{c}_K)\,;\,q_1,\cdots,q_K)\ , \tag{5}$$

is called a **contour approximation** of the data $\mathcal{D}$ if for all $\mathbf{x} \in \mathbb{R}^n$ and $k \in \overline{1,K}$,

$$D(\mathbf{x}) \leq d_k(\mathbf{x},\mathbf{c}_k)\ , \tag{6}$$

i.e., D captures the data in its lower level–sets.



(a) A data set of 200 points in $\mathbb{R}^2$ (b) Level sets of the JDF

Figure 1: Contour approximation of data, two equal clusters

Desirable properties of contour approximation are discussed in § 3. We see that $D(\mathbf{x})$ measures the uncertainty of classification, with low values of $D(\mathbf{x})$ indicating that it is easier to classify $\mathbf{x}$.

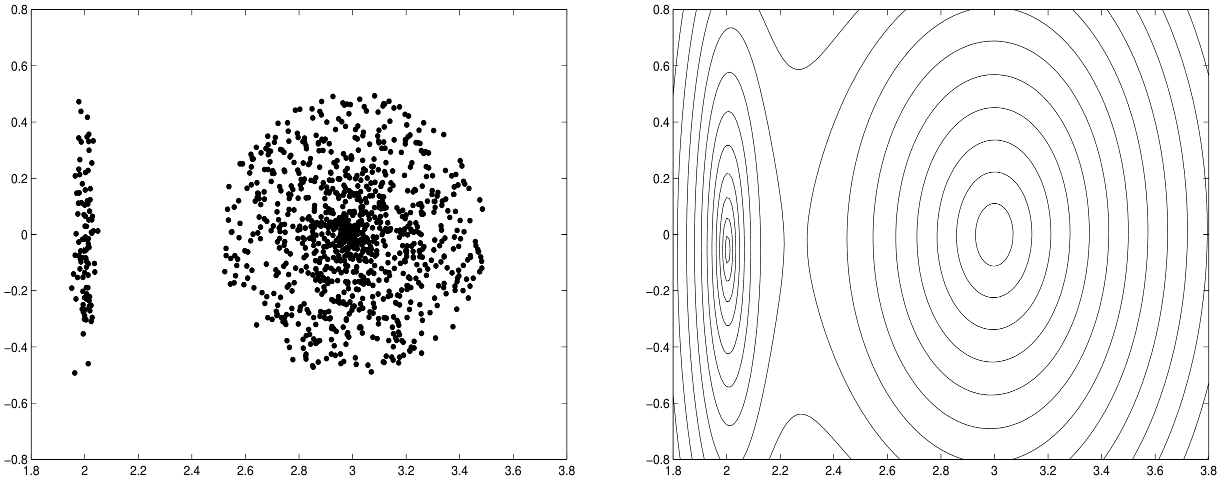
Section 2 is a detour, reviewing the **quasi–linear mean**, needed in the sequel. The main result is Theorem 1 in § 3, proving that the contour approximation of data is, up to a constant, the weighted harmonic mean of the distances to the cluster centers, with the cluster sizes as weights. This establishes the uniqueness of the JDF (3) as the contour approximation of interest.

2 The quasi–linear mean

Let f be a continuous function, mapping an interval $I = [a, b]$ into itself, and strictly monotonic on I. Let $r, s \geq 0;\ r + s > 0$. The **quasi–linear mean** of the numbers $x, y \in I$ is

$$F(x, y; r, s) = f\left(\frac{r f^{-1}(x) + s f^{-1}(y)}{r + s}\right) \tag{7}$$

see [1, § 5.3.2], and references therein. The numbers x, y are called the "variables" of (7), and r, s its "weights". The quasi–linear mean was characterized in [1, p. 242] by the following



(a) 1100 data points in two unequal clusters (b) Level sets of the JDF

Figure 2: Contour approximation of data, unequal clusters

properties, required for all $a \le x, y \le b$ and $r, s \ge 0$ with $r + s > 0$.

$$F(x, x;\, r, s) = x \quad \text{(reflexivity)}\ . \tag{8a}$$

$$a = F(a, b\,;\, 1, 0) < F(a, b\,;\, r, s) < F(a, b\,;\, 0, 1) = b\ ,\ \forall\ r, s > 0\ . \tag{8b}$$

$$F(a, b\,;\, rt, st) = F(a, b\,;\, r, s)\ ,\ \forall\ t > 0\ . \tag{8c}$$

$$\begin{aligned} &F(F(x, y;\, r, s), F(X, Y;\, R, S)\,;\, r + s, R + S) = \\ &\quad = F(F(x, X;\, r, R), F(y, Y;\, s, S)\,;\, r + R, s + S)\ ,\ \forall\ a \le X, Y \le b\,, R, S \ge 0\ , \end{aligned} \tag{8d}$$

$$F(a, b\,;\, r, s) < F(a, b\,;\, r, t)\ ,\ \forall\ s < t\ , \tag{8e}$$

$$F(x, y\,;\, r, s) < F(x, z\,;\, r, s)\ ,\ \forall\ y < z\ . \tag{8f}$$

The generator f can be expressed explicitly by F,

$$f(t) = F(a, b\,;\, 1 - t, t)\ ,\ \forall\ 0 \le t \le 1\ . \tag{9}$$

Remarks 1.
(a) From definition (7) it follows that F is continuous, and satisfies

$$F(x, y;\, r, s) = F(y, x;\, s, r) \quad \text{(symmetry)}\ , \tag{10a}$$

$$F(F(x, y;\, r, s)\,;\, r + s, t) = F(x, F(y, z\,;\, s, t)\,;\, r, s + t) \quad \text{(associativity)}\ , \tag{10b}$$

which imply (8d).
(b) The common value in (10b) is the quasi–linear mean of the three variables x, y, z,

$$F(x, y, z\,;\, r, s, t) = f\left(\frac{rf^{-1}(x) + sf^{-1}(y) + tf^{-1}(z)}{r + s + t}\right) \tag{11}$$

with weights r, s, t. The definition (7) is analogously extended to more than 3 variables.

3 A contour approximation of data

Let $\mathcal{D}$ be a given data set in $\mathbb{R}^n$ with 2 clusters, $\mathcal{C}_1$ and $\mathcal{C}_2$. The i^{th}–cluster is of **size** q_i, has **center** $\mathbf{c}_i$ and is endowed with a **distance** function $d_i(\cdot\,,\cdot)$, which may depend on the cluster (in particular, the Mahalanobis distance (1) depends on the cluster covariance.)

The centers $\mathbf{c}_1, \mathbf{c}_2$ are assumed distinct.

We list some desirable properties of a **contour approximation** of $\mathcal{D}$, i.e., a function $D(\mathbf{x})$ satisfying (6). In order to relate our results to the quasi–linear means of § 2, we express $D(\mathbf{x})$ as

$$D(\mathbf{x}) = \frac{F(d_1(\mathbf{x},\mathbf{c}_1), d_2(\mathbf{x},\mathbf{c}_2)\,;\,q_1,q_2)}{q_1+q_2}\,, \tag{12}$$

and study the function $F(d_1, d_2\,;\,q_1,q_2)$ in the numerator, identifying it with the quasi–linear mean $F(x, y\,;\,r, s)$ of § 2, see Remark 3(e). It suffices to study the case of two clusters, as the function, and its properties, extend to any number of clusters, see Remarks 3(c)–(d) below.

The desirable properties of F include (8a)–(8f), with $x = d_1, y = d_2, r = q_1$ and $s = q_2$. In addition we require that F is differentiable for $d_1, d_2 \geq 0$, and satisfies for all $d_1, d_2 \geq 0$, $q_1, q_2 \geq 0$ such that $q_1 + q_2 > 0$:

$$F(d_1, d_2\,;\,q_1,q_2) = 0 \text{ if, and only if, } d_1 = 0 \text{ or } d_2 = 0\,, \tag{13a}$$

$$F(\lambda\,d_1, \lambda\,d_2\,;\,q_1,q_2) = \lambda\,F(d_1, d_2\,;\,q_1,q_2)\,,\ \forall\,\lambda > 0\,, \tag{13b}$$

$$F_1'(0, d_2\,;\,q_1,q_2) = \frac{q_1+q_2}{q_1}\,,\ \forall\,q_1 > 0\,, \tag{13c}$$

$$F_2'(d_1, 0\,;\,q_1,q_2) = \frac{q_1+q_2}{q_2}\,,\ \forall\,q_2 > 0\,, \tag{13d}$$

where F_i' denotes the right derivative with respect to the i^{th} place, $i = 1, 2$.

Remarks 2.
(a) The argument $d_i = d_i(\mathbf{x},\mathbf{c}_i)$ vanishes if, and only if, $\mathbf{x} = \mathbf{c}_i$. Since the cluster centers are distinct, both arguments d_1, d_2 cannot vanish. Moreover, $d_1 = 0$ determines a unique $d_2 = d_2(\mathbf{c}_2, \mathbf{c}_1)$, and vice versa.
(b) A cluster center belongs unambiguously to its cluster. Since F measures the uncertainty of classification, it is reasonable to assume (13a).
(c) If d_1 and d_2 have a (physical) dimension of distance (say both are measured in meters), the homogeneity in (13b) means that $F(d_1, d_2\,;\,q_1,q_2)$ has the same dimension, i.e. it is a distance.
(d) The right derivatives are needed in (13c)–(13d) because F is defined only for $d_1, d_2 \geq 0$.
(e) The function F vanishes at the cluster centers, but its behavior near the centers depends on the cluster size: if $q_1 > q_2$ and $\alpha > 0$, the level set $\{\mathbf{x} : F(d_1(\mathbf{x},\mathbf{c}_1), d_2(\mathbf{x},\mathbf{c}_2)\,;\,q_1,q_2) \leq \alpha\}$ should include more points of the larger cluster, in other words, the graph of F is flatter near $d_1 = 0$ then near $d_2 = 0$. This explains the right hand sides of (13c)–(13d).

Theorem 1. Properties (8a)–(8f) and (13a)–(13d) characterize the function

$$F(d_1, d_2\,;\,q_1,q_2) = \frac{(q_1+q_2)\,d_1 d_2}{q_2 d_1 + q_1 d_2} \tag{14}$$

which, if $d_1, d_2 > 0$, is the weighted harmonic mean of d_1 and d_2,

$$F(d_1, d_2\,;\, q_1, q_2) = \frac{q_1 + q_2}{\dfrac{1}{d_1/q_1} + \dfrac{1}{d_2/q_2}} \tag{15}$$

Proof. Aczél proved that the the quasi–linear mean

$$F(d_1, d_2\,;\, q_1, q_2) = f\left(\frac{q_1 f^{-1}(d_1) + q_2 f^{-1}(d_2)}{q_1 + q_2}\right) , \tag{16}$$

with a suitable function f, is characterized by (8a)–(8f), see [1, p. 242]. We show that F has the form (14), and consequently that the function f in (16) is $f(t) = 1/t$.

From (13a) it follows that

$$F(d_1, d_2\,;\, q_1, q_2) = d_1\, \phi(d_1, d_2\,;\, q_1, q_2) = d_2\, \psi(d_1, d_2\,;\, q_1, q_2) , \tag{17}$$

for some functions ϕ , ψ that are homogenous of degree 0 by (13b), and satisfy for all $d_1 > 0$,

$$\phi(d_1, d_2\,;\, q_1, q_2) = \frac{d_2}{d_1}\, \psi(d_1, d_2\,;\, q_1, q_2) . \tag{18}$$

Differentiating (17) we verify

$$F_1'(0, d_2\,;\, q_1, q_2) = \phi(0, d_2\,;\, q_1, q_2) ,\ F_2'(d_1, 0\,;\, q_1, q_2) = \psi(d_1, 0\,;\, q_1, q_2) . \tag{19}$$

From (13c)–(13d), (18) and (19) it follows that

$$\phi(d_1, d_2\,;\, q_1, q_2) = \frac{(q_1 + q_2)d_2}{q_2 d_1 + q_1 d_2} \ , \text{ and } \psi(d_1, d_2\,;\, q_1, q_2) = \frac{(q_1 + q_2)d_1}{q_2 d_1 + q_1 d_2} , \tag{20}$$

proving (14). $\square$

Other properties of the function $F(d_1, d_2\,;\, q_1, q_2)$ follow easily from its definition. We list some below.

Corollary 1. The function F of (14) has the following properties for all $d_1, d_2 \geq 0$, $q_1, q_2 \geq 0$ such that $q_1 + q_2 > 0$:

$$F(d_1, d_2\,;\, q_1, q_2) = F(d_2, d_1\,;\, q_2, q_1) , \tag{21a}$$

$$F(F(d_1, d_2\,;\, q_1, q_2), d_3\,;\, q_1 + q_2, q_3) = F(d_1, F(d_2, d_3\,;\, q_2, q_3)\,;\, q_1, q_2 + q_3) ,\ \forall\ d_3, q_3 \geq 0 , \tag{21b}$$

and if $q_1, q_2 \geq 1$,

$$F(d_1, d_2\,;\, q_1, q_2) \leq (q_1 + q_2)\ \min\{d_1, d_2\} . \tag{21c}$$

Remarks 3.
(a) The symmetry relation (21a) is a rewriting of (10a). In the clustering context it implies impartiality between clusters.
(b) The common value of (21b) defines F for the case of 3 clusters,

$$F(d_1, d_2, d_3\,;\, q_1, q_2, q_3) = \frac{(q_1 + q_2 + q_3)\, d_1 d_2 d_3}{q_3 d_1 d_2 + q_2 d_1 d_3 + q_1 d_2 d_3}\,. \tag{22}$$

(c) The general case is defined analogously,

$$\begin{aligned} F(d_1, \cdots, d_k\,;\, q_1, \cdots, q_k) &:= F(F(d_1, \cdots, d_{k-1}\,;\, q_1, \cdots, q_{k-1}), d_k\,;\, q_1 + \cdots + q_{k-1}, q_k) \\ &= \frac{\left(\sum\limits_{i=1}^{k} q_i\right) \prod\limits_{i=1}^{k} d_i}{\sum\limits_{i=1}^{k} q_i \prod\limits_{j \neq i} d_j}\,. \end{aligned} \tag{23}$$

(d) Properties of $F(d_1, d_2\,;\, q_1, q_2)$ are easily translated to the general case (23). For example, the contour approximation D of (5) is

$$D(\mathbf{x}) = \frac{F(d_1(\mathbf{x}, \mathbf{c}_1), \cdots, d_K(\mathbf{x}, \mathbf{c}_K)\,;\, q_1, \cdots, q_K)}{q_1 + q_2 + \cdots + q_K}\,, \tag{24}$$

in analogy with (12), and the analog of (21c) is, for all $q_1, \cdots, q_K \geq 1$,

$$F(d_1, \cdots, d_K\,;\, q_1, \cdots, q_K) \leq (q_1 + \cdots + q_K)\, \min\{d_1, \cdots, d_K\} \tag{25}$$

guaranteeing the contour approximation inequality (6).
(e) Identifying $F(x, y\,;\, r, s)$ of (7) with $F(d_1, d_2\,;\, q_1, q_2)$ of (14), we note that the weights r, s are only required to be nonnegative with $r + s > 0$, while the cluster sizes q_1, q_2 are positive, and $q_1, q_2 \geq 1$ is required in (21c).

4 Discussion

(a) Some clustering methods use all pairwise distances between data points, not just distances to the cluster centers. These pairwise distances form a distance matrix, one for each data point. The matrix analog of the harmonic mean is the **parallel sum**, see, e.g., [2], [3], [4], [7], [9] and [10]. An analogous theory of contour approximation can be developed using the parallel sum of the distance matrices, but the computations involved are prohibitive.
(b) It is interesting that **home ranges** (areas of activity) of animals are based on the harmonic mean of areal moments in much the same way, and for similar reasons, that contour approximation uses the harmonic mean of distances, see, e.g., [6].
(c) A unified optimization framework for distance clustering, employing quasi–linear means and other functions of distances, is given in [11].

Acknowledgement

Thanks to Cem Iyigun for the graphics of Figs. 1–2, and permission to use them.

References

[1] J. Aczél, *Lectures on Functional Equations and their Applications*, Academic Press, 1966

[2] W.N. Anderson Jr., *Shorted operators*, SIAM J. Appl. Math. **20**(1971), 520–525

[3] W.N. Anderson Jr. and R.J. Duffin, *Series and parallel of matrices*, J. Math. Anal. Appl. **26**(1969), 576–594

[4] W.N. Anderson Jr., T.D. Morley and G.E. Trapp, *Ladder networks, fixpoints, and the geometric mean*, Circuits, Systems, and Signal Processing **2**(1983), 259–268

[5] A. Ben–Israel and C. Iyigun, *Probabilistic distance clustering*, Tech Report 2006-09, DIMACS–Center for Discrete Mathematics & Theoretical Computer Science, Rutgers University, May 2006

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

[7] R.J. Duffin, *Network Models*, Mathematical Aspects of Electrical Neteork Analysis, SIAM–AMS Proceedings, Vol. III, pp. 65–91, American Mathematical Society, 1971

[8] C. Iyigun and A. Ben–Israel, *Probabilistic distance clustering with cluster size*, DIMACS–Center for Discrete Mathematics & Theoretical Computer Science, Rutgers University, November 2006

[9] F. Kubo and T. Ando, *Means of positive linear operators*, Mathematische Annalen **246**(1980), 205–224

[10] T.D. Morley, *Parallel summation, Maxwell's principle and the infimum of projections*, J. Math. Anal. Appl. **70**(1979), 33–41

[11] M. Teboulle, *A unified continuous optimization framework to center-based clustering methods*, School of Math. Sci., Tel–Aviv University, 2006