The multivariate L_1-median and associated data depth

Yehuda Vardi[†] and Cun-Hui Zhang[‡]

Department of Statistics, Rutgers University, New Brunswick, NJ 08854

Communicated by Lawrence A. Shepp, Rutgers, The State University of New Jersey, Piscataway, NJ, November 17, 1999 (received for review October 15, 1999)

This paper gives three related results: (*i*) a new, simple, fast, monotonically converging algorithm for deriving the L_1-median of a data cloud in $\mathbb{R}^d$, a problem that can be traced to Fermat and has fascinated applied mathematicians for over three centuries; (*ii*) a new general definition for depth functions, as functions of multivariate medians, so that different definitions of medians will, correspondingly, give rise to different depth functions; and (*iii*) a simple closed-form formula of the L_1-depth function for a given data cloud in $\mathbb{R}^d$.

1. Introduction

In this paper, we derive three related results about multivariate median (MM) and data depth (DD): (*i*) a simple, but nontrivial, modification of the Weiszfeld (1) iterative algorithm for the computation of the multivariate L_1-median (L_1-MM); (*ii*) a general method for generating DD functions based on MMs, so that different MMs will, correspondingly, give rise to different DD functions and the MMs are always the points with the largest DD (= 1) they generate; and (*iii*) a simple closed-form formula for the L_1 data depth (L_1-DD), the DD function corresponding to the L_1-MM.

Consider the problem of minimizing the weighted sum of the Euclidean distances from m points, in $\mathbb{R}^d$. In industrial applications, this is known as the optimal location problem of Weber (2). In statistics, the solution of this optimization problem is the spatial median or L_1-MM, considered by Brown (3) and Small (4). As noted by Kuhn (5), the problem goes back to Fermat in the early seventeenth century and was generalized to the current form by Simpson in his *Doctrine and Application of Fluxions* (6). In the nineteenth century, Steiner made significant contributions to this problem and its extensions (cf. Courant and Robbins; ref. 7). Thus, the problem is known as the Fermat–Weber location problem and also as the Euclidean–Steiner problem. Dozens of papers have been written on variants of this problem, and most of them reproduced known results in different fields of applications. Weiszfeld (1) proposed a simple iterative algorithm, which has been rediscovered at least three times, and Kuhn (5) gave the algorithm a rigorous treatment. In particular, Kuhn corrected earlier claims and showed that the algorithm converges (monotonically) unless the starting point is inside the domain of attraction of the data points. Although Kuhn claimed that the domain of attraction of the data points is a denumerable set, Chandrasekaran and Tamir (8) pointed out an error in Kuhn's argument and showed that it could contain a continuum set. This set of bad starting points is not easy to identify, as demonstrated by Kuhn's example. Furthermore, Kuhn's proof does not preclude the possibility that this set of bad starting points is dense in an open region of $\mathbb{R}^d$, or perhaps the entire $\mathbb{R}^d$. In Section 2, we provide a new algorithm, a nontrivial modification of Weiszfeld's, which is guaranteed to converge monotonically to the L_1-MM of a data cloud from *any* starting point in $\mathbb{R}^d$.

Given the definition of an MM, say θ, and a distribution function in $\mathbb{R}^d$, say F, it is natural to treat F as data and define a DD function by asking what is the minimum incremental mass at location $\mathbf{y} \in \mathbb{R}^d$, say $w(\mathbf{y})$, needed for $\mathbf{y}$ to become the median of the resulting mixture distribution $(w\delta_{\mathbf{y}} + F)/(1 + w)$. We take $1 - w(\mathbf{y})$ to be the depth at $\mathbf{y}$. The value of the resulting DD function is always between zero and one, provided that θ satisfies the following condition: if F puts at least 1/2 of the probability mass at $\mathbf{y}$, then $\theta(F) = \mathbf{y}$ (cf. **3.2** below). Based on this concept of DD, different definitions of MM give rise to different DD functions, and the maximum depth 1 is always achieved at the MM that generates the DD function. This concept of depth is developed and applied to a number of examples in Section 3.

In Section 4, the above concept of DD-functions is applied to the L_1-MM to obtain a simple closed-form formula for the L_1-DD function.

2. The L_1-Median

Let $\mathbf{x}_1, \ldots, \mathbf{x}_m$ be m distinct points in $\mathbb{R}^d$ and $\eta_1, \ldots, \eta_m$ be m positive numbers. Think of the η_is as weights or, better yet, as *"multiplicities"* of the $\mathbf{x}_i$s, and let $C(\mathbf{y})$ denote the weighted sum of distances of $\mathbf{y}$ from $\mathbf{x}_1, \ldots, \mathbf{x}_m$:

$$C(\mathbf{y}) = \sum_i \eta_i d_i(\mathbf{y}), \qquad [2.1]$$

where $d_i(\mathbf{y}) = \|\mathbf{y} - \mathbf{x}_i\|$, the Euclidean distance between $\mathbf{y}$ and $\mathbf{x}_i$ in $\mathbb{R}^d$. The problem we consider is to find a point $\mathbf{y} \in \mathbb{R}^d$ (or a set of points) that minimizes the "cost function" $C(\mathbf{y})$, i.e., to find

$$\begin{aligned} \mathbf{M} &= \mathbf{M}(\mathbf{x}_1, \ldots, \mathbf{x}_m; \eta_1, \ldots, \eta_m) \\ &= \arg\min \{C(\mathbf{y}) : \ \mathbf{y} \in \mathbb{R}^d\}. \end{aligned} \qquad [2.2]$$

The solution $\mathbf{M}$ is called the L_1-median.

The problem has the following (idealized) interpretation of optimal selection of a location: a company wishes to choose an appropriate location for a warehouse that will service η_i customers at location $\mathbf{x}_i$, $i = 1, \ldots, m$. Each customer requires a daily trip from the warehouse, and the cost of travel from the warehouse to the customer is proportional to the Euclidean distance between their respective locations. Fraction customers incur the corresponding fraction of the cost. The company wants to locate the warehouse at a point $\mathbf{y}$ that minimizes the total cost of travel, $C(\mathbf{y})$, from the warehouse to its customers.

When the points $\mathbf{x}_1, \ldots, \mathbf{x}_m$ are not collinear, $C(\mathbf{y})$ is positive and strictly convex in $\mathbb{R}^d$, and hence the minimum is achieved at a unique point $\mathbf{M} \in \mathbb{R}^d$. In the collinear case where $\mathbf{x}_1, \ldots, \mathbf{x}_m$ lie in a straight line, the minimum of $C(\mathbf{y})$ is achieved at any one-dimensional median (always exists but may not be unique). We now consider only noncollinear problems (unless specified otherwise).

Abbreviations: MM, multivariate median; DD, data depth; L_1-MM, multivariate L_1-median; L_1-DD, L_1 data depth.

[†]E-mail: vardi@stat.rutgers.edu.

[‡]E-mail: czhang@stat.rutgers.edu.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Modified Weiszfeld Algorithm for 2.2. Define an $\mathbb{R}^d$ to $\mathbb{R}^d$ mapping

$$\widetilde{T}: \mathbf{y} \to \widetilde{T}(\mathbf{y}) = \sum_{\mathbf{x}_i \neq \mathbf{y}} w_i(\mathbf{y})\mathbf{x}_i, \qquad [2.3]$$

where $\{w_i(\mathbf{y}) : i = 1, \ldots, m, \mathbf{x}_i \neq \mathbf{y}\}$ are positive weights that sum to one and satisfy

$$w_i(\mathbf{y}) \propto \eta_i / d_i(\mathbf{y}).$$

Alternatively, because $d_i(\mathbf{y}) = \|\mathbf{y} - \mathbf{x}_i\|$, $\widetilde{T}(\mathbf{y})$ can be written as

$$\widetilde{T}(\mathbf{y}) = \left\{ \sum_{\mathbf{x}_i \neq \mathbf{y}} \frac{\eta_i}{\|\mathbf{y} - \mathbf{x}_i\|} \right\}^{-1} \sum_{\mathbf{x}_i \neq \mathbf{y}} \frac{\eta_i \mathbf{x}_i}{\|\mathbf{y} - \mathbf{x}_i\|}. \qquad [2.4]$$

The Weiszfeld algorithm is defined as

$$\mathbf{y} \to T_0(\mathbf{y}) = \begin{cases} \widetilde{T}(\mathbf{y}) & \text{if } \mathbf{y} \notin \{\mathbf{x}_1, \ldots, \mathbf{x}_m\} \\ \mathbf{x}_k & \text{if } \mathbf{y} = \mathbf{x}_k, k = 1, \ldots, m. \end{cases}$$

It converges to the L_1-median in **2.2** for a given initial point, if the algorithm never reaches the set $\{\mathbf{x}_k : k = 1, \ldots, m, \mathbf{x}_k \neq \mathbf{M}\}$. Thus, we shall modify the algorithm for $\mathbf{y} \in \{\mathbf{x}_k : k = 1, \ldots, m\}$. If $\mathbf{y} = \mathbf{x}_k$, then $\widetilde{T}(\mathbf{y}) = \widetilde{T}(\mathbf{x}_k)$ is a weighted average of data points other than $\mathbf{x}_k$, so that it makes sense to consider a weighted average of $\widetilde{T}(\mathbf{x}_k)$ and $\mathbf{x}_k$. But what should be the weights? Given $\mathbf{y} \in \mathbb{R}^d$, it is convenient to include $\mathbf{y}$ in the data and define the multiplicity at $\mathbf{y}$ to be

$$\eta(\mathbf{y}) = \begin{cases} \eta_k & \text{if } \mathbf{y} = \mathbf{x}_k, k = 1, \ldots, m, \\ 0 & \text{otherwise.} \end{cases} \qquad [2.5]$$

The new algorithm is

$$\mathbf{y} \to T(\mathbf{y}) = \left(1 - \frac{\eta(\mathbf{y})}{r(\mathbf{y})}\right)^+ \widetilde{T}(\mathbf{y}) + \min\left(1, \frac{\eta(\mathbf{y})}{r(\mathbf{y})}\right)\mathbf{y}, \qquad [2.6]$$

with the convention $0/0 = 0$ in the computation of $\eta(\mathbf{y})/r(\mathbf{y})$, where $\widetilde{T}(\mathbf{y})$ is as in **2.4**,

$$r(\mathbf{y}) = \|\widetilde{R}(\mathbf{y})\|, \quad \widetilde{R}(\mathbf{y}) = \sum_{\mathbf{x}_i \neq y} \eta_i \frac{\mathbf{x}_i - \mathbf{y}}{\|\mathbf{x}_i - \mathbf{y}\|}. \qquad [2.7]$$

For $\mathbf{y} \notin \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, $T(\mathbf{y}) = \widetilde{T}(\mathbf{y})$, by **2.6** with $\eta(\mathbf{y}) = 0$, as in the Weiszfeld algorithm. For $\mathbf{y} = \mathbf{x}_k$, $T(\mathbf{y})$ is a weighted average of $\widetilde{T}(\mathbf{y}) = \widetilde{T}(\mathbf{x}_k)$ and $\mathbf{y} = \mathbf{x}_k$, so that by **2.4** $T(\mathbf{y})$ is a weighted average of $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$. Also, for $\mathbf{y} \notin \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, $\widetilde{R}(\mathbf{y})$ of **2.7** is the negative of the gradient of $C(\mathbf{y})$. It follows from **2.3** and **2.4** that

$$\widetilde{R}(\mathbf{y}) = \{\widetilde{T}(\mathbf{y}) - \mathbf{y}\} \sum_{\mathbf{x}_i \neq \mathbf{y}} \eta_i / d_i(\mathbf{y}). \qquad [2.8]$$

This and **2.7** imply that $\widetilde{T}(\mathbf{y}) = \mathbf{y} = T(\mathbf{y})$ when $r(\mathbf{y}) = \|\widetilde{R}(\mathbf{y})\| = 0$.

Properties of the L_1-Median M in 2.2 and Algorithm 2.6.

$$\mathbf{y} = \mathbf{M} \text{ iff } T(\mathbf{y}) = \mathbf{y} \text{ iff } r(\mathbf{y}) \leq \eta(\mathbf{y}). \qquad [2.9]$$

In words: $\mathbf{y} \in \mathbb{R}^d$ is the L_1-median if and only if it is a fixed-point of our iterative algorithm **2.6**, if and only if $r(\mathbf{y}) \leq \eta(\mathbf{y})$, where $r(\mathbf{y})$ and $\eta(\mathbf{y})$ are given in **2.7** and **2.5** respectively.

Monotonicity of the Algorithm.

$$\text{If } \mathbf{y} \neq \mathbf{M}, \text{ then } C(T(\mathbf{y})) < C(\mathbf{y}). \qquad [2.10]$$

Convergence Theorem.

$$\lim_{n \to \infty} T^n(\mathbf{y}) = \mathbf{M} \text{ for all } \mathbf{y} \in \mathbb{R}^d. \qquad [2.11]$$

We note that the algorithm is extremely simple to program and our simulation results indicate very quick convergence. The proofs of **2.9–2.11** are given in Section 5. We proceed to describe the concept of DD based on multivariate median and define the L_1-DD based on the above.

3. Depth

Given a definition of a multivariate median, say θ, and a distribution function or, equivalently, data in $\mathbb{R}^d$, say F, we define the corresponding depth function (DD function) in $\mathbb{R}^d$ as follows:

$$D_{\theta,F}(\mathbf{y}) \equiv D_F(\mathbf{y}) = 1 - \inf\left\{w \geq 0 : \ \theta\left(\frac{w\delta_{\mathbf{y}} + F}{1+w}\right) = \mathbf{y}\right\}, \qquad [3.1]$$

where $\delta_{\mathbf{y}}$ is a point mass at $\mathbf{y}$. That is, $1 - D_F(\mathbf{y})$ is the minimum incremental mass w needed at position $\mathbf{y}$ for $\mathbf{y}$ to become the median of the mixture $(w\delta_{\mathbf{y}} + F)/(1+w)$. Because DD functions are defined by using MMs, it is natural that certain estimators such as the mean $\int \mathbf{x} dF$ are excluded. Throughout this paper, an MM must satisfy the following condition:

$$F(\{\mathbf{y}\}) \geq 1/2 \quad \Rightarrow \quad \theta(F) = \mathbf{y}, \qquad [3.2]$$

where $F(\{\mathbf{y}\})$ is the probability mass distributed to the point $\mathbf{y}$ by F. If **3.2** holds for $\theta(\cdot)$, then the depth function in **3.1** is nonnegative and well defined for all $\mathbf{y}$ and F.

It follows automatically from the definition of the depth in **3.1** that for all F in the domain of the multivariate median θ

$$\theta(F) \in \{\mathbf{y} : D_{\theta,F}(\mathbf{y}) = 1\}. \qquad [3.3]$$

In this sense, deeper points, with larger depth, are relatively closer to the prescribed median. The definition applies to data clouds by simply taking F to be the empirical distribution of the data cloud. We shall consider **3.1** for three examples in this section. The depth function associated with the L_1-MM, the fourth example, is discussed in detail in Section 4.

Example 1. The One-Dimensional Case. Let the median be denoted by $\theta^{(1)}(F)$ for univariate distributions F. A point y is the median of the mixture $(w\delta_y + F)/(1+w)$ iff both $(w + F(y))/(1+w)$ and $(w + \{1 - F(y-)\})/(1+w)$ are greater than or equal to 1/2, so that by **3.1** the depth function is

$$D_F^{(1)}(y) \equiv D_{\theta^{(1)},F}(y) = 2\min\{F(y), 1 - F(y-)\}. \qquad [3.4]$$

Example 2. Depth Based on the Marginal Median. A simple extension of the univariate median to multivariate distributions F in $\mathbb{R}^d$ is the marginal median

$$\theta^{(M)}(F) \equiv (\theta^{(1)}(F_1), \ldots, \theta^{(1)}(F_d)),$$

where F_j is the marginal distribution of the jth variable under F. The corresponding depth function, associated with $\theta^{(M)}$ by **3.1**, is then

$$\begin{aligned} D_F^{(M)}(\mathbf{y}) &= \min_{1 \leq j \leq d} \left\{D_{F_j}^{(1)}(y_j)\right\} \\ &= 2 \min_{1 \leq j \leq d} \min\{F_j(y_j), 1 - F_j(y_j-)\}, \end{aligned} \qquad [3.5]$$

where $D^{(1)}$ is as in **3.4** and y_j is the jth component of $\mathbf{y}$.

Example 3. The Tukey Depth. A more stringent extension, say $\theta^{(T)}$, of the univariate median requires that the projection of $\theta^{(T)}(F)$ be the median of the projection of the distribution F under any linear projection from $\mathbb{R}^d$ to $\mathbb{R}$. That is

$$\theta^{(T)}(F) = \mathbf{y} \quad \text{iff} \quad \mathbf{a}^{tr}\mathbf{y} = \theta^{(1)}\big(F_{\pi_\mathbf{a}}\big) \ \ \forall \mathbf{a} \in \mathbb{R}^d, \tag{3.6}$$

where $F_{\pi_\mathbf{a}}(t) = F\big(\{\mathbf{x} : \mathbf{a}^{tr}\mathbf{x} \le t\}\big)$ is the marginal distribution of F with the projection $\mathbf{x} \to \mathbf{a}^{tr}\mathbf{x}$. Although $\theta^{(T)}(F)$ is not defined for all F, the depth function **3.1** is always defined because $\theta^{(T)}\big(\{\delta_\mathbf{y} + F\}/2\big) = \mathbf{y}$ for all F and $\mathbf{y}$. The resulting depth function is

$$\begin{aligned} D_F^{(T)}(\mathbf{y}) &= \inf_{\mathbf{a}\in\mathbb{R}^d} D^{(1)}_{F_{\pi_\mathbf{a}}}\big(\mathbf{a}^{tr}\mathbf{y}\big) \\ &= 2 \inf_{\mathbf{a}\in\mathbb{R}^d} \min\big\{F_{\pi_\mathbf{a}}\big(\mathbf{a}^{tr}\mathbf{y}\big), 1 - F_{\pi_\mathbf{a}}\big(\mathbf{a}^{tr}\mathbf{y} - \big)\big\}, \end{aligned} \tag{3.7}$$

again with the $D^{(1)}$ in **3.4**. Hence, $D_F^{(T)}(\mathbf{y})/2$ is the Tukey (9) depth.

4. The L_1-depth.

Consider $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^d$ with multiplicities $\eta_1, \ldots, \eta_m$, respectively, and let $\mathbf{y} \in \mathbb{R}^d$. We shall apply the depth concept of **3.1** to this set-up using the L_1-MM.

Let $\mathbf{M}$ be the L_1-MM in **2.2**. Given a point $\mathbf{y}$ in $\mathbb{R}^d$, what is the minimum increment of multiplicity $\eta'(\mathbf{y})$ at $\mathbf{y}$, in addition to $\eta(\mathbf{y})$ in **2.5**, to ensure that $\mathbf{y}$ becomes the L_1-MM for the new data set with the new multiplicities? The answer to this question is an immediate consequence of **2.5**, **2.7**, and **2.9**: because $r(\mathbf{y})$ in **2.7** does not depend on the multiplicity $\mathbf{y}$ in **2.5**, by **2.8** $\mathbf{y}$ becomes the median for the modified data (multiplicities $\eta(\mathbf{y}) + \eta'(\mathbf{y})$ at $\mathbf{y}$ and η_k at $\mathbf{x}_k \ne \mathbf{y}$) if the incremental multiplicity $\eta'(\mathbf{y})$ is large enough to ensure that $\eta(\mathbf{y}) + \eta'(\mathbf{y}) \ge r(\mathbf{y})$. Thus, the minimum increment, to turn $\mathbf{y}$ into the L_1-MM, is

$$\eta'(\mathbf{y}) = \max\{r(\mathbf{y}) - \eta(\mathbf{y}), 0\}. \tag{4.1}$$

By **3.1**, the L_1-DD is the corresponding complementary proportion:

$$D(\mathbf{y}) = 1 - \frac{\eta'(\mathbf{y})}{\sum_{j=1}^m \eta_j} = 1 - \frac{\max\{r(\mathbf{y}) - \eta(\mathbf{y}), 0\}}{\eta_1 + \cdots + \eta_m}. \tag{4.2}$$

Alternatively, if $\mathbf{e}_i(\mathbf{y}) = (\mathbf{y} - \mathbf{x}_i)/\|\mathbf{y} - \mathbf{x}_i\|$ and $\bar{\mathbf{e}}(\mathbf{y}) = \sum_{\mathbf{x}_k \ne \mathbf{y}} \mathbf{e}_i(\mathbf{y}) f_i$ with $f_i = \eta_i / \sum_{j=1}^k \eta_j$, then

$$D(\mathbf{y}) = \begin{cases} 1 - \|\bar{\mathbf{e}}(\mathbf{y})\| & \text{if } \mathbf{y} \notin \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}, \\ 1 - \big(\|\bar{\mathbf{e}}(\mathbf{y})\| - f_k\big)^+ & \text{if } \mathbf{y} = \mathbf{x}_k, \ \forall k. \end{cases} \tag{4.3}$$

The function $\bar{\mathbf{e}}(\mathbf{y})$ is the spatial rank function considered in Möttönen and Oja (10) and Marden (11). Related multivariate quantiles were considered by Chaudhuri (12).

Note that, because $\mathbf{e}_i(\mathbf{y})$ are vectors of unit length for $\mathbf{y} \ne \mathbf{x}_i$, $\|\bar{\mathbf{e}}(\mathbf{y})\| \le \sum_{\mathbf{x}_k \ne \mathbf{y}} f_k \le 1$, so that by **4.3**

$$0 \le D(\mathbf{y}) \le 1. \tag{4.4}$$

Because $\lim_{c\to\infty} \mathbf{e}_i(c\mathbf{y}) = \mathbf{y}$ for $\|\mathbf{y}\| = 1$, $\lim_{\|\mathbf{y}\|\to\infty} \|\bar{\mathbf{e}}(\mathbf{y})\| = 1$ and

$$\lim_{\|\mathbf{y}\|\to\infty} D(\mathbf{y}) = 0. \tag{4.5}$$

Furthermore, by **4.3** and the fact that $\|\bar{\mathbf{e}}(\mathbf{y})\| \le \sum_{\mathbf{x}_i \ne \mathbf{y}} f_i$, $D(\mathbf{x}_k) = 1$ if $f_k \ge 1/2$; i.e., a data point $\mathbf{x}_k$ is the L_1-MM if it possesses half of the total multiplicity. Thus, **3.2** holds for the L_1-MM. The identity **4.5** is closely related to the well known fact that the breakdown of the L_1-MM is 1/2. Note that $D_{\theta,F} = 0$ iff the optimal w in **3.1** is one, that corresponds to the mixing probability $w/(1 + w) = 1/2$ for the mixture $(w\delta_\mathbf{y} + F)/(1 + w)$. See **(6.1–6.3)** in Section 6.

5. Proofs of 2.9–2.11

The key to the proof is the inequality

$$C(T(\mathbf{x}_k)) < C(\mathbf{x}_k) \quad \text{if } \mathbf{x}_k \ne \mathbf{M}, \quad k = 1, \ldots, m. \tag{5.1}$$

The monotonicity property **2.10** follows from **5.1**, because $C(T(\mathbf{y})) = C(\widetilde{T}(\mathbf{y}))$ for $\mathbf{y} \notin \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, and Kuhn (5) proved $C(\widetilde{T}(\mathbf{y})) < C(\mathbf{y})$ for the Weiszfeld algorithm in this case. Inequality **2.10** then implies that, starting from any initial point $\mathbf{y}$ in $\mathbb{R}^d$, the sequence $T^n(\mathbf{y})$ in the iterative algorithm **2.6** visits each $\mathbf{x}_k \ne \mathbf{M}$ at most once and it will not get stuck at $\mathbf{x}_k \ne \mathbf{M}$. After the last visit to the set $\{\mathbf{x}_k, k = 1, \ldots, m, \mathbf{x}_k \ne \mathbf{M}\}$, $T^n(\mathbf{y})$ converges to the L_1-MM $\mathbf{M}$ by Kuhn (5). This proves the convergence theorem **2.11** based on **5.1**. Moreover, Kuhn (5) proved that

$$\begin{aligned} \mathbf{y} = \mathbf{M} \quad &\text{iff} \quad R(\mathbf{y}) = 0, \\ R(\mathbf{y}) \equiv \{r(\mathbf{y}) - \eta(\mathbf{y})\}^+ &\widetilde{R}(\mathbf{y})/r(\mathbf{y}), \end{aligned} \tag{5.2}$$

with the convention $R(\mathbf{y}) \equiv 0$ for $r(\mathbf{y}) \equiv \|\widetilde{R}(\mathbf{y})\| = 0$. Thus, $\mathbf{y} = \mathbf{M}$ implies $r(\mathbf{y}) \le \eta(\mathbf{y})$. Finally, $r(\mathbf{y}) \le \eta(\mathbf{y})$ implies $T(\mathbf{y}) = \mathbf{y}$ by **2.6**, as $r(\mathbf{y}) = 0$ implies $T(\mathbf{y}) = \widetilde{T}(\mathbf{y}) = \mathbf{y}$, while $T(\mathbf{y}) = \mathbf{y}$ implies $\mathbf{y} = \mathbf{M}$ by the monotonicity **2.10**. Thus, **2.9** is also proved based on **5.1**.

It remains to prove **5.1**, because the rest of the proof has already been completed above under the assumption of **5.1**. Consider a fixed data point $\mathbf{x}_k \ne \mathbf{M}$. Define

$$g_k(\mathbf{x}) = 2\eta_k\|\mathbf{x} - \mathbf{x}_k\| + \sum_{i\ne k} \frac{\eta_i}{d_i(\mathbf{x}_k)} \|\mathbf{x} - \mathbf{x}_i\|^2. \tag{5.3}$$

Because $\widetilde{T}(\mathbf{x}_k)$ is the weighted average of $\{\mathbf{x}_i, i \ne k\}$ with weights proportional to $\eta_i/d_i(\mathbf{x}_k)$, as in **2.3**, $\widetilde{T}(\mathbf{x}_k)$ is the minimizer of $\sum_{i\ne k}\{\eta_i/d_i(\mathbf{x}_k)\}\|\mathbf{x} - \mathbf{x}_i\|^2$ in **5.3** and due to the cancelation of cross-product terms

$$\begin{aligned} g_k(\mathbf{x}) &= 2\eta_k\|\mathbf{x} - \mathbf{x}_k\| + \sum_{i\ne k} \frac{\eta_i}{d_i(\mathbf{x}_k)} \|\mathbf{x} - \widetilde{T}(\mathbf{x}_k)\|^2 \\ &\quad + \sum_{i\ne k} \frac{\eta_i}{d_i(\mathbf{x}_k)} \|\widetilde{T}(\mathbf{x}_k) - \mathbf{x}_i\|^2 \\ &= 2\eta_k\|\mathbf{x} - \mathbf{x}_k\| + A_k\|\mathbf{x} - \widetilde{T}(\mathbf{x}_k)\|^2 + B_k, \end{aligned} \tag{5.4}$$

where $A_k = \sum_{i\ne k} \eta_i/d_i(\mathbf{x}_k)$ and $B_k = \sum_{i\ne k}\{\eta_i/d_i(\mathbf{x}_k)\}\|\widetilde{T}(\mathbf{x}_k) - \mathbf{x}_i\|^2$. Because η_k, A_k and B_k in **5.4** do not depend on $\mathbf{x}$,

$$\min_\mathbf{z} g_k(\mathbf{z}) = \min_{0\le w\le 1} g_k((1 - w)\widetilde{T}(\mathbf{x}_k) + w\mathbf{x}_k). \tag{5.5}$$

i.e., $g_k(\mathbf{x}) \le g_k(\mathbf{z})$ if $\mathbf{x}$ is the projection of $\mathbf{z}$ to the line segment with endpoints $\mathbf{x}_k$ and $\widetilde{T}(\mathbf{x}_k)$. For the points in this segment and the above $g_k(\cdot)$,

$$\begin{aligned} &g_k((1 - w)\widetilde{T}(\mathbf{x}_k) + w\mathbf{x}_k) \\ &\quad = |1 - w|2\eta_k\|\widetilde{T}(\mathbf{x}_k) - \mathbf{x}_k\| + w^2 A_k\|\widetilde{T}(\mathbf{x}_k) - \mathbf{x}_k\|^2 + B_k, \end{aligned}$$

as a strictly convex function in w, is uniquely minimized at

$$\begin{aligned} w_* &\equiv \min\left\{1, \frac{\eta_k\|\widetilde{T}(\mathbf{x}_k) - \mathbf{x}_k\|}{A_k\|\widetilde{T}(\mathbf{x}_k) - \mathbf{x}_k\|^2}\right\} \\ &= \arg\min_w g_k((1 - w)\widetilde{T}(\mathbf{x}_k) + w\mathbf{x}_k). \end{aligned} \tag{5.6}$$

Because $\mathbf{x}_k \ne \mathbf{M}$, by **5.2** and **2.8**

$$r(\mathbf{x}_k) > \eta(\mathbf{x}_k) \ge 0, \qquad \widetilde{T}(\mathbf{x}_k) \ne \mathbf{x}_k. \tag{5.7}$$

By **2.7** and **2.8** and the definition of A_k in **5.4**, $r(\mathbf{x}_k) = \|\widetilde{T}(\mathbf{x}_k) - \mathbf{x}_k\|A_k$, so that $w_* = \eta(\mathbf{x}_k)/r(\mathbf{x}_k) \in (0, 1)$ by **2.5** and **5.7**. Thus,

by **5.5**, **5.6** and **2.6**, $T(\mathbf{x}_k) = (1 - w_*)\widetilde{T}(\mathbf{x}_k) + w_*\mathbf{x}_k$ is the unique minimizer of $g_k(\mathbf{x})$ and $T(\mathbf{x}_k) \neq \mathbf{x}_k$. These and **5.3** and the definition of $C(\mathbf{y})$ in **2.1** imply that

$$\begin{aligned} C(\mathbf{x}_k) &= g_k(\mathbf{x}_k) > g_k(T(\mathbf{x}_k)) \\ &= 2\eta_k\|T(\mathbf{x}_k) - \mathbf{x}_k\| + \sum_{i\neq k}\frac{\eta_i}{d_i(\mathbf{x}_k)}d_i^2(T(\mathbf{x}_k)) \\ &\geq 2\eta_k\|T(\mathbf{x}_k) - \mathbf{x}_k\| \\ &\quad + \sum_{i\neq k}\frac{\eta_i}{d_i(\mathbf{x}_k)}\left[d_i^2(\mathbf{x}_k) + 2d_i(\mathbf{x}_k)\left\{d_i(T(\mathbf{x}_k)) - d_i(\mathbf{x}_k)\right\}\right] \\ &= C(\mathbf{x}_k) + 2C(T(\mathbf{x}_k)) - 2C(\mathbf{x}_k). \end{aligned}$$

Hence, **5.1** holds for $\mathbf{x}_k \neq \mathbf{M}$ and the proof is complete.

6. Remark

An alternative definition of DD functions, based on the same idea as **3.1**, is

$$\begin{aligned} \widetilde{D}_{\theta,F}(\mathbf{y}) &\equiv D_F(\mathbf{y}) \\ &= 1 - \inf\{w \geq 0 :\ \theta(w\delta_{\mathbf{y}} + (1-w)F) = \mathbf{y}\}. \end{aligned} \qquad [\mathbf{6.1}]$$

That is, $1 - D_F(\mathbf{y})$ is the minimum mixing probability mass needed to be mixed into F, at position $\mathbf{y}$, for $\mathbf{y}$ to be the median of the mixture $w\delta_{\mathbf{y}} + (1-w)F$. Condition **3.2** is not required here, because **6.1** is well defined for all F and $\mathbf{y}$ if $\theta(\delta_{\mathbf{y}}) = \mathbf{y}$ for all $\mathbf{y} \in \mathbb{R}^d$. By simple algebra, **3.1** and **6.1** are monotone functions of each other, with

$$\widetilde{D}_{\theta,F}(\mathbf{y}) = \frac{1}{2 - D_{\theta,F}(\mathbf{y})}, \quad D_{\theta,F}(\mathbf{y}) = 2 - \frac{1}{\widetilde{D}_{\theta,F}(\mathbf{y})}. \qquad [\mathbf{6.2}]$$

Moreover, by **4.5**

$$\lim_{\|\mathbf{y}\|\to\infty}\widetilde{D}_{\theta,F}(\mathbf{y}) = 1/2. \qquad [\mathbf{6.3}]$$

Consider the mean $\mu(F) = \int \mathbf{x}F(d\mathbf{x})$. Because $\mu(\{w\delta_{\mathbf{y}} + F\}/\{1 + w\}) = \{w\mathbf{y} + \mu(F)\}/(1 + w)$ for the mixture in **3.1**, $D_{\mu,F}(\mathbf{y}) = 1$ for $\mathbf{y} = \mu(F)$ and $D_{\mu,F}(\mathbf{y}) = -\infty$ otherwise. Thus, the concept of depth function **3.1** is not useful for the mean $\mu(F)$. It seems reasonable to exclude such multivariate location estimators by considering only those $\theta(F)$ satisfying $D_{\theta,F}(\mathbf{y}) \geq 0$, or equivalently only those multivariate median θ satisfying **3.2**, as in **3.4**, **3.5**, **3.7** and **4.3**.

It is well known that the L_1-median is not affine equivariant, so that the L_1-depth in **4.2** is not affine invariant. It seems that we can make the L_1-median affine equivariant if a suitable data-dependent coordinate system is used, e.g., to replace $d_i(\mathbf{y}) = \|y - \mathbf{x}_i\|$ throughout by $\|A(\mathbf{y} - \mathbf{x}_i)\|$, where $A = V^{-1/2}$ for a suitable robust covariance-type matrix V based on data. See Chakraborty and Chaudhuri (13) for a similar approach. For general discussion of affine equivariant and related estimators of multivariate location and covariance, and their applications, we refer to Barnett (14); Brown and Hettmansperger (15); Donoho and Gasko (16); Eddy (17); Hettmansperger and Oja (18); Liu (19); Liu, Parelius, and Singh (20); Oja (21); Oja and Niinimaa (22); Rousseeuw (23); Rousseeuw and Leroy (24); Small (4); Tukey (9).

We thank J. De Leeuw and R. Vanderbei for references, and the National Security Agency and the National Science Foundation for grant support.

1. Weiszfeld, E. (1937) *Tôhoku Math. J.* **43,** 355–386.
2. Weber, A. (1909) *Über den Standort der Industrien*; English translation by Freidrich, C. J. (1929) *Alfred Weber's Theory of Location of Industries* Univ. of Chicago Press, Chicago).
3. Brown, B. (1983) *J. R. Stat. Soc. B* **45,** 25–30.
4. Small, C. G. (1990) *Int. Stat. Rev.* **58** (3), 263–277.
5. Kuhn, H. W. (1973) *Math. Program.* **4,** 98–107.
6. Simpson, R. (1750) *Doctrine and Application of Fluxions* (Printed for John Nourse, London).
7. Courant, R. & Robbins, H. (1941) *What Is Mathematics?* (Oxford Univ. Press, New York).
8. Chandrasekaran, R. & Tamir, A. (1989) *Math. Program.* **44,** 293–295.
9. Tukey, J. W. (1975) in *Proceedings of the International Congress of Mathematicians, Vancouver, 1974*, James, R. D. (Canadian Mathematical Congress), Vol. 2, pp. 523–531.
10. Möttönen, J. & Oja, H. (1995) *J. Nonparametr. Stat.* **5,** 201–213.
11. Marden, J. (1998) *Stat. Sinica* **8,** 813–826.
12. Chaudhuri, P. (1996) *J. Am. Stat. Assoc.* **91,** 862–872.
13. Chakraborty, B. & Chaudhuri, P. (1998) *J. R. Stat. Soc. B* **60,** 145–157.
14. Barnett, V. (1976) *J. R. Stat. Soc.* **A 139,** 319–354.
15. Brown, B. & Hettmansperger, T. (1989) *J. R. Stat. Soc.* **B 51,** 117–125.
16. Donoho, D. & Gasko, M. (1992) *Ann. Stat.* **20,** 1803–1827.
17. Eddy, W. (1982) *Convex Hull Peeling*, COMPSTAT, 42–47, eds. Caussinus, H., Ettinger, P. & Tomassone, R. (Physica, Vienna).
18. Hettmansperger, T. & Oja, H. (1992) *J. R. Stat. Soc. B* **56,** 235–249.
19. Liu, R. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 1732–1734.
20. Liu, R. Y., Parelius, J. M. & Singh, K. (1999) *Ann. Stat.* **27**, 783–858.
21. Oja, H. (1983) *Stat. Prob. Lett.* **1,** 327–332.
22. Oja, H. & Niinimaa, A. (1985) *J. R. Stat. Soc. B* **47,** 372–377.
23. Rousseeuw, P. J. (1984) *J. Am. Stat. Assoc.* **79,** 871–880.
24. Rousseeuw, P. J. & Leroy, A. M. (1987) *Robust Regression and Outlier Detection* (Wiley, New York).