Journal of Multivariate Analysis **69**, 135–153 (1999)
Article ID jmva.1998.1804, available online at http://www.idealibrary.com on IDEAL®

Halfspace Depth and Regression Depth Characterize the Empirical Distribution

Anja Struyf* and Peter J. Rousseeuw

Department of Mathematics and Computer Science, Universitaire Instelling Antwerpen, Universiteitsplein 1, B-2610 Antwerp, Belgium
http://win-www.uia.ac.be/u/statis/index.html

Received April 10, 1998; revised September 28, 1998

For multivariate data, the halfspace depth function can be seen as a natural and affine equivariant generalization of the univariate empirical cdf. For any multivariate data set, we show that the resulting halfspace depth function completely determines the empirical distribution. We do this by actually reconstructing the data points from the depth contours. The data need not be in general position. Moreover, we prove the same property for regression depth.

AMS 1991 subject classifications: 62G30, 62J05.

Key words and phrases: location depth, multivariate ranking, reconstruction algorithm, regression depth.

1. INTRODUCTION

Take any data set $X_n = \{\mathbf{x}_i;\ i = 1, ..., n\}$ with data points $\mathbf{x}_i = (x_{i1}, ..., x_{ip})' \in \mathbb{R}^p$. This data set determines an *empirical distribution* $\hat{P}_n$ which is the discrete probability distribution on $\mathbb{R}^p$ given by $\hat{P}_n(A) = \#\{\mathbf{x}_i \in A\}/n$. When the sample size n is given, $\hat{P}_n$ characterizes the data set X_n.

Tukey (1975) and Donoho and Gasko (1992) defined the *halfspace depth* of an arbitrary point $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)' \in \mathbb{R}^p$ relative to X_n as the smallest number of data points in any closed halfspace with boundary hyperplane through $\boldsymbol{\theta}$. We also call this the *location depth*, and it can be written as

$$ldepth(\boldsymbol{\theta}; X_n) = \min_{\|\mathbf{u}\| = 1} \#\{i;\ \mathbf{u}'\mathbf{x}_i \leqslant \mathbf{u}'\boldsymbol{\theta}\}, \tag{1.1}$$

where $\mathbf{u}$ ranges over all vectors in $\mathbb{R}^p$ with $\|\mathbf{u}\| = 1$. Interestingly, (1.1) is affine invariant. That is, if we consider a regular matrix $A \in \mathbb{R}^{p \times p}$ and some vector $\mathbf{b} \in \mathbb{R}^p$, it holds that

$$ldepth(A\boldsymbol{\theta} + \mathbf{b}; AX_n + \mathbf{b}) = ldepth(\boldsymbol{\theta}; X_n) \tag{1.2}$$

due to the fact that halfspaces are mapped to halfspaces.

* Aspirant Researcher with the FWO Belgium.

0047-259X/99 $30.00

Since (1.1) is defined for any $\boldsymbol{\theta} \in \mathbb{R}^p$ we call it the *depth function*. Its values are nonnegative integers. When $p = 1$ we have $u \in \{-1, 1\}$ so we can write (1.1) as

$$ldepth_1(\theta; X_n) = \min\{n\hat{F}_n(\theta; X_n), n\hat{F}_n(-\theta; -X_n)\}, \tag{1.3}$$

where $\hat{F}_n(\theta; X_n) = \#\{x_i \leqslant \theta\}/n$ is the usual empirical cdf. Figure 1(a) shows the depth function of a univariate data set with $n = 30$. The data values

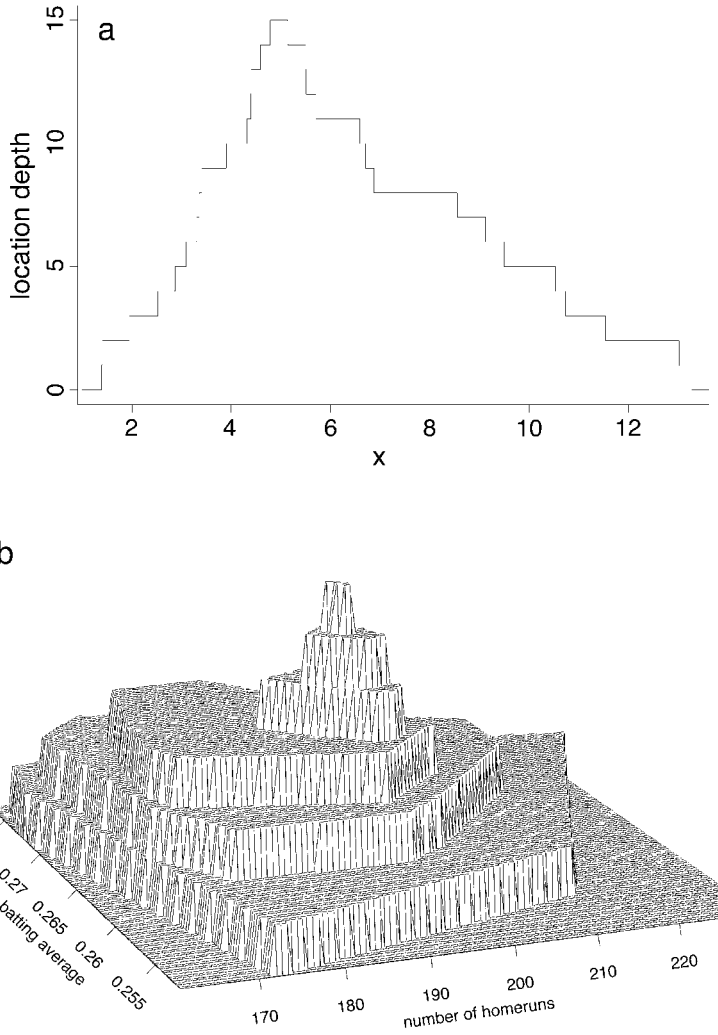

FIG. 1. Examples of a depth function for (a) $p = 1$; and (b) $p = 2$.

were randomly generated from a χ^2_6-distribution. The depth function clearly reflects the skewness. The increasing part of the function coincides with $\hat{F}_n$, whereas the decreasing part coincides with the empirical cdf of the image of the data under an affine transformation $x \to ax + b$ with $a < 0$. Figure 1(b) is the depth function of a bivariate data set ($p = 2$). The two variables are the batting average and the number of home runs of 14 baseball teams in the American League in 1987 (Moore and McCabe 1989; the data are also available at http://lib.stat.cmu.edu/DASL/). For any dimension $p \geqslant 2$ it holds that

$$ldepth(\boldsymbol{\theta}; X_n) = \min_{\|\mathbf{u}\| = 1} ldepth_1(\mathbf{u}'\boldsymbol{\theta}; \mathbf{u}'X_n), \tag{1.4}$$

which can also be written as

$$\min_{g \in \mathscr{A}} ldepth_1([\, g(\boldsymbol{\theta})\,]_1; [\, g(X_n)\,]_1) = \min_{g \in \mathscr{A}} n\hat{F}_n([\, g(\boldsymbol{\theta})\,]_1; [\, g(X_n)\,]_1),$$

where $\mathscr{A}$ is the set of all affine transformations of $\mathbb{R}^p$ and $[\, g(\boldsymbol{\theta})\,]_1$ denotes the first component of $g(\boldsymbol{\theta})$. Therefore, location depth can be seen as a natural affine equivariant generalization of the univariate empirical cdf. The usual multivariate empirical cdf is not affine equivariant because it depends on the coordinate system used.

The *depth contours* defined as

$$D_k = \{\boldsymbol{\theta} \in \mathbb{R}^p; ldepth(\boldsymbol{\theta}; X_n) \geqslant k\} \tag{1.5}$$

are convex, and $D_{k+1} \subseteq D_k$ for each k. The outermost contour D_1 is the convex hull of the data set. Each data set also has an innermost depth contour D_{k^*} where k^* is the maximum of the function $ldepth(\boldsymbol{\theta}; X_n)$ over all $\boldsymbol{\theta} \in \mathbb{R}^p$. Therefore, the complete set of contours is $D_{k^*} \subseteq D_{k^*-1} \subseteq \cdots \subseteq D_2 \subseteq D_1$. Figure 2 shows such a collection of contours. The depicted data set (from the Wall Street Journal of March 1, 1984, and provided at http://lib.stat.cmu.edu/DASL/) gives the 1983 TV advertising budget of several well-known companies, in millions of dollars. The second variable is based on a survey, where people had to cite a commercial for that product they had seen in the past week. The number of retained impressions (in millions) are plotted on the vertical axis. We see that the depth contours reflect the shape of the data set.

From definition (1.1) we can derive an equivalent expression for the kth depth contour:

LEMMA 1.

$$D_k = \bigcap_{A \in \mathscr{A}(n-k+1)} A,$$

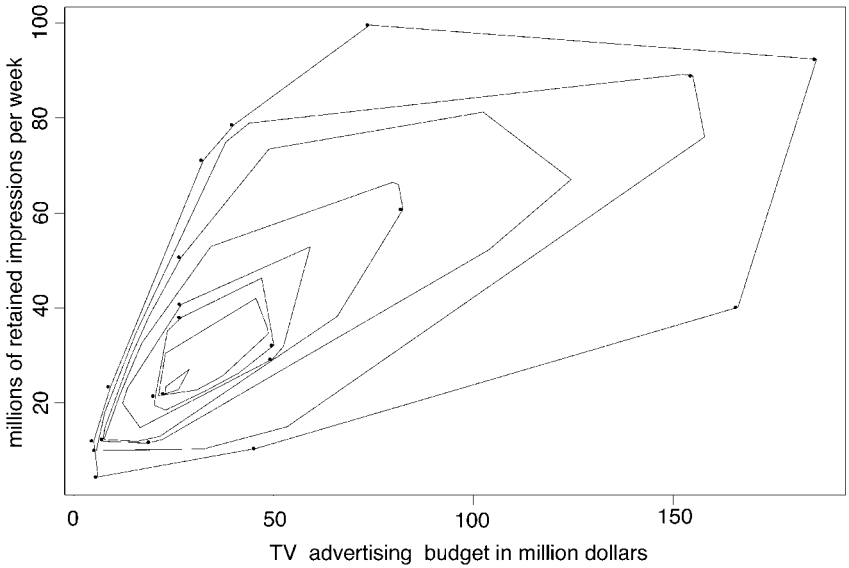

FIG. 2. Depth contours of the TV advertising data set.

where $\mathscr{A}(m)$ is the set of all closed halfspaces containing at least m data points.

Proof. We first prove $\subseteq$. If $\boldsymbol{\theta}$ does not belong to the intersection there exists a closed halfspace in $\mathscr{A}(n-k+1)$ which does not contain $\boldsymbol{\theta}$, hence $\boldsymbol{\theta}$ belongs to an open halfspace containing at most $n-(n-k+1)=k-1$ observations. Therefore $ldepth(\boldsymbol{\theta}; X_n) \leqslant k-1$, and thus $\boldsymbol{\theta} \notin D_k$. On the other hand, suppose that $\boldsymbol{\theta} \notin D_k$. Then $ldepth(\boldsymbol{\theta}; X_n) < k$, hence $\boldsymbol{\theta}$ belongs to a closed halfspace A containing fewer than k data points. The complement of A is an open halfspace containing at least $n-k+1$ data points, from which we immediately obtain a closed halfspace that does not contain $\boldsymbol{\theta}$ and has at least $n-k+1$ data points. Therefore also $\supseteq$ holds. ▌

Further properties about halfspace depth are given by Donoho and Gasko (1992) and Massé and Theodorescu (1994). Rousseeuw and Ruts (1996) constructed a fast algorithm to calculate (1.1) for a bivariate data set X_n. Based on this, Ruts and Rousseeuw (1996) developed an algorithm to compute the depth contours of X_n as in Fig. 2. The center of gravity of the innermost depth contour D_{k^*} is a multivariate generalization of the median, which is called the *deepest location* or the *Tukey median* of X_n. Rousseeuw and Ruts (1998) recently provided an algorithm for the bivariate Tukey median. The location depth of a point $\boldsymbol{\theta}$ measures how deep it lies inside the data cloud, and therefore it is sometimes called the

multivariate rank of $\boldsymbol{\theta}$ (Eddy 1985). Based on the halfspace depth, Rousseeuw and Ruts (1998) generalized the univariate boxplot, which is based on rank statistics, to the bivariate *bagplot*. The bagplot is a versatile graphical representation of a bivariate data set.

In (1.2)–(1.4) we have seen that the depth function is an affine equivariant generalization of the univariate empirical cdf. For instance, $ldepth(\boldsymbol{\theta}; X_n)$ depends on X_n in a global way whereas the data density is a local concept. Since the univariate ecdf characterizes the data, it would be interesting to know whether the depth function on $\mathbb{R}^p$ characterizes the data set as well. In this paper we will prove that the answer is affirmative:

THEOREM 1. *The empirical distribution of any data set $X_n \subset \mathbb{R}^p$ is uniquely determined by its halfspace depth function, i.e. the list of contours $\{D_1, ..., D_{k^*}\}$.*

An analogous property was already proved for the zonoid depth in (Koshevoy and Mosler 1997). Koshevoy (1997) proves the same property for the Oja depth (Oja 1983) and the simplicial depth (Liu 1990) when X_n is in general position. Together with the result of He and Wang (1997) that empirical depth contours converge to population depth contours, Theorem 1 suggests that one can use depth contours to understand distributional properties.

The proof of Theorem 1 will be given for a data set X_n of arbitrary dimension p, with data points in any position. Throughout this paper, the interior of a set A will be denoted as $\mathring{A}$ and its boundary as ∂A. We will often mention the dimension $\dim(C)$ of a convex set $C \in \mathbb{R}^p$, which is defined as the dimension of the affine span of C:

$$\begin{aligned}\text{affinespan}(C) &= \mathbf{c} + \text{linearspan}(C - \mathbf{c}) \qquad \text{with} \quad \mathbf{c} \in C \\ &= \left\{ \sum_{i=1}^{m} l_i \mathbf{x}_i;\ m \in \mathbb{N} \text{ and } \sum_{i=1}^{m} l_i = 1 \text{ and all } \mathbf{x}_i \in C \right\}.\end{aligned}$$

In Section 2 we prove Theorem 1 for data sets in general position, and then generalize this to arbitrary position in Section 3. Moreover, Section 3 gives an example showing that not every collection of nested convex contours originates from the halfspace depth function of an empirical distribution. Finally, Section 4 focuses on *regression depth*. This depth concept was introduced by Rousseeuw and Hubert (1996), who showed that its properties are similar to those of halfspace depth. In Section 4 we will prove that a property analogous to Theorem 1 also holds for regression depth.

2. PROOF FOR A DATA SET IN GENERAL POSITION

In this section we will prove Theorem 1 under the assumption that $X_n = \{\mathbf{x}_1, ..., \mathbf{x}_n\} \subset \mathbb{R}^p$ is in general position, i.e. that no p points lie in a $(p-1)$-dimensional affine subspace. In order to prove Theorem 1, we need some auxiliary results.

LEMMA 2. *Any data point* $\mathbf{x}_i$ *with* $ldepth(\mathbf{x}_i; X_n) = k$ *must be a vertex of* D_k.

Proof. Suppose that $\mathbf{x}_i$ is not a vertex of D_k. By definition (1.5), $\mathbf{x}_i$ must belong to D_k and hence only two possiblities are left:

(i) $\mathbf{x}_i \in \mathring{D}_k$. Because $ldepth(\mathbf{x}_i; X_n) = k$ there exists a closed halfspace A_H with boundary hyperplane H through $\mathbf{x}_i$ containing exactly k data points. Since $\mathbf{x}_i \in \mathring{D}_k$ we can shift H over a distance $\varepsilon \neq 0$, such that $A_{\tilde{H}} \subset A_H$ and $\tilde{H}$ has at least one point $\tilde{\mathbf{x}}$ in common with D_k. Because $A_{\tilde{H}}$ does not contain $\mathbf{x}_i$ it contains fewer than k data points, hence $ldepth(\tilde{\mathbf{x}}; X_n) < k$ which is impossible.

(ii) $\mathbf{x}_i \in \partial D_k$ but $\mathbf{x}_i$ is not a vertex of D_k. When H determines a closed halfspace A_H through $\mathbf{x}_i$ containing exactly k data points, H cannot contain the entire edge of D_k to which $\mathbf{x}_i$ belongs. Otherwise, we could consider an affine subspace l of H which does not contain $\mathbf{x}_i$ but passes through one of the vertices of D_k on the same edge. (For instance, l can be a point.) Next, we rotate H around l without passing any data points, such that $\tilde{H}$ still contains that vertex, but $A_{\tilde{H}}$ does not contain $\mathbf{x}_i$ anymore. Therefore $A_{\tilde{H}}$ contains fewer data points than A_H and the vertex of D_k would have depth $< k$, which is impossible. Therefore, the closed halfspace A_H will cut at least one vertex off D_k, and by shifting H to $\tilde{\tilde{H}}$ which passes through this vertex we can prove that this vertex must have depth less than k, which is also impossible.

Since both situations (i) and (ii) are impossible, $\mathbf{x}_i$ must be a vertex of D_k. ∎

LEMMA 3. *If* $\mathbf{x}_i$ *is a data point with* $ldepth(\mathbf{x}_i; X_n) = k$, *then there exists a closed halfspace* A *with boundary through* $\mathbf{x}_i$ *such that* $A \cap D_k = \{\mathbf{x}_i\}$, *and* A *contains exactly* k *data points.*

Proof. From Lemma 2 we already know that $\mathbf{x}_i$ must be a vertex of D_k. Suppose that all the closed halfspaces A with boundary through $\mathbf{x}_i$ and $A \cap D_k = \{\mathbf{x}_i\}$ contain at least $k+1$ data points. However, because $\mathbf{x}_i \in D_k$ and $\mathbf{x}_i \notin D_{k+1}$ there must exist a closed halfspace with boundary through $\mathbf{x}_i$ that contains exactly k data points. If A_H is such a halfspace with boundary hyperplane H, then H must fulfil one of the following two conditions:

(i) H is a limiting hyperplane of D_k and $\dim(H \cap D_k) \geqslant 1$. Take any vertex $\tilde{\mathbf{x}} \neq \mathbf{x}_i$ of D_k which lies in H. In H exists a $(p-2)$-dimensional affine subspace l containing $\tilde{\mathbf{x}}$, which does not pass through $\mathbf{x}_i$. At least one of the closed halfhyperplanes H^+ and H^- formed by l does not contain all of the data points in H. Then we rotate H around l to $\tilde{H}$ without passing any data points, such that the number of data points in $A_{\tilde{H}}$ is strictly smaller than the number of data points in A_H. In other words, we can "avoid" some of the data points on H. Because $A_{\tilde{H}}$ must contain at least k data points ($\tilde{\mathbf{x}} \in D_k$), A_H must contain more than k data points, which is a contradiction because we had chosen A_H such that it contained exactly k data points.

(ii) H passes through the interior of D_k. We can make the halfspace A_H smaller by shifting H over a distance ε to $\tilde{H}$, where $0 < |\varepsilon| < \min_{\mathbf{x}_j \notin H} d(H, \mathbf{x}_j)$ and such that $\tilde{H}$ still has a point $\tilde{\mathbf{x}}$ in common with D_k. Because $ldepth(\tilde{\mathbf{x}}; X_n) \geqslant k$ this $A_{\tilde{H}}$ will contain at least k data points. We also know that A_H contains at least one more data point (the point $\mathbf{x}_i$) than $A_{\tilde{H}}$ and therefore A_H contains at least $k+1$ data points, which is again a contradiction.

Neither (i) or (ii) are possible, and therefore the lemma is proved. ▮

Next, we will prove that for a data set in general position the depth contour D_k lies completely within the interior of D_{k-1} (for every $k \leqslant k^*$). This property is illustrated in Fig. 2, where every depth contour in the plot is strictly contained in all larger contours.

LEMMA 4. *Consider a data set $X_h \subset \mathbb{R}^p$ in general position with $h \leqslant p+1$. For any point $\mathbf{x} \in \mathbb{R}^p$ there exists a closed halfspace A which contains at most* 1 *data point and such that its boundary hyperplane ∂A passes through* $\mathbf{x}$.

Proof. The h points form a unique $(h-1)$-dimensional simplex $S = conhull(X_h)$ in $\mathbb{R}^p$, with vertices equal to the h data points. In the case that $\mathbf{x}$ equals one of the vertices of S we can separate that data point from the others. When $\mathbf{x} \notin S$ we can of course find a hyperplane $H_{\mathbf{x}}$ which separates $\mathbf{x}$ from S. When $\mathbf{x} \in \mathring{S}$ we choose a vertex $\mathbf{x}_i$ from S. Then we can find a hyperplane H_i through the other $h-1$ data points, such that $\mathbf{x}$ lies strictly between $\mathbf{x}_i$ and H_i. Then $H_{\mathbf{x}}$ should be chosen parallel to H_i. Finally, when $\mathbf{x} \in \partial S$ but $\mathbf{x} \neq \mathbf{x}_i$ for all i, we use induction on the dimension (for $p = 1$, the lemma is trivial). Let H_i be a hyperplane containing all vertices except for $\mathbf{x}_i$ such that $\mathbf{x} \in H_i$. By induction, we can find a $(p-2)$-dimensional affine subspace $l_{\mathbf{x}} \subset H_i$ through $\mathbf{x}$ that separates a vertex $\mathbf{x}_j$ from the other vertices in H_i. The hyperplane through $l_{\mathbf{x}}$ and the midpoint of $[\mathbf{x}_j, \mathbf{x}_i]$ then separates $\mathbf{x}_j$ from all other data points. ▮

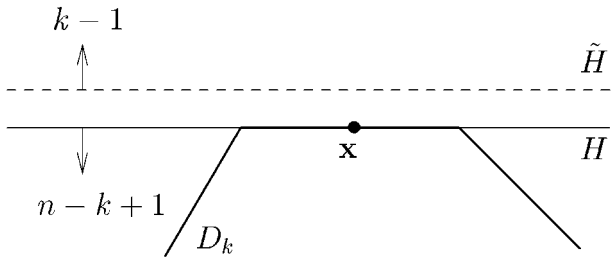

FIG. 3. Illustration of the proof of Lemma 5.

LEMMA 5. *For all values k of the depth function, it holds that*

$$\text{if} \quad D_k \subset \mathring{D}_{k-1} \qquad \text{then} \quad ldepth(\mathbf{x}; X_n) = k \qquad \text{for all} \quad \mathbf{x} \in \partial D_k.$$

Proof. Take a point $\mathbf{x} \in \partial D_k$. Due to lemma 1, $\mathbf{x}$ lies on a hyperplane H for which the closed halfspace $\tilde{A}_H$ with boundary H contains at least $n-k+1$ data points. Therefore $\mathring{A}_H := \mathbb{R}^p \backslash \tilde{A}_H$ contains at most $k-1$ data points. This is illustrated in Fig. 3. Suppose that $\mathring{A}_H$ contains fewer than $k-1$ points, for example $k-2$ points. We know that $D_k \subset \mathring{D}_{k-1}$ and hence we can shift H over a distance ε with $0 < |\varepsilon| < \min_{\mathbf{x}_i \notin H} d(\mathbf{x}_i, H)$, yielding $\tilde{H}$, and such that $\tilde{H}$ intersects $D_{k-1} \backslash D_k$. The closed halfspace $A_{\tilde{H}}$ then contains the same $k-2$ data points as $\mathring{A}_H$. However, this is impossible because $A_{\tilde{H}}$ passes through points with depth $k-1$. Hence $\mathring{A}_H$ must contain exactly $k-1$ data points.

Because H contains at most $p+1$ data points, we can find a $(p-2)$-dimensional affine subspace $l \subset H$ through $\mathbf{x}$ which separates at most 1 data point from the others in H (Lemma 4). When we rotate H around l without passing any data points we find a halfspace $A_{\tilde{H}}$ containing at most $k-1+1=k$ data points, and hence $ldepth(\mathbf{x}; X_n)$ must equal k because $ldepth(\mathbf{x}; X_n) \geqslant k$. ∎

COROLLARY 1. *For all values k of the depth function, it holds that*

$$\text{if} \quad D_k \subset \mathring{D}_{k-1} \qquad \text{and} \qquad k < k^* \qquad \text{then} \quad \mathring{D}_k \neq \varnothing.$$

Proof. Because $k < k^*$ there must be a point with depth $> k$. This point must belong to D_k and cannot lie on its boundary (Lemma 5), hence it must be in the interior of D_k. ∎

LEMMA 6. *For any value $k < k^*$ of the halfspace depth function, it holds that*

$$D_{k+1} \subset \mathring{D}_k.$$

Proof. The result holds for $k=0$ since $D_1 \subset \mathring{D}_0 = \mathbb{R}^p$. Let us now consider $k=1$. When $n>p$, we know that the interior of the convex hull D_1 of the data set cannot be empty. When $n \leqslant p$, Lemma 4 shows that $k^* \leqslant 1$, hence $k=1$ is impossible.

We can now use induction on k. Invoking Lemma 5 proves that any point on the boundary of D_k has depth equal to k, hence $D_{k+1} \subset \mathring{D}_k$. ∎

Now we are ready to formulate the main proposition on which Theorem 1 is based. This allows us to actually identify the data points. From Lemma 2 we know that every data point must be a vertex of a depth contour. Assume that all the data points on $D_1, ..., D_{k-1}$ for $k>1$ are already identified. The remaining question is whether the vertex $\mathbf{x}$ of D_k is a data point or not. Define the set

$$S_{\mathbf{x}}^k = \left(\bigcup_{i=1}^{k-1} \{\tilde{\mathbf{x}};\, \tilde{\mathbf{x}} \in X_n \text{ and } ldepth(\tilde{\mathbf{x}};\, X_n) = i\} \right) \cup \{\tilde{\mathbf{x}};\, \tilde{\mathbf{x}} \text{ is a vertex of } D_k\} \setminus \{\mathbf{x}\}. \tag{2.1}$$

PROPOSITION 1. *For any vertex* $\mathbf{x}$ *of* D_k *it holds that*

$$\mathbf{x} \text{ is a data point} \Leftrightarrow ldepth(\mathbf{x};\, S_{\mathbf{x}}^k) < k = ldepth(\mathbf{x};\, X_n).$$

Proof. Let us first prove $\Rightarrow$. Let $\mathbf{x}$ be a data point on ∂D_k. From Lemma 6 it follows that $ldepth(\mathbf{x};\, X_n) = k$, and hence Lemma 3 implies that there exists a closed halfspace A_H with boundary H through $\mathbf{x}$ which contains exactly k data points and such that $A_H \cap D_k = \{\mathbf{x}\}$. Because all data points lying outside D_k are already in $S_{\mathbf{x}}^k$ we know that all data points in A_H except for $\mathbf{x}$ are in $S_{\mathbf{x}}^k$, and so

$$\#\{\mathbf{x}_i;\, \mathbf{x}_i \in S_{\mathbf{x}}^k \cap A_H\} = k-1,$$

and this implies that

$$ldepth(\mathbf{x};\, S_{\mathbf{x}}^k) \leqslant k-1 < k.$$

Now we prove $\Leftarrow$. Let $ldepth(\mathbf{x};\, S_{\mathbf{x}}^k) < k$. Suppose that $\mathbf{x}$ is not a data point. We know there is a closed halfspace A_H with boundary H through $\mathbf{x}$ that contains fewer than k points in $S_{\mathbf{x}}^k$. Because $ldepth(\mathbf{x};\, X_n) = k$, this A_H contains at least k data points. Two situations can occur:

(i) $A_H \cap D_k = \{\mathbf{x}\}$. Then A_H contains at least k data points, which all lie on depth contours with depth $<k$, and these are all in $S_{\mathbf{x}}^k$. Therefore $A_H \cap S_{\mathbf{x}}^k$ contains at least k points, which is impossible due to the choice of A_H.

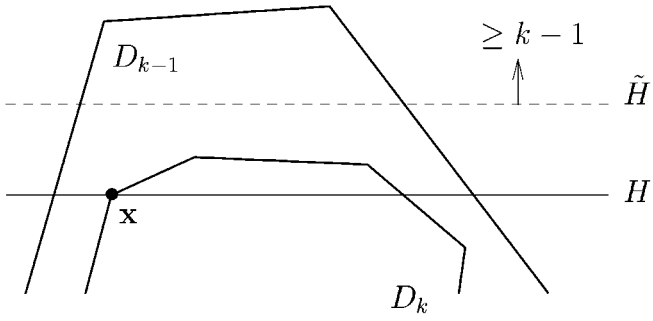

FIG. 4. Illustration of the proof of Proposition 1.

(ii) $\#(A_H \cap D_k) > 1$. Lemma 6 yields $D_k \subset \mathring{D}_{k-1}$. We can make A_H smaller by shifting H to $\tilde{H}$ in the direction of $H^{\perp}$ such that $\tilde{H}$ does not intersect D_k but still passes through at least one point of D_{k-1}, as in Fig. 4. This $A_{\tilde{H}}$ necessarily contains $k-1$ or more data points, which are also included in $S^k_{\mathbf{x}}$ because they all lie on depth contours with depth smaller than k. Because D_k is convex and H passes through $\mathbf{x}$ and at least one other point of D_k, it follows that A_H must contain at least one vertex of D_k different from $\mathbf{x}$. Therefore, by shifting H to $\tilde{H}$ we excluded at least one point in $S^k_{\mathbf{x}}$ from $A_{\tilde{H}}$. Hence A_H contains at least k points of $S^k_{\mathbf{x}}$ which again is a contradiction.

Since both (i) and (ii) are impossible, $\mathbf{x}$ must be a data point. ▮

For a data set in general position, we can now easily show that the depth function uniquely determines the data set.

Proof of Theorem 1. In Lemma 2 we saw that every data point is a vertex of one of the depth contours. It only remains to prove that we can distinguish between those vertices which are data points and those which are not. Because the depth contour D_1 is the convex hull of the data set, we know that any vertex of D_1 must be a data point. This also yields the set $S^k_{\mathbf{x}}$ defined in (2.1) for $k=2$. By sequentially applying Proposition 1 to increasing k, we identify all data points on subsequent contours. Finally, this yields all data points and hence the empirical distribution of X_n. ▮

The above proof is constructive since it amounts to an algorithm that reconstructs the original data set from the depth contours. We have actually implemented it as a program, for an additional verification of our results.

3. PROOF FOR A DATA SET IN ARBITRARY POSITION

We first observe that the proofs of Lemmas 2 and 3 remain valid for a data set X_n in arbitrary position. But when X_n is in non-general position,

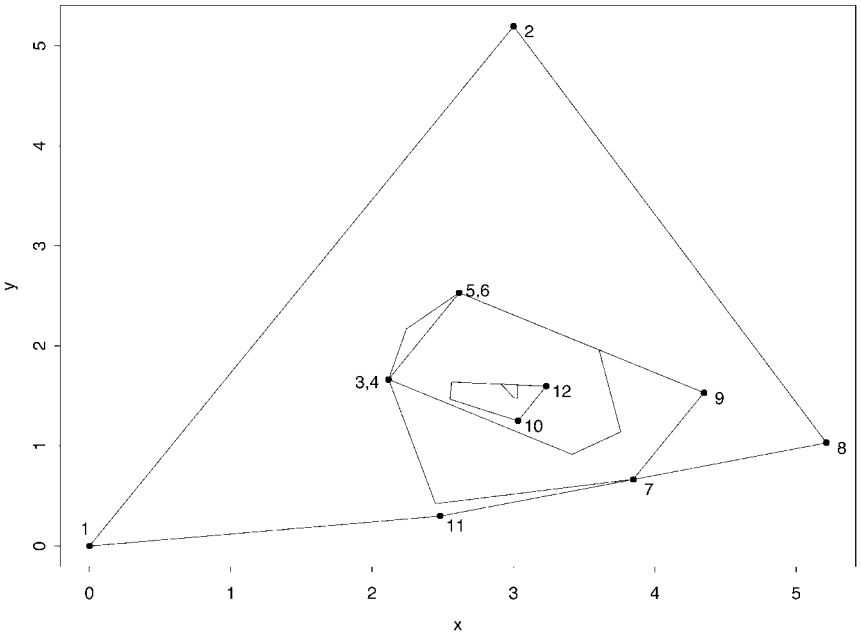

FIG. 5. Depth contours of a generated data set ($n = 12$) in non-general position. Several points are collinear or even coincident.

a depth contour is not necessarily contained in the interior of the previous contour. The generated data set in Fig. 5 illustrates this. This data set consists of 12 data points, of which several are collinear. Moreover, points 3 and 4 are coincident, as well as points 5 and 6.

First, let us assume that the data set X_n is in arbitrary position except that no two data points may coincide. Let $\mathbf{x}$ be a data point with depth k, and assume that all data points on $D_1, ..., D_{k-1}$ with depth smaller than k are already identified. Again take $S^k_{\mathbf{x}}$ as in (2.1).

We can repeat the formulation of Proposition 1, except that we now have to specify that the vertex $\mathbf{x}$ of D_k must have $ldepth(\mathbf{x}; X_n) = k$. In general position this was trivial because $D_k \subset \mathring{D}_{k-1}$ and hence a vertex of contour D_k always had depth k.

PROPOSITION 2. *For any vertex* $\mathbf{x}$ *of* D_k *which has* $ldepth(\mathbf{x}; X_n) = k$ *it holds that*

$$\mathbf{x} \text{ is a data point} \Leftrightarrow ldepth(\mathbf{x}; S^k_{\mathbf{x}}) < k = ldepth(\mathbf{x}; X_n).$$

Proof. When $\mathbf{x}$ is a data point with depth k, we can apply the same reasoning as in Proposition 1 to prove that $ldepth(\mathbf{x}; S^k_{\mathbf{x}}) < k$. It remains to

prove that $ldepth(\mathbf{x}; S^k_{\mathbf{x}}) < k$ implies that $\mathbf{x}$ is a data point. Suppose that $\mathbf{x}$ is not a data point. There exists a closed halfspace A_H with boundary H through $\mathbf{x}$ containing fewer than k points of $S^k_{\mathbf{x}}$. This A_H also contains at least k data points since $ldepth(\mathbf{x}; X_n) = k$. Three different situations can occur:

(i) $A_H \cap D_k = \{\mathbf{x}\}$. The k data points in A_H all lie outside of D_k and thus all belong to $S^k_{\mathbf{x}}$. Therefore $A_H \cap S^k_{\mathbf{x}}$ contains at least k points, which is impossible due to the choice of A_H.

(ii) H is a limiting hyperplane of D_k and $\dim(H \cap D_k) = \dim(A_H \cap D_k) \geq 1$. Denote $l = H \cap D_k$ which is a convex region of dimension $h \geq 1$. Because all data points in $\mathring{A}_H$ also belong to $S^k_{\mathbf{x}}$ we know that $\mathring{A}_H$ contains $k_H < k$ data points. Moreover, every vertex $\tilde{\mathbf{x}}$ of l has depth $\geq k$ in X_n, and therefore must have $ldepth(\tilde{\mathbf{x}}; H \cap X_n) \geq k - k_H$ (otherwise we could rotate H a little without passing data points such that the new halfspace would contain fewer than k data points). Because l is a convex region, at least $2(k - k_H - 1)$ data points lying outside of D_k must belong to $H \cap S^k_{\mathbf{x}}$. Moreover, l has at least $h+1$ vertices, of which at least h are also contained in $S^k_{\mathbf{x}}$ (all vertices except for $\mathbf{x}$). Therefore A_H contains at least $k_H + 2(k - k_H - 1) + h$ points of $S^k_{\mathbf{x}}$ as in Fig. 6, while the assumption says that A_H contains fewer than k points in $S^k_{\mathbf{x}}$. Therefore

$$k_H + 2(k - k_H - 1) + h < k \Leftrightarrow k < k_H - h + 2 \Rightarrow k \leqslant k_H$$

which is in contradiction with $k_H < k$. Therefore, this situation cannot occur.

(iii) $A_H \cap \mathring{D}_k \neq \varnothing$. We can make A_H smaller by shifting H to $\tilde{H}$ which is a limiting hyperplane of D_k. Because A_H contains fewer than k points in $S^k_{\mathbf{x}}$, also $A_{\tilde{H}}$ will. The hyperplane $\tilde{H}$ can take on two different positions relative to D_k.

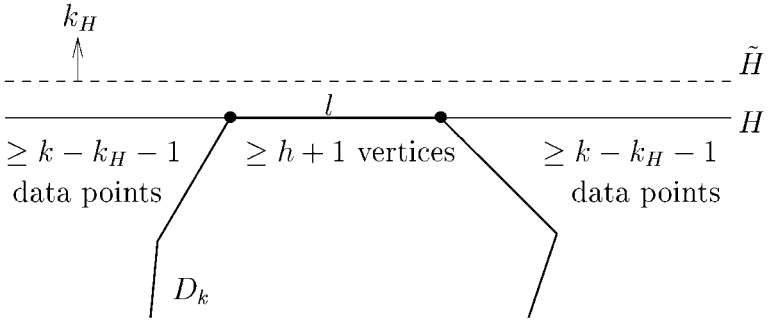

FIG. 6. Illustration of the proof of Proposition 2.

(a) $A_{\tilde{H}} \cap D_k = \{\tilde{\mathbf{x}}\}$. Here $\tilde{\mathbf{x}}$ is a vertex of D_k which differs from $\mathbf{x}$, and hence $\tilde{\mathbf{x}} \in S^k_{\mathbf{x}}$. Therefore, every data point in $X_n \cap A_{\tilde{H}}$ also belongs to $S^k_{\mathbf{x}}$ and hence $ldepth(\tilde{\mathbf{x}}; X_n) < k$, which is a contradiction because $\tilde{\mathbf{x}} \in D_k$.

(b) $\tilde{H}$ is a limiting hyperplane of D_k and $\dim(\tilde{H} \cap D_k) = \dim(A_{\tilde{H}} \cap D_k) \geqslant 1$. Denote $\tilde{l} = \tilde{H} \cap D_k$ which is of dimension $h \geqslant 1$. As in part (ii) of this proof, we can deduce that $A_{\tilde{H}}$ must contain at least $k_{\tilde{H}} + 2(k - k_{\tilde{H}} - 1) + h + 1$ points in $S^k_{\mathbf{x}}$ where $k_{\tilde{H}} < k$ (we do not need to subtract one vertex of $\tilde{l}$ because $\mathbf{x} \notin \tilde{l}$). We already knew that $A_{\tilde{H}}$ contains fewer than k points in $S^k_{\mathbf{x}}$ hence

$$k_{\tilde{H}} + 2(k - k_{\tilde{H}} - 1) + h + 1 < k \Leftrightarrow k < k_{\tilde{H}} - h + 1 \Rightarrow k < k_{\tilde{H}}$$

which is again in contradiction with $k_{\tilde{H}} < k$.

We have seen that the assumption that $\mathbf{x}$ is not a data point leads to a contradiction in all three cases (i)–(iii). Therefore $\mathbf{x}$ must be a data point, which proves the proposition. ▮

Proposition 2 thus extends Theorem 1 to a data set X_n in arbitrary position which does not contain any multiple points. However, a modified version of Proposition 2 will still apply when such points exist. In that case it is important to keep all copies of a multiple point in X_n which implies that we have to think of X_n as a *multiset* instead of a set (this terminology is used e.g. in Edelsbrunner 1987, p. 220). The following lemma makes it possible to prove Theorem 1 for multisets:

LEMMA 7. *If X_n contains m copies of a point, then that point will be a vertex of m subsequent depth contours.*

Proof. Suppose w.l.o.g. that $\mathbf{x}_1 = \mathbf{x}_2 = \cdots = \mathbf{x}_m$ and denote this point as $\mathbf{x}$. Put $k := ldepth(\mathbf{x}; X_n)$. Lemma 2 implies that $\mathbf{x}$ is a vertex of D_k. Lemma 3 tells us that there exists a hyperplane H through $\mathbf{x}$ such that $A_H \cap D_k = \{\mathbf{x}\}$ and A_H contains exactly k data points. When we reduce A_H to $A_{\tilde{H}}$ by shifting H over an arbitrary small distance $0 < |\varepsilon| < \min_{\mathbf{x}_j \neq \mathbf{x}} d(\mathbf{x}_j, H)$, we observe that $A_{\tilde{H}}$ contains only $k - m$ data points. Therefore, all points in $\mathring{A}_H$ have depth $\leqslant k - m$. Also the points $\tilde{\mathbf{x}}$ on $H \backslash \{\mathbf{x}\}$ have depth at most $k - m$ (rotate H a little around an affine plane through $\tilde{\mathbf{x}}$ without passing any data points, to exclude $\mathbf{x}$ and to get a halfspace containing $k - m$ data points). Therefore, none of the depth contours $D_{k-m+1}, \ldots, D_{k-1}$ can contain any other point of H than $\mathbf{x}$. Because they will also contain the complete contour D_k they must all have $\mathbf{x}$ as a vertex. ▮

Lemma 7 implies that an m-fold data point $\mathbf{x}_i$ with depth k is a vertex of the contours $D_{k-m+1}, \ldots, D_k$. Let us now consider an arbitrary vertex $\mathbf{x}$

of D_k with $ldepth(\mathbf{x}; X_n) = k$, and try to determine whether it is a data point or not. We generalize the definition of $S^k_{\mathbf{x}}$ by including all vertices $\tilde{\mathbf{x}}$ of $D_{k'}$ (where $k' < k$) which have $ldepth(\tilde{\mathbf{x}}; S^{k'}_{\tilde{\mathbf{x}}}) < k'$. Like X_n, also $S^k_{\mathbf{x}}$ becomes a multiset: when the same point is added m times, it has to be considered as being present m times in $S^k_{\mathbf{x}}$. Therefore the new definition of $S^k_{\mathbf{x}}$ is

$$S^k_{\mathbf{x}} = \Bigg(\biguplus_{i=1}^{k-1} \{\tilde{\mathbf{x}};\ \tilde{\mathbf{x}} \in X_n \text{ and } ldepth(\tilde{\mathbf{x}}; X_n) = i\} \uplus \biguplus_{j=1}^{k-1} \{\tilde{\mathbf{x}};\ ldepth(\tilde{\mathbf{x}}; X_n) \geq k \text{ and } ldepth(\tilde{\mathbf{x}}; S^j_{\tilde{\mathbf{x}}}) < j\} \uplus \{\tilde{\mathbf{x}};\ \tilde{\mathbf{x}} \text{is a vertex of } D_k\} \Bigg) \setminus \{\mathbf{x}\}. \tag{3.1}$$

Here $\uplus$ denotes the union of multisets, and $\setminus\{\mathbf{x}\}$ means that we delete all occurences of $\mathbf{x}$ from the multiset. For a data set X_n without multiple points, (3.1) reduces to (2.1). Using this new definition, we can now identify a multiple point by its exact depth relative to $S^k_{\mathbf{x}}$. The most general version of the proposition then becomes:

PROPOSITION 3. *Any vertex* $\mathbf{x}$ *of* D_k *with* $ldepth(\mathbf{x}; X_n) = k$ *occurs exactly* $m := k - ldepth(\mathbf{x}; S^k_{\mathbf{x}})$ *times in* X_n. *Here* $S^k_{\mathbf{x}}$ *is given by* (3.1). *Note that the integer* m *is zero iff* $\mathbf{x}$ *does not belong to* X_n.

Proof. As in the proof of Proposition 1 we find that for any m-fold data point with depth k on D_k there exists a halfspace A_H which contains exactly k data points and hence contains $k-m$ points of $S^k_{\mathbf{x}}$. Now suppose that another halfspace through $\mathbf{x}$ contains fewer than $k-m$ points of $S^k_{\mathbf{x}}$. Then this halfspace (or, if necessary, a shifted version which does not pass through the interior of D_k) contains fewer than k data points although it passes through at least 1 point on the boundary of D_k. So we conclude that $ldepth(\mathbf{x}; S^k_{\mathbf{x}}) = k - m$.

From Proposition 2 we know that any vertex of D_k with $ldepth(\mathbf{x}; S^k_{\mathbf{x}}) = k - m < k$ and $ldepth(\mathbf{x}; X_n) = k$ is a data point. It then must be an m-fold data point due to the first part of this proof, since any m'-fold data point should have depth $k - m'$ relative to $S^k_{\mathbf{x}}$. ∎

In conclusion, Proposition 3 proves Theorem 1 for any data set in arbitrary position. Therefore, the halfspace depth function characterizes the underlying empirical distribution. As in the general position case, the proof can again be written as an algorithm.

Note that not every collection of nested convex contours can be interpreted as a halfspace depth function. Consider the contours $\{D_1, D_2\}$ in Fig. 7. Clearly, all vertices of D_1 must be data points. Therefore the point $\mathbf{x}$ should have $ldepth(\mathbf{x}; X_n) \geq 2$, which is contradicted by its position

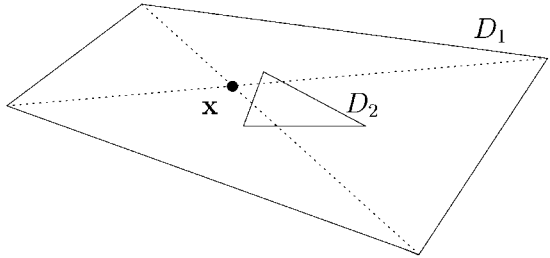

FIG. 7. Not every set of nested convex contours originates from a depth function.

outside D_2. In conclusion, $\{D_1, D_2\}$ cannot be the depth contours of any data set.

4. REGRESSION DEPTH

The *regression depth* (Rousseeuw and Hubert 1996) of a hyperplane measures how well that hyperplane fits a given data set $Z_n = \{\mathbf{z}_i = (x_{i1}, ..., x_{i,\,p-1}, y_i);\ i = 1, ..., n\} \subset \mathbb{R}^p$. Let a hyperplane $H_{\boldsymbol{\theta}} \subset \mathbb{R}^p$ be given by $y_i = \theta_1 x_{i1} + \cdots + \theta_{p-1} x_{i,\,p-1} + \theta_p$ where $\boldsymbol{\theta} \in \mathbb{R}^p$. The regression depth of $H_{\boldsymbol{\theta}}$ relative to the data set Z_n is defined as the smallest number of observations whose residual $r_i = y_i - \theta_1 x_{i1} - \cdots - \theta_{p-1} x_{i,\,p-1} - \theta_p$ needs to change sign to make $H_{\boldsymbol{\theta}}$ a *nonfit*. We call $H_{\boldsymbol{\theta}}$ a nonfit if there exists an affine hyperplane V in $\mathbf{x}$-space such that no $\mathbf{x}_i$ belongs to V, and such that $r_i > 0$ for all $\mathbf{x}_i$ in one of its open halfspaces and $r_i < 0$ for all $\mathbf{x}_i$ in the other open halfspace.

Figure 8 gives an example of a nonfit $H_{\boldsymbol{\theta}}$ in a three-dimensional data set Z. Here $\mathbf{x}$-space is the horizontal plane $y \equiv 0$, and the line V separates observations with positive and negative residuals. Note that $H_{\boldsymbol{\theta}}$ is called a nonfit because it can be tilted (rotated) around the line L in Fig. 8 until it becomes the vertical plane through V, without passing any observation. In this sense $H_{\boldsymbol{\theta}}$ is equivalent to the vertical plane, which is not a fit because it cannot be written in the form $\mathbf{y} = \theta_1 \mathbf{x}_1 + \theta_2 \mathbf{x}_2 + \theta_3$.

In words, the regression depth of a fit measures how far away it is from any nonfit. Therefore, a fit with large depth is well-balanced relative to the data, which is a good thing. Rousseeuw and Hubert (1996) constructed an algorithm to compute the exact regression depth of a line relative to a two-dimensional data set Z in $O(n \log n)$ time. Rousseeuw and Struyf (1998) constructed exact and approximate algorithms for higher dimensions.

Regression depth can equivalently be defined in the dual space, which is the set of all possible $\boldsymbol{\theta}$. To construct the dual space, each fit $H_{\boldsymbol{\theta}}$ is mapped to the point $D(H_{\boldsymbol{\theta}}) = \boldsymbol{\theta} \in \mathbb{R}^p$, and each data point $\mathbf{z}_i$ is mapped to the hyperplane $D(\mathbf{z}_i) := H_i$ given by $\theta_p = -x_{i1}\theta_1 - ... - x_{i,\,p-1}\theta_{p-1} + y_i$. This

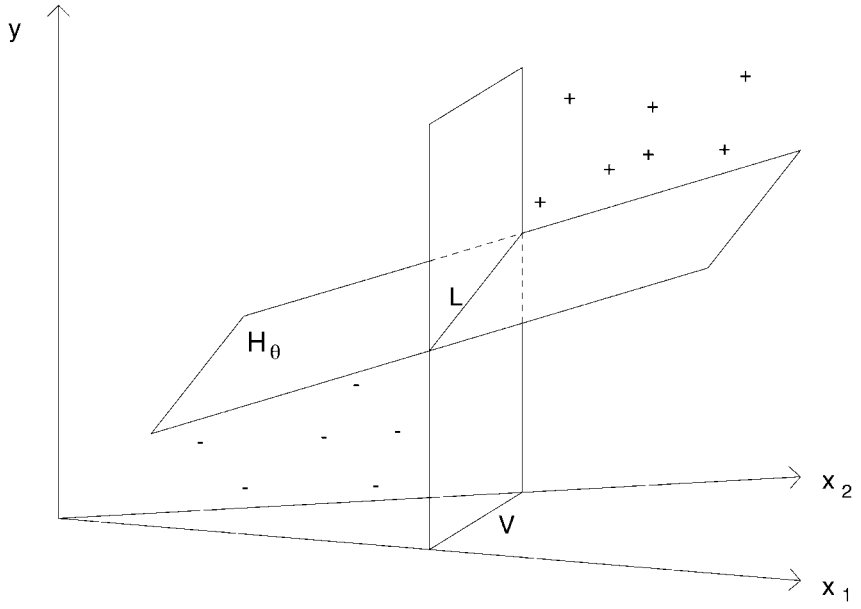

FIG. 8. A regression nonfit $H_{\boldsymbol{\theta}}$ in three dimensions.

definition ensures that a point $\mathbf{z}$ lying below/on/above a hyperplane H in primal space corresponds to a hyperplane $D(\mathbf{z})$ below/on/above the point $D(H)$ in dual space. Hence, a residual r_j changing sign in primal space corresponds to a hyperplane H_j in dual space moving from one side of $\boldsymbol{\theta}$ to the other. The point $\boldsymbol{\theta}$ corresponds to a nonfit $H_{\boldsymbol{\theta}}$ iff there exists a line $\langle \boldsymbol{\theta}, \boldsymbol{\theta} + \mathbf{u} \rangle$ (where $\mathbf{u}$ corresponds to a hyperplane V in primal space) that cuts all hyperplanes H_i on the same side of $\boldsymbol{\theta}$ (all residuals on the same side of V have the same sign). In general, $rdepth(\boldsymbol{\theta}; Z_n)$ is the smallest number of hyperplanes H_i that need to be removed to set $\boldsymbol{\theta}$ free. This means that we look for a direction $\mathbf{u}$ with $\|\mathbf{u}\| = 1$ (not parallel to any of the hyperplanes H_i) for which the halfline $[\boldsymbol{\theta}, \boldsymbol{\theta} + \mathbf{u}\rangle$ intersects the fewest hyperplanes H_i. (We assume throughout that a line parallel to a hyperplane H intersects H at infinity.)

Figure 9 illustrates regression depth in the primal and the dual. Figure 9(a) shows a two-dimensional data set of 6 observations in primal space. Two nonfits $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are indicated with their respective tilting points $v_{\boldsymbol{\theta}}$ and $v_{\boldsymbol{\eta}}$, i.e. the $\mathbf{x}$-coordinates at which they can be rotated to vertical lines. The fit $\boldsymbol{\xi}$ has regression depth 2 (we can remove e.g. points 4 and 5 to obtain a nonfit with tilting point $v_{\boldsymbol{\theta}}$). The dual plot is shown in Fig. 9(b). We clearly see that the nonfits $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are in unbounded regions of the arrangement of hyperplanes, and that two hyperplanes (e.g. 4 and 5) have

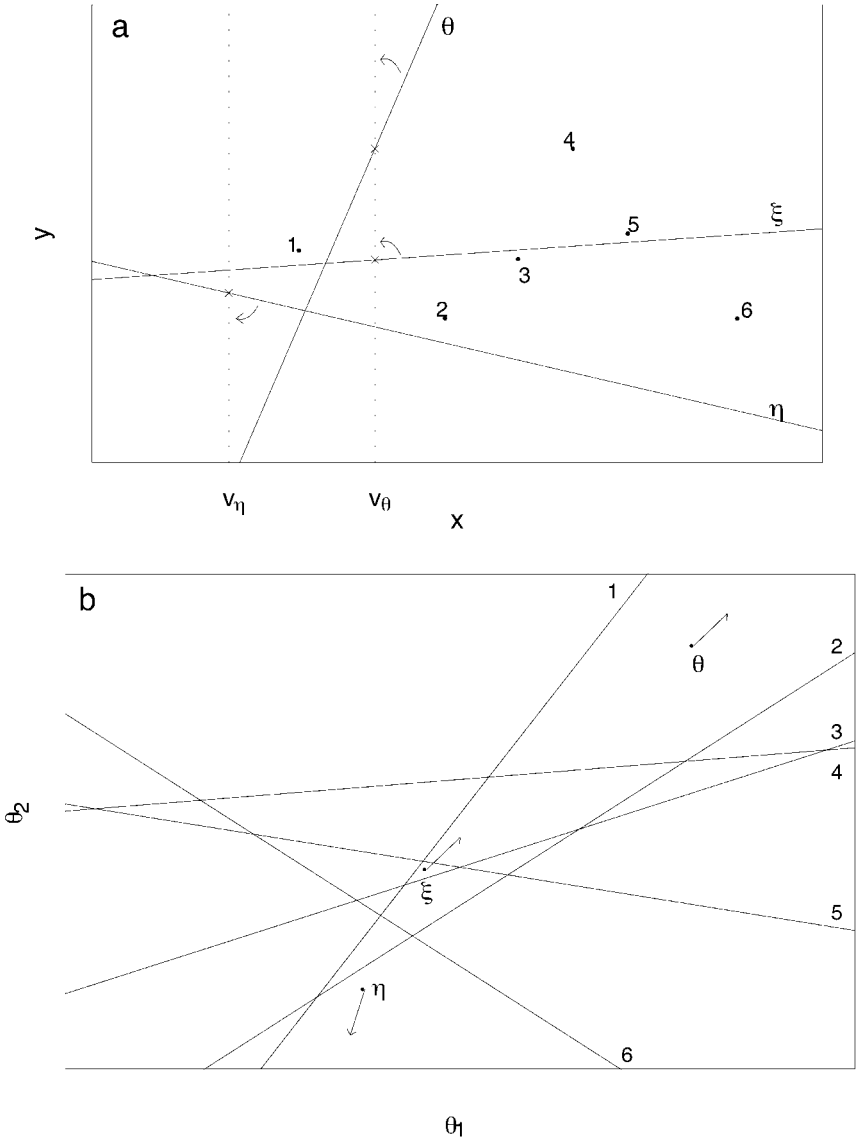

FIG. 9. Comparison of the regression depth in (a) primal space; and (b) dual space.

to be removed to set ξ free. More properties of regression depth in dual space are given in (Rousseeuw and Hubert 1996).

We will now prove that the regression depth characterizes the underlying empirical distribution for data sets in arbitrary position. The proof will be written down in dual space, and uses the fact that the regression depth is

constant on any open cell of the arrangement formed by the hyperplanes H_i. It is also constant on a common facet of two cells. Note that points with equal **x**-coordinates in primal space correspond to parallel hyperplanes in dual space.

THEOREM 2. *The empirical distribution of any data set $Z_n \subset \mathbb{R}^p$ is uniquely determined by its regression depth function.*

Proof. We will prove that for every data point $\mathbf{z}_i$ which belongs to the data set m times, the regression depth function makes a jump of m units between H_i and one of the unbounded regions which are separated by H_i. This is illustrated in Fig. 10: for three parallel hyperplanes we have indicated the depth in the unbounded regions on and near the hyperplanes. Clearly, there is a jump of $m=1$ units at each of these hyperplanes. Let $\boldsymbol{\theta}$ be a point on H_j lying in the relative interior of an unbounded common facet of two unbounded regions. Suppose that $\boldsymbol{\theta}$ has $rdepth(\boldsymbol{\theta}; Z_n) = k$. Then there exists a direction $\mathbf{u}$ such that the halfline $[\boldsymbol{\theta}, \boldsymbol{\theta} + \mathbf{u}>$ intersects exactly k hyperplanes (including H_j itself). Let $\boldsymbol{\theta}_\epsilon := \boldsymbol{\theta} + \epsilon\mathbf{u}$ be such that no hyperplane different from H_j passes between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_\epsilon$. Then $\boldsymbol{\theta}_\epsilon$ has depth at most $k-m$ where m is the number of times that the data point $\mathbf{z}_j$ occurs in the data set. Now suppose that a point $\tilde{\boldsymbol{\theta}}$ lying in one of the two open unbounded regions separated by H_j has depth smaller than $k-m$. Hence there exists a direction $\tilde{\mathbf{u}}$ such that the halfline $[\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}} + \tilde{\mathbf{u}}\rangle$ intersects fewer than $k-m$ hyperplanes H_i. But then the halfline $[\boldsymbol{\theta}, \boldsymbol{\theta} + \tilde{\mathbf{u}}\rangle$ intersects fewer than $(k-m)+m=k$ hyperplanes (there are exactly m hyperplanes lying

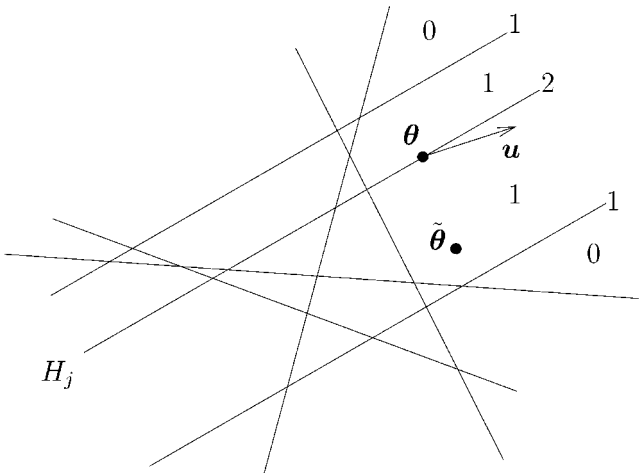

FIG. 10. Illustration in $\mathbb{R}^2$ of the proof of Theorem 2. For some unbounded regions and facets the regression depth is indicated, for example $rdepth(\boldsymbol{\theta}; Z_n) = 2$.

between $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$), which is a contradiction. Therefore, we know that there cannot be a jump larger than m units at either side of the unbounded facet which is part of H_j. Since we also found a point $\boldsymbol{\theta}_\epsilon$ in one of the unbounded regions separated by H_j with depth at most $k-m$, and the depth of $\boldsymbol{\theta}_\epsilon$ cannot be strictly smaller than $k-m$ by the same reasoning, there is a jump of exactly m units at H_j.

Therefore, when the regression depth function is given we can recover all the hyperplanes H_i by identifying jumps of the regression depth function between the open unbounded regions of the arrangement. ▮

REFERENCES

1. D. L. Donoho and M. Gasko, Breakdown properties of location estimates based on halfspace depth and projected outlyingness, *Ann. Statist.* **20** (1992), 1803–1827.
2. W. F. Eddy, Ordering of multivariate data, *in* "Computer Science and Statistics: Proceedings of the 16th Symposium on the Interface" (L. Billard, Ed.), pp. 25–30, North-Holland, Amsterdam, 1985.
3. H. Edelsbrunner, "Algorithms in Combinatorial Geometry," Springer-Verlag, Berlin, 1987.
4. X. He and G. Wang, Convergence of depth contours for multivariate datasets, *Ann. Statist.* **25** (1997), 495–504.
5. R. Liu, On a notion of data depth based on random simplices, *Ann. Statist.* **18** (1990), 405–414.
6. G. A. Koshevoy, Multivariate depths and underlying distributions: A uniqueness property, *Technical Report*, C.E.M.I. and Universität zu Köln, 1997.
7. G. A. Koshevoy and K. Mosler, Zonoid trimming for multivariate distributions, *Ann. Statist.* **25** (1997), 1998–2017.
8. J.-C. Massé and R. Theodorescu, Halfplane trimming for bivariate distributions, *J. Multivariate Anal.* **48** (1994), 188–202.
9. D. S. Moore and G. P. McCabe, "Introduction to the Practice of Statistics," Freeman, New York, 1989.
10. H. Oja, Descriptive statistics for multivariate distributions, *Statist. Probab. Lett.* **1** (1983), 327–332.
11. P. J. Rousseeuw and M. Hubert, Regression depth, *J. Amer. Statist. Assoc.*, to appear in the June 1999 issue.
12. P. J. Rousseeuw and I. Ruts, Algorithm AS 307: Bivariate location depth, *J. Roy. Statist. Soc. Ser. C* **45** (1996), 516–526.
13. P. J. Rousseeuw and I. Ruts, "The Bagplot: A Bivariate Box-and-Whiskers Plot," *Technical Report*, University of Antwerp, 1998.
14. P. J. Rousseeuw and I. Ruts, Constructing the bivariate Tukey median, *Statist. Sinica* **8** (1998), 827–839.
15. P. J. Rousseeuw and A. Struyf, Computing location depth and regression depth in higher dimensions, *Statist. and Comput.* **8** (1998), 193–203.
16. I. Ruts and P. J. Rousseeuw, Computing depth contours of bivariate point clouds, *Comput. Statist. Data Anal.* **23** (1996), 153–168.
17. J. W. Tukey, Mathematics and the picturing of data, *Proc. Internat. Congr. of Math.* **2** (1975), 523–531.

Printed in Belgium