JOURNAL OF MULTIVARIATE ANALYSIS **58**, 21–26 (1996)
ARTICLE NO. 0037

A Characterization of Halfspace Depth

Emilio Carrizosa*

Facultad de Matemáticas, Universidad de Sevilla, c/ Tarfia s/n. 41012 Seville, Spain

In this note we present a characterization of halfspace depth which relates it with well-known concepts of Locational Analysis. This characterization also leads to a natural extension of the concept of depth to noneuclidean location estimation as well as other settings like regression. © 1996 Academic Press, Inc.

1. INTRODUCTION

Given a d-dimensional dataset $X=\{X_1, X_2, ..., X_n\}$, the *depth* depth $(x; X)$ of a point $x \in \mathbb{R}^d$ is defined as

$$\text{depth}(x; X) = \min_{|u|=1} \#\{i\colon u'X_i \geqslant u'x\} \tag{1}$$

where $|\cdot|$ denotes the euclidean norm in $\mathbb{R}^d$ and $\#(A)$ stands for the cardinality of the set A. See, e.g., Tukey (1977) and Donoho and Gasko (1992)—hereafter referred to as [DG]—for further properties.

The notion of depth leads in a natural way to a robust location estimate, namely, the *deepest point* $T_*(X)$, defined in [DG] as

$$T_*(X) = \arg\max_x \text{depth}(x; X), \tag{2}$$

which has been shown to be affine equivariant, is a multivariate generalization of the median and enjoys good asymptotic properties.

In spite of these interesting properties, the original definition of depth$(\cdot; \cdot)$ does not seem to extend naturally from location estimation to other settings like regression fitting (see page 1815 of [DG]); moreover, its affine equivariance makes this concept of little use in anisotropic situations, as happens, e.g., when one deals with mixed variables. See also Arabie (1991).

Received April 21, 1994; revised September 1995.

AMS 1980 subject classifications: Primary 62F35; secondary 90A08.

Key words and phrases: Multivariate depth, location estimates, medians.

* E-mail address: ecarriz@obelix.cica.es.

0047-259X/96 $18.00

The purpose of this note is to present an equivalent definition of depth, taken from the context of Locational Analysis, which extends in a natural way to location estimation in anisotropic contexts and also to regression estimation.

2. A CHARACTERIZATION OF DEPTH

Definition 1. Let P be a probability distribution on $\mathbb{R}^d$. The normalized depth $ND(x; P)$ of a point $x \in \mathbb{R}^d$ in P is defined to be

$$ND(x; P) = \inf_{y \in \mathbb{R}^d} P(\{a: |y-a| \geqslant |x-a|\}) \tag{3}$$

This function ($1-ND$, to be more precise) has been introduced in the Operations Research literature to address facility location problems in a competitive framework: Suppose two firms, F_1 and F_2 want to enter into a market by sequentially locating one facility (e.g., shop) each. Consumers are assumed to be distributed in $\mathbb{R}^2$ according to a probability measure P, and will use their closest facility—ties allocated to the oldest firm, i.e., F_1; in other words, if F_1 (respec. F_2) locates its facility at $x \in \mathbb{R}^2$ (respec. at y), then F_1 (respec. F_2) captures the market consisting of those $a \in \mathbb{R}^2$ with $\|x-a\| \leqslant \|y-a\|$ (respec. $\|x-a\| > \|y-a\|$). The purpose of both firms is to find the facility locations maximizing their corresponding market share. Hence, if F_1 locates its facility at x, then F_2 will locale its facility at any $y^* \in \arg\max_y P(\|x-a\| > \|y-a\|)$, leading to a market share $1-ND(x; P)$, and $\arg\max_x ND(x; P)$, the so-called *Simpson points*, are just those locations for F_1 which maximize its market share (by minimizing the fraction of market captured by F_2).

ND also appears in facility location with voting: Suppose that the location of a certain facility is to be decided according to a voting process. The users, distributed following P, act as voters, and want the facility as close as possible. Hence, $P(\{a: \|x-a\| > \|y-a\|\})$ represents the weight of the coalition of voters which would agree in preferring y to x; $1-ND(x; P)$ gives the weight of the strongest possible coalition against x, and the Simpson points are then the least objectable locations. See, e.g., Carrizosa (1992), Durier (1989), Michelot (1993) and the references therein.

We will show that, when P represents the empirical probability measure of a dataset X of size n, ND$(x; P)$ equals $1/n$ depth$(x; X)$.

First, we give a more geometrical expression for ND. For any $x \in \mathbb{R}^d$, let $\mathscr{K}(x)$ denote the family of nonempty subsets A of $\mathbb{R}^d$ such that x does not belong to $\overline{\text{conv}}(A)$, the closure of the convex hull of A.

LEMMA 1. *For any probability measure P and $x \in \mathbb{R}^d$, one has*

$$ND(x; P) = 1 - \sup\{P(A) \colon A \in \mathscr{K}(x)\}. \tag{4}$$

Proof. Let $x \in \mathbb{R}^d$. Given $\alpha \in (0, 1]$,

$$\text{ND}(x; P) < \alpha \quad \text{iff } \exists y^* \in \mathbb{R}^d \quad \text{such that } P(\{a \colon |y^* - a| < |x - a|\}) > 1 - \alpha,$$

which is equivalent to saying that

$$\exists y^* \in \mathbb{R}^d, \exists A^*, \text{ s.t. } P(A^*) > 1 - \alpha \qquad \text{and} \quad |y^* - a| < |x - a| \; \forall a \in A^* \tag{5}$$

Now, by Proposition 1.3 of Durier and Michelot (1986), (5) turns out to be equivalent to

$$\exists A^*, \qquad \text{such that} \quad P(A^*) > 1 - \alpha, \qquad \text{and} \quad x \notin \overline{\text{conv}}(A^*)$$

or, in other words,

$$\exists A^* \in \mathscr{K}(x) \qquad \text{such that} \quad P(A^*) > 1 - \alpha. \tag{6}$$

Since (6) is equivalent to

$$\sup\{P(A) \colon A \in \mathscr{K}(x)\} > 1 - \alpha,$$

the result holds. ▮

PROPOSITION 1. *For any probability measure P and $x \in \mathbb{R}^d$, one has*

$$ND(x; P) = \inf_{|u| = 1} P(\{a \colon u'a \geqslant u'x\}) \tag{7}$$

Proof. Let $\alpha \in (0, 1]$. We will show that $\min_{|u|=1} P(\{a \colon u'a \geqslant u'x\}) < \alpha$ iff $\text{ND}(x; P) < \alpha$. Suppose first that $\text{ND}(x; P) < \alpha$; then, by Lemma above, there exists A^* such that $P(A^*) > 1 - \alpha$ and $x \notin \overline{\text{conv}}(A^*)$. Then (see, e.g., Luenberger (1984)), there exists a nonzero vector $\bar{u} \in \mathbb{R}^d$ such that $\bar{u}'x > \bar{u}'a$ for all $a \in A^*$. Hence, $\inf_{|u|=1} P(\{a \colon u'a \geqslant u'x\}) < \alpha$.

The proof of the converse is straightforward, and will not be given here. ▮

By Proposition 1 above, given a dataset $X = \{X_1, X_2, ..., X_n\} \subset \mathbb{R}^d$, if we denote by P_n its empirical probability measure (i.e., $P_n(S) = (1/n) \; \#\{i \colon X_i \in S\}$), one has that

$$\text{depth}(x; X) = n \, \text{ND}(x; P_n),$$

showing the equivalence between (1) and (3). This provides a new insight into the nature of depth$(\cdot; \cdot)$ and deepest points as estimators: asserting

$ND(x; P_n) > \alpha$ just means that, if we make a pairwise comparison between x and any other candidate to estimator y, the frequency of datapoints strictly closer to y never attains the value α, yielding deepest points as, in some sense, least-objectable estimates. This interpretation does not remain true when the parameter space Θ is forced to be a proper subset of $\mathbb{R}^d$ due to some prior knowledge on the location parameter—e.g., it has integer, or nonnegative coordinates—as the following example shows.

EXAMPLE 1. Let P be the bivariate probability measure given by

$$\begin{aligned} P(-1, -1) &= \tfrac{2}{9} \\ P(-1, 1) &= \tfrac{3}{9} \\ P(1, 1) &= \tfrac{2}{9} \\ P(1, -1) &= \tfrac{2}{9} \end{aligned}$$

and suppose that the parameter space is $\Theta = \{(\lambda, 0): \lambda \in \mathbb{R}\}$. It is not difficult to see that, for any $\lambda \in \mathbb{R}$,

$$ND((\lambda, 0); P) = \inf_{y \in \mathbb{R}^2} P(\{a: |y-a| \geqslant |(\lambda, 0) - a|\}) \tag{8}$$

$$= \begin{cases} 0, & \text{if } \ |\lambda| > 1 \\ \frac{2}{9}, & \text{if } \ |\lambda| \leqslant 1, \lambda \neq 0 \\ \frac{4}{9}, & \text{if } \ \lambda = 0 \end{cases} \tag{9}$$

yielding $(0, 0)$ as the unique deepest point, i.e.,

$$\{(0, 0)\} = \arg\max_{x \in \Theta} \inf_{y \in \mathbb{R}^2} P(\{a: |y-a| \geqslant |x-a|\}) \tag{10}$$

However, ND and deepest points cannot be calculated through pairwise comparisons among elements in Θ, since, as can be readily seen, for any $\lambda \in \mathbb{R}$ one has that

$$\inf_{y \in \Theta} P(\{a: |y-a| \geqslant |(\lambda, 0) - a|\}) = \begin{cases} 0, & \text{if } \ |\lambda| > 1 \\ \frac{4}{9}, & \text{if } \ -1 < \lambda \leqslant 1 \\ \frac{5}{9}, & \text{if } \ \lambda = -1, \end{cases} \tag{11}$$

thus

$$\{(-1, 0)\} = \arg\max_{x \in \Theta} \inf_{y \in \Theta} P(\{a: |y-a| \geqslant |x-a|\}), \tag{12}$$

yielding a different result than (11).

3. DISCUSSION

Now, let us discuss a few consequences of the characterization presented in the section above.

3.1. *Noneuclidean Depth*

The definitions (1) and (2) extend naturally to problems with non-euclidean metrics or dissimilarity measures (also applicable to binary or cathegorical data sets!), which is in agreement with the opinion that data-induced requirements may make other noneuclidean distances preferable for some data analysis problems. See, e.g. Arabie (1991) or Cuadras–Fortiana–Oliva (1994).

Indeed, given a dissimilarity measure Δ, one can define the normalized depth associated with Δ as

$$ND_{\Delta}(x; P) = \inf_{y} P(\{a: \Delta(z, y) \geqslant \Delta(a, x)\}), \tag{13}$$

and define the Δ-deepest points as the maximizers of ND_{Δ} in (13).

Whilst statistical properties of Δ-deepest points for noneuclidean metrics, such as l_p norms, seem to remain unexplored, a number of papers within the field of Locational Analysis have addressed this topic from an optimization or algorithmic viewpoint. See, e.g., the papers of McKelvey and Wendell (1976), Demange (1982), Durier (1989) or Michelot (1993) for theoretical properties, and Drezner (1982) for an $O(n^2 \log^2 n)$ algorithm to find $\|\cdot\|_2$-deepest points, shown in Durier (1989) to be also valid for any l_p-norm ($1 < p < \infty$).

3.2. *Depth in Regression*

Although much less geometrical than the original definition, the expression (1) proposed here is easily extended to regression settings such as linear regression. Indeed, given a probability measure P in $\mathbb{R}^d \times \mathbb{R}$ corresponding to a multivariate random variable (X, Y), we could define the normalized depth $\mathrm{ND}(H_{a,b}; P)$ of the hyperplane $H_{a,b} \equiv y = a'x + b$ as

$$\mathrm{ND}(H_{a,b}; P) = \inf_{(c,d) \in \mathbb{R}^d \times \mathbb{R}} P(\{(x, y) \in \mathbb{R}^d \times \mathbb{R} : |y - a'x - b| \leqslant |y - c'x - d\}) \tag{14}$$

Deepest hyperplanes (i.e., maximizers of (14)) are then those hyperplanes minimizing the highest mass of points for which some other hyperplane yields strictly lower residuals.

REFERENCES

Arabie, P. (1991). Was Euclid an unnecessarily sophisticated psycologist? *Psychometrika* **56** 567–587.

Carrizosa, E. (1992). *Problemas de localización multiobjetivo*, Ph.D. thesis. University of Seville, Spain.

Cuadras, C. M., Fortiana, J., and Oliva, F. (1994). The proximity approach of an individual to a population with applications in discriminant anlysis, Mathematics Preprints Series, No. 162. Universitat de Barcelona.

Demange, G. (1982). Spatial models of collective choice. In *Locational Analysis of Public Facilities* (J. F. Thisse and H. G. Zoller, Eds.), *Stud. Math. Managerial Econ.* **31** 153–182.

Donoho, D. L., and Gasko, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness, *Ann. Statist.* **20** 1803–1827.

Drezner, Z. (1982). Competitive location strategies for two facilities, *Regional Sci. Urban Econ.* **12** 485–493.

Durier, R. (1989). Continuous location under majority rule. *Math. Oper. Res.* **14** 258–274.

Durier, R., and Michelot, C. (1986). Set of efficient points in a normed space. *J. Math. Anal. Appl.* **117** 506–528.

Luenberger, D. G. (1984). "Linear and Nonlinear Programming." Addison–Wesley, Reading, MA.

McKelvey, R. D., and Wendell, R. E. (1976). Voting equilibria in multidimensional choice spaces, *Math. Oper. Res.* **1** 144–158.

Michelot, C. (1993). The mathematics of continuous location, *Stud. Locational Anal.* **5** 59–83.

Tukey, J. W. (1977). "Exploratory Data Analysis." Addison–Wesley, Reading, MA.