

Convex Hull Probability Depth: first results

Giovanni C. Porzio and Giancarlo Ragozini

Abstract In this work, we present a new depth function, the *convex hull probability depth*, that is based on the convex hull peeling notion. Given a point x , its depth is defined to be the expected value of (one minus) the probability content under F of the random convex hull to which x belongs in a random peeling sequence. For this depth, first theoretical results are offered. More specifically, we discuss how it properly induces inner-outward ordering when F is an absolutely continuous half-space symmetric distribution. In addition, we show that its deepest point is the half-space symmetry center (a proper multidimensional median notion), and we prove it is a statistical depth function of type A according to the Zuo and Serfling taxonomy.

Key words: Nonparametric multivariate data analysis, Robust statistics.

1 Introduction

Data depth is a function $D(x; F)$ that measures the centrality of a point $x \in \mathbb{R}^d$ with respect to a given multivariate distribution F . The deepest points lie at the core of the distribution, while points with lower depth values are located in the distribution tails. First applications of data depth have been multivariate center-outward ordering of data scatters, robust estimates of location and dispersion, multiple outlier detection, and multivariate data exploratory analysis [11, 1, 12, 3, 10]. More recently, robust regression analysis based on data depth have been introduced (see e.g. [9]). Data depth has also been used within a multivariate statistical process control setting

Giovanni C. Porzio

University of Cassino, Department of Economics, Via S. Angelo - Polo Folcara, 03043 Cassino (FR), Italy e-mail: porzio@eco.unicas.it

Giancarlo Ragozini

Federico II University of Naples, Department of Sociology, Vico Monte di Pietà 1, 81132 Naples, Italy e-mail: giragoz@unina.it

[2, 5, 4], while in a data mining framework it has been introduced as a tool for data cleaning.

Many depth functions are available in the literature (see e.g. [3, 13]). Among them, the half-space, the simplicial and the convex hull peeling depth are the most popular and used.

As known, the convex hull peeling depth is intuitive and computationally affordable in high dimensions. However, it is not a statistical depth function, essentially because its values strictly depend on the observed sample, and a population analogue is lacking.

For this reason, with this work we present a new depth notion, first introduced by Porzio and Ragozini in [6], that can be considered a population counterpart of the peeling depth. It has been called *convex hull probability depth*, as it joins the convex hull peeling idea with the probability contents of random convex hulls. It is worth noting this depth notion induces inner-outward ordering when F is an absolutely continuous half-space symmetric distribution. Furthermore, we note that its deepest point is the half-space symmetry center (a proper multidimensional median notion), and that it is a statistical depth function of type A according to the Zuo and Serfling taxonomy [13].

The paper is organized as follows. Section 2 provides some notations on convex hull peeling, while in Section 3 our new depth notion is defined. Section 4 offers some theoretical results on inner-outward ordering induced by convex hull probability depth and Section 5 shows our depth is a statistical depth function.

2 Convex hull peeling depth

Convex hull peeling depth was first introduced by Barnett [1] as a tool for ordering multivariate data. Given a finite set of points $\mathcal{Y} = \{y_1, \dots, y_r\}$, $\mathcal{Y} \in \mathbb{R}^d$, its convex hull $CH(\mathcal{Y})$ is the smallest convex set containing it:

$$CH(\mathcal{Y}) := \{y : y = \alpha_1 y_1 + \dots + \alpha_r y_r, 0 \leq \alpha_i \leq 1, \sum_i \alpha_i = 1\}.$$

Let $VCH(\mathcal{Y})$ be the function which provides the vertices of the convex hull of $\mathcal{Y}$. We have that a convex hull is completely defined by the set of its vertices $\mathcal{V} \subseteq \mathcal{Y}$:

$$\mathcal{V} = VCH(\mathcal{Y}) := \{y_i \in \mathcal{Y} : y_i \in \partial CH(\mathcal{Y})\},$$

with $\partial(\mathcal{S})$ the boundary of a set $\mathcal{S}$. In other words, the vertices are those y_i that lie on the convex hull boundary.

Consider now the sequence of the nested convex hulls $CH_k(\mathcal{Y})$, $k = 1, \dots, \tilde{K}$, where the index k refers to the layers. The sequence of the nested convex hulls is obtained by iteratively removing the vertices from the previous set in the sequence. In other words, the first element of the sequence is the convex hull of $\mathcal{Y}$. To obtain

the second element, remove the vertices from $\mathcal{V}$ and consider the convex hull of the peeled set, and so on. We call this sequence the convex hull peeling sequence.

The corresponding sequence of vertices will have elements $\mathcal{V}_1 = VCH(\mathcal{V})$, $\mathcal{V}_2 = VCH(\{\mathcal{V} - \mathcal{V}_1\})$, and generally

$$\mathcal{V}_k := VCH(\{\mathcal{V} - \bigcup_{j=1}^k \mathcal{V}_{j-1}\}),$$

with $\mathcal{V}_0 = \emptyset$. Note that the sequence ends when all the points in $\mathcal{V}$ are removed. That is, the last layer is given by $\tilde{K} = \min\{n | \{\mathcal{V} - \bigcup_{j=1}^{n+1} \mathcal{V}_{j-1}\} = \emptyset\}$.

The k -th element of the nested convex hull sequence will be then the set:

$$CH_k(\mathcal{V}) := CH(\{\mathcal{V} - \bigcup_{j=1}^k \mathcal{V}_{j-1}\}).$$

Finally, after Barnett [1], given an observed sample $\underline{y}_n = \{y_i\}_{i=1,\dots,n}$ drawn from a distribution F_Y in $\mathfrak{R}^d$, the convex hull peeling depth of a sample point y_i with respect to $\underline{y}_n$ is the layer to which it belongs in the peeling sequence. More formally, Barnett's depth $BD(y_i, \underline{y}_n)$ is given by:

$$BD(y_i, \underline{y}_n) := \{k : y_i \in \partial(CH_k(\underline{y}_n))\}, \quad \forall y_i \in \underline{y}_n. \quad (1)$$

3 Convex hull probability depth

Even if quite popular, Barnett's depth is not a statistical depth function [13]. First of all, it is not defined for all the points in the sample space but only for the observed points. Even more, it lacks a population analogue.

For these reasons, we consider a new depth notion that turns out to be a statistical depth function. As it joins the convex hull peeling idea and the probability contents of convex hulls, it has been called Convex Hull Probability Depth.

Let us first extend Barnett's depth to any point $x \in \mathfrak{R}^d$. Given a sample $y_1, \dots, y_n$ from a distribution F and a point x , in analogy with Equation (1), we define the layer $k^*(x, \underline{y}_n)$ to which x belongs in the convex hull peeling sequence as:

$$k^*(x, \underline{y}_n) := \{k : x \in \partial(CH_k(x, \underline{y}_n))\}, \quad \forall x \in \mathfrak{R}^d. \quad (2)$$

where $CH_k(x, y_1, \dots, y_n)$ is the k -th convex hull in the sequence of the nested convex hull peeling of the set $\{x, y_1, \dots, y_n\}$.

For our aims, let us consider also the probability content under F of the k -th convex hull $CH_k(x, y_1, \dots, y_n)$ in the peeling sequence. That is, let us consider the quantity $P(Y \in CH_k(x, y_1, \dots, y_n))$. Note this probability depends on the observed sample. Then, the Convex Hull Probability Depth is defined as follows.

Definition (Convex Hull Probability Depth).

Let $Y_1, \dots, Y_n$ be a random sample from a distribution F in $\mathbb{R}^d$, with $n \geq d + 1$. The Convex Hull Probability Depth of a point $x \in \mathbb{R}^d$ with respect to F is defined to be:

$$CHPD_n(x; F) := E[h_{CH}(x; Y_1, \dots, Y_n)], \quad (3)$$

with

$$h_{CH}(x; y_1, \dots, y_n) := 1 - P(Y \in CH_{k^*}(x, y_1, \dots, y_n)), \quad (4)$$

where $k^* = k^*(x, \underline{y}_n)$ as given by Equation (2), and $E[\cdot]$ is the expected value operator.

That is, the Convex Hull Probability Depth of a point x is the expected value of (one minus) the probability content under F of the convex hull to which x belongs in the peeling sequence. Rather than the probability itself, the complement of the probability content is considered in order to have a function that assigns higher values to deeper points.

Remark 3.1. We note that $CHPD_n(x; F)$ is a bounded function by definition, with $0 \leq CHPD_n(x; F) \leq 1$. In addition, its value depends on the sample size n .

Remark 3.2. The convex hull probability depth of a point x with respect to a distribution F combines two ideas. First, to each point x the probability content of the $CH_k(x, y_1, \dots, y_n)$ to which x belongs is associated, and not simply the number k^* of its layer (as in Barnett's depth). Then, the expected value over all the possible sample $\underline{Y}_n$ of size n is considered.

Remark 3.3. The $CHPD_n(x; F)$ definition involves the expected value of probabilities. We note that these latter are actually random numbers whose distribution depends on x , n and F through the random sample $(\underline{Y}_n)$. More specifically, the probabilities are function of the random sets $CH_{k^*}(x, \underline{Y}_n)$.

Remark 3.4. By definition, the Convex hull probability depth is a Type A depth function in the Zuo and Serfling taxonomy.

To illustrate this definition, we present a graphical example. Let it be of interest to evaluate $CHPD_{50}((1, 1)^T; F_Y)$, with $Y \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_2)$. That is, consider the value of the convex hull probability depth of the point $x^T = (1, 1)$ with respect to the bivariate normal distribution with zero means, unit variances and independent components, for $n = 50$.

We drew six samples $\underline{y}_{50}^s$ from $Y \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_2)$, $s = 1, \dots, 6$. Each of them is offered through a scatter plot in Figure 1. In addition, the point $x^T = (1, 1)$ is highlighted in each of the six plots through a large filled dot. Furthermore, the convex hull peeling sequences of the sets $\{x, \underline{y}_{50}^s\}$ is depicted through the nested series of the convex hull boundaries.

First of all, we note that the layer to which the point x belongs varies sample by sample. For instance, in the sample depicted in the upper left plot, x belongs to the fourth layer; in the upper right plot, it belongs to the second layer. How-

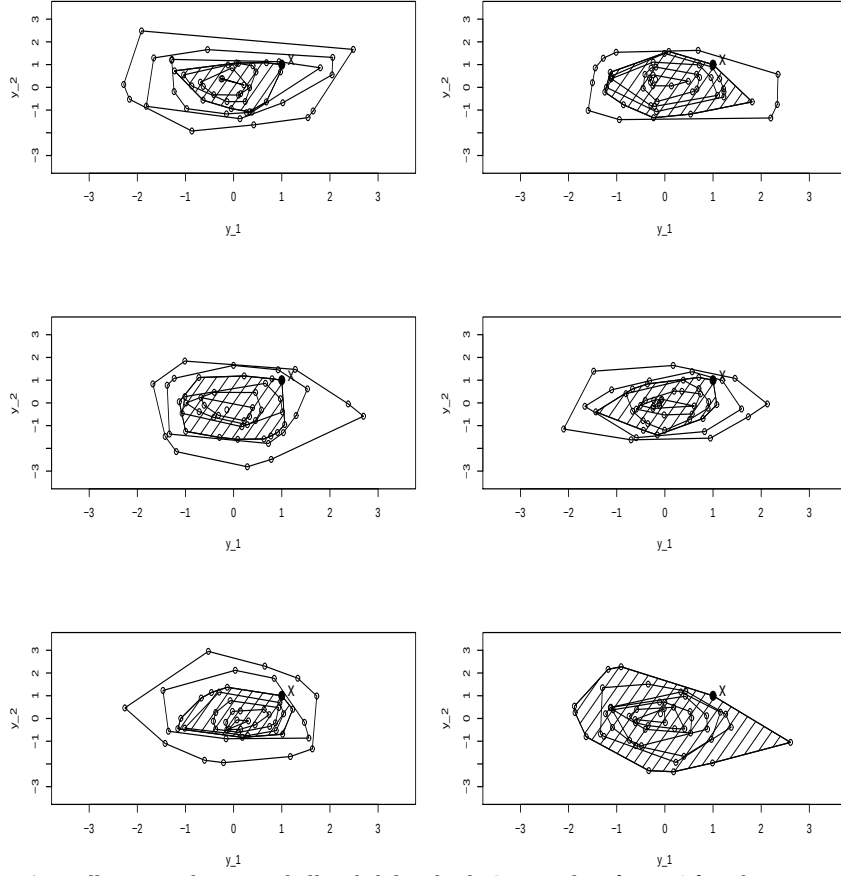


Fig. 1 Illustrating the convex hull probability depth. Six samples of size 50 from bivariate standard independent normal distributions and the corresponding convex hull peeling sequences of the sample plus the point $x = (1, 1)^T$ are depicted. Shaded areas highlight the convex hull layer to which x belongs in the peeling sequence.

ever, the layer itself is not of interest here. Rather, we care about the shaded area in each plot. That is, about the area included by the convex hull layer to which x belongs in the peeling sequence. Obviously, these areas are random sets: each sample defines a different area. The $CHPD_n$ is related to the probability content under F of these shaded areas. Given that the areas are random sets, the corresponding probability contents are random numbers. The $CHPD_n$ is then the expected value of (one minus) these random numbers. With respect to Equation (4), the function $h_{CH}(x; y_1, \dots, y_n) = 1 - P(Y \in CH_{k*}(x, y_1, \dots, y_n))$ yields the probability contents of (one minus) the shaded areas.

4 $CHPD_n$ inner-outward ordering

Depth functions have been generally introduced to provide an F -based center-outward ordering of points $x \in R^d$. Thus, investigating the inner-outward ordering induced by any depth function turns out to be at the core of its properties.

For this reason, $CHPD_n$'s inner-outward induced ordering is discussed. For the sake of clarity, we first illustrate the ordering induced in the univariate case. Then, we state the more general result.

Theorem 1 ($CHPD_n$ inner-outward ordering on the real line). *Let $Y_1, \dots, Y_n$ be a random sample from an absolutely continuous distribution F_Y in $\mathcal{R}^1$, θ be the distribution median (i.e. $F_Y(\theta) = 0.5$), x_1 and x_2 be two points in $\mathcal{R}^1$ with $|x_1 - \theta| \geq |x_2 - \theta|$. Then:*

$$CHPD_n(x_1; F) \leq CHPD_n(x_2; F) \quad \forall n. \quad (5)$$

Proof. The proof considers the random variable

$$k^*(x, \underline{Y}_n) - 1 = \min(R, n - R), \quad (6)$$

where $x \in \mathcal{R}^1$, $\underline{Y}_n$ is a random sample of size n , and R counts the Y_i 's less than x . Note that $k^*(x, \underline{Y}_n)$ is the (random) convex hull layer to which x belongs in the peeling sequence.

The random variable in Equation (6) is folded binomial distributed with probability parameter $p = \min(F_Y(x), 1 - F_Y(x)) = 1/2 - |F_Y(\theta) - F_Y(x)|$. This parameter measures thus the distance of x to the median θ , $|x - \theta|$, in terms of the distance $|F_Y(\theta) - F_Y(x)|$. Consequently, and given that folded binomial distributions are stochastically ordered with respect to the parameter p for a given m (Porzio and Ragozini, 2009), we have:

$$k^*(x_2, \underline{Y}_n) \leq_{st} k^*(x_1, \underline{Y}_n) \quad \forall n, \quad (7)$$

as $k^*(x_2, \underline{Y}_n) - 1 \sim fBin(n, p_2)$ and $k^*(x_1, \underline{Y}_n) - 1 \sim fBin(n, p_1)$, with $p_1 \leq p_2$, being $|x_1 - \theta| \geq |x_2 - \theta|$ by hypothesis. Finally, this stochastic ordering implies the $CHPD_n$ values are inner-outward ordered, as they are expected values of nondecreasing functions of k^* .

This theorem implies that in the univariate case the $CHPD_n$ deepest point is the median θ . In higher dimensional spaces, the multivariate median can be defined in several ways. One approach refers to some notions of multivariate symmetry, and among the possible notions we consider a very broad notion: the half-space symmetry. A distribution F_Y is half-space symmetric around θ if $P(Y \in H) \geq 0.5$ for every closed half-space H containing θ . In other words, we have $P(Y \in H_\theta) \geq 0.5$ for any closed half-space H with $\theta \in \partial H$. Note that the usual univariate median satisfies such symmetry notion. If you consider that elliptic distributions are all half-space symmetric, we have that half-space symmetry yields a quite broad centrality notion.

For our purposes, let us denote with $\mathcal{F}_\theta$ the class of the absolutely continuous distributions half-space symmetric around θ , and with density function non-zero everywhere. In such a case, we have that for $F_Y \in \mathcal{F}_\theta$, $\theta \in \mathbb{R}^d$ is the unique point for which $P(Y \in H_\theta) = 0.5$ [14]. We have that $CHPD_n$'s inner-outward ordering can be defined in $\mathbb{R}^d$ with respect to the half-space symmetry center θ . This in turns implies that, for $F_Y \in \mathcal{F}_\theta$, $\theta \in \mathbb{R}^d$, the half-space symmetry center θ is the $CHPD_n$ deepest point. Note that this property is shared with the simplicial and the Tukey's half-space depth.

Theorem 2 ($CHPD_n$ inner-outward ordering in $\mathbb{R}^d$).

Let $Y_1, \dots, Y_n$ be a random sample from a distribution $F_Y \in \mathcal{F}_\theta$ in $\mathbb{R}^d$. Let also $l_{\theta x_1}$ be the line passing through θ and the point $x_1 \in \mathbb{R}^d$, that is:

$$l_{\theta x_1} = \{x : x = \theta + \alpha(x_1 - \theta), \alpha \in \mathbb{R}\}.$$

For any point $x_2 = \theta + \alpha(x_1 - \theta)$, $0 \leq \alpha \leq 1$, i.e. $x_2 \in \mathbb{R}^d$ lies on $l_{\theta x_1}$ between θ and x_1 , it holds that:

$$CHPD_n(x_1; F) \leq CHPD_n(x_2; F) \quad \forall n. \quad (8)$$

The proof is available in [8].

Remark 4.1. As noted, the $CHPD_n$ value for a given x depends on the sample size n . However, the inner-outward ordering induced by this depth function is n invariant. Furthermore, Porzio and Ragozini [8] provided an asymptotic version of $CHPD_n$ that turns out to be n invariant.

5 The $CHPD_n$ as a statistical depth function

In this Section, we prove that the Convex Hull Probability Depth is a statistical depth function according to the desirable properties discussed by Zuo and Serfling [13].

First, we note that $CHPD_n$ is a bounded and non negative mapping. Furthermore, the following properties hold.

Theorem 3 ($CHPD_n$ affine invariance). For any random vector Y in $\mathbb{R}^d$, any $d \times d$ nonsingular matrix A , and any d -vector b it holds that:

$$CHPD_n(Ax + b; F_{AY+b}) = CHPD_n(x; F_Y).$$

Theorem 4 ($CHPD_n$ maximality at center). For any random vector Y in $\mathbb{R}^d$, with $F_Y \in \mathcal{F}_\theta$ (i.e. F_Y belongs to the class of absolutely continuous distributions half-space symmetric around θ and with density function non-zero everywhere) we have:

$$CHPD_n(\theta; F_Y) = \sup_{x \in \mathbb{R}^d} CHPD_n(x; F_Y) \quad \forall n.$$

Theorem 5 (CHPD_n monotonicity with respect to the deepest point). For any random vector Y in $\mathbb{R}^d$, with $F_Y \in \mathcal{F}_\theta$, and with deepest point θ ,

$$\text{CHPD}_n(x; F) \leq \text{CHPD}_n(\theta + \alpha(x - \theta); F) \quad \forall \alpha \in [0, 1], \forall n.$$

Theorem 6 (CHPD_n vanishing at infinity - weaker version). For any random vector Y in $\mathbb{R}^d$, with $F_Y \in \mathcal{F}_\theta$, as $\|x\| \rightarrow \infty$

$$P(\{y : \text{CHPD}_n(y; F) \leq \text{CHPD}_n(x; F)\}) \rightarrow 0 \quad \forall n.$$

CHPD_n affine invariance derives from the convex hull peeling affine invariance. Maximality at center and monotonicity are implied by the inner-outward ordering of CHPD_n given in Theorem (2). The last property, vanishing at infinity, holds as it is implied by Theorems (4) and (5) according to [13].

References

1. Barnett, V.: The ordering of multivariate data (with discussion). Journal of Royal Statistical Society, Ser. A. **139**: 318–354 (1976)
2. Liu, R.Y.: Control Charts for Multivariate Process. Journal of the American Statistical Association. **90**, 1380–1387 (1995)
3. Liu, R.Y., Parelius, J.M., Singh, K.: Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics and Inference. The Annals of Statistics. **27**, 783–858 (1999)
4. Messaoud, A., Weihs, C., Hering, F.: Detection of chatter vibration in a drilling process using multivariate control charts. Computational Statistics and Data Analysis. **52**, 3208–3219 (2008)
5. Porzio, G.C., Ragozini, G.: Multivariate Control Charts from a Data Mining Perspective. In: Recent Advances in Data Mining of Enterprise Data. Liao, T.W., Triantaphyllou, E. (Eds.), World Scientific, Singapore, 413–462 (2007)
6. Porzio, G.C., Ragozini, G.: Convex Hull Probability Depth. International Workshop on Robust and Nonparametric Statistical Inference. Hejnice, Czech Republic (2007)
7. Porzio, G.C., Ragozini, G.: Stochastic ordering of folded binomials. Statistics and Probability Letters. **79**, 1299–1304 (2009)
8. Porzio, G.C., Ragozini, G.: On Some Properties of the Convex Hull Probability Depth. Working Papers - Department of Economics, University of Cassino, Cassino, *submitted* (2010)
9. Rousseeuw, P.J., Hubert, M.: Regression depth (with discussion). Journal of the American Statistical Association. **94**, 388–433 (1999)
10. Rousseeuw, P.J., Ruts, I., Tukey, J.W.: The Bagplot: A Bivariate Boxplot. The American Statistician. **53**, 382–387 (1999)
11. Tukey, J.W.: Mathematics and the picturing of data. Proceedings of the International Congress of Mathematicians 2. Montreal, Canada, 523–531 (1975)
12. Zani, S., Riani, M., Corbellini, A.: Robust Bivariate Box-plots and Multiple Outlier Detection. Computational Statistics and Data Analysis. **28**, 257–270 (1998)
13. Zuo, Y., Serfling, R.: General notions of statistical depth function. Annals of Statistics. **28**, 461–482 (2000)
14. Zuo, Y., Serfling, R.: On the performance of some robust nonparametric location measures relative to a general notion of multivariate symmetry. Journal of Statistical Planning and Inference. **84**, 55–79 (2000)