Simplicial Depth: An Improved Definition, Analysis, and Efficiency for the Finite Sample Case*

Michael A. Burr   Eynat Rafalin   Diane L. Souvaine

Abstract

As proposed by Liu [8] the *simplicial depth* of a point x with respect to a data set $\mathcal{S}$ in $\mathbb{R}^d$ is the fraction of closed simplices given by $d+1$ of the data points containing x. We propose an alternative definition for simplicial depth which remains valid over a continuous probability field and fixes some problems in the finite sample case.

1 Introduction

A data depth measures how deep (central) a given point $x \in \mathbb{R}^d$ is relative to F, a probability distribution in $\mathbb{R}^d$ or relative to a given data cloud. Data depth provides an alternative analysis to classical staticstic because no assumption about the underlying distribution is needed, only the relative positions of the data points. However, many data depth functions are quite expensive to compute, thus study of these functions and related algorithms is essential for these functions to become more useful in statistics.

Most depth functions are defined with respect to a probability distribution F, considering $\{X_1, .., X_n\}$ random observations from F. The *sample version* of the depth function is obtained by replacing F by F_n, the empirical distribution of the sample $\{X_1, .., X_n\}$. We mainly discuss the finite sample version of *simplicial depth* [8], although some references are made to the continuous case.

To distinguish between the depth of points $\{X_i\}$ from the original data set and the depth of any other point of $\mathbb{R}^d$, points not part of the data set are referred to as *positions*. *Facet* is used to define the facets of a specific simplex defined by $d+1$ data points. The facets subdivide $\mathbb{R}^d$ into regions. A *cell* is the set of all positions connected by a path which does not intersect a facet.

1.1 Desirable properties

Several properties of depth functions were introduced by Liu [8]. Recently Serfling and Zuo [9] formulated a general definition of desirable properties of depth functions, based on Liu's work, and evaluated several depth functions according to these properties. The desirable properties:

P1. Affine invariance: The depth of a point x should not depend on the underlying coordinate system, or, in particular, on the scales of the underlying measurements.

P2. Maximality at center: For a distribution having a uniquely defined center (e.g. a point of symmetry), the depth function should attain maximum value at the center.

P3. Monotonicity Relative to Deepest Point: As a point x moves away from the 'deepest point' (the point at which the depth function attains a maximum) along any fixed ray through the center, the depth at x should decrease monotonically.

P4. Vanishing at Infinity: The depth of a point x should approach zero as $||x||$ approaches ∞.

For applications of data mining and classification of large data sets, consistency under dimensions change would be another desirable property. We propose:

P5. Invariance under dimensions change: The relative depth of any two positions should not depend on the dimension in which the depth was computed.

1.2 Simplicial Depth Background

Introduced by Liu [8], Simplicial depth is robust and affine invariant.

Definition 1 Simplicial depth (Liu [8]): *Given a probability distribution F in $\mathbb{R}^d$, the simplicial depth of x is the probability that x belongs to a random closed simplex in $\mathbb{R}^d$:*

$$SD_{Liu}(\mathcal{F}; x) = P_F(x \in S[X_1, \ldots X_{d+1}])$$

where $S[X_1, \ldots X_{d+1}]$ is a closed simplex formed by $d+1$ random observations from F[1].

Definition 2 Simplicial depth for the sample version (Liu [8]): *The* simplicial depth *of a point x with respect to a data set $\mathcal{S} = \{X_1, \ldots, X_n\}$ is the fraction of the closed simplices formed by $d+1$ points of $\mathcal{S}$ containing x, where I is the indicator function:*

$$SD_{Liu}(\mathcal{S}; x) = \binom{n}{d+1}^{-1} \sum I_{(x \in S[X_{i_1}, \ldots X_{i_{d+1}}])}.$$

1.3 Problems

Several problems arise in the finite sample case of simplicial depth under Liu's definition. Serfling and Zuo found that the function does not behave well in some finite sample cases (under the desirable properties) and may be unattractive for

*Department of Computer Science, Tufts University, Medford, MA 02155. {mburr, erafalin, dls}@cs.tufts.edu. Partially supported by NSF grant #EIA-99-96237

[1] $S[X_1, \ldots X_{d+1}]$ represents the convex hull of the $d+1$ points. If the point set $X_1, \ldots X_{d+1}$ is not affinely independent then $S[X_1, \ldots X_{d+1}]$ is not a simplex in $\mathbb{R}^d$ but rather a convex object contained within a k-flat for $k < d$.

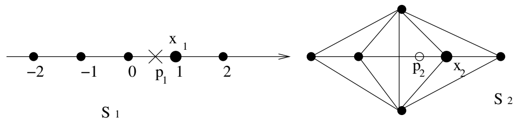


Figure 1: The Zuo-Serfling Counterexamples [9]: *Counterexample 1:* For $S_1 = \{-2,-1,0,1,2\}$ with center 0 in $\mathbb{R}^1$, $SD_{Liu}(p_1) = \frac{6}{10}$ and $SD_{Liu}(x_1) = \frac{7}{10}$. Violates P3. $SD_{BRS}(p_1) = \frac{6}{10}$ and $SD_{BRS}(x_1) = \frac{5}{10}$. Consistent with P3. *Counterexample 2:* For $S_2 = \{(\pm 1,0),(\pm 2,0),(0,\pm 1)\}$ with center $(0,0)$ in $\mathbb{R}^2$, $SD_{Liu}(p_2) = \frac{10}{20}$ and $SD_{Liu}(x_2) = \frac{12}{20}$. Violates P3. $SD_{BRS}(p_2) = \frac{12}{20}$ and $SD_{BRS}(x_2) = \frac{9}{20}$. Consistent with P3. [2]

some forms of statistical analysis. Our revised definition removes some of their concerns and alleviates other problems. Section 2.1 will describe how the revised definition alleviates these problems.

Maximality and Monotonicity: The Simplicial Depth function is a statistical depth function, in the sense of Serfling and Zuo's definition [9], for a continuous angularly symmetric distribution. Zuo and Serfling show, however, that the simplicial depth function for the finite sample case fails to satisfy the maximality (P2) and monotonicity (P3) properties using several counterexamples, two of them presented in a slightly modified form in Figure 1. The revised definition, as described below resolves the problems raised by these counterexamples. Nonetheless, as shown in Section 3, the maximality and monotonicity properties still do not hold.

Positions on Facets: Depth of positions on facets causes discontinuities in the depth function. The depth of all positions on the boundary of a cell is at least the depth of a position on its interior. In most cases the depth values on the boundaries are higher than the depth in each of the adjacent cells (e.g. Figure 3(a)).

2 Revised Definition

Definition 3 Revised Simplicial Depth: *Given a data set $\mathcal{S} = \{X_1, \ldots, X_n\}$ in $\mathbb{R}^d$, the simplicial depth of a point x is the average of the fraction of closed simplices containing x and the fraction of open simplices containing x:*

$$SD_{BRS}(\mathcal{S};x) = \frac{1}{2}\binom{n}{d+1}^{-1}\Big(\sum I_{(x \in S[X_{i_1},\ldots X_{i_{d+1}}])} + I_{(x \in int(S[X_{i_1},\ldots X_{i_{d+1}}]))}\Big)$$

where int refers to the open relative interior[3] of $S[X_{i_1},\ldots,X_{i_{d+1}}]$. Equivalently, this can be formulated as: $SD_{BRS}(\mathcal{S};x) = \rho(\mathcal{S},x) + \frac{1}{2}\sigma(\mathcal{S},x)$, where $\rho(\mathcal{S},x)$ is the number of which contain x in their open interior, and $\sigma(\mathcal{S},x)$ is the number of simplices which contain x in their boundary.

[2] Zuo and Serfling divide the number of simplices enclosing a query point by n^3, while we use $\binom{n}{3}$. Counterexample 3 compares the depth of degenerate, multiple points and is not described here because data points are currently only partially treated under the revised definition(Section 3).

[3] See Edelsbrunner [5], page 401.

2.1 Properties of the BRS Definition

For continuous distributions and for positions lying in the interior of cells, the revised definition reduces to Liu's original definition. Significantly, the revised definition corrects irregularity at boundaries of simplices (Section 1.3 and Figure 3(a)), assigning the depth of a point on the boundary between two cells the average of the depth of the cells (Proposition 2). The Zuo-Serfling counterexamples [9] are also all solved by the revised definition (see Figure 1).

Lemma 1 *The simplicial depth of any two positions in the same cell is equal* [4].

Proposition 2 *The simplicial depth of a position on a facet between two cells is equal to the average of the depths of a position in the two adjacent cells, assuming that only d points lie on the hyperplane defined by the facet.* [4]

Corollary 3 *For a data set $\mathcal{S} = \{X_1, \ldots X_n\}$ in general position, the median value is attained in the interior of a cell or at a data point.* [4]

Proposition 4 *The depth of the position at the intersection between two or more facets is equal to the average of the depths of two opposite cells* [5] *of the intersection point, assuming only d points lie on any hyperplane defined by the facets.* [4]

Proposition 5 *For a data set $\mathcal{S} = \{X_1, \ldots, X_n\}$ in $\mathbb{R}^d$, in general position, the ordering of data points by their simplicial depth due to Liu's definition is unchanged by the revised definition.* [4]

2.2 Invariance Under Dimensions Change - Comparing $\mathbb{R}^2$ and $\mathbb{R}$

Consider the case where a data set $\mathcal{S}_a = \{X_1, \ldots, X_n\}$ ($n \geq 3$) consisting of a set of collinear points is analyzed as an $\mathbb{R}^2$ data set instead of as an $\mathbb{R}^1$ data set. Assume w.l.o.g. that these n points lie on the x-axis (see Figure 2(a)). Table 1 compares the depth assigned by Liu's definition and by the revised definition for $\mathbb{R}$ and $\mathbb{R}^2$, where m represents the number of data points which lie to the left of the position or data point under consideration.

In $\mathbb{R}^2$, for a degenerate simplex BCD, with C between B and D (see Figure 2(b)), the revised definition assumes that the point C and a position A, between B and C, lie within the open (degenerate) simplex BCD. Both points B and

[4] For proofs see [3]

[5] Two cells whose boundaries both contain a position θ lying on the intersection of two or more hyperplanes induced by the data set are *opposite cells* if and only if the two cells lie on opposite sides of every facet that contains θ. It can be shown that at each intersection, every cell has a unique opposite.

Data Points	Dimension	Definition
$\binom{n}{2}^{-1}(m(n-m-1)+(n-1))$	$\mathbb{R}$	Liu
$\binom{n}{2}^{-1}\left(m(n-m-1)+\frac{1}{2}(n-1)\right)$	$\mathbb{R}$	Revised Definition
$\binom{n}{3}^{-1}\left((n-m-1)\binom{m}{2}+m\binom{n-m-1}{2}+\binom{n-1}{2}\right)$	$\mathbb{R}^2$	Liu
$\binom{n}{3}^{-1}\left((n-m-1)\binom{m}{2}+m\binom{n-m-1}{2}+m(n-m-1)+\frac{1}{2}\left(\binom{m}{2}+\binom{n-m-1}{2}\right)\right)$	$\mathbb{R}^2$	Revised Definition
Position	**Dimension**	**Definition**
$\binom{n}{2}^{-1}(m(n-m))$	$\mathbb{R}$	Both Definitions
$\binom{n}{3}^{-1}\left((n-m)\binom{m}{2}+m\binom{n-m}{2}\right)$	$\mathbb{R}^2$	Both Definitions

Table 1: Depth of data points and positions for set $\mathcal{S}_a$

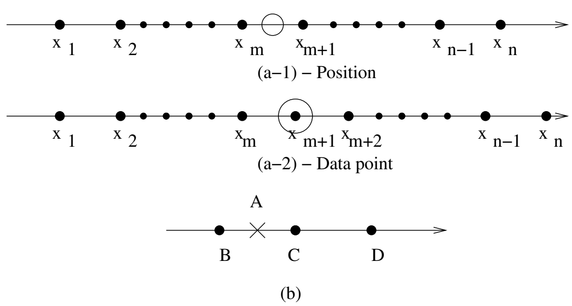


Figure 2: Data set S_a

D achieve a depth of $1/2$ as vertices of the simplex. Liu's definition assigns a depth of 1 to all points and positions on the interval $[B, D]$.

For both definitions, the ratio of the depths of a *position* when $\mathcal{S}_a$ is analyzed first as an $\mathbb{R}^1$ data set and second as an $\mathbb{R}^2$ data set is $\frac{2}{3}$. The same ratio for *points* rather than positions by Liu's definition is *not* identical to $\frac{2}{3}$. As desired, the ratio for data points under the *revised* definition is uniformly $\frac{2}{3}$, resulting in a scaling between the depths when analyzed by the $\mathbb{R}^1$ and $\mathbb{R}^2$. This property is desirable as discussed in Section 1, (P5). Although it holds for $\mathbb{R}^2/\mathbb{R}^1$, additional work is needed to generalize this for higher dimensions (Section 3).

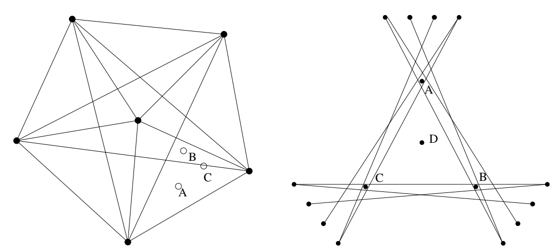
2.3 Existing Algorithms and the Effect of the Revised Definition

The revised definition maintains the time complexity in existing algorithms, after certain modifications.

- Existing $\Omega(n \log n)$ lower bound algorithms for computing the simplicial depth of a point or a position x relative to S in $\mathbb{R}^2$. [6, 1]

- Cheng and Ouyang's $O(n^2)$ algorithm [4] for computing the simplicial depth of x relative to S (based on [6]).

- Computing the depth of all n data points in $\mathbb{R}^2$ in $O(n^2)$ using the duality transform [6, 7].



(a) **(b)**

Figure 3: (a) Discontinuities in Simplicial depth at positions on facets: Position A, in an outer cell, has depth $SD(A) = \frac{4}{20}$; position B, in a neighboring cell, has depth $SD(B) = \frac{6}{20}$. However, position C on the boundary between the two cells has depth $SD_{Liu}(C) = \frac{7}{20}$, greater than the depth of A and B. (b) A problem with the revised definition. Data points A, B, and C all have depth $\frac{587}{1120}$, and data point D, at the center of the data set, has depth $\frac{355}{1120}$. For clarity not all cells are drawn.

- Aloupis *et al.* [2] algorithm for computing the simplicial median [6] in $\mathbb{R}^2$ in $O(n^4)$.

3 Open Problems

Maximality and Monotonicity Although the revised definition fixes many of the examples presented by Zuo and Serfling [9], it does not achieve all desired properties in the sample case. Figure 3(b) shows an example where the data set has a unique center, D, but it neither attains maximality at the center nor does it have monotonicity relative to the deepest point (properties P2, P3).

Data Points: The revised definition does not solve all the problems in sample data sets. For instance, each edge is inherently part of $n-2$ simplices while each point is part of $\binom{n-1}{2}$ simplices. A simple scaling does not take care of this problem as the depth of a point should be at least the minimum depth of all adjacent cells, which is not guaranteed simply by scaling. Thus the depth of a point should also depend on the geometry of the data set.

[6] The *simplicial median* is the point with the highest simplicial depth.

Higher Dimensions: For invariance under dimensions change (P5), the depth of a position in $\mathbb{R}^b$ when compared to $\mathbb{R}^d$ $(b > d)$ should be related by a multiple based on the number of b-dimensional simplices of which a d-dimensional simplex is a facet.

4 Summary

We present a modification to the definition of simplicial depth, that solves some of the problems raised in the past. We are currently investigating how to cope with the computation of the depth of data points in high dimensions, while maintaining the desirable properties, as described in Section 3. In addition we are working on approximation algorithms based on the local properties of the depth function, to enable efficient approximation for high dimensional data.

In [3] we present a connection between simplicial depth and halfspace depth. We believe that this relation can be further utilized to study the properties of the two depth functions and further improve algorithms' complexity.

References

[1] G. Aloupis, C. Cortes, F. Gomez, M. Soss, and G. Toussaint. Lower bounds for computing statistical depth. *Computational Statistics & Data Analysis*, 40(2):223–229, 2002.

[2] G. Aloupis, S. Langerman, M. Soss, and G. Toussaint. Algorithms for bivariate medians and a Fermat-Torricelli problem for lines. In *Proc. 13th Canadian Conf. on Comp. Geo.*, pages 21–24, 2001.

[3] M. Burr, E. Rafalin, and D. L. Souvaine. Simplicial depth: An improved definition, analysis, and efficiency for the sample case. Technical report 2003-28, DIMACS, 2003.

[4] A. Y. Cheng and M. Ouyang. On algorithms for simplicial depth. In *Proc. 13th Canadian Conference on Computational Geometry*, pages 53–56, 2001.

[5] H. Edelsbrunner. *Algorithms in combinatorial geometry*, volume 10 of *EATCS Monographs on Theoretical Computer Science*. Springer-Verlag, Berlin, 1987.

[6] J. Gil, W. Steiger, and A. Wigderson. Geometric medians. *Discrete Math.*, 108(1-3):37–51, 1992. Topological, algebraical and combinatorial structures. Frolík's memorial volume.

[7] S. Khuller and J. S. B. Mitchell. On a triangle counting problem. *Inform. Process. Lett.*, 33(6):319–321, 1990.

[8] R. Liu. On a notion of data depth based on random simplices. *The Annals of Statistics*, 18:405–414, 1990.

[9] Y. Zuo and R. Serfling. General notions of statistical depth function. *Ann. Statist.*, 28(2):461–482, 2000.