Research On Data Depth

Rasoul Shahsavarifar, David Bremner

Faculty of Computer Sceince, University of New Brunswick

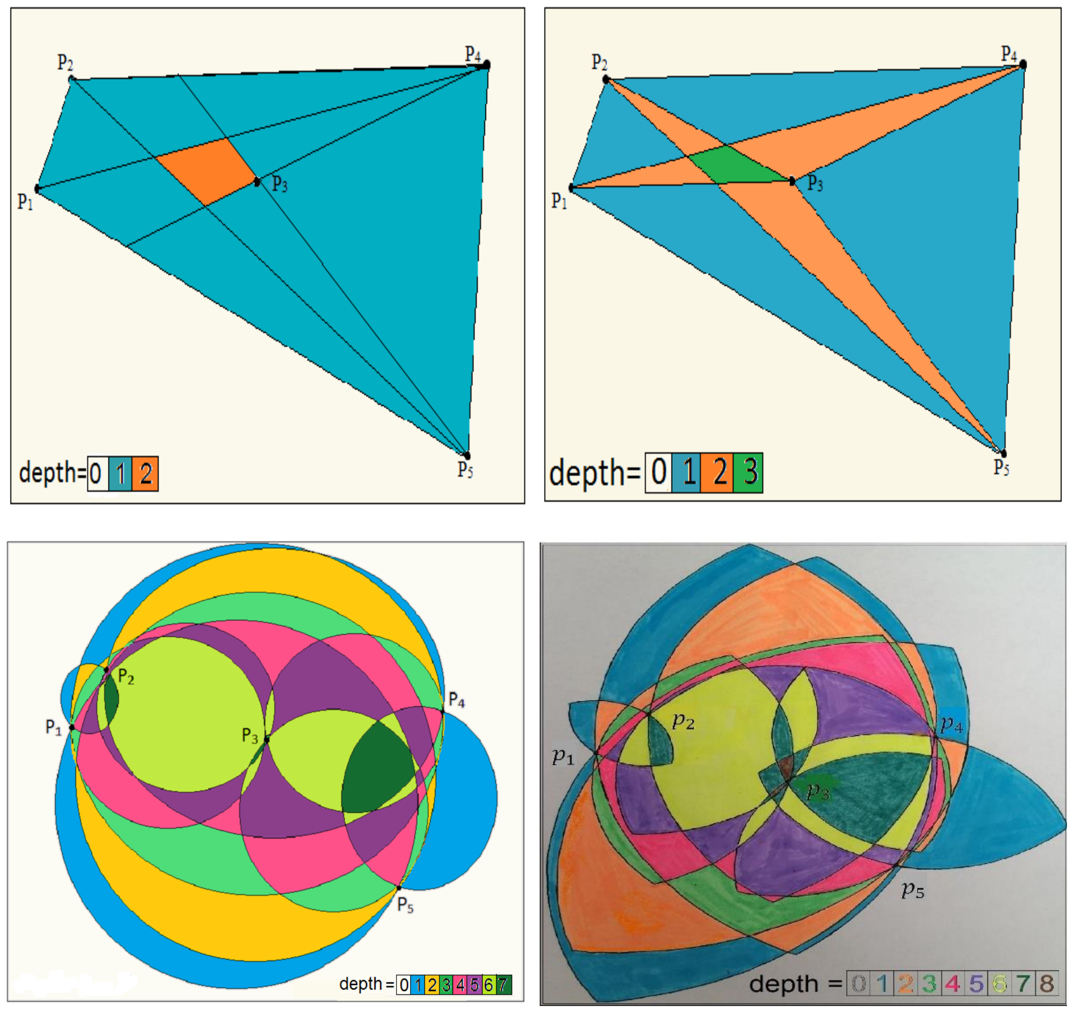
Introduction

Data depth, as a measure of centrality in multivariate data analysis, indicates how deep a point is located with respect to a given data set. The data depth is a convenient tool that can be applied to define an order among multi-variate data points, and classify them.

One major advantage of the data depth is that most of the depth functions are robust. This characteristic makes the data depth a remarkably appropriate tool in studying real life data, where there are a lot of outliers.

Different Notions

Over the last decades, various notions of data depth have emerged as powerful tools for multivariate data analysis. A few of them are: halfspace depth (Hotelling, 1929; Tukey, 1975), simplicial depth (Liu, 1990), spherical depth (Elmore, Hettmansperger, and Xuan, 2006), lens depth (Liu and Modarres, 2010), and others. The following figures respectively illustrate the halfspace depth, simplicial depth, spherical depth, and lens depth of points in the plane with respect to $S = \{p_1, p_{2,} p_{3,} p_{4,} p_5\}$.



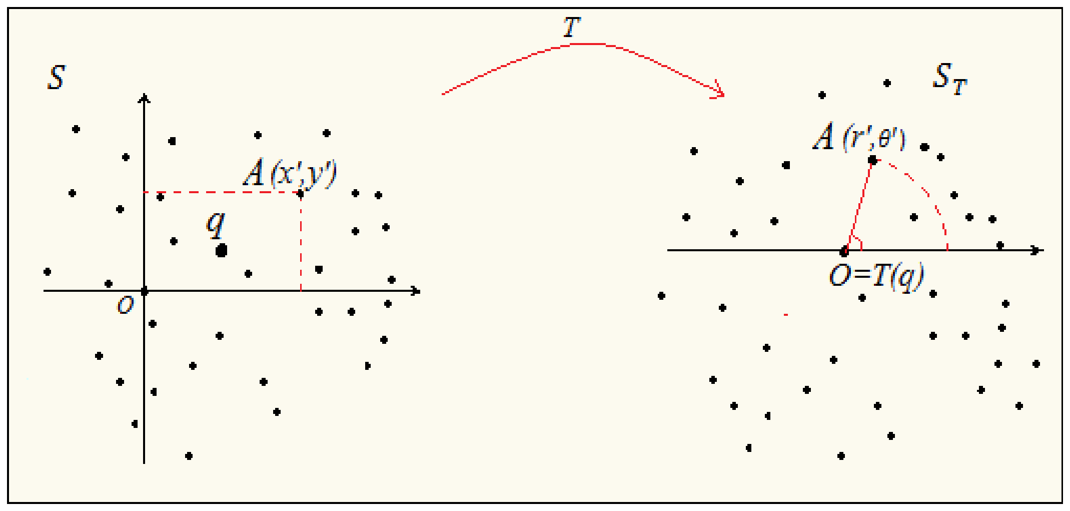
Ongoing Research

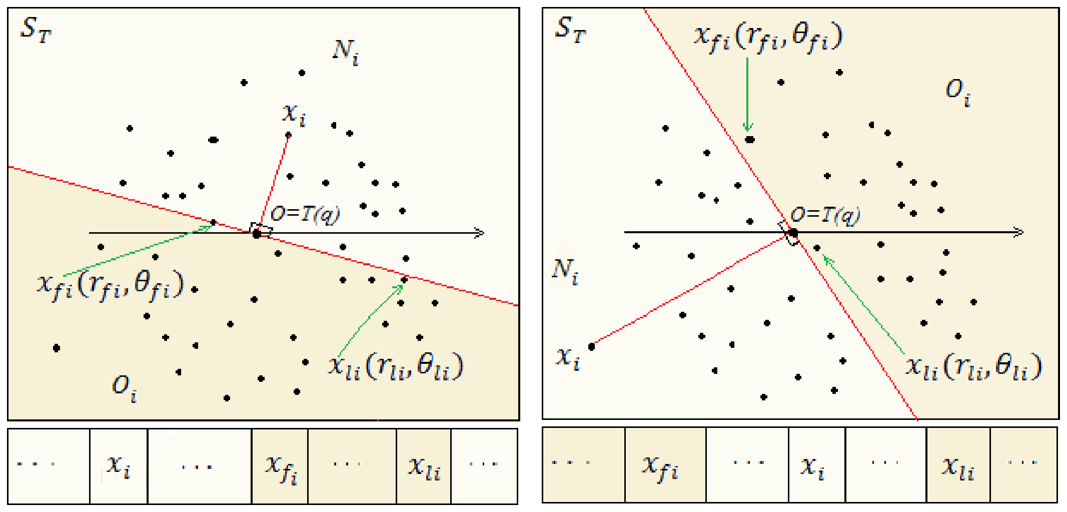
The ongoing research on data depth has different directions such as developing new algorithms, defining new depth functions, studying the data depth in high dimensions, studying the relationships between different data depth, etc.

We study data depth from both algorithmic and geometric points of view. We have achieved some results in both of these directions. Besides these results, we have done some experiments and we hope to make some theoretical progress to support our practical results. Our main result is developing an optimal algorithm for bivariate spherical depth of a point. The time complexity of this algorithm is $\theta(n\log n)$.

Optimal Algorithm for Spherical Depth

We develop an algorithm consisting of three procedures: Translating, Sorting, and Depth calculation which are illustrated in following figures.





S_T
N_i
x_i
$O=T(q)$
$x_{fi}(r_{fi},\theta_{fi})$
$x_{li}(r_{li},\theta_{li})$
O_i
N_i
$\cdots$ x_i $\cdots$ x_{f_i} $\cdots$ x_{li} $\cdots$
$\cdots$ x_{fi} $\cdots$ x_i $\cdots$ x_{li} $\cdots$

$\theta \notin (\frac{\pi}{2}, \frac{3\pi}{2}]$

$\theta \in (\frac{\pi}{2}, \frac{3\pi}{2}]$

The spherical depth of a query point $q \in R^2$ with respect to data set S in R^2 can be computed by:

$$SphD(q;S) = SphD(T(q);S_T) = SphD(O;S_T) = \frac{1}{2}\sum_{1\le i\le n}|O_i|,$$

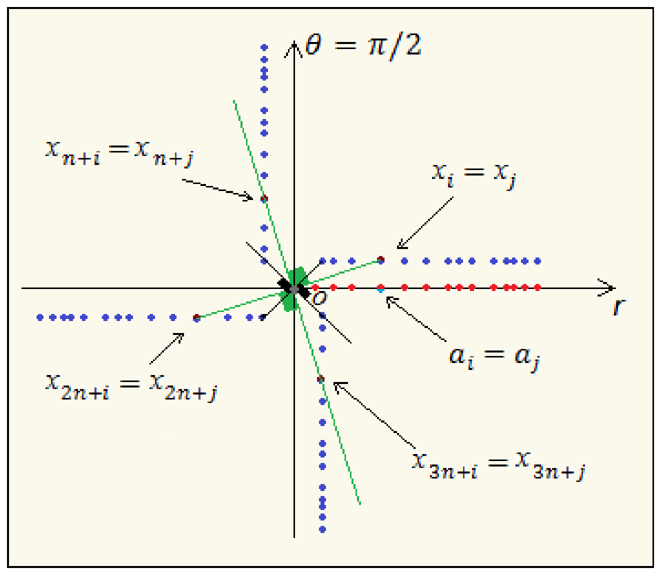
where

$$O_i = \left\{j \mid x_j \in S_T\,, \frac{\pi}{2} \le |\theta_i - \theta_j| \le \frac{3\pi}{2}\right\}.$$

The important part of the algorithm is figuring out the $|O_i|$, where we use two binary search calls to compute this desired value.

Lower Bound

By reducing the problem of Element Uniqueness, which has a lower bound of $\Omega(nlogn)$, to the problem of computing the spherical depth, we proved that the computation spherical depth of a point in the plane takes $\Omega(nlogn)$ time.



If all of the points in S are unique:

$SphD(q;S)$= $4n^2$+2n

If there exists some $i \neq j$ with $x_i = x_j$ in S:

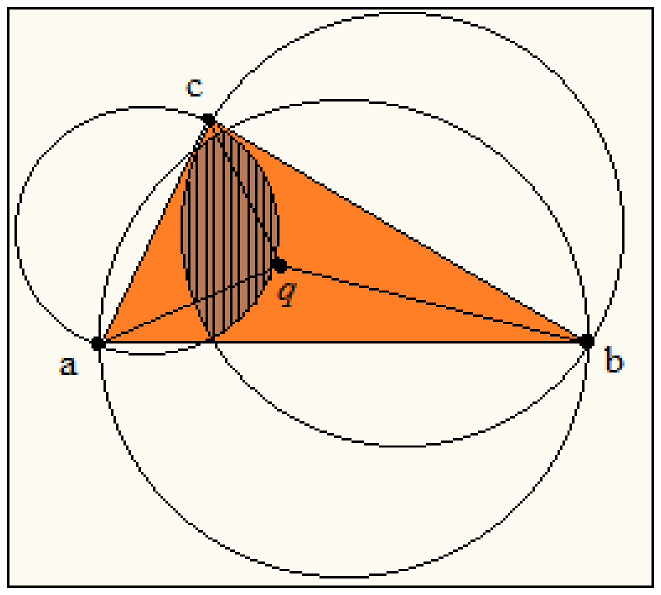
$SphD(q;S)$= $4n^2$+2n+4

Relationships

As a part of our research, we study the relationships between different notions of data depth. In this way, we bound one data depth in terms of another one. For example, we prove that there are some relationships between the influence regions of simplicial depth and spherical depth (see the following figure). We use these relationships and show that for every query point $q\epsilon R^2$ and data set S in R^2

$$SphD(q;S) \ge \frac{2}{3}SD(q;S),$$

where SD is stand for the simplicial depth.



Experiments

In this section, we present some experimental results to show the relationship between spherical depth and simplicial depth. The following table, indicates the spherical depth and the simplicial depth of the points in three random sets Q_1, Q_2, and Q_3 with respect to data sets S_1, S_2, and S_3, respectively. The elements of Q_i and S_i are some randomly generated points (double precision floating point) within the square $A = \{(x,y) | \; x,y\;\epsilon[-10,10]\}$. $|Q_1|$= 100, $|S_1| =$ 750, $|Q_2|$= 750, $|S_2| = 2500$, $|Q_3| = 2500$, and $|S_3| = 10000$. As can be seen, the experimental results are in consistent with the theoretical results. In fact, these results suggest a stronger bound (i.e. $SphD(q;S) \ge 2SD(q;S)$) than the theoretical bound.

	$(q_1;S_1)$		$(q_2;S_2)$		$(q_3;S_3)$	
	Min	Max	Min	Max	Min	Max
SD	0.00	0.25	0	0.25	0.00	0.24
$SphD$	0.01	0.50	0.00	0.50	0.00	0.50
$\frac{SphD}{SD}$	2.00	∞	2.00	∞	2.02	∞

Reference

- Bremner, D. and Shahsavarifar, R., 2017. An Optimal Algorithm for Computing the Spherical Depth of Points in the Plane. *arXiv preprint arXiv:1702.07399.*

Contact Info

Email: Ra.Shahsavari@unb.ca, Bremner@unb.ca

Faculty of Computer Science, University of New Brunswick,
Fredericton, NB, Canada