Data Depth and Computational Statistics

1 Introduction

Statistics gives specific information about collected data samples, such as height experiment, and patients' blood pressure results. If we have a graph that has patients' blood pressure and heart rate on each axis, what kind of questions can we ask? "what is the probability that a patient has blood pressure less than 90?" " what is the typical value of the graph?" could be some questions we can answer. So probability distribution, including concepts such as average, median, variance, explains information well.
There are two types of analysis on Statistics: one is **graphical analysis**, which involves plotting, looking at the data, and see how well it behaves (e.g. what are outliers), and the other is **quantitative analysis** that deals with numbers.
Classical analysis usually requires prior assumptions about the distribution model (models such as normal distribution, fibonacci function, and uniform distribution). Today we want to look at somewhat different way of analysis that doesn't require prior assumptions of the underlying distribution model.

2 Data Depth [1]

Definition 2.1 *Data depth is a way of measuring how deep (central) a given point is relative to the data set.*

Consider, for example, a data set for diabetic patients represented as a number of points on a plane, where the x-axis values represent their blood pressure and y-axis their heart rate. By looking at the points, we can see how a given point is central to the data set. The concept of data depth gives a way of both graphical and quantitative analysis.
Several ways of measuring data depth exist, and we will concentrate on three of them.

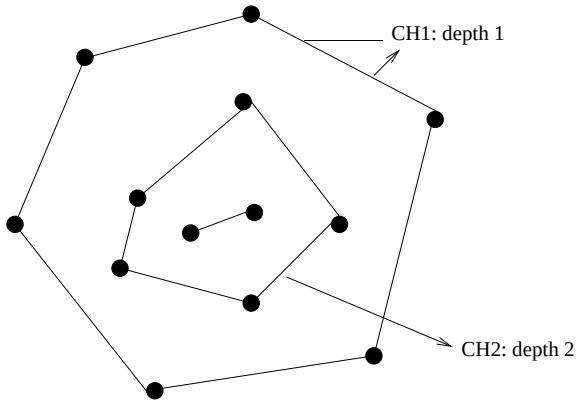

Figure 1: Convex Hull Peeling Depth

2.1 Convex hull peeling depth [7]

Given a number of points, we can compute a convex hull, CH1. The depth of the points that are on CH1 is 1. Now compute another convex hull CH2 on the points excluding the points on CH1. Then the depth of the points on CH2 will be 2(figure 1).
How many depths values can a data set have? The depths can range from 1 (all the points are on convex hull) to $n/3$ (each convex hull contains 3 points).
Some convex hull peeling depth-related questions are: "How much time does it take to compute all the convex hulls?" "How much time does it take to compute points with depth k or more?"

2.2 Half-space depth (location depth)

Definition 2.2 *Half space of a point p relative to a set $S = S_1, ..., S_n$ is the minimum number of points of S lying on any closed half space determined by a line through p.*

So the half-space(HS) depth of the points on the convex hull of the set is 0, the same as in Convex hull(CH) peeling depth. But for other points, the HS depth can be different from CH depth. For example, look at the figure 2. The CH depth of the point p is 2, but its HS depth is 1.

Think the half-space depth as following analogy: "I am trying to meet a king. As I approach to the king from each side, how many guards will I

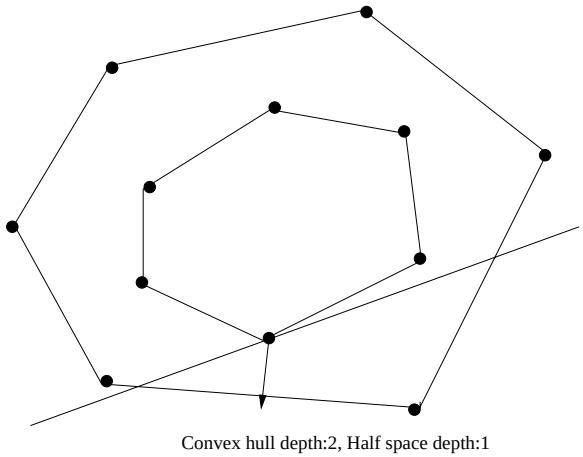

Figure 2: Half-space depth vs. Convex hull depth

encounter before I can meet the king?"
The minimum value of half-space depth is 0 (for any point outside the outer CH), and the maximum value is between $n/3$ to $n/2$ (for example, figure 3, where all the data points are on the CH).
Then, how much time does it take to compute a half-space depth for a given point p relative to a set of points $X_1, ..., X_n$? Note that we don't need to consider all the lines passing through point p, only those created by connecting p and the points of the data set, total of n lines. A brute-force algorithm would take $O(n^2)$, if we check how many points are on one side or other side for all the lines. But a better algorithm,$O(nlogn)$ exists for the problem.(homework8)

2.3 Simplicial depth

Definition 2.3 *Simplicial Depth of a point p relative to S in R^d(dimension d) is the number of d simplicies created by points of S that contain p.*

A simplex in R^2 is a triangle. So in $2-D$, a simplicial depth would be the number of triangles that a given point is included(figure 3). An analogy for $2-D$ simplicial depth is that "when I approach a king, how many triplets of guards does the king have?"

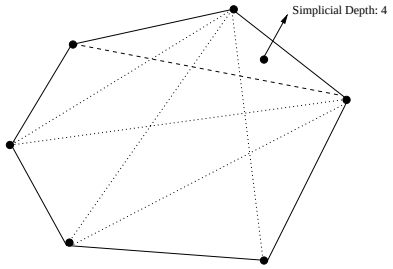

Figure 3: Simplicial Depth

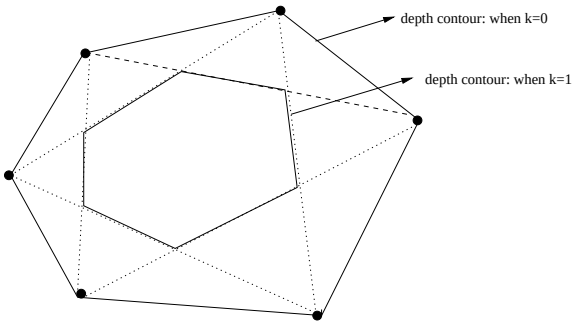

Figure 4: Depth Contour when k=0 and k=1

3 Depth Contours

Definition 3.1 *For a fixed positive integer k, the set of points in the plane with half-space depth ¿= k is a polygonal region whose boundary is the k^{t}h depth contour.*

For example, when $k = 0$, the depth contour of half-space depth would be the outer convex hull in figure 4. When $k = 1$, it would look like the inner convex hull in figure 4.

Note that the vertices of the depth contours are not necessarily points of the original data set.

Depth contours are especially powerful for visualizing and quantifying data. Simple (in many cases $2D$) graphs can be used to visualize these parameters for the data set. The concept of data depth has enormous potential for analysis of massive data sets.

1. Contours of a data set are good representation of the data set.

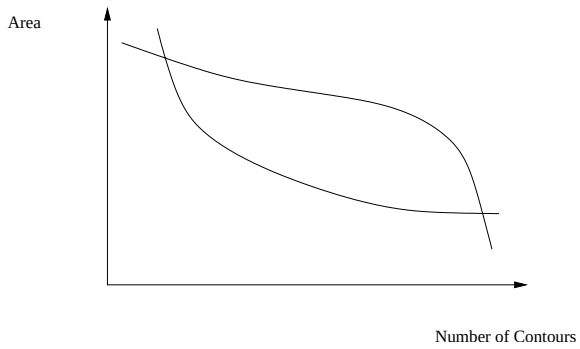

Figure 5: Data Representation

2. We can also ask how skewed our data using a graph of the area of each contours vs. its number (or the percentile of points it contains) (see figure 5). A graph that tends to 0 earlier means that the data set is more centered.

3. Outliers are mostly outside first 1-3 contours. We can compute the first few depth contours and remove their data points, as a tool to remove outliers from the data set.

4. DD-plots (Depth vs. Depth plots) are used to compare two data sets L and R. Given the two data sets, the DD-plot is the plot of the depth values of each point from the combined set LUR, relative to the contours of set L and relative to the contours of set R (figure 6). If both data sets are from the same distribution, we would expect to see a 45 degree line. Changes in the relationship between the two distributions will result in changes in the DD-plot [1].

4 Computing Half-Space depth contours

Lower bound: In the worst case, if all the points of the data set are on the CH, we will have contours of sizes $n, n-1, n-2$ to 1. The total size of the contours will be $O(n^2)$, which means it takes at least $O(n^2)$ to compute all contours.
By taking dual of each point in set S, we change the problem from finding the depth of a point o to finding a convex hull of line relative to an arrangement

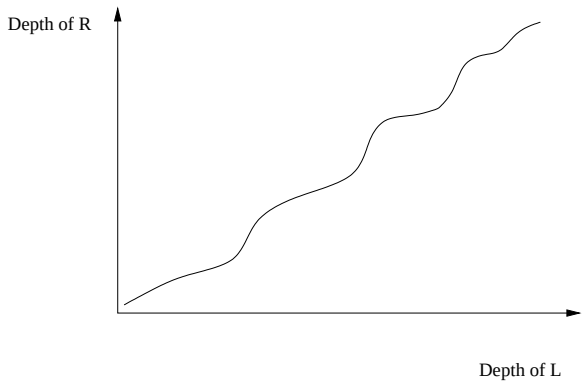

Figure 6: DD-plots

of lines. Figure 7 shows how we can compute half-space depth contours by using topological sweep. Figure 8 shows how topological sweep is done to get the candidate points of each level.

Algorithm [4]

1. Take dual of n points.
2. Use topological sweep $O(n^2)$ to get lines in sorted order by level.
3. Walk and just look at the points on the same level, Computing the convex hull of the primal points $O(n)$.
   1) Find upper chain.
   2) Find lower chain.
   3) Merge upper and lower chain, and check if the contour exists.

So it takes $O(n^2)$ to find all depth contours, which matches the lower bound. Step 3 is necessary because sometimes the contour doesn't exist because we are guaranteed to have $n/3$ contours, not $n/2$.

5 Simplicial depth contours

When it comes to Simplicial depth contours, the 'behavior' of the contours is not as nice as this of CH peeling depth or half-space depth contours. For example, Simplicial depth Contours may not be nested, as we can see in

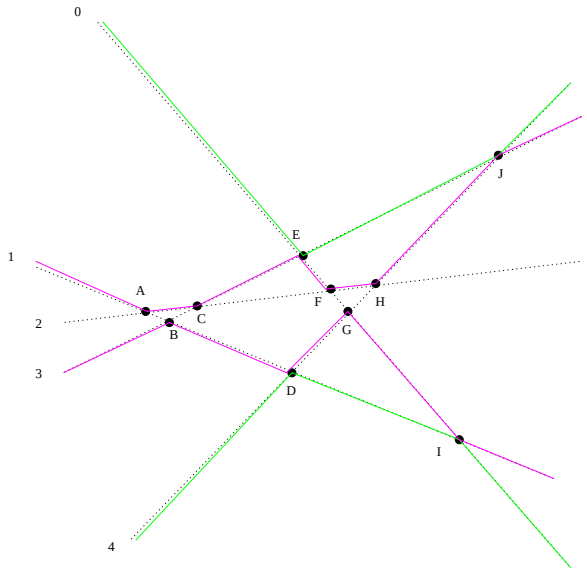

Figure 7: Computing Half-Space Depth Contours using topological sweep

level \ sweep order	A	B	C	E	F	D	G	H	J	I
0				E					J	
1	A		C	E	F			H	J	
2	A	B	C		F		D	H		
3		B				D	G			I
4						D				I

Figure 8: Computing Candidate Points

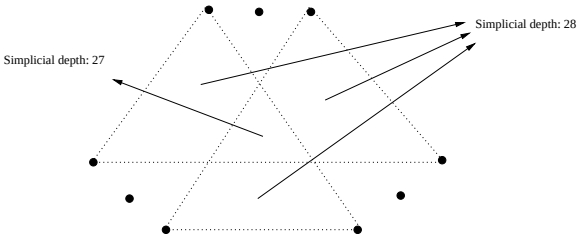

Figure 9: Simplicial contours may not be nested

figure 9 (the contour that encloses all points of depth 28 and up, forms a torus and is not nested in the contour enclosing depth 27 and up).

References

[1] website, *http*: $//www.eecs.tufts.edu/r/geometry/data-depth/$.

[2] *"On a notion of data depth based on random simplices"*, Liu, R. The Annals of Statistics (18) 405-414,1990.

[3] *" Efficient Computation of Location Depth Contours by Methods of Combinatorial Geometry"*, K. Miller, S. Ramaswami, P. Rousseeuw, T. Sellares, D. Souvaine, I. Streinu, A. Struyf. Statistics and Computing, 2003,.

[4] *"Fast implementation of depth contours using topological sweep"* K. Miller, S. Ramaswami, P. Rousseeuw, T. Sellares, D. Souvaine, I. Streinu, A. Struyf. Proceedings of the Twelfth ACM-SIAM Symposium on Discrete Algorithms, Washington, DC, January, 2001.

[5] *"Regression depth"*, Rousseeuw, P. J. and M. Hubert, J. Amer. Statist. Assoc. (94),388-433, 1999.

[6] *"Mathematics and the picturing of data"*, Tukey, John W.,Proceedings of the International Congress of Mathematicians,Vancouver, B. C., 1974, Vol. 2, 523–531.

[7] *"Convex hull peeling"*, Eddy, W in "COMPSTAT", 42-47, 1982.