

# CUSP Progress Report

Analyzing Data Depth of Multimodal Distributions

Robby Ramdin\*

January 22, 2008

The Computing Undergraduate Scholar Program (CUSP) is an endeavor by the National Science Foundation to encourage research by undergraduates in Computer Science and the related fields. A student participating in CUSP is required to select a research topic and attempt to find a solution. Other responsibilities include weekly meetings of all the students participating in the program with graduate advisors where research and classes are discussed.

For my CUSP research, I am investigating a facet of Computational Geometry called data depth. Data depth is a statistical device used for analyzing large multi-dimensional data sets. The Tufts Computational Geometry group defines it as follows:

Statisticians have recently developed the notion of data depth for non-parametric multivariate data analysis. This new concept provides center-outward orderings of points in Euclidean space of any dimension and leads to a new non-parametric multivariate statistical analysis in which no distributional assumptions are needed. [8]

The problems I am looking at are those related to multimodal data, that is data that does not conform to normal distributions (Gaussian distributions or, in two dimensions, bell curves). I am looking into building on existing algorithms to detect if data is multimodal in addition to investigating creating my own. Another compelling facet of my research is working on new ways of expressing data depth visually. A common and intuitive visualization is depth contours. As is evident in **figure 1**, contours work very well for normal distributions (and also uniform distributions), but fail in the multimodal case. I am also putting effort into properly drawing contour lines for multimodal data.

My advisor is Diane Souvaine, and I have also been working with John Hugg who has authored the Depth Explorer application.

## Motivation and Applications

Data Depth is a measure of centrality of a point in a given distribution. It can be thought of as a statistical metric similar to standard deviation. The standard deviation, however, has a number of shortcomings, namely that it only considers distance from the mean. Data depth, on the other hand, can not only handle data that does not conform to a strict normal distribution, but it also performs well with skewed distributions. When data is gathered in a situation where the

---

\*robby.ramdin@tufts.edu

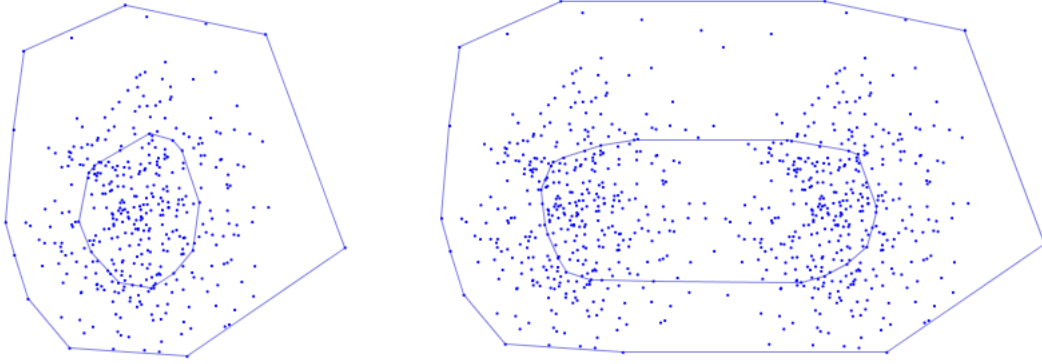


Figure 1: Contours showing depth in a normal distribution (left) and a multimodal distribution(right)

underlying distribution is unknown, depth measures perform very well, as they only consider the relative locations of instances [7].

Applications of data depth to machine learning are also appealing. Data depth allows for easy boolean classification of instances, and can give an indicator of confidence in its classification (the deeper a point is, the higher the confidence.). For example, given the body mass index and blood pressure of heart attack victims, we could calculate the depth for this set and predict whether future instances are heart attack risks.

Another interesting application is applying data depth to visualizations of large data sets. In a master's thesis for the University of Helsinki, Jukka Pekka Kontto suggests that data sets can be visualized more efficiently by representing the structure of the distribution rather than drawing each point individually. He accomplishes this by first applying a depth measure to ascertain the location and shape of the data, using Half-Space Depth [4].

## Background Research

I began this research by researching the background of data depth analysis and the principles of computational geometry, as I have had no experience in this field. I researched basic algorithms for calculating convex hulls and creating triangulations and Voronoi diagrams [6]. I also read up on different depth measures that are commonly used, including Convex-Hull Peeling Depth,  $L_1$  Depth, Half-space Depth, and Proximity Depth [3]. Of these depth measures, Proximity Depth seems like the best for analyzing depth in multimodal data (or at least bimodal data).

The other major piece of preexisting solutions is the Depth Explorer, an application written in Java at Tufts to aid in the visualization of depth measures. The advantage to having a visualization becomes clear when small parameters need fine tuning (such as changing  $\beta$  in  $\beta$  skeletons). It is helpful because a researcher can instantly estimate the efficacy of his algorithm.

## Fall Workshop on Computational and Combinatorial Geometry

In early November, I attended a conference on computational geometry sponsored by the NSF and IBM. While most of the material was over my head and unrelated to data depth, it was a valuable

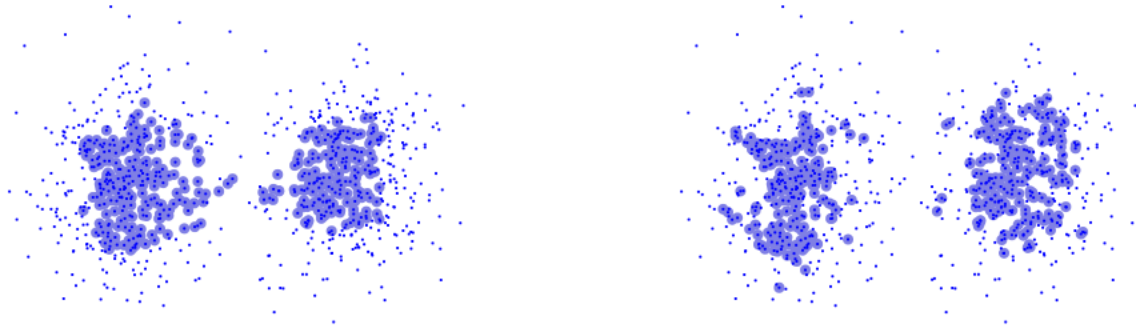


Figure 3: Proximity Depth (left) compared with my depth measure based on distances to neighboring points in the triangulation (right), with the deepest 50% of points highlighted. Proximity Depth is clearly superior.

learning experience. I got to see all the creative ways that concepts of computational geometry can be applied. I was also exposed to tons of new concepts.

## Current Progress

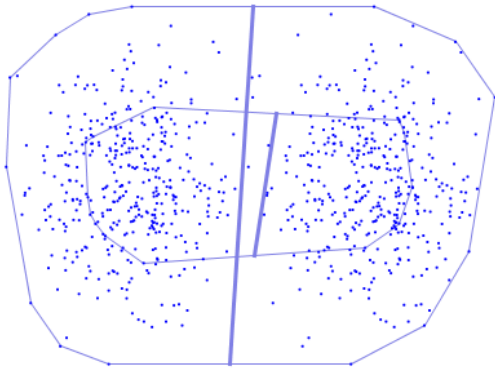


Figure 2: A line that bisects the data may be useful for drawing better contours.

exploring drawing a line through the data and dealing with points on either side independently. In a normal distribution, I have had moderate success with the line the bisects the two longest edges of a convex hull (figure 2).

Since I spent most of this semester getting oriented in both Computational Geometry and the Depth Explorer application, my progress thus far has been mostly intellectual, not tangible. I experimented with measuring depth based on areas of triangles and lengths of edges in a triangulation with moderate results. It has the significant problem that it fails on uniformly distributed data. While that exercise was not incredibly groundbreaking, it let me learn how to create new classes that work with Depth Explorer.

I decided to turn my work to better drawing contour lines on multimodal data. My aim is to find a way to split up each cluster and draw the contour lines more effectively. I am currently

## Future Plans

My plan is to spend a significant amount of time over winter break working on developing code, reading more about data depth, and writing about what I'm doing. In the spring, I will be taking Comp 163: Computational Geometry, taught by Diane. Taking a class in the material will probably

highlight different algorithms that I have dismissed as irrelevant. I am also planning to write a significant body of work on the subject, in the form of a Senior Honors Thesis. As one of the qualifications for that is to complete a chapter of my thesis by the end of the semester, I will turn my immediate attention to finishing that.

## References

- [1] Brad Barber. *QHull Manual*. Cambridge, MA, 2003.
- [2] Mark de Berg, Marc van Kreveld, Mark Overmars, and Otfried Schwarzkopf. *Computational Geometry: Algorithms and Applications*. Springer-Verlag Berlin Heidelberg, New York, 1988.
- [3] John Hugg, Eynat Rafalin, Kathryn Seyboth, and Diane Souvaine. An experimental study of old and new depth measures. 2006.
- [4] Jukka Pekka Kontto. Visualizing large epidemiological data sets using depth and density. Master's thesis, University of Helsinki, 2007.
- [5] Hyunsook Lee. *Two Topics: A Jackknife Maximum Likelihood Approach to Statistical Model Selection and a Convex Hull Peeling Depth Approach to Nonparametric Massive Multivariate Data Analysis*. PhD thesis, Pennsylvania State University, 2006.
- [6] James O'Rourke. *Computational Geometry in C*. Cambridge University Press, Cambridge, UK, 1998.
- [7] Eynat Rafalin, Kathryn Seyboth, and Diane Souvaine. Proximity graph depth, depth contours, and a new multimodal median. 2005.
- [8] Tufts University. Comp geo at tufts – data depth, December 2007.