Data Depth

Charlotte McFarland
&
Brendon Providence
Most analysis requires some assumption about the shape of the data
  i.e. uniform, bell curve etc.
It would be nice to not need to make such assumptions as they can be incorrect
We need a method that does not depend on the overall shape of the data
Ideally, Data Depth is a method for this type of data analysis.

- The behavior of the data should not affect the results of analysis.
- The depth of a point is a measure of how deep it is with a cloud of points.
What is Data Depth?

- Data Depth is measured by observing how "deep" a point is relative to the rest of the data.
- Taking the convex hull is a simple example.

To compute the convex hull, imagine putting a rubber band around all the points.
- The points that the band touches are removed and the next level will be found the same way.
Limitations of Data Depth

- Works great with two dimensions, however, efficiency severely drops with more dimensions.
- In higher dimensions, more and more of the points lie on the convex hull reducing the usefulness of the data.
- Not much use if 80% of the data is on one level.
Practicality

- Most of the research has been done with normal (bell curve) distributions
  - A normal distribution does not have any limits and trails off infinitely
- Need to study more practical applications like truncated normal
  - In a truncated normal distribution, the data is restricted to a certain range
What do we do?

- We are writing a program to test the percentage of points that appear on the convex hull for a given set.
- We would like to run this on both normal and truncated normal.
Our Current Results

Convex Hull Points in Various Dimensions

% of points on L

points
What Does this show?

- As seen in the previous slide, in 8 dimensions, over 95% of the data lies on the convex hull.
- However, we can see that with more points, the data approaches some constant asymptote.
Future Work

- More coding to test for more cases
- See if there is a way to predict the asymptotic value that the data approaches
- Look for other methods of computing data depth that don’t eliminate as much data at once