Geometric and Computational Aspects of Data Depth

David Bremner , Rasoul Shahsavarifar

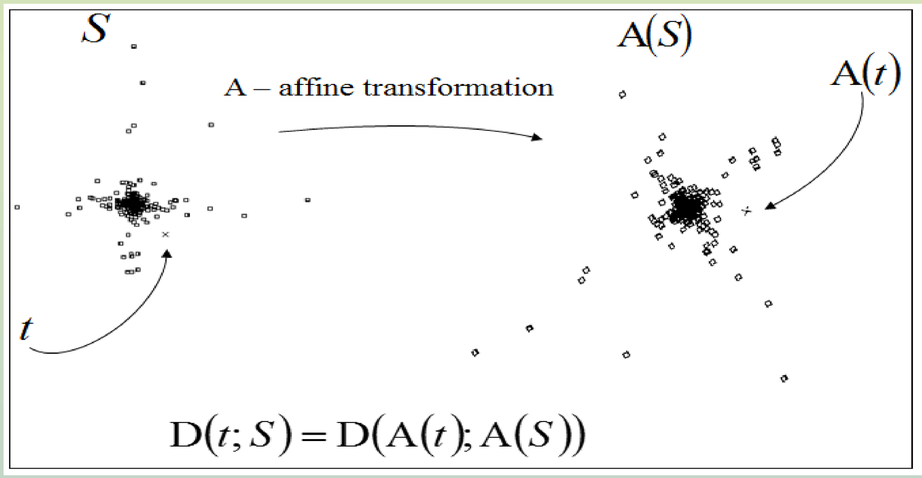
University of New Brunswick, Faculty of Computer Science

Introduction

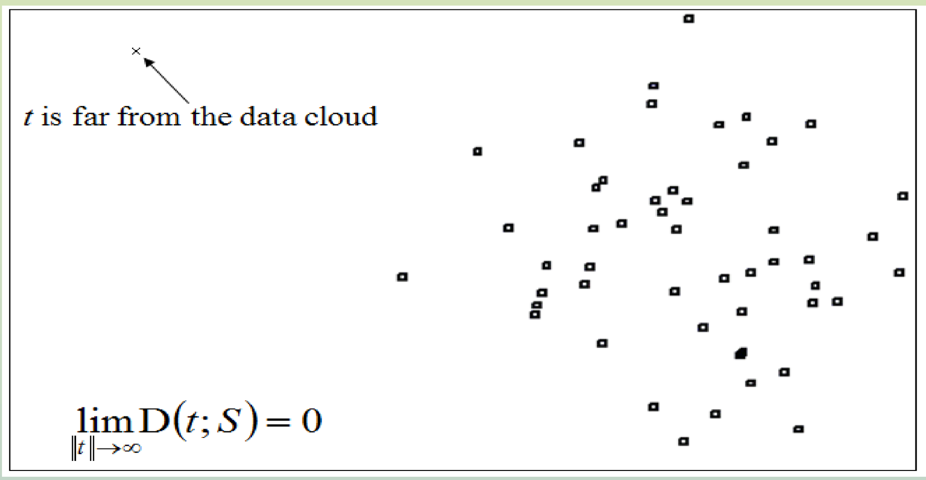
- Rank statistic tests play an important role in univariate nonparametric statistics. If one attempts to generalize rank tests to the multivariate case, the problem of defining a multivariate order statistic will occur. This problem can be attributed to the absence of a natural linear order on R^d. Thus is not clear how to define a multivariate order or rank statistic in a meaningful way.
- One approach to overcome this problem is to use the notion of data depth. Data depth measures the 'centrality' of a point in a given data cloud in non-parametric multivariate data analysis. In other words, it indicates, in some sense, how deep a point is located with respect to the data cloud. Using data depth, a multivariate order statistic can be defined by ordering the data points according to their depth in a suitable data cloud.

Framework of Data Depth

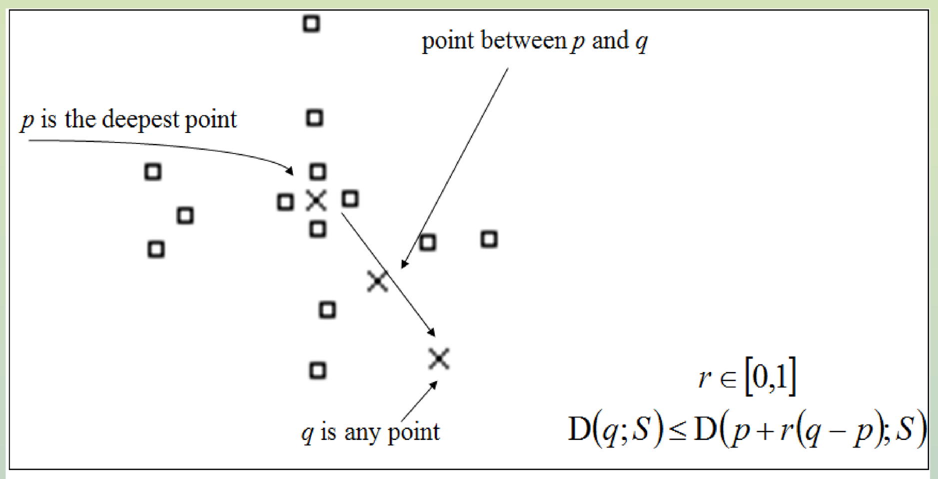
1. Affine invariance



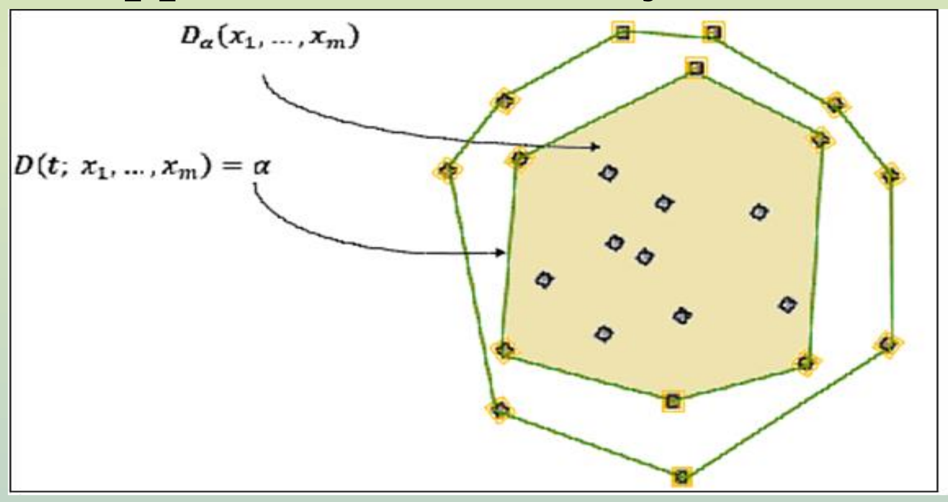
2. **Vanishing at infinity**



3. **Monotone on rays**



4. **Upper semi-continuity**



Different Notions of Data Depth : Among many different date depths, our research is mostly focused on the following data depths

1. Halfspace depth: $HD(t;S) = inf\{P(S \cap H);\ \forall H \text{ containing } t\} \rightarrow$ Figure1
2. Oja depth: $OjD(t;S) = \frac{1}{\binom{n}{d}} \sum_{(X_1,\dots,X_d)\in\binom{S}{d}} Vol_d(conv\ \{t, X_1, \dots, X_d\}) \rightarrow$ Figure2
3. Simplicial depth: $SD(t;S) = \frac{1}{\binom{n}{d+1}} \sum I(t \in S\,[X_1, \dots, X_{d+1}]) \rightarrow$ Figure3
4. Majority depth: $MjD(t;S) = \frac{1}{\binom{n}{d}} \sum I\,(t \in MjS\ (X_1, \dots, X_d)) \rightarrow$ Figure4
5. Lens depth: $LD(t;S) = \frac{1}{\binom{n}{2}} \sum_{i<j}^{n} I(t \in L(\,X_i, X_j)) \rightarrow$ Figure5

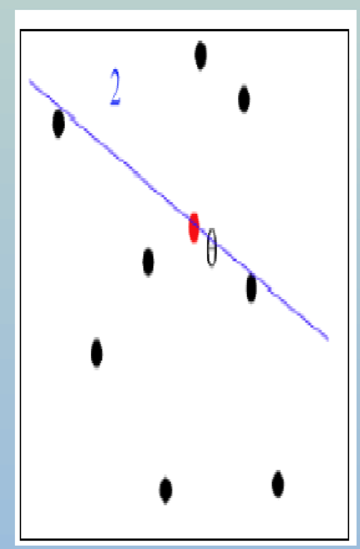


Figure1

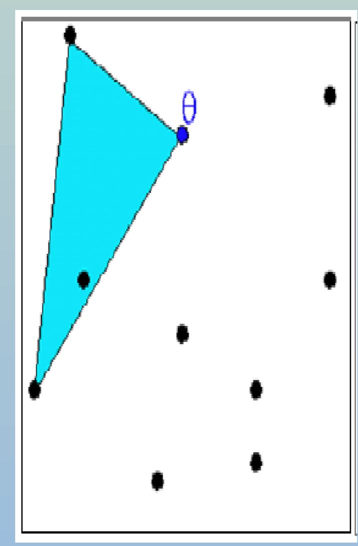


Figure2

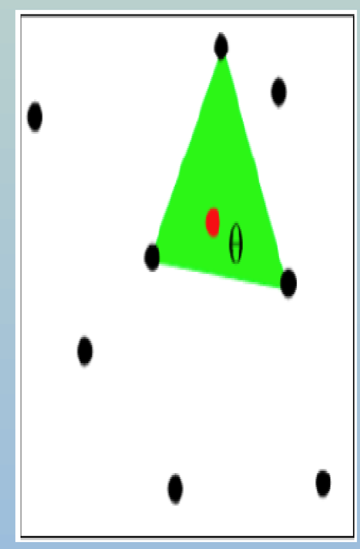


Figure3

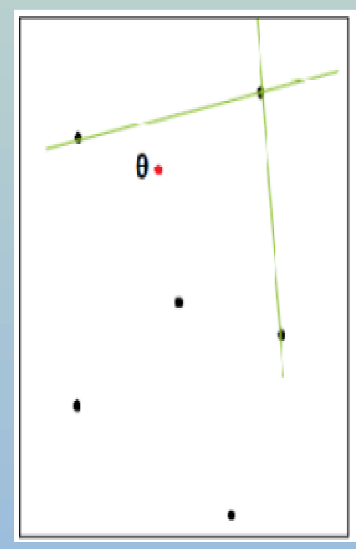


Figure4

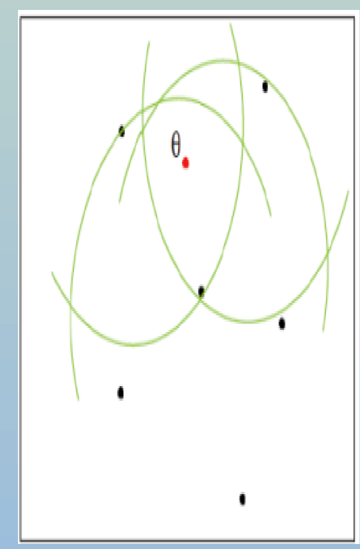


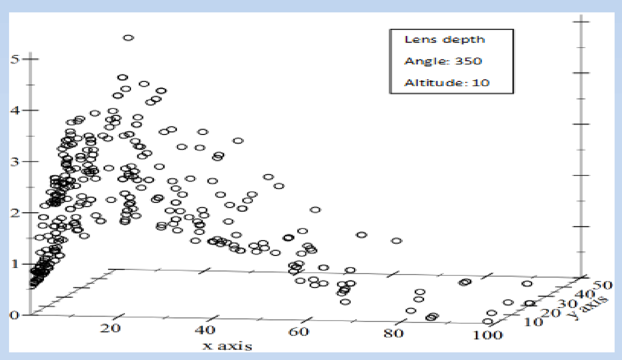
Figure5

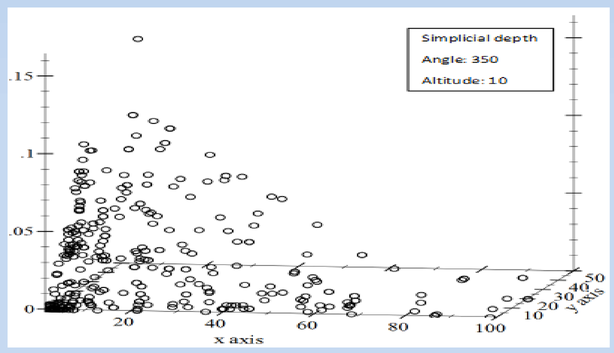
Time Complexity

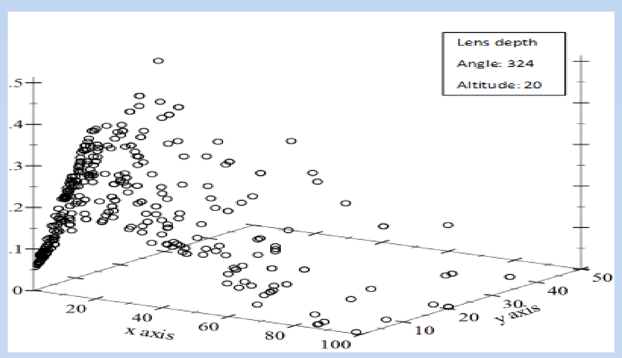
Depth Function	Depth Time Complexity	Bivariate Median Time Complexity
Halfspace depth	$O(m^{d-1})\log(m)$	$O(m \log^3 m)$
Oja	$O(m^d)$	$O(m^3 \log m)$
Majority		$O((n+m)\ logn)$ *(Brodal –Jacod model)* , $O((n+m)\ logn/\log(logn))$ *(RAM model)*
Simplicial	$O(m^{d+1})$	$O(m^4)$
Lens	$O(m^2 d)$	$O(m^2)$

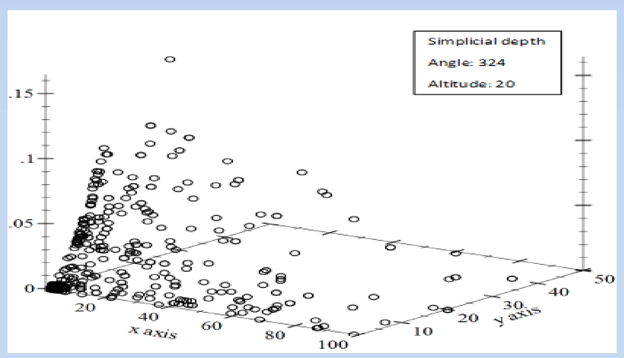
Implementation

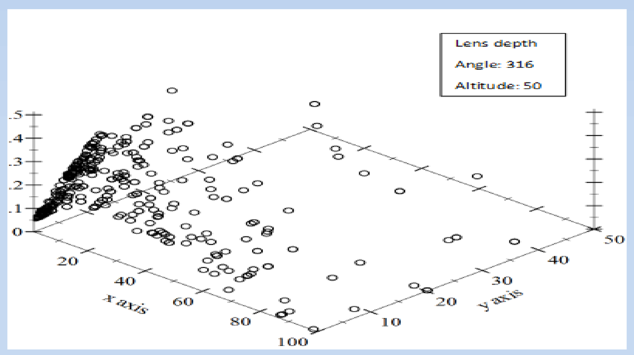
We implemented the Lens, Simplicial and Oja depth for a set of bivariate points Theta with respect to a given dataset S in order to compare different data depths . Following are the visualization of the output for random some data.

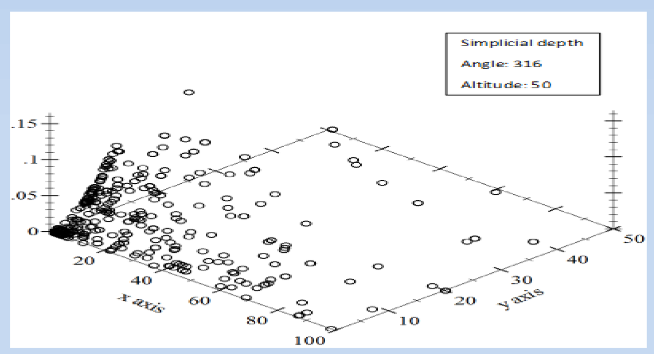












Future Work

Our research is currently focused on the following areas.

1- Understanding the geometrical aspects of the data depths.

2- Investigating some bounds and improving time complexity for computing data depths