A SIGNIFICANT STUDY ON ROBUST MEASURE OF LOCATION PARAMETERS USING DATA DEPTH APPROACHES

Kalaivani S
•
Assistant Professor
Department of Statistics and Data Science
Christ University
Bangalore, India
kalaivanistatistics1994@gmail.com

Abstract

Data depth procedures are statistical methods used to measure the centrality or depth of a point within a multivariate dataset. These procedures provide a way to quantify how deep or outlying a point is relative to the overall distribution of the data. This study explores various data depth procedures to find reliable location estimations in cases like with and without outliers. In this paper, various depth procedures, such as Mahalanobis depth, Halfspace depth, Euclidean depth, Simplicial depth, and Projection depth, are studied and compared. The efficiency of these depth functions is evaluated using real datasets and simulation studies with R software.

Keywords: data depth, robust procedures, inference, outliers

I. Introduction

Robust statistics is a fundamental branch of statistical theory and methodology designed to address the challenges posed by data containing deviations from standard assumptions. These deviations may include outliers or non-normality in the data. Robust statistics prioritizes methods that are insensitive to small outliers, which are a common occurrence in traditional statistical techniques. It aims to yield precise and reliable results even when the assumptions of classical statistics are not fulfilled. Robust statistical methods have been developed for many common problems such as estimating location, scale and regression parameters. The data depth approach is one of the robust statistical methods that measures the depth of a data point in a multivariate dataset. It determines the depth of a point by its distance from the center of the data, with points closer to the center having a higher depth value. This approach is useful for identifying outliers and robustly estimating location and scatter. These methods can be applied to both univariate and multivariate datasets, providing robust estimates of location and scatter [2].

The rest of the paper is structured as follows. The second section provides a concise overview and definitions of different data depth procedures. In the third section, the findings from numerical study conducted in both real datasets and simulated environments are presented. Finally, the paper concludes with a discussion the last section.

II. Data Depth Procedures

Data depth procedures are an innovative approach in robust statistics designed to measure the centrality of points within a data set, especially in multivariate contexts [3]. Depth assigns an integer to a candidate fit relative to a data set, enabling a center-outward ordering of sample points [6]. Unlike traditional order statistics, which rank data from smallest to largest, depth order statistics start from the center and move outward [8]. This center-outward approach is crucial for multivariate data sets, extending univariate concepts to multivariate analysis and allowing nonparametric methods to be used in multivariate data analysis [9]. This concept is particularly useful when dealing with complex data structures where classical methods may falter due to the presence of outliers or deviations from model assumptions.

The applications of data depth procedures are vast and varied, encompassing robust location estimation, multivariate outlier detection, classification, and data visualization. The data depth procedures used in this study is detailed below.

Mahalanobis Depth (MD)

Mahalanobis (1936) [7] introduced the Mahalanobis depth in robust statistics which measures the centrality of a point within a multivariate data set by using the Mahalanobis distance. Mahalanobis depth of a point x relative to a data set X is inversely related to the Mahalanobis distance from x to the mean of X. Mahalanobis depth function can be written as

$$M_hD\ (\mathrm{x}) = [1 + (x - \bar{x})^T S^{-1} (x - \bar{x})]^{-1} \tag{1}$$

where $\bar{x}$ and S are the mean vector and dispersion matrix.

This function lacks robustness because it relies on non-robust measures like the mean and the dispersion matrix, making it inadequate for handling outliers in a data set.

Halfspace Depth (HD)

In 1975, Tukey (1975) [10] introduced the concept of location depth, also known as halfspace depth or Tukey depth, as a tool for visually describing bivariate data sets. In p dimensions, the halfspace location depth of a point θ relative to a data set $x_n = (x_1, x_2, \ldots, x_n) \in R^{p \times n}$ is denoted as $ldepth(\theta; \backslash X_n)$. It is defined as the smallest number of observation in any closed halfspace with boundary through θ. In the univariate setting p, this definition becomes

$$ldepth_1(\theta; \backslash X_n) = \min\{\#(x_i \leq \theta), \#(x_i \geq \theta)\} \tag{2}$$

In the multivariate case, the concept of the median can be generalized to the point with the highest depth, known as the Tukey median. Numerous depth functions exist, all aiming to quantify how deep or central a point x is within the data cloud. A key advantage of halfspace depth is its affine invariance. The primary reason for employing the Tukey median as a multivariate location estimator is its robustness, which can be evaluated using the breakdown value ϵ^*. Halfspace depth provides a powerful, geometrically intuitive way to measure the centrality of points in multivariate data. It is widely used in robust statistics, particularly for identifying outliers and assessing data spread. By calculating how well a point is "enclosed" by the data, it provides a robust measure of centrality, independent of the data's distribution. However, the method's computational cost can be prohibitive in high-dimensional settings without efficient algorithms.

Euclidean Depth (L_2D)

The L_2-depth was introduced by Zuo and Serfling (2000) [11]. The L_2-depth D^{L_2} measures the outlyingness of a point based on its mean distance from a chosen center of the distribution, defined as

$$D^{L_2}(z/X) = (1 + E(\|z - X\|)^{-1} \tag{3}$$

For an empirical distribution of points x_i $(i = 1,2, \dots, n)$, it is given by

$$D^{L_2}(z/X^1, \dots . X^n) = (1 + \frac{1}{n}\sum_{i=1}^{n}(\|z - X^i\|)^{-1} \tag{4}$$

The L_2-depth vanishes at infinity and reaches its maximum at the spatial median of X, minimizing $E(\|z - X\|)$.

In centrally symmetric distributions, this maximum is at the center. The L_2-depth demonstrates properties such as monotonicity with respect to the deepest point, convexity, compactness of central regions, and continuous dependence on z. It also converges in probability for uniformly integrable and weakly convergent sequences. However, the L_2-depth is not a sensible ordering of dispersion as it contradicts the dilation order, increasing with the dilation of p.

The L_2-depth is invariant against rigid Euclidean motions but not affine invariant. An affine invariant version is constructed using a positive definite matrix M and the M-norm given by

$$\|Z\|_M = \sqrt{z' M^{-1} z}, z \in R^d \tag{5}$$

Let S_X be a positive definite d×d matrix that measures the dispersion of X in an affine equivariant way, such that $S_{XA+b} = AS_XA'$ holds for any matrix A of full rank and any b. Then an affine invariant L_2-depth is given by $(1 + E(\|z - X\|S_X)^{-1}$. Besides invariance, it has the same properties as the L_2-depth. A simple choice for S_X is the covariance matrix $\sum_X$ of X.

Simplicial Depth (SD)

Liu (1990) [4] introduced the concept of simplicial depth, which measures the centrality of a point x in a p-dimensional data set, x ϵ $S_n \subset \mathbb{R}^p$. Simplicial depth is defined as the number of closed simplices containing x and having p+1 vertices in S_n. In the bivariate case, it counts the number of triangles formed by sample points in S_n contain x. Simplicial depth is robust against outliers: if a set of sample points is represented by the point of maximum depth, up to a constant fraction of the sample points can be arbitrarily corrupted without significantly altering the location of the representative point. It is also invariant under affine transformations.

However, simplicial depth lacks some desirable properties for robust measures of central tendency. For centrally symmetric distributions, there is not always a unique point of maximum depth at the center of the distribution. Additionally, the simplicial depth does not necessarily decrease monotonically from the point of maximum depth. Despite these limitations, simplicial depth remains a useful measure in robust statistics and computational geometry, particularly for its robustness to outliers and its affine invariance. Simplicial depth is a robust, non-parametric method for measuring the centrality of a point in a multivariate dataset. By focusing on simplices (the convex hulls of subsets of data points), simplicial depth provides a geometric measure of how deep or central a point is within the distribution. It is particularly useful for outlier detection and robust estimation in multivariate data, but its computational complexity can be a limitation, particularly in high-dimensional datasets.

Projection Depth (PD)

The projection depth function initiated by Liu (1992) [5]. It is based on a measure of outlyingness and the idea behind the Donoho (1982) [1]. Further, this depth function was explored by Zuo and Serfling (2000) [11].

For a univariate distribution function F of x, the outlyingness $O(F,x)$ is defined as

$$O(F,x) = \sup\{Q(u,x,F)\} \tag{6}$$

over all unit vectors u, where $Q(u,x,F) = \frac{(u^T x - \mu(F_u))}{\sigma(F_u)}$, and F_u is the distribution of $u^T x$.

The projection depth $PD(x,F)$ is then given by

$$PD(x,F) = \frac{1}{1+O(x,F)} \tag{7}$$

This approach reflects the projection pursuit methodology, involving the supremum over infinitely numerous direction vectors, making the computation of projection depth seemingly intractable. Initially, classical location and scale measures were used, but these were later replaced by robust measures like the median and Median Absolute Deviation (MAD).

III. Numerical Study

This section evaluates the effectiveness of various data depth procedures by considering both real and simulation studies. The analysis includes a comprehensive assessment of different depth procedures, such as Mahalanobis, Halfspace, L_2, Simplicial, and Projection depths. These procedures are applied to both real datasets and simulated data to provide a robust evaluation of their performance. By calculating and comparing the depth values, the study aims to determine the efficiency and reliability of each method. This comparison helps identify which depth procedures are most effective in accurately determining centrality and handling outliers.

The experimental findings from two different real datasets, available in R packages, are presented in this section. These datasets contain one or more predictors. The depth values computed using various depth functions are presented in Tables 1 and 2.

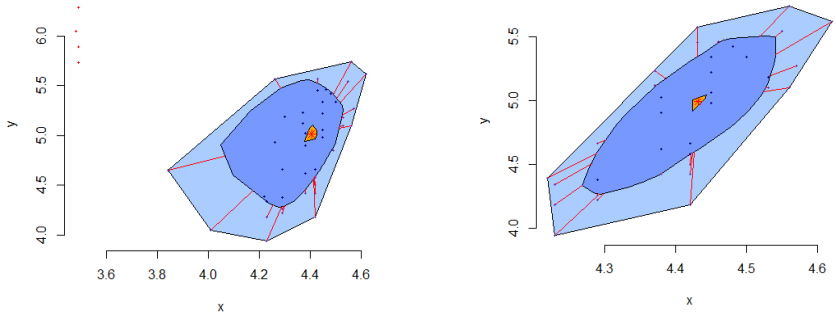
starsCYG dataset – It contains features of 47 stars in the Hertzsprung-Russell diagram of the Star Cluster CYG OB1. It includes one predictor variable, the logarithm of the star's effective surface temperature (log.Te), and one response variable, the logarithm of its light intensity (log.light). Cook's distance is used to identify the 9 outliers in the dataset.

Anscombe dataset – There are 51 observations in this dataset. The predictor variables are Income, Young, Urban and the response variable is Education. Cook's distance revealed 7 outliers in this dataset.

Table 1: *Computed depth values for starsCYG dataset*

Methods	MD	HD	L_2D	SD	PD
With Outliers	0.941	0.383	0.465	0.322	0.670
	(25)	(28)	(25)	(25)	(25)
Without	0.920	0.342	0.433	0.345	0.714
Outliers	(42)	(42)	(42)	(33)	(42)

(.) Observation number



(a). With Outliers (b). Without Outliers

Figure 1: *Bagplot for starsCYG dataset (with and without outliers)*

Table 2: *Computed depth values Anscombe dataset*

Methods	MD	HD	L2D	SD	PD
With Outliers	0.869 (14)	0.333 (25)	0.352 (25)	0.145 (25)	0.565 (25)
Without Outliers	0.941 (42)	0.341 (25)	0.346 (25)	0.169 (25)	0.583 (25)

Tables 1 and 2 reveal that, in both the presence and absence of outliers, the Mahalanobis depth consistently exhibits the highest depth values. This indicates that the Mahalanobis depth method is particularly effective at measuring centrality, regardless of whether outliers are present in the dataset. The real study compares the performance of five data depth methods MD, HD, L2D, SD, and PD with and without outliers. When outliers are present, MD is the most efficient method with a score of 0.869, showing its robustness in handling contamination. In contrast, SD and PD exhibit a significant drop in efficiency, with scores of 0.145 and 0.565, respectively, indicating their susceptibility to outliers. When the outliers are removed, the efficiency of all methods increases, with MD still leading at 0.941. HD and L2D show similar performance, with scores of 0.341 and 0.346, respectively. The efficiency of SD and PD improves somewhat after removing outliers, but they still remain the least efficient with scores of 0.169 and 0.583, respectively. The results highlight that MD is the most robust and efficient method, particularly when outliers are present, while HD and L2D offer a balanced performance.

The simulation study aims to assess and compare the efficiency of different data depth procedures in handling multivariate data. It investigates how each method performs under various contamination scenarios, such as location and scale contamination, which simulate real-world deviations from ideal data. The goal is to identify the most effective and reliable depth procedures, particularly in the presence of data contamination, which is common in practical applications. The data is simulated from normal distribution of sample size n=1000 with mean vector μ (0, 0) and unit covariance matrix $\sum = I_2$, and the simulated data is then contaminated in three different scenarios such as location contamination, scale contamination, and a combination of location and scale contamination. These contaminations are introduced at varying levels of 0%, 5%, 10%, 15%, 20%, and 25%.

For location contamination, the simulated data is contaminated by the mean vector μ (5, 5). In the case of scale contamination, the data is contaminated by altering the covariance matrix to $\sum = 2I_2$. In location and scale contamination, the simulated data is contaminated by the mean vector μ (3, 3) and $\sum = 1.5I_2$. These varying levels of contamination allows to evaluate the robustness and performance of different data depth procedures under different types and degrees of data contamination and are presented in Table 3.

Table 3: *Computed depth values for Simulation study*

Levels	MD	HD	L_2D	SD	PD
0%	0.869 (14)	0.333 (25)	0.352 (25)	0.145 (25)	0.565 (25)
Location Contamination					
5%	0.960 (45)	0.377 (45)	0.397 (45)	0.139 (45)	0.664 (45)
10%	0.958 (54)	0.388 (45)	0.399 (45)	0.155 (45)	0.706 (45)
15%	0.929 (54)	0.388 (45)	0.389 (45)	0.159 (43)	0.662 (45)
20%	0.936 (54)	0.377 (45)	0.390 (54)	0.157 (45)	0.667 (45)
25%	0.935 (54)	0.311 (43)	0.388 (54)	0.145 (43)	0.598 (43)
Scale Contamination					
5%	0.992 (52)	0.432 (52)	0.385 (52)	0.161 (52)	0.838 (52)
10%	0.890 (59)	0.344 (62)	0.386 (59)	0.136 (62)	0.609 (62)
15%	0.987 (52)	0.433 (52)	0.392 (52)	0.166 (52)	0.788 (52)
20%	0.888 (47)	0.352 (45)	0.385 (45)	0.152 (45)	0.673 (45)
25%	0.972 (47)	0.432 (47)	0.391 (47)	0.168 (47)	0.745 (47)
Location and Scale Contamination					
5%	0.916 (45)	0.344 (45)	0.381 (45)	0.145 (45)	0.583 (45)
10%	0.951 (45)	0.412 (45)	0.386 (45)	0.158 (45)	0.734 (45)
15%	0.950 (45)	0.382 (45)	0.387 (45)	0.156 (45)	0.676 (45)
20%	0.924 (45)	0.382 (45)	0.383 (45)	0.152 (45)	0.710 (45)
25%	0.922 (45)	0.381 (45)	0.382 (45)	0.158 (45)	0.707 (45)

Based on the results presented in Table 3, it can be concluded that the Mahalanobis depth consistently identifies the deepest location point among the different data depth procedures evaluated. This indicates that the Mahalanobis depth is particularly effective at determining the central point of the dataset, demonstrating its robustness and reliability in comparison to other depth measures. Even in the presence of outliers, Mahalanobis depth shows the smallest decrease in efficiency, highlighting its ability to maintain accuracy when the data is contaminated. The method's robustness is further emphasized by its superior performance under both location and scale contamination scenarios.

IV. Discussion

The study concludes that among the various data depth measures tested, Mahalanobis Depth consistently identifies the deepest points across different scenarios, both with and without outliers. This suggests that Mahalanobis Depth provides a stable measure of centrality, even when the data contains extreme values or deviates from standard assumptions. In contrast, other depth measures like Halfspace Depth and Projection Depth demonstrate sensitivity to outliers and complex distributions, sometimes shifting central points. L_2 Depth and Simplicial Depth also showed varied performance, especially in non-elliptical data structures. Overall, Mahalanobis Depth's consistent centrality assessment highlights its utility in robust statistical applications where a reliable measure of depth is crucial.

References

[1] Donoho, D. L. Breakdown Properties of Multivariate Location Estimators, Technical Report, Harvard University, Boston, 1982.
[2] Donoho, D. L. and Gasko, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics*, 1803-1827.
[3] Koshevoy, G and Mosler, K. (1997). Zonoid trimming for multivariate distributions. *The Annals of Statistics*, 25:1998–2017.
[4] Liu, R.Y. (1990). On a notion of data depth based on random simplicies. *The Annals of Statistics*, 18: 405–414.
[5] Liu, R.Y. (1992). Data depth and multivariate rank tests. In: Dodge, Y. (ed.), L1-Statistics and Related Methods. *North-Holland (Amsterdam)*, 279–294.
[6] Liu, R. Y., Parelius and Singh, K. (1999). Multivariate analysis by data depth: Descriptive Statistics, Graphics and Inference, *The Annals of Statistics*, 27:783-858.
[7] Mahalanobis J. (1936). On the generalized distance in statistics, *Proceedings of the National Academy, India*, 12:49–55.
[8] Muthukrishnan, R and Poonkuzhali, G. (2018). Robust Depth based weighted Estimator with Application in Discriminant Analysis, *International Journal of Scientific Research in Mathematical and Statistical Sciences*, 5:96-101.
[9] Muthukrishnan, R., Gowri, D and Ramkumar N. (2018). Measure of Location using Data Depth Procedures, *International Journal of Scientific Research in Mathematical and Statistical Sciences*, 5: 273–277.
[10] Tukey, J. W. (1975). Mathematics and the picturing of data. In: *Proceeding of the International Congress of Mathematicians*, Vancouver, 523–531.
[11] Zuo, Y. J. and Serfling R. (2000). General notions of statistical depth function, *The Annals of Statistics*, 28:461–482.