International Journal of Scientific Research in Mathematical and Statistical Sciences
Volume-5, Issue-6, pp.273-277, December (2018)
Research Paper
E-ISSN: 2348-4519

Measure of Location using Data Depth Procedures

R.Muthukrishnan[1], D.Gowri[2], N.Ramkumar[3]

[1, 2, 3] Department of Statistics, Bharathiar University, Coimbatore, Tamil Nadu

Available online at: www.isroset.org

Received: 02/Nov/2018, Accepted: 09/Dec/2018, Online: 31/Dec/2018

***Abstract*-** Data depth is used to measure the depth or outlyingness of a given multivariate sample with respect to its underlying distribution. It can be lead to a natural center-outward ordering of sample points. The essence of depth function in multivariate analyses is to measure the degree of centrality of point relative to a data set or probability distribution. This work explores data depth procedures in order to find the measure of location, namely deepest or center point. Further, the various depth procedures are examined under real and simulation environment with the help of R software. The efficiency of various data depth procedures have been studied by computing average misclassification error in the context of discriminant analysis with numerical illustration.

***Keywords*:** Data Depth, Location and Linear discriminant analysis.

I. INTRODUCTION

Depth is an integer assigned to be a candidate fit relative to a data set. This leads to outside-inward/ centre-outward ordering of the sample points. The usual order statistics is different from the depth order statistics. In usual order statistics the data are ordered from the smallest sample point to the largest, while the depth order statistics start from the middle sample point and move outwards in all directions.

Data depth is a function of measuring degree of centrality of a point relative to a probability distribution. Based on the depth functions, a lot of methods of signs and ranks, order statistics, quantiles, and outlyingness measures could be extended conveniently from their univariate counterparts in a unified way [14]. Data depth is a concept which plays an important role in many notable fields of statistics, namely; data exploration, ordering, asymptotic distributions and robust estimation [10].

The rest of the paper is organized as follows. Section 2 briefly summarizes the various data depth procedures. Section 3 presents the results obtained from the study based on real and simulation environment along with application in discriminant analysis. The paper ends with conclusion in the last section.

II. DATA DEPTH PROCEDURES

Many graphical and quantitative methods are fixed for analysing the measures such as location, scale and shape, as well as comparing inference methods based on data depth. Numerous depth notions have been proposed during the last few decades. The celebrated depth procedures, namely Mahalanobis Depth [1], Half Space Depth [2], Simplicial Depth [4], Simplicial Volume Depth [3], Spatial Depth [11], Zonoid Depth [8] and Projection Depth [12], [13] which are briefly summarized in this section

.

2.1 Half Space Depth

Tukey (1975) introduced the concept of half space depth (HSD). Half space Depth of a point $x=(x_1,....,x_p)\in S_n=\{x_i=(x_{i1},.....,x_{ip});i=1,...,n\}\subset R^p$ relative to a p - dimensional data set S_n is defined as the minimum number of data points in a closed half space with boundary through x. In univariate case, it is easy to see that the depth of a point is given by, $\min\{\#\{x_i\le x\},\#\{x_i\ge x\}\}$ the median is the point (or points) with maximal depth. In the multivariate case, the notation of median can be comprehensive, being the point with maximal depth. This multivariate median is called Tukey median. The Half Space depth is also called a location depth and Tukey depth.

2.2 Mahalanobis Depth

Mahalanobis (1936) introduced the concept of generalized distance in statistics. In 1975 Mahalanobis distance can be used as a measure to calculate the depth of a point. Mahalanobis depth (MD) of a point $x\in S_n\subset\mathbb{R}^p$ relative to a p -dimensional data set defined as:

$$MD(x;S_n)=\left[1+(x-\bar{x})^T S^{-1}(x-\bar{x})\right]^{-1} \quad (1)$$

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

This function fails by the side of being robust, since it is based on non robust measures such as the mean and the dispersion matrix. Another disadvantage of this function is that it depends on the continuation of second moments.

2.3 Projection Depth

Let μ(.) and σ(.) be univariate location and scale events, respectively. Then the outlyingness of a point with deference to the distribution function *F* of *x* defined by (Liu 1992)

$$O(x,F) = \sup_{\|u\|=1} \left| Q(u,x,F) \right| \qquad (2)$$

where, $Q(u,x,F) = \left(u^T x - \mu(F_u)\right) / \sigma(F_u)$ and FF_{uu} is the distribution of $u^T x$. Let, μ (.) and σ (.) be multivariate case used a point of a p-dimensional data set. The projection depth (PD) is defined by

$$PF_u D(x,F) = \frac{1}{1+O(x,F)} \qquad (3)$$

2.4 Simplicial Depth

Liu (1990) introduced the concept of Simplicial depth (SD). Simplicial depth of a point $x \in S_n \subset \mathbb{R}^p$ relative to a p-dimensional data set S_n defined as the number of closed simplex containing x and having $p+1$ vertices in S_n. In bivariate case, the simplicial depth of a point x is the number of triangles through vertices in S_n and containing x. Simplicial depth is counted as a probability that a point lies in a simplex, built on d + 1 data points.

$$D_S(x,F) = P_F\left(x \in S[X_1,...,X_{d+1}]\right), x \in R^d \quad (4)$$

Simplicial depth is robust against outliers. Since, if a set of sample points is represented by the point of maximum depth, then up to a constant fraction of the sample points can be arbitrarily corrupted without significantly changing the location of the representative point. It is also invariant under affine transformations of the plane. However, simplicial depth fails to have some other desirable properties for robust measures of central tendency. When applied to centrally symmetric distributions, it is not necessarily the case that there is a unique point of maximum depth in the center of the distribution. Also, from the point of maximum depth, it is not necessarily the case that the simplicial depth decreases monotonically.

2.5 Simplicial Volume Depth

Oja (1983) established a depth procedure using the concept of simplicial volume (SVD). Simplicial volume is a homotopy invariant of oriented closed associated manifolds that was introduced by Gromov (Gromov 1983). Intuitively, simplicial volume events are difficult to describe the manifold in question in terms of simplices (with real coefficients).

Let M be an oriented closed associated manifold of dimension n. Then the simplicial volume of M (also called the Gromov norm of M) is defined as,

$\|M\| := \|[M]\|_1 = \inf\{|c|_1 | c \in C_n(M;\Re)$ is a fundamental cycle of $M\} \in \Re_{\geq 0}$,

where, $[M] \in H_n(M;\mathbb{R})$ is the fundamental class of M with real coefficients.

Oja depth of a point $x \in S_n \subset \mathbb{R}^p$ relative to a p-dimensional data set S_n is defined as the sum of the volume of every closed simplex having a vertex at x and the others in any p points of the S_n data set. In the bivariate case, the Oja depth of a point x relative to a bivariate data set S_n is the sum of the areas of all triangles whose vertices are x, x_i, x_k with x_i and x_k belonging to S_n.

2.6 Zonoid Depth

Koshevoy and Mosler (1996) introduced a notion of data depth, called zonoid data depth (ZD). The zonoid data depth, $\text{depth}_\mu(x)$, of a point $X \in \mathbb{R}^d$ is defined by,

$$\text{depth}_\mu(x) = \begin{cases} \sup\{\alpha : x \in D_\alpha(\mu)\}, & \text{if } x \in D_\alpha(\mu) \text{ for some } \alpha, \\ 0, & \text{otherwise.} \end{cases} \qquad (5)$$

The data depth of a point x is the maximal height α at which $\alpha x \in proj_\alpha \hat{Z}(\mu)$. Here,

$$D_\alpha(\mu) = \frac{1}{\alpha} \text{proj}_\alpha\left(\hat{Z}(\mu)\right) \qquad (6)$$

where $0 < \alpha \leq 1$. Further, the depth of x equals zero if x lies outside $D_\alpha(\mu)$ for all α; it equals one if x is the expectation. If α > 0, $D_\alpha(\mu)$ is the set of all points that include data depth greater than or equal to α.

2.7 Spatial Depth

An implementation of the idea of spatial depth (SPD), established by Serfling (2002), which is defined as follows: Lt Y be d-dimensional random vectors have cumulative distribution function F. Then, the multivariate spatial depth of $x \in \mathbb{R}^d$ qualified F is defined as,

$$SD(x,F) = 1 - \left\| \int S(x-y) dF(y) \right\|_E = 1 - \|E[(x-y)]\|_E \qquad (7)$$

where $\|.\|_E$ is the Euclidean norm in $\mathbb{R}^d$. The spatial depth is a depth function that builds ahead the notion of spatial (also called geometric) quantiles for multivariate data, considered by Chaudhuri (1996) and Koltchinskii (1997), formulated by Vardi and Zhang (2000) and Serfling (2002). This Spatial depth also called L1-depth.

3. EXPERIMENTAL STUDY

This section presents the performance of various data depth procedures by providing the results of numerical illustration which was carried out under real/simulation data and by considering with/without outliers.

Further, various notions of data depth procedures have been studied by performing discriminant analysis in a real data set. The efficiency of these procedures is compared by computing misclassification probabilities.

3.1 Real data

For this study, a real data set was considered, namely NYC data. This data set contains two variables, with 23 observations. The variables are Manpower in percent, and percent change in weekly auto thefts. For the given data set, the 14th observation is identified as outliers. The deepest point is located by using various notions of depth procedures with and without outliers and is summarized in the form of table 1 and is given in appendix.

From the table 1, it is noticed that Simplicial (SD) and Spatial Depth (SPD) provides the same deepest point by considering the maximum depth value and also with and without outliers. These two methods equally perform well and better than the other methods. If the data cleaned (after removing outliers) almost all the methods represent the same data point as the deepest point (excludes simplicial volume depth).

3.2 Simulation data

A simulation study is performed to compare the efficiency of the various notions of data depth procedures. The data (n=100) are generated normal distribution, mean vector, μ= (0, 0) and unit covariance matrix, $\Sigma = I_2$. The various level of contaminations (mean vector, μ= (4, 4) and unit covariance matrix, $\sum = 1.5 I_2$) such as 0%, 1%, 2%, 5%, 10% and 20% are considered and the obtained results are summarized in the form of table 2 and is given in appendix

It is observed that, from the table 2, simplicial and spatial depths tolerates certain amount of contaminations and gives the same deepest point (measure of location). The other depth procedures fail to tolerate, even if the data contamination is very low (1%), does not provides the same deepest point. It is concluded that the simplicial and spatial depth is superior to other depth procedures.

3.3 Application (Discriminant Analysis)

This section demonstrates the efficiency of various notions of data depth procedures by applying in the multivariate technique, Discriminant analysis. For this, a real data set was considered, namely, anorexia data set. The data set contains 3 groups, each group two variables with a frame of 72 observations. The weight change data for young female anorexia patients. There are two variables, one is, prewt (weight of patients before study periods, in lbs) and second one is, postwt (weight of patients after study periods, in lbs), classified the three groups into the Cont (Control), CBT (Cognitive behavioral treatment), FT (Family treatment).

On comparing the average probability of misclassification values in the table 3 results given in appendix, simplicial and spatial Depth performs better than the other methods. Since these two procedures gives low misclassification probabilities when compared with other data depth procedures.

4. CONCLUSION

Measures of location play a vital role in almost all statistical data analysis. In this era of big data, it is to be estimated a good measure of location to perform any data analysis techniques, to understand the data. Many procedures are established to estimate the measure of location for the past two centuries. Data depth procedures are recent advances in statistics to locate a reliable location by considering deepest point in a data cloud. In this context, this paper demonstrates the various notions of data depth procedures which were established recent past. The efficiency of these procedures has been studied with application in the context of Discriminant analysis along with numerical study. From the study, it is suggested that simplicial and spatial depth performs equally good when compared with other depth procedures. The research communities can get more accuracy while using these procedures in order to find the good location by identifying the deepest point in a data cloud, instead of using conventional measure of location.

REFERENCES

[1] P. Mahalanobis, "On the generalized distance in statistics", Proceedings of the National Academy India Vol.12, pp.49–55, 1936.

[2] J.W. Tukey, "Mathematics and the picturing of data". In: Proceeding of the International Congress of Mathematicians, Vancouver, pp.523–531, 1975.

[3] H. Oja, "Descriptive statistics for multivariate distributions", Statistics & Probability Letters, Vol.1, pp.327–332, 1983.

[4] R. Y. Liu, "On a notion of data depth based on random simplicies", The Annals of Statistics, Vol.18, pp.405–414, 1990.

[5] R.Y. Liu, "Data depth and multivariate rank tests". In: Dodge, Y. (ed.), L1-Statistics and Related Methods, North-Holland (Amsterdam), pp.279–294, 1992.

[6] P. Chaudhuri, "On a geometric notion of quantiles for multivariate data", Journal of the Americal Statistical Association, Vol. 91, pp. 862–872, 1996.

[7] R. Dyckerhoff, G. Koshevoy, and K. Mosler, "Zonoid data depth: theory and computation". In: Prat A. (ed), COMPSTAT 1996. Proceedings in computational statistics, Physica-Verlag (Heidelberg), pp.235–240, 1996.

[8] G. Koshevoy, and K. Mosler, "Zonoid trimming for multivariate distributions", The Annals of Statistics, Vol.25, pp.1998–2017, 1997.

[9] Johnson and D.W. Wichern, "Applied Multivariate Statistical Analysis", 4th Edition. Prentice hall, Upper Saddle River, 1998.

[10] R.Y. Liu, J.M. Parelius and K. Singh, "Multivariate analysis by data depth: Descriptive Statistics, Graphics and Inference", The Annals of Statistics, Vol.27, pp.783-858, 1999.

[11] Y. Vardi, and C. Zhang, "The Multivariate L_1 Median and Associated Data Depth", Proceedings of the National Academy of Science USA, Vol.97, pp.1423-1426, 2000.
[12] Y.J. Zuo, and R. Serfling, "General notions of statistical depth function", The Annals of Statistics Vol.28, pp.461–482, 2000.
[13] Y. Zuo, "Projection-based depth functions and associated medians", The Annals of statistics, Vol.31, pp.1460-1490, 2003.
[14] R. Serfling, "Depth functions in nonparametric multivariate inference". In: Liu, R., Serfling, R., Souvaine, D. (eds.), Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications, American Mathematical Society, pp.1–16, 2006.
[15] X. Liu, and Y. Zuo, "Computing projection depth and its associated estimators", Statistics and Computing Vol.24, pp.51–63, 2014.
[16] R. Muthukrishnan, and G. Poonkuzhali, "Computing Median with Data Depth in Multivariate Data", Journal of Modern Sciences, Vol.7, Issue.2, pp.11-19, 2015.
[17] R. Muthukrishnan, M. Vadivel, and N. Ramkumar, "Projection based Data Depth Procedure with application in Discriminant Analysis". International Journal of Research in Advent Technology, Vol.6, Issue.5, pp.824-832, 2018.
[18] R. Muthukrishnan, and G. Poonkuzhali, "Robust Depth based weighted Estimator with Application in Discriminant Analysis", International Journal of Scientific Research in Mathematical and Statistical Sciences, Vol.5, Issue.3, pp.96-101, 2018.
[19] R Core Team, R: "A language and environment for statistical computing", Vienna, Austria, 2018.

AUTHORS PROFILE

Dr.R.Muthukrishnan was born in Tirunelveli. He holds M.Sc. (1993) and Ph.D. (2000) in Statistics from Manonmaniam Sundaranar University in 2000. He is currently working as Associate Professor in the Department of Statistics, Bharathiar University, Coimbatore, Tamil Nadu. He is a life member of various academic bodies such as ISPS, IBS, IISA and member in ISI and ISA. His main research work focuses on Robust Statistical Inference and Multivariate Analysis. He has more than 15 years of teaching and research experiences. He has published 45 research papers in reputed national / international journals and guided 21 research scholars for their research programmes.

D. Gowri graduated Master of degree in Statistics from Bharathiar University. Now she is doing Master of Philosophy at Bharathiar University. Her research interests are Robust Statistical Inference, Multivariate Analysis.

Appendix

Table 1: Measure of location and the associated depth value under various data depth procedures

Methods	MD	HSD	SD	SVD	SPD	ZD	PD
NYC data (With Outlier)	4 (-5.37, -1.01) **0.926396**	4 (-5.37, -1.01) **0.347826**	9 (-8.81, -0.76) **0.352908**	7 (-7.8, -0.4) **0.854811**	9 (-8.81, -0.76) **0.818959**	4 (-5.37, -1.01) **0.835277**	4 (-5.37, -1.01) **0.567126**
NYC data (Without Outlier)	9 (-8.81, -0.76) **0.900567**	9 (-8.81, -0.76) **0.363636**	9 (-8.81, -0.76) **0.361688**	5 (-10.23, -0.76) **0.844151**	9 (-8.81, -0.76) **0.834897**	9 (-8.81, -0.76) **0.815559**	9 (-8.81, -0.76) **0.617489**

. – Observation number**; (.)** – Location**; Bold** – Depth value

Table 2: Measure of location and the associated depth value under various data depth procedures

Error	MD	HSD	SD	SVD	SPD	ZD	PD
0%	57 (0.025383, 0.027475) **0.994756**	57 (0.025383, 0.027475) **0.4**	39 (0.143771, -0.11775) **0.273649**	91 (-0.07042, -0.43088) **0.687842**	39 (0.143771, -0.11775) **0.915621**	57 (0.025383, 0.027475) **0.956976**	39 (0.143771, -0.11775) **0.755757**
1%	48 (0.596259, 0.119718) **0.999663**	68 (0.689373, -0.95584) **0.41**	39 0.143771, -0.11775) **0.274286**	80 (-0.01253, -0.37485) **0.76151**	39 (0.143771, -0.11775) **0.920654**	48 (0.596259, 0.119718) **0.992166**	68 (0.689373, -0.95584) **0.760977**

2%	57 (0.025383, 0.027475) **0.995619**	57 (0.025383, 0.027475) **0.42**	39 (0.143771, -0.11775) **0.275603**	28 (-0.05488, 0.250141) **0.720274**	39 (0.143771, -0.11775) **0.932666**	57 (0.025383, 0.027475) **0.980722**	39 (0.143771, -0.11775) **0.775323**
5%	35 (0.248413, 0.065288) **0.988765**	39 (0.143771, -0.11775) **0.44**	39 (0.143771, -0.11775) **0.276982**	36 (0.019156, 0.257338) **0.739384**	39 (0.143771, -0.11775) **0.946419**	35 (0.248413, 0.065288) **0.950009**	39 (0.143771, -0.11775) **0.832403**
10%	83 (0.779584, 0.713241) **0.974027**	36 (0.248413, 0.065288) **0.42**	36 (0.248413, 0.065288) **0.276005**	60 (0.494796, 0.138053) **0.753897**	36 (0.248413, 0.065288) **0.947431**	36 (0.248413, 0.065288) **0.905542**	36 (0.248413, 0.065288) **0.812343**
20%	18 (0.726751, 1.151912) **0.972816**	96 (-0.01675, 0.161789) **0.41**	36 (0.019156, 0.257338) **0.271967**	35 (0.248413, 0.065288) **0.707513**	36 (0.019156, 0.257338) **0.894025**	83 (0.779584, 0.713241) **0.900445**	36 (0.019156, 0.257338) **0.774119**

. – Observation number**; (.)** – Location**; Bold** – Depth value

Table 3 Computed misclassification probabilities under various data depth procedures

Methods	**MD**	**HSD**	**SD**	**SVD**	**SPD**	**ZD**	**PD**
With outlier	0.4930	0.4930	**0.4507**	0.5352	**0.4507**	0.5070	0.5352
Without outlier	0.4853	0.4627	**0.4328**	0.4930	**0.4328**	0.4853	0.4507

