International Journal of Scientific Research in Mathematical and Statistical Sciences **Research Paper**
Volume-5, Issue-3, pp.96-101, June (2018) **E-ISSN:** 2348-4519

Robust Depth based weighted Estimator with Application in Discriminant Analysis

R. Muthukrishnan[1], G. Poonkuzhali[2*]

[1] Department of Statistics, Bharathiar University, Coimbatore 641 046, Tamil Nadu, India
[2] Department of Statistics, Bharathiar University, Coimbatore 641 046, Tamil Nadu, India

[*]*Corresponding Author: poostat18@gmail.com, Tel.: +91-9923368866*

Available online at: www.isroset.org

Received: 10/Mar/2018, Revised: 22/Mar/2018, Accepted: 13/Apr/2018, Online: 30/Jun2018

***Abstract*—** Data depth concept used to measure the deepness of a given point in the entire multivariate data cloud. It leads to center-outward ordering of sample points used rather than usual smallest to largest rank. The ordering starts from middle and moves in all directions. Multivariate location and scatter can be computed by using the depth value of each data point. Various depth procedures have been established by many authors. In this paper, a new depth procedure is proposed, namely Modified Mahalanobis Depth (MMD), which calculates depth based on robust distance with Minimum Covariance Determinant (MCD) approach and a weight function is established to determine the location and scale. The superiority of the proposed depth based procedure over existing depth procedures has been studied in simulated environment using R software with respect to application in discriminant analysis. The proposed depth procedure performs well when compared with the existing procedures even with higher contamination levels and larger sample sizes.

Keywords—*Data Depth, Mahalanobis Distance, MCD, Robust Distance, Weight Function, Discriminat Analysis, R Software*

I. INTRODUCTION

Location and scatter plays a very important part in multivariate statistical methods. Data depth is one of the main emerging concepts to determine such measures. Data depth measures deepness of a given point in the whole data cloud. This concept is essential since it leads to a center-outward ordering rather than usual smallest to largest rank. It means the ordering starts from the center and moves in all directions.

The depth value for every data point can be calculated by using various established depth procedures. The data point which has the highest depth value is considered as the deepest point and the lowest depth value as outlier. The data point which has the maximum depth value, which approaches to 1, is considered as the best location.

Various depth procedures have been developed in the literature [1]-[16]. Comprehensive surveys on data depth are described in [17]-[19].

In this paper, a new depth procedure and a new weight function associated with the study are proposed to estimate location and dispersion. The new procedure uses robust distance with the Minimum Covariance Determinant (MCD) approach instead of Mahalanobis distance.

In Section II, various existing depth procedures are defined. The proposed depth procedure, (MMD) is explained in section III. The new weight function to estimate location and scale is described in section IV. In section V, the performance of the proposed depth procedure is compared with existing procedures in a simulating environment with application in discriminant analysis in the form of apparent error rate (AER). The conclusion is presented in last section**.**

II. MATERIALS AND METHODS

The existing depth procedures, based on distances, projection pursuit, halfspaces, weighted mean and neighbourhoods are summarized in this section.

A. Mahalanobis Depth

Mahalanobis depth (MD) can be calculated from Mahalanobis distance [20]. Let $X = X_1, X_2, ..., X_p$ be a p dimensional multivariate data set X and let

$\bar{X} = \left[\bar{x}_1\ \bar{x}_2\ \ldots \bar{x}_p\right]$ be a sample mean vector and the covariance matrix

$$S = \begin{bmatrix} s_{11}\ s_{12}\ldots s_{1p} \\ s_{21}\ s_{22}\ldots s_{2p} \\ \vdots \\ s_{p1}\ s_{p2}\ldots s_{pp} \end{bmatrix}, \quad (1)$$

Then the Mahalanobis distance of each point can be estimated

$$d(\mathrm{x};\bar{\mathrm{x}},\mathrm{S}) = \left(\mathrm{x} - \bar{X}\right)' S^{-1}\left(\mathrm{x} - \bar{X}\right), \quad (2)$$

and from this distance, (MD) can be computed by

$$MD(\mathrm{x};\bar{\mathrm{x}},\mathrm{S}) = \left[1 + d(\mathrm{x};\bar{\mathrm{x}},\mathrm{S})\right]^{-1}, \quad (3)$$

The deepest point in the entire data has the largest depth value.

B. Halfspace Depth

The idea of halfspace depth was originally developed by [1]. For some number x, consider partition into two components: all points equal to or less than x are considered as a closed halfspace and all values less than x as open halfspace. Similarly, all values equal to or greater than x are considered as closed halfspace and all values greater than x as open halfspace. In general, for two dimensions, for any line, the values above this line shape the closed halfspace and the values below the line are considered as open halfspaces and for p=3, for any plane form two closed halfspaces and similarly two open halfspaces.

Let H be any closed halfspace have the value x and let $p(\mathrm{H})$ be the probability associated with H. Then Tukey's Halfspace depth (HD) is

$$HD = \inf_{\mathrm{H}}\left[p(\mathrm{H}) : \mathrm{H}\ \textit{is a closed halfspace having}\ x\right] \quad (4)$$

C. Zonoid Depth

Zonoid depth (ZD) was defined by [8] by using the Zonoid trimming concept introduced by Koshevoy and Mosler [5]. Multivariate trimming concept was used in this depth procedure. That is, the multivariate trimmed regions are centered about the mean instead of the median.

For probability distribution μ and $\alpha \in (0,1]$, ZD of a point x that belongs to multivariate data cloud is defined as

$$ZD = \begin{cases} Sup\{\alpha : \mathrm{x} \in \mathrm{D}_\alpha(\mu)\}, if\ \ \mathrm{x} \in \mathrm{D}_\alpha(\mu)\ for\ some\ \alpha, \\ 0 \qquad\qquad\qquad\quad , otherwise \end{cases}, \quad (5)$$

If x lies outside $D_\alpha(\mu)$ for all α, then the depth of x equals to zero; if depths of x equal to 1 then x is the expectation.

D. Spatial Depth

The idea of spatial quantiles was introduced by [21] and Spatial depth (SD) was formulated by [7] and extended by [22]. A generalization of L_1 norm in univariate case form spatial quantiles. Spatial depth is also known as L_1-depth.

For a distribution function F, spatial quantile function Q_F and the interpretation of $\left\|Q_F^{-1}(\mathrm{x})\right\|$ as a measure of outlyingness, SD is given by

$$SP = 1 - \left\|Q_F^{-1}(\mathrm{x})\right\|, \quad (6)$$

The point corresponding to maximal depth is considered as the spatial median and the point that has the lowest depth value is considered as the outlier.

E. Projection Depth

Projection depth (PD) has been proposed by [6]. Both Projection depth and halfspace depth are closely related, which reflects the projection pursuit methodology. This procedure involves supremum over infinitely numerous direction vectors hence the computation of PD appears intractable.

Let x be any point in the data cloud, u be any p dimensional vector having unit norm. M denotes the median of data cloud X, MAD represents the Median Absolute Deviation. Then PD is defined as

$$PD(\mathrm{x}, X) = \left[1 + \sup_{\|u\|=1} \frac{\left|u'\mathrm{x} - M(u'\mathrm{X})\right|}{MAD(u'\mathrm{X})}\right]^{-1}, \quad (7)$$

F. Local Depth

The concept of local depth (LD) was proposed by[12]. This procedure is a local extension of depth. The construction of this depth is obtained by conditioning the distribution to appropriate neighbourhoods. For defining a neighbourhood

of a point, depth of a point is calculated in the idea of symmetrisation of a distribution.

Let x be any point in the entire data cloud X, and instead of a distribution P^X a distribution $P_x = 1/2\,P^X + 1/2\,P^{2x-X}$ is used.

For any $\beta \in [0,1]$, the minimum depth region bigger or equal to β,

$$R^{\beta}(F) \underset{\alpha \in A(\beta)}{\cap} D_{\alpha}(F), \tag{8}$$

where $A(\beta) = \{\alpha \geq o : P(D_{\alpha}(F)) \geq \beta\}$ then for a locality parameter β, a neighbourhood of a point x as s $R_x^{\beta}(P)$.

Formally, let $D(\bullet, P)$ be a depth function, then LD with respect to a point x is defined as

$$LD : z \to D(z, P_x^{\beta}), \tag{9}$$

where $P_x^{\beta}(\bullet) = P(\bullet / R_x^{\beta}(P))$ is conditional distribution of P conditioned on $R_x^{\beta}(P)$.

III. Proposed Depth procedure

The proposed procedure is formulated by using the concept of MCD estimator instead of the conventional estimator of location and scale in the Mahalanobis distance, namely the MMD procedure. The robust MCD estimator was proposed by [23] to locate the robust measure of location and scatter. The computational depth procedure for MMD is as follows:

Let $X_1, X_2, ..., X_p$ be a p dimensional multivariate data set X and x be a numerical vector whose depth is to be calculated.

1) Find center (M_X) and the covariance matrix (Cov_X) using robust MCD estimator
2) Compute the distance
$$d_x = (x - M_X)' \operatorname{cov}_X^{-1} (x - M_X), \tag{10}$$
3) Sort the distance given in step 2 and denote it by S_{dx}
4) Find the median from the distance from step 3
$$MS_{dx} = Median(S_{dx}), \tag{11}$$
5) Find the difference between distance value and median from step 2 and step 4.
$$D_x = d_x - MS_{dx}, \tag{12}$$
6) Find the absolute value of difference given in step 5
$$abs_x = abs(D_x), \tag{13}$$
7) Finally the MMD depth can be computed by
$$MMD_x = \frac{1}{1 + abs(D_x)}, \tag{14}$$

Mahalanobis depth provides reliable results when the data is normal. But it gives unreliable results, when the data contain outliers, since this depth procedure uses traditional mean vector and covariance matrix which are very sensitive to outliers.

The proposed MMD procedure is used to compute depth by employing robust estimator and is also it is an advancement of MD, since it further calculates the absolute differences about median of distances. It gives the highest depth value to the best points and establishes reliable results.

IV. Proposed Weight Function

The weight function has been proposed to estimate location and scale after computing depth value of each point in a data cloud. Given a notion of depth described in section 2 and 3, the depth weight function is computed and is given below.

Consider the depth value for each data point denoted by $De_{(x)}$. Sort the depth values and find the median denoted by $md_{(x)}$. The depth weight function is given by

$$w_x = \begin{cases} 1 + \left[De_{(x)}^{\,2} / md_{(x)}\right]^2, & if\ De_{(x)} \geq md_{(x)} \\ \left[De_{(x)}^{\,2} / md_{(x)}\right]^2, & if\ De_{(x)} < md_{(x)} \end{cases} \tag{15}$$

,

From the above weight function, assign weights (w_x) fo each data point x. Then the location and scale can be computed by,

$$D_L(x) = \frac{\sum_x x w_x}{\sum_x w_x}, \tag{16}$$

and

$$D_S(x) = \frac{\sum_x w_x}{\left(\sum_x w_x\right)^2 - \sum_x w_x^2} \sum_x w_x \left(x - D_L(x)\right), \quad (17)$$

V. EXPERIMENTAL RESULTS

To study the performance of the proposed depth procedure (MMD), it was compared with the existing depth procedures which were given in the section 2. The experiments were carried out under simulation environment in the context of discriminant analysis by computing AER with the help of packages in R software and are summarised in this section.

The data were simulated with different dimensions p=2 and 5, the number of groups, g=2 and 3, the size of training sample, n=100,1000 and 5000, and various levels of contamination (0%,10%, 20%, 30% and 40%). In all the cases, the class distributions are normal, but the generated data sets have different mean vector and the same covariance matrix for 2 dimension,

$$\mu_1 = (0,0);\ \mu_2 = (2,2);\ \mu_3 = (4,4);\ \Sigma_1 = \Sigma_2 = \Sigma_3 = I_2.$$

Generated data were contaminated with different mean vectors along with various levels. The computed AER is considered to understand the efficiency of the proposed and the other depth procedures. The experimental results are summarised in the form of tables and is given in appendix.

Ttables 1 and 2 reveal that, all the depth procedures gave the same AER when the data were without contamination. But when the contamination level increased, all the depth procedures showed different results. Zonoid depth gets affected even when the contamination level increases to 10% and also with increases in sample size and dimensions. Halfspace depth gets affected at 20% level of contamination at both sample size and dimension increases. In some cases, SD gets affected at 30% level of contamination, but is fully affected at 40% contamination. At 40% level of contamination almost all depth procedures get affected except the proposed MMD procedure. It gives better results as compared with all other depth procedures. AER increases rapidly in all depth procedures except MMD procedure when the contamination level increases and also sample size and dimension increases. Projection depth also gives almost the same result as MMD, but in some cases it gives a little high AER compared with MMD. Finally, it is concluded that the, proposed MMD procedure performs well as compared with the other depth procedures.

VI. CONCLUSION

Location and scatter play a very important role in multivariate statistical methods. Data depth is one of the main emerging concepts to determine such measures. This paper proposed a new depth procedure called MMD and also proposed a weight function to estimate location and scatter. The performance of the proposed depth procedure is compared with the existing depth procedures. The experiments were carried out with application in discriminant analysis by computing AER. The computed AER under the proposed depth procedure MMD is lesser than the other depth procedures when contamination level increases and also when considering various sample sizes and dimensions. It is concluded that, the proposed MMD procedure can be applied in all multivariate statistical techniques, since it can tolerate certain levels of contamination and produces reliable results.

REFERENCES

[1]. T.W. Tukey, "Mathematics and picturing data," In *Proceedings of the International Congress on Mathematics (R. D. James, ed.)*, Canadian Math. Congress, 1975, Vol. 2, pp. 523–531

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

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

[4]. P.J. Rousseeuw, I. Ruts, "Bivariate location depth," *Journal of the Royal Statistical Society*, Vol. 45, Series C, pp. 516-526, 1996.

[5]. G. Koshevoy, K. Mosler," Zonoid trimming for multivariate distributions," *Annals of Statistics*, Vol. 25, pp. 1998-2017, 1997.

[6]. Y. Zuo, R. Serfling, "General notions of statistical depth function," *The Annals of Statistics*, Vol. 28, pp. 461–482, 2000.

[7]. Y. Vardi, C.H. Zhang,. "The multivariate L1-median and associated data depth," *Proceedings of the National Academy of Sciences*, U.S.A., 2000,Vol. 97, pp. 1423-1426.

[8]. K. Mosler, *Multivariate Dispersion, Central Regions and Depth : The Lift Zonoid Approach*. Springer, New York, 2002.

[9]. A. Cuevas, M. Febrero, R. Fraiman, "Robust estimation and classification for functional data via projection-based depth notions," *Computational Statistics*, Vol. 22, pp. 481-496, 2007.

[10]. J. Cuesta-Albertos, A. Nieto-Reyes, "The random Tukey depth," *Computational Statistics and Data Analysis*, Vol. 52, pp. 4979-4988, 2008.

[11]. Y. Hu, Q. Li, Y. Wang, Y. Wu, "Rayleigh Projection depth," *Computational Statistics*, Vol. 27, pp. 523-530, 2012.

[12]. D. Paindaveine, G. Van Bever, "From depth to local depth: a focus on centrality," *Journal of the American Statistical Association*, Vol. 105, pp. 1105-1119, 2013.

[13]. D. Chen, P. Morin, "Approximating majority depth," *Computational geometry: theory and applications*, Vol. 46, pp. 1059-1064, 2013.

[14]. R. Dyckerhoff, C. Ley, D. Paindaveine, "Depth-based runs tests for bivariate central symmetry," *Annals of the Institute of Statistical Mathematics*, Vol. 67, Issue 5, pp. 917-941, 2015.

[15]. M. Hubert, P.J. Rousseeuw, P. Segaert, "Multivariate and functional classification using depth and distance," *Advances in Data Analysis and Classification*.Vol. 11, Issue 3, pp. 445-466, 2017.

[16]. R. Muthukrishnan, G. Poonkuzhali, "Computing Robust Measure of Multivariate location – data depth approach," *Global Journal of Pure and Applied Mathematics*, Vol. 13, Issue 2, pp. 691-700, 2017.

[17]. I. Cascos, *Data depth: Multivariate statistics and geometry (W. Kendall, I. Molchanov, eds.) New Perspectives in Stochastic*

Geometry, Clarendon Press, Oxford University Press, Oxford, 2009.
[18]. K. Mosler, *Depth statistics (C. Becker, R. Fried, S. Kuhnt, eds.) Robustness and Complex Data Structures*, Festschrift in Honour of Ursula Gather, Springer, Berlin pp. 17–34, 2013.
[19]. R. Muthukrishnan, G. Poonkuzhali, "Computing median with data depth in multivariate data," *Journal of modern sciences*, Vol. 7, Issue 2, pp. 11-19, 2015.
[20]. P.C. Mahalanobis, "On the generalized distance in statistics", *Proceedings of National Institute of Sciences*. India, 1936,Vol. 12, pp. 49-55,
[21]. P. Chaudhuri, "On a geometric notion of quantiles for multivariate data," *Journal of the American Statistical Association*, Vol. 91, pp. 862-872, 1996.
[22]. R. Serfling, "Depth functions in nonparametric multivariate inference. In R. Liu, R. Serfling and D. Souvaine, eds., Data Depth: Robust Multivariate Analysis," *Computational Geometry and Applications, American Mathematical Society*, 1–16, 2006.
[23]. P.J. Rousseeuw, "Least median of squares regression," *Journal of the American Statistical Association*, Vol. 79, pp. 871-880, 1984.

AUTHORS PROFILE

R. Muthukrishnan graduated Doctor of Philosophy from Manonmaniam Sundaranar University in 2000. Now he is working as a professor in Bharathiar University. His research interests are Robust Statistical Inference, Sampling Techniques, Multivariate Analysis.

G. Poonkuzhali graduated Master of Philosophy in 2009 from Periyar University, India. She has been working as a Assistant Professor Since2009 in Indira Gandhi college of arts and science. Now she is doing PhD at Bharathiar University. Her research interests are Robust Inference, Multivariate Analysis.

Table 1 Apparent error rates under various depth procedures (p=2)

Depth Procedures	g=2 / (n1=n2=100)					g=3 / (n1=n2=n3=100)				
	Error					Error				
	0.00	**0.10**	**0.20**	**0.30**	**0.40**	**0.00**	**0.10**	**0.20**	**0.30**	**0.40**
MD	0.05	0.12	0.15	0.24	0.33	0.09	0.15	0.22	0.32	0.48
HD	0.07	0.12	0.20	0.29	0.35	0.09	0.16	0.26	0.41	0.54
PD	0.06	0.10	0.15	0.19	0.29	0.10	0.15	0.22	0.28	0.36
SD	0.06	0.10	0.16	0.26	0.35	0.09	0.15	0.22	0.37	0.53
ZD	0.05	0.14	0.19	0.27	0.35	0.22	0.3	0.33	0.40	0.51
LD	0.06	0.11	0.15	0.21	0.40	0.09	0.15	0.21	0.31	0.52
MMD	0.06	0.10	0.15	0.19	0.23	0.09	0.15	0.20	0.26	0.31
	(n1=n2=1000)					**(n1=n2=n3=1000)**				
MD	0.07	0.12	0.17	0.23	0.39	0.09	0.16	0.22	0.31	0.52
HD	0.07	0.12	0.18	0.30	0.44	0.09	0.15	0.24	0.41	0.59
PD	0.07	0.12	0.16	0.21	0.26	0.09	0.15	0.21	0.27	0.33
SD	0.07	0.12	0.17	0.27	0.42	0.09	0.15	0.22	0.36	0.56
ZD	0.07	0.13	0.19	0.28	0.41	0.16	0.24	0.31	0.41	0.56
LD	0.07	0.12	0.15	0.21	0.46	0.09	0.15	0.20		0.59
MMD	0.07	0.11	0.15	0.19	0.24	0.09	0.14	0.20	0.26	0.32
	(n1=n2=5000)					**(n1=n2=n3=5000)**				
MD	0.08	0.13	0.18	0.24	0.38	0.11	0.18	0.24	0.32	0.52
HD	0.08	0.13	0.18	0.30	0.44	0.11	0.17	0.25	0.41	0.59
PD	0.08	0.12	0.17	0.21	0.26	0.11	0.17	0.22	0.28	0.34
SD	0.08	0.12	0.18	0.27	0.41	0.11	0.17	0.24	0.36	0.57
ZD	0.08	0.14	0.20	0.28	0.41	0.11	0.18	0.26	0.38	0.56
LD	0.08	0.12	0.16	0.21	0.44	0.11	0.16	0.22	0.28	0.58
MMD	0.08	0.12	0.16	0.20	0.24	0.11	0.16	0.21	0.27	0.32

2 Apparent error rates under various depth procedures (p=5)

Depth Procedures	g=2 / (n1=n2=100)					g=3 / (n1=n2=n3=100)				
	Error					Error				
	0.00	**0.10**	**0.20**	**0.30**	**0.40**	**0.00**	**0.10**	**0.20**	**0.30**	**0.40**
MD	0.03	0.08	0.13	0.29	0.36	0.03	0.09	0.16	0.41	0.55
HD	0.03	0.08	0.16	0.32	0.38	0.03	0.09	0.24	0.43	0.58
PD	0.03	0.08	0.13	0.17	0.22	0.03	0.09	0.16	0.22	0.29
SD	0.03	0.08	0.13	0.3	0.37	0.03	0.09	0.19	0.44	0.55
ZD	0.03	0.12	0.27	0.24	0.41	0.04	0.14	0.23	0.37	0.55
LD	0.03	0.08	0.13	0.17	0.36	0.03	0.09	0.15	0.22	0.47
MMD	0.03	0.08	0.13	0.17	0.22	0.03	0.09	0.15	0.22	0.28
	(n1=n2=1000)					**(n1=n2=n3=1000)**				
MD	0.01	0.06	0.11	0.25	0.43	0.01	0.08	0.15	0.36	0.56
HD	0.01	0.07	0.15	0.28	0.45	0.01	0.09	0.21	0.39	0.57
PD	0.01	0.06	0.11	0.16	0.21	0.01	0.08	0.15	0.21	0.28
SD	0.01	0.06	0.12	0.29	0.45	0.01	0.08	0.17	0.41	0.58
ZD	0.01	0.08	0.14	0.23	0.4	0.01	0.09	0.19	0.32	0.51
LD	0.01	0.06	0.11	0.5	0.28	0.01	0.08	0.15	0.21	0.37
MMD	0.01	0.06	0.11	0.15	0.21	0.01	0.07	0.14	0.21	0.27
	(n1=n2=5000)					**(n1=n2=n3=5000)**				
MD	0.01	0.06	0.11	0.25	0.44	0.01	0.08	0.15	0.34	0.60
HD	0.01	0.07	0.14	0.29	0.45	0.01	0.09	0.19	0.40	0.61
PD	0.01	0.06	0.11	0.16	0.21	0.01	0.08	0.15	0.21	0.28
SD	0.01	0.06	0.12	0.28	0.46	0.01	0.08	0.16	0.39	0.62
ZD	0.01	0.06	0.13	0.23	0.39	0.01	0.09	0.17	0.31	0.53
LD	0.01	0.11	0.11	0.16	0.24	0.01	0.08	0.15	0.21	0.31
MMD	0.01	0.06	0.11	0.15	0.20	0.01	0.08	0.14	0.21	0.27

