Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 13, Number 2 (2017), pp. 691-700

http://www.ripublication.com

Computing Robust Measure of Multivariate Location – Data Depth Approach

R. Muthukrishnan[*] and G. Poonkuzhali[**]

[*] *Department of Statistics, Bharathiar University, Coimbatore 641046, Tamil Nadu, India.*

[**] *Department of Statistics, Bharathiar University, Coimbatore 641046, Tamil Nadu, India.*

*(Corresponding author: **)*

Abstract

Location parameter of distribution is to find a central value that best describes the data. The conventional methods of estimating location is used by all variables and cases in the given dataset, but it fails to produce reliable result when the dataset contains outliers. To overcome this problem the new method is proposed. In this paper, data depth approach was proposed for estimating location in multivariate data. Data depth means how deep a given point is in the entire data cloud. The proposed method combines the idea of Mahalanobis distance and robust measure of location namely, Minimum Covariance Determinant (MCD) estimator. The efficiency of the proposed method is compared with conventional method. The superiority of the method has been studied by computing multivariate location under real and simulating environment using R software. The conventional method of location, that is, classical location gets affected when the data contains outliers however the MCD based location is same under with/without outliers and also it is better than classical robust measure of location. The study is performed to find the location based on classical and robust depth procedures under real and simulating environment. The conventional procedure reliable results when the data meets the certain assumptions, no extreme points present in the data set and it fail when the data deviates and is contaminated. The proposed robust procedure performs well in both the situations, and it can tolerate up to certain level of contaminations.

Keywords and phrases: Location – Multivariate – Mahalanobis Distance – Outliers – Robust – MCD – Data Depth – Local Depth – Simulation.

2010 Mathematics Subject Classification: 62H12, 62G35

INTRODUCTION

Location measure plays a vital role in almost all univariate/multivariate statistical methods for analysing the data. Many graphical and computational methods have been established to estimate the measure of location for analysing data. The conventional methods such as sample mean, mode, etc., and robust methods such as MCD, MVE, M, etc., have been established to estimate the location. The conventional method of estimating location is used by all variables and cases in the given dataset, but it fails to produce reliable result when the dataset contains outliers. Consequently, Data depth approach is one of the approaches to find the true representative of the entire data cloud. Data depth means how deep a given point is in the entire data cloud. The concept of data depth is vital since it leads to centre-outward ordering of data points in multivariate data rather than ordinary ranking from smallest to largest ordering. The centre-outward ordering starts from the middle and moves in all direction. For each data point, depth value can be computed by using different notion of depth procedures. The data point with the highest depth value is being the deepest point and the data point with the smallest depth value is considered as the most outlying point in the data cloud. The maximum depth value approaches to 1. The data point with the maximum depth value is considered to be a good location. When there are more than one data points having same depth value then their average is considered as a deepest point.

Various notions of depth procedure have been established in the literature (see, Barnett (2003), Fraiman and Meloche (1999), Hu et al. (2012), Liu (1990), Liu et al. (2006), Oja (1983), Tukey (1975), Zuo and Serfling (2000). Comprehensive reviews on data depth are described in Cascos (2009), Mosler (2013), Muthukrishnan and Poonkuzhali (2015), Serfling (2006). Several researches have been performed in the data depth concept in recent days (Burr et al. (2011), Dyckerhoff and Mozharovskyi (2016), Dyckerhoff and Ley (2015), Lange et al. (2014(a), 2014(b)), Liu et al. (2013), Paindaveine and Van Bever (2013). The main objective of these depth procedures is to determine depth of each data point. Mahalanobis distance laid the foundation for computing depth of the data point. The depth procedure based on traditional measures of location and dispersion is termed as Mahalanobis depth and studied by Liu et al. (1999). Conventional estimates of mean vector and covariance matrix may provide unreliable results when the data set contains extreme points.

In this paper, a depth based procedure for estimating the measure of robust location and dispersion instead of classical location and dispersion in the computational procedure with Mahalanobis distance is proposed. Computes depth value for each

data point and then the deepest point (highest depth value) is considered as a location. The concept of local depth was proposed by Paidaveine and Van Bever (2013) is described and also the description of the proposed robust depth procedure is discussed in Section 2. The performance of the procedure is studied under real and simulation and the results are summarised in Section 3 and conclusion of the study is presented in the last section.

MATERIALS AND METHODS

Robust Depth Procedure

A concept of local depth was proposed by Paidaveine and Van Bever (2013). In this concept authors proposed using idea of symmetrisation of a distribution (a sample) with respect to a point in which depth is calculated. In their approach as an alternative of a distribution P^X ,a distribution $P_x = \frac{1}{2}P^X + \frac{1}{2}P^{2x-X}$ is used. For any $\beta \in [0,1]$, let introduce the smallest depth region bigger or equal to β ,

$$R^\beta(F) = \bigcap_{\alpha \in A(\beta)} D_\alpha(F), \tag{1}$$

Where $A(\beta) = \{\alpha \geq 0 : P[D_\alpha(F)] \geq \beta\}$. Then for a locality parameter β authors can take a neighbourhood of a point x as $R_x^\beta(P)$. (2)

Formally, let $D(.,P)$ be a depth function. Then the local depth with the locality parameter β and w.r.t a point x is defined as

$$LMD = LD^\beta(z,P) :\rightarrow (z, P_x^\beta), \tag{3}$$

Where $P_x^\beta = P(\cdot[R_x^\beta(P)])$ is conditional distribution of P conditioned on $R_x^\beta(P)$.

Let X be a p dimensional multivariate data cloud in R^p. The sample mean vector is defined and denoted by $\overline{X} = [\bar{x}_1\ \bar{x}_2\ ...\bar{x}_p]$, where, $\bar{x}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk}$, $k = 1,2,....p$ and the covariance matrix is

$$S = \begin{bmatrix} s_{11}\ s_{12}....s_{1p} \\ s_{21}\ s_{22}....s_{2p} \\ \vdots \\ s_{p1}\ s_{p2}....s_{pp} \end{bmatrix}, \text{ where, } s_{ik} = \frac{1}{n}\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k) \quad i = 1,2,...,p,\ k = 1,2,...,p$$

.

Mahalanobis distance (Mahalanobis 1936) of each point can be computed by

$$MD_d(x) = (x - \overline{X})' S^{-1}(x - \overline{X}) \tag{4}$$

and Mahalanobis depth of a point x is defined by

$$MD(x) = \left[1 + (x - \bar{X})' S^{-1} (x - \bar{X})\right]^{-1} \tag{5}$$

This depth procedure provides reliable depth value when the data set contains clean data, i.e., follow normality assumption. But if the data set contains outliers then Mahalanobis depth produces unreliable depth values of each data point, since the mean vector and covariance matrix itself unreliable one.

Mahalanobis depth procedure fails when the data encounters non-normal situations, since computational procedure is based on traditional mean and covariance matrix. Hence, it is proposed to use robust location and scatter instead of conventional mean vector and covariance. Among the robust procedures for estimating mean vector and covariance matrix, Minimum Covariance Determinant estimator (MCD) is the most reliable one. MCD was first introduced by Rousseeuw (1984) to find the robust measure of location and scatter. The objective of the method is obtained by choosing the halfset (h) points of the multivariate data whose covariance matrix has the lowest determinant. Initially the computational procedure is harder and applicable for limited number of data points and few dimensions. To overcome these difficulties Rousseeuw and Van Driessen (1999) proposed fast algorithm using c-step procedure, namely FAST-MCD.

Let M_{MCD} and S_{MCD} denote the sample mean and sample covariance matrix estimated by using MCD estimator. The robust depth can be computed by replacing the traditional mean vector and covariance matrix by robust alternatives in the computational procedure of mahalanobis depth. Robust depth is defined by,

$$RD(x) = \left[1 + (x - M_{MCD})' S_{MCD}{}^{-1} (x - M_{MCD})\right]^{-1} \tag{6}$$

RESULTS

To study the performance of the proposed depth procedure over the existing procedures, the experiments were carried out under real and simulation environment with the help of packages in R software and are summarised in this section.

Experiment 1

The efficiency of the robust (RD) over classical depth (MD), local depth (LMD) procedure has been studied with a real data. The data set (Anderson (2003)) contains 25 observations with 2 variables viz. head length and head breadth for first son. In fact, the data set with and without outliers are considered. The computed location based on traditional and robust procedure, and depth based location are summarised in the Table 1 and also see Figure 1.

TABLE 1

Computed Location under various depth procedure

Procedure	Location	
	Without outlier	With outlier
Classical	186,151	190,155
Robust	185,149	185,150
LMD	188,152 (0.970)	188,151 (0.976)
MD	188,152 (0.943)	191,155 (0.957)
RD	188,151 (0.925)	188,151 (0.962)

Source: See Anderson (2003) (p.109),

( ) indicates Depth value

Table 1 reveals that the classical location gets affected when the data contains outliers and The MCD based robust location provides almost stable in both with/without outliers. Local depth based on Mahalanobis depth (LMD) is almost similar to MCD based robust location. The robust depth based location is same under with/without outliers and also it is better than classical and robust measure of location. Further, it is also verified for the computed location based on depth procedure attains maximum depth value 1 while repeating the process.

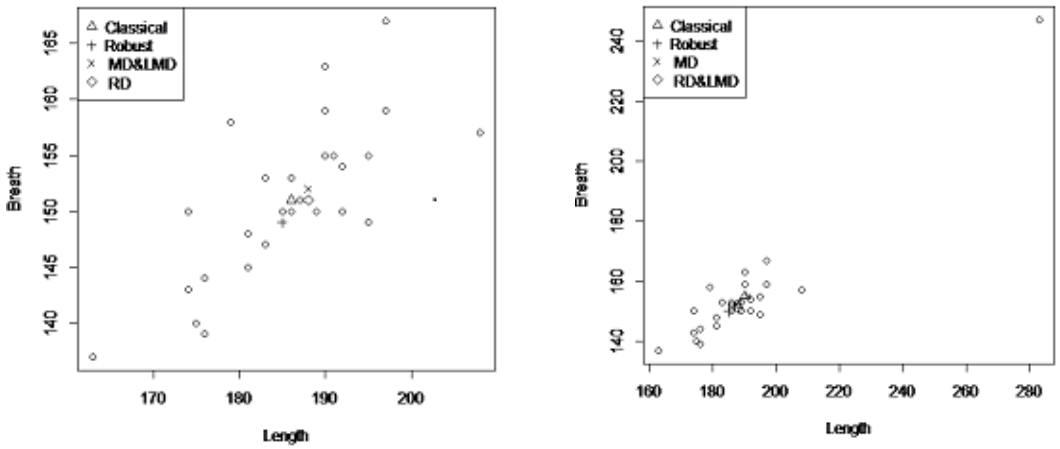

Figure 1. Location under classical, robust and local depth based procedures (without/with outliers)

Experiment 2

A simulation study is performed to find the location based on classical, local depth and robust depth procedures. The experiments were performed with the simulated data with mean vector (10,10) and the covariance matrix (1,0,0,1) under the sample

sizes 100, 1,000, 5,000 and 10,000. The same study was conducted under the various levels of location/scale/location and contaminations such as 10%, 20%, 30% and 40%. The computed locations are summarised in the Tables 2-4 and is given in appendix.

It is observed that, classical location and location induced by Mahalanobis depth, gets affected when the data contaminated has low level of location contamination and also location and scale contamination. MCD based location and location induced by robust depth represents the same location and tolerate certain amount of contaminations. Location induced by local depth produced different location point but not much affected by outliers like classical location and location induced by Mahalanobis depth. Location induced by Mahalanobis depth and local depth represents the different location point at different sample size and various level of contamination. It is observed that when sample size increases, specifically robust depth provides almost true location and tolerance limit goes to 40% of contaminations.

CONCLUSION

Measures of location play a prominent role in almost all statistical analyses. Many procedures have been to find a true location. But still it is a challenging task for the researchers to locate a true centre point in a data cloud, specifically in multivariate case. Data depth approach is emerging now-a-days to solve the issues. One of the depth based location procedure is local depth was proposed by Paidaveine and Van Bever (2013). This paper is proposed to find the location based on data depth value which is obtained by using robust measure of location and scale. The proposed robust procedure performs well and it can tolerate up to certain level of contaminations. The data depth approach finds the location, which is one among the entire data cloud in all situations but not MCD and classical location and local depth procedures. The limitation of the proposed procedures is that it takes lot of time, since the computational part of this procedure is uses MCD concept (need more iterations) and hence little difficult when compared with traditional procedure but it produces reliable results even in the non-normal situations.

REFERENCES

[1] Anderson, T.W., 2003, "An introduction to multivariate statistical analysis," A John Wiley and Sons, Inc., Publication.

[2] Barnett, V., 1976, "The ordering of multivariate data, Journal of the Royal Statistical Society," Series A 139**,** pp. 318-355.

[3] Burr, M.A., Rafalin, E., and Souvaine, D.L., 2011, "Dynamic maintenance of half-space depth for points and contours," arXiv:1109.1517 [cs.CG].

[4] Cascos, I., 2009, "Data depth: Multivariate statistics and geometry, in Kendall, W., and Molchanov, I.(eds.), New Perspectives in Stochastic Geometry," Clarendon Press, Oxford University Press, Oxford. DOI:10.1093/acprof:oso/9780199232574.003.0012.

[5] Dyckerhoff, R., and Mozharovskyi, P., 2016, "Exact computation of the halfspace depth," arxiv.org/abs/1411.6927v3.
[6] Dyckerhoff, R., and Ley, C., 2015, "Depth-based runs tests for bivariate central symmetry," Annals of the Institute of Statistical Mathematics, 67(5), pp. 917-941.
[7] Fraiman, R., and Meloche, J., 1999, "Multivariate L-estimation," Sociedad de Estadística e Investigación Operativa Test, 8, pp. 255-317.
[8] Hu, Y., Li, Q., Wang, Y., and Wu, Y., 2012, "Rayleigh Projection depth," Comput. Stat., 27, pp. 523-530.
[9] Lange, T., Mosler, K., and Mozharovskyi, P., 2014(a), "*DDα*-classification of asymmetric and fat-tailed data. In: Spiliopoulou, M., Schmidt-Thieme, L., and Janning, R. (eds.), Data Analysis," Machine Learning and Knowledge Discovery, Springer, Berlin, pp. 71–78.
[10] Lange, T., Mosler, K., and Mozharovskyi, P., 2014(b), "Fast nonparametric classification based on data depth," Statistical Papers, 55(1), pp. 49–69.
[11] Liu, R.Y., 1990, "On a notion of data depth based on random simplices," Annals of Statistics, 18, pp. 405–414.
[12] Liu, R.Y., Parelius, J.M., and Singh, K., 1999, "Multivariate analysis by data depth: descriptive statistics, graphics and inference," The Annals of Statistics, 27, pp. 783-858.
[13] Liu, R. Y., Serfling, R., and Souvaine, D. L., 2006, "Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications," American Mathematical Society.
[14] Liu, X., Zuo, Y., and Wang, Z., 2013, "Exactly computing bivariate projection depth contours and median," Computational Statistics and Data Analysis, 60, pp. 1-11.
[15] Mahalanobis, P.C., 1936, " On the generalized distance is statistics," Proc. Nat. Acad. Sci., 12, pp. 49-55.
[16] Mosler, K., 2013, "Depth statistics, in Becker, C., Fried, R., and Kuhnt, S., (eds.), Robustness and Complex Data Structures, Festschrift in Honour of Ursula Gather," Springer, Berlin.
[17] Muthukrishnan, R., and Poonkuzhali, G., 2015, "Computing median with data depth in multivariate data," Journal of modern sciences, 7, pp. 11-19.
[18] Oja, H., 1983, "Descriptive statistics for multivariate distributions, Statistics and Probability Letter," 1, pp. 327–332.
[19] Paindaveine, D., and Van Bever, G., 2013, "From depth to local depth : a focus on centrality," Journal of the American Statistical Association, 105, pp. 1105-1119.
[20] Rousseeuw, P., 1984, "Least median of squares regression," Journal of the American Statistical Association, 79, pp. 871-880.
[21] Rousseeuw, P., and Van Driessen, K., 1999, "A fast algorithm for the minimum covariance determinant estimator," Technometrics, 41, pp. 212-223.
[22] Serfling, R., 2006, "Depth functions in Nonparametric Multivariate Inference," DIMACS Series in Discrete Mathematics and Theoretical Computer Science, 72, American Mathematical Society.

[23] Tukey, J. W., 1975, "Mathematics and picturing data," in Proceedings of the International Congress on Mathematics, R. D. James (ed.), Canadian Math. Congress, 2, pp. 523–531.
[24] Zuo, Y., and Serfling, R., 2000, "General notions of statistical depth function," The Annals of Statistics, 28, pp. 461–482.

APPENDIX A

Table 2: *Computed Location under various depth procedure (Location contamination)*

Procedures	**n=100**				
	e=0.00	**e=0.10**	**e=0.20**	**e=0.30**	**e=0.40**
Classical	10.07,10.05	14.13,14.04	17.97,18.13	22.09,22.03	26.01,26.63
MCD	10.10,10.06	10.04,9.96	10.00,10.11	10.14,10.05	10.01,10.05
LMD	10.12,10.27(.987)	9.62,10.67(.983)	9.62,10.67(.984)	10.12,10.27(.967)	10.21,10.09(.946)
MD	10.21,10.09(.975)	11.20,11.32(.938)	11.20,11.32(.848)	11.20,11.32(.732)	10.41,10.40(.614)
RD	10.21,10.09(.970)	10.21,10.09(.966)	10.21,10.09(.972)	10.21,10.09(.996)	10.41,10.04(.999)
	n=1000				
Classical	10.04,9.99	14.02,13.98	18.05,18.01	22.00,22.02	26.03,26.01
MCD	10.03,9.98	10.02,9.98	10.07,9.98	10.03,10.00	10.05,10.00
LMD	9.99,9.96(.998)	50.62,50.76(.998)	10.29,9.94(.999)	10.03,10.00(.999)	10.03,10.00(.999)
MD	10.03,10.00(.999)	11.61,11.62(.961)	12.31,12.02(.999)	12.31,12.02(.753)	11.61,11.62(.649)
RD	10.03,10.00(.999)	10.03,10.00(.999)	10.03,10.00(.999)	10.03,10.00(.999)	10.03,10.00(.999)
	n=5000				
Classical	10.00,10.01	14.00,14.01	18.00,18.00	21.99,22.01	26.00,26.01
MCD	9.99,10.02	9.99,10.01	10.01,10.01	9.99,10.02	9.99,10.01
LMD	9.95,9.99(.999)	9.92,10.08(.999)	9.97,10.08(.999)	9.97,10.08(.999)	10.01,10.12(.999)
MD	10.00,10.01(.999)	12.27,12.20(.976)	12.27,12.20(.884)	12.30,12.21(.776)	12.30,12.21(.668)
RD	10.00,10.01(.999)	10.00,10.01(.999)	10.00,10.01(.999)	10.00,10.01(.999)	10.00,10.01(.999)
	n=10000				
Classical	10.00,10.01	14.00,14.01	18.00,18.01	22.01,22.01	25.99,26.01
MCD	10.00,10.01	10.00,10.02	9.99,10.01	10.02,10.02	10.01,10.01
LMD	9.99,9.99(.999)	11.27,10.05(.999)	10.02,10.01(.999)	10.02,10.01(.999)	10.02,10.01(.999)
MD	10.02,10.01(.999)	11.97,12.03(.971)	12.05,12.15(.877)	12.05,12.15(.770)	11.97,12.03(.661)
RD	10.02,10.01(.999)	10.02,10.01(.999)	10.02,10.01(.999)	10.02,10.01(.999)	10.02,10.01(.999)

( ) indicates depth value

Table 3 : *Computed Location under various depth procedure (Scale contamination)*

Procedures	**n=100**				
	e=0.00	**e=0.10**	**e=0.20**	**e=0.30**	**e=0.40**
Classical	10.14,10.05	10.18,9.88	10.14,10.09	9.92,10.08	9.90,10.12
MCD	10.19,9.90	10.15,10.02	10.14,10.11	10.17,9.98	9.95,10.14
LMD	10.33,9.93(.999)	10.05,10.04(.970)	10.05,10.04(.994)	10.27,9.97(.998)	10.05,10.04(.997)
MD	10.05,10.04(.993)	10.27,9.80(.987)	10.05,10.04(.995)	9.85,10.07(.998)	9.85,10.07(.998)
RD	10.05,10.04(.999)	10.05,10.04(.989)	10.05,10.04(.999)	10.05,10.04(.990)	10.05,10.04(.992)
n=1000					
Classical	10.01,9.99	10.01,9.98	10.01,10.00	10.00,9.98	9.96,9.93
MCD	10.02,10.00	10.01,9.99	10.00,9.98	10.01,10.01	10.03,10.00
LMD	10.07,10.11(.999)	10.07,10.11(.999)	9.89,10.11(.999)	10.06,10.07(.999)	10.00,10.00(.999)
MD	10.00,10.00(.999)	10.00,10.00(.999)	10.00,10.00(.999)	10.00,10.00(.999)	10.01,9.86(0.999)
RD	10.00,10.00(.999)	10.00,10.00(.999)	10.00,10.00(.999)	10.00,10.00(.999)	10.00,10.00(.999)
n=5000					
Classical	10.02,10.02	10.01,10.00	10.02,10.01	10.01,9.98	9.99,9.98
MCD	10.01,10.02	10.02,10.02	10.02,10.01	10.02,10.01	10.01,10.00
LMD	10.04,10.01(.999)	10.04,10.01(.999)	9.97,10.01(.999)	9.96,10.02(.999)	10.04,10.01(.999)
MD	10.04,10.01(.999)	10.02,9.98(.999)	10.04,10.01(.999)	9.99,9.99(.999)	9.98,9.96(.999)
RD	10.04,10.01(.999)	10.04,10.01(.999)	10.04,10.01(.999)	10.04,10.01(.999)	10.04,10.01(.999)
n=10000					
Classical	10.02,10.01	10.01,10.02	10.00,10.00	10.01,10.00	10.00,10.01
MCD	10.02,10.02	10.02,10.01	10.01,10.00	10.01,10.01	10.02,10.01
LMD	10.03,10.04(.999)	10.03,10.04(.999)	9.99,10.04(.999)	10.03,10.04(.999)	9.99,10.04(.999)
MD	10.03,10.01(.999)	10.03,10.04(.999)	10.03,10.01(.999)	10.03,10.01(.999)	10.02,10.01(.999)
RD	10.03,10.01(.999)	10.03,10.01(.999)	10.03,10.01(.999)	10.03,10.01(.999)	10.03,10.01(.999)

( ) indicates depth value

Table 4: *Computed Location under various depth procedure (Location and Scale contamination)*

Procedures	**n=100**				
	e=0.00	**e=0.10**	**e=0.20**	**e=0.30**	**e=0.40**
Classical	10.30,10.07	14.19,14.07	18.20,17.91	22.20,22.24	26.06,25.93
MCD	10.12,10.05	10.17,10.07	10.33,10.08	10.32,10.18	10.25,10.03
LMD	10.31,10.09(.999)	10.08,10.36(.999)	10.66,9.79(.997)	10.08,10.36(.991)	10.13,10.14(.999)
MD	10.31,10.10(.999)	10.97,10.77(.930)	11.40,10.90(.829)	11.37,11.49(.724)	11.37,11.49(.635)
RD	10.15,10.09(.998)	10.15,10.09(.999)	10.15,10.09(.983)	10.15,10.09(.984)	10.15,10.09(.996)
	n=1000				
Classical	9.99,10.04	13.97,14.02	18.02,18.04	22.02,22.00	25.93,26.00
MCD	10.00,10.04	10.00,10.02	9.97,10.04	10.01,10.04	9.99,10.01
LMD	10.05,10.01(.999)	10.05,10.01(.999)	10.05,10.01(.996)	10.05,10.01(.999)	10.05,10.01(.997)
MD	10.05,10.01(.996)	11.41,11.42(.955)	11.41,11.42(.854)	11.41,11.42(.750)	11.37,11.59(.644)
RD	10.05,10.01(.999)	10.05,10.01(.999)	10.05,10.01(.999)	10.05,10.01(.999)	10.05,10.01(.999)
LMD					
	n=5000				
Classical	10.02,10.01	14.01,13.98	18.01,18.00	23.03,22.03	26.01,26.00
MCD	10.03,10.01	10.02,10.01	10.04,10.01	10.02,10.02	10.03,10.01
LMD	10.04,10.05(.999)	10.11,9.93(.999)	10.11,9.96(.999)	10.07,9.97(.999)	10.05,10.02(.999)
MD	10.05,10.02(.999)	12.19,12.16(.977)	12.33,12.22(.844)	12.19,12.16(.777)	12.21,12.15(.668)
RD	10.05,10.02(.999)	10.05,10.02(.999)	10.05,10.02(.999)	10.05,10.02(.999)	10.05,10.02(.999)
LMD					
	n=10000				
Classical	9.99,10.01	13.99,14.01	17.98,18.00	22.00,22.02	25.98,26.02
MCD	10.00,10.01	10.00,10.01	9.98,10.01	10.00,10.01	9.99,10.00
LMD	9.99,9.96(.999)	10.05,9.97(.999)	10.00,10.02(.999)	10.00,10.02(.999)	10.00,10.02(.999)
MD	9.98,10.01(.999)	12.10,12.06(.974)	12.06,12.18(.880)	12.10,12.06(.774)	12.06,12.18(.666)
RD	10.00,10.02(.999)	10.00,10.02(.999)	10.00,10.02(.999)	10.00,10.02(.999)	10.00,10.02(.999)
LMD					

( ) indicates depth value