Available online at
http://www.ibnusina.utm.my/jfs

Journal of Fundamental Sciences

Article

A redefinition of mahalanobis depth function

Maman A. Djauhari* and Rian F. Umbara

Faculty of Mathematics and Natural Sciences, Institut Teknologi Bandung, Indonesia.
**To whom correspondence should be addressed. E-mail: maman@dns.math.itb.ac.id*

Received 5 October 2006
http://dx.doi.org/10.11113/mjfas.v3n1.23

ABSTRACT

Depth function is a new notion intensively developed in the last decade in the field of non-parametric statistics, computational geometry, algebra, and computer science. It is closely related to multivariate ordering, robust estimation, and outlier detection. One of the most widely used in statistics and related areas is the so-called Mahalanobis depth. In this paper we redefine that depth function by introducing a new one which is equivalent to the former, in the sense that they give the same multivariate ordering, less complicated to compute, and generalizes the "vanishing at infinity" property of depth function.

| center | covariance matrix | Mahalanobis depth | multivariate ordering |

1. Introduction

Suppose a random cloud in R^p or a probability distribution is given. A depth function measures how central a point is located in the cloud or the distribution. In the last decade, see, for example, Liu *et al*. (1999), Zuo and Serfling (2000), and Mosler (2004), the notion of depth function has been put into a general context of theory and applications. Among various depth functions, Mahalanobis depth is the oldest one dated 1936 (see Liu *et al*. (1999) for further information). Since then, various versions were proposed. For example, half-space depth proposed by Hodges in 1955 and by Tukey in 1975 as reported in Liu (1990), convex hull peeling depth proposed by Barnett (1976), Oja depth by Oja (1983), simplicial depth by Liu (1990), majority depth by Singh in 1991 as reported in Liu *et al*. (1999). In recent years, there are a lot of new depth functions available. Among them are regression depth proposed by Rousseeuw and Hubert (1999), tangent depth by Mizera (2002), projection depth by Zuo (2003), spherical depth by Elmore, Hettmansperger, and Xuan in 2004 as reported in Elmore (2005), and elliptical depth by Elmore (2005).

Theoretically, those depth functions are constructed in order to have a better one satisfying the following five key properties: affine invariant, monotone relative to deepest point, attain maximum value at the center, vanishes at infinity, and computationally efficient. Practically, the role of depth function in application is wider and wider. A comprehensive discussion on its wide applications such as in regression, confidence region, outlier identification, classification, discrimination, and multivariate control charts can be found in Mosler (2004). A specific

application in multivariate control charts is presented in Liu *et al*. (1999) and Dai *et al*. (2006) and an application in aviation safety analysis is presented in Cheng *et al*. (2000).

By definition, depth function is closely related to multivariate ordering in the sense of center-outward ordering in R^p, data outlyingness, and robust estimation. See, for example, Zuo and Serfling (2000) for the notion of these terminologies. In multivariate setting outlier region is defined as the complement of a depth central region. Furthermore, as in classical approach, the primary concern on robust estimation of location and covariance matrix lies in its property to have a high breakdown point. In classical approach, the most popular robust estimations are those constructed by minimizing the volume of ellipsoid (MVE) and by minimizing the determinant of covariance matrix (MCD) introduced by Rousseeuw (1985). Some improved versions of these two methods are proposed by many authors such as feasible solution algorithm in Hawkins (1994) and Hawkins and Olive (1999), fast MCD in Rousseeuw and van Driessen (1999), block adaptive computationally efficient outlier nominators (BACON) in Billor *et al*. (2000), and minimum vector variance in Herwindiati *et al*. (2006). It is to be noted that these versions are proposed in order to increase the computational efficiency.

The popularity of MVE- and MCD-based robust estimations is due to their commendable properties. They are affine-equivariant and have high breakdown point. See Lopuhaa and Rousseeuw (1991), Hadi (1992), Croux and Haesbroeck (1999), Rousseeuw and van Driessen (1999), Werner (2003), Hardin and Rocke (2004) for further discussion on these properties, and Jensen *et al*. (2005) for potential application in multivariate process control. However, because these robust estimations are constructed based on Mahalanobis depth, they are complicated to compute due to the need of inversion of covariance matrix.

The computational complexity of Mahalanobis depth, in terms of the number of operations in its computation, is still questionable especially for high dimensional data sets. The higher the dimension of the data sets the greater the number of operations in the computation of Mahalanobis distance the higher the computational complexity and the lower the computational efficiency. Can we redefine that depth function in a less complicated manner to compute? This is the problem that we intent to discuss in this paper. The main result consists of a new definition of Mahalanobis depth and a generalization of the "vanishing at infinity" property. The new definition will be formulated by introducing a new depth function which is equivalent to the former, i.e., they give the same multivariate ordering in the sense of center-outward ordering, and less complicated to compute.

This paper is organized as follows. In Section 2 we propose a new depth function and redefine the Mahalanobis depth. Section 3 will be focused on its computational complexity in terms of the number of operations in its computation. We show that asymptotically its relative complexity with respect to Mahalanobis depth is eight eleventh. This is a promising advantage. Additional remarks in Section 4 will close this presentation.

2. A Proposed Depth Function

Let Φ be the class of *p*-variate distributions and F_X be the distribution of a given random vector *X* in R^p. The following formal definition of depth function is given in Zuo and Serfling (2000).

Definition 1.

A non-negative and bounded mapping *D* from $R^p \times \Phi$ ® *R* is called depth function if it satisfies the following properties:

1. Affine invariance. $D(Ax + b, F_{Ax+b}) = D(x, F_X)$ for any random vector *X* in R^p, any non-singular matrix *A* of size $p \times p$, and any vectors *b* in R^p;

2. Maximality at center. $D(q,F) = \sup_x D(x,F)$, for any F in Φ and q in R^p called the center of F;
3. Monotonicity relative to deepest point. $D(x,F) \leq D(\lambda q + (1-\lambda)x, F)$ for any F in Φ having deepest point q and $0 \leq \lambda \leq 1$;
4. Vanishing at infinity. $D(x,F) \to 0$ if $\|x - q\| \to \infty$ and F in Φ.

A sample version of $D(x,F)$, denoted by $D_n(x,F_n)$ is defined by replacing F by a suitable empirical distribution F_n; n is the sample size. Thus, for a given random sample of size n, $D_n(x,F_n)$ is a function of x. In this setting, sample version of Mahalanobis depth is defined in the next paragraph.

Let X_1, X_2, ..., X_n be a random sample from p-variate distribution where the second moment exists. The sample mean vector and sample covariance matrix are, respectively,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ and } S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^t$$

Sample version of Mahalanobis depth of X_i is defined as (Liu *et al*. (1999))

$$MD_i = \frac{1}{1+(X_i - \bar{X})^t S^{-1}(X_i - \bar{X})}$$

This measures how depth X_i is with respect to the random cloud X_1, X_2, ..., X_n. The larger the value of MD_i the closer the point X_i to the center $\bar{X}$.

The second term of the denominator at the right hand side of MD_i is the so-called T^2-Hotelling's statistic or Mahalanobis distance. In the literature, see for example, Hadi (1992), Liu *et al*. (1999), Rousseeuw and van Driessen (1999), Werner (2003), and Herwindiati *et al*. (2006), that distance is computed directly from the definition. Thus, we need the inversion of sample covariance matrix S. This is a very tedious job especially for high dimensional data sets. Its computational complexity, in terms of the number of operations in its algorithm, is high. In what follows we redefine the Mahalanobis depth by introducing a new depth function with the following properties:

1. It is equivalent to Mahalanobis depth in the sense that they give the same multivariate ordering, i.e., the same center-outward ordering described by the second and third properties in Definition 1. See also, for example, Liu (1990) and Liu *et al*. (1999);
2. Its computation is less complicated than that of Mahalanobis depth.

A new definition of Mahalanobis depth will be formulated based on Proposition 1 which can be proved by using the property of the determinant of a partitioned matrix.

Proposition 1. Let X_1, X_2, ..., X_n be a random sample from p-variate distribution having the second moment and,

$$M_i = \begin{pmatrix} 1 & \left(X_i - \bar{X}\right)^t \\ \left(X_i - \bar{X}\right) & S \end{pmatrix}$$

a matrix of size (p+1)×(p+1) associated with X_i ; i = 1, 2, ... , n. If $|S|$ and $|M_i|$ are the determinant of S and M_i, respectively, then

$$MD_i = \frac{|S|}{2|S| - |M_i|}$$

Proof.

Let $T_i^2 = (X_i - \bar{X})^t S^{-1} (X_i - \bar{X})$. Then, by using the property of the determinant of a partitioned matrix, see Appendix A in Anderson (1966) or Mardia *et al*. (1979), we obtain,

$$T_i^2 = 1 - \frac{|M_i|}{|S|}.$$

From this equality we get,

$$MD_i = \frac{1}{1 + (X_i - \bar{X})^t S^{-1} (X_i - \bar{X})} = \frac{1}{1 + \left(1 - \frac{|M_i|}{|S|}\right)} = \frac{|S|}{2|S| - |M_i|}$$

as we have to prove.

In the following proposition we show that $|M_i|$ and MD_i define the same multivariate ordering.

Proposition 2. $MD_i \leq MD_j$ if and only if $|M_i| \leq |M_j|$.

Proof.

Proposition 1 gives us that $MD_i \leq MD_j$ if and only if $\frac{|S|}{2|S| - |M_i|} \leq \frac{|S|}{2|S| - |M_j|}$. This means that $MD_i \leq MD_j$ if only if $2|S| - |M_j| \leq 2|S| - |M_i|$ or if and only if $|M_i| \leq |M_j|$. Thus, we get the proof.

This proposition shows that the two functions $|M_i|$ and MD_i measure the depth of X_i and define the same multivariate ordering, i.e., the same center-outward ordering. Based on this result, we propose to use $|M_i|$ as a new depth function which is equivalent to Mahalanobis depth. Furthermore, as the value of MD_i is in (0, 1), the value of $|M_i|$ is in $(-\infty, |S|)$ because MD_i is proportional to the negative of the inverse of $|M_i|$. This suggests us to reformulate Definition 1 and generalize the fourth property of depth function described in that definition. The function D in Definition 1 does not need to be non-negative nor bounded. It just needs to be bounded above. Furthermore, the fourth property which says that $D(x, F) \to 0$ if $\|x - q\| \to \infty$ must be

generalized. As $|M_i|$ is in ($-\infty$, $|S|$), that property must be extended as follows: $D(x,F)$ tends to 0 or $-\infty$ at infinity.

3. Futher Result

An advantage of $|M_i|$ as a measure of the depth of X_i is that it does not need any matrix inversion in its computation. It only needs to compute the deteminant of a symmetric matrix. This means that $|M_i|$ is less complicated to compute than MD_i. Its computational complexity is certainly lower than that of Mahalanobis depth. More precisely, by using Cholesky decomposition to calculate the determinant of a symmetric matrix and the inverse of covariance matrix, the asymptotic relative computational complexity of $|M_i|$ with respect to MD_i, i.e., the ratio of the number of operations in their computations, is less than 1. This is given in Proposition 3.

Proposition 3. If $|M_i|$ and the inverse of covariance matrix are computed using Cholesky decomposition, then the asymptotic relative computational complexity of $|M_i|$ with respect to MD_i is $\frac{8}{11}$.

In fact, the number of operations in the algorithm to compute $|M_i|$ and MD_i are, respectively (substitute *m* by (*p*+1) in Appendix A and by *p* in Appendix B)

1. $\frac{2}{3}(p+1)^3+\frac{3}{2}(p+1)^2+\frac{5}{6}(p+1)+1$,
2. $\frac{11}{12}p^3+\frac{27}{8}p^2+\frac{39}{4}p+2$.

Thus, according to the number of operations in those algorithms, the asymptotic relative computational complexity of the proposed depth function with respect to Mahalanobis depth equals,

$$\lim_{p\to\infty}\frac{\frac{2}{3}(p+1)^3+\frac{3}{2}(p+1)^2+\frac{5}{6}(p+1)+1}{\frac{11}{12}p^3+\frac{27}{8}p^2+\frac{39}{4}p+2}=\frac{8}{11}.$$

Table 1 illustrates the difference of the number of operations in the computation of the two depth functions for various values of *p*. We see that, as *p* gets larger, the column $|M_i|$ divided by the column MD_i tends to eight eleventh.

Table 1. Number of operations in the computation of MD_i and $|M_i|$

p	MD_i	$\|M_i\|$
5	250	204
10	1354	1079
15	4001	3129
20	8880	6854
25	16678	12754
30	28082	21329
35	43780	33079
40	64459	48504
45	90806	68104
50	1.24E+05	92379
60	2.11E+05	1.57E+05
70	3.32E+05	2.46E+05

p	MD_i	$\|M_i\|$
80	4.92E+05	3.64E+05
90	6.96E+05	5.15E+05
100	9.51E+05	7.02E+05
150	3.17E+06	2.33E+06
200	7.47E+06	5.47E+06
250	1.45E+07	1.06E+07
300	2.51E+07	1.83E+07
350	3.97E+07	2.90E+07
400	5.92E+07	4.32E+07
450	8.42E+07	6.15E+07
500	1.15E+08	8.42E+07

Figure 1 is a graphical display of Table 1. The upper curve is for Mahalanobis depth and the lower for the proposed depth function. We see how the number of operations in the algorithm to compute the two depth functions differs considerably.

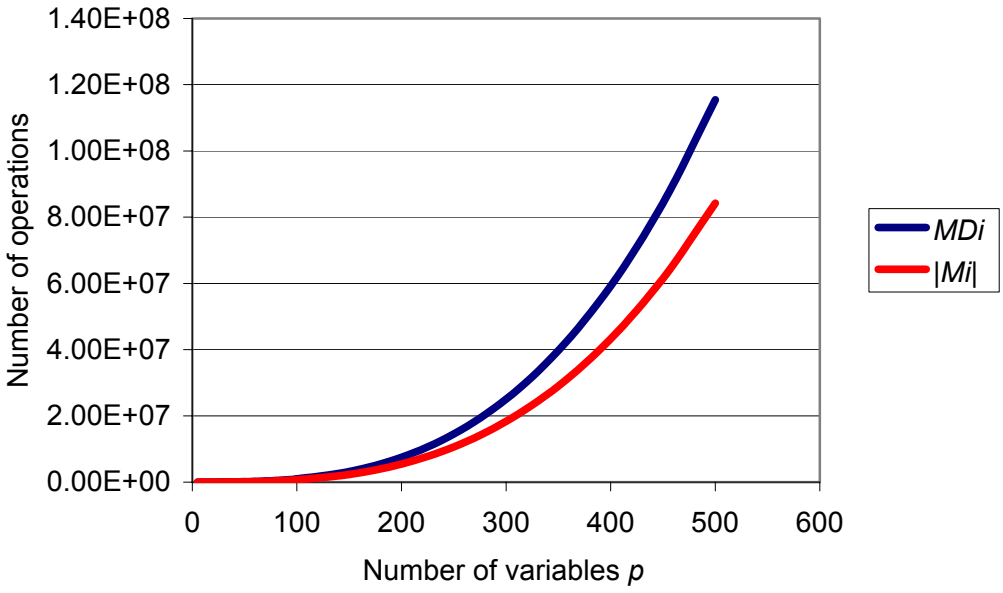

Figure 1. Number of operations in computing MD_i and $|M_i|$ as a function of p

4. Additional remarks

The advantage of the proposed depth function $|M_i|$ lies in its computation which is less complicated than that of Mahalanobis depth MD_i. Its computational complexity, i.e., the number of operations in the computation of $|M_i|$, is less than that of MD_i. Specifically, its asymptotic relative computational complexity is eight eleventh for p sufficiently large. However, $|M_i|$ has its own limitation with respect to MD_i. In the latter we need to compute the inverse of S once for all sample items, whereas in the former we need to involve S in $|M_i|$ for each sample item i. We also note that these two depth functions need the condition that the second moment of the population exists.

5. Acknowledgement

The authors gratefully acknowledge Institut Teknologi Bandung, Indonesia, for financial support under 2006 Competitive Research Grant, contract number: 0004/ K01.03.2/PL2.1.5/I/2006. They also appreciate the editor and the referees for their valuable suggestions.

6. References

[1] Anderson, T.W. (1966). *Introduction to multivariate analysis*. John Wiley & Sons, New York.
[2] Barnett, V. (1976). The ordering of multivariate data. *Journal of the Royal Statistical Society, Series A*, 139, pp. 318-352.
[3] Billor, N., Hadi, A.S., and Velleman, P.F. (2000). BACON: blocked adaptive computationally efficient outlier nominators. *Computational Statistics and Data Analysis*, 34, pp. 279-298.
[4] Cheng, A.Y., Liu, R.Y., and Luxhoj, J.T. (2000). Monitoring multivariate aviation safety data by data depth: control charts and threshold systems. *IIE Transaction*s, 32, pp. 861-872.
[5] Croux, C., and Haesbroeck, G. (1999). Influence Function and Efficiency of The Minimum Covariance Determinant Scatter matrix Estimator. *Journal of Multivariate Analysis*, 71, pp. 161-190.
[6] Dai, Y., Zhou, C., and Wang, Z. (2006). Multivariate Cusum control chart based on data depth for preliminary analysis. Department of Statistics, Nankai University, PR China. http://www.math.nankai.edu.cn/keyan/pre/preprint06/06-13.pdf
[7] Elmore, R.T. (2005). An affine-invariant data depth based on random hyperellipses. Workshop, Colorado State University, June 8 – 10.
[8] Hadi, A.S. (1992). Identifying multivariate outlier in multivariate data. *Journal of Royal Statistical Society B,* 53, pp. 761-771.
[9] Hardin, J., and Rocke, D.M. (2004). Outlier Detection in Multiple Cluster Setting Using Minimum Covariance Determinant Estimator. *Computational Statistics and Data Analysis*, 44, pp. 625-638.
[10] Hawkins, D.M. (1994). The feasible solution algorithm for the minimum covariance determinant estimator in multivariate data. *Computational Statistics and Data Analysis*, 17, pp. 197-210.
[11] Hawkins, D.M., and Olive, D.J. (1999). Improved feasible solution algorithm for high breakdown estimation. *Computational Statistics and Data Analysis*, 30, pp. 1-11.
[12] Herwindiati, D.E., Djauhari, M.A., and Mashuri, M. (2006). Robust multivariate outlier labeling. *Communication in Statistics* (conditionally accepted subject to revision).

[13] Jensen, W.A., Birch, J.B., and Woodall, W.H. (2005). High breakdown estimation methods for Phase Imultivariate control charts. Department of Statistics, Virginia Polytechnic Institute and State University. http://www.stat.org.vt.edu/dept/web-e/tech_reports/TechReport05-6.pdf
[14] Liu, R. (1990). On a notion of data depth based on random simplices. *Annals of Statistics*, 18, pp. 405-414.
[15] Liu, R.Y., Parelius, J.M., and Singh, K. (1999). Multivariate analysis by data depth: descriptive statistics, graphics and inference. Special Invited Paper. *Annals of Statistics*, 27, pp. 783-858.
[16] Lopuhaa, H.P., and Rousseeuw, P.J. (1991). Breakdown points of affine equivariance estimators of multivariate location and covariance matrices. *Annal of Statistics*, 19, pp. 229-248.
[17] Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979). *Multivariate Analysis.* Academic Press, London.
[18] Mizera, I. (2002). On depth and deep points: A calculus. *Annals of Statistics*, 30(6), pp.1681-1736.
[19] Mosler, K. (2004). Introduction: The geometry of data. *Allgemeines Statistisches Archiv*, 88, pp. 133-135, Physica-Verlag.
[20] Oja, H. (1983). Descriptive statistics for multivariate distributions. *Statistics and Probability Letters*, 1, pp. 327-332.
[21] Rousseeuw, P.J. (1985). Multivariate Estimation with High Breakdown Point. Paper appered in Grossman W., Pflug G., Vincze I. dan Wertz W., editors, *Mathematical Statistics and Applications*, **B**, pp. 283-297. D. Reidel Publishing Company.
[22] Rousseeuw, P.J., and Hubert, M. (1999). Regression depth. *Journal of the American Statistical Association*, 94, pp. 388-402.
[23] Rousseeuw, P.J., and van Driessen, K. (1999). A Fast Algorithm for The Minimum Covariance Determinant Estimator. *Technometrics*, 41, pp. 212-223.
[24] Werner, M. (2003). *Identification of Multivariate Outliers in Large Data Sets*. PhD Thesis, University of Colorado at Denver.
[25] Zuo, Y. (2003). Computing Projection Depth and Related Estimators. Michigan State University. http://dimacs.rutgers.edu/Workshops/Depth/Zuo.pdf
[26] Zuo, Y., and Serfling, R. (2000). General notions of statistical depth function. *Annals of Statistics*, 28, pp. 461-482.