Malaysian Journal of Mathematical Sciences 2(1): 1-24 (2008)

A Robust Estimation of Location and Scatter

Maman A. Djauhari
Faculty of Mathematics and Natural Sciences,
Institut Teknologi Bandung
Jalan Ganesa 10, Bandung 40132, Indonesia
E-mail: maman@dns.math.itb.ac.id

ABSTRACT

Statisticians face increasingly the task of analyzing large and high dimension multivariate data sets. This is due to the advances in computer technology which have facilitated greatly the collection of large data sets and, on the other hand, to the fact that most statistical experiments are multivariate in nature. One of the primary problems encountered in this task is robust estimation of location and scatter. In the literature the most popular and widely used robust parametric method for such parameter estimation is the so-called Fast MCD. However, although it is affine-equivariant and has high breakdown point, it is not apt when the data sets are of high dimension because its computational efficiency becomes lower. This is a direct consequence from the use of Mahalanobis distance or, equivalently, Mahalanobis depth in data ordering process which needs the inversion of covariance matrix and the use of MCD as the objective function. In this paper we propose a method which is as effective as Fast MCD but computationally more efficient. For this purpose, in multivariate ordering step, we use a new depth function which is equivalent to Mahalanobis depth and has lower computational complexity. Furthermore, in data concentration step, we use vector variance as the measure of multivariate scatter instead of covariance determinant and we replace the objective function MCD with minimum vector variance to reduce the complexity of this step. At the end of the paper we illustrate the effectiveness of this method using a simulation experiment.

Keywords: affine-equivariant, breakdown point, center-outward ordering, data depth, multivariate scatter, robust estimation of location and scatter.

INTRODUCTION

Suppose a random data cloud in R^p or a p-variate probability distribution is given. It is then natural to ask ‘How can we define an ordering in that cloud of random vectors or probability distribution?’ The ability to answer this question is essential in order to, for example, separate the ‘good’ from the ‘bad’ data, and more specifically, to construct a robust estimation of the parameters.

The idea of ordering in a space of dimension p = 1, 2, and 3 is as old as the history of human civilization which is 1.5 million years old. At the early history, human intellect was able to do the simplest ordering and to determine whether two objects differ to each other and which one is better.

Later on, in a more civilized artifact such as, for example, hieroglyphic writing dated 2500 BC we learn how phonetic characters, and characters representing ideas, were ordered to form historical documents. In modern era, when one says that Kuala Lumpur is geographically farther than Moscow from Paris, we understand that the circle with Moscow at its circumference and Paris as its center is inside the concentric circle with Kuala Lumpur at its circumference. This kind of ordering has similar meaning when in geocentric system of Ptolemy one says that the Moon is nearer than the Sun from the Earth. In Copernican heliocentric system, we say that Pluto is farther from the Sun than Venus because the elliptical orbit of Venus is entirely included in the elliptical orbit of Pluto. In these circumstances, objects in a space of dimension $p = 2$ and 3 are ordered in the sense of center-outward ordering. In recent years this notion of ordering, constructed based on the concept of depth function, has received very much attention in the statistical literature. Algebraic and geometric viewpoints of this concept have been developed and many approaches in multivariate analysis have been proposed. See, for example, half-space depth proposed by Hodges in 1955 and by Tukey in 1975 as reported in Liu (1990), convex hull peeling depth of Barnett (1976), Oja depth of Oja (1983), simplicial depth of Liu (1990), majority depth of Singh in 1991 as reported in Liu *et al.* (1999), regression depth of Rousseeuw and Hubert (1999), tangent depth of Mizera (2002), projection depth of Zuo (2003), spherical depth and elliptical depth of Elmore (2005).

The way people think and develop their viewpoint is fundamental in science. Kuhn (1997) has observed that the early developmental stages of most sciences have been characterized by continual competition between a number of distinct views of nature. 'View of nature' is a key success factor in exploring new frontier of science, the endless frontier. The ability to view the nature in distinct manner will guide us in producing a new paradigm, i.e., unprecedented and open-ended finding. An example of paradigm in conjunction with ordering problem is the so-called Cartesian coordinate system. Human civilization had to wait until Descartes, almost four centuries ago, investigated the ordering phenomenon and formulated it in that coordinate system to order objects in high dimensional space. The basis of that system is to order objects to make them easy to see clearly. Descartes, in his philosophical principles of scientific activities, addressed the following prophetic words which became the turning point of human civilization: "Never to accept anything for true which I did not clearly know to be such" In empirical sciences, the spirit of Descartes' words can be seen in the following wise phrase of George E.P. Box: "All models are wrong but some are useful" These phrases and Kuhn's 'view of nature' are in the spirit of this paper.

In the early history of modern statistics one of the great problems encountered was to measure how far a data point from the others in the space of any dimension. This was pioneered by French astronomer Pierce in 1852 when he had to separate anomaly data from the majority. See Kuwahara (1972) for more information. In modern statistics, following the advancement of probability theory, statisticians are interested in measuring how far a fact or evidence or a statistic differs from the hypothetical phenomenon. 'What do we mean by far?' and 'How far is far?' are among important problems that must be clarified. Statisticians realize that, in general, these questions push them to work in Hilbert space as their principal place and ordering objects in that space is their main tasks. In sampling context, the question becomes 'How do we define an ordering structure in a finite set of independent and identically distributed (i.i.d.) random vectors in R^p ?' To answer this question it is natural to consider the concept of center-outward ordering illustrated at the beginning of this section. One of the most attractive and widely used concepts to do that kind of ordering is the so-called depth function. Other concepts such as proposed, for example, by Wilks (1963), Rohlf (1975), Derquenne (1992), Pan *et al.* (2000), and Pena and Prieto (2001) are to order multivariate data in different ways. Due to its geometric structure which is easy to visualize the concept of depth function has received very much attention and thus it will be exploited in this paper.

The first idea of depth function came from parametric domain when in 1936 Mahalanobis introduced what we call now Mahalanobis depth. See Liu *et al.* (1999). It is the inverse of one plus the squared Mahalanobis distance. Consequently, it is sensitive to the presence of even one single outlier. This is the reason why at the early period of its development it had received far less attention than non-parametric depth function because the latter is robust. However, nowadays, there are many approaches available to handle the non-robustness of Mahalanobis distance. Among them, MVE (minimizing the volume of ellipsoid) and MCD (minimizing the covariance determinant) introduced by Rousseeuw (1985) are the most popular. Their popularity is due to their desirable properties, namely, affine-equivariant and 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. However, in recent years MCD receives much more attention than MVE due to its performance in estimating the true location and scatter. See Rousseeuw and van Driessen (1999), Werner (2003) and Hubert *et al.* (2005) for the details.

Some improved versions of MCD algorithm are available. For example, feasible solution algorithm in Hawkins (1994) and Hawkins and Olive (1999), Fast MCD (FMCD) algorithm in Rousseeuw and van Driessen (1999), block adaptive computationally-efficient outlier nominators (BACON) in Billor *et al.* (2000), improved FMCD algorithm in Hubert *et al.* (2005), and minimum vector variance (MVV) in our recent work Herwindiati *et al.* (2007). These versions are to increase the computational efficiency. Nowadays, FMCD is available in statistical packages such as S-Plus (function *cov.mcd*), R (*rrcov*), and SAS Version 9 (*PROC ROBUSTREG*). See Hubert *et al.* (2005) for the details. This shows that FMCD is very well accepted. However, its computational complexity increases exponentially when the dimension of the data sets increases. The larger the number of variables *p* the higher the complexity and thus the lower the computational efficiency. This is caused by the fact that FMCD involves the inversion of covariance matrix and covariance determinant.

The main concern of this paper is to find a new method which is able to reduce the level of complexity of FMCD and MVV, and maintain its effectiveness. An investigation to the structure of these algorithms will lead us to the conclusion that they consist of two main steps. The first step is the step of ordering the sample points in the sense of center-outward ordering. The second or data concentration step is to find a subset satisfying the objective function. In FMCD the objective function is to minimize covariance determinant while in MVV is to minimize vector variance. The method that we propose in this paper has the same structure as those algorithms but uses different concept. In the first step, instead of using Mahalanobis distance, we use a new data depth proposed in our recent work Djauhari and Umbara (2007) and, in the second step, we use vector variance as the scatter measure studied analytically in Djauhari (2007). See also Herwindiati *et al.* (2007) for its application in outliers labeling.

This paper is organized as follows. In the next section, some recent advancement in data depth will be discussed and a new data depth proposed in Djauhari and Umbara (2007) which is equivalent to Mahalanobis depth will briefly be presented. Later on, in Section 3 we discuss the so-called vector variance. Section 4 will be focused on the application of the results in Sections 2 and 3 for constructing a new method of robust estimation of location and scatter. We show that this method is computationally more efficient than FMCD and MVV. Furthermore, a simulation study in Section

5 shows that it is as effective as them. Additional remarks will close this presentation.

SOME RECENT ADVANCEMENT IN DATA DEPTH

Simple extensions of univariate statistics to the multivariate setting do not properly capture the higher-dimensional features of multivariate data because there is no natural and clear order principle in more than one dimension. For example, median and quantiles in univariate data analysis and inference have played important roles that their analogues in multivariate setting have been studied for years. In the last decade the concept of depth function has received considerable attention due to its ability to provide a center-outward ordering of vectors in the space of any dimension. More specifically, it is one of the fundamental concepts in the study of data depth which measures how deep a given point is with respect to a data cloud or a distribution. In other words, data depth is a centrality measure of a given random point. It then provides a new notion of multivariate location and scatter for the underlying distribution. In this regard, the most desirable data depth is of course the one which is invariant to the choice of coordinate system.

Although that notion of affine-invariance is desirable, most of the current depth functions which satisfy this property are quite cumbersome to compute in high dimension, e.g., Mahalanobis depth. In general, the desired depth functions are those which satisfy the following five key properties: affine-invariant, monotone relative to deepest point, attaint maximum value at the center, vanishes at infinity, and computationally efficient. Based on these properties the concept of depth function has been put into a general context of theory and applications in multivariate analysis. See, for example, Liu *et al.* (1999) and Zuo and Serfling (2000) for general theoretical discussion. Those who are interested in a comprehensive discussion on its applications in regression, confidence region, outlier identification, classification, and discrimination are suggested to consult Mosler (2004). Furthermore, in Liu *et al.* (1999), Dai *et al.* (2006) and Mosler (2004) one also can find an application of depth function in multivariate control charts. An application in aviation safety analysis is presented in Cheng *et al.* (2000). Other applications in nonparametric multivariate inference such as rank tests, quality control, robust estimation of location and scatter, multivariate goodness of fit, and outlier detection can easily be found in the literature. More importantly, from theoretical viewpoint, data depth offers new challenges in the interface of statistics, computer science, algebra, and computational geometry.

In what follows our attention will be focused on the Mahalanobis depth which is a principal base of FMCD and MVV and then a more efficient version will be presented.

A more efficient version of Mahalanobis depth

Let X_1 , X_2 , ..., X_n be a random sample from a *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 , see Liu *et al.* (1999), is defined as

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

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

We recognize that the second term of the denominator on the right hand side of MD_i is the squared 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.* (2007), that distance is computed directly from the definition. Thus, we need the inversion of sample covariance matrix *S*. This is a very tedious job with high computational complexity when the data sets are of high dimension. To eliminate this obstacle in our recent work, Djauhari and Umbara (2007), we define a new depth function which satisfies the following properties:

1. It is equivalent to Mahalanobis depth in the sense that they give the same multivariate ordering.
2. It is computationally more efficient than Mahalanobis depth.

Our definition is based on the following two propositions. The proof can be seen in Djauhari and Umbara (2007).

Proposition 1. Let X_1 , X_2 , ..., X_n be a random sample from a *p*-variate distribution where the second moment exists. If

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

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

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

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

Proposition 2 shows that $|M_i|$ is a depth function equivalent to MD_i . The two depth functions measure the depth of X_i and define the same ordering structure. The maximum value of $|M_i|$, which equals $|S|$, is attained at the center $\bar{X}$ and tends to $-\infty$ if X_i goes to infinity. This result indicates that the fourth property of depth function can be reformulated. Traditionally, that property says that a depth function must vanish at infinity. As $|M_i|$ is in ($-\infty$, $|S|$), that property should be extended: a depth function vanishes or tends to $-\infty$ at infinity.

Relative computational complexity

An advantage of $|M_i|$ compared with MD_i is that it does not need any matrix inversion in its computation. It only needs to compute the determinant of a symmetric matrix. Consequently, its computational complexity is lower than that of MD_i . More precisely, if $|M_i|$ is computed using Cholesky decompotition, then its asymptotic relative computational complexity with respect to MD_i , i.e., the ratio of the number of operations in the computation of $|M_i|$ and that of MD_i is $8/11$. See Djauhari and Umbara (2007) for the details. We learn how the complexity of $|M_i|$ differs significantly from MD_i . Due to this advantage in Section 4 $|M_i|$ will be exploited to construct a new method.

MEASURE OF SCATTER

Vector variance

Consider a random vector $X = \left(X_{(1)}^t \quad X_{(2)}^t\right)^t$ where $X_{(1)}$ and $X_{(2)}$ are of p and q dimensions, respectively. It is customary to write the covariance matrix Σ of X in the form of partitioned matrix

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

where $\Sigma_{ij} = E\left(\left(X_{(i)} - \mu_{(i)}\right)\left(X_{(j)} - \mu_{(j)}\right)^t\right)$ and $\mu_{(i)} = E\left(X_{(i)}\right)$; $i, j = 1, 2$.

Cleroux (1987) defines $Tr\left(\Sigma_{12}\,\Sigma_{21}\right)$ as a measure of linear relationship between the two random vectors $X_{(1)}$ and $X_{(2)}$, and he calls it vector covariance. It is equal to the trace, or the sum of all diagonal elements, of $\Sigma_{12}\Sigma_{21}$. By using the *vec* operator which transforms a matrix * of size ($r\times c$) into a column vector $vec(*)$ of dimension *rc* obtained by stacking the columns of * one underneath the other, see Muirhead (1982) and Schott (1997), vector covariance can be written as the scalar product $\left\langle vec\left(\Sigma_{12}\right), vec\left(\Sigma_{12}\right)\right\rangle$ or, equivalently, the norm $\left\|vec\left(\Sigma_{12}\right)\right\|^2$. Accordingly, we call the parameters $\left\|vec\left(\Sigma_{11}\right)\right\|^2 = Tr\left(\Sigma_{11}^2\right)$ and $\left\|vec\left(\Sigma_{22}\right)\right\|^2 = Tr\left(\Sigma_{22}^2\right)$ vector variance of $X_{(1)}$ and $X_{(2)}$, respectively. If $p = q = 1$, vector covariance is the square of the ordinary covariance and vector variance is the square of the ordinary variance.

In Djauhari (2007), we define vector variance (VV) as multivariate scatter measure. Let us consider an arbitrary random vector X of p dimension with covariance matrix Σ. By definition, VV of X is the sum of square of all elements of Σ. Heuristically, like covariance determinant (CD), VV is a non-negative real valued function of Σ which involves the covariance structure. Its value indicates the degree of how a multivariate distribution is scattered. The larger the value of VV the more scattered the distribution around its mean vector in a subspace of dimension $q \leq p$. It is equal to zero if and only if the distribution is degenerate at the mean vector. These properties are sufficient for VV to be considered heuristically as a scatter measure. Interestingly, this measure satisfies some desirable properties not possessed by CD.

In what follows we present an analytical approach. Let $X_1, X_2, \ldots,$ X_n be a random sample from a p-variate distribution where the second moment exists. Rousseeuw and van Driessen (1999, Theorem 1, p. 214) show an important result which is the basis of their notion of 'more scattered' data subset. Let H_1 and H_2 be two MCD subsets of **X** = { X_1, $X_2, \ldots, X_n$} of the same number of elements. Based on that theorem, they derive the criterion that H_1 is more scattered than H_2 if the covariance matrix S_2 of all sample items belonging to H_2 has smaller determinant than that of the covariance matrix S_1 of all sample items in H_1. They also show that if the procedure in that theorem is repeated several times, the results are convergent, i.e., there exist an index m such that $|S_m|$ = $|S_{m-1}|$ meaning that H_m is as scattered as H_{m-1}. This notion of 'more scattered' subset, defined by Rousseeuw and van Driessen (1999, p. 214), has the following implication. Let $\lambda_{k1} \geq \lambda_{k2} \geq \ldots \geq \lambda_{kp} > 0$ be the ordered eigenvalues of S_k ; k = 1, 2. Then, H_1 is more scattered than H_2 if $\lambda_{21} \cdot \lambda_{22} \ldots \lambda_{2p} < \lambda_{11} \cdot \lambda_{12} \cdots \lambda_{1p}$.

Another important result given by Rousseeuw and van Driessen (1999) is that an MCD subset H of **X** is separated from **X**\\$H$ by an ellipsoid. This implies that, if H_1 is more scattered than H_2, the smallest ellipsoid that covers H_2 has smaller volume than that of the smallest ellipsoid which covers H_1. There exists then an affine transformation such that the transformed former ellipsoid is contained entirely in the latter. This viewpoint will lead us to another notion of 'more scattered' subset which will be defined and exploited. We define that H_1 is more scattered than H_2 if $\lambda_{2i} < \lambda_{1i}$ for all $i = 1, 2, \ldots, p$.

This implies that, because all eigenvalues are positive and the covariance structure is involved, if H_1 is more scattered than H_2, then

1. $|S_2| < |S_1|$, and
2. $\|vec(S_2)\|^2 < \|vec(S_1)\|^2$.

The first conclusion is the necessary condition for H_1 to be more scattered than H_2 used by Rousseeuw and van Driessen (1999) and Hubert *et al.* (2005) to define MCD as the objective function in FMCD algorithm. The second one is the necessary condition for H_1 to be more scattered than H_2 that we use in our work Djauhari (2007). According to this necessary condition, the appropriate objective function is minimum vector variance.

Distributional properties

Let $X_1,\ X_2,\ ...\ ,\ X_n,\ ...$ be a sequence of random vectors of *p* components which converges in probability to a constant vector *c* in R^p and converges in distribution to a *p*-variate normal distribution $N_p(c,\Sigma)$. Let also $u(x)$ be a real valued function where u' exists and $u'(x) \neq 0$ for all *x* in the neighborhood of *c*. Then, $Y_n = u(X_n)$ can be written in the form

$$Y_n = u(c) + \left(\frac{\partial\, u(c)}{\partial\, X_n}\right)^t (X_n - c) + R_\xi \qquad (2)$$

where $R_\xi = \frac{1}{2}(X_n - c)^t A_\xi (X_n - c)$, the (*i,j*)-element of the symmetric matrix A_ξ is $a(i,j) = \frac{\partial^2 u(\xi)}{\partial\, X_n(i) \partial\, X_n(j)}$; $i, j = 1, 2, \ ...\ , p$, $X_n(i)$ is the *i*-th element of X_n, and ξ is in the neighborhood of *c* satisfying $\|\xi - c\| < \|X_n - c\|$. Since $X_n \xrightarrow{d} N_p(c,\Sigma)$ and the quadratic form

R_ξ converges faster than the linear form $\left(\frac{\partial u(c)}{\partial X_n}\right)^t \left(X_n - c\right)$ to 0, we have

$$Y_n \xrightarrow{d} N\left(\mu_Y, \sigma_Y^2\right) \tag{3}$$

where $\mu_Y = u(c)$ and $\sigma_Y^2 = \left(\frac{\partial u(c)}{\partial X_n}\right)^t \Sigma \left(\frac{\partial u(c)}{\partial X_n}\right)$.

In Djauhari (2007) we use the result (3) to investigate the asymptotic distribution of sample VV. Let $X_1,\ X_2,\ .\ .\ .\ ,\ X_n$ be a random sample from $N_p\left(\mu,\Sigma\right)$. Under this normality assumption, $vec\left(S\right)$ is the sum of (*n* – 1) i.i.d. random vectors. Thus, according to the central limit theorem $vec\left(S\right)$ converges in distribution to a p^2-variate normal. Its mean vector is $vec\left(\Sigma\right)$ and its covariance matrix is given in the following proposition. The proof can be seen, for example, in Muirhead (1982) and Schott (1997).

Proposition 3. Let *K* be the commutation matrix of size ($p^2 \times p^2$), i.e., $K = \sum_{i=1}^{p} \sum_{j=1}^{p} N_{ij} \otimes N_{ij}^t$ and N_{ij} is a (*p* × *p*) matrix having all elements equal 0 except its (*i*,*j*)-th element equals 1. The covariance matrix of $vec\left(S\right)$ is $\frac{1}{n-1}\left(I_{p^2} + K\right)\left(\Sigma \otimes \Sigma\right)$.

Corollary 1

Let $\Gamma = \left(I_{p^2} + K\right)\left(\Sigma \otimes \Sigma\right)$.

Then, we have $\sqrt{n-1}\left\{vec\left(S\right) - vec\left(\Sigma\right)\right\} \xrightarrow{d} N_{p^2}\left(0,\Gamma\right)$.

Corollary 2
Let $u\left(vec\left(S\right)\right)$ be a real valued function of $vec\left(S\right)$, u' exists and $u'\left(vec\left(S_0\right)\right) \neq 0$ for all S_0 in the neighborhood of Σ. Then, according to (3) and Corollary 1,

$$\sqrt{n-1}\left\{u\left(vec\left(S\right)\right)-u\left(vec\left(\Sigma\right)\right)\right\} \xrightarrow{d} N\left(0,\sigma^2\right) \tag{4}$$

where $\sigma^2 = \left(\frac{\partial\, u\left(vec\left(\Sigma\right)\right)}{\partial\, vec\left(S\right)}\right)^t \Gamma \left(\frac{\partial\, u\left(vec\left(\Sigma\right)\right)}{\partial\, vec\left(S\right)}\right)$.

Based on (4), if we define $u\left(vec\left(S\right)\right) = \left\|vec\left(S\right)\right\|^2$, we have the following result.

Proposition 4. $\sqrt{n-1}\left\{\left\|vec\left(S\right)\right\|^2 - \left\|vec\left(\Sigma\right)\right\|^2\right\} \xrightarrow{d} N\left(0,\sigma^2\right)$ with $\sigma^2 = 4vec\left(\Sigma\right)^t\, \Gamma\, vec\left(\Sigma\right)$.

The application of this proposition is seemingly complicated even for moderate value of *p* because σ^2 is a quadratic form of very high dimension, i.e., p^2. However, see Djauhari (2007), σ^2 can be written in a simple manner as follows.

Proposition 5. $\sigma^2 = 8\left\|vec\left(\Sigma^2\right)\right\|^2$.

The last two propositions show that $\sqrt{n-1}\left\{\left\|vec\left(S\right)\right\|^2 - \left\|vec\left(\Sigma\right)\right\|^2\right\}$ converges in distribution to a normal distribution with mean 0 and variance σ^2 which is equal to eight times the sum of square of all elements of Σ^2.

Some advantages

In what follows we underline some advantages of VV. First, we mention its speed of convergence. The asymptotic distribution in Proposition 4 is obtained by using the Taylor series of sample VV around Σ up to the second or linear term (2). Theoretically, because sample VV is a quadratic form, it can exactly be represented by a Taylor series up to the third or quadratic term. We point out in Djauhari (2007) that based on a simulation study the contribution of the quadratic terms into sample VV is very small, i.e., of order less than 10^{-5} even for small sample size n such as $n = 5$. This indicates that, for practical purpose, it is sufficient to approximate the distribution of sample VV by normal distribution given in Proposition 4. Second, see Djauhari (2007), the power of vector variance-based test is promising compared with likelihood-based test and covariance determinant-based test, when testing $H_0 : \Sigma = \Sigma_0$ versus $H_1 : \Sigma = k\ \Sigma_0$ with $k > 1$ and Σ_0 is a specified positive definite matrix. In general, the vector variance-based test is more sensitive than the likelihood-based test to small shift of covariance structure when n and k are small. Furthermore, the vector variance-based test and the covariance determinant-based test have similar performance when p is small. When p and n are large, the former is more sensitive than the latter to large shift of covariance structure. Third, VV is computationally more efficient than CD. Fourth, unlike CD, VV does not need the condition that the covariance matrix must be non-singular.

Those interesting properties of sample VV, together with those of $|M_i|$, will be exploited in the next section to construct a robust estimation method of location and scatter.

PROPOSED ROBUST METHOD

Covariance matrix is a principal source of information in multivariate analysis. Its estimation is at the primary concern of the analysts. However, the presence of even one single outlier can distort the classical estimate and making it useless. Outliers are data points that deviate from the usual assumptions and/or from the pattern suggested by the majority of data. They are more likely to occur in data sets with many observations and or variables, and often they do not show up by simple visual inspection. Detecting outliers in multivariate data is not trivial. A complicated problem often appears for two or more outliers. Moreover, outliers might be hard to detect when the number of variables exceeds two because it is not easy to get visual illustration. See Rousseeuw and van Zomeren (1990) for a deep discussion. The difficulty increases when the data set is of large size with large number of variables. Thus, robust estimation which incorporates the presence of outliers is needed. By robust estimation we mean the estimation we would have found without the outliers.

The concept of separating outlier suspects from the bulk of data has been developed in recent years. After separation process, outlier testing can be done on the group of suspects only. That concept has been introduced more than four decades and still received considerable attention. See, for example, Wilks's separation concept (1963) constructed based on the notion of closeness measure of two data subsets, Rohlf's (1975) based on the minimum spanning tree of inter-points Euclidean distance, and Derquenne's (1992) based on his definition of univariate transformation. However, these are not robust to the presence of outliers or too expensive. Later on, to mention some, Pan *et al.* (2000) and also Pena and Prieto (2001) propose to separate suspects from the group of 'good' data using projection methods. In particular, Pena and Prieto use projection on *2p* orthogonal directions; the first *p* maximizes the kurtosis and the remainder minimizes it. These projection methods are robust but very tedious and time consuming especially when the data sets are of high dimension. Rousseeuw (1985) proposes the MVE and MCD methods to construct robust estimates of location and scatter. Both methods possess the desirable properties, namely, affine-equivariant and high breakdown point. See also Rousseuw and Leroy (1987) and Lopuhaa and Rousseeuw (1991). However, as Hadi (1992) and Werner (2003) point out, they are computationally expensive. Rousseeuw and van Zomeren (1990) construct robust distance by modifying the number of data in each subset used in MVE. Hadi (1992) approximates MVE and modifies MCD by avoiding the possibility of covariance matrix to be singular in each subset. It was Rousseeuw and van Driessen (1999) who propose FMCD algorithm which becomes a very well accepted algorithm.

FMCD is an effective algorithm which is able to give high robust estimates of location and scatter. It is a very popular and widely used algorithm. Recently, Hubert *et al.* (2005) present an improvement in order to ensure that the final value of the objective function is as close as possible to the global minimum. They claim that: "It turn out that most of the currently available highly robust multivariate estimators are difficult to compute, which makes them unsuitable for the analysis of large data bases. Among a few exceptions is the MCD of Rousseeuw (1985). The MCD is a highly robust estimator of multivariate location and scatter that can be computed efficiently with the FMCD algorithm of Rousseeuw and van Driessen (1999)." In Herwindiati *et al.* (2007) we propose MVV algorithm to reduce the computational complexity of FMCD by replacing the objective function with minimum vector variance. Although FMCD and MVV are effective, they might not be computationally efficient for high dimension data sets because they involve the inversion of covariance matrix. Besides, as

Angiulli and Pizzuti (2005) point out, the computational efficiency is as important as the effectiveness of any algorithm.

In what follows we propose a new robust method, having the same structure as FMCD and MVV, by using the results discussed in the previous sections. It is as effective as FMCD and MVV, and computationally more efficient. To start with, we recall what FMCD and MVV are.

Separation process based on FMCD and MVV

Let $\{X_1, X_2, \cdots, X_n\}$ be a random data set of p-variate observations. FMCD algorithm is as follows.

1. Take a subset H_{old} containing $h = \left\lceil \frac{n+p+1}{2} \right\rceil$ data points where $\lceil x \rceil$ is the largest integer less than or equal to x.

2. Compute the mean vector $\overline{X}_{H_{old}}$ and covariance matrix $S_{H_{old}}$ of all observations belonging to H_{old}. Then, for all i = 1, 2, … , n, compute,

$$d^2_{H_{old}}(i) = d^2_{H_{old}}\left(X_i, \overline{X}_{H_{old}}\right) =$$

$$\left(X_i - \overline{X}_{H_{old}}\right)^t S^{-1}_{H_{old}} \left(X_i - \overline{X}_{H_{old}}\right)$$

3. Sort these squared distances in increasing order,

$$d^2_{H_{old}}(\pi(1)) \le d^2_{H_{old}}(\pi(2)) \le \ldots \le d^2_{H_{old}}(\pi(n))$$

where π is a permutation on $\{1, 2, \ldots, n\}$.

4. Define $H_{new} = \left\{X_{\pi(1)}, X_{\pi(2)}, \cdots, X_{\pi(h)}\right\}$

5. Compute $\overline{X}_{H_{new}}$, $S_{H_{new}}$ and $d^2_{H_{new}}\left(X_i, \overline{X}_{H_{new}}\right)$.

6. If $\left|S_{H_{new}}\right|$ = 0, repeat steps 1-5. If $\left|S_{H_{new}}\right|$ = $\left|S_{H_{old}}\right|$, the process is stopped. Otherwise, the process is continued until the k-th

iteration if $\left|S_{H_k}\right| = \left|S_{H_{k+1}}\right|$. Thus, we have $\left|S_{H_1}\right| \geq \left|S_{H_2}\right| \geq \ldots \geq \left|S_{H_k}\right| = \left|S_{H_{k+1}}\right|$.

Let $T_{MCD} = \bar{X}_{H_k}$ and $S_{MCD} = S_{H_k}$ be the location and covariance matrix issued from that algorithm. Robust squared Mahalanobis distance based on FMCD is defined as,

$$d^2_{RMCD}(X_i, T_{MCD}) =$$

$$(X_i - T_{MCD})^t \, S^{-1}_{MCD} \, (X_i - T_{MCD})$$

for all i = 1, 2, … , n. Observations having large $d^2_{RMCD}(X_i, T_{MCD})$ will be considered as suspects.

We recall also the MVV algorithm that we propose in Herwindiati *et al.* (2007). This algorithm is to replace the sixth point in the above algorithm with the following.

6*. If $Tr\left(S^2_{H_{new}}\right) = 0$, repeat steps 1-5. If $Tr\left(S^2_{H_{new}}\right) = Tr\left(S^2_{H_{old}}\right)$, the process is stopped. Otherwise, the process is continued until the k-th iteration if

$$Tr\left(S^2_{H_k}\right) = Tr\left(S^2_{H_{k+1}}\right).$$

We have $Tr\left(S^2_{H_1}\right) \geq Tr\left(S^2_{H_2}\right) \geq \ldots \geq Tr\left(S^2_{H_k}\right) = Tr\left(S^2_{H_{k+1}}\right)$.

Let $T_{MVV} = \bar{X}_{H_k}$ and $S_{MVV} = S_{H_k}$ be the location and covariance matrix given by MVV algorithm. Robust squared Mahalanobis distance based on MVV is,

$$d^2_{RMVV}(X_i, T_{MVV}) = (X_i - T_{MVV})^t \, S^{-1}_{MVV} \, (X_i - T_{MVV})$$

for all i = 1, 2, … , n. Large $d^2_{RMVV}\left(X_i\,, T_{MVV}\right)$ indicates that X_i is a suspect.

Proposed algorithm

The FMCD algorithm can be divided in two main steps. The first step which consists of the first three points can be considered as the step of multivariate ordering based on the Mahalanobis distance. The second one which consists of the remaining last three points is the step of data concentration with MCD as the objective function. The algorithm that we propose in the next paragraph is constructed based on the same structure as FMCD with the following criteria:

1. As the squared Mahalanobis distance is proportional to the inverse of Mahalanobis depth (1), in the first step we shall use the data depth $\left|M_i\right|$ instead of Mahalanobis distance.

2. In the second step, like MVV algorithm, instead of using MCD as the objective function we shall use MVV.

Thus, the algorithm that we propose is as follows.

1. Take a subset H_{old} containing $h = \left\lceil \frac{n+p+1}{2} \right\rceil$ data points.
2. Compute the mean vector $\bar{X}_{H_{old}}$ and covariance matrix $S_{H_{old}}$ of all observations belonging to H_{old}. Then compute,

$$\left|M_i\right| = \begin{vmatrix} 1 & \left(X_i - \bar{X}_{H_{old}}\right)^t \\ \left(X_i - \bar{X}_{H_{old}}\right) & S_{H_{old}} \end{vmatrix} \text{ for all } i = 1, 2, \ldots, n.$$

3. Sort these data depths in decreasing order,

$$\left|M_{\pi(1)}\right| \geq \left|M_{\pi(2)}\right| \geq \ldots \geq \left|M_{\pi(n)}\right|$$

where π is a permutation on {1, 2, … , n}.

4. Define $H_{new} = \left\{ X_{\pi(1)}, X_{\pi(2)}, \cdots, X_{\pi(h)} \right\}$
5. Compute $\bar{X}_{H_{new}}$, $S_{H_{new}}$ and $d^2_{H_{new}}\left(X_i, \bar{X}_{H_{new}} \right)$.
6. If $Tr\left(S^2_{H_{new}} \right) = 0$, repeat steps 1-5. If $Tr\left(S^2_{H_{new}} \right) = Tr\left(S^2_{H_{old}} \right)$, the process is stopped. Otherwise, the process is continued until the k-th iteration if $Tr\left(S^2_{H_k} \right) = Tr\left(S^2_{H_{k+1}} \right)$. Thus, we get

   $$Tr\left(S^2_{H_1} \right) \geq Tr\left(S^2_{H_2} \right) \geq \ldots \geq Tr\left(S^2_{H_k} \right) = Tr\left(S^2_{H_{k+1}} \right).$$

Let $T_{NEW} = \bar{X}_{H_k}$ and $S_{NEW} = S_{H_k}$ be the location and covariance matrix obtained from this algorithm. We compute $\left| RM_i \right|$ the robust version of $\left| M_i \right|$,

$$\left| RM_i \right| = \begin{vmatrix} 1 & \left(X_i - T_{NEW} \right)^t \\ \left(X_i - T_{NEW} \right) & S_{NEW} \end{vmatrix} \text{ for all } i = 1, 2, \ldots, n.$$

Based on this procedure, we consider as suspects all observations having small $\left| RM_i \right|$.

Based on this procedure, we consider as suspects all observations having small $\left| RM_i \right|$.

This algorithm, compared with FMCD, has certainly higher computational efficiency. In data ordering step, the asymptotic relative computational complexity of $\left| M_i \right|$ with respect to MD_i is 72.7 %. Furthermore, in data concentration step, the computational complexity of the quadratic form VV is lower than the multilinear form CD. As an illustration of this advantage, the computation speed of CD compared with VV for several values of p using MATLAB 7.0 is given in Table 1. See Herwindiati *et al.* (2007) for the details.

TABLE 1: Ratio between computation speed of CD and VV

p	Computation speed ratio (CD:VV)
10	6:1
25	13:1
50	34:1
75	67:1
100	95:1
150	127:1
200	231:1
250	326:1
300	443:1

ILLUSTRATIVE EXAMPLE

In Sections 2 and 3 we show analytically the advantages of the proposed method in terms of computational efficiency. In what follows, based on a simulation experiment, we show that the proposed method is as effective as FMCD of Rousseeuw and van Driessen (1999) and Hubert *et al.* (2005), and MVV of Herwindiati *et al.* (2007).

We generate 100 random data from a p-variate normal mixture model of two components $(1-\varepsilon)N_p\left(0,\mathrm{I}_p\right)$ + $\varepsilon N_p\left(\mathrm{e},\mathrm{I}_p\right)$, where p = 3, ε = 0.1, e = $\begin{pmatrix}5 & 5 & 5\end{pmatrix}^t$ and I_p is the identity matrix of dimension p. Thus, among them, we consider 10 contaminated data coming from different distribution as that of the majority of data. We then compute the squared Mahalanobis distance of all random data according to the classical method, FMCD, and MVV. Later on, we compute the robust data depth $|RM_i|$ proposed in Section 4. The results given by MATLAB 7.0 are visualized in Figures 1 – 4. In these figures the horizontal axis represents the number of observations. The vertical axis in Figures 1 – 3 is the squared Mahalanobis distance computed using the classical method, FMCD, and MVV, respectively, while in Figure 4 is the robust data depth $|RM_i|$.

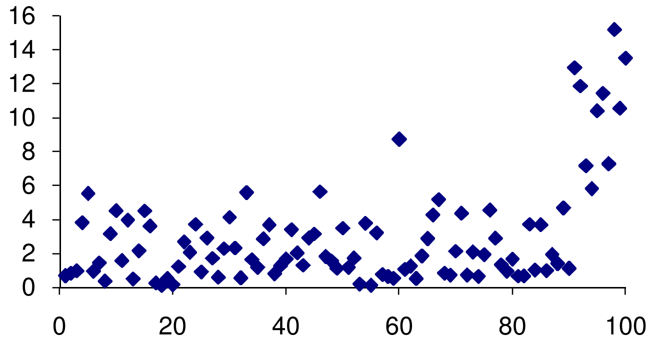

Figure 1: Scatter plot of classical squared Mahalanobis distance

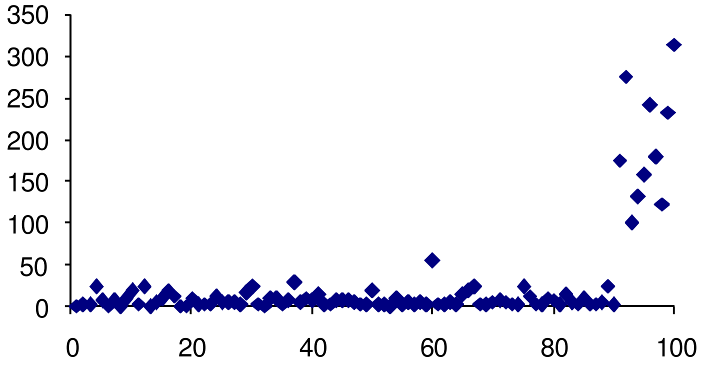

Figure 2: Scatter plot of robust squared Mahalanobis distance based on FMCD

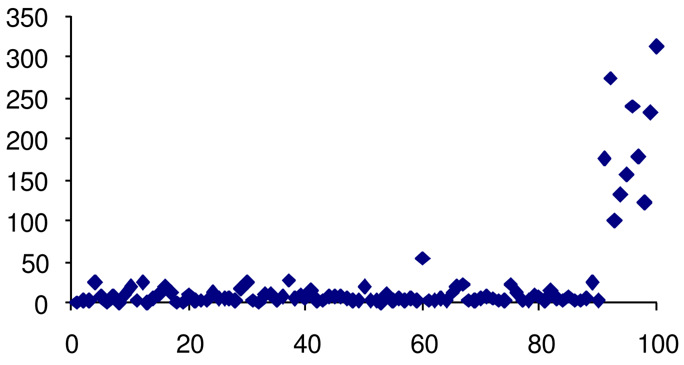

Figure 3: Scatter plot of robust squared Mahalanobis distance based on MVV

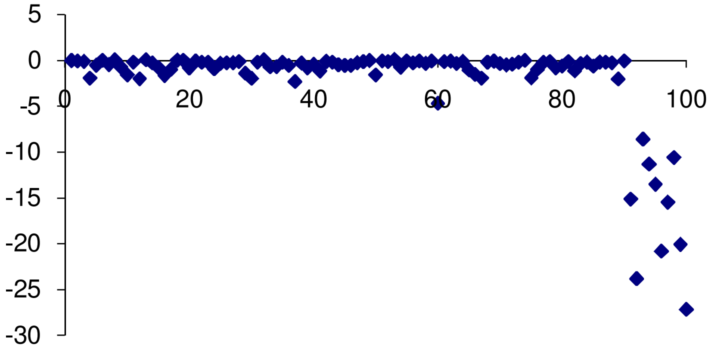

Figure 4: Scatter plot of $\left|RM_i\right|$ given by the proposed method

Figure 1 indicates that, based on the classical method, it is very hard to say that the 10 contaminated data are well separated. This method is not able to eliminate the presence of masking effect. On the other hand, Figures 2 and 3 show that the FMCD and MVV have, in the same manner, successfully separated all actual contaminated data. Furthermore, from Figure 4 we learn that the separation process based on the proposed method is as effective as FMCD and MVV but in reverse direction.

ADDITIONAL REMARKS

We consider that the structure of FMCD algorithm consists of two main steps. The first one is the step of multivariate ordering using Mahalanobis distance while the second is data concentration process based on MCD as the objective function. This viewpoint guides us to construct a new robust method where its algorithm has the same structure as FMCD but uses different criteria. In the first step, instead of Mahalanobis distance, we use the depth function $\left|M_i\right|$ which is able to reduce the computational complexity by 27.3 %. This is a significant gain. In the second step, we use VV as data concentration measure and MVV as the objective function. As VV is a quadratic form and CD is a multilinear form, the computational complexity of the former is lower than the latter. This is another computational advantage of the proposed method. A simulation experiment indicates that the proposed method is as effective as FMCD and MVV.

It is still in our investigation to define a depth function that can be used in multivariate data ordering step to replace $\left|M_i\right|$ where its computational complexity is as low as possible.

ACKNOWLEDGEMENT

It is a great honour for the author to be invited as a guest writer for this journal. He is very grateful to the Editor-in Chief for that invitation. As part of this work was supported by the Institut Teknologi Bandung 2006 Research Competitive Grant under contract number 0004/K01.03.2/PL2.1.5/I/2006, he also thanks the Institut Teknologi Bandung, Indonesia, for that support. His special thanks go to the Editorial Board for their help to improve this presentation.

REFERENCES

Angiulli, F., and Pizzuti, C. (2005). Outlier Mining and Large High-Dimensional Data Sets. *IEEE Transaction on Knowledge and Data Engineering,* 17(2), p. 203-215.

Barnett, V. (1976). The ordering of multivariate data. *Journal of the Royal Statistical Society, Series A*, 139, p. 318-352.

Billor, N., Hadi, A.S., and Velleman, P.F. (2000). BACON: blocked adaptive computationally efficient outlier nominators. *Computational Statistics and Data Analysis*, 34, p. 279-298.

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, p. 861-872.

Cleroux, R. (1987). Multivariate Association and Inference Problems in Data Analysis. Proceedings of the Fifth International Symposium on Data Analysis and Informatics, Vol. 1. Versailles, France.

Croux, C., and Haesbroeck, G. (1999). Influence Function and Efficiency of The Minimum Covariance Determinant Scatter matrix Estimator. *Journal of Multivariate Analysis*, 71, p. 161-190.

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

Derquenne, C. (1992). Outlier detection before running statistical methods. *Journal of SIAM*, 34, p. 323 – 326.

Djauhari, M.A. (2007). A measure of multivariate data concentration. *Journal of Applied Probability and Statistics* (to appear in November 2007 issue).

Djauhari, M.A., and Umbara, R.F. (2007). A redefinition of Mahalanobis depth function. *Journal of Fundamental Sciences*, 3(1), p. 150 – 157.

Elmore, R.T. (2005). An affine-invariant data depth based on random hyperellipses. Workshop, Colorado State University, June 8 – 10.

Hadi, A.S. (1992). Identifying multivariate outlier in multivariate data. *Journal of Royal Statistical Society B,* 53, p. 761-771.

Hardin, J., and Rocke, D.M. (2004). Outlier Detection in Multiple Cluster Setting Using Minimum Covariance Determinant Estimator. *Computational Statistics and Data Analysis*, 44, p. 625-638.

Hawkins, D.M. (1994). The feasible solution algorithm for the minimum covariance determinant estimator in multivariate data. *Computational Statistics and Data Analysis*, 17, p. 197-210.

Hawkins, D.M., and Olive, D.J. (1999). Improved feasible solution algorithm for high breakdown estimation. *Computational Statistics and Data Analysis*, 30, p. 1-11.

Herwindiati, D.E., Djauhari, M.A., and Mashuri, M. (2007). Robust Multivariate Outlier Labeling. *Communication in Statistics – Computation and Simulation* (in printing)

Hubert, M., Rousseeuw, P.J., and van Aelst, S. (2005). Multivariate Outlier Detection and Robustness. *Handbook of Statistics*, 24, Elsevier B.V., p. 263-302.

Kuhn, T.S. (1997). *The Structure of Scientific Revolution*. The University of Chicago Press, Ltd., London.

Kuwahara, S.S. (1997). Outlier testing: Its history and application. *BioPharm; The Technology and Business of Biopharmaceutical*, 10(4), p. 64-67.

Liu, R. (1990). On a notion of data depth based on random simplices. *Annals of Statistics*, 18, p. 405-414.

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, p. 783-858.

Lopuhaa, H.P., and Rousseeuw, P.J. (1991). Breakdown points of affine equivariance estimators of multivariate location and covariance matrices. *Annal of Statistics*, 19, p. 229-248.

Mizera, I. (2002). On depth and deep points: A calculus. *Annals of Statistics*, 30(6), p.1681-1736.

Mosler, K. (2004). Introduction: The geometry of data. *Allgemeines Statistisches Archiv*, 88, p. 133-135, Physica-Verlag.

Muirhead, R.J. (1982). *Aspect of Multivariate Statistical Theory*. John Wiley & Sons, Inc., New York.

Oja, H. (1983). Descriptive statistics for multivariate distributions. *Statistics and Probability Letters*, 1, p. 327-332.

Pan, J-X., Fung, W-K., and Fang, K-T. (2000). Multiple outlier detection in multivariate data using projection pursuit technique. *Journal of Statistical Planning and Inference*, 83, p. 153-167.

Pena, D., and Prieto, J.F. (2001). Multivariate outlier detection and robust covariance matrix estimation. *Technometrics*, 3, p. 286-322.

Rohlf, F.J. (1975). Generalization of the Gap Test for the Detection of Multivariate Outlier. *Biometrics*, 31, p. 93 – 101.

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, p. 283-297. D. Reidel Publishing Company.

Rousseeuw, P.J., and Hubert, M. (1999). Regression depth. *Journal of the American Statistical Association*, 94, p. 388-402.

Rousseeuw, P.J., and Leroy A. M., (1987). *Robust Regression and Outlier Detection*. John Wiley, New York.

Rousseeuw, P.J., and van Driessen, K. (1999). A Fast Algorithm for The Minimum Covariance Determinant Estimator. *Technometrics*, 41, p. 212-223.

Rousseeuw, P.J., and van Zomeren, B.C. (1990). Unmasking Multivariate Outliers and Leverage Points. *Journal of American Statistical Association*, 85(411), p. 633-639.

Schott, J.R. (1997). *Matrix Analysis for Statistics*. John Wiley & Sons, Inc., New York.

Werner, M. (2003). *Identification of Multivariate Outliers in Large Data Sets*. PhD Thesis, University of Colorado at Denver.

Wilks, S.S. (1963). Multivariate Statistical Outliers. *Shankya* A, 25, p. 407 – 426.

Zuo, Y. (2003). Computing Projection Depth and Related Estimators. Michigan State University. http://dimacs.rutgers.edu/Workshops/Depth/Zuo.pdf

Zuo, Y., and Serfling, R. (2000). General notions of statistical depth function. *Annals of Statistics*, 28, p. 461-482.