A NEW CRITERION IN ROBUST ESTIMATION FOR LOCATION AND COVARIANCE MATRIX, AND ITS APPLICATION FOR OUTLIER LABELING

DISSERTATION

**Submitted in partial satisfaction of
the requirement for the degree of
Doctor in Institut Teknologi Bandung**

By :

DYAH ERNY HERWINDIATI

NIM: 30101001

**INSTITUT TEKNOLOGI BANDUNG
2006**

A NEW CRITERION IN ROBUST ESTIMATION FOR LOCATION AND COVARIANCE MATRIX, AND ITS APPLICATION FOR OUTLIER LABELING

DISSERTATION

**Submitted in partial satisfaction of
the requirement for the degree of
Doctor in Institut Teknologi Bandung**

By :

DYAH ERNY HERWINDIATI

NIM: 30101001

INSTITUT TEKNOLOGI BANDUNG
2006

A NEW CRITERION IN ROBUST ESTIMATION FOR LOCATION AND COVARIANCE MATRIX, AND ITS APLICATION FOR OUTLIER LABELING

By

Dyah Erny Herwindiati

NIM : 30101001

Institut Teknologi Bandung

Approved

Date, 20 March 2006

Principal Supervisor

Prof. Dr. Maman A. Djauhari.

Supervisor I

Dr. Sutawanir Darwis

Supervisor II

Dr. Nihal Yatawara.

ABSTRAK

KRITERIA BARU DALAM PENAKSIRAN PARAMETER LOKASI DAN MATRIKS KOVARIANSI SECARA *ROBUST* DAN APLIKASINYA DALAM PELABELAN *OUTLIER*

Oleh

DYAH ERNY HERWINDIATI

NIM: 30101001

Disertasi ini mengemukakan metode penaksiran *robust* parameter lokasi dan matriks kovariansi serta aplikasinya dalam pelabelan *outlier* yang merupakan pengembangan dari metode-metode yang ada, khususnya metode MVE, MCD, *modified* MCD (MMCD), FSA dan FMCD. Semua metode tersebut memiliki *breakdown point* yang tinggi, mendekati 0,5 bila ukuran sampelnya semakin besar. Perkembangan mutakhir, bertarikh 2003, menunjukkan bahwa FMCD lebih unggul dibandingkan MVE, MMCD dan FSA. Namun demikian, penulis menilai bahwa efisiensi algoritma meminimumkan determinan matriks kovariansi, yang digunakan sebagai basis FMCD, menurun secara drastis tatkala banyaknya variabel p meningkat. Hal ini disebabkan karena perhitungan determinan matriks kovariansi yang membutuhkan waktu berorde $O(2^p)$. Oleh karena itu metode-metode tersebut kurang tepat bila diterapkan untuk matriks data berukuran besar dan berdimensi tinggi.

Mengingat hal tersebut, dalam disertasi ini penulis memperkenalkan metode *robust* yang baru, yang didasarkan kepada kriteria meminimumkan variansi vektor. Metode ini bersifat *robust* dan memiliki *breakdown point* yang sama dengan metode-metode yang telah disebut di depan. Keunggulan metode ini terletak kepada efisiensi algoritma yang tinggi, yakni berorde $O(p^2)$ dengan tingkat efektivitas yang sama dengan FMCD.

Kata kunci : *breakdown point, lokasi, matriks kovariansi, pelabelan outlier, penaksir robust, variansi vektor.*

ABSTRACT

A NEW CRITERION IN ROBUST ESTIMATION FOR LOCATION AND COVARIANCE MATRIX, AND ITS APPLICATION FOR OUTLIER LABELING

BY

DYAH ERNY HERWINDIATI

NIM: 30101001

In this dissertation we propose a method for robust estimation of location and covariance matrix and its application in outlier labeling, which is a development of the previous methods, especially MVE, MCD, modified MCD (MMCD), FSA and FMCD. All of these methods have high breakdown point (BP), close to 0.5 when sample size becomes large. Some recent developments, dated until the end of 2003, show that FMCD has better performance than MVE, MMCD and FSA. However, in our opinion, its algorithm efficiency to minimize the covariance determinant, the criterion used in FMCD, decreases drastically when the number of variables p increases. This is caused by the complexity of the computation of covariance matrix determinant which is of order $O(2^p)$. Thus, those methods are not appropriate for large and high dimension data set.

The above phenomenon motivates us to propose in this dissertation a new robust method using minimum vector variance as its criterion. This method is not only *robust* but also has the same *breakdown point* as those methods mentioned above. An advantage of our method lies in its algorithm efficiency which is of order $O(p^2)$ with the same effectiveness as FMCD.

keywords : *breakdown point, covariance matrix, location, outlier labeling, robust estimation, vector variance.*

GUIDELINES TO USE DISSERTATION

ACKNOWLEDGMENTS

In the name of Allah, the Most Gracious, the Most Merciful. The author expresses gratitude to Allah S.W.T for allowing her to finish this dissertation.

This dissertation is a result of a series of long period researches. The author realizes that without any helps and supports of various parts, morally or materially this dissertation is impossible to come out. With all due respects, the author would thank to Prof. Dr. Maman A. Djauhari as the Chief of Supervisors for his supports and encouragement. Special thank goes to the Team of Supervisors Dr. Sutawanir Darwis for valuable suggestions. The author also wishes thank to Dr. Nihal Yatawara for allowing the author to conduct a collaborative research at the Curtin University of Technology.

The author also thanks to all chairs of Universitas Tarumanagara, the dean of Faculty of Information Technology of UNTAR for allowing the author to finish this doctorate program. Also to Dr Pudji Astuti, and the whole members of Mathematics Department of ITB, the author wishes thank for every support and assistance.

Very special thank also goes to the author's husband Ferrial Sofyan, the author's father Bambang Legowo, the author's mother Soedijah and her beloved daughter Okky Wulandari for their patience, deep understanding and their prays.

Finally thank also goes to all friends in the Pasca Sarjana Matematika ITB, all staff of FTI UNTAR, for their supports and assistance during the study.

CONTENTS

FIGURES

TABLES

Chapter I Preliminary

I.1. Background

Identifying an outlier data in a larger group of data is a very important topic in statistical and data analysis. It is so, because the data we are identifying must be in clean condition, in other word, free of any influences of occurrence of outliers. On the other sides, an outlier, is an abstract concept which is not easy to define. There are a number of definitions which are often used in daily practices, for instance, one defined by Grubbs (1969), Hawkins (1980), Beckman and Cook (1983), Rousseeuw and van Zomeren (1990), and one from Barnett and Lewis (1984). In this dissertation the author follows the definition given by Barnett and Lewis (1984). They define an outlier to be one or more data which are not consistent among others.

The word 'not consistent' on the definition is not easy to be formulated in general situations. This reason makes people, up to now, develop better methods in identifying outliers. In the univariate case, we see various development of methods, for example Irwin (1925), Thomson (1935) and Pearson and Chandra Sekar (1936) in early XX century, Dixon (1950), Grubbs (1950), Tietjen and More (1972), Tukey (1977), Rosner (1975, 1983), Beckman and Cook (1983), Iglewicz and Hoaglin (1993), Barnett and Lewis (1984), Kuwahara (1997), and for much more recent ones Djauhari (1999, 2001, 2003).

Next, in the multivariate case, discussions about development of methods in identifying outliers most people use Wilks's (1963) as the starting point. As we can see in literatures, since then the problem of outlier identification for multivariate cases has became a challenging area of research and grows very rapidly. Today, its role can be found in every work based on multivariate data. Even for groups of large data and high dimension such as in data mining and knowledge discovery (Angiulli and Pizzuti (2005)), and intrusion detection (Ye et al.(2003)). In line with Angiulli and Pizzuti (2005), in the multivariate case there are two important problems need to take into account. The first one is the procedure of justification, and the second one is the efficiency or how fast algorithms work. These two problems are the main topics of this research.

There are many procedures to identify outliers. One of them is by using the outliers labeling approach as an important stage. This stage is very useful to separate data suspected as outliers from the group of main data. Researchers proposed different methods and terminology in outlier labeling for the same purpose. In the univariate case, it is known the Tukey labeling method (1977, p. 44) which says that data outside the fence as 'unclean data'. In the multivariate case, there are many ways of labeling, for example are ones proposed by Rousseeuw (1985), Hadi (1992), Rousseeuw and van Driesen (1999), Pan et al. (2000), and Pena and Prieto (2001).

Rousseeuw (1985) introduced two criterias to separate data into two groups, which they called as 'good' and 'not good' group. The first is the criteria of minimizing the determinant of covariance (minimum covariance determinant or is abbreviated as MCD) and the second is the criteria of minimizing the volume of ellipsoid (minimum volume ellipsoid or is abbreviated as MVE). Hadi (1992) uses the modified MVE or MCD (modified MCD abbreviated as MMCD) to ensure nonsingularity of covariance matrix. This criteria is used to separate the sets ‘basic’ and ‘non-basic’ data. In the development, MCD is appreciated broader and better than MVE because the effectivity and efficiency of the algorithm (Rousseeuw and van Driesen (1999)). Though, the efficiency of MCD is still unsatisfactory. This fact had brought Hadi (1992) to introduce the MMCD which ensured that the determinant of covariance matrix in every iteration is positive. Then, Hawkins (1994) offered an algorithm which is called the feasible solution algorithm (FSA) which ensured the optimal solution for MCD through a probabilistic approach. Next, Rousseeuw and van Driesen (1999) introduced an algorithm which is called the fast MCD (abbreviated as FMCD) which improves the performance of MCD. Almost in the same period, Billor et al. (2000) introduced the block adaptive computationally efficient outlier nominators (BACON) algorithm which improves the efficiency in time of computation. A couple years ago, Werner (2003) studied about MVE and MCD, he concluded that the FMCD in general is the best.

The labeling carried out by Pan et al. (2000) and Pena and Prieto (2001) gave a different direction. The approach used by Pan et al. (2000) is a projection along the axis generated by unit vectors, thus the results of projection spread out as uniform as they can. Meanwhile, Pena and Prieto (2001) proposed to separate the groups of suspected data among the ’good’ ones by using an orthogonal projection along $2p$ axis, where the first p orthogonal axis maximize the kurtosis and the second p orthogonal axis minimize the kurtosis.

The projection approximation method is not efficient compared to MVE and MCD, especially for large data of high dimension. Thus our attention will be focused to MCD, especially FMCD, because MCD has accepted more attention and good appreciation, as it has robust property of high breakdown point (BP). Eventhough, Werner (2003) showed that FMCD still takes longer time for large data of high dimension.

In this dissertation the author proposes a similar method to FMCD proposed by Rousseeuw and van Driesen (1999) with different criteria. Different with FMCD that uses the MCD criteria, the author proposes a criteria to minimize the vector variance (MVV). As a measure of multivariate dispersion, the vector variance was proposed by Suwanda and Djauhari (2002). This criteria will have a better efficiency than the FMCD, of the same effectivity.

I.2. Objective of the Research and Problem Formulation

The process of outlier identification consists of two stages, i.e. labeling and testing. The purpose of the labeling stage is to separate suspected data as outlier from the group of main data. Next, the purpose of testing stage is to find out whether the suspected data can be classified as outlier. The purpose of this research is to develop the procedure of outlier labeling which has robust property of high breakdown point (BP) and having high algorithm efficiency.

In this dissertation the author proposes the MVV criteria for labeling process. Suppose Σ is a covariance matrix of population where the data lies**.** The determinant of Σ, i.e. Det(Σ), and the sum of all diagonal elements of Σ^2, are two measures of multivariate dispersion. Det(Σ) is normally called a generalized variance or covariance determinant and $Tr\left(\Sigma^2\right)$, or the sum of all diagonal elements of Σ^2, is called a vector variance. Both measures of dispersion have their own advantages and weaknesses (Djauhari , 2005^b). As a measure of dispersion, the vector variance has much lower complexity level of time than $\text{Det}\left(\Sigma\right)$. Based on these facts, we hope that the using of MVV on the labeling process with the same effectivity level, will be more efficient than FMCD.

I.3. Literature Study

Studies about outlier have been a focus of many researchers for very long time, even according to Werner (2003), awareness on outlier occurrence had emerged since early XVI century, it was when Francis Bacon on 1620 wrote about the importance to know phenomenon of nature deviations.

A couple of researchers give various meaning to the outlier. For example, Grubbs (1969) defines an outlier to be an observation which seem to be clearly deviated among the others. Hawkins (1980) interprets an outlier as an observation which deviates quite away from the other observations so it gives a suspicion that the observation is generated by different mechanisms. Meanwhile Beckman and Cook (1983) interpret an outlier as data which is discordant to the researcher or contaminating data (contaminant), i.e. one come from distribution which is different with the distribution of the main group of data. Rousseuw and van Zomeren (1990) define an outlier to be contaminating data. Next, Barnett and Lewis (1984) define an outlier to be data which is inconsistent relative to the other group of data. In connection with modeling, Becker and Gather (1999) define an outlier to be observations which are away from the group of main data and possibly do not follow the assumed model. The study of outlier in this dissertation uses the definition given by Barnett and Lewis (1984).

Various procedures in identifying data which is considered to be 'inconsistent' are rapidly developing from time to time, both in the univariate case and in multivariate case. For instance, in the univariate case, Irwin (1925) proposed that the deviation of the mean as the criteria of outlier, Thomson (1935) developed Irwin's idea (1925) by proposing a new measuring tool, i.e. the ratio between the deviation from its mean and sample's standard deviation. The Statistics proposed by Thomson (1935) apparently has a very big impact to further development. Pearson and Chandra Sekar (1936) particularly conscientiously discussed that statistics, Dixon (1950) did an analysis about the extreme value for contamination data based on the statistics proposed by Thomson. Next, Grubbs (1950), Tietjen and Moore (1972) and Rosner (1975) built a measure to detect outlier based on the philosophy of Thomson's statistics. Grubbs (1950) proposed a statistic to test the largest or smallest data that is suspected as outliers. Tietjen and More (1972) developed Grubbs's research (1950) to test $k\ (k \geq 1)$ extreme data that deviated away from the group of the other $(n-k)$ data simultaneously through the gap, i.e. the distance between the $(n-k)^{\text{th}}$ extreme data and the $\left(n-k+1\right)^{\text{th}}$ extreme data. Next, Rosner (1975) introduced generalized extreme studentized deviation (GESD) which is a development of the idea of Tietjen and More (1972) to test several outliers simultaneously. For more general purposes, Rosner (1983) gave a table of critical values of GESD. In contrast to Beckman and Cook (1983) which figured out a direction of research for the univariate case, in Iglewicz and Hoaglin (1993) and Barnett and Lewis (1984) are brought a comparative study of various identification methods. Meanwhile in Kuwahara (1997) was proposed a history of the development and applications of outlier detection. The using of ESD is normally based on an approximation distribution. Having very good properties, ESD (Iglewicz and Hoaglin (1993)), Djauhari (1999) proposed the exact procedure. Five years ago, Djauhari (2001) perfected the ESD method. The exact critical points of ESD were given in Djauhari (2003) through beta inverse function.

In the multivariate case, say p-variate, problems encountered are not so simple as in the univariate case. For $p > 2$, different with the univariate case, a visual approach is more difficult to carry out. Researches in method of visualization for example was carried out by Shone and Fung (1987). Therefore, in the multivariate case, the analytical approach becomes a central approach. Here are a number of analytical approaches. Wilks (1963) introduced a method of test based on ratio of volume of a parallelotop. Because it is very difficult to define the critical points, Wilks only gave approximation value to the critical points for one and outliers. Besides that weakness, the Wilks's method has an advantage, i.e the candidates of outlier need not to group. Gnanadesikan and Kettenring (1972) detected several outliers consecutively through an analysis of principal components. They proposed a statistics test which is based on the Mahalanobis distance. The maximum value of statistics test is equivalent to the statistics test of single outlier of Wilks (1963). Rolfh (1975) tried to introduce a simultaneous testing of several outlier through the gap test. What is meant by a gap is the maximum distance between two groups of data which is measured by using the single linkage distance or the minimum spanning trees (MST). The use of MST here need to be looked at very carefully, due to possibility to be more than one MST. Djauhari (1996) gave a necessary and sufficient condition for the uniqueness of MST. In case that the MST is not unique, the affectivity of Rolfh’s method need to be investigated.

Basically, attempts to identify outliers in the multivariate case refer to the following philosophy. How to transform random vectors to be random variables so that candidates of outlier will be seen more clearly. This philosophy implicitly is used by Derquenne (1992). The most popular transformation is the Mahalanobis distance. This can be found in almost literatures of multivariate analysis, including in outlier studies. A very comprehensive book concerning outlier study is one written by Barnett and Lewis (1984).

On implementing the Mahalanobis distance, researchers are divided into two groups. The first one is ones who combine with the projection method, and the second one is ones who work directly in observation space without doing the projection. The purpose of the projection method is to find subspaces of low dimension, so that the data analysis is easier to carry out (Friedman (1987)). Some researchers who develop this method are Pan et al. (2000), Pena and Prieto (2001), and Hardin and Rocke (2004). They are classified to the second group, i.e. ones who use the Mahalanobis distance in observation rooms. But, this way is very sensitive to masking effects. To handle this problem, the method of robust estimator introduced by Huber (1964) is applicable as theoretical foundations of the construction of distance which is robust Mahalanobis.

Some researchers who did the robust distance on identifying outliers are Rousseeuw and van Zomeren (1990), Hadi (1992), Hawkins (1994), Becker and Gather (1999), Rousseeuw and van Driessen (1999), and Werner (2003). They proposed statistics test in form of robust Mahalanobis distance, by firstly finding the robust estimator of locations and covariance matrices. If the classic estimator is defined by involving the whole set of data, the robust estimator is built based on subsets consist of h data. The value of h is determined in such a way so that it is obtained an estimator of high BP estimator.

Various methods of robust estimation can be found in literatures. Rousseeuw (1985) introduced MVE and MCD methods with $h=\left[\frac{n}{2}+1\right]$ and n is the size of the sample. The notation $[z]$ here is the greatest integer but less than z. Rousseeuw and van Zomeren (1990) proposed the use of MVE to choose subsets having minimum volume of ellipsoid and covers at least h data. Hawkins (1994) introduced the feasible solution algorithm (FSA) to determine h data which give covariance matrices of minimum determinant. For the same sake, Rousseeuw and van Driessen (1999) proposed FMCD. The difference between the two methods lies in the process of determining h data. It was mentioned by Hardin and Rocke (2002), also Werner (2003) that FMCD is faster than MVE, MCD or even than FSA.

FMCD has a very impressive algortihm efficiency (Werner (2003)). But, according to the author, this thing happens only on multivariate data of low dimension. For large data of higher dimension, the efficiency of of the FMCD algorithm is worsening . This is due to computations of the determinant of covariance matrices, which takes time of order $O(2^p)$ by Cholesky's method. Here p is the number of variables. This has motivated the author to propose, in this dissertation, the use of the MVV criteria to estimate locations and covariances of robust property and high BP. It is clear that the trace computation of a matrix is much simpler than the computations of determinant. In contrast to the computation of determinant which takes time of order $O(2^p)$, the computation of trace only takes $O(p^2)$.

I.4. The Process of the Research

Eagerness to develop an outlier detection method of Pena and Prieto (2001) has initiated this research. They detect outliers through two stages, the outlier labeling and the outlier testing. Contrasted to their method which use the projection method that maximized and minimized the coefficient of kurtosis, for the same sake the author proposes a more efficient method, i.e the minimum spanning tree (MST) method. This method was inspired by one proposed by Rohlf (1975) and Djauhari (1996). Another result of this research is that the author improves the performances of critical points proposed by Pena and Prieto (2001). For the same sake, the author proposes the exact distribution, rather than the asymptotic distribution.

The method developed by the author in fact still has many weaknesses, particularly to masking effects. To reduce these weaknesses, the author has tried to develop a method in outlier labeling of robust property. Some robust approaches carried out to label outliers are MVE, MCD and FMCD. Study on the three methods, for small and medium size of data, concludes that MVE gives the longest time and FMCD is the most efficient method.

Next, the experiment is carried out to large data of high dimension. From the experiments it is obtained that FMCD still needs long enough time. Even for the case of dimension more than 100 and more than 1000 data, a computer of Pentium 3/1400 some times fail to compute the FMCD estimator. This fact has attracted the author to develop MVV.

Some properties of MVV will be further discussed on the next chapter. In the process of outlier labeling, MVV is more efficient than FMCD. From the results of simulation attached in Appendix E, it is concluded that MVV has the same effectivity level with FMCD. An open problem can be developed on further researches is the distribution of distance of robust property. This distribution will be used to test whether observations labeled as outliers are really outliers.

Chapter II Various Robust Mahalanobis Distance

II.1. Approaches in Outliers Identification

Identifying a multivariate outlier is not trivial as in the univariate case. Even, Rousseeuw and van Zomeren (1990) stated that it is not easy to do that when the number of variables p is larger than 2. In this case, a simple diagram such as scattering diagram is unable to figure out positions of every data in a p dimensional space. Further, a multivariate outlier need not to be an outlier on each variable involved, as seen in the illustration on Figure II.1 below.

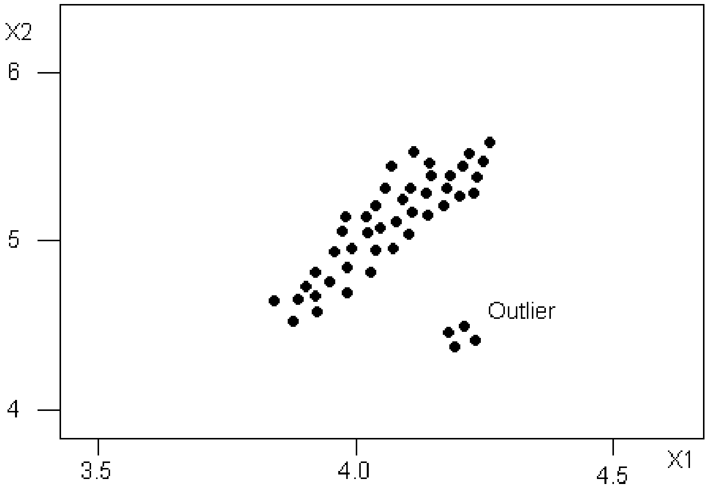


Figure II.1. An illustration of bivariate outlier phenomena

The same thing with the masking effect problem masking and swamping which frequently appears. Because of various complicated problems above , Gnanadesikan and Kettenring (1972) stressed that attempts in seeking procedures on outlier identification were fruitless. But, a good method must be specific and sensitive. Specific means that it is able to say that a 'good' data is really good, and sensitive means that it is able to say that a 'bad' data is really bad (Werner (2003)). The concept of sensitive developed more operational after Hampel et al. (1985) introduced the influential functions.

As it is discussed in the Preliminary chapter, the multivariate outliers identification is normally carried out by transforming random vectors to be random variables (Derquenne (1992)). The main tool is the Mahalanobis. See, for instance, Gnanadesikan and Kettenring (1972), Barnett and Lewis (1987), Pena and Prieto (2001), Werner (2003), and Djauhari (2004). Unfortunately, the distance is not suitable for groups of contaminated data. Therefore a Mahalanobis distance of robust property is very urgent to improve BP (see Lopuhaa and Rousseeuw (1991) and Becker and Gather (1999)).

Basically, there are three approaches on outlier identification. The first, is one based to distances including non-robust distance as stated by Derquenne (1992) and robust distance generated through MVE, MCD, MMCD, FSA, FMCD, and BACON. The usage of robust distance is to obtain location estimators and covariance matrices of robust property. Dealing with the robust estimator, Hampel (1974) introduced an estimator to both parameters based on the influencial function, Campbell (1980) estimates only the covariance matrix, Hampel et al. (1985) developed a robust estimator in more comprehensively than one introduced in Hampel (1974), and Woodruff and Rocke (1994) estimated a location parameter estimator and covariance matrix on matrix of large data. Dealing with MVE, Serfling (1980) gave a deep discussion about volume of ellipsoids, Hawkins (1993) and Grambow and Stromberg (1998) gave an algorithm, and and Werner (2003) gave a performance analysis of MVE. Further, dealing with MCD, Croux and Haesbroeck (1999) studied the efficiency of MCD, Rousseeuw and van Driessen (1999) gave the FMCD algorithm. Werner (2003) showed that in general, FMCD is better than MVE.

The second, is an approach based on labeling such as proposed by Rolfh (1975) by using MST. An efficient algorithm to determine MST was suggested by Djauhari (1996). Kitagawa (1979), Rocke and Woodruff (2000) and Rocke (2002) introduced labeling by using method of data grouping. Another labeling approach was proposed by Becker and Gather (1999) by defining the 'outlier area', Pan et al. (2000) and Pena and Prieto (2001) through the projection pursuit.

The third, is a non distance approach such as proposed by Wilks (1963). This method was then developed by Caroni and Prescott (1992) to test several outliers by sequentally using the statistics of Wilks. Another nondistance approach was introduced by Cleroux et al. (1986) and Lazraq and Cleroux (1989) which identify an outlier based on the RV coefficients, Shone and Fung (1987) who identify the candidates of outlier through graphic, and Viljoen and Venter (1999) who improve the performance of the method of Caroni and Prescott (1992) by using MCD.

Taking into account that FMCD is having very good properties in effectiveness on one side (Werner (2003)) and on the other side having low efficiency for data matrix of high dimension, the focus of this research are:

1. Development of the criteria of location estimator and covariance matrix of robust property.
2. Outlier labeling based on the Mahalanobis robust which is defined based on the estimator on the point 1 above. In the following sub chapter will be proposed various Mahalanobis distance approaches based on MVE and MCD.

II.2. Robust Mahalanobis Distance

Suppose $\vec{X}_1, \vec{X}_2, \cdots, \vec{X}_n$ are random sample of size n having $N_p(\vec{\mu}, \Sigma)$ where Σ is of positive definite. The vector of sample mean $\bar{\vec{X}}$ and sample covariance matrix **S** is,

$$\bar{\vec{X}} = \frac{1}{n}\sum_{i=1}^{n}\vec{X}_i \text{ dan } \mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\vec{X}_i - \bar{\vec{X}}\right)\left(\vec{X}_i - \bar{\vec{X}}\right)^t$$

The distance $d_{\mathrm{S}}\left(\vec{X}_i, \bar{\vec{X}}\right)$, where $d_{\mathrm{S}}^2\left(\vec{X}_i, \bar{\vec{X}}\right) = \left(\vec{X}_i - \bar{\vec{X}}\right)^t \mathbf{S}^{-1}\left(\vec{X}_i - \bar{\vec{X}}\right)$, is called the Mahalanobis distance of $\vec{X}_i$ to $\bar{\vec{X}}$.

Eventhough the Mahalanobis distance is very wellknown in practice, but it is not robust. Occurance of one or more outliers can significantly change the value $d_{\mathrm{S}}\left(\vec{X}_i, \bar{\vec{X}}\right)$. This happens, because $\bar{\vec{X}}$ and $\mathbf{S}$ as estimators of $\vec{\mu}$ and Σ are not robust estimators. Therefore the Mahalanobis distance $d_{\mathrm{S}}\left(\vec{X}_i, \bar{\vec{X}}\right)$ is also not robust. Hence, its usage in identifying outlier*s* is sensitive to the masking effect, and also probably the swamping effect.

Barnett and Lewis (1984, p. 114) says that a masking effect is an effect which causes outliers are undetectable due to covered by another outliers. A swamping effect is the converse, i.e. non outlier data detected as outlier. Masking effects are frequently found in the process of one by one identification of several outliers.

II.2.1. The Notion of the Robust Statistics

Since assumptions of normality, linearity and independence stick on the classic estimation methods frequently are not satisfied, Huber (1964) introduced the robust estimator.. One of the goal, as it was stated by Hampel et al. (1985), is to identify the deviation of data, or outlier. Compared to the classic methods, the robust statistics will give a clearer variability description between an outlier and ’good data‘, the classic statistics will vaguely the difference. Dealing with robustness of a statistics, some researchers give similar definitions, eventhough using different context, i.e., as an insensitivity to a small deviation of assumption (see Huber (1980, p. 1), Hoaglin, Mosteller and Tukey (1983, p. 2) and Hampel et al. (1985, p. 6). Measures of robustness are normally stated by the breakdown point (BP).

II.2.2 Affine Equivariant Property

The affine equivariant is very good property of an estimator, because it is not influenced by affine transformation. Consider random samples $\vec{X}_1, \vec{X}_2, \cdots, \vec{X}_n$ of random vectors $\vec{X}$ of location parameter $T \in \Re^p$ and scale parameter C in the space of $p \times p$ symmetric matrices. Suppose X defines an $n \times p$ matrix where the k-th row is $\vec{X}_k^t$. A location estimator $T_n(\mathrm{X}) \in \Re^p$ is said to have the affine equivariant property if for every vector $\vec{b} \in \Re^p$ and every nonsingular $p \times p$ matrix the condition

$$T_n\left(\mathrm{AX} + \vec{b}\right) = \mathrm{A}T_n(\mathrm{X}) + \vec{b},$$

holds (Rousseuw, 1985).

An estimator of scale $C_n(\mathrm{X})$, which is in form of an $p \times p$ matrix, symmetric and positive definite is affine equivariant if for every vector $\vec{b} \in \Re^p$ end every $p \times p$ non singular matrix A the following condition holds

$$C_n\left(\mathrm{AX} + \vec{b}\right) = \mathrm{A}\,C_n(X)\,\mathrm{A}^t.$$

It can be seen that when an estimator is having the affine equivariant property, it will not get influenced by an affine transformation. This good property will be a condition in searching of robust statistics.

The location estimator and scale estimator of maximum likelihood method,

$$T_n(\mathrm{X}) = \frac{1}{n}\sum_{i=1}^{n}\vec{X}_i \quad \text{and} \quad C_n(\mathrm{X}) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\vec{X}_i - T(X)\right)\left(\vec{X}_i - T(X)\right)^t$$

is affine equivariant but it is not robust, because occurrence of an outlier (even only one) is able to shift $T_n\left(X\right)$ far enough. In general, Rousseeuw (1985) says that M-estimator, which is a generalization of the maximum likelihood estimator, for multivariate data are mostly affine equivariant but is of small BP, i.e. at most $\frac{1}{p+1}$.

II.2.3. Two Basic Concepts of Robust Estimation

1. Breakdown Point

A quantitative measure to describe the concept of robustness is breakdown point (BP). This measures how many data can be changed to be infinity before they meaningless crushed to bits. Several researchers such as Hampel et al. (1985 p. 41), Huber (1980, p. 13), Rousseeuw (1985), Kotz and Johnson (1985 p.158), and Rousseeuw and Leroy (1987 p.10) gave interpretations of BP both from the context of population and from the context of sample. This dissertation refers to the interpretation given by Kotz and Johnson (1985 h.158) and Rousseeuw and Leroy (1987) from the context of sample. They define BP to be the smallest fraction of data which causes the value of estimator to be infinity when the value of all data in the fraction are changed to be infinity. Applying this definition, it is clear that in the univariate case, the median has BP = 0,5 and the mean of sample has BP = $\frac{1}{n}$.

The concept of BP is highly related to the concept of estimator bias. Concerning with the bias effect, Franklin and Brodeur (2005) say that the purpose of the robust estimation is to produce an estimator which is free of influence of occurrence of outliers by lessening the bias. The relation between BP and the value of bias will be discussed in the following paragraph.

Consider $T_n\left(\mathrm{X}\right)$ and $C_n\left(\mathrm{X}\right)$ on II.2.3. Suppose the estimator $T_n\left(\mathrm{X}\right)$ becomes $T_n\left(\mathrm{X}^*\right)$ if the value of m data are changed. Rousseeuw and Leroy (1987) define BP, for sample of size as follows

$$bias\left(m, T, \vec{X}\right) = \sup_{\mathrm{X}^*} \left\| \ T_n\left(\mathrm{X}^*\right) - T_n\left(\mathrm{X}\right) \right\|$$

which measures the greatest difference $T_n\left(\mathrm{X}^*\right)$ and $T_n\left(\mathrm{X}\right)$. Rousseeuw (1985) defines BP as follow ,

$$\varepsilon_n^*\left(T, \vec{X}\right) = \ \min\left\{\frac{m}{n} \middle| \ bias\left(m, T, \vec{X}\right) \text{infinite}\right\}.$$

Suppose the m data which the value are changed to be infinity imply that $bias\left(m, T, \vec{X}\right)$ is infinite. If the value of (m – 1) data among them are changed to be infinity do not imply $bias\left(m, T, \vec{X}\right)$ to be infinite, then BP = $\frac{m}{n}$. In the univariate case, the value of BP for some location estimators mentioned Hampel et al. (1985) the least is of the sample mean $\overline{X}$, i.e. $\frac{1}{n}$, and the greatest is of the median, i.e. 0.5. The BP value of the kurtosis and the studentized range are respectively 0.21 and 0.043. In the multivariate case, the vector $\vec{\overline{X}}$ of sample mean is having BP = $\varepsilon_n^*\left(\vec{\mu}, \vec{X}\right) = \frac{1}{n}$. Some literatures say that an estimator which is assumed to be good is one of BP ≥ 0.25 .

2. Influencial functions

Beside the concept of BP, another important concept which is used in seeking robust estimator is the concept of influential function, abbreviated as IF. The role of IF is to measure the magnitude of influence of disturbances on the estimator caused by existence of very small change on value of data. Hampel et al. (1985) introduced the concept of influential function as follow .

Suppose $X_1, X_2, \cdots, X_n$ are random sample of random variable X of distribution function *F*. If F_n is the function of empiric distribution and Δ_x is the degenerate distribution in x,

$$\Delta_x(t) = \begin{cases} 1\,, t = x \\ 0, t \neq x \end{cases}$$

then F_n can be written as $F_n = \frac{1}{n}\sum_{i=1}^{n} \Delta_{x_i}$ where $x_1, x_2, \cdots, x_n$ are realization $X_1, X_2, \cdots, X_n$. Consider the statistics,

$$T_n = T_n(\mathrm{X}) = T_n(F_n)$$

and the sequence of statistics $\{T_n, n \geq 1\}$ for every possibility of sample n. Here $T_n(F_n)$ is an estimator of a parameter on the distribution function F. Suppose the estimator $T_n(F_n)$ in the form of functional. This means that $T_n(F_n) = T(F_n)$ for every n and F_n where *T* is a functional where the domain in the set of all distributions in which *T* is defined so that ,

$$T_n(F_n) \xrightarrow{p} T(F)$$

converges in probability for $n \longrightarrow \infty$. Here $T(F)$ is the asymptotic value of the sequence of estimators $\{T_n, n \geq 1\}$. Under this assumption, IF of *T* on *F* is defined as,

$$\mathrm{IF}(x, T, F) = \lim_{\varepsilon \to 0} \frac{T((1-\varepsilon)F + \varepsilon\,\Delta_x) - T(F)}{\varepsilon}$$

provided the limit exists. For example, if F is replaced with F_{n-1} and we take $\varepsilon = \frac{1}{n}$,

then for $n \longrightarrow \infty$ it follows that,

$$\mathrm{IF}(x, T, F_{n-1}) \longrightarrow nT\left\{\left(\left(1-\frac{1}{n}\right)F_{n-1} + \frac{1}{n}\,\Delta_x\right) - T(F_{n-1})\right\}.$$

Therefore, IF measures (through approximation) *n* times of changes in the value of *T* which is caused by an addition of an observation *x* on a large sample of size (n–1).

In the case that x is an outlier, IF explains the influence of contamination of x in defining the estimator $T\left(F_{n-1}\right)$.

II.3. Some Robust Mahalanobis Distances

II.3.1 Minimum Volume Ellipsoid (MVE)

The minimum volume ellipsoid (MVE) method was introduced by Rousseeuw (1985) to estimate location parameters and covariance matrices. The concept of MVE was explained more clearly by Rousseeuw and Leroy (1987 p.258) as an attempt to determine the location estimator and covariance matrix based on $h=\left[\frac{n}{2}+1\right]$ data which give the minimum volume of ellipsoid among all of the sets of *h* possible data. Based on these *h* data, then it was carried out an estimation to the parameters. This estimator is then used to generate the robust Mahalanobis distance. In the development, the value *h* has not given a satisfied result yet. Next, Rousseeuw and van Zomeren (1990) showed that the optimal value of h is $h=\left[\frac{n+p+1}{2}\right]$. This is the value that is used until now.

Suppose $\vec{X}_1, \vec{X}_2, \cdots, \vec{X}_n$ are random samples of size n picked up from a *p*-variate distribution of location parameter $\vec{\mu}$ and positive definite covariance matrix Σ. The estimator MVE for the pair $\left(\vec{\mu}, \Sigma\right)$ is the pair $\left(T_{MVE}, C_{MVE}\right)$ which gives,

$$\text{Card}\left\{i \mid \left(\vec{X}_i - T_{MVE}\right)^t C_{MVE}^{-1}\left(\vec{X}_i - T_{MVE}\right) \le a^2\right\} \ge h$$

with $h = \left[\frac{n+p+1}{2} \right]$ and constant $a^2 = \chi^2_{p;0.5}$, i.e. the median of *chi-squared* distribution of degree of freedom p. The estimator MVE is an affine equivariant estimator, and of high BP i.e. $\frac{n-2(p-1)}{2n}$. See Rousseeuw and Leroy (1987).

Based on the estimator MVE the robust Mahalanobis distance of $\vec{X}_i$ with respect to T_{MVE}, written $dR_{MVE}\left(\vec{X}_i, T_{MVE}\right)$, is defined through the quadratic form as ,

$$dR^2_{MVE}\left(\vec{X}_i, T_{MVE}\right) = \left(\vec{X}_i - T_{MVE}\right)^t C^{-1}_{MVE}\left(\vec{X}_i - T_{MVE}\right)$$

The good property of MVE determined by the robust property of high BP, apparently does not guarantee the popularity. This is because of the algorithm efficiency which is not high (Werner (2003)), especially for large size data of high dimension. Therefore, in this dissertation the MVE will not be discussed too far.

II.3.2 Minimum Covariance Determinant (MCD)

Together with MVE, Rousseeuw (1985) also introduced the minimum covariance determinant (MCD) method. The purpose of both method are the same. The difference is only on the criteria they used. Contrasted to MVE which uses the minimizing volume of the ellipsoid criteria based on $h = \left[\frac{n+p+1}{2} \right]$ data, MCD uses the minimizing determinant of the covariance matrix criteria based on the h data. Just like MVE, the estimator MCD also is of affine equivariant property of the same BP, i.e. $\frac{n-2(p-1)}{2n}$. See Rousseeuw and Leroy (1987).

There are many proposed algorithms to determine the MCD estimator, for instance are by Hadi (1992), Hawkins (1992), Hawkins and Olive (1997), Rousseeuw and van

Driessen (1999), and Billor et al. (2000). Hadi (1992) introduced the MMCD algorithm, a modified MCD, which ensures that in every iteration, the determinant of covariance matrix is positive. Hawkins (1992) introduced the feasible solution algorithm (FSA) and Rousseeuw and van Driessen (1999) introduced the fast minimum covariance determinant (FMCD). Next, Billor et al. (2000) proposed the BACON algorithm. This number of proposed algorithm shows that the appreciation of researchers to MCD is very positive. This reason has attracted the author to focus on the development.

Both FSA and FMCD work on set consists of h data, but as mentioned in Hardin and Rocke (2002) and Werner (2003) that MCD has faster time process. The principal difference lies on the process of selection of data which are going to be entered to *h* sets of data. Contrasted to FSA which allows only single data to get in or to get out the set, FMCD allows simultaneously several data to get in and to get out. The difference between FMCD and MCD is on one that Rousseeuw and van Driessen (1999) called as the *C*- step algorithm, as we will figure out in the following.

Just like in MVE, suppose $\vec{X}_1, \vec{X}_2, \cdots, \vec{X}_n$ are random samples of size *n* picked up from a *p*-variate distribution having location parameter $\vec{\mu}$ and positive definite covariance matrix Σ. The MCD estimator for the pair $(\vec{\mu}, \Sigma)$ is the pair (T_{MCD}, C_{MCD}) where,

$$T_{MCD} = \frac{1}{h} \sum\nolimits_{i \in H} \vec{X}_i$$

$$C_{MCD} = \frac{1}{h} \sum\nolimits_{i \in H} \left(\vec{X}_i - T_{MCD} \right) \left(\vec{X}_i - T_{MCD} \right)^{\mathrm{t}}$$

and the determinant C_{MCD} is minimum among all possible $h=\left[\frac{n+p+1}{2}\right]$ sets *H*. The C-step algorithm proposed by Rousseeuw and van Driessen (1999) is as follow form an arbitrary set H_{old} consists of $h=\left[\frac{n+p+1}{2}\right]$ data.

1. Compute the mean vector $\bar{\vec{X}}_{H_{old}}$ and covariance matrix $S_{H_{old}}$ of all available data in H_{old}. Then, for *i* = 1, 2, … , *n* , compute

   $d^2_{H_{old}}(i) = d^2_{H_{old}}\left(\vec{X}_i, \bar{\vec{X}}_{H_{old}}\right) = \left(\vec{X}_i - \bar{\vec{X}}_{H_{old}}\right)^t S^{-1}_{H_{old}} \left(\vec{X}_i - \bar{\vec{X}}_{H_{old}}\right).$
2. Sort the results of computations, from the smallest to the greatest. This order gives a permutation π on the observations index. Suppose the result in order is $d^2_{H_{old}}(\pi_1) \le d^2_{H_{old}}(\pi_2) \le \cdots \le d^2_{H_{old}}(\pi_n)$.
3. Form a set H_{new} consists of h observations of index $\pi_1, \pi_2, \cdots, \pi_3$.
4. Compute $\bar{\vec{X}}_{H_{new}}$, $S_{H_{new}}$ and $d^2_{H_{new}}\left(\vec{X}_i, \bar{\vec{X}}_{H_{new}}\right)$ as on the item 2.
5. If $\mathrm{Det}\left(S_{H_{new}}\right)=0$, repeat step $1-5$. If $\mathrm{Det}\left(S_{H_{new}}\right)=\mathrm{Det}\left(S_{H_{old}}\right)$, the process is done. If $\mathrm{Det}\left(S_{H_{new}}\right)<\mathrm{Det}\left(S_{H_{old}}\right)$, the process is resumed until the *k*-th iteration when $\mathrm{Det}\left(S_{H_{new}}\right)=\mathrm{Det}\left(S_{H_{old}}\right)$.
6. Suppose S_{H_i} is the covariance matrix got from the *i*-th iteration. At the end of the *k*- iteration we get

   $$\mathrm{Det}\left(S_{H_1}\right) \ge \mathrm{Det}\left(S_{H_2}\right) \ge \ldots \ge \mathrm{Det}\left(S_{H_{k-1}}\right) = \mathrm{Det}\left(S_{H_k}\right).$$

Suppose T_{MCD} and C_{MCD} state the MCD estimator for the location parameter and covariance matrix. Therefore, $T_{MCD} = \bar{\vec{X}}_{H_{new}}$ and $C_{MCD} = S_{H_{new}}$ on the *k*-th iteration. The robust Mahalanobis distance between $\vec{X}_i$ and T_{MCD} based on MCD, is written $dR_{MCD}\left(\vec{X}_i, T_{MCD}\right)$, and is defined on the quadratic form as

$$dR^2_{MCD}\left(\vec{X}_i, T_{MCD}\right) = \left(\vec{X}_i - T_{MCD}\right)^t C^{-1}_{MCD}\left(\vec{X}_i - T_{MCD}\right)$$

for i = 1, 2, … , n. Data which gives the greatest value of $dR_{MCD}\left(\vec{X}_i, T_{MCD}\right)$ will be labeled as an outlier (labeled outlier) and is considered as a candidate of an outlier.

Eventhough on the labeling process that FMCD is much better than MVE, but it is still not practical for large data of high dimension (Werner (2003)). This phenomenon is one that motivated the author to propose, on the next chapter, the use of more efficient criteria of the same effectivity.

II.3.3 BACON

BACON is the short of blocked adaptive computationally efficient outlier nominator, which was proposed by Billor et al. (2000). BACON is a method of fast robust property in identifying the set of *'basic'* data considered free of an outlier. The BACON algorithm for the multivariate case consists of the early stage and implementation stage is described as follow.

1. Early Stage

On the early stage we form a set of *'basic'* data. Billor et al.(2000) gave the following two choices.

1. Compute the square of the classic Mahalanobis $d^2_S\left(i\right) = d^2_S\left(\vec{X}_i, \bar{\vec{X}}\right)$ for $i = 1, 2, \cdots, n$. Next, sort from the smallest to the greatest. This order defines a permutation π on the observation index. Suppose the result of the sorting is

$d_{\mathrm{S}}^2(\pi_1) \le d_{\mathrm{S}}^2(\pi_2) \le \cdots \le d_{\mathrm{S}}^2(\pi_n)$. Form a set consists of $m = cp$ observations of index $\pi(1), \pi(2), \cdots, \pi(m)$. This set is one that is used as the *'basic'* set on the early stage. Suppose $\vec{\tilde{X}}$ is a vector of dimension *p* where the *k*-th component is the sample's median of the *k*-th variable (coordinatewise median). Compute $d(i) = \left\| \vec{X}_i - \vec{\tilde{X}} \right\|$ for $i = 1, 2, \cdots, n$. Next sort from the smallest to the greatest. This order defines a permutation π on the observation index. Suppose the result of the sorting is $d((\pi_1)) \le d((\pi_2)) \le \cdots \le d((\pi_n))$. Form a set consists of $m = cp$ observations of index $\pi(1), \pi(2), \cdots, \pi(m)$. This set is one that is used as the set of *'basic'* on the early stage.

For the two choices, Billor et al. (2000) and Werner (2003) suggested c = 4 or 5. Werner (2003) says that the two choices are better, because is more effective in identifying many outliers.

2. Implementation Stage

1. Based on the set of 'basic' data in the early stage, compute the sample mean T_{BACON} and the sample covariance matrix C_{BACON} .
2. Compute $d_{BACON}(i) = d_{C_{BACON}}\left(\vec{X}_i , T_{BACON}\right) = \sqrt{\left(\vec{X}_i - T_{BACON}\right)^t C_{BACON}^{-1} \left(\vec{X}_i - T_{BACON}\right)}$, i.e. the Mahalanobis distance of $\vec{X}_i$ to T_{BACON} based on the BACON method.
3. Sort the result from the smallest to the greatest. This order gives a permutation π on the observation index. Suppose the result of the sort is $d_{BACON}(\pi_1) \le d_{BACON}(\pi_2) \le \cdots \le d_{BACON}(\pi_n)$.
4. Form a new 'basic' set consists of r observations of index $\pi(1), \pi(2), \cdots, \pi(r)$ where $d_{BACON}(\pi_r) \le c_{npr} \sqrt{\chi^2_{\left(1-\frac{\alpha}{n}\right), p}}$ and $d_{BACON}(\pi_{r+1}) > c_{npr} \sqrt{\chi^2_{\left(1-\frac{\alpha}{n}\right), p}}$. The

constant c_{npr} is a correction factor with $c_{npr} = c_{np} + c_{hr}$ where $h = \left[\frac{n+p+1}{2} \right]$,

r is the number of observations on the new ’basic’ set

$$c_{np} = 1 + \frac{p+1}{n-p} + \frac{2}{n-1-3p} ,$$

and $c_{hr} = \max \left\{ 0 \, , \frac{h-r}{h+r} \right\}$.

5. Repeat the step 2 and 3 so that the number of observations in the ’basic’ set remains unchanged.
6. The last data outside the ’basic’ set is considered as a candidate of an outlier or a labeled outlier.

Chapter III The Proposed Method

III. 1. Motivation

The identification process of anomalous data or multivariate outlier is a complicated process. Basically, there are two main problems to tackle. The first is, the efficiency of outlier labeling algorithm, and the second is the hypothesis testing (Angiulli and Pizzuti (2005)). Particularly, in case of high dimensional large data such as in the data mining or intrusion detection (ID), the algorithm efficiency is the first priority to handle (Werner (2003)).

As it has already brought on the first two chapters, the main focus of this dissertation is the development of criteria on the C-step (FMCD) algorithm. The background and things motivated the focus are facts that :

1. FMCD is having good properties. It is robust of high BP and gives affine equivariant location estimator and covariance matrix.
2. FMCD is having weaknesses. The efficiency of the algorithm is getting lower when the dimension of data is going higher.

On this chapter the author proposes a modification of C-step by using new criteria. In contrast to C-step (FMCD) which uses minimization criteria of determinant of the covariance matrix (covariance determinant abbreviated as CD), the author proposes to modify the C-step method by square of covariance matrix minimization criteria. The latter criteria, in literatures is known as the vector variance, and is abbreviated as VV. CD and VV are two measures of multivariate dispersion of their advantages and weaknesses. Two advantages of VV (Djauhari (2005^{b})) are:

1. Able to measure multivariate dispersions, although the covariance matrix is singular.

2. Its computation process is very efficient because it is only the sum of square. First, squaring every element of covariance matrix, and then add them up.

By taking those advantages the author modifies C-step by virtue of a criteria that the author calls the minimum vector variance (MVV). Just like MCD, MMCD, FMCD, and BACON, the MVV method is also of purpose to determine the robust estimator for the location parameter T_{MVV} and covariance matrix C_{MVV} based on the set of $h = \left[\frac{n+p+1}{2}\right]$ data by means minimizing $Trace\left(C_{MVV}^2\right)$. Werner (2003) showed that MVE and MCD have the same BP, i.e. $\frac{n-2(p-1)}{2n}$. On the last section of this chapter the author will explain that MVV is also having the same BP with MVE and MCD.

III. 2. Vector Variances (VV)

There are two famous measures of dispersions in the study of multivariate, the total variance (abbreviated as TV) and the determinant of covariance (covariance determinant abbreviated as CD). Suppose $\vec{X}$ is a random vector of covariance matrix Σ. Then TV = $Tr(\Sigma)$ while CD = $|\Sigma|$. CD has a much more general use than TV, including its use in various robust method proposed in III.1. Therefore, if $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ are eigen values of Σ of size $(p \times p)$, then TV = $Tr(\Sigma)$ = $\lambda_1 + \lambda_2 + \cdots + \lambda_p$ and CD = $|\Sigma|$ = $\lambda_1\ \lambda_2 \cdots \lambda_p$. Concerning the role of TV and CD in measuring the spread of multivariate data, Pena and Rodriguez (2003) gave a very comprehensive discussion.

The role of TV generally can be found on the reduction problem of data dimension such as in the principal component analysis (Anderson (1984), Jolliffe (1986) and

Johnson and Wichern (1988)), analysis of discriminant (Anderson (1984) and Johnson and Wichern (1988)), canonic analysis (Anderson (1984)). Meanwhile the role of CD can be found in every literature of multivariate analysis. Particularly, the role in multivariate dispersion monitoring can be found, for example, in Kotz and Johnson (1985), Alt and Smith (1988), Montgomery (2001) and Djauhari (2005[a]) and related references.

Lack of TV's role is understandable, because TV involves the variance only without involving the structure of covariance. Thus it is simply involving the diagonal elements of the covariance matrices, meanwhile CD involves both the matrix structure and the covariance. This reason makes CD has a wider role in applications (Djauhari (2005[a])).

Although CD has a wider applications than TV, but it is not coming without lack. The main lack lies on the property of having CD = 0 when there is a variable of variance 0 or when there is a variable which is a linear combination of any other variables. In fact, that CD = 0 is not certainly implies that $\vec{X}$ is of degenerate distribution in the vector $\vec{\mu}$. There is probably a subspace of low dimension where $\vec{X}$ is of non degenerate distribution. In the context of sample, that CD = 0 shows that there is low dimension subspaces where data spreads around the mean vector. Because of this lack, the author proposes another measure of multivariate dispersion, which is about to show in the following paragraphs.

Suppose $\vec{X}$ and $\vec{Y}$ are two random vectors of arbitrary finite dimension having joint covariance matrix

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

where Σ_{11} and Σ_{22} are respectively the covariance matrix of $\vec{X}$ and $\vec{Y}$, and Σ_{12} = Σ_{21}^{t} is the covariance matrix between $\vec{X}$ and $\vec{Y}$. Lazraq and Cleroux (1989) define the measure of correlation between the two random vectors $\vec{X}$ and $\vec{Y}$ as follow.

$$\rho_V\left(\vec{X}\,,\vec{Y}\right)=\frac{Tr\left(\Sigma_{12}\,\Sigma_{21}\right)}{\sqrt{Tr\left(\Sigma_{11}^{2}\right)Tr\left(\Sigma_{22}^{2}\right)}}\,.$$

In line with this definition, the author uses $Tr\left(\Sigma_{11}^{2}\right)$ and $Tr\left(\Sigma_{22}^{2}\right)$ respectively as measures of random vector variance $\vec{X}$ and $\vec{Y}$ which is later called as the vector variance (VV).

In general, if random vector $\vec{X}$ has covariance matrix Σ, then VV of $\vec{X}$, ie. $Tr\left(\Sigma^{2}\right)$, measures the spread of multivariate data around $\vec{\mu}$. See Suwanda and Djauhari (2003). This measure has different properties with CD, but complete each other. VV = 0 is exactly shows that $\vec{X}$ is of degenerate distribution in $\vec{\mu}$ (see Appendix A). Another good properties of VV are:

1. Different with CD which requires a condition that the covariance matrix must be non singular, VV does not.
2. The computation of VV is very efficient. Different with CD which uses Cholesky's decomposition of order $O\left(2^{p}\right)$, VV is of order $O\left(p^{2}\right)$. For p = 100, as an example, CD is of order O(1.26765E+30) meanwhile VV is of order O(1.0E+4). This is an advantage which is very significant.

The second property is one of the reasons of why VV is exploited in this dissertation.

III.3. The MVV Criteria and Modification of C-step

III. 3. 1. The MVV Criteria

Recall that MVE and MCD use the minimization of ellipsoid's volume criteria and minimization of the determinant of covariance matrix to determine the location estimator and covariance matrix. But, in this dissertation the author proposes to use the minimization of vector variance (MVV) criteria. Consider a data set $X = \left\{\vec{X}_1, \vec{X}_2, \cdots, \vec{X}_n\right\}$ of $p-$variate observations and let $H \subseteq X$. Suppose T_{MVV} and C_{MVV} are MVV estimator for the location parameter and covariance matrix. This two estimators are determined based on the set *H* consists of $h = \left[\frac{n+p+1}{2}\right]$ data which give covariance matrix C_{MVV} of minimum $Tr\left(C_{MVV}^2\right)$ among all possible sets of *h* data. Therefore,

$$T_{MVV} = \frac{1}{h}\sum_{i\in H}\vec{X}_i$$

$$C_{MVV} = \frac{1}{h}\sum_{i\in H}\left(\vec{X}_i - T_{MVV}\right)\left(\vec{X}_i - T_{MVV}\right)^{\mathrm{t}}$$

Like T_{MVE} and C_{MVE}, T_{MCD} and C_{MCD} are, T_{MVV} and C_{MVV} are also of affine equivariant property and of the same BP i.e. $\frac{n-2(p-1)}{2n}$. The affine equivariant property of MVV is guaranteed, because,

1. $T_{MVV}\left(\mathrm{AX}+\vec{\mathrm{b}}\right) = \frac{1}{h}\sum_{i\in H} \left(\mathrm{A}\vec{X}_i+\vec{b}\right) = \mathrm{A}\left(\frac{1}{h}\sum_{i\in H}\vec{X}_i\right)+\vec{b}$

$$= \mathrm{A}T_{MVV}+\vec{b}$$

2. $C_{MVV}\left(\mathrm{AX}+\vec{b}\right) = \frac{1}{h}\sum_{i\in H}\left(\mathrm{A}\vec{X}_i+\vec{b}-\mathrm{A}T_{MVV}-\vec{b}\right)\left(\mathrm{A}\vec{X}_i+\vec{b}-\mathrm{A}T_{MVV}-\vec{b}\right)^{\mathrm{t}}$

$$= \frac{1}{h}\sum_{i\in H}\left(\mathrm{A}\vec{X}_i-\mathrm{A}T_{MVV}\right)\left(\mathrm{A}\vec{X}_i-\mathrm{A}T_{MVV}\right)^{\mathrm{t}}$$

$$= \mathrm{A}C_{MVV}\mathrm{A}^{\mathrm{t}}$$

Next, that BP of the MVV estimator is the same value with BP of MVE and MCD estimators, i.e. $\frac{n-2(p-1)}{2n}$, can be explained as follow.

Suppose that,

$$\left(\vec{X}_i-T\right)^t C^{-1}\left(\vec{X}_i-T\right)=d^2$$

is an arbitrary ellipsoid. Let $\lambda_1,\lambda_2,\ldots,\lambda_p$ be eigen values of C_{MCD} and λ_{*1}, λ_{*2}, … , λ_{*p} be eigen values of C_{MVV}. Then the value of VE, CD and VV obtained based on n MVE, MCD and MVV respectively are,

$$\mathrm{VE} = \frac{d^p\sqrt{\pi^p}}{\Gamma\left(\frac{p}{2}+1\right)}\sqrt{\left|C_{MVE}\right|} = \frac{d^p\sqrt{\pi^p}}{\Gamma\left(\frac{p}{2}+1\right)}\sqrt{\left|C_{MCD}\right|} \text{ (see Serfling (1980))}$$

$$= \frac{d^p\sqrt{\pi^p}}{\Gamma\left(\frac{p}{2}+1\right)}\sqrt{\lambda_1.\lambda_2\ldots\lambda_p}\ .$$

$$\mathrm{CD} = \left|C_{MCD}\right| = \lambda_1.\lambda_2\ldots\lambda_p\,.$$

$$\mathrm{VV} = Tr\left(C_{MVV}^2\right) = \lambda_1^2 + \lambda_2^2 + \ldots + \lambda_p^2\,.$$

This means that,

1. VE is a multiple product of standard deviations of all principal components of C_{MCD}.
2. CD equals to multiple product of variances of all principal components of C_{MCD}.
3. VV is a quadratic sum of all variances of all principal components of C_{MVV}.

Taking into account the eigen values, those of C_{MCD} and of C_{MVV}, it is clear that BP of MVV are equal with BP of MVE and MCD, i.e. $\frac{n-h}{n} = \frac{n-2(p-1)}{2n}$. The eigen values will be finite when one of their components of at most $\frac{n-2(p-1)}{2n}$ data is changed to infinity. The eigen values turn to *breakdown* (the value becomes infinity) when one of its component of $\left[\frac{n-2(p-1)}{2n}+1\right]$ vector data are changed to be infinity. The simulation result will be given in Appendix E.

Taking into account the advantages of VV above, in the following discussion, the author presents a modification of the C-step algorithm. The modification is on the use of criteria. Contrasted to C-step (FMCD) which uses a minimization of the covariance matrix determinant criteria, on the modified C-step, it is used the minimization of variance vector criteria.

III. 3. 2. Modified C-step

The MVV algorithm is a modification of the C-step algorithm, precisely it is as follow:

1. Form an arbitrary set H_{old} consists of $h = \left[\frac{n+p+1}{2}\right]$ data.

2. Compute mean vector $\vec{\bar{\bar{X}}}_{H_{old}}$ and covariance matrix $S_{H_{old}}$ of all data in H_{old}. Next, for $i = 1, 2, \dots, n$, compute $d^2_{H_{old}}(i) = d^2_{H_{old}}\left(\vec{X}_i, \vec{\bar{\bar{X}}}_{H_{old}}\right) = \left(\vec{X}_i - \vec{\bar{\bar{X}}}_{H_{old}}\right)^t S^{-1}_{H_{old}}\left(\vec{X}_i - \vec{\bar{\bar{X}}}_{H_{old}}\right)$.
3. Sort the computations from the smallest to the largest. The order gives a permutation π on the index of observations. Suppose that the result of sorting is $d^2_{H_{old}}(\pi_1) \le d^2_{H_{old}}(\pi_2) \le \cdots \le d^2_{H_{old}}(\pi_n)$.
4. Form a set H_{new} consists of h observations of index $\pi(1), \pi(2), \cdots, \pi(h)$.
5. Compute $\vec{\bar{\bar{X}}}_{H_{new}}$, $S_{H_{new}}$ and $d^2_{H_{new}}\left(\vec{X}_i, \vec{\bar{\bar{X}}}_{H_{new}}\right)$ like in the point 2.
6. If $Tr\left(S^2_{H_{new}}\right) = Tr\left(S^2_{H_{old}}\right)$, the process is done. If $Tr\left(S^2_{H_{new}}\right) < Tr\left(S^2_{H_{old}}\right)$, the process is continued until the *k*-th iteration when $Tr\left(S^2_{H_{new}}\right) = Tr\left(S^2_{H_{old}}\right)$.
7. Suppose that S_{H_i} is the covariance matrix obtained from the *k*-th iteration. At the end of the *k*-th iteration we obtain $Tr\left(S^2_{H_1}\right) \ge Tr\left(S^2_{H_2}\right) \ge \ . \ . \ . \ \ge Tr\left(S^2_{H_{k-1}}\right) = Tr\left(S^2_{H_k}\right)$.

The MVV estimator for location parameters and covariance matrices respectively are $T_{MVV} = \vec{\bar{\bar{X}}}_{H_{new}}$ and $C_{MVV} = S_{H_{new}}$ on the *k*-th iteration. The robust Mahalanobis distance between $\vec{X}_i$ and T_{MVV} based on MVV, is written as $dR_{MVV}\left(\vec{X}_i, T_{MVV}\right)$, and it is defined on the quadratic form as,

$$dR^2_{MVV}\left(\vec{X}_i, T_{MVV}\right) = \left(\vec{X}_i - T_{MVV}\right)^t C^{-1}_{MVV}\left(\vec{X}_i - T_{MVV}\right)$$

for i = 1, 2, … , n. The data give large $dR_{MVV}\left(\vec{X}_i, T_{MVV}\right)$ value will be labeled as outlier (labeled outlier) and are assumed as candidates of outliers.

III. 4. The MVV Algorithm for the Univariate Scheme

In the univariate case, simply like in the multivariate, Iglewicz and Hoaglin (1993) stated the importance of outlier labeling as the first stage in identification process of outlier candidates. They presented methods which are frequently used, such as Z-scored, Boxplot, and Extreme Studentized Deviation (ESD). ESD which was first introduced by Rosner(1975), is very popular for practitioners. One basic weakness of ESD is that the critical point is obtained from an approximation through simulations. This weakness was improved by Djauhari (2001) by proposing the exact distribution. Tietjen and Moore (1972) offered a method for the case where there is more than one outliers. Also Rosner (1983) who proposed the generalized ESD (GESD) method.

At the end of this chapter the author introduces a method robust property and of maximal BP i.e. 0.5 for the univariate outlier labeling. This is the same with BP belongs to the median. The labeling method which the author proposes is a univariate version of MVV. In the univariate case, this criteria is equivalent to the predecessors i.e. MVE and MCD. The following is the algorithm for the outlier labeling of the univariate case.

1. Sort the data $x_1, x_2, ..., x_n$ from the smallest to the largest. Suppose the sorted data is $x_{\pi(1)}, x_{\pi(2)}, ..., x_{\pi(n)}$. Once again, π is a permutation on the index set {1, 2, … , n}.
2. Determine a set H consists of $h = \left[\frac{n}{2} + 1\right]$ data of minimum variance among all set of h possible data. The steps are as follow.

2.1. Compute the variance s_1^2 of $x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(h)}$, the variance s_2^2 of $x_{\pi(2)}, x_{\pi(3)}, \ldots, x_{\pi(h+1)}$, …. , and the variance $s_{(n-h+1)}^2$ of $x_{\pi(n-h+1)}, x_{\pi(n-h)}, \ldots, x_{\pi(n)}$.

2.2. Compute the minimum of s_1^2, s_2^2, …. , s_{h-1}^2. Suppose s_k^2 is the minimum. Then *H* consists of $x_{\pi(k)}, x_{\pi(k+1)}, \ldots, x_{\pi(k+h-1)}$.

3. Suppose that T_{Uni} and C_{Uni} are the mean and variance of *H*. Compute $d_{C_{Uni}}^2\left(x_i, C_{Uni}\right) = \dfrac{\left(x_i - T_{Uni}\right)^2}{C_{Uni}}$; *i* = 1, 2, ... , *n*.

4. Write $x_i = d_{C_{Uni}}^2\left(x_i, C_{Uni}\right)$; *i* = 1, 2, ... , *n*.

5. Repeat the step 1 – 4 until it is obtained a set *H* which is equal with one given in the previous iteration.

Chapter IV Examples of Outlier Labeling

IV. 1 Preface

In this chapter will be discussed benefit of the MVV method in outlier labeling for univariate cases and multivariate cases. Several examples on univariate and multivariate data will be given to show performance of the MVV in separating candidates of outlier.

IV. 2. Examples in the Univariate Case

The following examples describe advantages of the MVV method in separating 'suspects' on univariate data. Compared to the others four well known methods, i.e. the classic Mahalanobis method, boxplot, Z-scored, and the ESD method, apparently MVV gives better results.

a. The classic Mahalanobis distance method

This distance is often used to measure of how far a point from a mean sample with respect to a covariance matrix sample. Suppose $\vec{X}_1, \vec{X}_2, \cdots, \vec{X}_n$ are random sample of size n having $N_p(\vec{\mu}, \Sigma)$ where Σ is of positive definite. If $\bar{\vec{X}}$ is the vector of sample mean and **S** is sample covariance matrix , then the distance $d_{\mathrm{S}}\left(\vec{X}_i, \bar{\vec{X}}\right)$, where $d_{\mathrm{S}}^2\left(\vec{X}_i, \bar{\vec{X}}\right) = \left(\vec{X}_i - \bar{\vec{X}}\right)^t \mathbf{S}^{-1}\left(\vec{X}_i - \bar{\vec{X}}\right)$, is called the Mahalanobis distance of $\vec{X}_i$ to $\bar{\vec{X}}$. A point will be declared as labeled outlier if $d_{\mathrm{S}}^2\left(\vec{X}_i, \bar{\vec{X}}\right) > \mathrm{C}$, where $C = \frac{(n-1)}{n} F_{1,(n-1)}^{-1}$.

b. The boxplot method

This method is the popular graphical method to identify labeled outlier. Tukey (1977) proposed boxplot to separate the outlier candidates. Observations beyond the fences are labeled as outlier. The fences are determined by the upper bound – outlier (UBO) = $Q_1 + 1.5(Q_3 - Q_1)$ and the lower bound-outlier (LBO) = $Q_1 - 1.5(Q_3 - Q_1)$. Here, Q_1 and Q_3 are the first and the third quartile.

c. The Z-scored method

Iglewicz and Hoaglin (1993) proposed a modified Z -scored. The observations will be outlier candidates when $|M_i| > D$, where $M_i = \frac{0.6745\left(X_i - \bar{X}\right)}{\text{MAD}}$ and MAD = $Median\left\{\left|X_i - \bar{X}\right| \ \middle| \ i = 1, 2, \ldots, n\right\}$. Based on the simulation study, they suggest D = 3.5.

d. The ESD method with exact distribution

Djauhari (2001) improved the extreme studentized deviate (ESD) method proposed by Rosner (1975) by deriving the exact distribution. The critical point is

$$C = \frac{(n-1)^2}{n} Beta^{-1}\left((0.95) \ ; \frac{p}{2} \ ; \ \frac{(n-2)}{2}\right) .$$

e. The MVV method

Author proposes the method to separate suspects on univariate data(see section III.4). The observations are labeled outlier if $d^2_{C_{Uni}}\left(x_i, C_{Uni}\right) > CR$. Hadi (1992) proposed

$$CR = \frac{c_{npr} \ h}{\chi^2_{1,0.975}} \text{ and } c_{npr} = \left\{1 + \frac{(n-h)}{(n-p)}\right\}^2, \quad p = 1.$$

Example 1.

The following ordered data is about the strength of gears taken from Iglewicz and Hoaglin (1993 p.19).

1958, 2185, 2210, 2250, 2251, 2263, 2275, 2311, 2329, 2353, 2431

The data spread is shown on the dotplot on Figure IV.1.

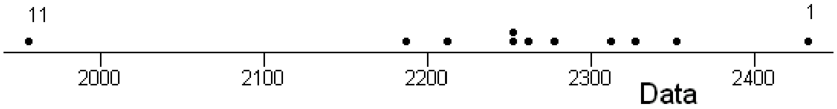


Figure IV.1. Data dotplot of gears strength

The figure indicates that the 11-th and the 1-st data are suspected as candidates of anomalies data. How the status of both data really are? The Figure IV.2 illustrates the result. The classic Mahalanobis distance, Z-scored and ESD only label the11-th data as labeled outlier. Because of masking effect, the 1-st data can not be identified as an labeled outlier. On other hand, boxplot and MVV apparently give the same result to the characteristic of data spread. They can detect that the 11-th and the 1-st data are candidates of outliers

GAMBAR 1AA (39)

Example 2.

The following are data of cholesterol level of a group of healthy people, courtesy of Bolton, taken from Djauhari (2001). Ordered data of 15 normal people is,

165, 194, 197, 200, 202, 205, 210, 214, 215, 227, 231, 239, 249, 297

The data spread is shown in dotplot on the following figure.

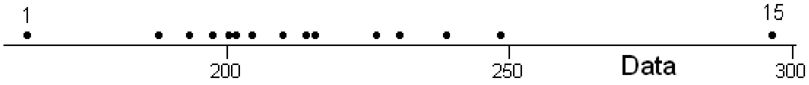


Figure IV.3. The dotplot of serum cholesterol data

Bolton, as cited by Djauhari (2001), stressed that "*without the presence of an obvious error, one would probably be remiss if these two values (165 and 297) were omitted from a report of normal cholesterol values in these normal subjects*". Next he added that "*with the knowledge that plasma cholesterol levels are approximately normally distributed, a statistical test can be applied to determine whether the extreme values (165 and 297) should be rejected*".

As in Example 1, the five methods will be used to identify the suspects. Figures IV.4 shows the result from five approaches.

GAMBAR 1BB (41)

Example 3.

It will be shown the advantages of the robust method in labeling outliers on data resulted from simulation. The numbers of 50 data are generated randomly from mixed normal univariate model where the 40 data are of standard normal distribution $N(0,1)$ and the remaining 10 date are of $N(5,1)$ distribution. The generating processes are done 50 times. The data spread is shown on the following dotplot.

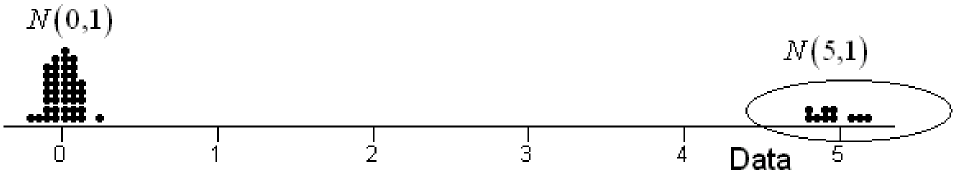


Figure IV.5. The dotplot of 50 univariate data from simulation

As it is shown on Figure IV.6., the Mahalanobis distance approach is unable to give correct outlier label. A masking effect has appeared. The masking effect is also found in the boxplot, the ESD and the Z-scored method. To avoid the masking and swamping effect, it was proposed the MVV method, a method of robust property. Based on the proposed method, all outliers can be labeled correctly (see Figure IV.6e)

GAMBAR 1CC (43)

IV. 3. Examples in the Multivariate case

This section will discuss performance of MVV and performance of the famous robust methods, MVE and FMCD. Compared to MVE and FMCD, MVV gives faster time. Furthermore, MVV gives better result than MVE, and MVV has the same effectivity from FMCD. Advantages of the MVV method on the outlier labeling process are given on the following examples.

Example 1.

This example uses data of physical dimension of Iris Virginica a kind of spider lily flower; taken from Mardia et al. (1979, p. 5-7).

Figures IV.7 illustrates the outlier labeling from data of Iris Virginica. Based on the approach of classic Mahalanobis distance it is apparent that there is no data can be identified as outlier candidates. The same result is also found on the robust approaches (MVE, FMCD and MVV). Even though analysis carried out by Rolfh (1975) gave 4 outliers, they are data no. 7, 20, 10, 15, and Wilks (1963) found data no. 19, 35, 7, 32 as outliers, it is clear that there is no data labeled as outlier.

Example 2.

This example uses data borrowed from Hawkins, Bradu and Kass (1984, p. 205, Table 4). The sample size is n = 75 and number of variables involved is p = 3, they are $X_{1,}$ X_2, and X_3.

GAMBAR 2AA

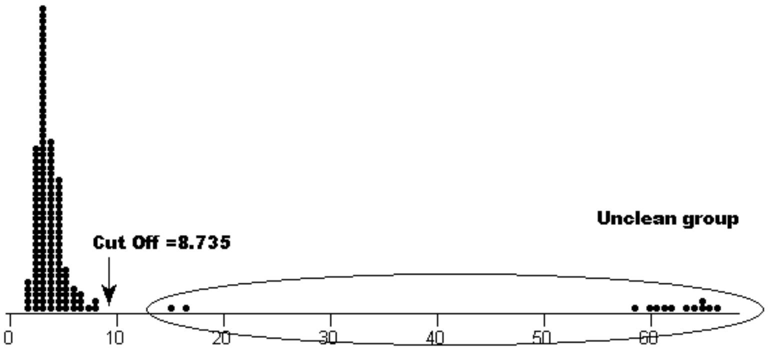
GAMBAR 2BB

Figure IV.8 illustrates the outlier labeling from data of HBK. Based on the classic Mahalanobis distance, it is apparent that there is no data can be identified as suspect. In the other hand MVE and FMCD give similar patterns. Candidates of outliers form a significantly separate group from clean data. From observations on unclean data, it is known that there are 14 observations labelled as outliers. The MVV method gives result of similar quality to the previous two methods FMCD and MVE. From the labeling result using MVV, it is apparent that the clean data and the unclean data are separated very clearly.

Example 3.

The numbers of 300 data are generated from the multivariate normal mixture model of low dimension, i.e. p =15. The model is $(1-\varepsilon)\, N_{15}(\vec{\mu}_1, \mathrm{I}_{15}) + \varepsilon\, N_{15}(\vec{\mu}_2, \mathrm{I}_{15})$, with $\varepsilon = 0.05$, $\vec{\mu}_1 = \vec{0}$, $\vec{\mu}_2 = 4\,\vec{e}$, and $\vec{e}$ is a vector of dimension 15 and all of its components are having value 1. From the model, there are 15 contaminant data out of 300 data as a result of small shift of the mean vector. Three robust methods, i.e. MVE, FMCD and MVV are used to identify suspects. The projection approach, which is proposed by Pena and Prieto (2001), is also done in this example.

Figure IV.9 gives a visualization of outlier labeling based on Pena and Prieto's method.



**Each dot represents up to 2 observations.

Figure IV.9. Outlier labeling by the method of Pena and Prieto of data resulted from simulation for p=15

In the Figure IV.9, it is seen that there are 17 data assumed as outliers and are grouped into the unclean group or contaminants. This number is more than it is supposed to, i.e. 15. This shows that the projection method of labeling above allows masking effects, which have produced another data to the group of contaminants

The outlier labeling from this case can be seen in Figure IV.10. The figure shows the classics and the robust distance approach. The pattern of classic Mahalanobis distance does not show any difference between the group of clean data and the contaminant. This is not surprising because the classic Mahalanobis distance is not robust. The MVE and FMCD give a better result than the classical approach. We see that the two approaches, MVE and FMCD are able to separate the group of clean data and the contaminants well. MVE and FMCD are able to label fairly the outliers.

Besides of advantage on the labeling aspect, FMCD also has an advantage on the aspect of time. A comparison of time processing the 300 data by using FMCD, MVE and Pena and Prieto’s method is given on Table IV.1.

Table IV.1 Comparison of time processing

Method	**Time (seconds)**
FMCD	32.5
MVE	673.9
Pena and Prieto’s method	715.6

GAMBAR 2cc

.

Even though FMCD gives the shortest time among the three methods, but for large *p*, for example hundreds, FMCD also requires long time. This is because the computation of determinant of covariance matrix of size $p \times p$. Through the Cholesky's decomposition method, required time to compute the determinant is of order $O\left(2^p\right)$. Relating with the aspect of processing time, MVV shows an advantage with time complexity of order $O\left(p^2\right)$ for VV computation. The following Table IV.2 shows differences on processing time of the determinant of covariance matrix $\left|C_{MVV}\right|$ of order $O\left(2^p\right)$and the computation of $Tr\left(C_{MVV}^2\right)$ of order $O\left(p^2\right)$.

Table IV.2 Comparison of computation time of VV and CD for *p* =15

Method	**Time (seconds)**
VV	1.10E-04
CD	1.31E-04

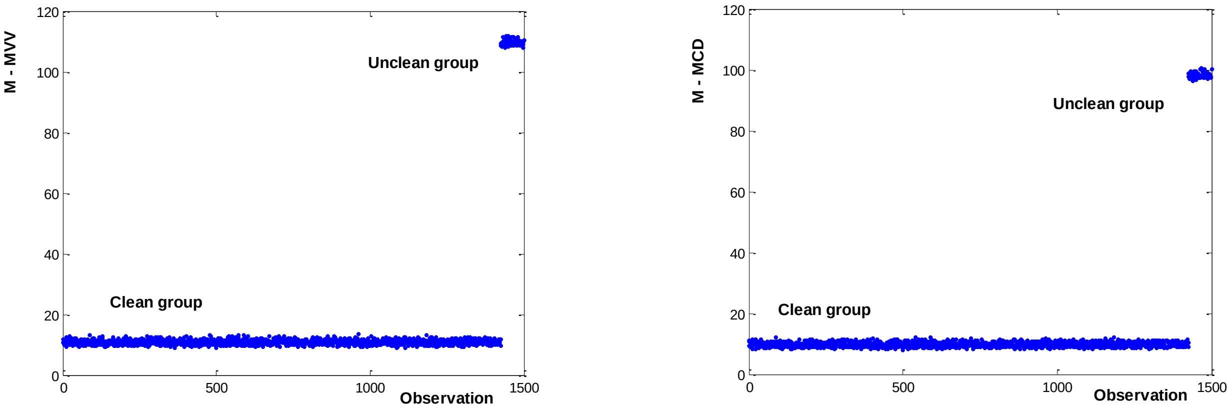
From Figure IV.10 also seen that the effectivity of MVV is equal with MVE and FMCD. Further more, as it is seen on Table IV.2, from the aspect of time MVV is much more efficient than FMCD.

Example 4

The numbers of 1500 random data are generated from a mixture model of high dimension, i.e. 100. The model is $(1-\varepsilon)\, N_{100}\,(\vec{\mu}_1, I_{100}) + (\varepsilon)\, N_{100}\,(\vec{\mu}_2, I_{100})$, with $\varepsilon = 0.05$, $\vec{\mu}_1 = \vec{0}$, $\vec{\mu}_2 = 10\,\vec{e}$, and $\vec{e}$ is a vector of dimension 100 where all of its components are 1.Thus, there are 75 contaminants data out of the 1500 data.

Next it will be found out the performance of FMCD and MVV on separating the 75 contaminants data from the group of clean data.

Figure IV.11 illustrates the effectivity of MVV and FMCD from this case, it shows that in labeling MVV has the same effectivity with FMCD. The computation time $Tr\left(C_{MVV}^{2}\right)$ is much faster than the computation time $\left|C_{MVV}\right|$. Table IV.3 presents a comparison of time between the two computations. Figure IV.11 and Table IV.3 strengthen reasons of usage of MVV on labeling outlier of *robust* property.



(a)

(b)

Figure IV. 11 The scatter plot based on (a) MVV Mahalanobis distance, (b) FMCD Mahalanobis distance

Table IV.3 The time comparison between VV and CD

Method	**Time (seconds)**
VV	5.03E-04
CD	4.76E-02

Chapter V Conclusions and Direction of Further Research

V.1. Conclusions

1. On the aspect of effectivity on labeling of outlier data, the examples in Chapter IV highly indicate that MVV is having the same effectivity with FMCD. This is a nice property of MVV which turns out to be a consideration for using it.
2. On the aspect of algorithm efficiency, MVV is much better than FMCD. Complexity on the computation of covariance matrix determinant (CD) by using the Cholesky's method is of order $O\left(2^p\right)$, meanwhile the *trace* computation merely needs $\operatorname{Det}(\mathbf{S})$.
3. Through a simulation study using Matlab 6.15, the illustration of comparison between required time in the computation of VV, $\operatorname{Tr}\left(\mathbf{S}^2\right)$, and CD, $\operatorname{Det}(\mathbf{S})$ for various size of positive definite symmetric matricex **S** of size $p \times p$ is presented on Table V.1.

Table V.1 Comparison between time of computation of CD and VV

p	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

4. Table V.1 shows clearly the advantages on the aspect of time given by MVV.

V.2. Direction of Further Research

The most fundamental problem which provide the basis of all robust methods such as MVE, MCD, MMCD, FSA, FMCD, and BACON is the definition of multivariate dispersion measure. Different with MVE which uses the least volume of ellipsoid that covers the whole data as the dispersion measure and BACON uses the corrected quantile of chi-squared distribution, the others use the determinant of covariance matrix. MVV that the author proposes is based on VV as the measure of multivariate dispersion. As measures of dispersion, VE, CD and VV, come with their own weakness. Two different structures of covariance could have the same measure.

Another fundamental problem is investigation on the distribution of $dR_{MVV}\left(\vec{X}_i, T_{MVV}\right)$. The (1 – α)-th quantile of the distribution is required to test the hypothesis, whether labeled outliers are really outliers.

Those are fundamental problems on the multivariate outlier identification that author stresses as further directions of research. Specifically, the directions are (1) how to build a measure of dispersion that has a better performance in explaining the situation of spreading of multivariate data, and (2) what is the distribution of $dR_{MVV}\left(\vec{X}_i, T_{MVV}\right)$.

REFERENCES

1.Alt, F.B. and Smith, N.D. (1988), Multivariate Process Control, *Handbook of Statistics*, 7, 333-351.

2.Anderson, T.W. (1984), *An Introduction to Multivariate Statistical Analysis*, Second Edition, John Wiley, New York

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

4.Barnett, V. and Lewis, T. (1984), *Outliers in Statistical Data*, Second Edition, John Wiley , New York.

5.Becker, C. and Gather, U. (1999), The Masking Breakdown Point of Multivariate Outlier Identification Rules, *Journal of the American Statistical Association*, 94, 947 – 955.

6.Beckman, R.J. and Cook, R.D. (1983), Outlier …s, *Technometrics*, 25, 119 - 149 .

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

8.Brassard, G. and Bratley, P. (1982), *Fundamentals of Algorithmics*, Prentice Hall, New Jersey

9.Campbell, N.A. (1980), Robust Procedures in Multivariate Analysis 1: Robust Covariance Estimation, *Applied Statistics*, **29**, 231 – 237.

10.Caroni, C. and Prescott, P. (1992), Sequential Application of Wilks's Multivariate Outlier Test, *Applied Statistics*, **41** (2), 355-364.

11.Cleroux, R., Helbling, J.M. and Ranger, N. (1986), Some Methods of Detecting Multivariate Outliers, *Journal Computational Statistics Quarterly*, **3**, 177 – 195.

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

13.Derquenne, C. (1992), Outlier Detection Before Running Statistical Methods, *Siam*, **34** (2), 323 – 326.

14.Dixon, W.J. (1950), Analysis of Extreme Values, *Annals of Mathematical Statistics*, 21, 488 - 506.

15.Djauhari, M.A. (1996), A Necessary and Sufficient Condition for the Uniqueness of Minimum Spanning Tree, *Proceedings Institut Teknologi Bandung*, **29** (1/2), 11-18.

16.Djauhari, M.A. (1999), An Exact Test for Outlier Detection, *BioPharm; The Applied Technology of Biopharmaceutical Development, Milwaukee*, **12** (6).

17.Djauhari, M.A. (2001), Improving the Extreme Studentized Deviation (ESD) Procedure for Outlier Testing, *BioPharm; The Applied Technology of Biopharmaceutical Development, Milwaukee*, **14** (3).

18.Djauhari, M.A. (2003), Statistical Testing for Outliers: Calculating the Critical Point of the Extreme Studentized Deviation Using the Beta Inverse Function, *BioPharm International; The Applied Technology of Biopharmaceutical Development, Milwauke*, **16** (10), 60 – 68.

19.Djauhari, M.A. (2004), Mahalanobis Distance of Two Complementary Groups of Observations, *Proceedings of the 12th National Symposium on Mathematical Sciences*, International Islamic University Malaysia.

20.Djauhari, M.A. (2005a), Improved Monitoring of Multivariate Process Variability, *Journal of Quality Technology*, **37** (1), 32-39.

21.Djauhari, M.A. (2005b), Outlier Detection: Some Challenging Problem for Future Research, *Proceedings of the Second International Conference on Research and Education of Mathematics*, Universiti Putra Malaysia.

22.Franklin, S. and Broudeur, M. (2005), *A practical Application of a Robust Multivariate Outlier Detection Method*, Statistics Canada, BSMD, R.H. Coats Bldg, 11th floor, Ottawa, Ontario, Canada, K1A 0T6

23.Friedman, J.H. (1987), Exploratory Projection Pursuit, *Journal of the American Statistical Association*, Vol 82, No 397, 249-266.

24.Grambow, S. and Stromberg, A.J. (1998), *Combining the EID and FSA for Computing the Minimum Volume Ellipsoid*, Department of Statistics, University of Kentucky.

25.Grubbs, F.E. (1950), Sample Criteria for Testing Outlying Observations, *Annals of Mathematical Statistics*, 21, 28-58.

26.Grubbs, F.E., (1969), Procedures for Detecting Outlying Observations Samples, *Technometrics*, 11, 1-21.

27.Gnanadesikan, P. and Kettenring, J.R. (1972), Robust Estimates, Residuals, and Outlier Detection with Multirespon Data, *Biometrics*, 23, 81-124.

28.Hadi, A.S. (1992), Identifying Multivariate Outlier in Multivariate Data, *Journal of Royal Statistical Society B,* **53** (3), 761-771.

29.Hampel, F.R. (1974), The Influence Curve and Its Role in Robust Estimation. *Journal of the American Statistical Association*, **69** (346), 383 – 393.

30.Hampel, F.R., Ronchetti, E. M., Rousseuw, P.J. and Stahel, W.A. (1985), *Robust Statistics,* John Wiley , New York.

31.Hardin, J. and Rocke, D.M. (2002), The Distribution of Robust Distance, http://www.cipic.ucdavis.edu/~dmrocke/preprints.html.

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

33.Hawkins, D.M. (1980), *Identification of Outliers*, Second Edition, Chapman and Hall , New York.

34.Hawkins, D.M., (1994), The Feasible Solution Algorithm for the Minimum Covariance Determinant Estimator in Multivariate Data, *Computational Statistics and Data Analysis*, 17, 197-210.

35.Hawkins, D.M., Bradu, D. and Kass, G.V. (1984), Location of Several Outliers in Multiple Regression Data Using Elemental sets, *Technometrics*, **26** (3), 197-208.

36.Hawkins, D.M. and Olive D.J. (1999), Improved Feasible Solution Algorithm for High Breakdown Estimation, *Computational Statistics and Data Analysis*, 30, 1-11.

37.Hoaglin, D.C., Mosteller, F. and Tukey, J.W. (1983), *Understanding Robust and Exploratory Data Analysis*, John Wiley, New York.

38. Huber, P.J. (1964), Robust Estimation of Location Parameter, *Annals of Mathematical Statistics*, 35, 73-101.

39. Huber, P.J. (1980), *Robust Statistics*, Massachusetts, Wiley Series in Probability and Mathematical Statistics.

40. Iglewicz, B. and Hoaglin, D.C. (1993), *How to Detect and Handle with Outliers*, American Society for Quality, Statistics Division, **16**, Milwaukee.

41. Irwin, J.O. (1925), On a Criterion for the Rejection of Outlying Observations, *Biometrics*, **17** (3/4), 238-250.

42. Johnson, R.A. and Wichern, D.W. (1988), *Applied Multivariate Statistical Analysis*, Second Edition, John Wiley, New York.

43. Jolliffe, I.T. (1986), *Principal Component Analysis*, Springer Verlag.

44. Kitagawa, G. (1979), On the Use of AIC for the Detection of Outliers, *Technometrics*, **21** (2), 193 – 199.

45. Kotz, S. and Johnson, N.L (1985), *Encyclopedia of Statistical Sciences,* **6** (110-122), John Wiley, New York.

46. Kuwahara, S.S. (1997), Outlier Testing: Its History and Application, *BioPharm; The Technology and Business of Biopharmaceutical*, **10** (4), 64 – 67.

47. Lazraq, A. and Cleroux, R. (1989), On the Detection of Multivariate Data Outliers and Regression Outliers, *Data Analysis, Learning Symbolic and Numerical Knowledge* (E. Diday eds), Inria, Nova Science Publishers, Inc., 133 – 140.

48. Lopuhaa, H.P. and Rousseeuw, P.J. (1991), Breakdown Points of Affine Equivariance Estimators of Multivariate Location and Covariance Matrices, *Annals of Statistics*, **19** (1), 229 – 248.

49. Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979), *Multivariate Analysis,* Academic Press, London.

50. Montgomery, D.C. (2001), *Introduction to statistical Quality Control*, Fourth Edition, John Wiley, New York.

51. 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, 153-167.

52. Pearson, E.S. and Chandra Sekar, C. (1936), The Efficiency of the Statistical Tool and a Criterion for the Rejection of Outlying Observations, *Biometrika*, 28, 308-320.

53.Pena, D. and Prieto, J.F. (2001), Multivariate Outlier Detection and Robust Covariance Matrix Estimation, *Technometrics*, **43** (3), 286- 300.

54. Pena, D. and Rodriguez, J. (2003), Descriptive Measures of Multivariate Scatter and Linear Dependence, http://halweb.uc3m.es/esp/personal/dpena/article/JMVA03.PDF

55.Rocke, D.M. (2002), Multivariate Outlier Detection and Cluster Identification. *International Conference on Robust Statistics (ICORS)*, May 13, University of British Columbia.

56.Rocke, D.M. and Woodruff, D.L. (2000), A Synthesis of Outlier Detection and Cluster Identification. http://handel.cipic.ucdavis.edu/dmrocke/Synth5.pdf.

57.Rolfh, F.J. (1975), Generalization of the Gap Test for the Detection of Multivariate Outliers, *Biometrics*, 31, 93-101.

58.Rosner, B. (1975), On the Detection of Many Outliers, *Technometrics*, **17** (2), 221-227.

59.Rosner, B. (1983), Percentage Points for a Generalized ESD Many-Outlier Procedure, *Technometrics*, **25** (2), 165 – 172.

60.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**, 283-297. D. Reidel Publishing Company.

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

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

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

64.Shone, J.B. and Fung, W.K. (1987), A New Graphical Method for Detecting Single and Multiple Outliers in Univariate and Multivariate Data, *Applied Statistics*, **36** (2), 153-162.

65. Serfling, R.J. (1980), *Approximation Theorems of Mathematical Statistics*, John Wiley , New York.

66. Suwanda and Djauhari, M.A. (2003), A New Concept In Monitoring Multivariate Process Variability, *Newsletter*, *Data Analysis Research Group*, Department of Mathematics Institut Teknologi Bandung, **2** (2).

67.Tietjen, G.L. and Moore, R.H. (1972), Some Grubbs-type Statistics for The Detection of Several Outliers, *Technometric*, 55, 583 – 598.

68. Thompson, W. R. (1935), On a Criterion for the Rejection of Observation and the Distribution of the Ratio of Deviation to Sample Standard Deviation, *The Annals of Mathematical Statistics*, **6** (4), 214-219.

69. Tukey, J.W. (1977), *Exploratory Data Analysis*, Addison - Wesley, Canada.

70.Viljoen, H. and Venter, J.H. (1999), A Computer Intensive Approach to Find Multivariate Outliers, http://www.stat.fi/isi99/proceedings/arkisto/varasto/vilj0153.pdf .

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

72.Wilks, S.S. (1963), Multivariate Statistical Outliers, *Sankya* A, 25, 407-426.

73.Woodruff, D.L. and Rocke, D.M. (1994), Computable Robust Estimation of Multivariate Location and Shape in High Dimension Using Compound Estimators, *Journal of the American Statistical Association*, **89** (427), 888 – 896.

74.Ye, N., Borror, C.M., and Parmar, D. (2003), Scalable Chi-Square Distance versus Conventional Statistical Distance for Process Monitoring with Uncorrelated Data Variables, *Quality and Reliability Engineering International*, **19**, 505-515.