ROBUST MAHALANOBIS DEPTH BASED ON MINIMUM REGULARIZED COVARIANCE DETERMINANT ESTIMATOR FOR HIGH-DIMENSIONAL DATA

R Muthukrishnan and Surabhi S Nair
•
Department of Statistics, Bharathiar University, Coimbatore 641046, Tamil Nadu, India
muthukrishnan1970@gmail.com, surabhinair93@gmail.com

Abstract

Handling of high-dimensional data is an important issue in robust literature. For analyzing data, location measure plays a vital role in almost all statistical methods. The location parameter of a distribution is used to find the central value. Many computational methods are used to find the measure of location for analyzing data. The data depth procedure is one approach to finding the true representative of the entire data and it is one of the key concepts in multivariate data analysis. Data depth is a term used to describe how deep a particular point is inside the broad multivariate data cloud. Instead of the typical smallest to biggest rank, the sample points can be ordered from the center outward. Mahalanobis depth is one of the popular depth procedures. The traditional approach used to find Mahalanobis depth is based on Mahalanobis distance, it is based on the classical sample mean vector and covariance matrix. So the conventional Mahalanobis depth is sensitive to outliers and may fail when the data is contaminated. To solve this problem, the Minimum Covariance Determinant (MCD) estimators are used instead of classical estimators. However, the MCD estimators cannot be calculated in high dimensional data sets, in which the variable number is higher than the subset size. To calculate Mahalanobis depth values in high dimensional data, propose a new depth function namely the Robust Regularized Mahalanobis Depth (RRMD), which can be calculated in high dimensional data sets. The proposed procedure is based on Minimum Regularized Covariance Determinants (MRCD) estimators, this study shows that the proposed depth function is successful in finding the deepest point in high dimensional data sets with real and simulation studies up to a certain level of contamination.

Keywords: mahalanobis depth, outliers, robust distance, minimum covariance determinant estimator, minimum regularized covariance determinant estimators

I. Introduction

Data depth is a key concept in the nonparametric method of multidimensional analysis of data. The idea of data depth was proposed by Tukey [13] as a graphical tool for displaying two-dimensional data sets, and it has now been expanded to include multivariate data sets, Donoho and Gasko [3]. Data depth is a statistical technique that describes data distribution in accordance with center-outward ordering instead of density or linear ordering, Liu [6], Modarres [10]. According to Liu et al. [8], the idea of data depth is being used for statistical analysis with multiple variables since it offers a nonparametric method. Several researchers have established various notions of depth preliminaries in the literature.

Researchers are looking for solutions in robust depth procedures to address the sensitivity issue in high-dimensional data analysis. Also can build robust depth procedures to handle the

presence of outliers by replacing the classical estimators with robust estimators such as M-estimators, Minimum Covariance Determinant (MCD) estimators, Mia Hubert and Michiel Debruyne [4], Minimum Volume Ellipsoid (MVE) estimators, Stefan Van Aelst and Peter Rousseeuw [12], and Minimum Regularized Covariance Determinants (MRCD), Kris Boudt, Peter Rousseeuw [1].

A very reliable estimator of multivariate location and scatter is the Minimum Covariance Determinant (MCD) approach. Given an $n \times p$ data matrix $Y = (Y_2, \dots, Y_p)'$ with $y_i = (y_{i1}, y_{i2}, \dots, Y_{ip})'$, its objective is to find h observations whose sample covariance matrix has the lowest possible determinant. Here $h < n$ is fixed. The average of these h points serves as the MCD estimate of location, while the scatter estimate represents a multiple of the covariance matrix. The maximum possible breakdown value is found in the MCD, which also possesses a bounded influence function (i.e. 50%) when $h = [(n + p + 1)/2]$. This covariance matrix of any h-subset must be non-singular for the dimension p to meet $p < h$, which is a significant restriction of the MCD technique. In fact, taking n > 5p is frequently advised for the estimator's accuracy. This restriction leaves a hole in the selection of high-breakdown techniques. In order to reduce this gap, Boudt et al. (2020) suggested modifying the MCD so that it can be used for high dimensions. A regularized covariance estimate, which is a weighted mean of the sample covariance of the h-subset and a preset positive definite target matrix, is intended to replace the subset-based covariance. The regularized covariance based on the h-subset that results in the least overall determinant is then the Minimum Regularized Covariance Determinant (MRCD) estimate. In addition to supporting high dimensions, the MRCD estimator's key characteristics include maintaining the MCD estimator's excellent breakdown qualities and being highly conditioned by construction.

This study proposes a new depth procedure, namely Robust Regularized Mahalanobis Depth (RRMD), which can be used to find the location measure in high dimensional datasets by comparing the existing method. In the suggested algorithm, Mahalanobis depth is obtained based on MRCD estimators instead of classical mean and covariance matrix. Using various kinds of simulation tests and two real datasets, it is examined if the recommended algorithm for location estimation in high dimensional data produces accurate results even if the data is contaminated.

The rest of this paper is structured as follows. Section 2 describes traditional Mahalanobis depth, the robust estimator used, and the proposed method. The results and discussions based on the real data and simulation study will be given in section 3. The conclusion will be provided in the last section.

II. Methods

Today, in numerous domains, huge amounts of data are produced and tainted by noise. It is important to establish training methods that are resilient to data instabilities and disruptions. In this section, the foundations of controlled learning are discussed, including the traditional Mahalanobis Depth, the estimator used in this study – Minimum Regularized Covariance Determinant Estimator, and the proposed depth procedure - Robust Regularized Mahalanobis Depth.

I. Mahalanobis Depth

Mahalanobis depth (MD) was first described by Liu et al. [7] from Mahalanobis distance. Mahalanobis [9] established the statistical idea of generalized distance which is calculated by using a classical mean vector and covariance matrix. For determining the Mahalanobis depth of an observation, the Mahalanobis distance is used. The positive inverse of Mahalanobis distance is termed as Mahalanobis depth. For an observation $y \in S \subset R^d$ about a d- dimensional data,
The squared Mahalanobis distance (D) and Mahalanobis depth (MD) are given by

$$D(Y, \bar{Y}, S) = (Y - \bar{Y})' S^{-1} (Y - \bar{Y})$$

$$MD(Y,\bar{Y},S) = [1 + D(Y,\bar{Y},S)]^{-1}$$

where $\bar{Y}$ and S are the mean vector and sample covariance matrix. Since it is reliant on non-robust parameters like the mean and dispersion matrix, this algorithm lacks to be reliable.

To get a reliable result MD is calculated using a robust location vector and covariance matrix using MCD estimator instead of classical mean vector and covariance matrix. Generally, the Minimum Covariance Determinants (MCD) estimators are used for this aim. Due to the failure of MCD estimators to be generated for high-dimensional datasets, this approach is not applicable in these cases. So, the Mahalanobis depth using MCD estimator can't be applicable when want to analyze a high dimensional data set.

II. Minimum Regularized Covariance Determinant Estimator (MRCD)

The MRCD estimator is a modified version of NCD estimators for high-dimensional data and was proposed by Boudt et. al. [1]. To guarantee that the MRCD scatter estimator is scale equivariant and location unvarying, as is common in the literature, first, standardize the variables. The use of a trustworthy univariate location and scale estimate is required for standardization. For this, the median of each subset is calculated and placed in a location vector called m_x. Additionally, each variable's scale using the Qn estimator of Rousseeuw and Croux (1993) is calculated, then insert these scales into d_x, the diagonal matrix.

Let $X = (x_1, x_2, \dots, x_n)'$ be an $n \times p$ matrix with $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})'$; its goal is to identify the h observations with the sample covariance matrix with the lowest determinant. The term H refers to a set of h variables denoting the data contained in the subset, while the term $\mathcal{H}_h$refers to the compilation of all of these sets. The corresponding $h \times p$ submatrix of X is denoted by X_H for a particular H in $\mathcal{H}_h$. The mean and sample scatter matrix of X_H are given by

$$M_X(H) = h^{-1} X_H' I_h \tag{1}$$

$$S_X(H) = (h-1)^{-1}(X_H - M_X(H))'(X_H - M_X(H)) \tag{2}$$

After that, the MCD method seeks to minimize the determinant of $S_X(H)$ for all $H \in \mathcal{H}_h$.

$$H_{MCD} = \underset{H \in \mathcal{H}_h}{argmin} \left(det(S_X(H))^{1/p} \right) \tag{3}$$

For statistical considerations, eqn (3) takes the p^{th} root of the determinant. The geometric average of its eigenvalues is the p^{th} root of the determinant of the scatter matrix. It is referred to as the standardized generalized variance.

The mean of the h-subset is used to define the MCD estimate of location M_{MCD}, while the MCD scatter estimate is expressed as a multiple of the sample scatter matrix, and is given by

$$M_{MCD} = M_X(H_{MCD}) \tag{4}$$

$$S_{MCD} = C_\alpha S_X(H_{MCD}) \tag{5}$$

where C_α is a consistency factor that is based on the trimming percentage = $(n-h)/n$ and is similar to the one provided by Rousseeuw and Croux [11].

The standardized observation is given by

$$Z_i = d_x^{-1}(x_i - m_x) \tag{6}$$

The regularized scatter matrix of the standardized observation is

$\mathcal{S}(H) = \rho T + (1-\rho)C_\alpha S_Z(H)$
where $S_Z(H)$ is defined in (2), however, in the case of Z, c is the same consistency parameter as in (5).

Let A be the diagonal matrix containing eigenvalues of T, and the orthogonal matrix Q contains the relevant eigenvectors. Utilizing the spectral decomposition $T = QAQ'$ will be practical.

Now,
$$\mathcal{S}(H) = QA^{1/2}[\rho I + (1-\rho)C_\alpha S_W(H)]AA^{1/2}Q' \quad (7)$$

where W is the $n \times p$ matrix consisting of the transformed standardized observations $w_i = A^{-1/2}Q'Z_i$, and $S_W(H) = A^{-1/2}Q'S_ZQA^{-1/2}$

The MRCD subset is given by

$$H_{MRCD} = \underset{H\in\mathcal{H}_h}{argmin}\left(det(\rho I + (1-\rho)C_\alpha S_W(H))^{1/p}\right) \quad (8)$$

The MRCD location and scatter estimations of the initial data matrix X are defined as follows

$M_{MRCD} = m_X + d_x M_Z(H_{MRCD})$

$S_{MRCD} = C_\alpha S_X(H_{MRCD})$.

III. Robust Regularized Mahalanobis Depth

The proposed depth procedure namely Robust Regularized Mahalanobis Depth (RRMD), it is based on the Minimum Regularized Covariance Determinant estimator, which can be calculated in high dimensional data sets. ie., Mahalanobis depth can be calculated using robust location and covariance matrix calculated, the MRCD estimator instead of classical mean vector and covariance matrix. The robust MRCD estimator was proposed by Boudt et. al. [1] to locate the robust measure of location and scatter for high dimensional data. The computational depth procedure for RRMD is as follows.

Let M_{MRCD}, and S_{MRCD} be the location and scatter matrix using the MRCD estimator, D_{MRCD} diagonal matrix, which consists of the diagonal elements of S_{MRCD} matrix. The Robust Regularized Mahalanobis Depth obtained from the regularized squared mahalanobis distance, $D(Y, M_{MRCD}, S_{MRCD}) = (Y - M_{MRCD})' D_{MRCD}{}^{-1}(Y - M_{MRCD})$,, and is given by

$$MD(Y, M_{MRCD}, S_{MRCD}) = [1 + D(Y, M_{MRCD}, S_{MRCD})]^{-1}$$

Let $Y = (Y_2, \dots, Y_d)$ be a d dimensional multivariate data set and y be a numerical vector whose depth is to be calculated. The following steps are carried out to find the proposed method.

i. By using the dataset calculate robust MRCD location (M_{MRCD}) and scatter estimators (S_{MRCD}). Obtain D_{MRCD} diagonal matrix, which consists of the diagonal elements of S_{MRCD} matrix.

ii. The Regularized Squared Mahalanobis distance can be calculated from (i)

ie., $D(Y, M_{MRCD}, S_{MRCD}) = (Y - \mu_{MRCD})' D_{MRCD}{}^{-1}(Y - \mu_{MRCD})$

iii. S_D be the sorted distance given in (ii)

iv. M_{S_D} be the median from the distance from (iii),

$M_{S_D} = Median(S_D)$

v. D_y be the difference between Regularized Squared Mahalanobis distance value from (ii) and median from (iv), ie., $D_y = D(Y, \mu_{MRCD}, \Sigma_{MRCD}) - M_{S_D}$

vi. $Abs(D_y)$ be the absolute value of the difference in (v)

vii. Now, the proposed depth procedure, Robust Regularized Mahalanobis Depth can be calculated by $RRMD_y = [1 + Abs(D_y)]^{-1}$

III. Experimental Results

The performance of the proposed RRMD procedure over the classical MD procedure, the experiments were conducted under actual and simulation environments by computing location measure corresponding to the deepest point in high-dimensional data and thus the obtained results are demonstrated in this sections.

I. Real data study

The proposed depth function can be used to find the location parameter in high-dimensional datasets. Two real data set is used here to evaluate the performance of the suggested algorithm compared with the existing method. The First one is the brain data which is also from the "rda" package of R software. The brain data contains two objects, namely the microarray expression data for 42 brain cancer samples, and the class labels for these samples. An expression data matrix (42x5597) and a class label vector for 42 samples. The second one is the NCI60 data which is obtained from the "ISLR" package of R software. The data contains expression levels on 6830 genes from 64 cancer cell lines. Due to the enormous dimensions of these datasets, this study only used the first p ($p > 3 * n$) variables for convenience. The results obtained from the real data study are summarized in the form of Table 1.

Tab le 1: *Deepest point and observation number of brain data and NCI60 data*

Methods		MD	RRMD
Brain data	Highest Depth Value	0.148 (0.103)	0.148 (0.103)
	Deepest Point	42 (16)	4 (4)
NCI60 data	Highest Depth Value	0.097 (0.086)	0.031 (0.029)
	Deepest Point	23 (5)	49 (49)

(.) – Without outlier

The suggested algorithm, Robust Regularized Mahalanobis Depth (RRMD) is obtained based on MRCD estimators and then the location parameter is calculated using the depth values for the high dimensional data set. Table 1 shows that the proposed method gives a same location under with/without outlier conditions, but the conventional method differs from it.

II. Simulation Study

Simulation study were carried out under two different dimensions along with various amounts of contamination are employed to compare the performance of the proposed methodology to the current approach. The experiments were carried out by computing the maximum depth values that correspond to location measure.

First generated data with dimension, 100×300, with mean vector $\mu = (0,0,\dots,0)_{1\times 300}$and covariance matrix$\Sigma = I_{300}$. Here n=100, and p=300. Further same experiments were performed under various levels of contaminations, such as ε= 10%, 15%, 20%, 30% (For Location, $\mu = (1.5, 1.5, \dots, 1.5)_{1\times 300}$ and $\Sigma = I_{300}$, Scale, $\mu = (0,0,\dots,0)_{1\times 300}$, and $\Sigma = 1.5I_{300}$, Location and Scale, $\mu = (1.5, 1.5, \dots, 1.5)_{1\times 300}\ and\ \Sigma = 1.5I_{300}$) are taken into account. Also, the same experiment is repeated for 200×400 dimensional data. The results obtained from the simulation study is given in table 2 and 3 respectively.

From the simulation study it is concluded that the suggested depth procedure, RRMD can tolerate and gives the same deepest point up to a certain level of contamination. The MD method fails to provide identical location measurements even if the data contamination is very low.

Table 2: *Deepest point and observation number* under various contamination models

Simulation Study 1 Dimension:100 × 300; n=100, p=300						
Highest Depth Values and the corresponding observation						
ε	**Location Contamination**		**Scale Contamination**		**Location-Scale Contamination**	
	MD	***RRMD***	***MD***	***RRMD***	***MD***	***RRMD***
0.10	0.099 (96)	0.019 (33)	0.870 (41)	0.018 (33)	0.044 (93)	0.019 (33)
0.15	0.034 (81)	0.019 (33)	0.056 (92)	0.018 (33)	0.086 (79)	0.019 (33)
0.20	0.126 (10)	0.019 (33)	0.108 (77)	0.018 (33)	0.032 (88)	0.019 (33)
0.30	0.014 (38)	0.019 (28)	0.082 (30)	0.018 (22)	0.046 (90)	0.019 (22)
(.) – Observation number						

Table 3: *Deepest point and observation number* under various contamination models

Simulation Study 2 Dimension:200 × 400; n=200, p=400						
Highest Depth Values and the corresponding observation						
ε	**Location Contamination**		**Scale Contamination**		**Location-Scale Contamination**	
	MD	RRMD	MD	RRMD	MD	RRMD
0.10	0.013 (5)	0.009 (17)	0.100 (20)	0.009 (17)	0.124 (86)	0.009 (17)
0.15	0.151(61)	0.009 (17)	0.162 (18)	0.009 (17)	0.073 (130)	0.010 (17)
0.20	0.356 (33)	0.010 (17)	0.406 (155)	0.009 (17)	0.042 (128)	0.010 (124)
0.30	0.126 (21)	0.010 (17)	0.035 (80)	0.009 (131)	0.112 (80)	0.010 (128)
(.) – Observation number						

IV. Conclusion

Conventional methods should work reasonably well if certain assumptions are true, however, they may not be trustworthy if one or more of these assumptions are erroneous. Both sample mean vector and covariance matrix are extremely susceptible to anomalies. As a result, when the data contains anomalies, the traditional Mahalanobis depth fails to generate reliable results. For non-normal conditions, a robust alternative is required to improve accuracy even when the data somewhat depart from the model assumptions. When robust estimators such as MCD and MRCD are used the analysis performs well compared to traditional methods. To find the location measure in high dimensional data, this paper proposed a new robust depth procedure namely RRMD. The proposed method is compared with the existing procedure and gives reliable results up to certain levels of contamination. Robust methods perform well even when the model assumptions are not met. The study came to the conclusion that for robust and affine equivariant location, the proposed depth procedure gives better results followed by the existing method. By finding the deepest point in a dataset instead of relying on a more conventional method of determining location, the research groups can find the best location with greater precision when using these methods.

References

[1] Boudt K., Rousseeuw P.J. and Vanduffel S. (2020). The minimum regularized covariance determinant estimator. *Statistics and Computing*, 30:113–128.

[2] Choral C.Y. and Rousseeuw P.J. (1992) Integrating a high-breakdown option into discriminant analysis in exploration geochemistry. Journal of Geochemical Exploration, 43:191-203.

[3] Donoho D.L. and Gasko M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness, *Annals of Statistics*, 2:1803–1827.

[4] Hubert M. and Debruyne M. (2009). Minimum covariance determinant. *WIREs Computational Statistics*, 2:36-43.

[5] Hubert M. Debruyne M. and Rousseeuw P.J. (2017). Minimum covariance determinant and extensions. *WIREs Computational Statistics*, 10: e-1421.

[6] Liu R.Y. (1992). Data depth and multivariate rank tests. *In: Dodge, Y. (ed.), L1-Statistics and Related Methods*, North-Holland (Amsterdam). 279–294.

[7] Liu R.Y., Parelius J.M. and Singh K. (1993.) Multivariate analysis by data depth: Descriptive Statistics, Graphics and Inference. *The Annals of Statistics*, 27:783-858.

[8] Liu R. Y. and Singh K. A. (2012). Quality Index Based on Data Depth and Multivariate Rank Tests. *Journal of the American Statistical Association,* 88:252-260.

[9] Mahalanobis P. (1936). On the generalized distance in statistics. *Proceedings of the National Academy India*, 12:49–55.

[10] Modarres R. (2014). Data Depth. *In: Lovric, M. (eds) International Encyclopedia of Statistical Science*. *Springer, Berlin, Heidelberg*, 334-336.

[11] Rousseeuw P. J. and Croux C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88:1273–1283.

[12] Stefan V.A. and Rousseeuw P. J. (2009). Minimum volume ellipsoid. *WIREs Computational Statistics*, 1:71-82.

[13] Tukey J. W. (1975). Mathematics and the picturing of data. *In: Proceeding of the International Congress of Mathematicians*, Vancouver, 523–531.