Journal of Computer and Mathematical Sciences, *Vol.9(12),1875-1881 December 2018* *ISSN 0976-5727 (Print)*
(An International Research Journal), *www.compmath-journal.org* *ISSN 2319-8133 (Online)*

Detection of Multidimensional Outlier Using Multivariate Spatial Median

Sajana O. K and Sajesh T. A

Department of Statistics,
St Thomas' College (Autonomous), Thrissur District, Kerala, INDIA.
Department of Statistics,
St Thomas' College (Autonomous), Thrissur District, Kerala, INDIA.
email: sajana.kunjunni@gmail.com, sajesh.t.abraham@gmail.com

(Received on: November 29, 2018)

ABSTRACT

Robust estimators of location vector and dispersion matrix are significant part of outlier detection in multidimensional data. This paper attempt to propose an outlier detection method by estimation of dispersion matrix based on spatial median. The distributional properties of the estimator are studied using simulation. Some empirical properties of the proposed estimator in outlier detection are discussed. Applications and diagnostic plots are included.

Keywords: Outlier detection, robust estimation, spatial median, dispersion.

1. INTRODUCTION

Statistical inferences are generally based on some general assumptions about the underlying characteristics of the data. In practice these assumptions can be violated since observations may deviate from the specified model. Single or set of observations deviating from the common behavior of rest of the dataset is called as outlier (contaminated observations). Here focus is on the estimation part of statistical inference is carried on. The conventional method of estimating location and scatter of multivariate data uses all the observations in the given data set, but it fails to produce reliable estimates when the data contains outliers. It becomes necessary to analyze given data set if it is contaminated or not.

Obviously, the classical methods for estimating Multivariate location and scatter are highly influenced by the outlying observations. One of the remedy to the outlier problem is the robust estimation of the parameters. Many robust methods are developed over these years for the robust estimation of Multivariate location and scatter. Minimum Volume Ellipsoid

(MVE) and Minimum Covariance Determinant (MCD) method proposed by[1] it deals with computing subset of observations with smallest covariance determinant that would hold at least half of the observation. More improved Fast MCD algorithm was proposed by[2]. For large variable dimensions Fast MCD requires substantial running time consequently the procedure become computationally expensive. There are robust estimates based on optimizing an objective function like S-estimator proposed by[3 and 4] introduced Multivariate τ-estimators. One can use positive definite and approximately affine equivariant Orthogonalized Gnanadesikan-Kettenring (OGK) estimate presented by[5] estimates when the dimension of data set exceeds number of samples.

Considering the fact that median is a robust measure of centrality, there are several robust location estimates depending on component wise median and multivariate spatial median. An overview about multidimensional median is given[5]. Orthogonal equivariant versions of multivariate medians are suggested in[7]. Other median location estimators are halfspace median by[8], Oja median introduced[9] Simplicial depth median developed by[10]. Transformation- retransformation median established by[11] and practical affine equivariant multivariate median[12] More recently techniques for robust covariance matrix estimation based on different sign and rank concepts are proposed by[13] sign and rank covariance matrices. Function form for Geometric median using stochastic gradient algorithm suggested and applied in robust estimation[14]. In this article, a simple method for multivariate outlier detection based on robust Mahalanobis distance (RMD) is proposed. RMD is estimated based on robust estimator of location vector and covariance matrix. Spatial Median considered as an initial location estimate in iterative procedure for outlier detection. Bivariate location measure which minimises absolute distance is explained in[15] and studied the uses o the defined spatial median. Asymptotic elliptical properties are investigated by[16].

Median absolute difference from median (MAD) is becomes a special case of a robust measure of covariance between X and Y established by[17]. As for robust estimate of initial covariance matrix, spatial median is used instead of coordinate wise median. Not only the estimates but also suitable cut offs for separating outlier are important. RMD does not necessarily fit a chi-squared distribution except for Gaussian distribution. An adjusted threshold approach is adapted in order to decide whether an observation is outlier. Properties of proposed method are discussed by means of simulation studies. The simulations are done using statistical software R-programming. Sufficient graphical plots to enhance efficiency of the modified method are given.

2. MULTIVARIATE OUTLIER DETECTION

The term dispersion contained with two aspects, viz. Variances of the responses and inter-correlations among the responses. For some reasons it may be desirable to consider the aspects separately. An estimate of same behaviour and generalises the separate estimation is suggested. At the same time minimizes limitations of case wise median for estimating multivariate location. A simple estimate for estimating covariance between pair of random variables X, Y based on spatial median.

$$\tilde{\mathrm{C}}(\mathrm{X},\mathrm{Y}) = \mathrm{med}\left(\left(\mathrm{X} - \widehat{\mathrm{M}}(\mathrm{X})\right)\left(\mathrm{Y} - \widehat{\mathrm{M}}(\mathrm{Y})\right)\right) \quad (1)$$

Next extend this definition in to multivariate setting. Assuming that $\mathbf{X}= (\mathbf{x}_1, \ldots, \mathbf{x}_n)^t$ be an $n \times p$ data matrix with $\mathbf{x_i}=(x_{i1}, \ldots, x_{ip})^t$. Hence n stands for number of observations and p stands for the dimension of variable.

$$\hat{\mathrm{C}}(\mathbf{x}_\mathrm{j}\,\mathbf{x}_\mathrm{k}) = med\left(\left(\mathrm{x_{ij}} - \widehat{\mathrm{M}}(\mathbf{x}_\mathrm{j})\right)\left(\mathrm{x_{ik}} - \widehat{\mathrm{M}}(\mathbf{x}_\mathrm{k})\right)\right);\ \mathrm{j,k} = 1,2 \ldots p\ , \quad (2)$$

where *med* indicate median and $\widehat{\mathrm{M}}$ is the any orthogonally equivariant robust estimate of multivariate location. The estimate $\hat{\mathrm{C}}(\mathrm{X})$ represents single element in dispersion matrix, arrange each single estimate in ij[th] order gives complete matrix estimate. Note that regular covariance corresponds to mean instead of median. Also note that this estimate turn out to an alternative estimate for median based covariance termed as comedian[17] when location is coordinate wise median.

However, resulting matrix is symmetric but not necessarily positive definite for slightly higher dimensions. To overcome lack of positive definite, propose a modification in (2) for nearest positive definite matrix and based on the algorithm formulated[18] to real symmetric matrix.

It is needed to define an estimator for location to covariance estimate, here consider multivariate version of univariate median called spatial median[15]. The spatial median for multivariate random variable X is defined as.

$$\widehat{\mathrm{M}}(\mathrm{X}) = \underset{m \in\, \mathbb{R}^p}{argmin} \sum_{i=1}^{n} \| x_i - m \| \quad (3)$$

Where, $|\mathrm{X}|$ is euclidean norm. It has asymptotic breakdown point 0.5 like univariate median and orthogonally affine equivariant.

The main focus is on detecting multivariate outlier using robust Mahalanobis distance defined as

$$\mathrm{RMD_i}(\widehat{\mathrm{M}}(\mathrm{X}), \hat{\mathrm{C}}(\mathrm{X})) = \left(\mathrm{x_i} - \widehat{\mathrm{M}}(\mathrm{X})\right)\ \hat{\mathrm{C}}(\mathrm{X})^{-1}\left(\mathrm{x_i} - \widehat{\mathrm{M}}(\mathrm{X})\right)^T;\ i = 1,2, \ldots n \quad (4)$$

Where $\hat{\mathrm{C}}(\mathrm{X})$ and $\widehat{\mathrm{M}}(\mathrm{X})$ are defined in (2) and (3). An adjusted threshold is adapted to ensure the effectiveness of outlier detection is define as

$$cv = \frac{Gamma_{0.99}(\frac{p}{2}, p)\ med\left(\mathrm{RMD_i}(\hat{\mu}(\mathrm{X}), \hat{\mathrm{C}}(\mathrm{X}))\right)}{Beta_{0.5}(p,p)} \quad (5)$$

Where $Gamma_{0.99}(\frac{p}{2}, p)$ and $Beta_{0.5}(p,p)$ are 0.99[th] and 0.5[th] quantile of gamma and beta distributions respectively. The weight function of i[th] (i=1,2,..*n*)observation is then defined according to any observation with $\mathrm{RMD_i}\left(\hat{\mu}(\mathrm{X}), \hat{\mathrm{C}}(\mathrm{X})\right) > cv$ can be consider as an outlier. The robust estimate of location and dispersion can be calculates for the weight function. Reweighting can be done to increase the efficiency of the outlier detection procedure.

3. SIMULATION STUDIES

During the empirical study hundred of samples each of size 100 were simulated for studying chances of correct outlier detection. There are mainly two types of wrong outlier detection or misidentification of outliers one of them is masking i.e. not detecting the real outliers and the other one is swamping i.e. detect the good observations (inliers) in the sense of outliers. To check whether the proposed method has a wrong use in terms of outlier detection some empirical measures from the simulated data set are calculated. Usually, these values are termed as success rate (SR) represents rate of complete detection of true outlier from the data set. Similarly failure rate (FR) represents rate of incorrect detection of inlier observations as outliers these measures ensure the good performance of the method.

Traditionally, simulation studies consist of generation of multivariate data of same characteristic or same distribution and insert some outliers (i.e. data follows different distribution or similar distribution with different parameters). Such method is called Tukey-Huber contamination model (THCM). THCM mechanism is formed as follows:

$$\mathrm{H} = (1 - \mathrm{t})\mathrm{F} + \mathrm{tQ} \tag{16}$$

where F is a central parametric distribution such as multivariate normal $\mathrm{N}_p(\mu, \Sigma)$, Q is an undefined outlier generating distribution and *t* is the percentage of contamination. On the other hand, if the is clean (data without outlier) sample mean and sample covariance are most efficient estimators for Multivariate Normal distribution. Empirical investigation in contaminated data set begin with generation of Multivariate standard Normal data of size *n*=100 and dimension *p* (4, 6, 10, 12, 15); $\mathrm{N}_p(0, \mathrm{I})$, were I stands for identity matrix . In order to create real life contaminated situation *t*% of contaminated sample specifically $\mathrm{N}_p(\delta \mathrm{u}, \lambda \mathrm{I})$ were, $\mathrm{u} = (1, \ldots 1)^{\mathrm{T}}$. Different choices of t, δ and λ (i.e. *t* = 10, 20, 25, δ=6, 10 and λ=0.01, 0.1, 0.25) are conducted for the study, the results were arranged in tables. Simulation is repeated for 100 times and estimated the rate of successes and failure in outlier detection. Table 1-2 shows rate of success in detecting true outliers and failures in the detection, here the central distribution is $\mathrm{N}_p(0, \mathrm{I})$ and the outlier part is $N_p(\delta \mathrm{u}, \lambda \mathrm{I})$. Different percentages of contaminations are tested for *t*=10, 20 and 25, the method is succeeded in detecting all the outliers in almost all the case in different dimensions.

Table 1: Rate of success and failures of proposed outlier detection method for n=100 with contamination Normal distribution: δ =6.

p	λ =0.01						λ =0.1						λ =0.25					
	t=10		*t*=20		*t*=25		*t*=10		*t*=20		*t*=25		*t*=10		*t*=20		*t*=25	
	SR	FR	SR	FR	SR	FR	SR	FR	SR	FR	SR	FR	SR	FR	SR	FR	SR	FR
4	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
6	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
10	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
12	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
15	0.9	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0

Table 1: Rate of success and failures of proposed outlier detection method for n=100 with contamination Normal distribution: δ =10.

p	λ =0.01						λ =0.1						λ =0.25					
	t=10		t=20		t=25		t=10		t=20		t=25		t=10		t=20		t=25	
	SR	FR	SR	FR	SR	FR	SR	FR	SR	FR	SR	FR	SR	FR	SR	FR	SR	FR
4	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
6	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
10	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
12	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
15	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0

Figure 1: Index plot for Bushfire data

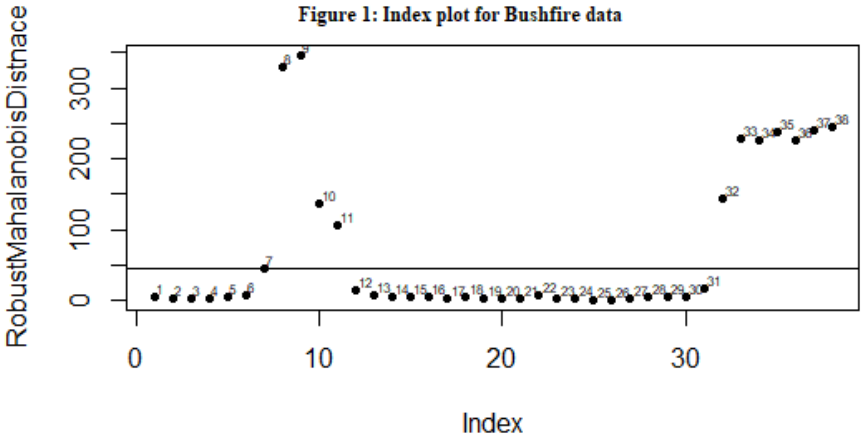

Figure 2 Index plot for Crime data

4. ILLUSTRATION OF EXAMPLE

The first illustration consists of 38 observations on 5 measurements of Bushfire data studied by[19]. Collected the data set for the purpose of examine Bushfire scars[20]. Experimented

by establishing Stahel- Donoho estimators with Huber weight function and modified median absolute deviation[19]. They found that the observations 7-11 and 32-38 are outliers. From Figure1, it is easy to see that the proposed method identifies the similar extreme observations 7-11 and 32-38 are potential outliers.

Crime data 2016 available in http://ncrb.gov.in/StatPublications contains state wise population and other 6 variables such as violent crimes, crime against women, crime against children, crime committed by juveniles, economic offences and environment related offences. The data considers 36 observations which include 29 states and 7 union territories in India. Proposed multivariate outlier detection procedure is applied in this data set to find feasible outliers. Figure 2 shows extreme observations in the crime data and identifies observations 9, 22, 27 and 34 corresponding states Himachal Pradesh, Rajasthan and Delhi are outlying samples.

5. CONCLUSIONS

Outlier detection and robust estimation perform a crucial part of multivariate analysis. Several authors are proposed estimation method for robust estimation. In this paper gives an attempt to a new methodology for robust dispersion estimation and multivariate outlier detection. Estimation procedures consider a multivariate location rather than case wise measure of central tendency. The study of distributional property of RMD is helped to adjust the suitable cut off for the outlier identification. Simulated standard Gaussian examples are created by THCM model, and the efficiencies of proposed method are investigated by rate of success and failures in true outliers in the given data. Some real data situations are applied to test the possibility of proposed method in real life condition. Adequacy of proposed method is fair in the described situations. Further investigations are needed to improve the results.

REFERENCES

1. Rousseeuw, P. J. "Least median of squares regression." *Journal of the American Statistical Association*, Vo. 79, No. 388, pp. 871–880 (1984).
2. Rousseeuw, P. J and Van Driessen, K. "A fast algorithm for the minimum covariance determinant estimator." *Technometrics*, Vol. 41, No. 3, pp. 212–223 (1999).
3. Rousseeuw, P. J. and Leroy. A.M. Robust regression and outlier detection. *Wiley-Interscience*, New York, (1987).
4. Lopuha¨a, H. "Multivariate τ-estimators for location and scatter." *The Canadian Journal of Statistics*, Vol. 19, No. 3, 1991, pp. 307–321.
5. Small C.G. "A survey of multidimensional medians." *International Statistical Review*, Vo. 58, No. 3, pp. 263–277 (1990).
6. Maronna, R. A. and Zamar, R. "Robust estimates of location and dispersion for high-dimensional data sets." *Technometrics*, Vol. 44, No. 4, pp. 307–317 (2002).
7. Grübel, R. "Orthogonalization of multivariate location estimators: the orthomedian." *The Annals of Statistics*, Vol. 24, No. 4, pp. 1457–1473 (1996).

8. Tukey, J.W. "Mathematics and the picturing of data." *Proceedings of the International Conference of Mathematicians*, Vol. 2, pp. 523–531 (1975).
9. Oja, H. "Descriptive statistics for multivariate distributions." *Statistics and Probability Letters*, Vol. 1, No. 6, pp. 327–332 (1983).
10. Liu, R.Y. "On a notion of data depth based on random simplices." *The Annals of Statistics*, Vol. 18, No. 1, pp. 405–414 (1990).
11. Chakraborty, B. and Chaudhuri, P. "On a transformation and re-transformation technique for constructing an affine equivariant multivariate median." *Proceedings of the American Mathematical Society*, Vol. 124, No. 8, pp. 2539–2547 (1996).
12. Hettmansperger, T. P. and Randles, R.H. "A practical affine equivariant multivariate median." *Biometrika*, Vol. 89, No.4, pp. 851–860 (2002).
13. Visuri, S., Koivunen, V. and Oja, H. "Sign and rank covariance matrices." *Journal of Statistical Planning and Inference*, Vol. 91, No. 2, pp. 557–575 (2000).
14. Cardot, H., Cénac P. and Zitt P. A. "Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm." *Bernoulli*, Vol. 1, No. 1, pp. 18–43 (2013).
15. Brown, B. M. "Statistical uses of the spatial median." *Journal of the Royal Statistical Society*, Series B 45 (1), 25-30 (1983).
16. Andrew Magyar, David E. Tayler. "The asymptotic efficiency of spatial median for elliptically symmetric distributions." *Sankhya B*, Vol. 73, No. 2, pp.165-192 (2011).
17. Falk, M. "On MAD and Comedians." *Annals of the Institute of Statistical Mathematics*, Vol. 49, No. 4, pp. 615–644 (1997).
18. Higam, N.J. "Computing a nearest symmetric positive semi definite matrix." *Linear Algebra and its Applications*, Vol. 103, pp. 103-118 (1988).
19. Maronna, R. A. and Yohai, V. J. "The Behavior of the Stahel-Donoho Robust Multivariate Estimator." *Journal of the American Statistical Association*, Vol. 90, No. 429, pp. 330-341 (1995).
20. Campbell, N. A. "Robust Bushfire mapping using NOAA AVHRR data.", *Technical Report, CSIRO*, North Ryde, Australia (1989).