Journal of Computer and Mathematical Sciences, *Vol.9(12),1847-1850 December 2018* *ISSN 0976-5727 (Print)*
(An International Research Journal), *www.compmath-journal.org* *ISSN 2319-8133 (Online)*

Location Model for Mixed Data Using Winsorization with Comedian

LathaV[1*] and P Rajalakshmi[2]

[1]Department of Statistics,
Jyoti Nivas College, Autonomous, Bangalore, INDIA.
[2]Professor (Retd.), Department of Statistics,
Bangalore University, Bangalore, INDIA.
*Corresponding author email: latha.v.1968@gmail.com

(Received on: November 29, 2018)

ABSTRACT

Location model is used for obtaining classification rules for mixed data. It does not perform well when outliers are present in the data. A method has been proposed to obtain robust classification rules when outliers are present in mixed data. In this method outliers are identified and trimmed values are obtained by winsorization using comedian, so that the information from extreme values can be effectively used for computing location and scatter measures in the construction of robust classification rule.

Keywords: Location model, mixed data, winsorization, comedian, breaking point.

INTRODUCTION

Discriminant analysis has been used as a measure of classification for continuous variables, categorical variables and for mixtures of continuous and categorical variables. In literature we can find discrimination procedures using parametric, semi-parametric and non parametric approaches. Kernel based approaches are non-parametric, methods like logistic discrimination are semi-parametric and linear discriminant analysis is a parametric approach based on continuous variables. The location model by Olkin and Tate (1975) was used by Krzanowski (1975) for constructing classification rules for mixed data.

Location model does not perform well in the presence of outliers since the location and scatter estimates are affected by outliers. Hence robust methods are used to reduce the influence of outliers. Currently the problem of outliers is addressed through winsorization so

that information from outlier units are not lost completely and some information is retained by replacing outlier values with the corresponding trimmed values.

In this paper an attempt is made to identify outliers using the comedian method and then use trimmed values in their place to retain as much information as possible about the variables under study.

It is well known that outliers affect the process of estimation and inference. The measures of location and scatter are unduly affected by outliers. Hence there is a need to study their influence and identify methods to reduce their effect or eliminate them from the data sets. This is a challenging aspect of data analysis. Various methods and techniques have been devised to detect multivariate outliers. Some of these methods are distance based. Mahalanobis squared distance is one such measure used in multivariate settings. A large value of this measure indicates that the corresponding value is an outlier. The problem of 'masking' and 'swamping' also exist in the sense that there are outliers who have a very small value for the Mahalanobis distance and a large value of the distance measure need not necessarily indicate an outlier. Hence the classical location and dispersion estimates are not robust to outliers. Therefore there isa need to tackle this problem by using robust distances which are obtained by replacing the classic estimates by the robust ones. Some of the robust estimators proposed are the affine equivariant M estimators proposed by Maronna (1976), Stahel Donoho estimators (1982) which are the weighted mean vector and the covariance matric with weights depending on the outlyingness of an observation, Minimum covariance determinant estimator by Rosseeuw (1984), a fast outlier detection procedure proposed by Pena and Prieto (2001) using the direction of the projections that maximise and minimize the coefficient of kurtosis of the projected data, the orthogonalised Gnanadesikan-Kettenring estimator proposed by Maronna and Zamar (2002) to obtain affine equivariant robust scatter matrices beginning with any pairwise robust scatter matrix, which performs well under high collinearity.

A method to detect multivariate outliers has been proposed by Sajesh and Srinivasan (2012) using the measure Comedian defined by Falk (1997). This method can detect a large number of outliers. Falk introduced a dependence measure called the comedian which is a robust measure of the covariance between random variables X and Y. For any two random variables X and Y, comedian is defined as,

$COM(X,Y)= med((X-med(X))(Y-med(Y)))$

where med(X) and med(Y) are the medians of X and Y respectively. It is equal to the square of the median absolute deviation (MAD) when X=Y and has the highest breakdown point. Further COM(X,Y) always exists, is symmetric and location and scale invariant. An alternative to the coefficient of correlation based on the median is called the correlation median and is given by

$\delta(X,Y)= COM(X,Y)/(MAD(X)MAD(Y)$

where MAD(X) and MAD(Y) are the median absolute deviations of X and Y respectively. Falk proposed this measure as a measure of dependence.

The method proposed by Sajesh and Srinivasan for outlier detection is as follows-
Let X be a nxp matrix with $\mathbf{x_j}$ j=1,2…p as the columns and $\mathbf{x_i}$', i=1,2..n as the rows.

The comedian matrix is COM(X) =COM(X_i,X_j)I,j=1,2,…p and the multivariate correlation median matrix is δ(X)= DCOM(X)D' where D is a diagonal matrix with diagonal elements being the reciprocal of MAD(X_i), i=1,2...p. As the comedian matrix is not positive definite, the following steps are used so that the estimators obtained are robust.

The eigen values λ_i and eigen vectors e_j of the scatter matrix such that δ(X)=EΛE' where E is the matrix with columns as ej's and Λ=diag(λ_1, λ_2,… λp). Then define $Q=D(X)^{-1}E$ where D is as defined above. Let $\mathbf{z}_i= Q^{-1}x_i$, i=1,2…n where z_i' is the i^{th} row of the orthogonal matrix Z. Then the robust location and scatter estimates are

m(X)= QF where F=(med($\mathbf{z}_1$),med($\mathbf{z}_2$),…med(z_j))

S(X)= QΓQ' where Γ=diag(t_1^2, t_2^2,…t_p^2) where t_j= MAD(z_j), j=1,2,…p

The robust Mahalanobis distance is given by

RD(x_i, m)= rd_i= (x_i-m)'S^{-1}(x_i-m) where m and S are as defined above.

The cutoff value for identifying outliers is given by

$$cv= \frac{1.4826(\chi^2_{p(0.95)})}{\chi^2_{p(0.5)}} med(rd1, rd2, \dots rdn)$$

If any RD(x_i, m)>cv, then x_i is an outlier. The expression for cv is obtained following Maronna and Zamar(2002) and holds for non-normal original data.

Consider a set of n_1 observations from group π_1 and another set of n_2 observations from a second group π_2. A classification rule is required to be set up based on a vector **x** of b binary variables and vector **y** of c continuous variables observed on every individual. The binary variables are multinomial with 2^b states and each binary structure represents a distinct multinomial cell with $m=1+\sum_1^b x_q 2^{q-1}$. The vector y of c continuous variables follows multinomial distribution with mean μ_{im} in cell m of π_i and the common covariance matrix Σ. p_{im} is the probability of obtaining an observation in cell m of π_i. The allocation rule to classify a new observation requires the determination, from **x**, of the cell into which this observation falls. If the cell is assumed to be 'm' then the observation is classified into group π_1 if

$(\mu_{1m}- \mu_{2m})' \Sigma^{-1}(y-\frac{1}{2}(\mu_{1m}+ \mu_{2m})) \geq \log(p_{2m}/p_{1m}) +\log a$

Else it is assigned to group π_2.

'a' is a constant based on misclassification rates and prior probabilities for the two groups and it approaches zero when the costs and prior probabilities are equal for both the groups. The values of the unknowns (μ_{im}, p_{im}),i=1,2, and Σ are estimated using information obtained from the samples.

Estimation of parameters is done using the trimmed values obtained after the winsorization process. The given observations are arranged in ascending order to identify outliers and the outliers are replaced by the corresponding trimmed values. This is done by replacing the outliers less than the smallest value by the smallest value retained and the outliers greater than the largest value by the largest value retained. Hence the winsorized sample is given by

$$W_{ij} = \begin{cases} x_{(i_1+1)j,} & if \quad x_{ij} \le x_{(i_1+1)j,} \\ x_{ij} & if \ x_{(i+1)j,} \le x_{ij} \le x_{(n-i_2)j,} \\ x_{(n_j-i_2)j,} & if \ x_{ij} \ge x_{(n-i_2)j,} \end{cases} \qquad i=1,2..n;\ j=1,2\ldots p$$

Then the estimate of the winsorized location measure is m(W)= QF where F=(med($\mathbf{w}_1$),med($\mathbf{w}_2$),…med($\mathbf{w}_p$)) and the scatter estimate is given by S(W)= QΓQ' where where Γ=diag(t_1^2, t_2^2,…t_p^2) where t_j= MAD($\mathbf{w}_j$), j=1,2,…p

The classification rule is then obtained by replacing the location and scatter measures by the corresponding winsorized estimates. The classification rule so obtained is robust and the estimators are constructed using comedian which has a high breakdown value of 0.5. A comparison of this method with available methods is being studied using simulated and real data to examine its efficiency.

REFERENCES

1. Anderson T.W, An Introduction to Multivariate Statistics, New York, Wiley (1968).
2. Devlin S.J, Gnanadesikan R and Kettenring J.R, Robust estimation of dispersion matrices and principal components, *Journal of the American Statistical Association*, 76:354-362 (1981)
3. Falk. M, On MAD and Comedians, 1997, *Ann. Inst. Statist. Math*., 615-644 (1997).
4. Kokic P.N, Bell P.A, Optimal winsorizing cutoffs for a stratified finite population estimator, *Journal of Official Statistics*, 10,419-435 (1994).
5. Krzanowski W.J, Discrimination and classification using both binary and continuous variables, *Journal of the American Statistical Association*,70,782-790 (1975).
6. Krzanowski W.J, Some linear transformations for mixtures of binary and continuous variables with particular reference to linear discriminant analysis, *Biometrika*, 66,33-39 (1979).
7. Lachenbruch P.A and Goldstein M, Discriminant Analysis, *Biometrics*, 35, 69-85 (1979).
8. Maronna R.A and Zamar R. H, Robust estimates of location and dispersionfor high dimensional data sets, *Technometrics*, 44:307-317 (2002).
9. Martinoz F.C, Haziza D and Beaumont J.F, A method of determining the winsorization threshold, with an application to domain estimation, *Survey Methodology*, 41-1:57-77 (2015).
10. Olkin I , Tate R.F, Multivariate correlation models with mixed, discrete and continuous variables, *Annals of Mathematical Statistics*, 32,448-465 (1961).
11. Pena D and Preito F.J, Multivariate outlier detection and robust covariance matrix estimation, *Technometrics*, 43:286-300 (2001).
12. Rousseeuw P and K. Van Driesen, A fast algorithm for the minimum covariance determinant estimator, *Technometrics*, 41:212-223 (1999).
13. Sajesh T.A and Srinivasan M.R, Outlier detection for high dimensional data using the Comedian approach, *Journal of Statistical Computation and Simulation*, 82:5, 745-757 (2012).