BEST: International Journal of Management, Information Technology and Engineering (BEST: IJMITE)
ISSN (P): 2348-0513, ISSN (E): 2454-471X
Vol. 5, Issue 02, Feb 2017, 23-30
© BEST Journals

COMPARATIVE STUDY OF ROBUST ESTIMATORS OF LOCATION USING MAHALANOBIS DEPTH

OKEKE, EVELYN NKIRUKA & OKEKE, JOSEPH UCHENNA
Department of Mathematics and Statistics, Federal University Wukari, Nigeria

ABSTRACT

The sample mean is not a robust estimator of location and thus can produce misleading information when the data are not normally distributed. In this paper we studied three different robust estimators of location and compared their performance in providing better Mahalanobis depth with that of median (another robust measure) and mean. Comparing the performance of the five different estimators of location we have that M-estimator is more efficient than the other four methods in providing better depth for the significant test of equal population mean vectors when the data is free from outlier. In the presence of outlier median and trimmed mean seems to be better than the others. The whole five measures produced inconsistent result when the data are normally distributed.

KEYWORDS: Trimmed Mean, MCD Estimator, M-Estimator of Mean, Mahalanobis Depth, Kruskal-Wallis Test

INTRODUCTION

Measures that characterize a distribution, such as measures of location and scale, are said to be robust if any little change in a distribution have relative little effect on their values. The population mean (μ) and standard deviation (σ) as well as the sample mean ($\bar{X}$) and sample standard deviation (S^2) are not robust. Robustness is particularly important when there is a possibility that our data set contains “contaminated” or “corrupt” data. If we take the mean as an estimator for instance, outlying points carry more “weight” than points near the mean. Deleting the outlying point would have a greater impact on the location of the mean than deleting a point in the dense region. A single point is enough to greatly influence the mean of a data set. In the contrary, 50% of a data set must be moved to infinity in order to force the median to do the same. This suggests a robust estimator of location. High breakdown point of the median is a reason for choosing it over the mean as an estimator of location especially when the data is contaminated.

The finite breakdown point of a statistic, which is a technical device for judging an estimator is the smallest proportion of observations that, when altered sufficiently, can render the statistic meaningless. More precisely the finite sample breakdown point of an estimator refers to the smallest proportion of observations which when altered can cause the value of the statistic to be arbitrary large or small. The breakdown point of the mean is $\frac{1}{n}$ in all dimensions, which is one of the smallest among other estimators of location. Apart from mean and median, we have other estimator of location like trimmed mean which is based on the predetermined amount of trimming required in a data set. Also we have minimum covariance estimator of the mean and M-estimator of location. M-estimator approach, unlike trimmed mean, determines empirically the amount of trimming that is required of a data set.

A quantity such as the population mean which characterizes a distribution is said to be a measure of location if it satisfies the following four conditions, and a fifth is sometimes added. To describe these conditions, let X be a random variable with distribution F, and let $\theta(X)$ be some descriptive measure of F. Then $\theta(X)$ is said to be a measure of location if for any constant a or b

1. $\theta(X+b)=\theta(X)+b$
2. $\theta(-X)=-\theta(X)$
3. $X \geq 0 \; implies \; \theta(X) \geq 0$ 1
4. $\theta(aX)=a\theta(X)$
5. Let $F_x(x)=P(X\leq x)$ and $F_y(x)=P(Y\leq x)$ be the distribution corresponding to the random variable X and Y. Then X is said to be stochastically larger than Y if for any x, $F_x(x)\leq F_y(x)$ with strict inequality for some x. If all the quantiles of X are greater that the corresponding quantiles of Y, then X is stochastically larger than Y. Bickel and Lehmann argue (1975) that if X is stochastically larger than Y, then it should be the case that $\theta(X)\geq\theta(Y)$ if θ is to be qualified as a measure of location. The population mean has this property.

Each estimator of location works well under certain condition of data. For example, if sampling is from a light-tailed distribution or even a normal distribution, it might be desirable to trim very few observations or none at all. If a distribution is skewed to the right, a natural reaction is to trim more observation from the right versus the left tail of the empirical distribution. When the distribution is normal, mean as a location may not produce misleading result. Then if the distribution is skewed M-estimator of location may produce a better result than the mean. In this paper we wish to test the efficiency of different estimators of location in testing for the significance difference between the mean vectors of different populations through Mahalanobis depth which provides the ranks we used in Kruslal-Wallis non-parametric test statistic for such analyses. We will as well, if possible, state the estimator that is better used for different data distributions.

LOCATION ESTIMATORS

In this section we are going to study different robust estimator of location and their properties. Among the measures of location we have mean, median, mode, trimmed mean, geometric mean, MCD location estimator, M-estimator of location etc. Robust estimators include trimmed mean, MCD location estimator, and M-estimator of location. In this section we decided to study the following estimators:

The Sample Trimmed Mean

The standard error of the sample mean can be relatively larger when sampling from a heavy-tailed distribution. Sample mean estimator is a non-robust measure of location, μ. A more robust estimator, the sample trimmed mean has come as remedy to the problems of sample mean. The sample trimmed mean, which estimates μ, is computed as follows. Let $x_1, x_2, \dots, x_n$ be a random sample and let $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ be the observations arranged in ascending order of magnitude. The ith largest value, $x_{(i)}$, is called the ith order statistics. Supposed the desired amount of trimming has been chosen to be γ, $0\leq\gamma\leq 0.5$. Let $k=(\gamma n)$, where (γn) is the value of γn rounded down to the nearest integer. For example, (10.9)=10. The sample trimmed mean is computed by removing the k largest and k smallest observations and averaging the values that remains. In symbols, the sample trimmed mean is

$$\bar{X}_{tm}=\frac{x_{(k+1)}+\cdots+x_{n-k}}{n-2k} \tag{2}$$

(Wilcox 2004)

The empirical distribution is trimmed in a manner consistent with how the probability density function was trimmed when defining μ. for the trimmed mean to have any practical importance, a value for γ must be chosen and this depends on the number of contamination that is the data. For various reasons, a good choice for general use is $\gamma = 0.2$. If γ is too small, the standard error of the trimmed mean, $\sqrt{VAR(\bar{X}_{tm})}$, can be drastically inflated by outliers or sampling from a heavy-tailed distribution. If γ is too large, the standard error can be relatively large compared to the standard error of the sample mean when sampling from a normal distribution. An empirical investigation based on data from actual studies suggests that the optimal amount of trimming, in terms of minimizing the standard error, is usually between 0 and 0.25 (see, e.g., Hill and Dixon, 1982). In this we are going to consider 0.01 and 0.02 amount of trimming.

Minimum Covariance Determinant (MCD) Estimator

Hubert and Van Driessen (2004) used the re-weighted MCD estimator of multivariate location and scale because of its good statistical properties and FAST MCD algorithm which provides an efficient algorithm of computing estimates for large data set. For the X sample the MCD estimator is defined as the mean $\hat{\mu}_{x,0}$ and the covariance matrix $S_{x,0}$ of h_x observations out of n_x observations whose covariance matrix has the lowest determinant. The quantity should be larger than $[(n_x\text{-}p+1) =2]$ and $n_x\text{-}h_x$ should be smaller than the number of outliers in the X population. With this choice the MCD attain its maximum breakdown value $[(n_x\text{-}p+1) =2] = 50\%$.The breakdown value of an estimator is defined as the largest percentage of contamination it can withstand. If one suspects less than 25% contamination in the X sample, it is advised to use $h_x \approx 0.75n_x$ as this yields higher finite sample efficiency. Based on the initial estimates $\hat{\mu}_{x,0}$ and $S_{x,0}$, one computes for each observation x_i, its (preliminary) robust distance.

$$RD^0 = \sqrt{(x_i - \mu_{X,0}) S_{x,0}^{-1} (x_i - \mu_{X,0})} \quad i=1, 2,\ldots, n_{x\,3} \tag{3}$$

Then assign weight 1 to x_i if $RD^0 \leq \sqrt{\chi^2_{p,0.975}}$ and weight 0 otherwise. The re-weighed MCD estimator is then obtained as the mean $\hat{\mu}_x$ MCD and the covariance matrix $\hat{\Sigma}_x$ MCD of those observations with weight 1. This re-weighing step increases the finite sample efficiency of the MCD estimator considerably, whereas the breakdown value remains the same. This can be used to flag off outlier and so can be used to detect outliers.

Through extensive simulation studies we observed that h_x observations out of n_x observations whose covariance matrix has the lowest determinant were those observations with smaller Mahalanobis distance. The first thing we did here was to calculate the Mahalanobis distance of each data point. With the assistance of this distance we were able to get dataset whose estimators attain their maximum breakdown value in a few numbers of iterations. For more details (see Okonkwo, Okeke, and Nwabueze 2014).

M-Estimator for Location

The trimmed mean is based on a predetermined amount of trimming. That is, you first specify the amount of trimming that is desired, after which the sample trimmed mean, $\bar{X}_{tm}$, can be computed. Another approach is to determine empirically the amount of trimming. For example, if sampling is from a light-tailed distribution, or even a normal distribution, or even a normal distribution, it might be desirable to trim very few observations or none at all. If the distribution is skewed to the right, a natural reaction is to trim more observation from the right of the empirical distribution. In essence this is what M-estimator does. When searching for a measure of location, one strategy is to use some value, say, c, that is in some sense close, on average, to all the possible values of the variable X. One way of quantifying how close a value c is from all possible of X is in terms of its expected squared distance. In symbols,

$$E(X-c)^2 \tag{4}$$

Represents the expected squared distance from c. If c is intended to characterize the typical subject or thing under study, a natural approach is to use the value c that minimizes$E(X-c)^2$. Taking $E(X-c)^2$ as a function of c, the value of c minimizing this function is obtained by differentiating, setting the result equal to 0, and solving for c. That is, c is given by the equation

$$E(X-c)=0$$

so$c=\mu$. In other words, μ is the closest point to all possible values of X in terms of expected squared distance? But μ is not robust and $E(X-c)^2$ gives an inordinate amount of weight to values of X that are far from c. In other words, the function $(x-c)^2$ increases rapidly as x moves away from c. This approach described for deriving a measure of location can be improved by considering a class of function for measuring the distance from a point and then searching for a function within this class that has desirable properties. To this end, let

$$\xi(X-\mu_m) \tag{5}$$

Be some function that measures the distance from μ_m and let φ be its derivatives with respect to μ_m. Attention is restricted to those functions for which$E(\xi(X-\mu_m))$, viewed as a function of μ_m, has a derivative. Taking the derivatives of (5) to be

$$E[\varphi(X-\mu_m)]=0 \tag{6}$$

where the function φ is assumed to be odd, meaning that $\varphi(-x)=-\varphi(x)$ for any x

An M-estimator of location is the value μ_m such that

$$\xi\left(\frac{X-\hat{\mu}_m}{\tau}\right)=(X-\hat{\mu}_m)^2 \tag{7}$$

yielding $\hat{\mu}_m=\bar{X}$, which is optimal under normality. The problem with this function is that it can cause practical problem when sampling from non-normal distributions for which extreme values can occur. But because this choice of ξ is optimal under normality, a natural strategy is to search for some approximation of (7) that gives nearly the same results when sampling from a normal distribution. In particular, consider functions that are identical to (7) provided X_i is not too extreme. For simplicity, consider a standard normal distribution, and take τ to be σ, the standard deviation which in this case is 1. Then the optimal choice of $is\ (X-\hat{\mu}_m)^2$. Suppose instead that ξ is taken to be

$$\xi(X-\hat{\mu}_m)=\begin{cases}-2K(X-\hat{\mu}_m)\ if\ x<-K\\(X-\hat{\mu}_m)^2\ if-K\le x\le K\\2K(X-\hat{\mu}_m),if\ x>K\end{cases} \tag{8}$$

Where K (some constant to be determined) is a tuning constant that determines the degree of robustness. Thus when sampling from a normal distribution, the optimal choice for ξ is being used provided an observation is not extreme, meaning that its value does not exceed K or is not less than –K. If it is extreme, ξ becomes a linear function, rather than a quadratic function, and this linear function increases less rapidly than (7) , so extreme values are having less of influence on $\hat{\mu}_m$. Now $\hat{\mu}_m$ is the value minimizing$\sum \xi \left(\frac{X-\hat{\mu}_m}{\tau}\right)$. Taking the derivatives of this equation, with ξ given by (8), and setting result equal to zero, $\hat{\mu}_m$ is determined by

$$2\sum \varphi(X-\mu_m) = 0 \tag{9}$$

Where

$$\varphi(x) = \max\left[-K, \min(K, x)\right]$$

is Huber's φ. There remains how to get K. A common strategy is to choose the constant K so that when estimating $\hat{\mu}_m$, the estimator has reasonably high efficiency when sampling from a normal distribution, but continues to have efficiency when sampling from a heavy-tailed distribution instead. A common choice is $K = 1.28$, the 0.9 quantile of the standard normal distribution. Other choice might be more optimal, but $K = 1.28$ guards against relatively larger standard errors while sacrificing very little when sampling from a normal distribution. A more efficacious choice might be made based on knowledge about the distribution being sampled.

The computation of M-estimator of location, $\hat{\mu}_m$ is done iteratively, but one step M-estimator is obtained using the formula

$$\hat{\mu}_m = \frac{1.28(MADN)(i_2-i_1)+\sum_{i=i+1}^{n-i_2} X_{(i)}}{n-i_1-i_2} \tag{10}$$

Where MADN $= \frac{\text{MAD}}{\text{z}_{.75}} \approx \frac{\text{MAD}}{0.6745}$ and $MAD = MED\{|X_1 - M|, \ldots, |X_n - M|\}$ with M being the usual sample median. i_1 is the number of observationX_i for which $\left(\frac{X_i-M}{MADN}\right) < -1.28$ and i_2 the number of observations for which $\left(\frac{X_i-M}{MADN}\right) > 1.28$.

Iterative Method of Finding M-Estimator of Location, $\hat{\mu}_m$.

Set $k = 0, \hat{\mu}_k = M$, the sample median, and choose a value for K. A common choice is $K = 1.28$

Step 1: Set $A = \sum \varphi\left(\frac{X_i-\mu_k}{MADN}\right)$ and calculate it using $\varphi(x) = \begin{cases} x(1-x^2)^2, if\ |x| < 1 \\ 0, if\ |x| \geq 1 \end{cases}$

Step 2: Set $B = \sum \varphi'\left(\frac{X_i-\mu_k}{MADN}\right)$ and calculate it using the derivative of φ which is $\varphi(x) = \begin{cases} 1, if\ -K \leq x \leq K \\ 0, otherwise \end{cases}$

B is just the number of observations X_i satisfying$-K \leq \left(\frac{X_i-\mu_k}{MADN}\right) \leq K$.

Step 3: set $\hat{\mu}_{k+1} = \hat{\mu}_k + \frac{MADN \times A}{B}$ and solve.

Step 4: If $|\hat{\mu}_{k+1} - \hat{\mu}_k| < 0.0001$, stop and set $\hat{\mu}_m = \hat{\mu}_{k+1}$. Otherwise, increase k by one and repeat steps 1 through 4. (Wilcox 2004)

METHODOLOGY

The two main methods of analyses we considered in this study includes;

Mahalanobis Depths

Mahalanobis depth (MD) is obtained from little adjustment of Mahalanobis distance. Recall the mahalanobis distance $d^2 = (y_i - \bar{y})' S^{-1} (y_i - \bar{y})$, if m_x is the vector that measures the location of X in a continuous and affine equivariant way and $C_x$the matrix that measures the scatter of the distribution such that $C_{XA+c} = AC_X A'$ holds for any matrix A of full rank and any c. Then based on these parameters a simple depth fuction called the Mahalanobis depth is constructed as

$$M_D(x) = \left(1 + \|x - \mathrm{m_x}\|^2_{C_x}\right)^{-1} \tag{11}$$

$M_D(x)$ Takes its unique maximum at the center m_x. Mahalanobis depth is continuous on x and in the distribution of X. In particular, with $\mathrm{m_x} = \mathrm{E(X)}$ and $C_x = \Sigma_x$ the moment Mahalanobis distance is given as

$$M_{mD}(x) = [1 + (x - E(X))' \Sigma_x^{-1} (x - E(X))]^{-1}$$

The sample version is

$$MD(x_1, \dots, x_n;\ F_x) = [1 + (x - \bar{x})' S_x^{-1} (x - \bar{x})]^{-1} \tag{12}$$

Where $\bar{x}$ is the mean vector and S_x^{-1} is the empirical covariance matrix

Kruskal –Wallis Test

Kruskal-Wallis (1952) test is a useful tool for testing the equality of k independent populations. Suppose we have k samples of sizes $n_1, n_2, \dots, n_k$, with the total size of all samples $n = \sum_{i=1}^{k} n_i$. Suppose further that the data from all the samples taken are ranked and that the sums of the ranks for the k samples are $R_1, R_2, \dots, R_k$, respectively. If we define the Kruskal-Wallis H test as

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1) \tag{13}$$

then it can be shown that the sampling distribution of H is very nearly a chi-square distribution with k-1 degrees of freedom, provided that $n_1, n_2, \dots, n_k$ are all at least 5.

The H test provides a non-parametric method in the analysis of variance for one-way classification, or one-factor experiments, and generalization can be made.

DISCUSSIONS

MD, just like Mahalanobis distance reveals a lot of information about the distribution of dataset. It has long been in use in various statistical procedures due to its intuitive appeal and mathematical tractability (despite the restriction to elliptical contours that it imposes) (see Serfling 2004). In this our paper we studied the effect of different robust location estimators on the Mahalanobis depth. The Mahalanobis depth provided us with the ranks we used in testing the null hypothesis of equal population mean vectors by its application to Kruskal-Wallis non-parametric test statistic. MD in this case allowed us to convert p-dimensional (where$p \geq 2$) dataset to one-dimensional dataset where a univariate test statistics can be applied. In this study four different simulated datasets that follow hypergeometric distribution were generated using different specifications for analyses. For each data set we ensured that there exists difference within the groups in the set. Out of the four datasets, one is highly affected by outliers. Also two real life data that follow normal distribution were also studied to see how the method works when the data is normal, of course we know that Kruskal-wallis test is not supposed to be used when the data is normally distributed. The first normally distributed dataset is available at http:www.real-statistics.com/multivariate-statistics/hotellings-t-squ… and is on tropical disease characterized by fever, low blood pressure and body ache. A pharmaceutical company who is working on a new drug to treat this type of disease wanted to determine whether the drug is effective. They took a random sample of 20 people treated with the new drug and 18 with a placebo and wanted to determine whether the drug is effective at reducing these three symptoms. The result using Hotel lings T^2 test showed that the test in not significant. The second normally distributed dataset is obtained from Methods of Multivariate Analysis, second edition by Rencher (2002) or http://www.amazon.com/methods-multivariate-Analysis...Rencher/dp/0470178965 and the data contains 2 types of coating for resistance to corrosion on 15 pieces of pipe. Two pipes, one with each type of coating were buried together and left for the same length of time at 15 different locations, providing a natural pairing of the observations. Corrosion for each type of coating was measured on $p = 2$ variable (maximum depth of pit in thousandths of an inch, and number of pits.) The Hotellings T^2 result has it that the two coatings differ in their effect on corrosion.

During the analysis the Mahalanobis depth of each dataset were computed with different robust estimators of location. The computed depths are then used as ranks in Kruskal-Wallis test statistic and results were obtained.

RESULTS

The result of the analyses in Table 1 showed the p-value of each location estimator at different data distribution. The performance of the estimators is assessed by their P-value. Estimator with very low value indicates that the test is highly significant and thus there exist difference in the populations mean vectors.

Table 1: Performance of Six Different Estimators of Location and their P-Values

	Data Distribution					
Location Estimator	**Hypergeometic Without Outlier**	**Hypergeometic Without Outlier**	**Hypergeometic Without Outlier**	**Hypergeometic with ~~Outlier~~**	**Normal**	**Normal**
Mean	0.033	0.023	0.082	0.102	0.520	0.501
Median	0.248	0.041	0.049	0.047	0.694	0.273
Trimmed mean (0.1 trimming)	0.419	0.028	0.049	0.070	0.520	0.483
Trimmed mean (0.2 trimming)	0.419	0.034	0.049	0.031	0.548	0.501
MCD	0.326	0.028	0.041	1.000	0.330	0.397
M-Estimator	0.050	0.023	0.034	0.122	0.633	0.682

CONCLUSIONS

Comparing the performance of the six different estimators of location we have that M-estimator is more efficient than the other five methods in providing better depth for the significant test of equal population mean vectors when the data is free from outlier. In the presence of outlier median and trimmed mean seems to be better than others. The whole six estimators produced inconsistent result when the data are normally distributed of course we know that our method could not work well with normally distributed data.

REFERENCES

1. Bickel, P.J and Lehmann, E.L (1975).Descriptive statistics for nonparametric models II. Location. Annals of Statistics 3, 1045-1069.
2. Kruskal, W.H. and Wallis, W.A. (1952).Use of rank in one criterion variance analysis, J*ournal of the American Statistical Association*, 47, 583-621.
3. Hill, M. and Dixon, W.J. (1982). Robustness in real life: A study of clinical laboratory data. Biometrics 38, 377-396.
4. Hubert, M. and Van Driessen, K. (2004). Fast and robust discriminant analysis. *Comput. Statist. Data Anal.*, 45(2):301-320.
5. Okonkwo, E.N., Okeke, J.U., and Nwabueze, J.C.(2014), A comparative Study of Nonparametric and Robust Linear Discriminant Procedures, *International Journal of Statistics and Systems*, Research Indian Publication, ISSN 0973-2675, 9(1): 61-74. www.ijss.org.
6. Serfling, R. (2004). Nonparametric multivariate descriptive measures based on spatial quantiles. *Journal of Statistical Planning and Inference* 123:259-278
7. Rencher, A.C. (2002). *Method of multivariate analysis,* 2nd Ed., John Wiley and Sons, Canada, 280-281.
8. Wilcox, R. R (2004). *Introduction to robust estimation and hypothesis testing*, Academic Press, New York, 20-22, 46-51