



Robust Procedure for Estimating Multivariate Location and Scatter

Muthukrishnan. R¹ and K.Mahesh²

Department of Statistics, Bharathiar University
Coimbatore-46, Tamilnadu, INDIA

Abstract: The most of the multivariate statistical methods heavily depend on two estimators of location and scatter. The sample mean and covariance which are most commonly used efficient estimators when the data belong to multivariate normal distribution. But they are extremely sensitive to outliers. Many multivariate datasets contain outliers and often they do not show up by simple visual inspection. It is quite easy to detect a single outlier by Mahalanobis distances. But, it is no longer valid for multiple outliers, because the multiple outliers do not necessarily have large Mahalanobis distance. Many robust estimators have been developed to obtain the reliable results while analysing the high dimensional data during the past decades. The main objective of such robust estimator is to find the optimal (h) data points out of n and then estimate the location and scatter. This paper proposes a similar procedure for the selection of optimal data points to estimate multivariate location and scatter along with MCD. The efficiency of the proposed procedure is compared with the most widely used robust estimators such as M, OGK, MVE, MCD and S with the classical estimator MLE in a real data set and the simulating environment in R is also presented.

Keywords: Robust estimator – Location and Scatter – R.

I. Introduction

Many researchers facing difficult to handling multivariate datasets contains vulnerable, the measures of location and scatter are get affected in turn it leads to false interpretation of the data. Robust alternative are used to get estimates of location and scatter if the data contains vulnerable ones. Robust procedures are used only half of the data points to estimate the location and scatter. It is very difficult to get half of the data points, the optimal points to estimate the same. Many procedures are established in the literature and still it is challenging task in order to estimate the estimators which is very close to the true value of the estimate. In a large set of data the selection of optimal points or subset selection is most important task. Once the set of best points selected, then one can estimate the sample mean vector and covariance matrix which is very close to the true estimator. The results based on these estimators will give more reliable one, in the context of multivariate tools such as principle component analysis, clustering, factor analysis, discriminant analysis or classification techniques and etc. Many authors have already studied subset selection procedures, Gnanadesikan and Kettenring (1972), Rousseeuw and Leroy (1987), Davies (1987), Hendrik, Lopuhaa and Rousseeuw (1991), Hawkins (1994), Mia Hubert, Rousseeuw and Stefan van Aelst (2008), Todorov (1992, 2009), Todorov and Filzmoser (2010). All the established procedure uses a half of the data points in order to estimate a multivariate location and scatter.

This research paper provides the algorithm for the selection of the best data points from the given observations, in order to estimate the location and scatter. Section 2 provides a brief description of the most widely used estimators such as Maximum Likelihood Estimator (MLE), Minimum Volume Ellipsoid (MVE), Minimum Covariance Determinant (MCD), S- Estimator, M-estimator, Orthogonalized Gnanadesikan/Kettenring (OGK). The proposed algorithm for Best Points Selection (BPS) with MCD is presented in the section 3. The efficiency of the proposed algorithm along with the other methods is executed in a real and simulating environment involving the multivariate technique, discriminant analysis and the results are discussed in the section 4 and 5 respectively. Section 6 concludes with summary of the results.

II. Classical and robust methods

The classical method namely MLE and the robust procedures MVE, MCD, S, M, OGK are briefly summarized in this section.

A. Maximum Likelihood Estimator (MLE)

The principle of MLE, originally developed by R.A Fisher in 1920. Assuming that the data is drawn from a population whose distribution is multivariate normal, then the optimal estimators for location and dispersion are

found, respectively, as the $p \times 1$ sample mean vector, $m = \frac{\sum_{i=1}^n Z_{y,i}}{n}$, and $p \times p$ sample covariance matrix $C = \frac{\sum_{i=1}^n (Z_{y,i} - m)(Z_{y,i} - m)'}{n-1}$.

These are obviously, mean-based estimators, so any unusual or extreme observation can arbitrarily inflate either of them.

B. M-ESTIMATOR

M-estimator was originally constructed by Huber (1964) to estimate a one-dimensional location parameter. Later, Huber (1967) considered a very general framework in which consistency and asymptotic normality was proved under relatively weak conditions.

Maronna (1976) was the first to define M-estimator for multivariate location and covariance. Huber (1981) extended Maronna's definition by defining M-estimates based upon $x_1, \dots, x_n \in \mathbb{R}^p$ as solution of the simultaneous equations.

$$(1/n) \sum_{i=1}^n v_1(d_i)(x_i - t) = 0 \quad (2.1)$$

$$(1/n) \sum_{i=1}^n \{v_2(d_i^2)(x_i - t)(x_i - t)^T - v_3(d_i)C\} = 0, \quad (2.2)$$

where $d_i = d(x_i - t, C)$ and v_1, v_2 and v_3 are real-valued functions on $[0, \alpha]$.

C. Orthogonalized Gnanadesikan/ Kettenring (OGK)

Gnanadesikan and Kettenring (1972) proposed positive definite and approximately affine equivariant robust scatter matrices starting from any pair wise robust scatter matrix and then applied for robust covariance estimate the resulting of multivariate location and scatter estimates are called OGK, the general procedure for OGK is as follows,

Step 1. Let $m(\cdot)$ and $s(\cdot)$ be robust univariate estimators of location and scale.

Step 2. Construct $y_i = D^{-1}x_i$ for $i = 1, \dots, n$ with $D = \text{diag}(s(X_1), \dots, s(X_p))$.

Step 3. Compute the correlation matrix U of the variable of $Y = (Y_1, \dots, Y_p)$, given by $u_{jk} = 1/4(s(Y_j + Y_k)^2 - s(Y_j - Y_k)^2)$.

Step 4. Compute the matrix E of eigenvectors of U and 1.) Project the data on these eigenvectors, i.e. $V = YE$; 2.) Compute robust variances of $V = (V_1, \dots, V_p)$, i.e. $\wedge = \text{diag}(s^2(V_1), \dots, s^2(V_p))$; 3.) Set the $p \times 1$ vector $\hat{\mu}(Y) = Em$ where $m = (m(V_1), \dots, m(V_p))^T$, and compute the positive definite matrix $\hat{\Sigma}(Y) = E \wedge E^T$.

Step 5. Transform back to X , i.e. $\hat{\mu}_{\text{RAWOGK}} = D\hat{\mu}(Y)$ and $\hat{\Sigma}_{\text{RAWOGK}} = D\hat{\Sigma}(Y)D^T$. Here, $m(\cdot)$ is a weighted mean and $s(\cdot)$ is the τ -scale of Yohai and Zamar (1988). Step 2 makes the estimate scale equivariant, whereas the following steps are kind of principal components that replace the Eigen values of U by robust variances.

D. Minimum Volume Ellipsoid Estimator (MVE)

The Minimum Volume Ellipsoid (MVE) estimator, first proposed by Rousseeuw (1984), has been studied extensively and frequently used in the detection of multivariate outliers. It seeks to find the ellipsoid of minimum volume that covers a subset of at least h data points. Subsets of size h are called half sets because h is often chosen to be just more than half of the data points. The location estimator is the geometrical center of the ellipsoid and the estimator of the variance – covariance matrix is the matrix defining the ellipsoid itself, Valentin Todorov et al. (2009) have discussed the minimum volume ellipsoid estimator searches for the ellipsoid of minimal volume containing at least half of the points in the data set X . Then the location estimate is defined as the center of this ellipsoid and the covariance estimate is provided by its shape.

The search for the approximate solution is made over ellipsoids determined by the covariance matrix of $p+1$ of the data points and by applying a simple but effective improvement of the sub-sampling procedure as described in Maronna et al. (2006). The MVE was the first popular high breakdown point estimator of location and scatter but latter it was replaced by the MCD, mainly because of the availability of an efficient algorithm for its

computations (Rousseeuw and van Driessen (1999)). Recently the MVE gained importance as initial estimator for S estimation.

E. Minimum Covariance Determination Estimators (MCD)

Rousseeuw (1985) introduced the minimum covariance determinant estimator (MCD) method to estimate the mean vector and covariance matrix along with detection of outliers in multidimensional data. The multivariate location and diffusion estimation in high breakdown principles is based on the determinant of the covariance matrix. If the covariance matrix $n \times n$ positive semi-definite matrix, p Eigen values are positive, the determinant of covariance matrix equals the product of Eigen values. Thus, a small value in the determinant reflects some linear patterns in the data. Consider all C_h^n subsets, and compute the determinant of the covariance matrix for each subset. The subset with smallest determinant is used to calculate the usual $p \times 1$ mean vector, and corresponding $p \times p$ covariance matrix, these estimators are called minimum covariance determinant estimators.

F. S-Estimator

S-estimator of μ and Σ were introduced by Davies (1987) and further studied by Lopuhaa (1989). For the data set of p -variate observations $(x_1, x_2, \dots, x_n)$ an S estimate (T, C) is defined as the solution of $\sigma(d_1, d_2, \dots, d_n) = \min$ where $d_i = (x - T)^T C^{-1} (x - T)$ and $\det(C) = 1$. Here $\sigma = \sigma(z)$ is the M-scale estimate of a data set $z = \{z_1, z_2, \dots, z_n\}$ defined as the solution of $\frac{1}{n} \sum \rho(z / \sigma) = \delta$ where ρ is non-decreasing $\rho(0) = 0$ and $\rho(\infty) = 1$ and $\delta \in (0, 1)$. An equivalent definition is to find the vector T and a positive definite symmetric matrix C that maximize $\det(C)$ subject to $\frac{1}{n} \sum \rho(d_i) = b_0$, with the above

d_i and ρ . As shown by Lopuhaa (1989), S estimates have a close connection to the M-estimators and the solution (T, C) is also a solution to equation defining a M-estimator as well as weight sample mean and covariance matrix

$$d_i^{(j)} = [(x_i - T^{(j-1)})^T (C^{(j-1)})^{-1} (x_i - T^{(j-1)})]^{1/2} \quad (2.3)$$

$$T^{(j)} = \frac{\sum w(d_i^{(j)}) x_i}{\sum w(d_i^{(j)})} \quad (2.4)$$

$$C^{(j)} = \frac{\sum w(d_i^{(j)}) (x_i - T^{(j)}) (x_i - T^{(j)})^T}{\sum w(d_i^{(j)})} \quad (2.5)$$

III. Best Points Selection based on MCD (BPS-MCD)

The proposed Best Points Selection (BPS) algorithm based on minimum covariance determinant (MCD) is to find the weighted mean and weighted covariance matrix based on the selected points. The detailed description of this proposed algorithm is as follows:

Let S be a set with $|S| = n$. If $1 \leq k \leq n$, then a k -permutation of S is a sequence $(s_1, s_2, \dots, s_k)$ with no repetitions of elements from S ; so note that such a permutation (k) is more succinctly described as a one-to-one map $\sigma: \{1, 2, \dots, k\} \rightarrow S$. Simply a permutation of S is an n -permutation of s , the number of k -permutation of S by $P(n, k)$.

To specify an n -permutation $(s_1, s_2, \dots, s_n)$ of S , we have the full n choice for S_1 , then only $n-1$ choice for S_2 , $n-2$ choices for s_3 , and so on until finally there is only one choice left for S_n , so by the product rule we have $P(n, n) = n(n-1)(n-2), \dots, 2, 1 = n!$

is the number of permutation of S . Then note that a k -permutation, $(s_1, s_2, \dots, s_k)$, for $1 \leq k < n$, so the number of such k -permutation is

$$P(n, k) = n(n-1)(n-2) \dots (n-k+1) = \frac{n!}{(n-k)!}$$

Hence, the method has k -permutation is $P(n, n) = n!$

$$P(n, k) = \frac{n!}{(n-k)!} ; 1 \leq k < n$$

Therefore,

$$P(n, k) = \begin{cases} n! & ; k = n \\ \frac{n!}{(n-k)!} & ; 1 \leq k < n \end{cases}$$

Here, replace n by $k = \frac{(g_1 + g_2 + \dots + g_n)}{([m + p + 1]/2)}$, then

$$P\left(\frac{(g_1 + g_2 + \dots + g_n)}{([m + p + 1]/2)}, k\right) = \begin{cases} \left(\frac{(g_1 + g_2 + \dots + g_n)}{([m + p + 1]/2)}\right)! & ; k = \frac{(g_1 + g_2 + \dots + g_n)}{([m + p + 1]/2)} \\ \frac{\left(\frac{(g_1 + g_2 + \dots + g_n)}{([m + p + 1]/2)}\right)!}{\left(\frac{(g_1 + g_2 + \dots + g_n)}{([m + p + 1]/2)} - k\right)!} & ; 1 \leq k < \frac{(g_1 + g_2 + \dots + g_n)}{([m + p + 1]/2)} \end{cases}$$

where $h = \left(\frac{(g_1 + g_2 + \dots + g_n)}{([m + p + 1]/2)}\right)$, and then selecting the h points by minimum covariance determinant

estimator (MCD) procedure. The characteristic of distance values or mean values are $(d_1, d_2, \dots, d_k)$ are obtained and such distance, the following optimality criteria can be calculated:

D_A = The minimum of distance values. $D_A = \{\min d_i\}$

D_M = The minimum of the mean values for distance $D_M = \min\{\bar{X} = \sum_{i=1}^N w_i x_i\}$

The proposed BPS algorithm summarized in the following steps.

Step 1: Let $X = (x_1, x_2, \dots, x_m)$ be a set of m points in $\mathcal{R}^p$, let h be a natural number such

that $m/2 \leq h = \frac{(g_1 + g_2 + \dots + g_n)}{([m + p + 1]/2)} < m$.

Step 2: The $p+1$ data points $\{x_1, x_2, \dots, x_{p+1}\}$, that satisfy the two optimality criteria were selected and use to obtain the location and scatter matrix

$$\bar{x} = \frac{1}{p+1} \sum_{i=1}^N w_i x_i, \text{ and } s_{jk} = \frac{1}{p+1} \frac{\sum_{i=1}^N w_i}{\left(\sum_{i=1}^N w_i\right) - \sum_{i=1}^N w_i} \frac{\sum_{i=1}^N w_i (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{2}$$

Step 3: The Mahalanobis distance for all the k observations using $\bar{y}$ and S can be calculated as

$$d_i^2 = (y - \bar{y}) S^{-1} (y - \bar{y})^T,$$

Step 4: The $d_i^2 (i=1, 2, \dots, k)$ is arranged in order of magnitude from the least to the highest. The first p+j (j=2, 3, 4...h-p-1) distances are selected and their corresponding sample units are used to compute the next $\bar{x}$ and s as follows

$$\bar{x} = \frac{1}{p+j} \sum_{i=1}^N w_i x_i \text{ and } s_{jk} = \frac{1}{p+j} \frac{\sum_{i=1}^N w_i}{\left(\sum_{i=1}^N w_i\right) - \sum_{i=1}^N w_i} \frac{\sum_{i=1}^N w_i (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{2}$$

The new set of $\bar{x}$ and s are then used to obtain the Mahalanobis distances for all the observations.

Step 5: Step 3 and 4 are repeated until the number of units selected is $h = \left(\frac{(g_1 + g_2 + \dots + g_n)}{([m + p + 1]/2)}\right)$.

Then the location and scatter can be computed by using,

$$\bar{x}_{(BPS)} = \frac{1}{h} \sum_{i=1}^N w_i x_i, \text{ and } s_{jk(BPS)} = \frac{1}{h} \frac{\sum_{i=1}^N w_i}{\left(\sum_{i=1}^N w_i\right) - \sum_{i=1}^N w_i^2} \sum_{i=1}^N w_i (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k).$$

IV. Numerical Study

The efficiency of the proposed BPS algorithm along with other procedures, is executed in a real data set (Narayan C. Giri (2004), p.469). The considered data set consists of two populations of plants of two distinct varieties of wheat. The measurements for each member of these two populations are x_1 -Plant height (cm), x_2 -Number of effective tillers, x_3 -Length of ear (cm), x_4 -Number of fertile spikelets per 10 ears, x_5 -Number of grains per 10 ears, x_6 -Weight of grains per 10 ears. Assuming that there are six-dimensional normal populations with different unknown mean vectors μ_1, μ_2 and with the same unknown covariance matrix Σ , there are each 27 observations. The estimated mean vectors, covariance matrices, apparent error rate and confusion matrices under various methods are displayed in the tables 4.1- 4.4 respectively.

Table 4.1 - Estimated Mean Vectors

MLE	M	OGK	MVE	MCD	S	BPS
79.06	78.43	78.69	78.47	78.62	78.64	78.57
171.33	169.80	170.13	169.37	170.19	169.46	170.17
10.15	10.14	10.15	10.09	10.13	10.14	10.16
12.96	12.93	12.99	12.89	12.93	12.97	12.91
333.93	336.61	338.30	331.54	335.88	337.80	338.03
14.70	14.57	14.70	14.49	14.63	14.64	14.73

Table 4.2 - Estimated Covariance Matrices

MLE	M	OGK
$\begin{pmatrix} 46.81 & 82.81 & 0.86 & 2.96 & 2.86 & 3.23 \\ 82.81 & 505.88 & -0.54 & 2.39 & -155.31 & -5.81 \\ 0.86 & -0.52 & 0.38 & 0.37 & 21.93 & 1.15 \\ 2.96 & 2.39 & 0.37 & 0.74 & 19.15 & 1.28 \\ 2.86 & -155.31 & 21.93 & 19.15 & 2044.78 & 77.82 \\ 3.23 & -5.81 & 1.15 & 1.28 & 77.82 & 4.61 \end{pmatrix}$	$\begin{pmatrix} 59.16 & 95.09 & 1.39 & 4.66 & 27.83 & 5.25 \\ 95.09 & 593.20 & 0.91 & 4.79 & -109.04 & -1.98 \\ 1.39 & 0.91 & 0.46 & 0.45 & 28.67 & 1.36 \\ 4.66 & 4.79 & 0.45 & 0.79 & 24.36 & 1.50 \\ 27.83 & -109.04 & 28.67 & 24.36 & 2229.32 & 100.37 \\ 5.25 & -1.98 & 1.36 & 1.50 & 100.37 & 5.32 \end{pmatrix}$	$\begin{pmatrix} 44.60 & 73.38 & 1.08 & 3.59 & 20.31 & 3.58 \\ 73.38 & 477.72 & 0.52 & 3.74 & -109.12 & -2.84 \\ 1.08 & 0.52 & 0.37 & 0.36 & 23.88 & 1.11 \\ 3.59 & 3.74 & 0.36 & 0.69 & 21.27 & 1.25 \\ 20.31 & -109.12 & 23.88 & 21.27 & 1920.12 & 85.01 \\ 3.58 & -2.84 & 1.11 & 1.25 & 85.01 & 4.18 \end{pmatrix}$
MVE	MCD	S
$\begin{pmatrix} 69.51 & 110.33 & 1.56 & 5.50 & 20.98 & 5.53 \\ 110.33 & 708.64 & 0.76 & 5.12 & -144.09 & -2.31 \\ 1.56 & 0.76 & 0.51 & 0.50 & 30.13 & 1.52 \\ 5.50 & 5.12 & 0.50 & 0.91 & 25.55 & 1.69 \\ 20.98 & -144.09 & 30.13 & 25.55 & 2244.18 & 107.66 \\ 5.53 & -2.31 & 1.52 & 1.69 & 107.66 & 5.84 \end{pmatrix}$	$\begin{pmatrix} 72.11 & 107.39 & 1.59 & 5.62 & 20.25 & 5.81 \\ 107.39 & 672.44 & -0.12 & 3.84 & -221.60 & -4.05 \\ 1.59 & -0.12 & 0.57 & 0.55 & 35.47 & 1.74 \\ 5.62 & 3.84 & 0.55 & 0.97 & 29.16 & 1.89 \\ 20.25 & -221.60 & 35.47 & 29.16 & 2786.64 & 128.64 \\ 5.81 & -4.05 & 1.74 & 1.89 & 128.64 & 6.71 \end{pmatrix}$	$\begin{pmatrix} 57.94 & 96.13 & 1.24 & 4.45 & 23.34 & 4.67 \\ 96.13 & 602.87 & -0.03 & 4.48 & -153.23 & -5.10 \\ 1.24 & -0.03 & 0.48 & 0.45 & 30.38 & 1.43 \\ 4.45 & 4.48 & 0.45 & 0.83 & 26.07 & 1.56 \\ 23.34 & -153.23 & 30.38 & 26.07 & 2387.47 & 107.59 \\ 4.67 & -5.10 & 1.43 & 1.56 & 107.59 & 5.63 \end{pmatrix}$
BPS		
$\begin{pmatrix} 58.21 & 64.59 & 1.01 & 4.25 & -0.40 & 4.49 \\ 64.59 & 345.59 & -3.14 & 0.19 & -319.90 & -10.66 \\ 1.01 & -3.14 & 0.36 & 0.35 & 22.13 & 1.15 \\ 4.25 & 0.19 & 0.35 & 0.68 & 18.93 & 1.40 \\ -0.40 & -319.90 & 22.13 & 18.93 & 1787.82 & 85.48 \\ 4.49 & -10.66 & 1.15 & 1.40 & 85.48 & 4.73 \end{pmatrix}$		

Table 4.3 - Apparent Error Rates

MLE	M	OGK	MVE	MCD	S	BPS
0.055	0.129	0.092	0.074	0.111	0.074	0.058

Table 4.4 - Confusion Matrices

MLE	M	OGK	MVE	MCD	S	BPS
$\begin{pmatrix} 0.888 & 0.112 \\ 0.000 & 1.000 \end{pmatrix}$	$\begin{pmatrix} 0.852 & 0.148 \\ 0.111 & 0.889 \end{pmatrix}$	$\begin{pmatrix} 0.889 & 0.111 \\ 0.074 & 0.926 \end{pmatrix}$	$\begin{pmatrix} 0.888 & 0.112 \\ 0.037 & 0.962 \end{pmatrix}$	$\begin{pmatrix} 0.852 & 0.148 \\ 0.074 & 0.926 \end{pmatrix}$	$\begin{pmatrix} 0.926 & 0.074 \\ 0.074 & 0.926 \end{pmatrix}$	$\begin{pmatrix} 0.882 & 0.118 \\ 0.000 & 1.000 \end{pmatrix}$

It is observed from the above tables almost the proposed algorithm gives reliable results while comparing with the other procedures. The apparent error rate of the BPS algorithm is 5.8%, which is very low while comparing with the other robust procedures. From the confusion matrix, one can concluded that BPS and MLE procedures misclassified around 12% in the first group and no misclassification in the second group. All the other methods misclassified in both the groups. Hence, the BPS procedure gives an estimate which is very nearer to the true estimator for the given real data.

V. Simulation Study

The simulation study also performed to compare the efficiency of the proposed BPS algorithm along with the other procedures. The data points are simulated from multivariate normal distribution. Two groups of multivariate data were simulated each with 100 observations with mean $\mu_1 = (0,0,0)$, $\mu_2 = (1,1,1)$ and the common covariance matrix Σ is identity matrix is $(1,0,0,0,1,0,0,0,1)$. The estimated mean vectors, covariance matrices, apparent error rate and confusion matrices under various methods are displayed in the tables 5.1-5.4 respectively.

Table 5.1 Estimated Location Vector

MLE	M	OGK	MVE	MCD	S	BPS
$\begin{pmatrix} 0.5732 \\ 0.5655 \\ 0.5984 \end{pmatrix}$	$\begin{pmatrix} 0.5472 \\ 0.5760 \\ 0.6093 \end{pmatrix}$	$\begin{pmatrix} 0.5685 \\ 0.5831 \\ 0.6471 \end{pmatrix}$	$\begin{pmatrix} 0.5761 \\ 0.5457 \\ 0.6010 \end{pmatrix}$	$\begin{pmatrix} 0.5794 \\ 0.5450 \\ 0.6043 \end{pmatrix}$	$\begin{pmatrix} 0.5360 \\ 0.5779 \\ 0.6153 \end{pmatrix}$	$\begin{pmatrix} 0.5351 \\ 0.5926 \\ 0.6056 \end{pmatrix}$

Table 5.2 Estimated Scatter matrices

<i>MLE</i>	<i>M</i>	<i>OGK</i>	<i>MVE</i>
$\begin{pmatrix} 1.1966 & 0.3261 & 0.2472 \\ 0.3261 & 1.5115 & 0.2459 \\ 0.2472 & 0.2459 & 1.1871 \end{pmatrix}$	$\begin{pmatrix} 1.2591 & 0.4044 & 0.4040 \\ 0.4044 & 1.6453 & 0.2266 \\ 0.4040 & 0.2266 & 1.0923 \end{pmatrix}$	$\begin{pmatrix} 1.1078 & 0.3253 & 0.2902 \\ 0.3253 & 1.3679 & 0.1958 \\ 0.2902 & 0.1958 & 1.0235 \end{pmatrix}$	$\begin{pmatrix} 1.0547 & 0.1868 & 0.3961 \\ 0.1868 & 1.2916 & 0.2104 \\ 0.3960 & 0.2697 & 1.0966 \end{pmatrix}$
<i>MCD</i>	<i>S</i>	<i>BPS</i>	
$\begin{pmatrix} 1.0812 & 0.2104 & 0.3735 \\ 0.2104 & 1.2958 & 0.2532 \\ 0.3735 & 0.2532 & 1.0959 \end{pmatrix}$	$\begin{pmatrix} 1.2529 & 0.3721 & 0.3803 \\ 0.3721 & 1.6122 & 0.2142 \\ 0.3803 & 0.2142 & 1.1266 \end{pmatrix}$	$\begin{pmatrix} 0.7658 & 0.1664 & 0.2669 \\ 0.1664 & 0.7269 & 0.2210 \\ 0.2669 & 0.2210 & 0.5393 \end{pmatrix}$	

Table 5.3 Apparent Error Rates

MLE	M	OGK	MVE	MCD	S	BPS
0.01	0.015	0.025	0.03	0.015	0.01	0.00

Table 5.4 Confusion Matrices

MLE	M	OGK	MVE	MCD	S	BPS
$\begin{pmatrix} 1.00 & 0.00 \\ 0.02 & 0.98 \end{pmatrix}$	$\begin{pmatrix} 0.99 & 0.01 \\ 0.02 & 0.98 \end{pmatrix}$	$\begin{pmatrix} 0.97 & 0.03 \\ 0.02 & 0.98 \end{pmatrix}$	$\begin{pmatrix} 0.99 & 0.01 \\ 0.05 & 0.95 \end{pmatrix}$	$\begin{pmatrix} 1.00 & 0.00 \\ 0.03 & 0.97 \end{pmatrix}$	$\begin{pmatrix} 0.98 & 0.02 \\ 0.00 & 1.00 \end{pmatrix}$	$\begin{pmatrix} 1.00 & 0.00 \\ 0.00 & 1.00 \end{pmatrix}$

From the above table, the apparent error rate is zero for the proposed algorithm and even MLE have 0.01. BPS algorithm gives zero percent misclassification in both the groups. Almost all the other methods misclassification occurs in both the groups except MLE and MCD. Hence, it is concluded that, the proposed algorithm gives the reliable estimate of multivariate location and scatter in turn it gives the reliable results while using it in multivariate analyses.

VI. Summary and Conclusion

The estimate of multivariate location scatter plays a vital role in the context of multivariate analysis. Many authors have proposed estimators to estimate the same. In this paper we propose a new algorithm namely, BPS algorithm with MCD is used to estimate the location and scatter matrices by selecting best points from the given multivariate data. Also, the efficiency of the proposed BPS algorithm is compared with classical and most widely used robust procedures by computing location and scatter matrices with involving classification aspect in a real and simulated data. It is observed from the results the BPS algorithm gives an estimate which is very nearer to true estimate and usage this algorithm in the multivariate techniques shows a reliable results.

Acknowledgements

This research work was partially supported by the Department of Science and Technology (DST), under grants Promotion of University Research and Scientific Excellence (PURSE) programme, New Delhi, at Bharathiar University, Tamil Nadu and India.

Reference

- Anderson, T.W. (1984). An Introduction to Multivariate Statistical Analysis. *John Wiley and Sons Inc.* New York.
- Abhimanyu Das and David Kempe. (2008). Algorithm for Subset Selection in Linear Regression. *Symposium on the Theory of Computing*, May 17, 20, Victoria, British Columbia, Canada.
- Butler R. Davies P. Jhun M. (1993). Asymptotic for the Minimum Covariance Determinant Estimator. *The Annals of Statistics* **21**:1385-1401.
- Bjorn K. Alsberg et al. (1998). Variable Selection in Discriminant Partial Least-Squares Analysis. *Anal. Chem.*, Vol. **70**, 4126-4133.
- Cook R.D., et al. (1993). Exact iterative computation of the robust multivariate minimum volume ellipsoid estimator. *Statistics and Probability*, Vol. **16**, 213-218.
- Davies P. (1987). Asymptotic Behavior of S-Estimators of Multivariate Location Parameters and Dispersion, *The Annals of Statistics*, **15**, 1269-1292.
- Devlin SJ, Gnanadesikan R, Kettenring JR (1981). Robust Estimation of Dispersion Matrices and Principal Components, *Journal of the American statistical Association*, **76**, 354-362.
- Gnanadesikan R. Kettenring JR. (1972). Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data, *Biometrics*, **28**, 81-124.
- Hawkins DM. (1994). The Feasible Solution Algorithm for Minimum Covariance Determinant Estimators in Multivariate Data, *computational Statistics & Data Analysis*, **17**(2), 197-210.
- Hendrik P. Lopuhaa. (1989). On the relation between S-estimators and M-estimators of Multivariate Location and Covariance, *The Annals of Statistics*, Vol. **17**, No.4, pp.1662-1683.
- Hendrik P. Lopuhaa and Peter J. Rousseeuw. (1991). Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices, *The Annals of Statistics*, Vol. **19**, No.1, pp.229-248.
- Huber, P. J. (1964). Robust estimation of a location parameter, *Ann. Math. Stat.*, **35**, 73-101.
- Huber, P. J. (1967). The behavior of Maximum likelihood estimates under nonstandard conditions, *Proc. Fifth Berkeley Symp. Math. Statist. Prob.*, **1**, 221-233.
- Huber, P.J. (1972). *Robust Statistics*, Wiley New York.
- Huber, P.J. (1972). Robust Statistics: A review, *Ann. Math. Stat.*, **43**, 1041-1067.
- Huber, P. J (1981). *Robust Statistics*, John Wiley & Sons, New York.
- Lopuhaa HP. (1989). On the Relation between S-Estimators and M-estimators of Multivariate Location and Covariance, *The Annals of Statistics*, **17**, 1662-1683.
- Macro Rani and Anthony C. Atkinson, (2000). Robust Diagnostic Data Analysis: Transformations in Regression. *American Statistical Association and the American Society for Quality, Technometrics*, November, Vol. **42**, No. 4.
- Maronna RA, Martin D, Yohai V. (2006). Robust Statistics: Theory and Methods, *John Wiley & Sons, New York*.
- Maronna RA, Zamar RH. (2002). Robust Estimation of Location and Dispersion for High Dimensional Datasets, *Technometrics*, **44**, 307-317.
- Maranna, R.C. (1976). Robust M estimators Multivariate Location and Scatter. *Annals of Statistics* **4**: 51-67.
- McHenry C.E. (1978). Computation of a best Subset in Multivariate Analysis, *Journal of the Royal Statistical Society*. Vol. **27**, No.3, PP.291-296.
- Mia Hubert, Peter J. Rousseeuw and Stefan Van Aelst. (2008). High-Breakdown Robust Multivariate Methods. *Statistical Science*, Vol. **23**, No.1, 92-119.
- Muthukrishnan.R and K.Mahesh. (2012). Performance of Classical Robust Linear Discriminant Analysis, *International Journal Statistics and Analysis* ISSN 2248-9959 Vol.2, Number 3, PP.239-243.
- Muthukrishnan.R, and K.Mahesh. (2013). Evaluation of classical and robust discriminant methods under apparent error rate. *International journal of Current Research*, ISSN: 0975-833X- Vol.5, Issue.10, PP.2817-2820.
- Narayan C.Giri. (2004). Multivariate Statistical Analysis. Second edition, Marcel Dekker, Inc. New York.
- Oyeyemi G.M and R.A. Ipinyomi. (2009). A robust method of estimating covariance matrix in multivariate data analysis. *African Journal of mathematics and computer science Research*. Vol. **3**(1), 001-018.
- Rousseeuw, P. J (1984). Least median of squares regression, *J. Am. Stat. Association*, **79**, 871-880.
- Rousseeuw, P.J. (1985). Multivariate estimation with high breakdown point, *Mathematical Statistics and Applications*, Vol. B (Grossmann et al., eds.), 283-297, Reidel, Dordrecht.
- Rousseeuw, P.J., Leroy .AM. (1987). Robust Regression and Outlier Detection. *New York: John Wiley & Sons*.
- Rousseeuw PJ, Van Deriessen K. (1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator, *Technometrics* **41**, 212-223.
- Rousseeuw PJ et al. (2009). Robust base: Basic Robust Statistics, R package version 04-5, URL: <http://CRAN.R-project.org/package=robust.base>.
- Todorov V. (1992). Computing Minimum Covariance Determinant Estimator (MCD) by Simulated Annealing, *Computational Statistics and Data Analysis*, **14**, 515-525.
- Todorov V. (2009). rrcov: Scalable Robust Estimators with High Breakdown Points, R package version 0.5-03, URL <http://CRAN.R-project.org/package=rrcov>.
- Todorov V, Filizmoser P. (2010). Robust Statistic for the one-way MANOVA, *computational Statistics & Data Analysis*, **54**:37-48.
- Yohai V.J. (1987). High Breakdown Point and High Efficiency Robust Estimates for Regression, *Annals of Statistics*, **15**, 642-656.
- Yohai V.J, and Zamar R.H. (1988). High breakdown estimates of regression by means of the minimization of an efficient Scale, *J. Amer. Statist. Assoc.*, vol. **83**, 406-413.