ISSN 2347-1921

Best Points Selection Procedure for Estimating Location and Scatter in Multivariate Data with Application to Discriminant Analysis

Muthukrishnan. R and K.Mahesh
Department of Statistics, Bharathiar University
Coimbatore-46. Tamilnadu, INDIA

Abstract

Multivariate data analysis is rely on the two measures namely location and scatter. The most widely used such estimators; sample mean and covariance matrix are extremely sensitive to outliers, then the results obtained with these estimators are inaccurate. Many robust alternatives are established and perform well while handling the data with outliers. But even still a challenging task while handling the large number of cases and/or variables with reference to the features such as dimensionality of data, heterogeneous of data, computing time, adequacy of the results and applications. This paper provides a procedure for the selection of best data points in order to estimate multivariate location and scatter. The obtained results also compared with the established robust procedures such as various MCD algorithmic techniques and MVE by a real environment. The application aspect of the procedure is also executed in the context of discriminant analysis of multivariate grouped data. The results such as apparent error rate, confusion matrix of classical and various robust discriminant procedures are also provided.

Keywords: Multivariate Location and Scatter; MCD; MVE; Discriminant Analysis.

Council for Innovative Research
Peer Review Research Publishing System
Journal: Journal of Advances in Mathematics

Vol 7, No. 3
editor@cirjam.com
www.cirjam.com, www.cirworld.com

ISSN 2347-1921

1.0 Introduction

The computation of location and scatter is plays a vital role in multivariate data analysis. There are many methods have already established and proved their efficiency in order to compute the location and scale matrix. Rousseeuw and Leroy (1985) was introduced minimum volume ellipsoid estimator with high break down point. Rousseeuw and Leroy (1987) classified a general procedure for S estimator. Many authors have contributed in various ways in order to estimate location and scatter matrix and studied the efficiency, some of them are mentioned here. Rousseeuw and van Driessen (1999), Hawkins (1994), Hawkins and McLachlan (1997), He and Fung (2000), Hubert and Van Driessen (2004), Valentin Todorov and Ana M.Pires (2007), Todorov and Filizmosor (2009, 2010). Almost 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 descriptions of the most widely used methods such as Maximum Likelihood Estimator (MLE), Minimum Volume Ellipsoid (MVE), Stahel Donoha Estimator (SDE) and Minimum Covariance Determinant estimator (MCD). The proposed algorithm for Best Points Selection (BPS) along with MVE is presented in the Section 3. The efficiency of the proposed algorithm along with the other methods is computed to a real data set involving the multivariate techniques, discriminant analysis and the results are presented in the section 4. Section 5 concludes with summary of the results.

2.0 Classical and Robust Procedures

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

2.1 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 x 1 sample mean vector, $m = \frac{\sum_{i=1}^{n} Z_{y,i}}{n}$, and p x 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 an arbitrarily inflate either of them.

2.2 Stahel Donoha Estimator (SDE)

A projection-based estimation procedure was developed independently by Stahel (1981) and Donoho (1982) and also discussed Rousseeuw and Leroy (1987). Simplistically, the idea here is that an outlier or high leverage point will separate out and away from the bulk of the data when viewed from the right perspective. There are two stages to the formation of the robust multivariate location and dispersion estimators. First, robust distances are determined via a projection computation. These distances become the arguments in a weight function that is used to calculate a weighted mean vector and weighted covariance matrix. This procedure is affine equivariant, and attains a 50% (asymptotic) breakdown point when $n > 2p+1$ and the data are in general position (Donoho (1982), Rousseeuw and Leroy (1987)), which means that no more than p +1 points lie in any p dimensional affine subspace.

While the definition of the Stahel-Donoho estimator requires the supremum of over all possible directional vectors, Rousseeuw and van Zomeren (1990) propose a shortcut method which uses just n directional vectors, one vector in the direction of each centered observation (centered by the coordinatewise median, vector starting from the origin). The projections of the original data on these n directional vectors produce the robust distances.

2.3 Minimum Volume Ellipsoid (MVE)

Rousseeuw (1985) introduced the robust minimum volume ellipsoid (MVE) method for detection of outliers in multidimensional data. Subsets of approximately 50% of the observations are examined, and to find a subset which minimizes the volume of ellipsoid occupied by the data. The best subset smallest volume is then used to calculate the covariance matrix. An appropriate cut-off value is then estimated, and the observations with distances that exceed that cut-off are declared to be outliers. The MVE estimator is the center and the covariance of a subsample size h $(h \leq n)$ that minimizes the volume of the covariance matrix associated to the subsample. Formally, MVE= $(\overline{X_i^*}, S_i^*)$, where i={set of h instances $Vol(S_i^*) \leq Vol(S_k^*)$ for all k such that #(k)=h} $Vol(S_k)=\{|S_k|med_{j=1,2\ldots,h}d_j^2\}^{1/2}$, d_j-represents the Mahalanobis distance of the j-th instance in S_k.

ISSN 2347-1921

The ellipsoid is defined by $(X-\overline{X})' S^{-1}(X-\overline{X}) \le a^2$, the value of h can be treated as the minimum number of instances which must not be outlying and $h=[(n+p+1)/2]$, is the greatest integer function and p is the number of predictors.

2.4 Minimum Covariance Determinant 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 nxn positive semi-definite matrix, p eigen values are positive, the determinant of covariance matrix equals the product of eigenvalues. 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 $p\,x1$ mean vector, and corresponding pxp covariance matrix, these estimators are called minimum covariance determinant estimators. Further, to improve the efficiency of the MCD, Rousseeuw P.J. Driessen K. (1999) developed algorithm namely Fast MCD algorithm.

Todorov and Flizmoser (2010) extended the MCD algorithm with in the context of pooled algorithm then the results are namely MCD A, MCD B, MCD C. Also, Hawkins (1994) established a Feasible solution algorithm (FSA) for MCD approach and these are summarized below.

Minimum covariance determinant procedures was studied by Todorov et al. (1990), Rousseeuw et al. (1992), Croux and Dehon (2001) and all are applied for robustifying multivariate analyses based on S estimates. A drawback of this method is that the same trimming proportions are applied to all groups which could lead to a loss of efficiency if some groups are outliers free, namely MCD A.

Minimum covariance determinant analysis in each one of the group was proposed by He and Fung (2000) for the S estimates and modified algorithm was developed in Hubert and VanDriessen (2004), as a replacement for pooling the group covariance matrices, the observations are centered and to obtain the each one of the individual group location estimates as the reweighted minimum covariance determinant location estimates of each group, namely MCD B.

The next approach is to modify the algorithm for high breakdown points estimation itself in order to accommodate the pooled sample. He and Fung (2000) studied the modified algorithm for S estimation in case of two groups. Hawkins and McLachan (1997) defined the minimum within group covariance determinant estimator which does not apply the same trimming proportion to each group but minimizes directly the determinant of the common within group's covariance matrix by pairwise swaps of observations. This type of approach is namely MCD C.

Douglas M. Hawkins (1994) has proposed the feasible solution algorithm for MCD method. It is clear from the definition that the MCD estimator for μ and Σ is the sample mean vector and covariance matrix of a subset of size $n-h$, the determinant of $\hat{\Sigma}$ cannot be decreased by any case wise exchange exchanging one of the trimmed cases for one of the retained cases,

3.0 Best Points Selection based on MVE (BPS - MVE)

The proposed Best Points Selection (BPS) algorithm based on minimum volume ellipsoid (MVE) 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:

Suppose that a sequence $(S_1, S_2,\ldots,S_k)$ is chosen recursively with, there are n_1 choices for s_1. For each $1<j<k$ once $s_1, s_2,...,s_{j-1}$ have been chosen, there are n_j choices for s_j then there are $n_1, n_2,...,n_k$ choices for the sequence.

Suppose that $S_1, S_2,...,S_p$ are finite sets. Then their product set is $S_1.S_2\ldots S_k$ is the set of all sequence $(s_1, s_2,...,s_p)$ with each $s_i \varepsilon S_i$. In forming the sequence in this product the choice at any point is independent of the previous choices, so the number of choice of any s_i is $|S_i|$. Thus, from the product rule we have the familiar identity $|S_1 \times S_2 \times ... \times S_p| = |S_1||S_2|...|S_k|$.

So, in particular, for any set S the number of sequence of length k from S is $|S^p| = |S^k|$.

Let S be a set with $|S| = n$. If $1 \le k \le n$, then a k- permutation of S is a sequence $(s_1, s_2,...,s_k)$ with no repetitions of elements from S; so note that such a permutation (k) is more succinctly described as a one-to-one map

$\sigma : \{1,2,...,k\} \to S$. 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 $(s_1, s_2, ..., s_n)$ 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

$$P(n,n) = n(n-1)(n-2),...,2.1 = n!$$

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

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

Hence, the method has $k-$ permutation is given below

$$P(n,n) = n!$$

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

Therefore,

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

replace n by $k = \dfrac{(g_1 + g_2 + ... + g_n)}{([m+p+1]/2)}$

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

where $h = \left(\dfrac{(g_1+g_2+...+g_n)}{([m+p+1]/2)}\right)$, and then selecting the h points by minimum volume ellipsoid (MVE) procedure.

Step 1: Let $X = (x_1, x_2, ..., x_m)$ be a set of m points in $\Re^p$, let h be a natural number such that $m/2 \le h = \dfrac{(g_1+g_2+...+g_n)}{([m+p+1]/2)} < m$.

Step 2: The $p+1$ data points $\{x_1, x_2, ..., x_{p+1}\}$, were selected and to obtain the location and scatter matrix

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

Step 3: The mahalanobis distance for all the k observations using $\bar{y}$ and S can be calculated as $d_i^2 = (x-\bar{x})S^{-1}(x-\bar{x})^T$,

Step 4: The $d_i^2 (i = 1,2,...,k)$ 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

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

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 $h = \left(\frac{(g_1 + g_2 + ... + g_n}{([m+p+1]/2)}\right)$.

Then based on the BPS algorithm the location and scatter are given as:

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

4.0 Numerical Study

Case 1

For the purpose of comparing the efficiency of the proposed BPS algorithm with other methods, the famous iris data set was considered. The iris data set contains 3 groups each with 50 observations of the 4 variables. The groups are namely setosa, versicolor and verginica, the variables are Sepal.Length, Sepal.Width, Petal.Length and Petal.Width. The first two groups of the Iris data were considered in this first case.

Table 4.1 Estimated of Mean vector based on BPS along with the other procedures

MLE	*MVE*	*SDE*	*MCD*	*MCDA*	*MCDB*	*MCDC*	*FSA*	*BPS*
5.471	5.461	5.305	5.006	5.444	5.508	5.501	5.415	5.396
3.099	3.113	3.193	3.369	3.095	3.132	3.125	3.092	3.107
2.861	2.743	2.406	1.560	2.836	2.879	1.560	2.807	2.680
0.786	0.741	0.602	0.291	0.754	0.782	0.291	0.756	0.689

Table 4.2 Estimated of covariance matrices based on BPS along with the other procedures

MLE

0.412	−0.063	0.756	0.286
−0.063	0.229	−0.418	−0.155
0.756	−0.418	2.101	0.802
0.286	−0.155	0.802	0.319

MVE

0.534	−0.091	0.962	0.364
−0.091	0.278	−0.555	−0.205
−0.962	−0.555	2.548	0.962
0.364	−0.205	0.962	0.375

SDE

0.555	−0.104	1.065	0.415
−0.104	0.211	−0.507	−0.194
1.065	−0.507	2.828	1.098
0.360	−0.194	1.098	0.436

MCD

0.237	0.191	0.029	0.019
0.191	0.393	−0.176	−0.071
0.029	−0.176	0.394	0.162
0.019	−0.071	0.162	0.090

MCDA

0.359	0.165	0.148	0.043
0.165	0.172	0.081	0.045
0.148	0.081	0.188	0.056
0.043	0.045	0.056	0.033

MCDB

0.234	0.099	0.080	0.023
0.099	0.118	0.027	0.016
0.080	0.027	0.086	0.015
0.023	0.016	0.015	0.013

MCDC

0.159	0.070	0.050	0.015
0.070	0.085	0.018	0.011
0.050	0.018	0.057	0.010
0.015	0.011	0.010	0.009

FSA

0.134	0.078	0.042	0.017
0.078	0.103	0.028	0.019
0.042	0.028	0.051	0.014
0.017	0.019	0.014	0.014

BPS

0.746	−0.099	1.552	0.693
−0.099	0.142	−0.420	−0.179
1.552	−0.420	3.892	1.757
0.693	−0.179	1.757	0.810

Table 4.3 Confusion Matrices based on various procedures

$$MLE \quad \begin{pmatrix} 1.00 & 0.00 \\ 0.00 & 1.00 \end{pmatrix} \qquad MVE \quad \begin{pmatrix} 1.00 & 0.00 \\ 0.00 & 1.00 \end{pmatrix} \qquad SDE \quad \begin{pmatrix} 1.00 & 0.00 \\ 0.00 & 1.00 \end{pmatrix}$$

$$MCD \quad \begin{pmatrix} 1.00 & 0.00 \\ 0.00 & 1.00 \end{pmatrix} \qquad MCDA \quad \begin{pmatrix} 1.00 & 0.00 \\ 0.00 & 1.00 \end{pmatrix} \qquad MCDB \quad \begin{pmatrix} 1.00 & 0.00 \\ 0.00 & 1.00 \end{pmatrix}$$

$$MCDC \quad \begin{pmatrix} 1.00 & 0.00 \\ 0.00 & 1.00 \end{pmatrix} \qquad FSA \quad \begin{pmatrix} 1.00 & 0.00 \\ 0.00 & 1.00 \end{pmatrix} \qquad BPS \quad \begin{pmatrix} 1.00 & 0.00 \\ 0.00 & 1.00 \end{pmatrix}$$

The estimated location and scatter matrices are displayed in table 4.1 & 4.2. Table 4.3 shows the confusion matrix. It is observed that the apparent error rate is 0.00 for all the classical, robust procedures and the proposed BPS procedure, almost all the methods including the BPS algorithm classified the given data in to two groups correctly.

Case 2

In this case, more than two groups, i.e., multi group classification was considered. Hence, we considered all the three groups of Iris data. The computed location and scatter matrices, apparent error rate, confusion matrices are listed in the following tables 4.4, 4.5, 4.6 and 4.7 respectively.

Table 4.4 Estimated of Mean vector based on BPS along with the other procedures

MLE	*MVE*	*SDE*	*MCD*	*MCDA*	*MCDB*	*MCDC*	*FSA*	*BPS*
5.843	5.497	5.561	5.503	5.775	5.792	5.804	5.771	5.683
3.057	3.076	2.945	3.107	3.050	3.077	3.073	3.054	3.066
3.758	2.946	3.311	2.897	3.691	3.711	3.704	3.684	3.465
1.199	0.817	0.975	0.793	1.182	1.207	1.211	1.185	1.089

Table 4.5 Estimated of covariance matrices based on BPS along with the other procedures

$$MLE \quad \begin{pmatrix} 0.685 & -0.042 & 1.274 & 0.516 \\ -0.042 & 0.189 & -0.329 & -0.121 \\ 1.274 & -0.329 & 3.116 & 1.295 \\ 0.516 & -0.121 & 1.295 & 0.581 \end{pmatrix} \qquad MVE \quad \begin{pmatrix} 0.744 & -0.126 & 1.367 & 0.516 \\ -0.126 & 0.296 & -0.646 & -0.234 \\ 1.367 & -0.646 & 3.594 & 1.369 \\ 0.516 & -0.234 & 1.369 & 0.542 \end{pmatrix}$$

$$SDE \quad \begin{pmatrix} 1.092 & -0.087 & 2.043 & 0.834 \\ -0.087 & 0.188 & -0.402 & -0.151 \\ 2.043 & -0.402 & 4.586 & 1.884 \\ 0.834 & -0.151 & 1.884 & 0.810 \end{pmatrix}$$

$$MCD \quad \begin{pmatrix} 0.728 & -0.138 & 1.336 & 0.507 \\ -0.138 & 0.359 & -0.735 & -0.268 \\ 1.336 & -0.735 & 3.577 & 1.358 \\ 0.507 & -0.268 & 1.358 & 0.534 \end{pmatrix} \qquad MCD\,A \quad \begin{pmatrix} 0.332 & 0.136 & 0.158 & 0.054 \\ 0.136 & 0.151 & 0.079 & 0.057 \\ 0.158 & 0.079 & 0.195 & 0.060 \\ 0.054 & 0.057 & 0.060 & 0.065 \end{pmatrix}$$

$$MCD\,B \quad \begin{pmatrix} 0.239 & 0.104 & 0.123 & 0.044 \\ 0.104 & 0.129 & 0.061 & 0.039 \\ 0.123 & 0.061 & 0.147 & 0.052 \\ 0.044 & 0.039 & 0.052 & 0.040 \end{pmatrix}$$

$$MCD\,C \quad \begin{pmatrix} 0.195 & 0.079 & 0.101 & 0.039 \\ 0.079 & 0.098 & 0.044 & 0.027 \\ 0.101 & 0.044 & 0.111 & 0.039 \\ 0.039 & 0.027 & 0.039 & 0.029 \end{pmatrix} \qquad FSA \quad \begin{pmatrix} 0.188 & 0.079 & 0.088 & 0.032 \\ 0.079 & 0.094 & 0.035 & 0.019 \\ 0.088 & 0.035 & 0.097 & 0.030 \\ 0.032 & 0.019 & 0.030 & 0.021 \end{pmatrix}$$

$$BPS \quad \begin{pmatrix} 0.746 & -0.099 & 1.552 & 0.693 \\ -0.099 & 0.142 & -0.420 & -0.179 \\ 1.552 & -0.420 & 3.892 & 1.757 \\ 0.693 & -0.179 & 1.757 & 0.810 \end{pmatrix}$$

Table 4.6 Apparent Error rates under various procedures

MLE	MVE	SDE	MCD	MCD A	MCD B	MCD C	FSA	BPS
0.020	0.013	0.020	0.026	0.030	0.020	0.040	0.040	0.012

Table 4.7 Confusion Matrix under various procedures

MLE	MVE	SDE
$\begin{pmatrix}1.000 & 0.000 & 0.000\\ 0.000 & 0.963 & 0.037\\ 0.000 & 0.037 & 0.963\end{pmatrix}$	$\begin{pmatrix}1.000 & 0.000 & 0.000\\ 0.000 & 0.926 & 0.037\\ 0.000 & 0.074 & 0.963\end{pmatrix}$	$\begin{pmatrix}1.000 & 0.000 & 0.000\\ 0.000 & 0.960 & 0.040\\ 0.000 & 0.020 & 0.098\end{pmatrix}$
MCD	**MCD A**	**MCD B**
$\begin{pmatrix}1.000 & 0.000 & 0.000\\ 0.000 & 0.960 & 0.040\\ 0.000 & 0.040 & 0.960\end{pmatrix}$	$\begin{pmatrix}1.000 & 0.000 & 0.000\\ 0.000 & 0.920 & 0.080\\ 0.000 & 0.020 & 0.980\end{pmatrix}$	$\begin{pmatrix}1.000 & 0.000 & 0.000\\ 0.000 & 0.960 & 0.040\\ 0.000 & 0.020 & 0.980\end{pmatrix}$
MCD C	**FSA**	**BPS**
$\begin{pmatrix}1.000 & 0.000 & 0.000\\ 0.000 & 0.960 & 0.040\\ 0.000 & 0.080 & 0.920\end{pmatrix}$	$\begin{pmatrix}1.000 & 0.000 & 0.000\\ 0.000 & 0.960 & 0.040\\ 0.000 & 0.080 & 0.920\end{pmatrix}$	$\begin{pmatrix}1.000 & 0.000 & 0.000\\ 0.000 & 0.963 & 0.037\\ 0.000 & 0.000 & 1.000\end{pmatrix}$

It is observed from the table the estimated mean vector under BPS algorithm is lies between the classical estimates MLE and the robust procedure MVE. It is not far away from the classical and robust procedure. While comparing with the other robust procedures its shows a better result. From the apparent error rate, BPS procedure gives very low (0.012) compared to the classical as well as robust procedures. It is observed from the above table, the classical procedure MLE and other robust procedures classified correctly the first group and not on second and third group, they are misclassified between 1% and 4%. But the proposed BPS algorithm the first and last group data were classified correctly and 1% were misclassified in the second group. It is conclude that the proposed BPS procedure performs better than the classical and the other robust procedures for the given real data set.

4. Summary and Conclusion

The proposed BPS algorithm is used to estimate the location and scatter matrices by selecting best points from the given data set. The efficiency of the proposed BPS algorithm is compared along with classical and the most widely used robust procedures by computing the location and scatter matrices with a real data set. Also, the application aspect of the proposed BPS algorithm with the other procedures is also carried out in the context of classification problem. It is observed that the performance of the proposed algorithm equally well with other procedures and in some situation it is more efficient than the others. In high dimensional data analysis, BPS algorithm with MVE is more robust and time consuming.

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.

References

[1] Anderson, T.W. (1984). An Introduction to Multivariate Statistical Analysis. *New York: Wiley.*

[2] Akritas, M.G. (1978). Robust M-estimation in the two sample problem. *Jour. Amer. Statist. Assoc.* **86**, 201-204.

[3] Croux.C and Dehon.C. (2001). Robust linear discriminant analysis using S-estimators, *The Canadian Journal of Statistics,* Vol. **29**. 473-492.

[4] Donoho,D.L.(1982). Breakdown properties of multivariate location estimators, Ph.D Thesis. D.O.Statistics, Harvard University.

[5] Fung, W.K. (1992). Some diagnostic measures in discriminant analysis, *Statistics and Probability Letters,* Vol. **13**, 279-285.

[6] Hawkins DM. (1994). The feasible solution algorithm for the minimum covariance determinant estimator in multivariate data, *Computational Statistics & Data Analysis,* Vol.**17(2)**, 197-210, ISSN 0167-9473.

[7] Hawkins DM. McLachlan G.J. (1997). High breakdown linear discriminant analysis, *joutnal of the American Statistical Association,* Vol.**92**, 136-143.

[8] He.X and Fung W.K. (2000). High breakdown estimation for multiple populations with applications to discriminant analysis, *Journal of Multivariate Analysis,* Vol. **72**. 151-162.

[9] Hubert M. Van Driseen K. (2004). Fast and robust discriminant analysis, , *Computat.Statist Data Anal*, vol. **45**. 301-320

[10] Indrani Basak. (1998). Robust M-estimation in Discriminant Analysis. *The Indian Journal of Statistics,* Vol.**60**, Series B, Pt. 2, PP.246-268.

[11] Maronna, R.A. (1976). Robust M-estimators of multivariate location and scatter, *Annals of Statistics,* Vol.**4**, 51-67.

[12] Mia Hubert, Peter J. Rousseeuw and Stefan Van Aelst. (2008). High-Breakdown Robust Multivariate Methods, *Institute of Mathematical Statistics.* Vol.**23**, No.1, 92-119.

[13] 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.

[14] 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.

[15] Rousseeuw, P.J. (1984). Least median square regression, *Journal of the American Statistical Association,* Vol. **79**, 871-880.

[16] Rousseeuw, P.J. (1985). Multivariate estimation with high breakdown point, Mathematical Statistics and Applications, Vol. B (Grossmann et al., eds.), 283-297, Reidel, Dordrecht.

[17] Rousseeuw, P.J. and A.M.Leroy. (1987). Robust regression and outlier detection, *New York, John Wiley and Sons.*

[18] Rousseeuw, P.J. and B.Van Zomeren. (1990). Unmasking multivariate outliers and leverage points, *Jour. Amer. Statist. Assoc.* Vol. **85**, 633-651.

[19] Rousseeuw,P.J. and Van Driessen.K. (1999). A fast algorithm for the minimum covariance determinant estimator, *Technometrics,* Vol.**41**. 212-223.

[20] Lopuhaa, H.P. (1989). On the relation between S-estimators and M-estimators of multivariate location and covariance, *Annals of Statistics,* Vol.**17**, 1662-1683.

[21] Narayan C.Giri. (2004). Multivariate Statistical Analysis. Second edition, Marcel Dekker,Inc. Newyork.

[22] Stahel, W. A. (1981). .Robuste Schätzungen: Infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen., Ph.D. Thesis, ETH Zürich, Switzerland.

[23] Todorov.V, Neykov N. Neytchev P1. (1994). Robust two group discrimination by bounded influence regression, *Computat.Statist Data Anal,* Vol. **17**. 289-302.

[24] Valentin Todorov and Ana M.Pires. (2007). Comparative Performance of Several Robust Linear Discriminant Analysis methods, *REVSTAT- Statistical Journal* 5(1), 63-83.

[25] Todorov V. Filizmosor P. (2009). An Object oriented Frame work for Robust Multivariate Analysis, *Journal of Statistical Software, 32* (3), 1-47.

[26] Todorov V. Filizmosor P. (2010). Robust Statistics for the One-way MANOVA, *Computational Statistics and Data Analysis,* **54**, 37-48.