# A New Procedure of Clustering Based on Multivariate Outlier Detection

G. S. David Sam Jayakumar* and Bejoy John Thomas
*Jamal Mohamed College*

*Abstract*: Clustering is an extremely important task in a wide variety of application domains especially in management and social science research. In this paper, an iterative procedure of clustering method based on multivariate outlier detection was proposed by using the famous Mahalanobis distance. At first, Mahalanobis distance should be calculated for the entire sample, then using $T^2$-statistic fix a UCL. Above the UCL are treated as outliers which are grouped as outlier cluster and repeat the same procedure for the remaining inliers, until the variance-covariance matrix for the variables in the last cluster achieved singularity. At each iteration, multivariate test of mean used to check the discrimination between the outlier clusters and the inliers. Moreover, multivariate control charts also used to graphically visualizes the iterations and outlier clustering process. Finally multivariate test of means helps to firmly establish the cluster discrimination and validity. This paper employed this procedure for clustering 275 customers of a famous two-wheeler in India based on 19 different attributes of the two wheeler and its company. The result of the proposed technique confirms there exist 5 and 7 outlier clusters of customers in the entire sample at 5% and 1% significance level respectively.

*Key words*: Mahalanobis distance, multivariate outliers, multivariate test, upper control limit, variance-covariance matrix.

## 1. Introduction and Related Work

Outliers are the set of objects that are considerably dissimilar from the remainder of the data (Han, 2006). Outlier detection is an extremely important problem with a direct application in a wide variety of application domains, including fraud detection (Bolton, 2002), identifying computer network intrusions and bottlenecks (Lane, 1999), criminal activities in e-commerce and detecting suspicious activities (Chiu, 2003). Different approaches have been proposed to

---

*Corresponding author.

detect outliers, and a good survey can be found in (Knorr, 1998; Knorr, 2000; Hodge, 2004). Clustering is a popular technique used to group similar data points or objects in groups or clusters (Jain and Dubes, 1988). Clustering is an important tool for outlier analysis. Several clustering-based outlier detection techniques have been developed. Most of these techniques rely on the key assumption that normal objects belong to large and dense clusters, while outliers form very small clusters (Loureiro, 2004; Niu, 2007). It has been argued by many researchers whether clustering algorithms are an appropriate choice for outlier detection. For example, in (Zhang and Wang, 2006), the authors reported that clustering algorithms should not be considered as outlier detection methods. This might be true for some of the clustering algorithms, such as the $k$-means clustering algorithm (MacQueen, 1967). This is because the cluster means produced by the $k$-means algorithm is sensitive to noise and outliers (Laan, 2003). Similarly, that the case is different for the Partitioning Around Medoids (PAM) algorithm (Kaufman and Rousseeuw, 1990). PAM attempts to determine $k$ partitions for $n$ objects. The algorithm uses the most centrally located object in a cluster (called medoid) instead of the cluster mean. PAM is more robust than the $k$-means algorithm in the presence of noise and outliers. This is because the medoids produced by PAM are robust representations of the cluster centers and are less influenced by outliers and other extreme values than the means (Laan, 2003; Kaufman and Rousseeuw, 1990; Dudoit and Fridlyand, 2002). Furthermore, PAM is a data-order independent algorithm (Hodge, 2004), and it was shown in (Bradley, 1999) that the medoids produced by PAM provide better class separation than the means produced by the $k$-means clustering algorithm. PAM starts by selecting an initial set of medoids (cluster centers) and iteratively replaces each one of the selected medoids by one of the none-selected medoids in the data set as long as the sum of dissimilarities of the objects to their closest medoids is improved. The process is iterated until the criterion function converges. In this paper, a new method of clustering was proposed based on multivariate outlier detection. Note that our approach can be easily implemented when compare to other clustering algorithms that are based on PAM, such as CLARA (Kaufman and Rousseeuw, 1990), CLARANS (Ng and Han, 1994) and CLATIN (Zhang and Couloigner, 2005).

As discussed in (Loureiro, 2004; Niu, 2007; Zhang and Wang, 2006), there is no single universally applicable or generic outlier detection approach. Therefore, many approaches have been proposed to detect outliers. These approaches can be classified into four major categories based on the techniques used (Zhang and Wang, 2006), which are: distribution-based, distance-based, density-based and clustering-based approaches. Distribution-based approaches (Hawkins, 1980; Barnett and Lewis, 1994; Rousseeuw and Leroy, 1996) develop statistical models

(typically for the normal behavior) from the given data and then apply a statistical test to determine if an object belongs to this model or not. Objects that have low probability to belong to the statistical model are declared as outliers. However, distribution-based approaches cannot be applied in multidimensional scenarios because they are univariate in nature. In addition, a prior knowledge of the data distribution is required, making the distribution-based approaches difficult to be used in practical applications (Zhang and Wang, 2006). In the distance-based approach (Knorr, 1998; Knorr, 2000; Ramaswami, 2000; Angiulli and Pizzut, 2005), outliers are detected as follows. Given a distance measure on a feature space, a point $q$ in a data set is an outlier with respect to the parameters $M$ and $d$, if there are less than $M$ points within the distance $d$ from $q$, where the values of $M$ and $d$ are decided by the user. The problem with this approach is that it is difficult to determine the values of $M$ and $d$. Density-based approaches (Breunig, 2000; Papadimitriou, 2003) compute the density of regions in the data and declare the objects in low dense regions as outliers. In (Breunig, 2000), the authors assign an outlier score to any given data point, known as Local Outlier Factor (LOF), depending on its distance from its local neighborhood. A similar work is reported in (Papadimitriou, 2003). Clustering-based approaches (Loureiro, 2004; Gath and Geva, 1989; Cutsem and Gath, 1993; Jiang, 2001; Acuna and Rodriguez, 2004), consider clusters of small sizes as clustered outliers. In these approaches, small clusters (i.e., clusters containing significantly less points than other clusters) are considered outliers. The advantage of the clustering-based approaches is that they do not have to be supervised. Moreover, clustering-based techniques are capable of being used in an incremental mode (i.e., after learning the clusters, new points can be inserted into the system and tested for outliers). (Cutsem and Gath, 1993) present a method based on fuzzy clustering. In order to test the absence or presence of outliers, two hypotheses are used. However, the hypotheses do not account for the possibility of multiple clusters of outliers. Jiang *et al.* (Jiang, 2001) presented a two-phase method to detect outliers. In the first phase, the authors proposed a modified $k$-means algorithm to cluster the data, and then, in the second phase, an Outlier-Finding Process (OFP) is proposed. The small clusters are selected and regarded as outliers by using minimum spanning trees. In (Loureiro, 2004) clustering methods have been applied. The key idea is to use the size of the resulting clusters as indicators of the presence of outliers. The authors use a hierarchical clustering technique. A similar approach was reported in (Almeida, 2007). Acuna and Rodriguez (Acuna and Rodriguez, 2004) performed the PAM algorithm followed by the technique (henceforth, the method will be termed PAMST). The separation of a cluster A is defined as the smallest dissimilarity between two objects; one belongs to cluster A and the other does not. If the separation is large enough, then all objects that

belong to that cluster are considered outliers. In order to detect the clustered outliers, one must vary the number $k$ of clusters until obtaining clusters of small size and with a large separation from other clusters. In (Yoon, 2007), the authors proposed a clustering-based approach to detect outliers. The $k$-means clustering algorithm is used. As mentioned in (Laan, 2003), the $k$-means is sensitive to outliers, and hence may not give accurate results.

## 2. Proposed Approach

In this paper we proposed a new approach of outlier based clustering based on Mahalanobis distance. In statistics, Mahalanobis distance is a measure introduced by P. C. Mahalanobis (1936), which is based on correlations between variables by which different patterns can be identified and analyzed. It gauges similarity of an unknown sample set to a known one. It differs from Euclidean distance which takes the correlations of the data set and it is scale-invariant. In other words, it is a multivariate size. Formally, the multivariate distance of a multivariate vector $X = (X_1, X_2, X_3, \cdots, X_n)^T$ from group of values with mean $\bar{X} = (\bar{X}_1, \bar{X}_2, \bar{X}_3, \cdots, \bar{X}_n)^T$ and the co-variance matrix $S$ is defined as

$$\text{(Mahalanobis distance)}_i = \sqrt{(X_i - \bar{X})^T S^{-1} (X_i - \bar{X})}. \tag{1}$$

From (1), $\overline{X}$ is the sample mean matrix of order $p \times 1$ and $S$ is the sample variance-covariance matrix of order $p \times p$. The test statistic for the Mahalanobis distance is the squared Mahalanobis distance defined as $T$-square was first proposed by Harold Hotelling (1951) and it is given as

$$T_i^2 = (X_i - \bar{X})^T S^{-1} (X_i - \bar{X}). \tag{2}$$

From (2), Hotelling derived the (UCL) upper control limit of $T$-square statistic as UCL $= ((n-1)^2/n)\beta(\alpha, p/2, (n-p-1)/2)$, where $n$ is the sample size, $p$ is the number of variables, $\alpha$ is the level of significance and $\beta(\alpha, p/2, (n-p-1)/2)$ followed a beta distribution. Based on the above said distance measures, first, assume all the variables follows a multivariate normality and calculate the Mahalanobis distance from (1) for the $n$ observations based on $p$ variables, where $n > p$. Secondly, from (2) fix a UCL for $T$-square statistic, observations above the UCL are consider as outlier cluster and named as cluster 1. Repeat the same procedure for remaining observations excluding the observations in cluster 1. Repeat the process, until the nature of variance-covariance matrix for the variables in the last cluster achieves singularity. Moreover, the cut-off $T$-square value can fixed by using the beta distribution and the identification of individual outlier observation can be done with the help of 1% or 5% significance points of $T$-square test statistic. The basic structure of the proposed method is as follows:

**Step 1:** Calculate the Mahalanobis distance for $n$ observations based on $p$ variables.

**Step 2:** Determine the observations which are above the UCL of $T$-square statistic and consider those observations are outlier cluster 1.

**Step 3:** Using multivariate test of means, check the equality of means for the variables in cluster 1 and remaining observations. If the means are equal, then stop the iteration and it shows there are no clusters in the sample. If the means are not equal, there exists some discrimination between the variables in cluster 1 and for the remaining variables. Then repeat the process of Step 3.

**Step 4:** Repeat step No.1 and 2 for the remaining observations and ascertain the cluster 2.

**Step 5:** Continue the iteration process, until the nature of variance-covariance matrix of the $p$ variables in the last cluster is singular.

**Step 6:** In order to scrutinize the overall discriminant validity of the clusters, multivariate test of means should use with the assumption of the homogenous variance-covariance matrix.

## 3. Results and Discussion

In this section, we investigated the effectiveness of our proposed approach on the survey data collected from the famous two wheeler users' in India. The data comprised of 19 different attributes about the two wheeler company and the data was collected from 275 two wheeler users. A well-structured questionnaire was prepared and distributed to 300 two wheeler customers and the questions were anchored at five point likert scale from 1 to 5. After the data collection is over, only 275 completed questionnaires were used for analysis. The aim of this article is to describe the proposed clustering approach not the application of the theoretical concept. The following table shows the results extracted from the analysis by using SAS JMP v9.0 and STATA v11.2.

Table 1 visualizes the iteration summary of the identification of the multivariate outlier detection by using the $T$-square distance or the squared Mahalanobis distance. At first iteration, 275 observation and 19 variables were used to calculate the Mahalanobis distance for all observation. Among 275 observations, the value $T$-square statistic for 220 observations were below the UCL of $T$-square test statistic (29.53) at 5% significance level and the remaining No. of observations (55) are above the cut-off. Therefore, we consider the 55 observations

as first outlier cluster. Then repeat the iteration process to the next stage for calculating the $T$-square distance or the squared Mahalanobis distance based on 220 observations $(275 - 55)$ for the same 19 variables in iteration 2. Likewise, if we continue the iteration process for the remaining stages, the iteration reached the limit in the fifth step with 111 observations as outlier cluster No.5. At the iteration No.5, the variance-covariance matrix of 19 variables for 111 observations is singular, therefore it is not possible to calculate the $T$-square distance or the squared Mahalanobis distance for the observations. Hence based on 5 iterations, we identified five different outlier cluster at 5% significance level with $(n = 55)$, $(n = 44)$, $(n = 34)$, $(n = 31)$ and $(n = 111)$ observations respectively.

Table 1: Iteration summary for multivariate outlier identification

| Iteration | $(n)$ | UCL of $T$-square statistic* | Below UCL of $T$-square statistic $(n)$ | Above UCL of $T$-square statistic $(n)$ | Outlier clusters $(n)$ at 5% level |
|---|---|---|---|---|---|
| 1 | 275 | 29.53 | 220 | 55 | **55** |
| 2 | 220 | 29.37 | 176 | 44 | **44** |
| 3 | 176 | 29.17 | 142 | 34 | **34** |
| 4 | 142 | 28.93 | 111 | 31 | **31** |
| 5 | **111** | - | - | - | - |
| 6 | **-** | - | - | - | - |
| 7 | **-** | - | - | - | - |

| Iteration | $(n)$ | UCL of $T$-square statistic** | Below UCL of $T$-square statistic $(n)$ | Above UCL of $T$-square statistic $(n)$ | Outlier clusters $(n)$ at 1% level |
|---|---|---|---|---|---|
| 1 | 275 | 35.05 | 237 | 38 | **38** |
| 2 | 237 | 34.87 | 207 | 30 | **30** |
| 3 | 207 | 34.68 | 177 | 30 | **30** |
| 4 | 177 | 34.42 | 155 | 22 | **22** |
| 5 | 155 | 34.17 | 129 | 26 | **26** |
| 6 | 129 | 33.75 | 115 | 14 | **14** |
| 7 | **115** | - | - | - | - |

**$P$(number of variables) = 19    \*p-value < 0.05    \*\*p-value < 0.01**

Moreover we also identified the outlier clusters at 1% level. In iteration 1, 275 observations and 19 variables were used to calculate the Mahalanobis distance. Among the 275 observation, the value of $T$-square statistic for 237 observations was below the UCL of $T$-square test statistic (35.05) at 1% significance level and the remaining observations (38) are above the cut-off value. Therefore we finalize 38 observations as $1^{st}$ outlier cluster. By repeating this iteration process, finally we reached the final iteration No.7 with 115 observations as outlier cluster No.7. In the final iteration, it is not possible to calculate the $T$-square statistic

or the squared Mahalanobis distance because of the singularity of the variance-co-variance matrix. Hence, based on 7 iterations we identified 7 different outlier clusters at 1% significance level with ($n = 38$), ($n = 30$), ($n = 30$), ($n = 22$), ($n = 26$), ($n = 14$) and ($n = 115$) observations respectively. The iteration and identification of multivariate outlier clusters were explained with the help of the following multivariate control charts.

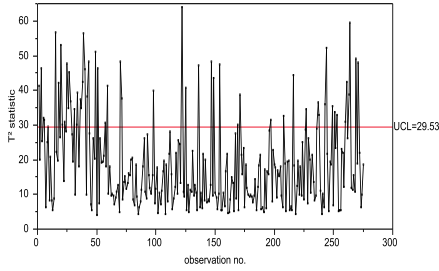**Multivariate control chart shows the classification of outliers with UCL of $T$-square statistic at 5% level**
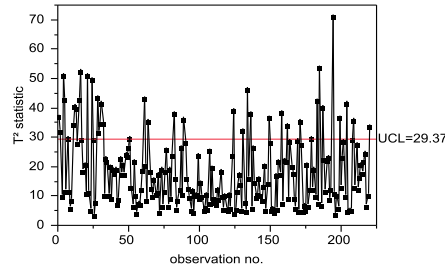


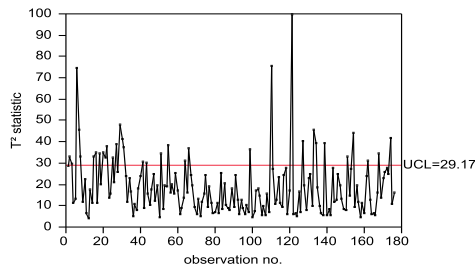Figure 1: Iteration 1 ($n = 275$)



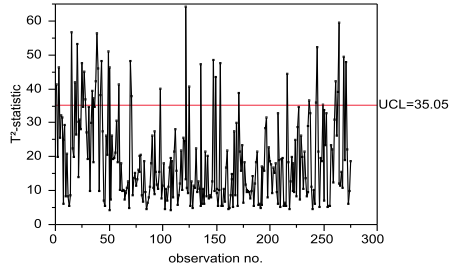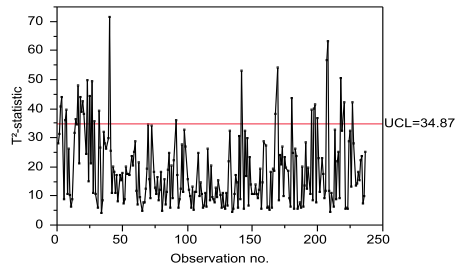Figure 2: Iteration 2 ($n = 220$)



Figure 3: Iteration 3 ($n = 176$)
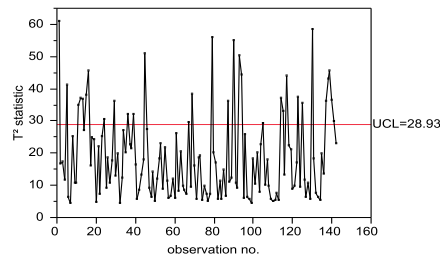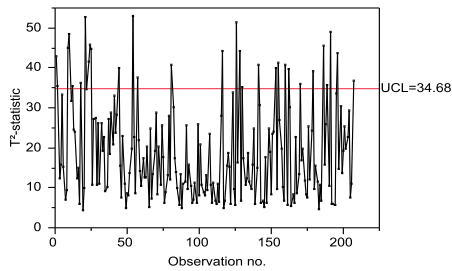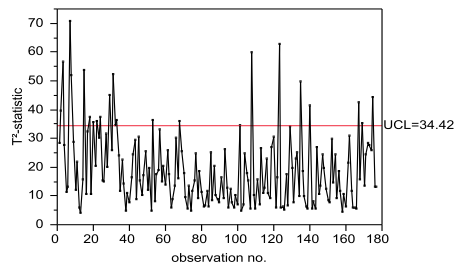


Figure 4: Iteration 4 ($n = 142$)

Table 2 describes the results of the five different test statistics such as Wilk's lambda, Pilla's trace, Lawley-Hotelling trace, Roy's largest root test and the traditional $F$-statistic which helps us to finalize the discriminant validity of the clusters based on the 19 variables at each iteration. In the first iteration, out of 275 observations, 55 are treated as outlier cluster 1 and remaining are inliers. The result of the test statistic confirms that the means of 19 variables are significantly differed at 1% between the outlier cluster 1 and the inliers. This indicates the variables in the outlier cluster are different from the inliers. This process is done at each iteration and we achieve a positive indication of attaining the discriminant validity between the outlier cluster and the remaining inliers. Finally, in the last iteration, it is not possible to segregate the new outlier cluster, because the variance-covariance matrix of the 19 variables for 111 observations is singular. So, the iteration is stopped and we treat the 111 observations as outlier cluster No.5.

**Multivariate control chart shows the classification of outliers with UCL of $T$-square statistic at 1% level**



Figure 5: Iteration 1 ($n = 275$)



Figure 6: Iteration 2 ($n = 237$)



Figure 7: Iteration 3 ($n = 207$)



Figure 8: Iteration 4 ($n = 177$)
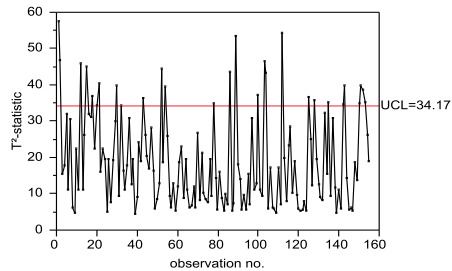


Figure 9: Iteration 5 ($n = 155$)
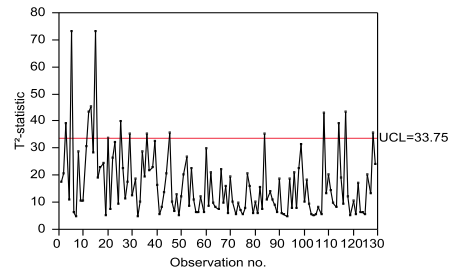


Figure 10: Iteration 6 ($n = 129$)

Similarly, the above said test statistic were also used to finalize the discriminant validity of the outlier clusters at 1% level. The result of the test statistic confirms that in all iterations, the means of the 19 variables are significantly differed between the outliers and the inliers at 1% significant level. Finally, in last iteration it is not possible to classify the new outlier cluster with ($n = 115$) observations because of a singularity of variance-covariance matrix. Hence the iteration was stopped here and we treat 115 observations as outlier cluster No.7. The following table shows the cluster wise means of the variables.

Table 3 exhibits the cluster wise centroids of the 19 variables. In order to test the equality of multivariate means of 19 variables among 5 outlier clusters, five

Table 2: Iteration summary for test of equality of means

| Iteration (n) | Outlier cluster (n) at 5% | Inliers (n) | Test statistic* | | | | F-ratio | d.f |
|---|---|---|---|---|---|---|---|---|
| | | | Wilk's lambda | Pillai's trace | Lawley-Hotelling trace | Roy's Largest root | | |
| 1 | 275 | 55 | 220 | 0.5404 | 0.4596 | 0.8505 | 0.8505 | 11.41 | (19,255) |
| 2 | 220 | 44 | 176 | 0.4721 | 0.5279 | 1.1180 | 1.1180 | 11.77 | (19,200) |
| 3 | 176 | 34 | 142 | 0.4334 | 0.5666 | 1.3074 | 1.3074 | 10.73 | (19,156) |
| 4 | 142 | 31 | 111 | 0.4090 | 0.5910 | 1.4452 | 1.4452 | 9.28 | (19,122) |
| 5 | 111 | - | - | - | - | - | - | - | - |
| 6 | - | - | - | - | - | - | - | - | - |
| 7 | - | - | - | - | - | - | - | - | - |

| Iteration (n) | Outlier cluster (n) at 1% | Inliers (n) | Test statistic* | | | | F-ratio | d.f |
|---|---|---|---|---|---|---|---|---|
| | | | Wilk's lambda | Pillai's trace | Lawley-Hotelling trace | Roy's Largest root | | |
| 1 | 275 | **38** | 237 | 0.6039 | 0.3961 | 0.6560 | 0.6560 | 8.80 | (19,255) |
| 2 | 237 | **30** | 207 | 0.6820 | 0.3180 | 0.4664 | 0.4664 | 5.33 | (19,217) |
| 3 | 207 | **30** | 177 | 0.5505 | 0.4495 | 0.8166 | 0.8166 | 8.04 | (19,187) |
| 4 | 177 | **22** | 155 | 0.4384 | 0.5616 | 1.2809 | 1.2809 | 10.58 | (19,157) |
| 5 | 155 | **26** | 129 | 0.5332 | 0.4668 | 0.8755 | 0.8755 | 6.22 | (19,135) |
| 6 | 129 | **14** | 115 | - | - | - | - | - | - |
| 7 | **115** | - | - | - | - | - | - | - | - |

*P(number of variables) = 19*          **p-value < 0.01*

different test statistic such as Wilk's lambda, Pilla's trace, Lawley-Hotelling trace, Roy' largest root test and the traditional $F$-statistic which helps to strongly establish the discriminant validity among the clusters. From Table 4, the result of the battery of multivariate test confirms that the means of variables among the 5 outlier clusters are significantly differed at 1% level. This indicates all clusters are different and each outlier cluster conveys different meaning which leads to achieve the overall discriminant validity among the clusters.

Similarly the above said test of multivariate means also utilized to check the differences among the means of 19 variables for the outlier clusters at 1% level. The result of the test confirms the means of the variables among the 7 outlier clusters are significantly differed at 1% level. This indicates the entire outlier clusters at 1% level are different and each clusters conveys different meaning which leads to achieve the overall discriminant validity among the clusters. The following graph visualizes the summary of membership of each observation in each outlier cluster.

Table 3: Cluster-wise means of variables

| S.No. | Variable Name | outlier clusters at 5% level | | | | | outlier clusters at 1% level | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | cluster 1 (n=55) | cluster 2 (n=44) | cluster 3 (n=34) | cluster 4 (n=31) | cluster 5 (n=111) | cluster 1 (n=38) | cluster 2 (n=30) | cluster 3 (n=30) | cluster 4 (n=22) | cluster 5 (n=26) | cluster 6 (n=14) | cluster 7 (n=115) |
| 1 | Top of the mind awareness | 4.07 | 4.27 | 4.59 | 4.74 | 4.88 | 4.11 | 4.17 | 4.27 | 4.59 | 4.54 | 4.79 | 4.89 |
| 2 | Fuel efficient bike | 3.25 | 3.68 | 4.68 | 4.77 | 4.94 | 3.00 | 3.70 | 3.90 | 4.27 | 4.85 | 4.93 | 4.91 |
| 3 | Four stroke bike | 3.64 | 3.8 | 4.44 | 4.74 | 4.98 | 3.61 | 3.80 | 3.83 | 4.09 | 4.73 | 4.64 | 4.99 |
| 4 | Brand familiarity | 3.96 | 4.34 | 4.56 | 4.61 | 5 | 4.00 | 4.07 | 4.33 | 4.41 | 4.81 | 4.43 | 4.99 |
| 5 | Satisfaction | 3.76 | 4.32 | 4.44 | 4.81 | 5 | 3.66 | 4.10 | 4.37 | 4.36 | 4.69 | 4.99 | 5.00 |
| 6 | Commitment | 3.73 | 4.11 | 4.32 | 4.68 | 5 | 3.53 | 3.93 | 4.30 | 4.14 | 4.69 | 4.57 | 5.00 |
| 7 | Liking | 3.51 | 3.7 | 4.24 | 4.71 | 4.93 | 3.47 | 3.50 | 3.90 | 4.05 | 4.54 | 4.57 | 4.94 |
| 8 | Price premium | 2.38 | 2.64 | 2.5 | 2.16 | 1.91 | 2.34 | 2.47 | 2.40 | 3.23 | 2.12 | 2.29 | 1.90 |
| 9 | Best in category | 3.67 | 4.07 | 4.21 | 4.68 | 5 | 3.55 | 3.93 | 3.97 | 4.05 | 4.69 | 4.71 | 5.00 |
| 10 | Popularity | 3.55 | 4.27 | 4.38 | 4.58 | 5 | 3.42 | 3.93 | 4.33 | 4.18 | 4.54 | 4.71 | 5.00 |
| 11 | Band leader | 3.45 | 3.89 | 4.06 | 4.13 | 4.55 | 3.21 | 3.93 | 3.90 | 3.82 | 4.08 | 4.50 | 4.54 |
| 12 | Innovation | 3.6 | 3.98 | 3.74 | 4.26 | 4.3 | 3.53 | 3.93 | 3.83 | 3.64 | 4.15 | 4.36 | 4.29 |
| 13 | Esteem | 3.91 | 3.82 | 3.97 | 4.26 | 4.46 | 3.95 | 3.57 | 3.90 | 4.05 | 4.35 | 4.43 | 4.41 |
| 14 | Performance | 3.84 | 3.64 | 4.03 | 4.13 | 4.37 | 3.84 | 3.60 | 3.80 | 3.86 | 4.15 | 4.14 | 4.37 |
| 15 | Value association | 3.87 | 3.98 | 4.09 | 4.32 | 4.27 | 3.89 | 3.87 | 3.97 | 4.09 | 4.08 | 4.43 | 4.30 |
| 16 | Organizational association | 3.73 | 3.59 | 3.56 | 3.84 | 3.73 | 3.74 | 3.70 | 3.47 | 3.82 | 3.73 | 3.79 | 3.70 |
| 17 | Brand differentiation | 3.64 | 3.98 | 3.94 | 4.45 | 4.84 | 3.58 | 4.00 | 3.93 | 3.95 | 4.31 | 4.07 | 4.82 |
| 18 | Celebrity association | 3.24 | 3.41 | 2.68 | 3.45 | 3.59 | 3.37 | 3.27 | 3.50 | 2.73 | 2.69 | 3.57 | 3.59 |
| 19 | Animal association | 2.69 | 2.59 | 2.09 | 3.1 | 2.88 | 2.68 | 2.80 | 2.43 | 2.18 | 2.46 | 3.21 | 2.90 |

Table 4: Test of equality of cluster means with homogenous variance covariance matrix

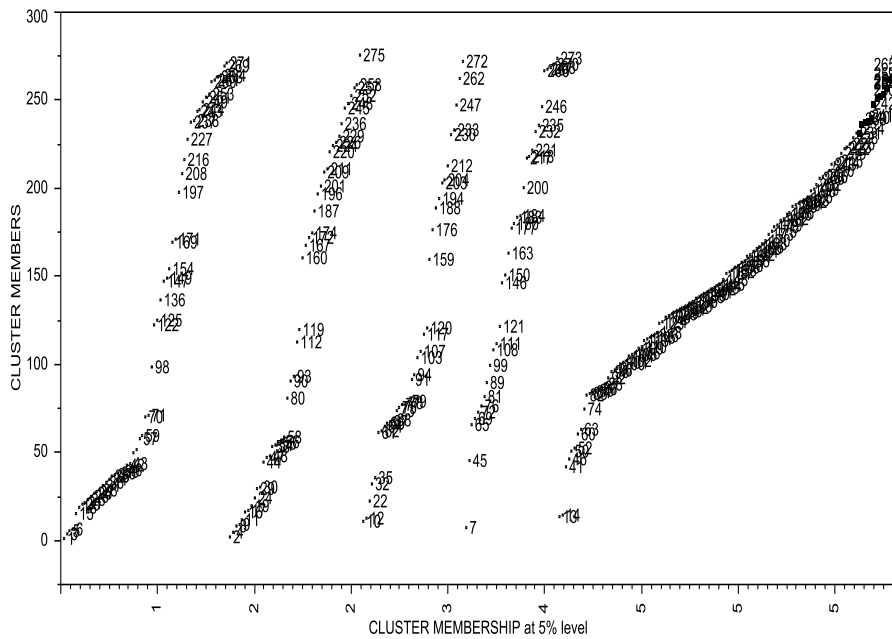| Multivariate Test | outlier clusters at 5% level | | | | outlier clusters at 1% level | | | |
|---|---|---|---|---|---|---|---|---|
| | Test value | F-ratio | d.f | p-value | Test value | F-ratio | d.f | p-value |
| Wilk's lambda | 0.2474 | 5.57 | (76,995) | < 0.01 | 0.2124 | 3.92 | (114,1446) | < 0.01 |
| Pillai's trace | 0.9016 | 3.91 | (76,1020) | < 0.01 | 1.0074 | 2.71 | (114,1530) | < 0.01 |
| Lawley-Hotelling trace | 2.4762 | 8.16 | (76,1002) | < 0.01 | 2.7783 | 6.05 | (114,1490) | < 0.01 |
| Roy's Largest root | 2.2477 | 30.17 | (19,255) | < 0.01 | 2.4552 | 32.95 | (19,255) | < 0.01 |

$P(number\ of\ variables) = 19$



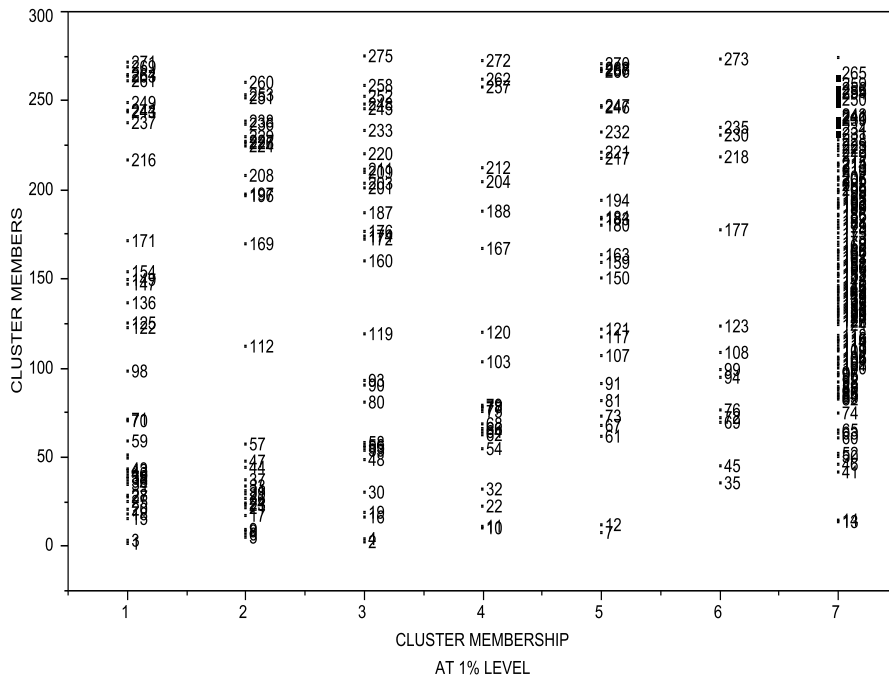Figure 11: Cluster membership for outlier clusters at 5% level

Figure 12: Cluster membership for outlier clusters at 1% level

## 4. Conclusion

In this paper a new method of clustering was proposed based on Multivariate outlier detection. Though several clustering procedures available in the literature, the proposed technique gives a unique idea to cluster the sample observations in a survey study based on the multivariate outliers. The feature of the proposed clustering technique was elaborately discussed and the authors also highlighted the application of the technique in a survey research. Based on the results derived, the proposed technique gives more insights to the researcher to cluster the sample observation at 5% and 1% significance level. Finally the authors enlighten an idea for further research by conducting simulation experiments for testing relationship between the significance level and the number of outlier clusters extracted. Moreover more rigorous experiments may conduct to identify the Multivariate outliers' inside the outlier clusters.

## Acknowledgements

Authors would like to thank the referee for his/her insightful comments, which help to improve the earlier version of this manuscript.

## References

Acuna, E. and Rodriguez, C. (2004). A meta analysis study of outlier detection methods in classification. Technical paper, Department of Mathematics, University of Puerto Rico at Mayaguez. http://academic.uprm.edu eacuna/paperout.pdf

Almeida, J. A. S., Barbosa, L. M. S., Pais, A. A. C. C. and Formosinho, S. J. (2007). Improving hierarchical cluster analysis: a new method with outlier detection and automatic clustering. *Chemometrics and Intelligent Laboratory Systems* **87**, 208-217.

Angiulli, F. and Pizzuti, C. (2005). Outlier mining in large high-dimensional data sets. *IEEE Transactions on Knowledge and Data Engineering* **17**, 203-215.

Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data*, 3rd edition. Wiley, New York.

Blake, C. L. and Merz, C. J. (1998). UCI repository of machine learning databases. University of California, Irvine. http://mlearn.ics.uci.edu/ML-Repository.html

Bolton, R. J. and Hand, D. J. (2002). Statistical fraud detection: a review. *Statistical Science* **17**, 235-255.

Bradley, P. S., Fayyad, U. M. and Mangasarian, O. L. (1999). Mathematical programming for data mining: formulations and challenges. *INFORMS Journal on Computing* **11**, 217-238.

Breunig, M., Kriegel, H. P., Ng, R. T. and Sander, J. (2000). Identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data.* ACM Press, 93-104. Dallas, Texas

Chiang, J. H. and Yin, Z. X. (2007). Unsupervised minor prototype detection using an adaptive population partitioning algorithm. *Pattern Recognition* **40**, 3132-3145.

Chiu, A. L. and Fu, A. W. (2003). Enhancement on local outlier detection. In *Proceedings of the Seventh International Database Engineering and Applications Symposium (IDEAS'03)*, 298-307.

Cutsem, B. V. and Gath, I. (1993). Detection of outliers and robust estimation using fuzzy clustering. *Computational Statistics and Data Analyses* **15**, 47-61.

Dudoit, S. and Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology* **3**, 1-21.

Gath, I. and Geva, A. B. (1989). Fuzzy clustering for the estimation of the parameters of the components of mixtures of normal distributions. *Pattern Recognition Letters* **9**, 77-86.

Han, J. and Kamber, M. (2006). *Data Mining*: *Concepts and Techniques*, 2nd edition. Morgan Kaufmann, New York.

Hawkins, D. (1980). *Identifications of Outliers.* Chapman and Hall, London.

Hodge, V. J. and Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review* **22**, 85-126.

Hotelling, H. (1951). A generalized $T$ test and measure of multivariate dispersion. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability* (Edited by J. Neyman), 23-41. University of California Press, Berkeley.

Jiang, M. F., Tseng, S. S. and Su, C. M. (2001). Two-phase clustering process for outlier detection. *Pattern recognition Letters* **22**, 691-700.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for Clustering Data.* Prentice-Hall, New Jersey.

Knorr, E. M. and Ng, R. T. (1998). Algorithms for mining distance-based outliers in large datasets. In *Proceedings of the 24th International Conference on Very Large Databases* (*VLDB*) *Conference*, 392-403. New York.

Knorr, E. M. and Ng, R. T. and Tucakov, V. (2000). Distance-based outliers: algorithms and applications. *VLDB Journal* **8**, 237-253.

Van der Laan, M., Pollard, K. S. and Bryan, J. (2003). A new partitioning around medoids algorithms. *Journal of Statistical Computation and Simulation* **73**, 575-584.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data*: *An Introduction to Cluster Analysis.* Wiley, New York.

Lane, T. and Brodley, C. E. (1999). Temporal sequence learning and data reduction for anomaly detection. *ACM Transactions on Information and System Security* **2**, 295-331.

Loureiro, A., Torgo, L. and Soares, C. (2004). Outlier detection using clustering methods: a data cleaning application. In *Proceedings of KDNet Symposium on Knowledge-based Systems for the Public Sector*. Bonn, Germany.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability* **1** (Edited by L. M. Le Cam and J. Neyman), 281-297. University of California Press, Berkeley, California.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. In *Proceedings of the National Institute of Sciences of India* **2**, 49-55.

Niu, K., Huang, C., Zhang S. and Chen, J. (2007). ODDC: outlier detection using distance distribution clustering. *PAKDD 2007 Workshops, Lecture Notes in Artificial Intelligence* (*LNAI*) 4819 (Edited by T. Washio *et al.*), 332-343. Springer-Verlag, New York.

Ng, R. and Han, J. (1994). Efficient and effective clustering methods for spatial data mining. In *Proceedings of the 20th International Conference on Very Large Data Bases* (*VLDB*), 144-155. Santiago, Chile.

Papadimitriou, S., Kitagawa, H., Gibbons, P. B. and Faloutsos, C. (2003). LOCI: fast outlier detection using the local correlation integral. In *Proceedings of the 19th International Conference on Data Engineering* (*ICDE*), 315-326. Bangalore, India.

Ramaswami, S., Rastogi, R. and Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 427-438. Dallas, Texas.

Rousseeuw, P. and Leroy, A. (1996). *Robust Regression and Outlier Detection*, 3rd edition. Wiley, New York.

Yoon, K. A., Kwon, O. S. and Bae, D. H. (2007). An approach to outlier detection of software measurement data using the k-means clustering method. *First International Symposium on Empirical Software Engineering and Measurement*, 443-445. Madrid.

Zhang, Q. and Couloigner, I. (2005). A new and efficient k-medoid algorithm for spatial clustering. In *ICCSA, Lecture Notes in Computer Science* (*LNCS*) 3482 (Edited by O. Gervasi *et al.*), 181-189. Springer-Verlag, Singapore.

Zhang, J. and Wang, H. (2006). Detecting outlying subspaces for high-dimensional data: the new task, algorithms, and performance. *Knowledge and Information Systems* **10**, 333-355.

G. S. David Sam Jayakumar
Jamal Institute of Management
Jamal Mohamed College
Tiruchirappalli-620 020, South India, India
samjaya77@gmail.com

Bejoy John Thomas
Jamal Institute of Management
Jamal Mohamed College
Tiruchirappalli-620 020, South India, India
bejoy66@gmail.com