Statistics and Applications {ISSN 2452-7395(online)}
Volume 18, No. 2, 2020 (New Series), pp 249–261

Some Computational Issues Related to Common Clustering Techniques

Asis Kumar Chattopadhyay
Department of Statistics
University of Calcutta, Kolkata, India

Received: 29 March 2020; Revised: 02 August 2020; Accepted: 04 August 2020

Abstract

In the present work attempts have been made to highlight different computational problems related to common clustering and dimensionality reduction techniques depending on input data type and underlying model assumptions of the different statistical methods. As clustering and dimensionality reduction techniques are widely used under machine learning and big data analysis, it is very much necessary to highlight the limitations to the user community (especially for the software industry). The effects of directional and missing data have also been considered.

Key words: Clustering; Computation; Directional Data; Missing Data.

1. Introduction

Cluster Analysis, also called data segmentation, has a variety of goals. All relate to grouping or segmenting a collection of objects (also called observations,individuals, cases, or data rows) into subsets or "clusters", such that those within each cluster are more closely related to one another than objects assigned to different clusters. Central to all of the goals of cluster analysis is the notion of degree of similarity (or dissimilarity) between the individual objects being clustered. There are two major methods of clustering - hierarchical clustering and k-means clustering.

Statistical techniques for classification are essentially of two types. Members of the first type are used to construct a sensible and informative classification of an initially unclassified set of data; these are known as cluster analysis methods. The information on which the derived classification is based is generally a set of variable values recorded for each object or individual in the investigation, and clusters are constructed so that individuals within clusters are similar with respect to their variable values and different from individuals in other clusters. The second set of statistical techniques concerned with classification is known as discriminant or assignment methods. Here the classification scheme is known a priori and

Correponding Author: Asis Kumar Chattopadhyay
Email: akcstat@caluniv.ac.in

the problem is how to devise rules for allocating unclassified individuals to one or other of the known classes.

Different Statistical techniques are available for clustering and classification (Fraix Burnet *et al.* (2015), De *et al.* (2013) and references there in). But depending on the nature of the different types of data the following problems often arise and in some cases a proper solution is still not available.

1. Sometimes the data set under consideration has a distributional form (usually normal) and sometimes it is of non normal nature. Based on the above point, there is a justification needed about which clustering or classification technique should be used so that it reflects the proper nature of the data set provided. This problem is more relevant for classification as most of the classification methods are model based. For clustering most of the methods are non parametric in nature and as such the above problem is not very serious. But here also basic assumption is that the nature of the variables under study are continuous where as under practical situations these may be categorical like binary, nominal, ordinal and even directional (particularly for environmental and Astronomical data). Under such situations standard similarity/ dissimilarity measures will not work.

2. The clustering techniques which require an inherent model assumption are known as Model Based Methods, whereas the clustering technique where no modelling assumption or distributional form is needed may be termed as Non-Model based Methods. Hence based on the nature of data set, one has to decide about proper application of the two types of techniques.

3. Even if one decides about the proper methods for the data set at hand, there are several techniques available under both the categories and no predefined criteria can be set to judge which technique is the best for the situation under consideration.

4. The above point arises the need of a comparative study among various available techniques and a computational analysis of all the methods.Once all the methods are implemented, it requires a criterion to decide upon the best technique based on a post classifier. So an appropriate post classification approach is also needed in this regard. For a post classification approach, a pre-classifier or training sample is required. Since in this type of techniques a prior knowledge of classification is provided, these are called Supervised Learning. All other techniques where no prior classification is provided are known as Unsupervised Learning.

5. A comparative validity algorithm may be helpful for predicting the superiority of different techniques.

6. At present big data issues related to data size is quite common. In statistical terms this problems may be tackled in terms of both the number of observations and the variables considered. Many standard clustering techniques fails to deal with such big data sets. Thus some dimension reduction methods may be applied at first and then clustering may be performed on the reduced data set. Some data mining techniques are very helpful under such situations.

9. The above criteria also needs to be validated depending on whether the data is Gaussian or non-Gaussian. That means the dimension reduction techniques may vary according as the data set has a distributional form or not.

10. Finally and most importantly after all these considerations, the similarity of grouping of objects obtained from different methods should be checked in terms of some physical properties .

2. Hierarchical Clustering Technique

Central to all of the goals of cluster analysis is the notion of degree of similarity (or dissimilarity) between the individual objects being clustered. There are two major methods of clustering -hierarchical clustering and k-means clustering. In hierarchical clustering the data are not partitioned into a particular cluster in a single step. Instead, a series of partitions takes place, which may run from a single cluster containing all objects to n clusters each containing a single object. Hierarchical Clustering is subdivided into agglomerative methods, which proceed by series of fusions of the n objects into groups, and divisive methods, which separate n objects successively into finer groups. Agglomerative techniques are more commonly used. Hierarchical clustering may be represented by a two dimensional diagram known as dendrogram which illustrates the fusions or divisions made at each successive stage of analysis.

2.1. Agglomerative method

An agglomerative hierarchical clustering procedure produces a series of partitions of the data, $C_n, C_{n-1}, \ldots.., C_1$. The first C_n consists of n single object 'clusters', the last C_1, consists of single group containing all n cases.

At each particular stage the method joins together the two clusters which are closest together (most similar). (At the first stage, of course, this amounts to joining together the two objects that are closest together, since at the initial stage each cluster has one object.) Differences between methods arise because of the different ways of defining distance (or similarity) between clusters.

A key step in a hierarchical clustering is to select a distance measure. A simple measure is Manhattan distance, equal to the sum of absolute distances for each variable. The name comes from the fact that in a two-variable case, the variables can be plotted on a grid that can be compared to city streets, and the distance between two points is the number of blocks a person would walk.

A more common measure is Euclidean distance, computed by finding the square of the distance between each variable, summing the squares, and finding the square root of that sum. In the two-variable case, the distance is analogous to finding the length of the hypotenuse in a triangle; that is, it is the distance as the crow flies. A review of cluster analysis in health psychology research found that the most common distance measure in published studies in that research area is the Euclidean distance or the squared Euclidean distance.

To calculate distance between two clusters it is required to define two representative points from the two clusters. Different linkage measures like “single linkage”, “complete linkage”, “average linkage” *etc* have been proposed for this purpose.

2.2. Similarity measure for mixed type data

The above mentioned dissimilarity/similarity measures are applicable to continuous type data only. But generally we work with mixed type data sets which includes different types like continuous, discrete, binary, nominal, ordinal *etc.* Gower (1971) has proposed a general measure known as Gower’s coefficient of similarity. Two individuals i and j may be compared on a character k and assigned a score s_{ijk}. There are many ways of calculating s_{ijk}, some of which are described below.

Corresponding to n individuals and p variables, Gower’s similarity index S_{ij} is defined as

$$S_{ij} = \Sigma_{k=1}^{p} s_{ijk} / \Sigma_{k=1}^{p} \delta_{ijk} (i, j = 1, 2, \ldots n)$$

$$\begin{aligned} \text{where } \delta_{ijk} &= 1 \quad \text{when character } k \text{ can be compared for} \\ & \qquad\; \text{observations } i \text{ and } j \\ &= 0 \quad \text{otherwise} \end{aligned}$$

For continuous (quantitative) variables with values $x_{1k}, x_{2k}, \ldots, x_{nk}$ for the kth variable

$$s_{ijk} = 1 - \mid x_{ik} - x_{jk} \mid / R_k$$

where R_k is the range of the variable k and may be the total range in population or the range in the sample.

For a categorical (qualitative) character with m categories ($m = 2$ for binary variable)

$$\begin{aligned} s_{ijk} &= 0 \text{ if } i \text{ and } j \text{ are totally different} \\ &= q \text{ (positive fraction) if there is some degree of agreement} \\ &= 1 \text{ when } i \text{ and } j \text{ are same} \end{aligned}$$

2.3. Linkage measures

To calculate distance between two clusters it is required to define two representative points from the two clusters. Different methods have been proposed for this purpose. Some of them are listed below.

Single linkage: One of the simplest methods is single linkage, also known as the nearest neighbor technique. The defining feature of the method is that distance between clusters is defined as the distance between the closest pair of objects, where only pairs consisting of one object from each cluster are considered.

In the single linkage method, d_{rs} is computed as $d_{rs} =$ Min d_{ij}, where object i is in cluster r and object j is in cluster s and d_{ij} is the distance between the objects I and j. Here the distance between every possible object pair (i, j) is computed, where object i is in cluster r and object j is in cluster s. The minimum value of these distances is said to be the distance between clusters r and s. In other words, the distance between two clusters is given by the value of the shortest link between the clusters. At each stage of hierarchical clustering, the clusters r and s, for which d_{rs} is minimum, are merged.

Complete linkage: The complete linkage, also called farthest neighbor, clustering method is the opposite of single linkage. Distance between clusters is now defined as the distance between the most distant pair of objects, one from each cluster. In the complete linkage method, $d - rs$ is computed as $d_{rs} =$ Max d_{ij}, where object i is in cluster r and object j is cluster s. Here the distance between every possible object pair (i, j) is computed, where object i is in cluster r and object j is in cluster s and the maximum value of these distances is said to be the distance between clusters r and s. In other words, the distance between two clusters is given by the value of the largest distance between the clusters. At each stage of hierarchical clustering, the clusters r and s, for which d_{rs} is minimum, are merged.

Average linkage: Here the distance between two clusters is defined as the average of distances between all pairs of observations, where each pair is composed of one object from each group. In the average linkage method, d_{rs} is computed as $d_{rs} = Trs/(Nr \times Ns)$ where Trs is the sum of all pairwise distances between cluster r and cluster s. Nr and Ns are the sizes of the clusters r and s respectively. At each stage of hierarchical clustering, the clusters r and s, for which d_{rs} is the minimum, are merged.

Minimax Linkage: This was introduced by Bien and Tibshirani (2011). For any point x and cluster G, define

$$d_{\max}(x, G) = \max_{y \in G} d(x, y)$$

as the distance to the farthest point in G from x. Define the minimax radius of the cluster G as

$$r(G) = \min_{x \in G} d_{\max}(x, G)$$

that is, find the point $x \in G$ from which all points in G are as close as possible. This minimizing point is called the prototype for G. It may be noted that a closed ball of radius

$r(G)$ centered at the prototype covers all of G. Finally we define the minimax linkage between two clusters G and H as

$$d(G, H) = r(GUH)$$

that is, we measure the distance between clusters G and H by the minimax radius of the resulting merged cluster.

It is very important to choose a proper linkage measure in a particular situation. A liberal attitude always leads to single linkage whereas a conservative attitude leads to complete linkage. Minimax is a good choice when one tries to avoid a wrong decision (loss is more important than gain) and without any prior belief, average linkage may give the best answer.

2.4. Optimum number of clusters

Usually the number of clusters are determined from the dendrogram and validated by the physical properties. We specify a horizontal line for a particular similarity/dissimilarity value and the clusters below this line are selected as optimum. But some mathematical rules (thumb rules) are also available which are based on between cluster and within cluster sum of squares values. If we denote by k, the number of clusters and define by $W(k)$ the sum of the within cluster sum of squares for k clusters then the values of $W(k)$ will gradually decrease with increase in k and that "k" may be taken as optimum where $W(k)$ stabilizes. For detailed discussion on may follow the link http://www.cc.gatech.edu/∼hpark/papers/cluster JOGO.pdf (by Jung *et al.* (2002)).

3. Partitioning Clustering - k-means Method

The k-means algorithm (MacQueen, 1967) assigns each point to the cluster whose center (also called centroid) is nearest. The center is the average of all the points in the cluster that is, its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster. This method can be used for clustering of objects and not variables.

This method starts with a value of k. We will discuss later the method of selection of the value of k. Then we randomly generate k clusters and determine the cluster centers, or directly generate k seed points as cluster centers. Assign each point to the nearest cluster center in terms of Euclidean distance. Re-compute the new cluster centers. Repeat until some convergence criterion is met *i.e.* there is no reassignment. The main advantages of this algorithm are its simplicity and speed which allows it to run on large data sets. Its disadvantage is that it is highly dependent on the initial choice of clusters. It does not yield the same result with each run, since the resulting clusters depend on the initial random assignments. It maximizes inter-cluster variance and minimizes intra-cluster variance.

The advantages of partitioning method are as follows:

(a) A partitioning method tries to select best clustering with k groups which is not the goal of hierarchical method.

(b) A hierarchical method can never repair what was done in previous steps.

(c) Partitioning methods are designed to group items rather than variables into a collection of k clusters.

(d) Since a matrix of distances (similarities) does not have to be determined and the basic data do not have to be stored during the computer run partitioning methods can be applied to much larger data sets.

For k-means algorithms (Hartigan, 1975) the optimum value of k can be obtained in different ways. On the basis of the method proposed by Sugar and James (2003), by using k-means algorithm first determine the structures of clusters for varying number of clusters taking $k = 2, 3, 4$ *etc.* For each such cluster formation compute the values of a distance measure

$$d_K = (1/p)\min_x E[(x_k - c_k)'(x_k - c_k)]$$

which is defined as the distance of the x_k vector (values of the parameters) from the center c_k (which is estimated as mean value), p is the order of the x_k vector. Then the algorithm for determining the optimum number of clusters is as follows. Let us denote by d'_k the estimate of d_k at the kth point which is actually the sum of within cluster sum of squares over all k clusters. Then d'_k is the minimum achievable distortion associated with fitting k centers to the data. A natural way of choosing the number of clusters is plot d'_k versus k and look for the resulting distortion curve. This curve is always monotonic decreasing. Initially one would expect much smaller drops *i.e.* a levelling off for k greater than the true number of clusters because past this point adding more centers simply partitions within groups rather than between groups.

According to Sugar and James (2003) for a large number of item versus transformed d'_k. Then calculate the jumps in the transformed distortion as

$$J_k = (d_k'^{(-(p/2))} - d_{k-1}'^{(-(p/2))})$$

Another way of choosing the number of clusters is plot J_k versus k and look for the resulting jump curve. The optimum number of clusters is the value of k at which the distortion curve levels off as well as its value associated with the largest jump.

The k-means clustering technique depends on the choice of initial cluster centers (Chattopadhyay *et al.*, 2012). But this effect can be minimized if one chooses the cluster centers through group average method (Milligan, 1980). As a result, the formation of the final groups will not depend heavily on the initial choice and hence will remain almost the same according to physical properties irrespective of initial centers. In MINITAB package, the k-means method is almost free from the effect of initial choice of centers as they have used the group average method.

3.1. Advantages and disadvantages of k-means algorithm

The main advantages of this algorithm is that it is very fast (in terms of computational speed), robust, easy to understand and interpret. In fact the algorithm has been modified by Hartigan and Wong (1979) which speeds up the algorithm and is used most commonly in the community. The open-source statistical computing environment R (https://cran.r-project.org/); the software which is used in this entire work; has the built-in function kmeans() which implements the above discussed version of the k-means algorithm as its default. The algorithm is very much well suited for the data which are distinct and well-separated from each other. The clusters thus formed are tight and often tighter than the Hierarchical Clustering method, especially when the clusters are globular. But the algorithm suffers due to a number of reasons. k -means depends heavily on the initialization/ seeds. The algorithm assumes the joint distribution of the features within each cluster to have equal variance and to be independent of each other. This assumption is hard to satisfy more than often. Correlation between the features breaks this assumption. k-means cannot find non-convex clusters or the clusters with unusual shapes or overlapping clusters. Finally, this algorithm requires a priori knowledge on the number of clusters/groups to be formed. This is most commonly tackled by using the method proposed by Sugar and James (2003) which has been discussed earlier. Jump-Statistic as a mean of determining the number of clusters "k" is very popular and widely accepted measure. Other possibilities are the uses of gap statistic or silhouette index.

3.2. Example using k-means algorithm

The Fisher's *Iris* data set is a multivariate data set introduced by R.A. Fisher (Fisher (1936)). It is also known as Anderson's *Iris* data set because Edgar Anderson collected the data to quantify the morphologic variation of Iris flowers of three related species. The data set consists of 50 samples from each of three species of Iris (Iris setosa (type-3), Iris versicolor (type-2) and Iris virginica (type-1)). Four features were measured from each sample: the length and the width of the sepals and petals, in centimetres.

We have performed k-means clustering of the data on the basis of the four variables viz. sepal length, sepal width, petal length and petal width. Choosing $k = 3$, we have divided the 150 observations into three groups in order to verify whether we can identify three groups corresponding to three species. From our analysis it is clear that k-means method has correctly identified Iris setosa (type-3) species for all the 50 cases where as there are some errors corresponding to types 1 and 2. For type 2 three cases and for type 1 fourteen cases had wrongly identified. The summary result for k-means clustering is given below:

Number of clusters: 3

	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	39	25.414	0.732	1.552
Cluster2	61	38.291	0.731	1.647
Cluster3	50	15.151	0.482	1.248

4. Clustering of Variables

The hierarchical clustering method can also be used for clustering of variables on the basis of the observations. Here instead of the distance matrix one may start with the correlation matrix (higher correlation indicating similarity of variables). The linkage measures as listed in the previous section will not be applicable for variable clustering. In order to measure similarity/dissimilarity between two clusters of variables, one may either use the correlation between first principal components corresponding to the two clusters or the canonical correlations.

Dimensionality reduction techniques like Principal Component Analysis (PCA) or Independent Component Analysis (ICA) could alternatively be used for variable clustering. The variables with larger loading belonging to a particular component may be considered to be in the same cluster.

4.1. Principal Component Analysis (PCA)

In this technique, given a data set of observations on correlated variables, an orthogonal transformation is performed to convert it into a set of uncorrelated variables called the principal components. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance. One rule of thumb is to consider those components whose variances are greater than one in the reduced space. Principal components are guaranteed to be independent only if the variables are jointly normally distributed.

4.2. Independent Component Analysis (ICA)

One of the most recent powerful statistical techniques for analyzing large data sets is independent component analysis (ICA), see Comon (1994) for the original description of ICA. Such data sets are generally multivariate in nature. The common problem is to find a suitable representation of the multivariate data. For the sake of computational and conceptual simplicity such representation is sought as a linear transformation of the original data. Principal component analysis, factor analysis, projection pursuit are some popular methods for linear transformation. But ICA is different from other methods, because it looks for the components in the representation that are both statistically independent and non-Gaussian.

In essence, ICA separates statistically independent component data, which is the original source data, from an observed set of data mixtures. All information in the multivariate data sets are not equally important. We need to extract the most useful information. Independent component analysis extracts and reveals useful hidden factors from the whole data sets. ICA defines a generative model for the observed multivariate data, which is typically given as a large database of samples. See Hyvarinen *et al.* (2001), Comon and Jutten (2010) and Lee (1998) for book length discussions on ICA. ICA can be applied in various fields like neural network (Fiori, 2003), studying EEG data (Bartlett *et al.*, 1995), speech processing (Kumaran *et al.* 2005), brain imaging (McKeown *et al.*, 1997), signal separation (Adali *et al.*, 2009), telecommunications (Hyvarinen *et al.*, 2002), econometrics (Bonhomme and Robin, 2009), *etc.* Chattopadhyay *et al.* (2012) has applied ICA for astronomical data set.

4.3. Conversion of directional data to linear

Note that PCA or ICA has been developed for linear continuous data but if one variable, is circular in nature then the method will not work.. But it is not immediate how to include this type of data for clustering directly or through PCA or ICA. If a density plot of the data show the circular variable has a bimodal distribution and the two modes are near 0^o and 200^o, we may be motivated to consider two main directions, say east and west (approximately), which correspond to 0^o and 180^o.

Chattopadhyay *et al.* (2015) proposed a method of conversion from circular to linear where they considered standard cosine angular distance of an angle θ from a fixed angle ϕ, defined by $d_\phi = 1 - \cos(\theta - \phi)$, which is in the linear scale, and $d \in [0, 2]$. Thus, for a circular variable θ , we may consider two distances $d_0 = 1 - \cos(\theta - 0^o)$ and $d_{180} = 1 - \cos(\theta - 180^o)$, both of which are linear. So, instead of taking θ in our analysis, we may consider the pair $(d_{\max}, d_{sign})$, where $d_{\max} = \max(d_0, d_{180})$ and $d_{sign} = +1$ if $d_{\max} = d_0$ and $d_{sign} = -1$ if $d_{\max} = d_{180}$. Alternately, if we want to ignore the sign we can work with $\theta^* = 2 \times \theta$, which is approximately unimodal with mode near 45^o. We may work with $d^* = 1 - \cos(\theta^* - 45^o)$.

5. Incomplete Data problems

Statistical analysis with missing data is an important problem as the problem of missing observation is very common in many situations. During the last two decades different methods have been developed to tackle the situation.One possible way to handle missing values is to remove either all features or all objects that contain missing values. Another possibility is imputation where we fill in the missing values by inferring new values for them. The imputation method may not be applicable to some astronomical data sets (Chattopadhyay, 2017) as the missing value may arise from physical process and imputing missing values is misleading and can skew subsequent analysis of data. For example, the Lyman break technique (Giavalisco, 2002) can identify high-redshift galaxies based on the absence of detectable emissions in bands corresponding to the FUV rest frame of the objects. Such high-redshift galaxies were previously unobservable.

Missing values occur for a variety of reasons, from recording problems to instrument limitations to unfavorable observing conditions. In particular, when data are combined from multiple archives or instruments, it is virtually certain that some objects will not be present in all of the contributing sources. Little and Rubin (1987) identified three models for missing data. When values are Missing At Random (MAR, MCAR), imputation may be a reasonable approach since the values may be predicted from the observed values. The third type of missing values are Not Missing at Random (NMAR), when the value itself determines whether it is missing. This is precisely the case when objects fall below a detector's sensitivity threshold. There is no way to impute these values reliably, because they are never observed.

Under the regression set-up with predictor X and response Y, missing value problems often arise. To decide how to handle missing value problems, primarily we need to know why these values are missing. We may explain the above three general missing mechanisms in the following manner.

A variable value is missing completely at random (MCAR) if the probability of missingness is the same for all units. Under the regression set-up if the missing values are independent of both response and predictor then these are called missing completely at random. Most missingness is not completely at random. A more general assumption,missing at random (MAR), is that the probability of a variable value is missing depends only on variable information. Under the regression set up, if the missing value depends on predictor but not on response then these are called missing at random.

Missingness is no longer at random if it depends on information that has not been recorded and this information also predicts the missing values. In particular, a difficult situation arises when the probability of missingness depends on the variable itself. Under the regression set-up this type of situation arises when probability of response depends on both response and predictor.For statistical inference with missing information, we usually assume that the missingness pattern is MCAR or MAR. But in many situations these assumptions are not valid.

In clustering algorithms, different packages use different types of imputation techniques like mean imputation, hot deck imputation etc. In order to estimate the missing values properly one should take care of this fact. Use of EM algorithm is usually recommended.

6. Conclusion

From the above discussions it is very clear that although clustering and dimension reduction problems are widely used under different disciplines by scientists from several areas, one should always take care of the nature of data in order to apply the methods successfully. In the introduction we have listed several such problems and only a few are discussed in latter sections. It is quite expected that one may identify many other computation based problems which are not listed here.

References

Adali, T., Jutten, C., Romano, J. M. T. and Barros, A. K. (2009). Independent component analysis and signal separation. In: *Proceedings of 8th international conference, ICA 2009,* Conference held in Paraty, Brazil, March 15–18, 2009. Springer,Berlin.

Bien., J. and Tibshirani R. (2011). Hierarchical Clustering with prototypes via minimax linkage. *Journal of the American Statistical Association,* **106(495)**, 1075-1084.

Bonhomme, S. and Robin, J. M. (2009). Consistent noisy independent component analysis. *Journal of Econometrics,* **149**, 12–25.

Chattopadhyay, A. K. (2017). Incomplete data in Astrostatistics. *Wiley StatsRef: Statistics Reference Online,* 1-12. John Wiley & Sons.

Chattopadhyay, A. K., Mondal, S. and Biswas, A. (2015). Independent component analysis and clustering for pollution data. *Environmental and Ecological Statistics,* **22**, 33-43.

Chattopadhyay, A. K., Mondal, S. and Chattopadhyay, T. (2012). Independent component analysis for the objective Classification of globular clusters of the galaxy NGC 5128. *Computational Statistics and Data Analysis,* **56(1)**, 17-32.

Comon, P. (1994). Independent component analysis, a new concept? *Signal Process,* **36**, 287–314.

Comon, P. and Jutten, C. (2010). *Handbook of Blind Source Separation, Independent Component Analysis and Applications.* Academic Press, Oxford, UK

De, T., Chattopadhyay, T. and Chattopadhyay, A. K. (2013). Comparison among clustering and classification techniques on the basis of galaxy data. *Calcutta Statistical Association Bulletin,* **65**, 257–260.

Fiori, S. (2003). Overview of independent component analysis technique with an application to synthetic aperture radar (sar) imagery processing. *Neural Networks,* **16(3–4)**, 453–467.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics,* **7(2)**, 179-188.

Fraix-Burnet, D., Thuillard, M. and Chattopadhyay, A. K. (2015). Multivariate approaches to classification in extragalactic Astronomy. *Frontiers in Astronomy and Space Science,* **2**, 1-17.

Giavalisco, M. (2002). Lyman-Break Galaxies. *Annual Reviews of Astronomy and Astrophysics,* **40**, 579-641.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics,* **27(4)**, 857-871.

Hartigan, J. A. (1975). *Clustering Algorithms.* John Wiley & Sons.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A *k*-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics),* **28**, 100-108.

Hyvarinen, A., Karhunen, J. and Oja, E. (2001). *Independent Component Analysis.* Wiley, New York.

Hyvarinen, A., Karhunen, J. and Oja, E. (2002). *Telecommunications. In: Independent Component Analysis,* Ch 2, Wiley, New York.

Jung, Y., Park, H., Du, D. and Drake, B. L. (2003). A decision criterion for the optimal number of clusters in hierarchical clustering. *Journal of Global Optimization,* **25(1)**, 91-111.

Kumaran, R. S., Narayanan, K. and Gowdy, J. N. (2005). Myoelectric signals for multimodal speech recognition. *Interspeech*, 1189–1192.

Lee, T. W. (1998). *Independent Component Analysis: Theory and Applications.* Kluwer, Boston, MA.

Little, J. A. R. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data.* John Wiley & Sons, New York.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In: *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability,* 281-297.

McKeown, M. J., Makeig, S., Brown, G. G., *et al.* (1997). Analysis of fmri data by decomposition into independent components. *American Academy of Neurology Abstract,* **48**, A417.

Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika,* **45(3)**, 325-342.

Sugar, A. S. and James, G. M. (2003). Finding the number of clusters in a data set: An information theoretic approach. *Journal of the American Statistical Association,* **98**, 750-763.