Journal of Advances in Mathematics & Computer Science
X(X): XX-XX, 20XX

SCIENCEDOMAIN *international*
www.sciencedomain.org

Heart Disease Diagnosis via Nonparametric Mixture Models

Chipo Mufudza[*1] **and Hamza Erol**[2]

[1] *Applied Mathematics Department, National University of Science and Technology, Corner Cecil Avenue and Gwanda Road, Bulawayo*

[2] *Department Computer Engineering, Mersin University, Çiftlikkoy Campus, TR-33343, Mersin, Turkey*

Research Article

Received: XX December 20XX
Accepted: XX December 20XX
Online Ready: XX December 20XX

Abstract

Aims/objectives: Effective and efficient heart disease prediction via nonparametric mixture regression models.
Data Source: Data used in this paper is from the UCI database of the Cleveland Clinic Foundation for heart disease. The original data source contains 76 raw attributes with 303 observations each. For the purpose of this paper only 14 attributes were used as explained in section 4.
Methodology: Cluster analysis was applied via mixture models in the form of Nonparametric Density-based models. The clusters were identified using a graph theory based technique. Voronoi diagrams were used and and their distributions were estimated nonparametrically through a mixture model with Gaussian kernels. The optimal number of clusters and components of the identified clusters were determined, analysed and diagnosed using a density based silhouette information criteria. All the data analysis and model diagnosis were performed in R using the PdfCluster package.
Results: Different number of components resulted in different number of clusters when nonparametric mixture are used on heart disease. However, the optimal number of clusters under heart disease risks were found to be represented by two clusters with two components using density based silhouette information criteria. These were both well separated and classified as indicated by lack of spurious clusters and high positive density based silhouette values (See Figure 3.1 and 3.3). Their properties are given in table 3.2. The result is irregardless of the flexible conditions which are assumptions free on: shape of the distribution, number of components and number of clusters.
Conclusion: When nonparametric mixture models are used, individuals under risks of heart diseases can be diagnosed either under high or low risk depending on the dominant characteristics on a given individual. Those under high risk progresses have attributes that makes them progresses to heart diseases faster compared to those under low risk. Therefore by classifying individuals into these categories, medical personnel can quickly diagonise heart disease and efficiently identify characteristics associated with each category.

**Corresponding author: E-mail: chipo.mufudza@nust.ac.zw*

Journal of Advances in Mathematics & Computer Science
X(X): XX-XX, 20XX

SCIENCEDOMAIN *international*
www.sciencedomain.org

Keywords: Density based silhouette information, Heart disease, Kernel Density Estimator, Nonparametric Mixture Models

2010 Mathematics Subject Classification: 53C25; 83C05; 57N16

1 Introduction

Data clustering aims to partition data points given in a certain space. There exist no universal rule to define clusters, thus many methods which include both parametric and nonparametric has been proposed. In particular density based clustering methods also referred to as mixture models has been of interest due to their ability to capture diverse heterogeneous properties of the data. Parametric mixture models although very informative and easy to interpret, include a lot of restrictions and assumptions on the shape and distribution of clusters which can be very misleading hence wrong interpretations. Overcoming these restrictions can involve use of algorithms that are assumption free or minimise assumptions, a direction which make use of nonparametric density estimations. It is the aim of this paper to concentrate on these nonparametric mixture models due to their flexibility. Nonparametric methods derive their strength from the sample data given thus reducing the assumption deficit posed by the parametric methods as they limit number of assumptions. The nonparametric densities used in nonparametric mixture models can be represented by histograms or density estimation methods which include a range of varieties although we will concentrate on kernel density estimators (kde). Cluster formed from nonparametric mixtures are derived from own cluster density functions that is unknown but estimated rather than assuming a shape and distribution of a cluster. It is therefore this view for improved flexibility that we are going to focus on nonparametric clustering methods through mixture models. Given the general parametric mixture model shown by equation (1.1):

$$f(\mathbf{x}) = \sum_{i}^{n} \pi_j f_j(\mathbf{x}) \tag{1.1}$$

It is evident that clusters are associated with the components f_j and hence the shape and properties are assumed to follow suit, whilst no assumptions are made nonparametric clusters as they are usually associated with regions of high density. Although the two approaches may sometimes lead to the same results, they are totally very different and it is our ultimate purpose in this paper to dwell on the nonparametric without imposing any assumptions on shape, density and cluster-component factor. This sound a more robust way of identifying and inferring subpopulations as it can use modality inference on the number of modalities as alluded by [1] over the parametric and hierarchical approach which estimates the number of cluster output. Nonparametric density based clustering can also be done via smooth polynomials which model the different cluster density estimations as shown by [2]. They implemented the use Legendre polynomials, Gamma and Beta mixtures to approximate nonparametric mixture models. In this paper we focus on nonparametric mixture models on multivariate mixed count data for heart disease using Gaussian kernel density estimators. Although most nonparametric mixture models have been built under the assumption of identically independent components for identifiability, latent variable model with Gaussian kernels mixtures distributions can be built from observed data of mixed type which can be ordinary, binary, continuous and with component independence etc as explained in [3, 4], respectively.

2 Materials and Methods

2.1 Nonparametric Density Based Clustering

Density based clustering can be done both nonparametrically and parametrically as explained by [5, 6]. In particular it was observed that heart disease can be predicted via two risk groups namely

high and low risk groups via Poisson mixture regression model as explained in [7]. Nonparametric mixtures models however, identify clusters as regions of high density separated by regions of low densities which can be done by identifying local maximum values of the estimated density or modes of the data. Whilst nonparametric methods associate the clusters to the regions around the modes of the probability distribution of data, clusters in the model-based parametric approach correspond to the components of a mixture of distributions. The difference merges clearly since the number of the modes in a mixture of distributions does not necessarily match the number of components as explained by [8].

There has been great interest on nonparametric density based clustering methods due to their flexibility in cluster determination and inference and for any given data type. Different approaches has also been applied making use of both nonparametric density estimations and sometimes distances. Nonparametric density based clustering estimation techniques also allow data to model relationships among variables, thus making them robust to functional form specification and hence the ability to detect structure which sometimes remains undetected by traditional parametric estimation techniques, [9]. The whole idea of parametric mixture model (1.1) is to assume that each component is identified by f_j a parametric density function, thus shape of each cluster is approximated by the same distribution. The clustering problem is now an approximation of the mixing parameters π_j and parameters associated with the density function f_j. This is done under some conditions which makes the model (1.1) identifiable. The most commonly used densities are Gaussian although a lot of variations and shapes have also been considered including using skewed t distributions to capture for a more flexible way for the shapes of clusters. Nonparametric density based clustering can come in as a relief, on the need to free individual clusters from a given density shape hence explore density assumption free clusters. In many cases the parametric mixture modeling comes with serious disparity between a component and a cluster caused by compliance to geometric heuristics as alluded by [10]. If the cluster shapes do not match the shapes of the density f_j, the parametric mixture approach may face difficulties, a motivation for considering completely assumption free cluster shapes densities via nonparametric formulation. The second challenge may also involve variability in the case of the mixing proportions $\pi_j's$ instead of having them as constants.

Nonparametric clustering involves different approaches although the general idea uses kernel density estimators (KDE) which are a representation of mixture functions. Although KDE can use either fixed or adaptive smoothing/bandwidth, the later is usually used to cater for the adaptive and changing mixing proportions of the different clusters. Therefore a nonparametric representation which replaces model (1.1) can be achieved via clusters with kernel density functions. Researchers which include [10] used these kernel density functions as a mixture densities for clusters to identify a point of local maximum of density called a hilltop using the Modal EM (MEM) to identify clusters. Different clusters were linked via a ridgelines linking two hilltops developed, creating a more accurate way of nonparametric clustering through geometric heuristics. Practical procedures on nonparametric cluster quality diagnosis have been developed through the use of mean integrated errors as well as nonparametric density based silhouettes information by [8]. In [11], they used nonparametric estimates for finite mixtures from data on repeated measurements with the aim of advancing statistical inference on such models.

The density based nonparametric clustering methods normally use clusters driven by density estimates which determines associated connected regions. Unlike the parametric methods nonparametric density based clustering methods do not chose particular density function but normally use kernel density estimate with kernels one one's choice. The Gaussian kernel has so far proved to be the commonly used, [19, 20, 21, 12]. General researchers use the multidimensional kernel density estimator with adaptive bandwidth suggested by [13] and represented as equation (2.1):

$$\hat{f}(x) = \sum_{i=1}^{n} \frac{1}{nh_{i,j}.....h_{i,d}} \prod K\left(\frac{x^j - x_i^j}{h_{i,j}}\right) \tag{2.1}$$

where h is the bandwidth and K id the kernel function. Thus the general use of nonparametric density

based clustering represent a mixture model.

Several authors have represented nonparametric mixture models in different ways with the general assumption that each cluster is generated by its own unknown density function, [12]. Nonparametric mixture models can therefore also means that no assumptions are made about the form of the density fj of model (1.1), even though the weights π_j maybe scalars parameters as alluded by [14]. It should however, be noted that the weights are not only restricted to scalar weights since they can be variables. [15], defines nonparametric mixture modeling in a different sense such that the family F from which the component densities come is fully specified up to a parameter Θ, but the mixing distribution from which they are drawn is assumed to be completely unspecified and unknown, rather than having finite support of known cardinality. Thus nonparametric mixtures models can therefore be used to describe the case in which no assumptions are made about the distribution form the mixture model, even though the parameter mixing parameter can be Euclidean. Semi parametric models normally refers to cases where the distribution form is partly specified by a finite-valued parameter, such as the case in which $f_j(x) = f(x - \mu_j)$ for a symmetric but otherwise completely unspecified density $f()$, as proposed by [16] an idea used in [14, 17] for the multivariate cases. The assumption that component distributions come from a family of densities that may be indexed by a finite-dimensional parameter vector is normally ignored when dealing with nonparametric mixtures. It is still however, necessary to restrict the family of multivariate density functions from which the component densities are drawn in order to avoid the problem of model non-identifiability, [17].

A number of nonparametric mixture models have considered that the observed variables are jointly conditionally independent given the latent class and use kernel methods to identify the finite mixtures of nonparametric distributions, [18]. Suppose θ is a vector of parameters, including the mixing proportions $\lambda_1, ..., \lambda_k$ and the univariate densities f_{jk} where, j indexes the component and k indexes the coordinate, so $1 < k < r$ and $1 < j < m$. Thus, under the assumption of conditional independence, the nonparametric mixture density evaluated at $x_j = (x_{i1}, ..., x_{ir})^t$ is given as:

$$f_\theta(\mathbf{x_i}) = \sum_{j=1}^{m} \lambda_j \left\{ \prod_{k=1}^{r} f_{jk}(x_{ik}) \right\} \tag{2.2}$$

If we let z is the latent random variable with probability mass function (PMF) : $\chi \rightarrow [0; 1]$, where $\chi = \{z_1, z_2, ..., z_k\}$ for an integer K. $\mathbf{x} = (x_1; x_2; .....x_r)$ be a vector of repeated measurements on an observable outcome variable x whose marginal probability density function (PDF) takes the form equation (2.2) and f_{jk} denotes the PDF of $\mathbf{x_i}$ conditional on $z = z_k$. It should be noted that $\mathbf{x_i}$ sometimes are not only conditionally independent but identically distributed as well which can be represented in terms of blocks as eluded in the work of [17], where each block represents coordinates that follows the same univariate distribution.

Graph theory techniques such as voronoi diagrams can also be used to identify clusters by observing closely concentrated connections then estimation of density function along an interval of connections has been explored in [19]. This method suggests that any existence of a valley between point intervals is a disconnection using Voronoi diagram partition and Delaunay graph as edge connection between points. The same idea was developed by [20, 21] although the later further made an improvement on the measure on which the valleys exhibited by use of minimum mass probability necessary to fill the valley. In this paper we will focus heart disease diagnosis using graph theory and KDE to estimate the nonparametric mixture models.

2.2 Nonparametric Mixtures Using Graph Theory

When graph theory methods are used in mixture models, clusters are connected components estimated using graph based algorithms. The number of clusters can then be chosen based on nearest neighbourhood graph $X_i : \hat{K_h} \geq k$ where K_h is the kernel density estimator and k is the number of clusters. The number of clusters is normally done on the basis of the number of connected

components of a level set $f > c$ in the given graph a scenario explained by [22]. [23] proposed that a combined kde with single linkage graph can be used to estimate the number of clusters. A lot more authors who included [24] have also used density clustering via connected regions using data driven bandwidth selection measures and stability of the identified clusters. In this sense mixtures are formed by kde's where an edge between two points indicate that they belong to the same ball ρ, in the same cluster. The number of connected components will then corresponds to the number of clusters found in the sample. The same idea was used by [25] when they investigated the stability of the density based clustering by altering different tuning parameters to the mixture models including the kernels.

Graph theory can also be used to identify cluster tress and pruning the trees to do away with spurious clusters in cases where they exist. [26] explored a plug-in approach to cluster tree estimation: estimate the cluster tree of the feature density by the cluster tree of a density estimate. In [19, 20] they described nonparametric mixture models differently in an interesting way via graph theory by use of the KDE. They identified clusters using connected points by observing the density function along a given interval in a multidimensional space. Clusters are then identified by an existence of some disconnections between Voronoi diagram indicated through a valley using a mode function which is associated with both components and probability. Thus points of high connections indicates a mode of the kernel density and hence a single cluster. Connected regions are then identified using Delaunay diagrams or pair in a high dimensional space. They demonstrated this via the PdfCluster package in R, [20, 21], an algorithm which makes use of the density based silhouette information to differentiate clusters and inferences on them as described by [8]. Thus nonparametric mixture models has that flexibility to identify clusters effectively through so many methods. This is the reason why we are going to concentrate on some of these methods on real data set to help alleviate and easy cardiological problems and hence better heart disease diagnosis.

2.2.1 Nonparametric Mixtures: Graph Theory Procedure

We will explore some of the procedures used by [19] that if the estimated function is unimodal then the we have a connected cluster otherwise it is disconnected. Thus, if we have $\chi = (x_1, ..., x_n),\ x_i \in \Re^d$ to be clustered in a d dimensional space with unknown bounded and differentiable density function, f. Then for each constant $c,\ R(c) = \{x : x \in \Re^d, f(x) \geq c\}$. In this method a mode function is described such that the function f is replaced by its nonparametric estimate $\hat{f}$. Therefore in multidimensional space where there is no obvious way to determine the connected region the interest is to focus on the sample set of $\aleph$ such that we have

$$S(c) = \{x : x \in \Re^d, \hat{f}(x) \geq c\} \tag{2.3}$$

Thus we have a multidimensional problem of identifying connected regions. Graph theory is then used to identify connected components as detection of connected components of a given graph **G** whose elements are vertices of $S(c)$ where the edge is key. The task is then done via Voronoi diagram using either Delaunay triangulation for $d \leq 3$ and pairing for $d > 3$ and according to some measure of distance. The Voronoi diagram is a partition of $\Re^d$ given χ into k regions $\Upsilon(x_1), ..., \Upsilon(x_n)$ where $\Upsilon(x_i)$ is the set of all points of $\Re^d$ closer to x_i than to any other point in the set according to some measure of distance. The connected components are identified as union of connected pairs that share at least one vertex and in this way cluster cores are determined. The clusters cores are formed by data lying in the regions around the detected modes. Clusters can also be detected by a minimum spanning tree as explained by [27] which is a subgraph of the Delaunay triangulation, [21]. Figure (2.2.1) shows an example of how the connected points share one facet of Delaunay triangulation as the belong to adjacent Voronoi regions. When dimension is high then pairwise connections are implemented as in our case where $d = 14$. The basic idea is to examine the behaviour of $\hat{f}(x)$, the kde when we move along a segment $[x_1,\ x_2]$, since it depends on whether the sample values x_1 and x_2 belong to the same connected set of $S(c)$ or not. Thus we can view set $S(c)$ as a union of the

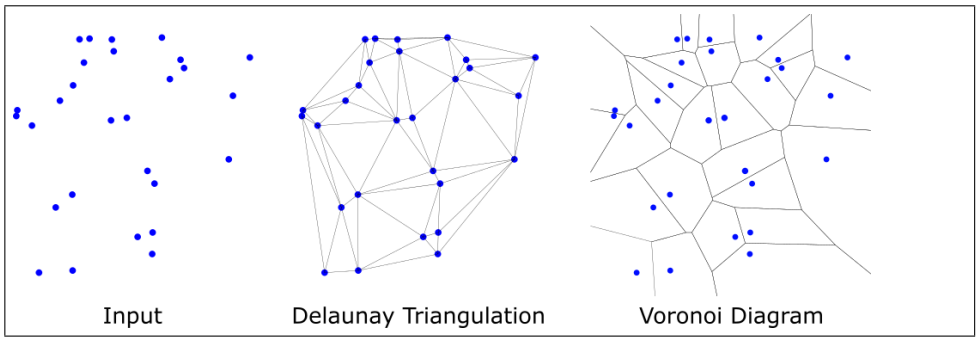

Figure 2.1: Voronoi tessellation and superimposed Delaunay triangulation

two intervals of the group. If x_1 and x_2 belong to the same interval, then the corresponding portion of density along the segment $(x_1,\ x_2)$ has no local minimum. On the contrary, if x_1 and x_2 belong to different subsets of $S(c)$, then at some point along $[x_1,\ x_2]$ the density exhibits a local minimum, which we shall refer to as presence of a valley, [20]. The density estimates which determines the connected regions are not linked to particular density function but are purely estimated. The commonly used density estimate is the multidimensional kernel density estimate given before in equation (2.1). The choice of the kernel normally does not have an effect on the estimate unlike the smoothing parameter. The bandwidth can be fixed and adaptive as suggested according to [26] . In this paper although, both the fixed and adaptive bandwidth were developed and analysed for heart disease, only adaptive bandwidth will be represented. Nonparametric mixture models via graph theory can be described by use of the KDE where clusters are identified by an existence of some disconnections between voronoi diagram indicated through a valley. Thus points of high connections indicates a mode of the kernel density and hence a single cluster. They demonstrated this via the ***PdfCluster***, [20, 21] algorithm which makes use of the density based silhouette information to differentiate clusters and inferences on them as described in [8].

2.2.2 Density Based Silhouette Information (DBS) Criteria

In general silhouette analysis is a form of internal cluster validation which can be used to study the separation distance between and within the resulting clusters. It validates the clustering performance based on the pairwise difference of between and within cluster distances. This index can also be used to determine the optimal cluster number through maximizing the value of this index, [28]. The silhouette plot displays a measure of range $[-1,\ 1]$ which show how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. Silhouette index (as these values are referred to as) near $+1$ indicate that the sample is far away from the neighboring clusters. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters and negative values indicate that those samples might have been assigned to the wrong cluster. The thickness of the silhouette plot can also give an idea of the cluster size. Silhoutte information can also be used to describe cluster compactness. This is a description of how objects are closely related in a single cluster although this can be done via variance analysis. therefore, silhouette information can be used for both within and between cluster diagnosis.

The between and within distances are not always easy to calculate and apply with nonparametric methods. An idea of silhouette information usage in nonparametric density-based clustering procedure was developed by [8]. This was made possible via incorporating the idea of both within cluster distances as well as cluster posterior probabilities same idea employed by [20, 21]. The dbs values

are calculated from the fact that if $x_i \in \chi$ is drawn from a probability density function f, one can evaluate the posterior probability that it belongs to group $\upsilon_m,\ m = 1 : M$ as:

$$\tau_m(x_i) = \frac{\pi_m f_m(x_i)}{\sum \pi_m f_m(x_i)} \tag{2.4}$$

where π_m is the prior probability of υ_m and f_m is the density of group υ_m at x_i. Then the dbs is defined as follows

$$dbs_i = \frac{log\left(\frac{\tau_{m_0}(x_i)}{\tau_{m_1}(x_i)}\right)}{max_{j=1:n}\left|\ log\left(\frac{\tau_{m_0}(x_j)}{\tau_{m_1}(x_j)}\right)\ \right|} \tag{2.5}$$

where m_0 is such that x_i has been classified in υ_{m_0} and m_1 is the group into which τ_m is maximum $m \neq m_0$. Thus a dbs is vector reporting the density-based silhouette information of the clustered data clusters, [8, 20, 21]. It can be used as a diagnostic tool for cluster quality evaluation as proposed where a high positive dbs value indicates well classified observations and clusters that are well separated whilst a negative dbs indicates observations might have been assigned to wrong clusters. The partitioning of clusters may indicates existence of hidden clusters (spurious), [8].

2.2.3 Bandwidth Selection

The selection of an appropriate value of h the bandwidth parameter which affects the density estimations dramatically as explained in nonparametric methods is vital for best cluster results and estimations. The choice of kernel since it is not as influential we are going to use Gaussian kernels for all the estimations. Bandwidth selection methods have been studied intensively especially in no mixture structures by authors who include [13, 29]. Bandwidth selection in nonparametric mixtures can be a very challenging procedure although some standard ideas can still be incorporated, there seems to be challenges on under or over smoothing as explained by [30] in the case of conditional independence. However, when Gaussian kernels are used under the assumption of multivariate normality, [19] suggested that a bandwidth of the following format equation (2.6) can be used to the adaptive Gaussian kernel in equation (2.1) namely

$$h_j = \sigma_j \left(\frac{4}{(d+2)^n}\right)^{\frac{1}{(d+4)}},\ \ j = 1, ..., d, \tag{2.6}$$

where the standard deviation σ_j of the jth variable is replaced by an estimate. The bandwidth, h_j are then multiplied by a shrinkage factor of $3/4$ in order to relieve the over smoothing determined by computing the bandwidths under the assumption of multivariate normality. A square root law can be used to allow variation for each dimension and on different scales. This brings flexibility in such a way that different components are allowed to have different properties. Moreover, since the bandwidth is iterative, the bandwidth estimations can be done prior to knowledge of the mixture structure.

3 Results and Discussion

3.1 Data

The data used in this paper is found in the UCI database for the Cleveland Clinic Foundation for heart disease as given by [32]. Original data is mixed data with 76 raw attributes and 303 observations for each attribute of which only 14 attributes are used in this paper. These include number of diagnosis (num), Age, sex, chest pain, cholesterol, fasting blood sugar level (fbs), rest blood pressure (trestbp), maximum heart rate (thalcd), resting electrocariographic (restecg), exercise induced angina (exang), depression by exercise (oldpeak), slope of peak exercise (slope), vessels (ca) and defects (thal).

The data had some missing values which we replaced using the mean response method. In the analysis continuous variables were taken to be log normal variables and sometimes changed to log-normal whilst discrete variables with more than two levels were considered ordinal. Nonparametric cluster analysis is done without prior information assumed or known. The cluster density estimations, clusters and inferences on the clusters are solely based on the data and hence the use of nonparametric methods. Kernel estimation methods with a Gaussian kernel were considered in this paper to estimate the density distributions. Thus the ultimate distribution of the data is a nonparametric mixture model which compromises of Gaussian kernel density estimates.

3.2 Packages Used

Clustering of the data was nonparametrically done using graph theory methods as explained in section 2.2.1. The dissimilarities distances among the points were calculated using the Gower coefficient by [27] using the **cluster** package in R. The **cluster** package is also used to classify multidimensional scaling of the data under the **cmdscale** function as described by [33] since the data used was mixed data. Nonparametric density approximations are then implemented to the mixed heart disease data using Gaussian kernel density estimators. This was implemented through the **PdfCluster** package by Azzalini and Menardi, [20]. The PdfCluster automatically selects the procedure to be used for detecting connected components of the density level sets, depending on the data dimensionality. This is enabled by making an internal call to function *kepdf* both to estimate the density underlying the data and to build the connection network when the pairwise connection criterion is selected. In this paper a kernel density estimation with Gaussian kernel is chosen and built, with the vector of smoothing parameters set to the one asymptotically optimal under the assumption of multivariate normality. We used both a fixed shrinked smoothing parameter and adaptive bandwidth since the data is highly multidimensional. Cluster diagnosis in this method uses the dbs criteria described before in section (2.2.2). The number of groups here is identified as the number of modes of the estimated density used. Analysis was done under different number of components and diagnosis of the clusters was done using dbs criteria from the **PdfCluster** package. Therefore, use of the dbs method is applied to evaluate the cluster quality of Cleveland heart disease data under nonparametric density based clustering using different principal components for the same data.

3.3 Analysis Covered

Analysis on the Cleveland heart disease data is done using **PdfCluster** to determine the number of groups possible for heart disease prediction in a nonparametric way. Each of the cluster is represented by a Gaussian kernel density funtion and hence the whole sample data as a mixture. Both the fixed smoothing parameter (here not shown) and adaptive smoothing parameter via the adaptive bandwidth as explained in section (2.2.3). Different number of components were used to determine the number of clusters nonparametrically via connected regions as explained in section (2.2.1). Results of the analysis is shown in the next sections (3.4 - 3.6).

3.4 Cluster Diagnosis

In order to determine cluster optimal number of clusters and their quality, dbs criteria is used for cluster diagnosis as explained in section 2.2.2. Different models with different number of components result in a variety of clusters produced and hence the need to determine the quality and optimality of the clusters. Figure 3.1 shows dbs plot and values for models with different principal components resulting in varying median dbs values for each given cluster. The cluster information for a model with principal components i.e. $p = 2, 3, 4,$ and 5 are given by Fig. 3.1(a), (b), (c) and (d), respectively. The 2 component model has 2 clusters both with positive dbs median values, and no partitions within

each cluster. Both clusters have high positive median dbs values which indicates that observations have been well classified. They are also well separated due absence of within dbs plot partition an evidence of no hidden clusters (spurious) within each of the clusters. The quality of clusters seems to decrease as the number of principal component increases to 3 as shown by Fig. 3.1(b) which shows the existence of 3 spurious clusters out of the 5 clusters. Furthermore, the 2 other clusters have negative dbs median values a pointer that most of the observations were wrongly classified. Whilst

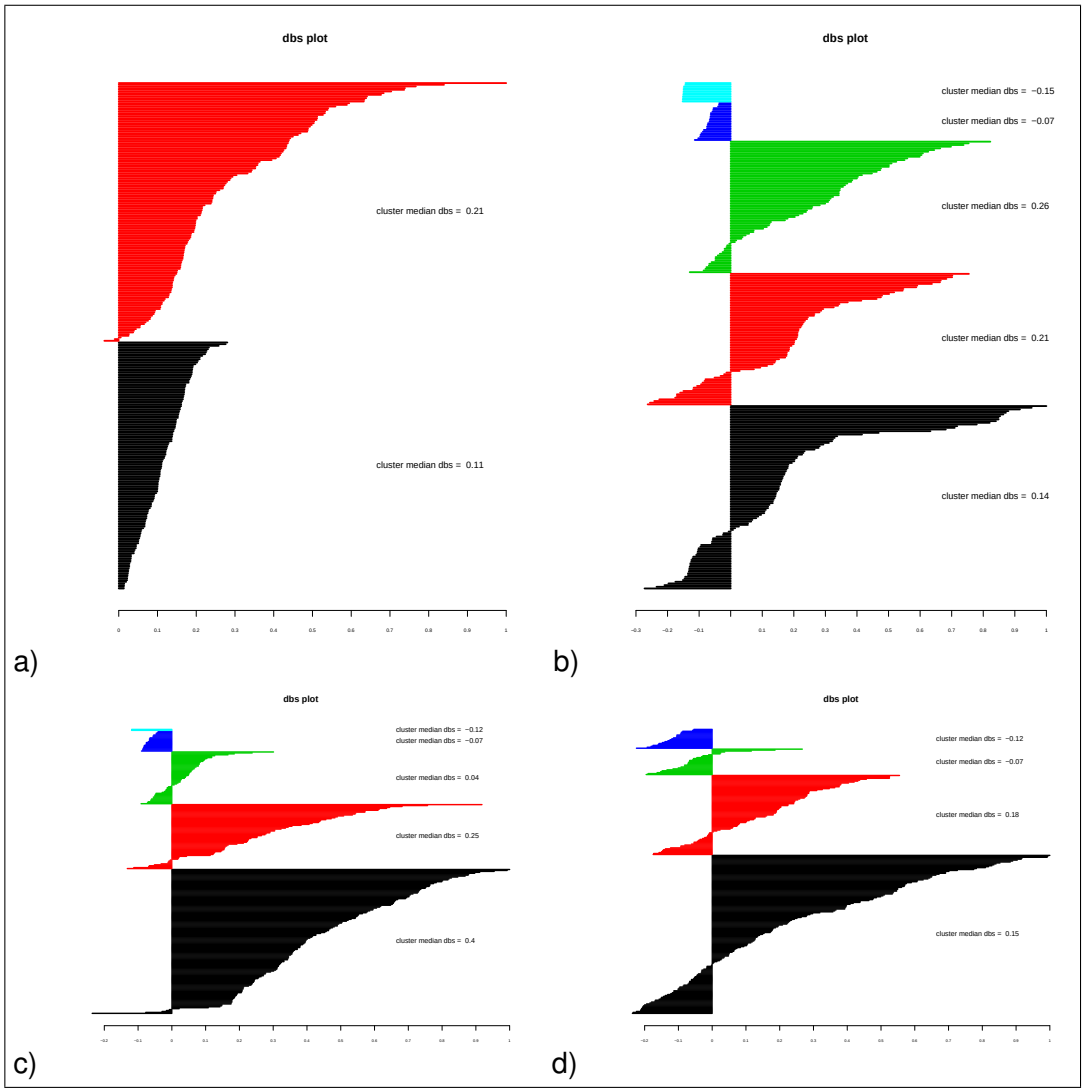

Figure 3.1: Dbs diagrams for different p principal components: a) $p = 2$, b) $p = 3$, c) $p = 4$ and d) $p = 5$

the 4 component model in Fig. 3.1(c) has 5 clusters, the 5 component in Fig. 3.1(d) has 4 clusters most of them which has hidden clusters and with wrongly classified observation as indicated by dbs plot partitions and with negative dbs median values.) has 3 spurious clusters with 2 cluster with

negative dbs a sign of poor quality clusters.

The dbs plot and dbs median values for models with 8 and 10 components are are given by Fig. 3.2(a) and (b), respectively. The 8 component model has 5 clusters whilst there are only 2 clusters with a ten component model. Although the 10 component model gives the same number of clusters as in 2 principal component one, the dbs median values for the clusters are both negative a high indication that observations in these clusters were wrongly classified and ill partitioned. Therefore

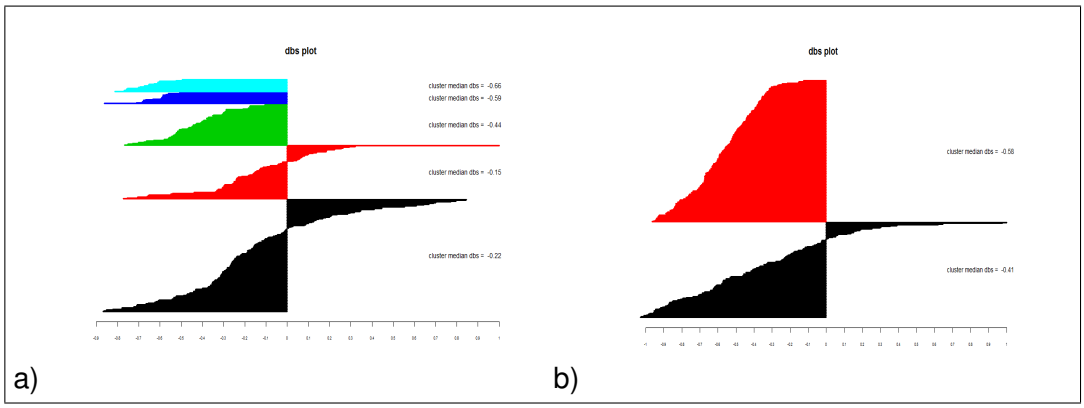

Figure 3.2: Dbs diagrams for different p principal components: a) $p = 8$, b) $p = 10$.

given the mean dbs values for each cluster we conclude that data fits well in model with 2 components and 2 clusters. The width of the dbs graphs for the 2 component 2 cluster model are also reasonably bigger than any other an indication that most observations are clustered in these clusters. This is because the clusters are well separated, have no hidden clusters and with well classified observations. Thus, number of principal component corresponds to the number of clusters although it should be noted that different results can be found in other situations. It can be deduced that individuals exposed to heart disease risks can be classified into 2 categories depending on which risks are they exposed to at a given time. This can be high risk and low risk individuals as observed by [7] using Poisson mixture models.

3.4.1 Cluster Scatter Plots

The cluster scatter plots represent how observations are classified, separated and partitioned within and between clusters. It can also help to view cluster partitioning besides within a component and cluster besides using the dbs diagnosis plot. In order to view these cluster partitions and classifications, we are going to use Fig. 3.3, which represents the cluster scatter plots for models with different components. The cluster scatter plot for a 2 and 3 component models are shown by Fig. 3.3(a) and (b), respectively. They show that a 2 component model has 2 clusters which are well separated in the scatter graph as groupings can easily be identified. A same 2 component model here not shown using fixed bandwidth had 3 clusters which were poorly separated. However, a 3 component model on the other hand has 5 clusters which are very difficult to identify on the scatter graph. The same condition is repeated as we increase the number of components to 4 as shown by Fig.3.3(c). It shows that it is very difficult to discretely separate observations among the different clusters. This highly indicates clusters that are ill separated and observations are most likely to be wrongly classified. A 10 component model although it has the same number of clusters as the 2 component model, Fig. 3.3(d) shows that their elements can rarely be distinguished from each other. Thus again using cluster scatter plots we deduce that 2 component 2 cluster model proves to be the best model for heart disease diagnosis.

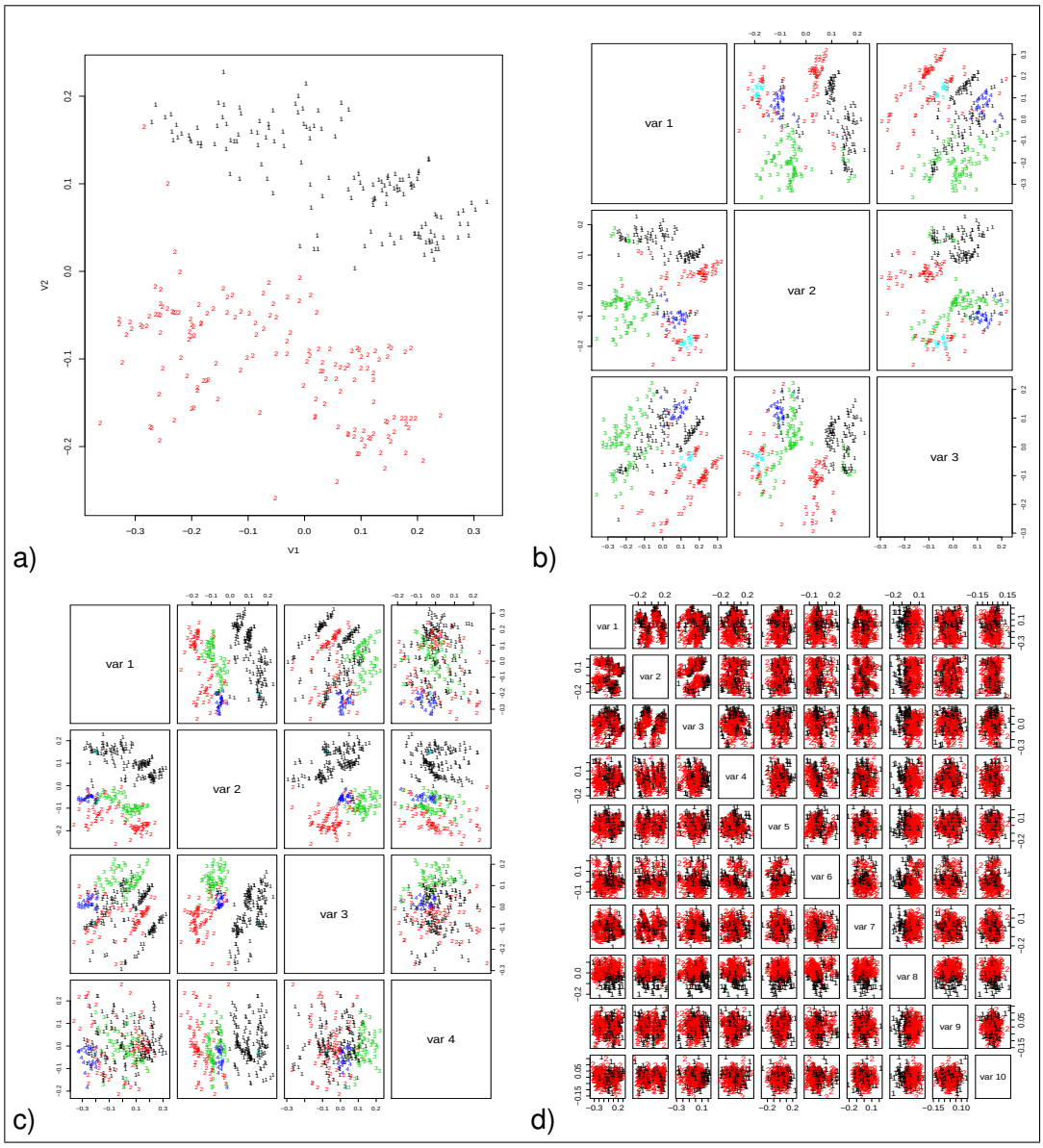

Figure 3.3: Cluster scatter plots for model with: a) 2 components, b) 3 components, c) 4 components and d)10 components

3.5 Cluster Summary for Different Models

In this section we represent a summary of all the models considered in the analysis with different number of components and clusters as given in table 3.1. It shows the different cluster properties of the nonparametric mixture models analysed using **PdfCluster** package in R. The clusters alternates from 2-5 with increase in the number of components. However, this seem to result in an increase in the

number of spurious clusters as well as wrong observation classifications as indicated by increase in clusters with negative dbs median values thus reducing the cluster quality. Deducing from the cluster properties in the table, best clusters are produced by model with 2 components as it has no spurious clusters as well as all positive dbs median values as we have already seen. Although the 7 and 10 componential models produced the same number of clusters as 2, both clusters from a 7 component model are spurious and with very low positive dbs median values, compared to a 2 component model. This indicates poor cluster separation and relatively high observation misclassification. In the same manner the 10 component model have all negative dbs values for the 2 clusters a high indication that clusters were wrongly classified. We therefore conclude that individual under risks of heart disease be classified by 2 component nonparametric density model with 2 clusters.

Table 3.1: Cluster Summary Under Adaptive Bandwidth

Components	**Total Clusters**	**Spurious Clusters**	**Negative dbs Clusters**
2	**2**	**0**	**0**
3	5	3	2
4	5	3	2
5	4	3	2
6	5	3	4
7	2	2	0
8	5	2	3
9	5	2	5
10	2	2	2

3.6 Best Model Properties (2 Component 2 Cluster Model)

The best model for the heart disease is a 2 component 2 cluster model as it proved to be best model for the heart disease data. In this section we will explore some of the properties of the choosen model to have a better under understanding of the clusters. This may involve getting the properties of each of the clusters produced by the **PdfCluster** algorithm. Table 3.2 shows a summary of some of the statistical measures for the model : The table summarises the dbs information for each cluster as well as for the whole data. It shows that cluster 2 has high median dbs and mean values of 0.21 and 0.29, respectively compared to cluster 1 which has a value even lower than the whole sample data. Cluster 2 also has a high max positive dbs of 1 indicating well compacted cluster quality clustered data.

Table 3.2: 2-2 Model DBS Cluster Summary

Cluster	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	0.01311	0.06579	0.11020	0.11450	0.16130	0.28030
2	-0.03721	0.14310	**0.21280**	**0.28590**	0.43170	1.00000
Data	-0.03721	0.09534	0.15800	0.20220	0.23520	1.00000

Discussion

In parametric mixture models, components and clusters are assumed to be same with the same parametric distribution. Although the best model nonparametrically for heart disease has 2 clusters and 2 components, same results as produced by [7] parametrically, the nonparametric model analysis in section 3.4 shows that the number of components is not always the same with the number of clusters. Thus, the number of components does not always correspond to the number of clusters as normally assumed in parametric mixture models but different number of components can correspond to different number of clusters being produced. As expected with change in the smoothing parameter results in a change in the model outcomes, there were drastic changes in the model output with change from fixed bandwidth to adaptive iterative bandwidth. The adaptive bandwidth usage resulted in rigorous classifications where clusters are well differentiated. Whilst fixed band width seems to have little effects on the optimal clusters from different components, when adaptive bandwidth is used, increase in the number of components result in increase in number of clusters. An interesting point to note is how nonparametric mixture models differentiate the heterogeneity among observations by grouping them into clusters and their ability to infer more on the clusters compactness, within and between correlations unlike parametric mixture which make use of property assumptions. In this paper the density based silhouette information was used to determine cluster properties by using the median dbs values, dbs graph orientation as well as cluster scatter plot for the different components. Whilst high positive mean dbs values of up to one shows compactness the negative values indicate wrong classification. The different clusters have different level of risk depending on which characteristics are more evident than the other. It was noted that individuals in cluster 2 were at more risk compared to those in cluster 1 for the chosen model. Thus nonparametric mixture models explains the heterogeneous properties on the observations and groups. Further research on this work can also involve incorporating analysis of the stability of the clusters when different tuning parameters in the estimations are used, i.e How stable are the clusters under different smoothing parameter and kernels. An example is incorporating different kernel estimators to the nonparametric clustering.

Conclusion

In this paper we presented how nonparametric mixture models can be used to diagnose heart disease efficiently via the use of graph theory techniques and estimation of the cluster densities using Gaussian kernel density estimators. This was achieved by use of voronoi diagrams and pairwise connection to identify high connections point (modes) and low connection points (valley). This enabled us to capture the heterogeneity in the heart disease data which is a mixed count data as connection points(clusters) can then be represented by a mixture density via kernel density functions. Analysis of these combined processes were done via cluster and PdfCluster packages in R. In this case the cluster and component behaviour are solely determined and estimated from the data, thus bringing the flexibility and robustness to cater for heterogeneous qualities of any given population. The use of density based silhouette information enabled us to determine the compactness, within and between cluster correlations and hence quality of the required number of clusters. The analysis and results therefore enabled us to clearly classify individuals under heart disease risks into 2 clusters using 2 components. These clusters has different risk levels, either a cluster for those who are most likely to have the disease (high risk) and those unlikely to get the disease (low risk). A closer look on within and between cluster correlations through dbs criteria and paired connections enabled us to have a clear picture of relations on observations within a cluster. We noted that cluster 2 had a more compact structure than cluster 1 due to its highly positive mean dbs value. Individuals in cluster 2 where those considered to be at high risk, thus we can deduce that those individual who are at high risk are more likely to show tightly related symptoms than those who are unlike to develop it as shown by the cluster compactness. By identifying individuals under heart disease risk into either a

low or high risk group, this can help effectively and efficient diagnose heart disease and hence reduce cardiological challenges that can be brought by wrong diagnosis. Treatment can then be administered accordingly. Thus we can conclude that when nonparametric mixture models are used to diagnose heart disease, a 2 cluster model with 2 components can be used as it can diagnose individuals with different heart disease risk factors properly.

Recommendations

The model can be very useful to physicians and medical personnel because it helps to diagonise heart diseases not only in a quick and effective fast way but also efficiently. Since the model can be used to classify a patient under high or low risk category, by administering a machine learning programme that captures the 14 attributes used in this work from any given patient, physicians can easily limit the time needed for diagnosis. This is possible because the clustering is performed at individual level capturing individual characteristics hence individuals can be classified. It is of great interest to note that since each cluster has common individual characteristics, it will be therefore be faster for a physician to identify the attributes associated with those in high risk or low risk categories directly therefore minimising the time to diagonise heart diseases. Consequently, effective monitoring strategies can be invented for each cluster depending on the common characteristics identified.
In this paper, clusters were identified nonparametrically using graph theory producing same results with [7] were Poisson mixture regression models were used. We recommend to data scientists therefore that a different model either with more attributes or that uses different clustering methods like splines be implemented to check the possible number of clusters that can be produced.

Acknowledgment

Gratitude goes to the Turkish Government Scholarship for the financial provision to the corresponding author.

References

[1] Cheng, Y.S., Ray, S. (2014). Multivariate Modality Inference Using Gaussian Kernel, Open Journal of Statistics., 4, 419–434. http://dx.doi.org/10.4236/ojs.2014.45041.

[2] Rebafka, T., Roueff, F. (2015). Nonparametric Estimation of the Mixing Density Using Polynomials, Mathematical Methods of Statistics, Allerton Press, Springer (link)., 3, hal-00458648v2.

[3] McParland, D., Gormley, I.C. (2016). Model Based Clustering for Mixed Data: clustMD Adv Data Anal Classif., 10, 155. doi:10.1007/s11634-016-0238-x.

[4] Zhu, X., Hunter D.R. (2015). Clustering Via Finite Nonparametric ICA Mixture Models, Annals of Applied Statistics., 1–27.

[5] McLachlan, G., Peel, J. (2000). Finite Mixture Models. Wiley Series.

[6] Fraley, C., A.E Raftery, A.E. (2002). Model-based clustering, discriminant analysis and density estimation. Journal of the American Statistical Association. ABI/INFORM Global., 97, 611.

[7] Mufudza, C., Erol, H. (2016). Poisson Mixture Regression Models for Heart Disease Prediction. Computational and Mathematical Methods in Medicine. Article ID 4083089, 10 pages. http://dx.doi.org/10.1155/2016/4083089.

[8] Menardi, G. (2011). Density-based Silhouette diagnostics for clustering methods. Stat Comput. 21, 295–308. DOI 10.1007/s11222-010-9169-0.

[9] Raccine, J., Li, Q. (2004). Nonparametric Estimation of Regression Functions with Both Categorical and Continuous Data. Journal of Econometrics., 119, 99–130.

[10] Li, J., Ray, S., Lindsay, B.G. A Nonparametric Statistical Approach to Clustering via Mode Identification. Journal of Machine Learning Research., 8, 1687–1723.

[11] Bonhomme, S., Jochmans, K., Robin, J.M. (2016). Nonparametric estimation of finite mixtures from repeated measurements. Journal of the Royal Statistical Society: Series B (Statistical Methodology)., 78, 211–229. DOI: 10.1111/rssb.12110.

[12] Mallapragada, P.K., Jin, R., Jain, A. (2010). Nonparametric Mixture Models for Clustering, Chapter: Structural, Syntactic and Statistical Pattern Recognition: Lecture Notes in Computer Science., 6218, 334–343.

[13] Scott, D.W., Sain, S. (2005). Multi-dimensional Density Estimation, Hand Book of Statistics., 24, 229–261.

[14] Chauveau, D., Hoang, V.T.L. (2015). Nonparametric Mixture Models with Conditionally Independent Multivariate Component Densities., hal-01094837v2.

[15] Lindsay, B.G., Lesperance, M.L. (1995). A review of semiparametric mixture models. Journal of Statistical Planning and Inference., 47, 29–39.

[16] Hall, P., Zhou, X. (2003). Nonparametric Estimation of Component Distributions in a Multivariate Mixture. The Annals of Statistics., 31, 201–224.

[17] Benaglia, T., Chauveau, D., Hunter, R. (2009a) An EM-like Algorithm for Semi- and Nonparametric Estimation in Multivariate Mixtures. Journal of Computational and Graphical Statistics., 18, 505–526.

[18] Sgouritsa, E., et al. (2013). Identifying Finite Mixtures of Nonparametric Product Distributions and Causal Inference of Confounders. Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence (UAI)., 556–565.

[19] Azzalini, A., Torelli, N. (2007). Clustering via nonparametric density estimation, Stat Comput., 17, 71–80. DOI 10.1007/s11222-006-9010-y.

[20] Azzalini, A., Menardi, G. (2014). Clustering via Nonparametric Density Estimation: The R Package PdfCluster, Journal of Statistical Software., 57, 1–26.

[21] Menardi, G., Azzalini, A. (2014). An Advancement in Clustering via Nonparametric Density Estimation, Stat Comput., 24, 753–767. DOI 10.1007/s11222-013-9400-x.

[22] Cuevas, A., et al. (2000). Estimating the Number of Clusters, The Canadian Journal of Statistics., 28, 1–17.

[23] Biau, G., Cadre, B., Pelletier, B. (2007). A Graph Based Estimator of the Number of Clusters. ESAIM: Probability and Statistics., 11, 272–280. DOI: 10.1051/ps:2007019.

[24] Rinaldo, A., Wasserman, L. (2010). Generalised Density Clustering, The Annals of Statistics., 38, 678–2722. DOI: 10.1214/10-AOS797.

[25] Rinaldo, A. Singh, A., Nugent, L. (2012). Stability of Density Based Clustering, Journal of Machine Learning Research., 13, 905–948.

[26] Stuetzle, I., Nugent, R. (2010). A Generalised Single Linkage Method for Estimating the Cluster Tree of a Density. J. Comput. Graph Statistics., 19, 397–418.

[27] Gower, J.C., Ross, J.S.G. (1969). Minimum Spanning Trees and Single linkage Cluster Analysis, J.R.Stat. Soc., Series C (Applied Statistics)., 18, 54–64.

[28] Liu, Y., et al. (2010). Understanding of Internal Clustering Validation Measures, IEEE International Conference on Data Mining., 911–916.

[29] Härdle, W., et al. (2004). Nonparametric and Semiparametric Models. Springer.

[30] Benaglia, T., Chauveau, D., Hunter, R. (2009b). An EM-like algorithm for semi- and nonparametric estimation in multivariate mixtures, Journal of Computational and Graphical Statistics., 18, 505–526.

[31] Benaglia, T., Chauveau, D., Hunter, R. (2011). Bandwidth Selection in an EM-like algorithm for nonparametric multivariate mixtures, World Scientific Publishing Co. Nonparametric Statistics and Mixture Models: A Festschrift in Honor of Thomas P. Hettmansperger, World Scientific Publishing Co., 15–27. hal-00353297.

[32] Detrano R., (M.D., Ph.D.), Long Beach and Cleveland Clinic Foundation, V.A. Medical Center. http://archive.ics.uci.edu/ml/datasets/Heart+Disease.

[33] Maechler, M., et al. (2016). cluster: Cluster Analysis Basics and Extensions. R package version 2.0.5.

[34] Chauveau, D., Hoang, V.T.L. (2016). Nonparametric mixture models with conditionally independent multivariate component densities. Computational Statistics & Data Analysis., 103, 1–16. http://dx.doi.org/10.1016/j.csda.2016.04.013.

©2018 Chipo Mufudza & Hamza Erol; This is an Open Access article distributed under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/2.0, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Disclaimer: - This manuscript was presented in the conference “ICANAS, 2nd INTERNATIONAL CONFERENCE ON ADVANCES IN NATURAL AND APPLIED SCIENCES, ANTALYA, TURKEY, At Antalya, Turkey” available link is “https://www.researchgate.net/publication/316550105_Heart_Disease_Diagnosis_via_Nonparametric_Mixture_Modelsdate” April 2017.