Department of Management Science & Technology

MSc in Business Analytics

"Clustering Mixed Mode Data"

By

Eleftheria Apostolaki

Student ID Number: P2821803

Name of Supervisor: Dimitris Karlis

June 2021

Athens, Greece

Table of Contents

List of Tables

List of Figures

I. Introduction

A. Clustering & Applications

Clustering is a division of data into groups of similar objects. Each group - called cluster - consists of objects that are similar amongst themselves and dissimilar compared to objects of other groups. Cluster analysis is the organization of a collection of patterns (usually represented as a vector of measurements, or a point in a multidimensional space) into clusters based on similarity. Patterns within a valid cluster are more similar to one another than they are to a pattern belonging to a different cluster.

Cluster analysis is a main task of exploratory data mining and a common technique for statistical data analysis used in many fields, including pattern recognition, image analysis, information retrieval, bioinformatics, data compression, computer graphics and machine learning. Depending on the application, what is meant by a "cluster" may differ a lot, and cluster definition and methodology have to be adapted to the specific aim of clustering in the application of interest. Different aims served through clustering can indicatively be:

- delimitation of species of plants or animals in biology,
- medical classification of diseases,
- discovery and segmentation of settlements and periods in archaeology,
- image segmentation and object recognition,
- social stratification,
- market segmentation,
- efficient organization of databases for search queries

There are also various general tasks for which clustering is applied in many subject areas such as:

- exploratory data analysis looking for "interesting parties" without prescribing any specific interpretation, potentially thus creating new research questions and hypotheses,
- information reduction and structuring of sets of entities from any subject area for simplification, effective communication, or effective

access/action such as either complexity reduction for further data analysis or classification systems,

- the investigation of the correspondence of a clustering in specific data with other groupings or characteristics, whether hypothesized or derived from other data.

B. Traditional Clustering Algorithms

Although the presentation of a complete list of all the existing clustering algorithms is challenging due to the diversity of information, the intersection of research fields and the development of modern computer technology, the traditional clustering algorithms can be divided into 9 categories within which some of the most used clustering methods are included. As it will be analyzed below, these algorithms are explicitly capable of handling datasets that consist of a single type of variable (either numeric or categorical).

Clustering Algorithms Based on Partition

The basic idea of this kind of clustering algorithms is to regard the centre of data points as the centre of the corresponding cluster. K-means and K-medoids are the two most famous of their kind of clustering algorithms. The core idea of K-means is to update the center of cluster represented by the center of data points by iterative computation and the iterative process will then be continued until some criteria for convergence are met. K-mediods is an improvement of K-means aiming at dealing with discrete data, which takes the data point, most near the center of data points, as the representative of the corresponding cluster. The typical clustering algorithms based on partition also include PAM, CLARA, CLARANS.

Clustering Algorithms Based on Hierarchy

The basic idea of this kind of clustering algorithms is the construction of the hierarchical relationship among data in order to cluster. Typical algorithms of this kind of clustering include BIRCH, CURE, ROCK and Chameleon. BIRCH registers

the clustering result by constructing the feature tree of clustering, CF tree, one node of which stands for a subcluster. CF tree will dynamically grow when a new data point comes. CURE, suitable for large-scale clustering, takes a random sampling technique to cluster samples separately and integrates the results in the end. ROCK is an improvement of CURE for dealing with data of enumeration type, taking into consideration the effect on the similarity from the data around the cluster. Chameleon, initially, divides the original data into clusters of smaller size based on the nearest graph, and then the clusters of small size are merged into a cluster of a bigger size, based on agglomerative algorithm.

Clustering Algorithms Based on Fuzzy Theory

The basic principle behind this kind of clustering algorithms is that the discrete value of belonging label, {0, 1}, is changed into the continuous interval [0, 1], with a view to describing the belonging relationship among objects more rationally. Typical algorithms of this kind of clustering include FCM, FCS and MM. The core idea of FCM is to get membership of each data point to every cluster by optimizing the object function. FCS, different from the traditional fuzzy clustering algorithms, takes the multidimensional hypersphere as the prototype of each cluster, so as to cluster with the distance function based on the hypersphere. MM, based on the Mountain Function, is used to find the centre of cluster.

Clustering Algorithms Based on Distribution

The basic concept is that the data, generated from the same distribution, belongs to the same cluster if several distributions in the original data exist. The typical algorithms are DBCLASD and GMM. The core idea of DBCLASD, a dynamic incremental algorithm, is that if the distance between a cluster and its nearest data point satisfies the distribution of expected distance generated from the existing data points of that cluster then the nearest data point should belong to this cluster. The core idea of GMM is that GMM consists of several Gaussian distributions from which the original data is generated and the data, obeying the same independent Gaussian distribution, is considered to belong to the same cluster.

Clustering Algorithms Based on Density

The basic idea of this kind of clustering algorithms is that the data located in the region with high density of the data space is considered to belong to the same cluster. The typical ones include DBSCAN, OPTICS and Mean-shift. DBSCAN is the most well known density-based clustering algorithm, generated from the basic idea of this kind of clustering algorithms directly. OPTICS is an improvement of DBSCAN and it overcomes the shortcoming of DBSCAN of being sensitive to two parameters - the radius of the neighborhood and the minimum number of points in a neighbourhood. In the process of Mean-shift, the means of offsetting the current data point is calculated at first, then the next data point is figured out based on the current data point and the offset, and last, the iteration will be continued until some criteria are met.

Clustering Algorithms Based on Graph Theory

According to this kind of clustering algorithms, clustering is realized on the graph where the node is regarded as the data point and the edge is regarded as the relationship among data points. Typical algorithms of this kind of clustering are CLICK and MST-based. The core idea of CLICK is to carry out the minimum weight division of the graph with iteration in order to generate the clusters. Generating the minimum spanning tree from the data graph is the key step to do the cluster analysis for the MST-based clustering algorithm.

Clustering Algorithms Based on Grid

The basic notion of this kind of clustering algorithms is that the original data space is changed into a grid structure with definite size for clustering. Typical algorithms of this kind of clustering are STING and CLIQUE. The core concept of STING which can be used for parallel processing is that the data space is divided into many rectangular units through constructing the hierarchical structure and the data within different structure levels is clustered respectively. CLIQUE takes advantage of the grid-based clustering algorithms and the density-based clustering ones.

Clustering Algorithms Based on Fractal Theory

Fractal stands for the geometry which can be divided into several parts sharing some common characteristics with the whole. The typical algorithm of this kind of clustering is FC, the core idea of which is that the change of any inner data of a cluster does not have any influence on the intrinsic quality of the fractal dimension.

Clustering Algorithms Based on Model

The basic intention is to select a particular model for each cluster and find the best fitting for that model. There are mainly two kinds of model-based clustering algorithms, one is based on a statistical learning method and the other is based on a neural network learning method. The typical algorithms based on the statistical learning method are COBWEB and GMM. The core idea of COBWEB is to build a classification tree based on some heuristic criteria in order to realize hierarchical clustering on the assumption that the probability distribution of each attribute is independent. The typical algorithms based on neural network learning method are SOM and ART. The core idea of SOM is to build a mapping of dimension reduction from the input space of high dimension to output space of low dimension on the assumption that there is topology in the input data. The core idea of ART, an incremental algorithm is to generate a new neuron dynamically to match a new pattern which will then create a new cluster when the current neurons are not enough.

The detailed and comprehensive comparisons of the afore mentioned clustering algorithms are summarized in the table below as follows:

Category	Typical Algorithm
Clustering algorithm based on partition	K-means, K-medoids, PAM, CLARA, CLARANS
Clustering algorithm based on hierarchy	BIRCH, CURE, ROCK, Chameleon
Clustering algorithm based on fuzzy theory	FCM, FCS, MM

Clustering algorithm based on distribution	DBCLASD, GMM
Clustering algorithm based on density	DBSCAN, OPTICS, Mean-shift
Clustering algorithm based on graph theory	CLICK, MST
Clustering algorithm based on grid	STING, CLIQUE
Clustering algorithm based on fractal theory	FC
Clustering algorithm based on model	COBWEB, GMM, SOM, ART

Table 1 *Traditional Algorithms*

C. Clustering Mixed Mode Data

The advent of sophisticated tools of measurement has given rise to new modes of data collection for cluster analysis. As a result, data often come along with complex dependence structures. These complex structures typically require non-standard statistical approaches that usually entail computationally intensive methodologies. Conventional tools generally rely on the assumption that data or some suitable transformations of them, follow a normal distribution. This assumption no longer applies to these contexts directly. Over the past 20 years, there have been remarkable advancements in statistical methodology for the analysis of such data. The development of statistical software and packages has unfortunately not kept pace with these methodological advances, but practitioners nonetheless now have a host of increasingly sophisticated tools available to them for handling complex data. This has made their adoption as well as their application in the solution of important substantive problems across several disciplines made possible particularly in engineering, finance, medicine and health.

Multivariate data comprising mixtures of discrete (i.e., categorical, binary, count) and continuous measurements (also referred to as "non-commensurate" outcomes) are a particularly common example of non-standard correlated data in practice. And since most real data contain different types of variables, an important area in cluster analysis deals with clustering mixed data. In practice, mixed data sets arise when the variables observed consist of heterogeneous sets of variables indicating

several variable types, e.g., numeric, categorical, etc. Numeric features can take real values such as height, weight, and distance. Categorical features represent data that can be divided into a fixed number of categories such as colour, race, gender, profession, and blood group. The capability of dealing with datasets containing both numeric and categorical attributes is undoubtedly crucial since datasets with mixed types of characteristics are the most common case in real life data mining applications.

As it is already acknowledged, clustering algorithms group data points into clusters using some notion of "similarity" which can be as simple as the Euclidean distance. To compute the similarity between numeric feature values mathematical operations (such as distances, angles, summation, or mean) are applied to them. Distance-based similarity measures are mostly utilised for numeric data points. In general, categorical feature values are not inherently ordered (e.g., red and blue) and consequently it is not possible to directly compute the distance between two categorical feature values. Therefore, computing distance-based similarity measures for categorical data is a demanding task partly justifying the existence of several options for clustering mixed data. The fundamental challenges encountered when dealing with this kind of data is to equitably balance the contribution from continuous and categorical variables (considering that the current clustering algorithms are unable to properly handle data sets in which only a subset of variables is related to the underlying cluster structure of interest) and to characterise the nature of relationships between measurements of different and/or the same subjects either over time or cross-sectionally in one or more spatial dimensions. Furthermore, the ad hoc approach of carrying out separate analyses for the numeric and continuous variables in the data is not only clearly deficient in many applications, but also not a straightforward undertaking.

It should be highlighted that the primary focus in the literature for the last decades has been on clustering data sets that are comprised of a single type, that is all variables are either numeric, or categorical (nominal or ordinal). For cluster analysis based on mixed-type data, comparatively few clustering methods are available. As such, the analysts working with data sets containing a mix of numeric and categorical-valued data apply clustering algorithms that usually adopt the following approaches: convert the data set to a single data type (by either coding the categorical variables as numbers and applying methods designed for continuous variables or converting the continuous variables into categorical variables via interval-based bucketing) or

directly handle mixed data clustering. And the fact that few researchers have developed methods for converting a mixed dataset to a pure numeric dataset so that clustering algorithms meant for pure numeric datasets can be employed is of interest. So, this is indeed a new perspective on the challenging problem of mixed data clustering. In general, the clustering methods for mixed-type data which belong to the first approach (i.e. that convert the data sets to a single data type) suffer from significant drawbacks including loss of information due to discretization thus raising an open demand to the research community to develop algorithms able to reduce the adverse effects of data transformation, computation of distances between any observations not feasible for large data sets, arbitrary weighting of continuous versus categorical variables (e.g., as in dummy coding or the similarity metric of Gower 1971) and inability to equitably balance the contribution of continuous and categorical variables without strong parametric assumptions or difficult-to-specify tuning parameters.

Regarding the area of big data mentioned, it is common knowledge that most successful machine learning algorithms lose their interpretability and may be treated as a black box when datasets increase in size and domains become complex. Hence, mixed data clustering algorithms are no exception to this. The idea of clustering models that are easy to explain is appealing to practitioners such as clinicians, business analysts, geologists, and biologists - interpretable models can assist them in making informed decisions. Unfortunately, only a few researchers have explored this area of developing interpretable mixed data clustering methods to address critical aspects of the models: e.g., why a certain set of data points forms a cluster or how different clusters can be distinguished from one another. Novel research in this area will produce outcomes outside the realms of the research community. Many clustering algorithms may benefit from reducing the dimensions of multivariate mixed data as a result of reducing their execution time and model complexity. Recent research in the field of feature selection for mixed data has been carried out; however, combining such results with clustering has not been satisfactorily explored. The selection of a subset of relevant features has the potential of enhancing the interpretability of clustering algorithms as well. Another repercussion of big data is ensuring the scalability of clustering algorithms so as to make them useful in real-world scenarios. Parallelization of mixed data clustering algorithms is a viable approach to allow scaling with increasing data size and maintaining linear time complexity (especially

for partitional clustering). Active research in this area is in progress to keep the field in synchronization with big data challenges. Similarly, developing fast and accurate online clustering algorithms to handle large streams of mixed data requires attention to address shortcomings which include low clustering quality, evaluation of new concepts and concept drift in the underlying data, difficulties in determining cluster centers, and insufficient ability to deal with outliers.

Significant Application Areas

Employing mixed data clustering in multiple domains is of paramount importance. Mixed data frequently occur in many applications such as health, marketing, business, finance, and social studies. For instance, in the field of health and biology, McParland, Gormley, and McParland et al. developed a mixed data clustering algorithm to study high dimensional numeric phenotypic data and categorical genotypic data. The study led to a better understanding of metabolic syndrome (MetS). Malo et al. used mixed data clustering to study people who died of cancer in Hijuelas between 1994 and 2006. Storlie et al. developed a model-based clustering for mixed datasets with missing feature values to cluster autism spectrum disorder. Aerts et al. gave several examples from developmental toxicology where fetal data from laboratory animals include binary, categorical, and continuous outcomes. More recently, Daniels and Normand analysed mixed patient data to profile the performance of regional networks of health care providers in the United States. Researchers have used various types of clustering approaches for mixed data for heart disease, occupational medicine, digital mammograms, acute inflammations, age of abalone snails, human life span, dermatology, medical diagnosis, toxicogenomic, genetic regulation, analysis of biomedical datasets, and cancer.

Another example in the field of business and marketing is Hennig and Liao who applied mixed data clustering techniques for socio-economic stratification by using a 2007 US data survey of consumer finances. Kassi et al. developed a mixed data clustering algorithm to segment gasoline services stations in Morocco to determine important features able to influence the profit of these service stations. Mixed data clustering has also been used in credit approval, income prediction (adult data), marketing research, customer behaviour discovery, customer segmentation and catalogue marketing, customer behaviour pattern discovery, motor insurance, and

construction management. Other applications could be these of Moustaki and Papageorgiou who applied mixed data clustering in archaeometry for the classification of archaeological findings into groups. Philip and Ottaway used mixed data clustering to cluster Cypriot hooked-tang weapons. Chiodi used mixed data clustering for andrological data and Iam-On and Boongoen for student dropout prediction in a Thai university. Mixed data clustering has also been used in teaching assistant evaluation, class examination, petroleum recovery, intrusion detection, forest cover type, online learning systems, automobiles, printing process delays and country flags mining.

Impact Areas

As discussed in the previous subsection, mixed data clustering algorithms have been applied in various application domains. Although employing mixed data clustering in multiple domains is very important, areas like these of health and business informatics will have more impact because they attempt to solve real-world problems related to people.

More specifically, most of the data for health applications are based on either electronic health records (EHR) or sensors. EHR data can contain a patient's medical history, diagnoses, medications, treatment plans, immunization dates, allergies, radiology images as well as laboratory and test results. EHR is a great resource to allow the deployment of evidence based machine learning tools (supervised and unsupervised) to make decisions about patients' care. Therefore, EHR data is a good example of mixed data with high-impact real-world applications. Data from sensors can be either numeric (e.g., motion or physiology) or categorical (e.g., door open or closed). These datasets are important in building machine learning driven applications for rehabilitation, assessment of medical conditions, and detection and prediction of health-related events. Application of mixed data clustering on these datasets is crucial in identifying medical conditions among people with disability, morbidity or cognitive disorders. Clustering on these diverse datasets can also help in performing sex and gender-based research, vulnerable populations and older adults.

Business analytics is another domain in which a large number of mixed datasets is created. Market research is an important area in this domain. Analysis of customer datasets containing categorical features (e.g., type of a customer, preference, and

income group) and numeric features (e.g., age and the number of transactions) provide managers with insights into customer behaviour. Credit card data analysis is used for the prediction of the financial health of an individual. Typically, credit card datasets are mixed datasets on which various clustering algorithms have been applied. The financial statements of a company are analysed to assess the company's financial health; the datasets consisting of categorical features (e.g., the type, products, and the region of the company) along with numeric features (e.g., financial ratios) present better information about a company. People analytics is also an emerging area: companies are interested in knowing about present and potential employees to improve both productivity and satisfaction. Employee datasets consisting of categorical features (e.g., education, department, and job type) and numeric features (e.g., age, years in job, and salary) can capture information about employees better than datasets containing only one type of feature.

In the current thesis, the research problem that will be approached covers the clustering of mixed mode data, its benefits and applications. In the subsequent Chapter 2, the literature review for clustering mixed mode data is detailed including the methodologies that will be used as part of the thesis and any additional methodologies that are available for this type of clustering according to the bibliography. In Chapter 3, a detailed overview and analysis is presented for the dataset on which the selected clustering methods will be applied while in Chapter 4 the clustering results and their interpretation are provided. In chapter 5 that follows, the conclusions drawn from this research are described along with any future work required for the clustering of mixed mode data.

II. Literature Review: Methodologies for clustering mixed mode data

In this chapter, the clustering algorithms to be used as part of the current thesis are analysed and compared and other clustering algorithms for mixed mode data included in the relevant bibliography are also presented.

A. Comparison of used clustering methods

Within the scope of this thesis, three clustering algorithms are used: K-Prototypes, Kamila, and Latent Variable Model, each one of which can cluster mixed-type data.

The K-Prototypes algorithm, proposed by Huang, belongs to the family of distance-based and partitional cluster algorithms and it is an extension to the K-means and K-modes algorithms. The algorithm depends on the concept of K-means algorithm and, furthermore, it eliminates the numeric data constraint of that algorithm. In this type of algorithms, the centre of data point becomes the centre of the consistent cluster. The cluster centres are represented by mean values for numeric features and mode values for categorical features. The cluster objects are subsequently divided, and the clusters are updated depending on the partition. However, as proposed by the bibliography and will be further examined as part of the thesis, there are some shortcomings in this clustering process: (1) The random selection of the initial cluster centres results in the uncertainty and randomness of the clustering results, and the number of clusters should be manually determined; (2) the simple Hamming distance is used to calculate the dissimilarity between the categorical data (0 or 1 depending on whether the feature values are same or different) and the cluster centres resulting in the loss of information and the inability to objectively reflect the real situation between the data objects and the clusters resulting in inaccurate clustering results; (3) the parameter used to adjust the proportion between categorical data and numerical data needs to be manually determined; and (4) the structural characteristics of categorical data and numerical ones and the overall distribution of datasets have not been fully considered. This algorithm can be found in the (kproto, Kproto) function of the *clustMixType* R package.

The KAMILA (KAy-means for MIxed LArge data) algorithm, proposed by Foss, A., Markatou, M. and Ray, B., is a model-based adaptation of the K-means algorithm and the Gaussian multinomial mixture models. It overcomes the challenges inherent in the various extant methods for clustering mixed continuous and categorical data. More specifically, in this type of clustering, the variables (i.e., interval, nominal, or categorical scale) are used in their original measurement scale and hence they are not transformed to either all numeric or all categorical ones thus avoiding a loss of information. Moreover, this algorithm ensures: (1) a sensible balancing between continuous and categorical variables by using the properties of Gaussian-multinomial mixture models and (2) the avoidance of overly restrictive parametric assumptions for numeric features generalizing the form of the clusters to a broad class of elliptical distributions as it happens in K-means. The sample of continuous variables is assumed to follow a mixture distribution with arbitrary spherical clusters (where the density of the data is only dependent on the distance to the centre of the distribution) by using a kernel density estimation technique, the categorical variables are supposed to be sampled from a mixture of multinomial variables and the Modha-Spangler weighting of variables is also used in which categorical variables must be transformed into indicator variables in advance. The latter results in not requiring the user to specify any variable weights or use coding schemes which facilitate algorithm reproducibility and ease of use. This algorithm begins with a set of centroids for the continuous variables and a set of parameters for the categorical variables. For continuous variables, the Euclidean distance with the closest centroid is computed. This set of minimal distances is used for the estimation of the mixture distribution of continuous variables. For categorical variables, the probabilities of observing the data in the corresponding cluster are computed. The log-likelihood of the sum of these components is then used to find the most appropriate cluster for each subject. Based on this temporary partition, the centroids and the parameters are updated to best represent the clusters. These steps are repeated until the clusters are stable. Finally, multiple runs of this process are performed with various initializations and the partition maximizing the sum of the best final likelihoods is retained. The R package *Kamila* is a direct implementation of this technique by its authors.

The Latent Variable Model, proposed by McParland, D. and Gormley, I.C., is a model-based clustering procedure which uses a latent variable model. It is proposed that a latent variable, following a mixture of Gaussian distributions, generates the

observed data of mixed type. The observed data may be any combination of continuous, binary, ordinal or nominal variables. A parsimonious covariance structure is employed for the latent variables, leading to a suite of six clustering models that vary in complexity and provide an elegant and unified approach to clustering mixed data. An expectation maximisation (EM) algorithm is used to estimate clustMD; in the presence of nominal data a Monte Carlo EM algorithm is required. The Latent Variable Model algorithm is implemented by its authors in the *clustMD* R package.

The Latent Class Model, proposed by Marbac, M., Sedki, M. and Patin, T., is model-based clustering focusing on variable selection. The proposed model considered two types of variables (relevant - having a different distribution among components - and irrelevant - having the same distribution among components) and assumes independence within components. Additionally, it presents two methods of model selection: the first one simultaneously performs model/variable selection and parameter inference and optimizes the Bayesian Information Criterion (BIC) with a modified version of the standard expectation – maximization algorithm (EM) to permit the maximization of the penalized likelihood while the second one selects variables of a diagonal Gaussian mixture model without requiring parameter inference by maximizing the Maximum Integrated Complete-data Likelihood criterion (MICL). An important difference between the previous cluster analysis techniques and LCM is that the latter is a model-based approach which means that it is postulated for the population the data sample is obtained from. More precisely, it is assumed that a mixture of underlying probability distributions generates the data. The specific clustering approach is similar to standard non-hierarchical cluster techniques such as k-means clustering, in which the allocation of objects to clusters should be optimal according to some criteria. These criteria typically involve minimizing the within-cluster variation or, equivalently, maximizing the between-cluster one. An advantage of using a statistical model is that the choice of the cluster criterion is less arbitrary and the approach includes rigorous statistical tests. Therefore, instead of finding clusters with a chosen distance measure as happens in the other algorithms, a model is used that describes the data distribution and depending on this model, the probabilities that certain cases are members of certain latent classes are assessed. Moreover, LC clustering is very flexible as both simple and complicated distributional forms can be used for the observed variables within clusters. As in any statistical model, restrictions can be imposed on the parameters to obtain more parsimony and

formal tests can be used to check their validity. Another advantage of the model-based clustering approach is that several criteria can be used to assess the optimal number of clusters- a direct consequence of the statistical model used to describe the data- and no decisions must be made about the scaling of the observed variables. The first one is a large advantage compared to, e.g., hierarchical clustering methods where a cut-off value must be chosen by the user and in most cases, no clear criteria exist for such a choice. For the second one, when working with normal distributions with unknown variances, the results will be the same irrespective of whether the variables are normalized or not. This is very different from standard non-hierarchical cluster methods like k-means, where scaling is always an issue. Other advantages are that it is relatively easy to deal with variables of mixed measurement levels (different scale types) and that there are more formal criteria to be met in order to make decisions about the number of clusters (i.e., Bayesian information criterion (BIC) and dissimilarity-based ones) and other model characteristics. The Latent Class Model algorithm is implemented by its authors in the *VarSelLCM* R package.

After thoroughly analysing the afore mentioned clustering methods, the subsequent Table 2 summarizes these selected algorithms as far as three design areas are concerned in the specific setting of mixed data clustering. The first area relates to how these algorithms calculate similarities/distances for categorical and numeric data when they are distance-based algorithms or how they transform the data when they are model-based methods. The second area refers to the methodology used to merge numerical and categorical parts. The last area concerns the algorithm choice regarding the approach to be used to build optimal clusters.

As seen below, algorithms are grouped as distance-based or model-based - K-Prototypes algorithm belongs to the first category while Kamila, Latent Variable Model and Latent Class Model belong to the second category. It is concluded that the dissimilarity criterion used by the K-Prototypes algorithm is different between numeric and categorical data as Euclidean distance is calculated for the numeric variables and Hamming distance for the categorical ones. For this method, merging is also required by using the weighted sum of the distances with the optimization algorithm to be this of K-means. For Latent Class Model, the distributions of both numeric and categorical variables are transformed into probabilities as opposed to Kamila where Euclidean distance is used to handle numeric variables and probabilities for categorical ones. Due to this different variable handling, Kamila needs to set up

an ensemble-like approach to merge both types of data using both K-means and Expectation Maximization (EM) as optimization algorithms. On the other hand, Latent Class Model does not need any merging procedure as both types of variables are included in a unique probabilistic model and EM algorithm is used with specific variants.

Clustering Method	Distance or transformation		Merge Mode	Optimization Algorithm
	Numeric	Categorical		
Distance-based methods				
K-Prototypes	Euclidean	Hamming	Weighted sum	K-means
Model-based methods				
Kamila	Euclidean	Probabilities	Ensemble-like approach	K-means & EM
Latent Class Model (LCM)	Probabilities	Probabilities	NA	EM & feature selection

Table 2 *Description of selected methods with regards to design questions related to (1) similarities/distances or data transformation, (2) methodology to merge numerical and categorical parts and (3) algorithm choice*

B. Analysis of used clustering methods

Kamila Clustering

Detailed material for the Kamila clustering method can be found in the bibliography:

- Foss, A., Markatou, M., Ray, B. and Heching, A. (2016). A semiparametric method for clustering mixed data. Machine Learning, Volume 105(Issue 3), pp.419–458.
- Foss, A.H. and Markatou, M. (2018). kamila: Clustering Mixed-Type Data in R and Hadoop. Journal of Statistical Software, Volume 83(Issue 13).

Notation & Definitions

Consider a data set that consists of N independent and identically distributed observations of a (P + Q) - dimensional vector of random variables $(V^T, W^T)^T$ that follow a finite mixture distribution with g components. V is a P-dimensional vector of continuous random variables and W is a vector of Q categorical random variables where the q^{th} element of W has L_q categorical levels denoted 1, 2, . . ., L_q with q = 1, 2, . . ., Q. Note that vector V is conditionally independent of W, given population membership.

Given membership in the g^{th} cluster, the following are considered:

- V is modelled as a vector following a mixture distribution with arbitrary spherical clusters with individual component density functions $fv_{,g}(v; \mu_g, \Sigma_g)$, where g is the number of clusters in the mixture, μ_g denotes the g^{th} centroid, and Σ_g the g^{th} scaling matrix.
- W is modelled as a vector following a mixture of multinomial random variables with individual component probability mass functions $fw_{,g}(w) = \Pi_q^Q =_1 m(w_q; \theta_{gq})$, where $m(\cdot; \cdot)$ denotes the multinomial probability mass function, and θ_{gq} is the parameter vector of the multinomial mass function of the q^{th} categorical variable for the g^{th} cluster.
- Taking into account the local independence assumption, the joint density of $(V^T, W^T)^T$ is

$$f_{\mathbf{V},\mathbf{W},g}(\mathbf{v},\mathbf{w};\ \mu_g,\Sigma_g,\theta_{gq}) = f_{\mathbf{V},g}(\mathbf{v};\mu_g,\Sigma_g)\prod_{q=1}^{Q} m(w_q;\theta_{gq}),$$

with the overall density unconditional on cluster membership to be

$$f_{\mathbf{V},\mathbf{W}}(\mathbf{v},\mathbf{w}) = \sum_{g=1}^{g} \pi_g f_{\mathbf{V},\mathbf{W},g}(\mathbf{v},\mathbf{w};\ \mu_g,\Sigma_g,\theta_{gq}), \quad (1)$$

where π_g is the prior probability of observing the g^{th} cluster.

Kernel Density Estimation

The Proposition 2 states that if $X \in \mathbb{R}^p$ follows a spherically symmetric distribution with centre μ, then

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{f_R(r)\,\Gamma(\frac{p}{2}+1)}{p r^{p-1}\pi^{p/2}}, \tag{2}$$

where r = $\sqrt{(x-\mu)^T(x-\mu)}$, R = $\sqrt{(X-\mu)^T(\mathrm{X}-\mu)}$ and f_R is the probability density of R. The $\widehat{f_R}$ is constructed by using a kernel density estimation scheme, which is then substituted into (2) in place of f_R. Note that X corresponds to vector V above within a particular cluster, and that by using a scaling matrix Σ_g, the result can be extended to elliptical distributions. The KAMILA function currently uses Σ_g equal to the identity matrix.

Algorithm Description

Kamila proceeds by estimating the unknown parameters of (1) through an iterative process. For each initialization, the algorithm runs iteratively for multiple times until either a pre-specified maximum number of iterations is reached or until population membership is unaltered from the previous iteration.

At the t^{th} iteration of the algorithm, let $\hat{\mu}_g^{(t)}$ denote the estimator for the centroid of population g and $\hat{\theta}_{gq}^{(t)}$ denote the estimator for the parameters of the multinomial distribution corresponding to the q^{th} discrete random variable drawn from population g. The estimation procedure proceeds iteratively, with each iteration consisting of two broad steps: the **partition** and the **estimation** ones. The partition step assigns each of N observations to one of g clusters and the estimation step re-estimates the parameters of interest by using the memberships of the new cluster.

Given a complete set of and $\hat{\mu}_g^{(t)}$ and $\hat{\theta}_{gq}^{(t)}$'s at the t^{th} iteration, the Euclidean distance from observation *i* to each of the $\hat{\mu}_g^{(t)}$'s is calculated as:

$$d_{ig}^{(t)} = \sqrt{\sum_{p=1}^{P}[\xi_p(v_{ip} - \hat{\mu}_{gp}^{(t)})]^2},$$

where ξ_p is an optional weight for the variable p. The minimum distance for the i^{th} observation is $r_i^{(t)} = \min_g(d_{ig}^{(t)})$ and the kernel density estimate of the minimum distances is calculated as follows:

$$\hat{f}_R^{(t)}(r) = \frac{1}{Nh^{(t)}} \sum_{\ell=1}^{N} k\left(\frac{r - r_\ell^{(t)}}{h^{(t)}}\right), \tag{3}$$

where $k(\cdot)$ is a kernel function and $h^{(t)}$ is the corresponding bandwidth parameter at iteration t. The Gaussian kernel is currently used with bandwidth $h = 0.9An{-}1/5$, where $A = \min(\hat{\sigma}, \hat{q}/1.34)$, $\hat{\sigma}$ is the sample standard deviation and $\hat{q}$ is the sample interquartile range. The function $\widehat{f_R}^{(t)}$ is used to construct $\widehat{f_x}^{(x)}$, as shown above. Taking into account the independence between the Q categorical variables within a given cluster g, the log probability of observing the i^{th} categorical vector given population membership is calculated as $\log(c_{ig}^{(t)}) = \Sigma_q^Q = {}_1\, \xi_q \cdot \log(m(w_{iq}; \hat{\theta}_{gq}^{(t)}))$, where $m(\cdot;\cdot)$ is the multinomial probability mass function and ξ_q is an optional weight for the variable q.

The i^{th} object is assigned to the cluster g when the following function is maximized:

$$H_i^{(t)}(g) = \log\left[\hat{f}_{\mathbf{V}}^{(t)}(d_{ig}^{(t)})\right] + \log\left[c_{ig}^{(t)}\right], \tag{4}$$

Given a partition of the N observations at iteration (t), the estimation step calculates $\hat{\mu}_{gp}^{(t+1)}$ and $\hat{\mu}_{gq}^{(t+1)}$ for all g, p, and q. Given that $\Omega_g^{(t)}$ denotes the set of indices of observations assigned to cluster g at iteration t, the parameter estimates are calculated as:

$$\hat{\boldsymbol{\mu}}_g^{(t+1)} = \frac{1}{\left|\Omega_g^{(t)}\right|} \sum_{i \in \Omega_g^{(t)}} \mathbf{v}_i$$

$$\hat{\theta}_{gq\ell}^{(t+1)} = \frac{1}{\left|\Omega_g^{(t)}\right|} \sum_{i \in \Omega_g^{(t)}} I\{w_{iq} = \ell\}$$

where $I\{\cdot\}$ denotes the indicator function and $|A|$ denotes the cardinality of the set A. The partition and estimation steps are repeated until a stable solution is reached (or

the maximum number of iterations is reached. For each initialization, we calculate the objective function as follows:

$$\sum_{i=1}^{N} \max_{g} \{H_i^{(final)}(g)\}. \tag{5}$$

After all runs are executed, the best partitioning which maximizes the objective function of equation (5) over all runs is selected.

K-Prototypes Clustering

Detailed material for the K-Prototypes clustering method can be found in the following bibliography:

- Huang, Z. (1998). Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery 2, pp.283–304.
- Szepannek, G. (2019). clustMixType: User-Friendly Clustering of Mixed-Type Data in R. The R Journal, Volume 10(Issue 2), p.200.

The Model

The current algorithm is almost identical to the k-means: initial prototypes are selected as temporary centres of the clusters and then each subject is allocated to the closest prototypes. When all subjects are allocated, the prototypes are updated to represent their optimal class. The subjects are then reallocated to the updated prototypes if needed and the process is repeated until the partition is stable. The objective function in K-Prototypes clustering is given by:

$$E = \sum_{i=1}^{N} \sum_{j=1}^{k} u_{ij} d\left(x_i, \mu_j\right), \tag{1}$$

where x_i, $i = 1, \ldots, N$ are the observations in the sample, μ_j, $j = 1, \ldots, k$ are the cluster prototype observations and u_{ij} are the elements of the binary partition matrix U_{Nxk} satisfying $\Sigma_j^k = {}_1\, u_{ij} = 1, \forall i$.

The distance function between two points is given by:

$$d(x_i, \mu_j) = \sum_{m=1}^{q} \left(x_i^m - \mu_j^m\right)^2 + \lambda \sum_{m=q+1}^{p} \delta\left(x_i^m, \mu_j^m\right). \quad (2)$$

As seen above, *d()* is the weighted sum of Euclidean distance between two points in the metric space and simple matching distance for categorical variables where δ(a, b) = 0 for a = b and δ(a, b) = 1 for a <> b. In equation (2), m is an index over all variables in the data set where the first q variables are numeric and the remaining p - q variables are categorical. The parameter λ defines the trade-off between both terms and has to be specified in advance as the number of clusters k does as well. Note that the impact of the categorical variables increases as the value of λ increases while for λ = 0, the impact of the categorical variables vanishes, and only numeric variables are taken into account.

As in the traditional K-means algorithm, the means for the numeric variables and the mode for the categorical variables minimizes the total distance within the cluster. More specifically, the following steps are applied in the iterations of the algorithm:

1. Random cluster prototypes are initialized.
2. For each point (observation):
   (a) The closest prototype according to d() is assigned.
   (b) Upon their allocation, the cluster prototypes are updated by cluster - specific means/modes for all variables.
3. In case any observations have swapped their cluster assignment in step (2), or the maximum number of iterations has not been reached, then step (2) is repeated.

Latent Variable Model

Detailed material for the Latent Variable Model method can be found in the following bibliography:

- Mcparland, D. and Gormley, I.C. (2015). Model Based Clustering for Mixed Data: clustMD. Advances in Data Analysis and Classification, [online] Volume 10(Issue 2), pp.155–169.

The Model

This model employs a mixture of latent variable models to cluster mixed type data. It assumes the observed J mixed type variables in each observation vector y_i are a manifestation of an underlying latent continuous vector, z_i (for $i = 1, \ldots, N$), which follows a Gaussian mixture distribution.

Modelling Continuous Data

Under the clustMD model, continuous variables follow a multivariate Gaussian distribution, i.e., if variable j is continuous, $y_{ij} = z_{ij} \sim N(\mu_j, \sigma_j^2)$.

Modelling Ordinal Data

In the case of an ordinal variable, it is supposed that the observed response, y_{ij} is a categorical manifestation of the latent continuous variable, z^{ij}, as is typical in item response theory models (Johnson and Albert 1999; Fox 2010). For ordinal variable *j* with K_j levels let γ_j denote a $K_j + 1$ vector of thresholds that partition the real line. The value of the latent z_{ij} in relation to γ_j determines the observed ordinal response y_{ij}. The threshold parameters are constrained such that $-\infty = \gamma_{j,0} \leq \gamma_{j,1} \leq \cdots \leq \gamma_{j,Kj} = \infty$. If the latent z_{ij} is such that $\gamma_{j,k-1} < z_{ij} < \gamma_{j,k}$ then the observed ordinal response, $y_{ij} = k$. The latent z_{ij} follows a Gaussian distribution i.e., $z_{ij} \sim N(\mu_j, \sigma_j^2)$. Thus, the probability of observing level *k* can be expressed as the difference between two Gaussian cumulative distribution functions (CDF) denoted by Φ: $P(y_{ij} = k) = \Phi$

$\left(\frac{\gamma_{j,k}-\mu_j}{\sigma_j}\right) - \Phi\left(\frac{\gamma_{j,k-1}-\mu_j}{\sigma_j}\right)$. The threshold parameters are invariant under translation and their values are not of primary interest in clustMD. Thus, for reasons of identifiability and efficiency, $\gamma_{j,k}$ is fixed such that $\gamma_{j,k} = \Phi^{-1}(\delta_k)$, where δ_k is the proportion of the observed values of variable *J* which are less than or equal to level *k*. A binary variable can be thought of as an ordinal variable with two levels, denoted 1 and 2. Thus if variable j is binary, then $P(y_{ij} = 2) = 1 - \Phi\left(\frac{\gamma_{j,1}-\mu_j}{\sigma_j}\right)$.

Modelling Nominal Data

Nominal variables are more difficult to model since the set of possible responses is unordered. In this case, a multivariate latent vector is assumed to underlie the observed nominal variable. For nominal variable *j* with K_j possible responses, the underlying continuous vector has $K_j - 1$ dimensions, i.e., $z_{i\,j} = (z_{ij}^1, \ldots, z_{ij}^{K_j-1}) \sim MVNKj-1(\mu_j, \Sigma_j)$, where MVN denotes the multivariate Gaussian distribution. The observed nominal response y_{ij} is a manifestation of the values of the elements of z_{ij} relative to each other and to a threshold, assumed to be 0. That is,

$$y_{ij} = \begin{cases} 1 \text{ if } \max_s\{z_{ij}^s\} < 0; \\ k \text{ if } z_{ij}^{k-1} = \max_s\{z_{ij}^s\} \text{ and } z_{ij}^{k-1} > 0 \text{ for } s = 2, \ldots, K_j. \end{cases}$$

Binary data can be considered as nominal with two unordered responses. This model for nominal data is equivalent to the proposed ordinal data model in such a case. A similar latent variable approach to modelling nominal data is the mutinomial probit model (Geweke et al. 1994).

A Joint Model for Mixed Data

Let Y denote a data matrix with *N* rows and *J* columns. Without loss of generality, suppose that the continuous variables are in the first C columns, the ordinal and binary variables are in the following *O* columns and the nominal variables are in the final $J - (C + O)$ columns. The latent continuous data underlying both the ordinal and

nominal data are assumed to be Gaussian, as are any observed continuous data. Thus, the joint vector of observed and latent continuous data is assumed to follow a multivariate Gaussian distribution $z_i \sim MVN_P(\mu, \Sigma)$. Since more than one latent dimension is required to model each nominal variable $P = C + O + \sum_{j=C+O+1}^{J} (K_j - 1)$. This model provides a unified way to simultaneously model continuous, ordinal and nominal data.

A Mixture Model for Mixed Data

The joint model for mixed data is embedded in a finite mixture model, facilitating the clustering of mixed data. This model, clustMD, is closely related to the parsimonious mixture of Gaussian distributions (Banfield and Raftery 1993; Celeux and Govaert 1995). In clustMD, it is assumed that z_i follows a mixture of G Gaussian distributions i.e., $z_i \sim \sum_{g=1}^{G} \pi_g MVN_p(\mu_g, \sum g)$ where π_g is the marginal probability of belonging to cluster g and μ_g and $\sum g$ denote the mean and covariance for cluster g respectively.

Decomposing the Covariance Matrix

Gaussian parsimonious mixture models utilise an eigenvalue decomposition of the cluster covariance matrix $\sum g = \lambda_g D_g A_g D_g$ where $|A_g| = 1$. The λ_g parameter controls the cluster volume, D_g is a matrix of eigenvectors of $\sum g$ that controls the orientation of the cluster and A_g is a diagonal matrix of eigenvalues of $\sum g$ that controls the shape of the cluster. The decomposed covariance is constrained in various ways to produce parsimonious models. The covariance matrix for the clustMD model is assumed to be diagonal, meaning that $D_g = I$, the identity matrix. This assumption implies that variables are conditionally independent given their cluster membership. Thus, under clustMD $\sum g = \lambda_g A_g$. These parameters can then be constrained to be different or equal across groups and A can also be constrained to be the identity matrix. This gives rise to a suite of 6 clustMD models with varying levels of parsimony. The 6 clustMD models and corresponding constraints are detailed in below.

Model	λ	A	D	# Covariance parameters	
				No nominal variables	Nominal variables
EII	C	I	I	1	1
VII	U	I	I	G	$2G-1$
EEI	C	C	I	$1+P$	$C+O$
VEI	U	C	I	$G+P$	$2G+C+O-2$
EVI	C	U	I	$1+GP$	$G(P-2)+C+O-P+2$
VVI	U	U	I	$G(1+P)$	$P(G-1)+O$

Parameters are unconstrained (U), constrained (C) to be equal across groups or equal to the identity (I)

Identifying clustMD in the Presence of Nominal Variables

If no nominal variables are present in the data, then the clustMD model is identified because the threshold parameters are fixed. However, in the presence of nominal variables, the model as it stands is not identified. Infinitely many combinations of the model parameters give rise to the same likelihood. Constraints must be placed on the parameters relating to nominal variables to obtain consistent parameter estimates. As in Cagnone and Viroli (2012), the constraint $\sum g\ \pi_g\mu_{gp} = 0$ for each dimension p corresponding to a nominal variable is applied across the suite of models, which amounts to insisting that $E(z_{ip}) = 0$ for p = C +O +1, . . ., P. Further, a separate volume parameter $\widetilde{\lambda_g}$ which applies only to the latent dimensions corresponding to nominal variables is also required. The diagonal elements of $\sum g$ corresponding to these dimensions are $\tilde{\lambda}_g\alpha_{gp}$, where α_{gp} is the p^{th} diagonal element of A_g.

Different constraints on $\widetilde{\lambda_g}$ are required in the different clustMDmodels. For example, the EII model is identified by fixing $\tilde{\lambda} = 1$, meaning that the diagonal elements of Σ corresponding to nominal variables are simply set to 1. The VII model is identified by insisting that $\sum g\ \widetilde{\lambda_g} = 1$. This may be accomplished by dividing each $\widetilde{\lambda_g}$ by $\sum g\ \widetilde{\lambda_g}$ after each iteration of the model fitting algorithm. To identify the EEI model $\tilde{\lambda}$ is set to 1, as is a_p for p corresponding to nominal variables. The VEI model is constrained so that $\sum g\ \widetilde{\lambda_g} = 1$ and $a_p = 1$ for nominal dimensions p. Thus, the nominal portions of the EEI and VEI models are the same as the EII and VII models respectively.

The EVI model is identified by fixing $\tilde{\lambda} = 1$ and constraining agp so that $\sum g\ a_{gp} = 1$ for nominal dimensions p. This constraint on a_{gp} is implemented by dividing each a_{gp} term by $\sum g\ a_{gp}$ after each iteration of the model fitting algorithm. Finally, the VVI model is identified by constraining $\widetilde{\lambda_g}$ and a_{gp} so that $\sum g\ a_{gp} = 1$ for each nominal dimension p. It is possible to fit all 6 clustMD models, even in the presence of nominal data. However, there are only 4 models for the nominal portion of the clustMD model.

Fitting the model

The clustMD model is fitted using an EM algorithm. If nominal data are present, then a Monte Carlo approximation is required for the expectation step and hence the algorithm is a Monte Carlo EM (MCEM) algorithm.

Deriving the Complete Data Log Likelihood

The categorical part of each observation can be thought of as one of a (possibly large) number, M, of response patterns. Let y_i^{β} be a binary vector of length M indicating which response pattern is observed, i.e., if response pattern m is observed, write $y_{im} = 1$; all other entries are 0. Thus, $y_i^{\beta} \sim$ Multinomial (1, q) where $q = (q_1, \ldots, q_M)$ and $q_m = \int_{\Omega_m} f(z_i)dz_i$. The portion of R^{P-C} that generates pattern m is denoted Ω_m. Let z_i^{β} denote the latent continuous vector corresponding to the observed categorical variables and the superscript β denote the portions of the model parameters corresponding to these data. A binary latent variable, l_i is introduced that indicates the cluster membership of observation i, i.e., $l_{ig} = 1$ if observation i belongs to cluster g; all other entries are 0. Thus, the joint density of y_i^{β} and z_i^{β} can be written as

$$f(\mathbf{z}_i^{\beta}, \mathbf{y}_i^{\beta}, \boldsymbol{\ell}_i) = f(\mathbf{z}_i^{\beta} | \mathbf{y}_i^{\beta}, \ell_{ig} = 1) f(\mathbf{y}_i^{\beta} | \ell_{ig} = 1) f(\boldsymbol{\ell}_i)$$

where $l_i \sim$ Multinomial (1, π), where $\pi = (\pi 1, \ldots, \pi g)$, $y_i^{\beta} | l_{ig} = 1 \sim$ Multinomial (1, q_g) where $q_g = (q_{g1}, \ldots, q_{gM})$ and $q_{gM} = \int_{\Omega_m} f\left(z_i^{\beta} \middle| l_{ig} = 1\right) dz_i^{\beta} = \int_{\Omega_m} MVN(z_i^{\beta} | \mu_g^{\beta}, \sum_g^{\beta}) dz_i^{\beta}, z_i^{\beta} | y_i^{\beta}, l_{ig} = 1 \sim MVN^T(z_i^{\beta} | \mu_g^{\beta}, \sum_g^{\beta})$, a truncated multivariate Gaussian

distribution. The points of truncation are those which satisfy the ordinal and/or nominal conditions given y_i^{β}. Thus,

$$f(\mathbf{z}_i^{\beta}, \mathbf{y}_i^{\beta}, \boldsymbol{\ell}_i) \propto \left\{ \prod_{g=1}^{g} \left[\frac{\text{MVN}(\mathbf{z}_i^{\beta} | \boldsymbol{\mu}_g^{\beta}, \Sigma_g^{\beta})}{\prod_{m=1}^{M} q_{gm}^{y_{im}^{\beta}}} \right]^{\ell_{ig}} \right\} \left\{ \prod_{g=1}^{g} \prod_{m=1}^{M} \left[q_{gm}^{y_{im}^{\beta}} \right]^{\ell_{ig}} \right\} \left\{ \prod_{g=1}^{g} \pi_g^{\ell_{ig}} \right\}$$
$$= \prod_{g=1}^{g} \left[\pi_g \text{MVN}(\mathbf{z}_i^{\beta} | \boldsymbol{\mu}_g^{\beta}, \Sigma_g^{\beta}) \right]^{\ell_{ig}}.$$

Let $y_i^{a} = z_i^{\alpha}$ denote the observed continuous variables and the superscript α denote the portions of the model parameters that apply to continuous variables. Since $\sum g$ is assumed to be diagonal, the complete data likelihood is the product of the likelihood of the continuous variables and the likelihood of the latent variables relating to the observed categorical variables:

$$\mathcal{L}_c = \prod_{i=1}^{N} \prod_{g=1}^{g} \left[\pi_g \text{MVN}(\mathbf{z}_i^{\alpha} | \boldsymbol{\mu}_g^{\alpha}, \Sigma_g^{\alpha}) \times \text{MVN}(\mathbf{z}_i^{\beta} | \boldsymbol{\mu}_g^{\beta}, \Sigma_g^{\beta}) \right]^{\ell_{ig}} \quad (1)$$

$$\Rightarrow \log \mathcal{L}_c = \sum_{i=1}^{N} \sum_{g=1}^{g} \left[\ell_{ig} \log \pi_g + B - \frac{\ell_{ig}}{2} \log |\Sigma_g| - \frac{\ell_{ig}}{2} \left(\mathbf{z}_i^{\alpha T} \Sigma_g^{\alpha -1} \mathbf{z}_i^{\alpha} + \mathbf{z}_i^{\beta T} \Sigma_g^{\beta -1} \mathbf{z}_i^{\beta} \right) \right.$$
$$\left. + \ell_{ig} \left(\boldsymbol{\mu}^{\alpha T} \Sigma_g^{\alpha -1} \mathbf{z}_i^{\alpha} + \boldsymbol{\mu}^{\beta T} \Sigma_g^{\beta -1} \mathbf{z}_i^{\beta} \right) - \frac{\ell_{ig}}{2} \left(\boldsymbol{\mu}^{\alpha T} \Sigma_g^{\alpha -1} \boldsymbol{\mu}^{\alpha} + \boldsymbol{\mu}^{\beta T} \Sigma_g^{\beta -1} \boldsymbol{\mu}^{\beta} \right) \right]$$

where B denotes a constant.

The Expectation Step

The expectation step (E-step) of the EM algorithm consists of computing the expectation of the complete log likelihood with respect to the latent data z_i^{β} and the latent cluster labels l_{ig}. The below three expectations are therefore required:

$$\mathbb{E}(\ell_{ig} | \mathbf{y}_i, \boldsymbol{\mu}_g, \Sigma_g, \pi_g)$$

$$\mathbb{E}(\ell_{ig} \mathbf{z}_i^{\beta} | \mathbf{y}_i, \boldsymbol{\mu}_g^{\beta}, \Sigma_g^{\beta}, \pi_g)$$

$$\mathbb{E}(\ell_{ig}\mathbf{z}_i^{\beta^T}\mathbf{z}_i^{\bar{\beta}}|\mathbf{y}_i, \boldsymbol{\mu}_g^{\beta}, \Sigma_g^{\beta}, \pi_g)$$

For the first expectation, since l_{ig} takes the values 0 or 1, then:

$$\mathbb{E}(\ell_{ig}|\ldots) = \frac{\pi_g \text{MVN}(\mathbf{z}_i^{\alpha}|\boldsymbol{\mu}_g^{\alpha}, \Sigma_g^{\alpha}) \int_{\Omega_m} \text{MVN}(\mathbf{z}_i^{\beta}|\boldsymbol{\mu}_g^{\beta}, \Sigma_g^{\beta}) d\mathbf{z}_i^{\beta}}{\sum_{g'=1}^{G} \pi_{g'} \text{MVN}(\mathbf{z}_i^{\alpha}|\boldsymbol{\mu}_{g'}^{\alpha}, \Sigma_{g'}^{\alpha}) \int_{\Omega_m} \text{MVN}(\mathbf{z}_i^{\beta}|\boldsymbol{\mu}_{g'}^{\beta}, \Sigma_{g'}^{\beta}) d\mathbf{z}_i^{\beta}} = \tau_{ig} \quad (2)$$

Since the covariance matrix $\sum_g^{\beta}$ *is* assumed to be diagonal, the integrals in (2) can be expressed as a product of probabilities. The probabilities corresponding to ordinal variables are easily approximated given the threshold parameters. However, in the presence of nominal variables, calculating the probabilities is more challenging, due to the way in which the latent data generate a nominal response. Thus, for each cluster, a Monte Carlo approximation of the probability of each possible response is obtained by simulating a large number of continuous vectors from a multivariate Gaussian distribution with mean μ_g^j and covariance $\sum_g^j$ *are* the portions of the mean vector and covariance matrix for cluster g, corresponding to nominal variable j. The probability of each response is approximated by the proportion of these simulations that generate each response. The Monte Carlo approximations can then be used to estimate τ_{ig} above. Like Karlis and Santourian (2009), the second expectation is

$$\begin{aligned}&\mathbb{E}(\ell_{ig}\mathbf{z}_i^{\beta}|\mathbf{y}_i^{\beta}, \boldsymbol{\mu}_g^{\beta}, \Sigma_g^{\beta}, \pi_g)\\ &= \mathbb{P}(\ell_{ig} = 1|\boldsymbol{\mu}_g^{\beta}, \Sigma_g^{\beta}, \pi_g)\mathbb{E}(\mathbf{z}_i^{\beta}|\ell_{ig} = 1, \boldsymbol{\mu}_g^{\beta}, \Sigma_g^{\beta}, \pi_g) = \tau_{ig}\mathbf{m}_{ig}\end{aligned}$$

and the third expectation is

$$\begin{aligned}\mathbb{E}(\ell_{ig}\mathbf{z}_i^{\beta^T}\mathbf{z}_i^{\beta}|\boldsymbol{\mu}_g^{\beta}, \Sigma_g^{\beta}, \pi_g) &= \mathbb{P}(\ell_{ig} = 1|\boldsymbol{\mu}_g^{\beta}, \Sigma_g^{\beta}, \pi_g)\mathbb{E}(\mathbf{z}_i^{\beta^T}\mathbf{z}_i^{\beta}|\ell_{ig} = 1, \boldsymbol{\mu}_g^{\beta}, \Sigma_g^{\beta}, \pi_g)\\ &= \tau_{ig}\sum_p \mathbb{E}(z_{ip}^2|\ell_{ig} = 1, \ldots) = \tau_{ig}\sum_p s_{igp}.\end{aligned}$$

The computation of m_{igp} and s_{igp} corresponding to ordinal variables is straightforward: given the relevant threshold parameters, they are simply the first and second moments of a truncated Gaussian distribution. In the case of dimensions relating to nominal variables, m_{igp} and s_{igp} are also related to the first and second moments of a truncated

multivariate Gaussian, but they are difficult to calculate given the outlined truncations outlined. A Monte Carlo approximation again is used in these cases. Suppose that y_{ij} = k for nominal variable j, then the below

$$\mathbb{E}(\mathbf{z}_i^j | y_{ij} = k, \ell_{ig} = 1, \boldsymbol{\mu}_g, \Sigma_g, \pi_g)$$

$$\mathbb{E}(\mathbf{z}_i^{j^T} \mathbf{z}_i^j | y_{ij} = k, \ell_{ig} = 1, \boldsymbol{\mu}_g, \Sigma_g, \pi_g)$$

must be calculated. The Monte Carlo samples generated to calculate the probabilities τ_{ig} for the first expectation can be reused to this end. For each possible response *k* and each cluster *g* the first moment can be approximated by calculating the sample mean of those Monte Carlo samples which generate response k. The second moment can be approximated by calculating the inner product of the vectors that generate response k and then calculating the sample mean of these inner products. The second expectation can then be approximated by summing the elements of this sample mean vector.

The Maximisation Step

The maximisation (M-step) of the algorithm maximises the expected value, *Q*, of the complete log likelihood based on the current values of the model parameters. The M-step in the case of the VVI model is derived below. The VVI model is the most general of the 6 clustMD models i.e., $\sum g = \lambda_g A_g$. The M-step maximises:

$$\begin{aligned}
Q = &\sum_g \log \pi_g \sum_i \tau_{ig} - \frac{C+O}{2} \sum_g \log \lambda_g \sum_i \tau_{ig} - \frac{P-C-O}{2} \sum_g \log \tilde{\lambda}_g \sum_i \tau_{ig} \\
&- \frac{1}{2} \sum_g \sum_p \log a_{gp} \sum_i \tau_{ig} - \frac{1}{2} \sum_g \sum_{p=1}^{C} \sum_i \frac{z_{ip}^2 \tau_{ig}}{\lambda_g a_{gp}} - \frac{1}{2} \sum_g \sum_{p=C+1}^{C+O} \sum_i \frac{s_{igp} \tau_{ig}}{\lambda_g a_{gp}} \\
&- \frac{1}{2} \sum_g \sum_{p=C+O+1}^{P} \sum_i \frac{s_{igp} \tau_{ig}}{\tilde{\lambda}_g a_{gp}} + \sum_g \sum_{p=1}^{C+O} \sum_i \frac{\mu_{gp} z_{igp}^* \tau_{ig}}{\lambda_g a_{gp}} \\
&+ \sum_g \sum_{p=C+O+1}^{P} \sum_i \frac{\mu_{gp} z_{igp}^* \tau_{ig}}{\tilde{\lambda}_g a_{gp}} - \frac{1}{2} \sum_g \sum_{p=1}^{C+O} \sum_i \frac{\mu_{gp}^2 \tau_{ig}}{\lambda_g a_{gp}} \\
&- \frac{1}{2} \sum_g \sum_{p=C+O+1}^{P} \sum_i \frac{\mu_{gp}^2 \tau_{ig}}{\tilde{\lambda}_g a_{gp}} + R
\end{aligned}$$

where R denotes a constant and $z_{ig}^{*}= (z_i^{\alpha}, m_{ig})^{T}$. Maximising Q with respect to λ_g yields

$$\hat{\lambda}_g = \frac{\sum_{p=1}^{C} \sum_i \frac{z_{ip}^2 \tau_{ig}}{a_{gp}} + \sum_{p=C+1}^{C+O} \sum_i \frac{s_{igp} \tau_{ig}}{a_{gp}} - \sum_{p=1}^{C+O} \frac{\mu_{gp}}{a_{gp}} \left[2 \sum_i z_{igp}^{*} \tau_{ig} - \mu_{gp} \sum_i \tau_{ig} \right]}{(C+O) \sum_i \tau_{ig}}$$

and, if nominal variables are present, maximising Q with respect to $\widetilde{\lambda g}$ yields

$$\hat{\tilde{\lambda}}_g = \frac{\sum_{p=C+O+1}^{P} \sum_i \frac{s_{igp} \tau_{ig}}{a_{gp}} - 2 \sum_{p=C+O+1}^{P} \frac{\mu_{gp}}{a_{gp}} \sum_i z_{igp}^{*} \tau_{ig} + \sum_{p=C+O+1}^{P} \frac{\mu_{gp}^2}{a_{gp}} \sum_i \tau_{ig}}{(P-C-O) \sum_i \tau_{ig}}.$$

Maximising Q with respect to a_{gp} yields

$$\hat{a}_{gp} = \frac{\zeta_{gp} - 2\mu_{gp} \sum_i z_{igp}^{*} \tau_{ig} + \mu_{gp}^2 \sum_i \tau_{ig}}{\lambda_g \xi_g \sum_i \tau_{ig}}$$

where $\lambda_g = \tilde{\lambda}_g$ if $p > (C+O)$, $\zeta_{gp} = \sum_i z_{ip}^2 \tau_{ig}$ if $p \leq C$ and $\zeta_{gp} = \sum_i s_{igp} \tau_{ig}$ if $p > C$, $\xi_g = (\prod_{p=1}^{C+O} a_{gp})^{\frac{1}{C+O}}$ if $p \leq C+O$ and $\xi_g = 1$ if $p > C+O$.

The (Monte Carlo) E and M steps are iterated until convergence is reached. Convergence is guaranteed even though a Monte Carlo approximation is used. However, the monotone increase in the likelihood at each iteration, which a standard EM algorithm guarantees, does not apply here. The example of Wei and Tanner (1990) is followed, and the algorithm is terminated when a plot of the parameter estimates against the iteration number show that the process has stabilised. For more detail on convergence and the Monte Carlo EM algorithm see McLachlan and Krishnan (2008). The algorithm is initialised by obtaining an initial clustering and estimating model parameters based on that clustering. To avoid local minima a number of different initialisations are used; namely K means, hierarchical and random clustering. The sensitivity of the EM algorithm to initialising values is a known problem. Recent work

on this issue includes that of O'Hagan et al. (2012). However, for the data sets analysed in this paper, the (MC)EM algorithm has not displayed particular sensitivity.

Model Selection

The best fitting covariance structure and number of components is selected using an approximation of the Bayesian information criterion (BIC) (Schwarz 1978; Kass and Raftery 1995). The BIC cannot be evaluated for clustMD models since the observed likelihood relies on the calculation of intractable integrals. However, the observed likelihood may be estimated as follows. The observed data vector $y_i = (y_i^{\alpha}, y_i^{\beta})$ where $y_i^{\alpha} = z_i^{\alpha} \sim \sum_{g=1}^{G} \pi_g \text{MVN}(\mu_g^{\alpha}, \sum_g^{\alpha})$ and $y_i^{\beta} \sim$ Multinomial (1, q). Treating these random variables as independent, the joint density can be approximated as the product of their marginals:

$$f(\mathbf{y}_i) \approx \left[\sum_{g=1}^{G} \pi_g \text{MVN}(\mathbf{z}_i^{\alpha} | \boldsymbol{\mu}_g^{\alpha}, \Sigma_g^{\alpha})\right]\left[\prod_{m=1}^{M} q_m^{y_{im}^{\beta}}\right] \qquad (3)$$

The first term in (3) is easily evaluated but the second term requires the probability of the observed categorical response pattern for observation *i*. i.e.,

$$\begin{aligned} q_m &= \int_{\Omega_m} \sum_g \pi_g \text{MVN}(\mathbf{z}_i^{\beta} | \boldsymbol{\mu}_g^{\beta}, \Sigma_g^{\beta}) d\mathbf{z}_i^{\beta} = \sum_g \pi_g \int_{\Omega_m} \text{MVN}(\mathbf{z}_i^{\beta} | \boldsymbol{\mu}_g^{\beta}, \Sigma_g^{\beta}) d\mathbf{z}_i^{\beta} \\ &= \sum_g \pi_g \left[\prod_{j=C+1}^{O} \int_{\Omega_{mj}} N(z_{ij} | \mu_{gj}, \sigma_{gj}^2) dz_{ij}\right]\left[\prod_{j=C+O+1}^{J} \int_{\Omega_{mj}} MVN(\mathbf{z}_i^j | \boldsymbol{\mu}_{gj}, \Sigma_{gj}) d\mathbf{z}_i^j\right] \end{aligned} \qquad (4)$$

The products in (4) consist of probabilities which were estimated in order to calculate τ_{ig} during the model fitting process. The products in the first term in (4) are easily obtained from a normal distribution while the probabilities in the second are obtained by the Monte Carlo approximation. It should be noted that q_m need only be estimated for the observed response patterns and not all M possible response patterns. Thus, the observed likelihood is approximated by:

$$\hat{\mathcal{L}} = \prod_{i=1}^{N} \left[\sum_{g=1}^{g} \pi_g \text{MVN}(\mathbf{z}_i^{\alpha} | \mu_g, \Sigma_g) \right] \left[\prod_{m=1}^{M} \hat{q}_m^{\gamma_{im}^{\beta}} \right]$$

The approximated BIC is then $\widehat{BIC} = 2\hat{L} - \nu \log(N)$ where ν is the number of free parameters in the model.

Latent Class Model

Detailed material for the Latent Class Model method can be found in the following bibliography:

- Matthieu, M. and Mohammed, S. (2017). Variable selection for model-based clustering using the integrated complete-data likelihood. Statistics and Computing, Volume 27(Issue 4), pp.1049–1063.
- Marbac, M., Mohammed, S. and Patin, T. (2019). Variable Selection for Mixed Data Clustering: Application in Human Population Genomics. Journal of Classification, Volume 37(Issue 2).

The Model

Data to analyse consist of N observations $x = (x_1, ..., x_N)$, where each observation $x_i = (x_{i1}, ..., x_{id})$ is defined by space $X_1 \times ... \times X_d$, X_j depending on the nature of variable j. Hence, $X_j = R$ (respectively N, $\{1, ..., m_j\}$) if variable j is continuous (respectively integer and categorical with m_j levels). Observations are assumed to arise independently from the mixture of g components defined by its probability distribution function (pdf):

$$f(\boldsymbol{x}_i | g, \boldsymbol{\theta}) = \sum_{k=1}^{g} \tau_k f_k(\boldsymbol{x}_i | \boldsymbol{\alpha}_k) \text{ with } f_k(\boldsymbol{x}_i | \boldsymbol{\alpha}_k) = \prod_{j=1}^{d} f_{kj}(x_{ij} | \boldsymbol{\alpha}_{kj}), \quad (1)$$

where, $\theta = \{\tau_k, \alpha_k; k=1, ..., g\}$ groups the model parameters, τ_k is the proportion of component k such that $0 < \tau_k \leq 1$ and $\sum_k^g = {}_1\tau_k = 1$, f_k is the pdf of component k parametrized by $\alpha_k = (\alpha_{k1}, ..., \alpha_{kd})$, and f_{kj} is the pdf of variable j for component k parametrized by α_{kj}. The univariate marginal distribution of variable j depends on its

definition space; therefore, f_{kj} is the pdf of a Gaussian distribution N (μ_{kj}, σ_{kj}^2) (Poisson P (α_{kj}) and multinomial M (α_{kj1}, ..., α_{kjmj})) if variable *j* is continuous (respectively integer and categorical) with $\alpha_{kj} = (\mu_{kj}, \sigma_{kj})$ (respectively $\boldsymbol{\alpha}_{kj} = \alpha_{kj}$ and $\boldsymbol{\alpha}_{kj} = (\alpha_{kj1}, ..., \alpha_{kjmj})$).

In clustering, a variable is said to be irrelevant if its univariate margins are invariant over the mixture constituents. Considering the model defined by equation (1), variable *j* is irrelevant if $\alpha_{1j} = ... = \alpha_{gj}$, and it is relevant otherwise. The role of the variables is defined by the binary vector $\boldsymbol{\omega} = (\omega_1, ..., \omega_g)$, since $\omega_j = 0$ if variable *j* is irrelevant and $\omega_j = 1$ otherwise. Consequently, the couple $\mathbf{m} = (g, \boldsymbol{\omega})$ defines the model at hand, because it defines the parameter space. Therefore, for a model **m**, the pdf of $\mathbf{x_i}$ is

$$f(\boldsymbol{x}_i|\boldsymbol{m},\boldsymbol{\theta}) = \prod_{j\in\boldsymbol{\Omega}^c} f_{1j}(x_{ij}|\boldsymbol{\alpha}_{1j}) \sum_{k=1}^{g} \tau_k \prod_{j\in\boldsymbol{\Omega}} f_{kj}(x_{ij}|\boldsymbol{\alpha}_{kj}), \tag{2}$$

where $\boldsymbol{\Omega} = \{j: \omega_j = 1\}$ and $\boldsymbol{\Omega}^C = \{1, ..., d\} \setminus \boldsymbol{\Omega}$.

Maximum Likelihood Inference

The general form of the observed-data log-likelihood of model **m** is defined by $\ell(\theta|m, x) = \sum_{i}^{N} = {}_1 \ln (\sum_{k}^{g} = {}_1 \tau_k \prod_{j}^{d} = {}_1 f_{kj} (x_{ij} |\alpha_{kj}))$. Therefore, equalities between the parameters defined by $\boldsymbol{\omega}$ imply that:

$$\ell(\boldsymbol{\theta}|\boldsymbol{m},\mathbf{x}) = \left(\sum_{j\in\boldsymbol{\Omega}^c}\sum_{i=1}^{N} \ln f_{1j}(x_{ij}|\boldsymbol{\alpha}_{1j})\right) + \left(\sum_{i=1}^{N} \ln\left(\sum_{k=1}^{g} \tau_k \prod_{j\in\boldsymbol{\Omega}} f_{kj}(x_{ij}|\boldsymbol{\alpha}_{kj})\right)\right) \tag{3}$$

The MLE of the parameters corresponding to the irrelevant variables are explicit, but not those of the proportions and the relevant variables. Thus, it is standard to use an expectation– maximization (EM) algorithm to maximize the observed-data log-likelihood. Here, the partition among the observations is unobserved. This partition is denoted by $\mathbf{z} = (z_1, ..., z_N)$ with $z_i = (z_{i1}, ..., z_{ig})$, where $z_{ik} = 1$ if observation *i* arises

from component k and $z_{ik} = 0$ otherwise. Accordingly, the complete-data likelihood of model **m** (log-likelihood computed on the observed and unobserved variables) is defined by:

$$\ell(\boldsymbol{\theta}|\boldsymbol{m}, \mathbf{x}, \mathbf{z}) = \sum_{j\in\Omega^c}\sum_{i=1}^{N} \ln f_{1j}(x_{ij}|\boldsymbol{\alpha}_{1j}) + \sum_{k=1}^{g}\sum_{i=1}^{N} z_{ik} \ln \tau_k + \sum_{j\in\Omega}\sum_{k=1}^{g}\sum_{i=1}^{N} z_{ik} \ln f_{kj}(x_{ij}|\boldsymbol{\alpha}_{kj}) \tag{4}$$

The EM algorithm alternates between two steps: The Expectation step (E-step) consisting of computing the expectation of the complete-data likelihood under the current parameters, and the maximization step (M-step) consisting of maximizing this expectation over the model parameters. Therefore, this algorithm starts from the initial value of the model parameter $\boldsymbol{\theta}^{[0]}$ randomly sampled and its iteration [r] is defined by:

E-step: Computation of the fuzzy partition $t_{ik}^{[r]} := \mathbb{E}[Z_{ik}|\mathbf{x}_i, \mathbf{m}, \boldsymbol{\theta}^{[r-1]}]$; hence,

$$t_{ik}^{[r]} := \frac{\tau_k^{[r-1]} \prod_{j=1}^{d} f_{kj}(x_{ij}|\boldsymbol{\alpha}_{kj}^{[r-1]})}{\sum_{\ell=1}^{g} \tau_\ell^{[r-1]} \prod_{j=1}^{d} f_{\ell j}(x_{ij}|\boldsymbol{\alpha}_{\ell j}^{[r-1]})},$$

M-step: Maximization of the expected value of the complete-data log-likelihood over the parameters:

$$\tau_k^{[r]} = \frac{n_k^{[r]}}{N} \text{ and } \boldsymbol{\alpha}_{kj}^{[r]} = \begin{cases} \boldsymbol{\alpha}_{jk}^{*[r]} & \text{if } \omega_j = 1 \\ \tilde{\boldsymbol{\alpha}}_{1j} & \text{otherwise} \end{cases},$$

where, $N_k^{[r]} = \sum_{i}^{N} =_1 t_{ik}^{[r]}$, $\tilde{a}_{1j} = \text{argmax}_{\boldsymbol{\alpha}1j} \sum_{i}^{N} =_1 \ln f_{1j}(x_{ij}|\boldsymbol{\alpha}_{1j})$ is the MLE for an irrelevant variable, and $\boldsymbol{\alpha}_{jk}^{*[r]} = \arg\max_{\boldsymbol{\alpha}kj} \sum_{i}^{N} =_1 t_{ik}^{[r]} \ln f_{kj}(x_{ij}|\boldsymbol{\alpha}_{kj})$ is the estimate for a relevant variable. This algorithm converges to a local optimum of the observed-data log-likelihood. Accordingly, the MLE for the model **m**, denoted by $\hat{\theta}_m$, is obtained by performing many different random initializations of $\boldsymbol{\theta}^{[0]}$.

Model Selection by Optimizing the BIC

Information Criterion for Data Modeling

Model selection generally aims at finding the model $\hat{\mathbf{m}}$ which obtains the greatest posterior probability, among a collection of competing models $\mathbb{M}$. The number of components of the competing models is usually bounded by a value g_{max}. Thus,

$$\mathcal{M} = \left\{ m = (g, \omega) : \; g \in \{1, \ldots, g_{\max}\} \text{ and } \omega \in \{0, 1\}^d \right\} \tag{5}$$

By assuming uniformity for the prior distribution of $\mathbf{m}$, $\hat{\mathbf{m}}$ maximizes the integrated likelihood defined by:

$$\hat{m} - \arg\max_{m \in \mathcal{M}} p(\mathbf{x}|m) \text{ with } p(\mathbf{x}|m) - \int_{\Theta_m} p(\mathbf{x}|m, \theta) p(\theta|m) d\theta, \tag{6}$$

where $\mathbf{\Theta_M}$ is the parameter space of model $\mathbf{m}$, $p(\mathbf{x}|\mathbf{m}, \boldsymbol{\theta}) = \prod_i^N = {}_1 \text{ f } (x_i|\mathbf{m}, \boldsymbol{\theta})$ is the likelihood function, and $p(\boldsymbol{\theta}|\mathbf{m})$ is the pdf of the prior distribution of the parameters. Regrettably, the integrated likelihood is intractable, nevertheless, many methods allow for approximations of its value. The most popular approach comprises using the BIC, which approximates the logarithm of the integrated likelihood by Laplace approximation, and consequently requires MLE. The BIC is defined by:

$$\text{BIC}(m) = \ln p(\mathbf{x}|m, \hat{\theta}_m) - \frac{\nu_m}{2} \ln(N), \tag{7}$$

where, ν_m is the number of independent parameters required by $\mathbf{m}$.

Optimizing the Penalized Likelihood

For a fixed number of components g, selecting the variables necessitates the comparison of 2^d models. Therefore, an exhaustive approach approximating the integrated likelihood for each competing model is not deemed feasible. Instead, Raftery and Dean perform model selection by deterministic algorithms like a stepwise method. Yet, this approach cannot safeguard the acquisition of the model maximizing the BIC. Furthermore, it can be computationally expensive if many variables are

observed. The component assumption permits the direct maximization of any penalized log-likelihood function defined by

$$\ell_{\text{pen}}(\boldsymbol{\theta}|\boldsymbol{m}, \mathbf{x}) = \ell(\boldsymbol{\theta}|\boldsymbol{m}, \mathbf{x}) - \nu_{\boldsymbol{m}} c, \tag{8}$$

for any constant c. This function is maximized by means of a modified version of the EM algorithm. Thereupon, the penalized complete-data log-likelihood function is introduced:

$$\ell_{\text{pen}}(\boldsymbol{\theta}|\boldsymbol{m}, \mathbf{x}, \mathbf{z}) = \ell(\boldsymbol{\theta}|\boldsymbol{m}, \mathbf{x}, \mathbf{z}) - (g-1)c - c\sum_{j=1}^{d} \nu_j (g\omega_j + 1 - \omega_j), \tag{9}$$

where, ν_j is the number of parameters for one univariate marginal distribution of variable j (i.e., $\nu_j = 2$ if the variable is continuous, $\nu_j = 1$ if the variable is integer, and $\nu_j = m_j - 1$ if the variable is categorical with m_j levels). This modified version of the EM algorithm finds the model maximizing the penalized log-likelihood for a fixed number of components. It starts at an initial point ($\mathbf{m}^{[0]}$, $\boldsymbol{\theta}^{[0]}$) randomly sampled with $\mathbf{m}^{[0]} = (g, \boldsymbol{\omega}^{[0]})$, and its iteration [r] is constituted of two steps:

E-step: Computation of the fuzzy partition

$$t_{ik}^{[r]} := \frac{\tau_k^{[r-1]} \prod_{j=1}^{d} f_{kj}(x_{ij}|\boldsymbol{\alpha}_{kj}^{[r-1]})}{\sum_{\ell=1}^{g} \tau_\ell^{[r-1]} \prod_{j=1}^{d} f_{\ell j}(x_{ij}|\boldsymbol{\alpha}_{\ell j}^{[r-1]})},$$

M-step: Maximization of the expectation of the penalized complete-data log-likelihood over ($\boldsymbol{\omega}$, $\boldsymbol{\theta}$), hence $\mathbf{m}^{[r]} = (g, \omega^{[r]})$ with

$$\omega_j^{[r]} = \begin{cases} 1 \text{ if } \Delta_j^{[r]} > 0 \\ 0 \text{ otherwise} \end{cases}, \quad \tau_k^{[r]} = \frac{n_k^{[r]}}{n} \text{ and } \boldsymbol{\alpha}_{jk}^{[r]} = \begin{cases} \boldsymbol{\alpha}_{kj}^{\star[r]} \text{ if } \omega_j^{[r]} = 1 \\ \tilde{\boldsymbol{\alpha}}_{kj} \text{ otherwise} \end{cases},$$

where $\Delta_j = \sum_k^g = \sum_i^N = 1\ t_{ik}^{[r]}$ (ln f_{kj} (x_{ij} $|a_{kj}^{*[r]}$) − ln f_{1j} (x_{ij} $|\tilde{a}_{1j}$) − (g − 1)ν_jc is the difference between the maximum of the expected value of the penalized complete-data loglikelihood obtained when variable *j* is relevant and when it is not. To obtain the couple (**ω**, **θ**) maximizing the penalized observed-data log-likelihood for a fixed number of components, many random initializations of this algorithm should be done. So, the couple (**m**, **θ**) maximizing the penalized observed-data log-likelihood is acquired through performing this algorithm for every value of *g* between 1 and g_{max}. By considering c = $\frac{1}{2}$ ln(N), this algorithm carries out the model selection according to the BIC. Moreover, other criteria like the AIC by setting *c* = 1 can also be considered.

Model Selection by Optimizing the MICL

Information Criterion

Although the BIC has good properties of consistency, it does not focus on the clustering goal. Moreover, it involves an approximation in $\mathbb{O}(1)$ which can deteriorate its performances, especially when *N* is small or when $\mathbb{M}$ is large. Criteria based on the complete-data likelihood like the ICL or the MICL have been introduced. The integrated complete-data likelihood is defined by:

$$p(\mathbf{x}, \mathbf{z}|\boldsymbol{m}) = \int_{\boldsymbol{\Theta_m}} p(\mathbf{x}, \mathbf{z}|\boldsymbol{m}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{m}) d\boldsymbol{\theta}. \tag{10}$$

where, p(**x**, **z**|**m**, ***θ***) = $\prod_N^i =1\ \prod_k^g =1[\tau_k f_k(\mathbf{x}_i|\boldsymbol{\alpha}_k)]^{z_{ik}}$ is the complete-data likelihood. When conjugate prior distributions are used, the integrated complete-data likelihood has the following closed form. Thus, independence is assumed between the prior distributions, such that:

$$p(\boldsymbol{\theta}|\boldsymbol{m}) = p(\boldsymbol{\tau}|\boldsymbol{m}) \prod_{j=1}^{d} p(\boldsymbol{\alpha}_{\bullet j}|g, \omega_j) \text{ with } p(\boldsymbol{\alpha}_{\bullet j}|g, \omega_j)$$
$$= \begin{cases} \prod_{k=1}^{g} p(\boldsymbol{\alpha}_{kj}) & \text{if } \omega_j = 1 \\ p(\boldsymbol{\alpha}_{1j}) \prod_{k=1}^{g} \mathbb{1}_{\{\boldsymbol{\alpha}_{kj}=\boldsymbol{\alpha}_{1j}\}} & \text{if } \omega_j = 0, \end{cases}, \tag{11}$$

where, $\alpha_{\bullet j} = (\alpha_{1j}, \ldots, \alpha_{gj})$. Conjugate prior distributions are used, thus $\boldsymbol{\tau}|\mathbf{m}$ follows a Dirichlet distribution $D_g(u, \ldots, u)$. If variable *j* is continuous, $p(\boldsymbol{\alpha}_{kj}) = p(\sigma_{kj}^2)\, p(\mu_{kj}|\sigma_{kj}^2)$ where σ_{kj}^2 follows an Inverse-Gamma distribution $\mathbb{IG}(a_j / 2, b_j^2 / 2)$ and $\mu_{kj}|\mathbf{m}, \sigma_{kj}^2$ follows a Gaussian distribution $N(c_j, \sigma_{kj}^2 / d_j)$. If variable *j* is an integer, then α_{kj} follows a Gamma distribution $\mathbb{G}a(a_j, b_j)$ while $\boldsymbol{\alpha}_{kj}$ follows a Dirichlet distribution $\mathbb{D}_{mj}(a_j, \ldots, a_j)$ in case variable *j* is categorical with m_j levels. If there is no information on the parameters *a priori*, the Jeffreys non-informative prior distributions for the proportions (i.e., $u_k = 1/2$) are used for the hyperparameters of a categorical variable (i.e., $a_{jk} = 1/2$) as well. Such prior distributions do not exist for the parameters of the Gaussian and Poisson distributions, so flat prior distributions are used. The conjugate prior distribution implies the following closed form of the integrated complete-data likelihood:

$$p(\mathbf{x}, \mathbf{z}|\boldsymbol{m}) = \frac{\Gamma\left(\frac{g}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^g} \frac{\prod_{k=1}^{g} \Gamma\left(n_k + \frac{1}{2}\right)}{\Gamma\left(n + \frac{g}{2}\right)} \prod_{j=1}^{d} p(\mathbf{x}_{\bullet j}|g, \omega_j, \mathbf{z}), \tag{12}$$

where $x_{\bullet j} = (x_{ij}; i = 1, \ldots, N)$, $n_k = \sum_i^N = 1 z_{ik}$ and

$$p(\mathbf{x}_{\bullet j}|g, \omega_j, \mathbf{z}) = \int p(\boldsymbol{\alpha}_{\bullet j}|g, \omega_j) \prod_{k=1}^{g} \prod_{i=1}^{N} f_{kj}(x_{ij}|\boldsymbol{\alpha}_{kj})^{z_{ik}} d\boldsymbol{\alpha}_{\bullet j}. \tag{13}$$

The integral defined by equation (13) is explicit, thus providing a closed form of the integrated complete-data likelihood. The MICL corresponds to the greatest value of the integrated complete-data likelihood among all the possible partitions. Thus, the MICL is defined by:

$$\text{MICL}(\boldsymbol{m}) = \ln p(\mathbf{x}, \mathbf{z}_{\boldsymbol{m}}^{\star}|\boldsymbol{m}) \text{ with } \mathbf{z}_{\boldsymbol{m}}^{\star} = \arg\max_{\mathbf{z}} \ln p(\mathbf{x}, \mathbf{z}|\boldsymbol{m}). \quad (14)$$

This criterion is similar to the ICL and inherits its main properties. More specifically, it is less sensitive to model misspecification than the BIC. Unlike the ICL and the BIC, it does not require the MLE and thus avoids the multiple calls to the EM algorithm. Because $\boldsymbol{\omega}$ does not impact the dimension of $\mathbf{z}$, we can maximize the integrated complete data likelihood over $(\boldsymbol{\omega}, \mathbf{z})$, and thus the best model according the MICL can be obtained for a fixed number of components.

Optimizing the MICL

An iterative algorithm is used for finding the model maximizing the MICL for a fixed number of components. Starting at the initial point $(\mathbf{z}^{[0]}, \mathbf{m}^{[0]})$ with $\mathbf{m}^{[0]} = (g, \boldsymbol{\omega}^{[0]})$, the algorithm alternates between two optimizations of the integrated complete-data likelihood: optimization on z given $(\mathbf{x}, \mathbf{m})$, and maximization on $\boldsymbol{\omega}$ given $(\mathbf{x}, \mathbf{z})$. The algorithm is initialized as follows: $\omega_j^{[0]} = 1$ with probability 0.5 then $\mathbf{z}^{[0]} = \hat{\mathbf{z}}_{\mathbf{m}[0]}$ is the partition provided by a MAP rule associated to model $\mathbf{m}^{[0]}$ and to its MLE $\widehat{\boldsymbol{\theta}}_{\mathbf{m}[0]}$. Iteration [r] of the algorithm is written as:

Partition step: Find $\mathbf{z}^{[r]}$ such that

$$\ln p(\mathbf{x}, \mathbf{z}^{[r]}|\boldsymbol{m}^{[r]}) \geq \ln p(\mathbf{x}, \mathbf{z}^{[r-1]}|\boldsymbol{m}^{[r]}).$$

Model step: Find $\mathbf{m}^{[r+1]} = \arg\max_{\mathbf{m}\in Mg} \ln p(\mathbf{x}, \mathbf{z}^{[r]}|\mathbf{m})$ such that

$$\begin{aligned} \boldsymbol{m}^{[r+1]} &= (g, \boldsymbol{\omega}^{[r+1]}) \text{ with } \omega_j^{[r+1]} \\ &= \begin{cases} 1 \text{ if } p(\mathbf{x}_{\bullet j}|g, \omega_j = 1, \mathbf{z}^{[r]}) > p(\mathbf{x}_{\bullet j}|g, \omega_j = 0, \mathbf{z}^{[r]}) \\ 0 \text{ otherwise} \end{cases}. \end{aligned}$$

At iteration [r], the model step consists of finding the vector $\mathbf{m}^{[r+1]}$ maximizing the integrated completed-data likelihood for the current partition $\mathbf{z}^{[r]}$. This optimization can be performed independently for each element ω_j, due to the within-component independence assumption. The partition step is more complex; therefore, $\mathbf{z}^{[r]}$ is defined as a partition increasing the value of the integrated complete-data likelihood for the current model. It is obtained by an iterative method initialized at the partition $\mathbf{z}^{[r-1]}$. Each iteration consists of sampling an individual uniformly, which is affiliated to the component maximizing the integrated complete-data likelihood while the other component memberships are unchanged. Like the EM algorithm, the proposed algorithm converges to a local optimum of ln $p(\mathbf{x}, \mathbf{z}|\mathbf{m})$, so many different initializations should be done.

C. Other clustering algorithms for mixed data

According to the extensive taxonomy for mixed data clustering proposed by A. Ahmad and S. S. Khan, there are five major research themes of clustering methods for heterogeneous data - *partitional*, *hierarchical*, *model based*, *neural network-based*, and *other*. The "other" category encompasses several minor groups of clustering algorithms that either do not fit into the other major research themes or have not been extensively studied. In the current sub-section, a subset of clustering algorithms that belong to some of these categories but are out of scope in the thesis, is presented.

According to the bibliography, the algorithms below are available in the category of hierarchical clustering methods, as shown in Table 3. As already mentioned, hierarchical clustering methods create a hierarchy of clusters organized in a top to bottom (or bottom to top) order. To create clusters, the hierarchical algorithms need both of the following: (i) Similarity matrix - this is constructed by finding the similarity between each pair of mixed data points, (ii) Linkage criterion - this determines the distance between sets of observations.

Algorithm	**Clustering Algorithm**
Philip and Ottaway Chiu *et al.* Li and Biswas	Agglomerative hierarchical clustering method BIRCH algorithm

Hsu *et al.*	Agglomerative hierarchical clustering with group-average method
Hsu and Chen	Agglomerative hierarchical clustering
Hsu and Huang	Incremental clustering
Shih *et al.*	Adaptive resonance theory network
Lim *et al.*	Agglomerative hierarchical clustering algorithm
Chae *et al.*	Agglomerative hierarchical clustering method
	Agglomerative hierarchical clustering method

Table 3 *Some hierarchical clustering algorithms for mixed datasets*

Philip and Ottaway used Gower's similarity measure to compute the similarity matrix for mixed datasets. Gower's similarity measure computes the similarity by dividing features into two subsets - one for categorical features and the other for numeric features. Hamming distance is applied to compute the similarity between two data points for a categorical feature. A weighted average of similarities for all categorical features is the similarity between two data points in a categorical feature space. For numeric features, the similarity is computed so that the one between the same feature values is 1, whereas if the difference between the values is the maximum possible difference (the difference between maximum and minimum values of the feature), the similarity is 0. The sum of the similarity values for all numeric features is the one of two data points in a numeric feature space. The similarity in the categorical feature space and the numeric feature sone are added to compute the similarity between two data points. Hierarchical agglomerative clustering is then used to create clusters.

Fang et al. developed a similarity measure to compute the similarity between two clusters for mixed data. This similarity measure is related to the decrease in the log-likelihood function when two clusters are merged. Zhang et al. combined the BIRCH clustering algorithm, which uses hierarchical clustering algorithm with their proposed similarity measure to develop a clustering algorithm that can handle mixed datasets. Li and Biswas proposed a similarity-based agglomerative clustering (SBAC)

algorithm for mixed data clustering. SBAC used the Goodall similarity measure and applied a hierarchical agglomerative approach to build cluster hierarchies. Hsu et al. proposed a distance measure based on a concept hierarchy consisting of concept nodes and links. More general concepts are represented by higher level nodes whereas more specific concepts are represented by lower-level nodes. The categorical values are represented by a tree structure in a way that each leaf is represented by a categorical value. Each feature of a data point is associated with a distance hierarchy. The distances between two data points are calculated by using their associated distance hierarchies. An agglomerative hierarchical clustering algorithm is applied to a distance matrix to obtain the clusters.

Hsu and Chen proposed a new similarity measure to cluster mixed data. The algorithm uses variance for computing the similarity of numeric values. For similarity between categorical values, they utilized entropy with distance hierarchies. The similarities are then aggregated to compute the similarity matrix for a mixed dataset. Incremental clustering is used on the similarity matrix to obtain the clusters. In an extended work, Hsu and Huang applied an adaptive resonance theory network (ART) to cluster data points by using the distance hierarchies as the input of the network. A better interpretation of clusters is possible with the proposed algorithm as compared to the K-prototypes algorithm. Shih et al. converted categorical features of a mixed dataset into numeric features by using frequencies of co-occurrence of categorical feature values. The dataset with all numeric features is then clustered by using a hierarchical agglomerative clustering algorithm. Lim et al. partitioned the data into two sections: categorical and numeric data. The two types are clustered separately. The clustering results are combined by using a weighted scheme to obtain a similarity matrix. The agglomerative hierarchical clustering method is applied on the similarity matrix to obtain the final clusters. Gower's similarity measure assigns equal weights to both types of features in computing the similarity between two data points. The similarity matrices may be dominated by one feature type. Chae and Yang assigned weights to the different feature types to overcome this problem. Improved clustering results are shown with these weighted similarity matrices.

In the category of model-based clustering algorithms, a great number of algorithms is also included, as shown in Table 4. As already mentioned, model-based clustering methods assume that a data point matches a model, in many cases, a statistical distribution. The models are generally user-defined, so they are prone to

yielding undesirable clustering outcomes if inappropriate models (or their parameters) are chosen.

Algorithm	Model
Cheeseman and Stutz	Bayesian methods
Everitt	Model-based clustering with the use of thresholds for the categorical features
Lawrence and Krzanowski	Extension of homogeneous conditional Gaussian model to the finite mixture situation
Moustaki and Papageorgiou	Latent class mixture model
Browne and McNicholas	A mixture of latent variables with the expectation-maximization framework
Andreopoulos *et al*.	Pseudo-Bayesian process with categorical data clustering to guide clustering of numeric data
Hunt and Jorgensen	A finite mixture of multivariate distributions is fitted to data
McParland *et al*.	Bayesian finite mixture model
Saadaoui *et al*.	A projection of the categorical features on the subspaces spanned by numeric features and then the application of the Gaussian Mixture Model

Table 4 *Some model-based clustering algorithms for mixed datasets*

AUTOCLASS performed clustering by integrating finite mixture distribution and Bayesian methods with prior distribution of each feature. AUTOCLASS can cluster data containing both categorical and numeric features. Everitt proposed a clustering algorithm by using model-based clustering for datasets consisting of both numeric and binary or ordinal features. The normal model is extended to handle mixed datasets by using thresholds for the categorical features. Because of high computational cost, the method is only useful for datasets containing very few categorical features. To overcome this problem, Lawrence and Krzanowski extended

the homogeneous Conditional Gaussian model to the finite mixture case, to compute maximum likelihood estimates for the parameters in a sample population. They suggest that their method works for an arbitrary number of features. Moustaki and Papageorgiou used a latent class mixture model for mixed data clustering. Categorical features are converted to binary ones by a 1-in-q representation. A multinomial distribution is used for categorical features and a normal distribution for a numeric feature. Features are considered independent in each cluster. Browne and McNicholas proposed a mixture of latent features model for clustering and the expectation-maximization (EM) framework is used for model fitting.

Andreopoulos et al. presented a clustering algorithm - Bi-level clustering of mixed categorical and numeric data types (BILCOM) for mixed datasets. The algorithm uses categorical data clustering to guide the clustering of numeric data. Hunt and Jorgensen proposed a mixture model clustering approach for mixed data. In this approach, a finite mixture of multivariate distributions is fitted to data and then the membership of each data point is calculated by computing the conditional probabilities of cluster membership. A local independence assumption can be used to reduce the model parameters. They further show that the method can also be applied for clustering mixed datasets with missing values. McParland et al. proposed a clustering algorithm for high-dimensional mixed data by using a Bayesian finite mixture model. In this algorithm, the estimation is done by using the Gibbs sampling algorithm. To select the optimal model, they also proposed an approximate Bayesian Information Criterion-Markov chain Monte Carlo criterion. They showed that the method works well on a mixed medical dataset consisting of high-dimensional numeric phenotypic features and categorical genotypic features. Saâdaoui et al. proposed a projection of the categorical features on the subspaces spanned by numeric features while an optimal Gaussian mixture model is obtained from the resulting principal component analysis regressed subspaces.

III. Dataset

A. Context Background

This thesis focuses on the analysis of data for prostate cancer to determine areas such as whether a personalized treatment based on the individual characteristics of each patient could be of significant benefit for their health improvement. Before thoroughly analysing the dataset utilized to perform this kind of exploration, let us firstly acquire the domain knowledge and comprehend the scientific framework within which this research moves.

Prostate cancer is a form of cancer marked by an uncontrolled (malignant) growth of cells in the prostate gland. The prostate is a walnut-sized gland located behind the base of the penis, in front of the rectum, and below the bladder. It surrounds the urethra, the tube-like channel that carries urine and semen through the penis. The prostate's main function is to make seminal fluid, the liquid in semen that protects, supports, and helps transport sperm. Prostate cancer can be so slow-growing that a lot of people die of other diseases before the prostate cancer causes significant problems. Some prostate cancer cases may not even cause symptoms or problems for years or ever. Even when prostate cancer has spread to other parts of the body, it can often be under control for a long time. So, people with prostate cancer, and even those at an advanced stage, may live in good health and enjoy quality of life for many years to come. On the other hand, other prostate cancer cases could be more aggressive and can spread outside the confines of the prostate gland, which can be deadly. The prostate cancer survival rate is greatly improved with early detection and personalized treatment.

Based on the statistics, prostate cancer is the second most commonly occurring cancer in men and the fourth most commonly occurring cancer overall. Age-adjusted incidence rates of prostate cancer have increased dramatically, and this is largely because of the increased availability of screening for prostate-specific antigen (PSA) in men without symptoms of the disease. Prostate-specific antigen (PSA or serum prostatic acid phosphatase) is a protein produced by cells in the prostate gland and released into the bloodstream. Although there is no specific indication for a "normal PSA" for anyone at any given age, a higher-than-normal level of PSA can be found in

people with prostate cancer. Therefore, such blood test may lead to the detection of many prostate cancers that are small and/or would otherwise remain unrecognised, and which may or may not develop further into higher stage disease.

Some people with several known risk factors may never develop cancer, while others with no known risk factors do. The following indicative factors may raise a person's risk of developing prostate cancer:

- Age - The risk of prostate cancer increases with age, especially after the age of 50. Around 60% of prostate cancers are diagnosed in people who are 65 or older. Older adults who are diagnosed with prostate cancer can face unique challenges, specifically regarding cancer treatment.
- Race - Black men in the United States and other men of African ancestry are diagnosed with prostate cancer more than men of other races. Black men are more likely to die from prostate cancer than white men.
- North American or northern European location - Prostate cancer occurs most often in North America and northern Europe. It also appears that prostate cancer is increasing among Asian people living in urbanized environments, such as Hong Kong, Singapore, and North American as well as European cities, particularly among those with a lifestyle of less physical activity and a less healthy diet.
- Family history - Prostate cancer that runs in a family-called familial prostate cancer- makes up of about 20% of all prostate cancers. This type of prostate cancer develops because of a combination of shared genes and shared environmental or lifestyle factors. Hereditary prostate cancer, which is inheriting the risk from a relative, is rare and accounts for about 5% of all cases. Hereditary prostate cancer occurs when changes in genes, or mutations, are passed down within a family from one generation to the next. Hereditary prostate cancer may be suspected if a family history includes any of the following characteristics:

    - 3 or more first-degree relatives with prostate cancer
    - Prostate cancer in 3 generations on the same side of the family

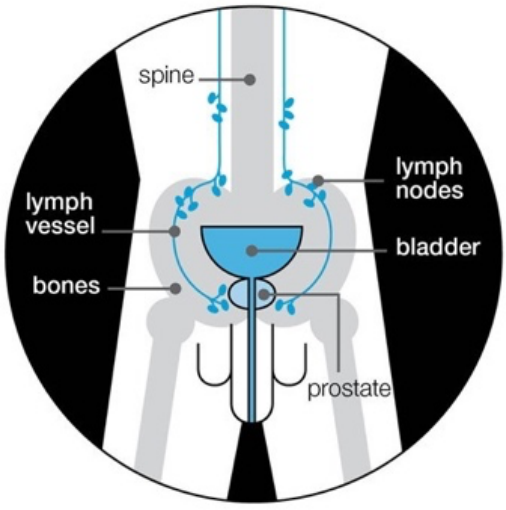
- 2 or more close relatives, such as a parent, sibling, child, grandparent, uncle, or nephew, on the same side of the family diagnosed with prostate cancer before the age of 55

- Hereditary breast and ovarian cancer (HBOC) syndrome – HBOC is associated with germline, or inherited, DNA-repair mutations to the *BRCA1* and/or *BRCA2* genes. BRCA stands for "BReast Cancer". HBOC is most associated with an increased risk of breast and ovarian cancers in women. However, people with HBOC also have an increased risk of developing breast cancer and a more aggressive form of prostate cancer, as well. Moreover, mutations in the *BRCA1* and *BRCA2* genes are thought to cause only a small percentage of inherited prostate cancers. Those who have *BRCA1* or *BRCA2* mutations should consider screening for prostate cancer at an earlier age. Genetic testing may only be appropriate for families with prostate cancer that may also have HBOC.
- Other genetic changes – Other genes that may carry an increased risk of developing prostate cancer include *HPCI*, *HPC2*, *HPCX*, *CAPB*, *ATM*, *FANCA*, *HOXB13*, and mismatch repair genes. However, none of them has been directly shown to cause prostate cancer or be specific to this disease. Research to identify genes associated with an increased risk of prostate cancer is ongoing, and researchers are constantly learning more about how specific genetic changes can influence the development of prostate cancer.
- Eating habits - No study has proven that diet and nutrition can directly cause or prevent the development of prostate cancer. However, many studies that look at links between certain eating behaviours and cancer suggest there may be a connection. For example, obesity is associated with many cancers, including prostate cancer.

There are several types of treatments that are the standardised types of care for prostate cancer including active surveillance and watchful waiting, local treatments and systemic treatments. *Active surveillance and watchful waiting* are usually preferred for patients with very low and low risk prostate cancer and for older adults besides those with other serious or life-threatening illnesses expected to live less than

5 years. *Local treatments* (e.g., surgery, radiation therapy, focal therapy, etc.) intend to remove cancer from a specific, limited area of the body. If the cancer has spread outside the prostate gland, the *systemic treatments* (e.g., hormonal therapy, targeted therapy, chemotherapy, immunotherapy, radiopharmaceuticals, bone-modifying drugs, etc.) may be needed to destroy cancer cells located in other parts of the body. In any case, regardless of which treatment may be chosen for the specific patient, several key parameters regarding the current state of the cancer should also be taken into consideration-parameters such as whether symptoms exist or PSA levels are rising rapidly or cancer has spread to the bones, the health history, the quality of life, the current urinary and sexual function and any other medical conditions the patient may have.

However, there are times where the cancer may be spread from the prostate to other parts of the body, most commonly to the bones and lymph nodes, and in that case the prostate cancer is described as metastatic, secondary, secondaries, metastases or mets. As depicted below, lymph nodes (sometimes called lymph glands) are part of the lymphatic system, which is part of the body's immune system. Lymph nodes are found throughout the body including in the pelvic area, near the prostate. This kind of spread may be developed when prostate cancer cells move through the blood stream or lymphatic system. Unfortunately, if a patient is diagnosed with advanced prostate cancer, although there are some treatment options like chemotherapy with hormone therapy, hormone therapy alone and clinical trials, these may not cure the cancer but only keep it under control and manage any symptoms.



B. Dataset Overview

The dataset used (provided in the *clustMD* R package) contains information for a group of 475 prostate cancer patients. These patients have either stage 3 or stage 4 prostate cancer. The specific data consist of variables of mixed type (continuous, binary, ordinal or nominal). Most of them are continuous followed by binary, nominal and then ordinal variables. The nine continuous variables consist of age, weight, systolic blood pressure, diastolic blood pressure, serum haemoglobin, size of primary tumour, index of tumour stage and histolic grade, serum prostatic acid phosphatase and patient observation ID. The categorical variables consist of 3 binary ones: cardiovascular disease history, bone metastases and cancer stage**,** two nominal ones: electrocardiogram code and patient post trial survival status, and one ordinal one: patient performance rating. It is also of importance to emphasise that different measurement units are used for the continuous variables (e.g., g/100ml, centimetres squared, King-Armstrong units, etc.).

Variable Name	Description	Type
Age	A numeric vector indicating the age of the patient.	Continuous
Weight	A numeric vector indicating the weight of the patient.	Continuous
Performance.rating	An ordinal variable indicating how active the patient is: 0 - normal activity, 1 - in bed less than 50% of daytime, 2 - in bed more than 50% of daytime, 3 - confined to bed	Ordinal

Cardiovascular.disease.history	A binary variable indicating if the patient has a history of cardiovascular disease: 0 - no, 1 – yes	Binary
Systolic.Blood.pressure	A numeric vector indicating the systolic blood pressure of the patient in units of ten.	Continuous
Diastolic.blood.pressure	A numeric vector indicating the diastolic blood pressure of the patient in units of ten.	Continuous
Electrocardiogram.code	A nominal variable indicating the electrocardiogram code: 0 - normal, 1 - benign, 2 - rhythmic disturbances and electrolyte changes, 3 - heart blocks or conduction defects, 4 - heart strain, 5 - old myocardial infarct, 6 - recent myocardial infarct	Nominal
Serum.haemoglobin	A numeric vector indicating the serum haemoglobin levels of the patient measured in g/100ml.	Continuous

Size.of.primary.tumour	A numeric vector indicating the estimated size of the patient's primary tumour in centimetres squared.	Continuous
Index.of.tumour.stage.and.histolic.grade	A numeric vector indicating the combined index of tumour stage and histolic grade of the patient.	Continuous
Serum.prostatic.acid.phosphatase	A numeric vector indicating the serum prostatic acid phosphatase levels of the patient in King-Armstrong units.	Continuous
Bone.metastases	A binary vector indicating the presence of bone metastasis: 0 - no, 1 - yes	Binary
Stage	The stage of the patient's prostate cancer.	Binary
Observation	A patient ID number.	Continuous
SurvStat	The post trial survival status of the patient: 0 - alive, 1 - dead from prostatic cancer,	Nominal

	2 - dead from heart or vascular disease, 3 - dead from cerebrovascular accident, 4 - dead form pulmonary embolism, 5 - dead from other cancer, 6 - dead from respiratory disease, 7 - dead from other specific non-cancer cause, 8 - dead from other unspecified non-cancer cause, 9 - dead from unknown cause	

Table 5 *Dataset Variables*

C. Data Insights

In Table 6, the summary statistics for the continuous variables are presented. It is observed that the patients are middle-aged and elderly people whose weight may vary depending on their age. Moreover, it seems that for most quantitative characteristics, the range and the standard deviation are relatively low (e.g., the systolic blood pressure of the patients varies from 8 to 30 units with the standard deviation to be 2,43 units) with the latter metric to generally indicate that the data points tend to be close to the mean. Another interesting inference from the output is that the mean is not equal to the median for any of the variables which suggests an indication on the lack of symmetric and thus normal distribution; a conclusion that will be further investigated considered that this may affect the effectiveness of the discrimination mechanism for the clusters. Moreover, a variable that needs to be remarked on as it draws the statistical attention is this of the serum prostatic acid phosphatase, known to be highly related with the development of prostate cancer. It is the characteristic that, both in terms of comparison and absolutely, presents an

extremely high standard deviation equal to 638,48 King-Armstrong units implying a spread of the values over a large range of values, a wide range of 9.998 King-Armstrong units and a significant difference of 118,7 King-Armstrong units between the mean and the median, as well. The above findings verify a sharp distribution skew for the variable of serum prostatic acid phosphatase with this skewness to be quantified in Table 7 and visualized in Figures 1,2 and 3.

Variable Name	Min	1st Quart	Median	Mean	3rd Quart	Max	St. Dev.
Age	48,0	70,0	73,0	71,56	76,0	89,0	6,92
Weight	69,0	90,0	98,0	99,01	107,0	152,0	13,34
Systolic.Blood.pressure	8,0	13,0	14,0	14,38	16,0	30,0	2,43
Diastolic.blood.pressure	4,0	7,00	8,00	8,15	9,0	18,0	1,46
Serum.haemoglobin	59,0	122,5	137,0	134,2	147,0	182,0	19,38
Size.of.primary.tumour	0,0	5,0	10,0	14,29	21,0	69,0	12,23
Index.of.tumour.stage.and. histolic.grade	5,0	9,0	10,0	10,3	11,0	15,0	2,01
Serum.prostatic.acid. phosphatase	1.0	5,0	7,0	125,7	29,5	9.999,0	638,48

Table 6 *Continuous Variables - Summary Statistics*

The skewness, a significant metric of symmetry, may exist due to either the presence of extreme abnormal outliers, that may not be important to us, or due to the natural distribution itself that is skewed with the tail to be important to us. Considering that a highly skewed distribution, a moderately skewed distribution and an approximately symmetric distribution exists when the skewness value is less than −1 or greater than +1, is between −1 and −½ or between +½ and +1 and is between −½ and +½ respectively, significant conclusions may be drawn for the continuous variables of this dataset.

As shown in Table 7 in which different colors are used for the distinction of the various distributions (*red* color - highly skewed distribution, *yellow* color - moderately used distribution, *green* color - approximately normal distribution), most

of the characteristics such as weight, systolic blood pressure and diastolic blood pressure follow a moderately skewed distribution as the skewness calculated is within the boundaries of −½ and +½. On the contrary, other characteristics like these of age, size of primary tumour and serum prostatic acid phosphatase exceed -1 or +1 verifying the existence of a highly skewed distribution. Again, it is of interest that the serum prostatic acid phosphatase presents the highest skewness (that equals to 10,66) among all variables, a remark that is all but expected considering the non-normal statistical behavior identified and analyzed above.

.

Variable Name	Skewness Index
Age	- 1,068929
Weight	0,5418303
Systolic.Blood.pressure	0,9795097
Diastolic.blood.pressure	0,8027419
Serum.haemoglobin	- 0,5402281
Size.of.primary.tumour	1,418378
Index.of.tumour.stage.and.histolic.grade	0,2714751
Serum.prostatic.acid.phosphatase	**10,66378**

***Table** 7 Continuous Variables – Skewness Index*

Next, in Figure 1 the representation of the distribution of numerical data is thoroughly examined. Through this visualization, interesting insights are gained into the profile of the under-research prostate cancer patients. This profile is aligned with the factors of age, obesity and poor nutrition which may increase the probability of developing prostate cancer as it is revealed that the majority of patients

(a) are elderly people at the age of about 70 - 75,

(b) are at relatively high weight levels ranging between 90 - 100 kg,

(c) have marginally high pressure considering that the optimal blood pressure in healthy adults is below 120 for systolic and below 80 for diastolic pressure,

(c) have a rather small size of primary tumour up to approximately 5 square centimeters.

Moreover, through Figure 1 and Figure 2 that follows, the tendency for skewness is once again identified. Variables such as *age* and *serum haemoglobin*

present a negative skewness, which reveals that the mean of the values is less than the median, which in its turn means that the data distribution is left-skewed. On the other hand, the positive skewness of *weight*, *systolic* and *diastolic blood pressure*, *primary tumour size*, *tumour stage and histolic grade index* and *serum prostatic acid phosphatase* suggests that the mean of the data values is larger than the median, and the data distribution is right-skewed.

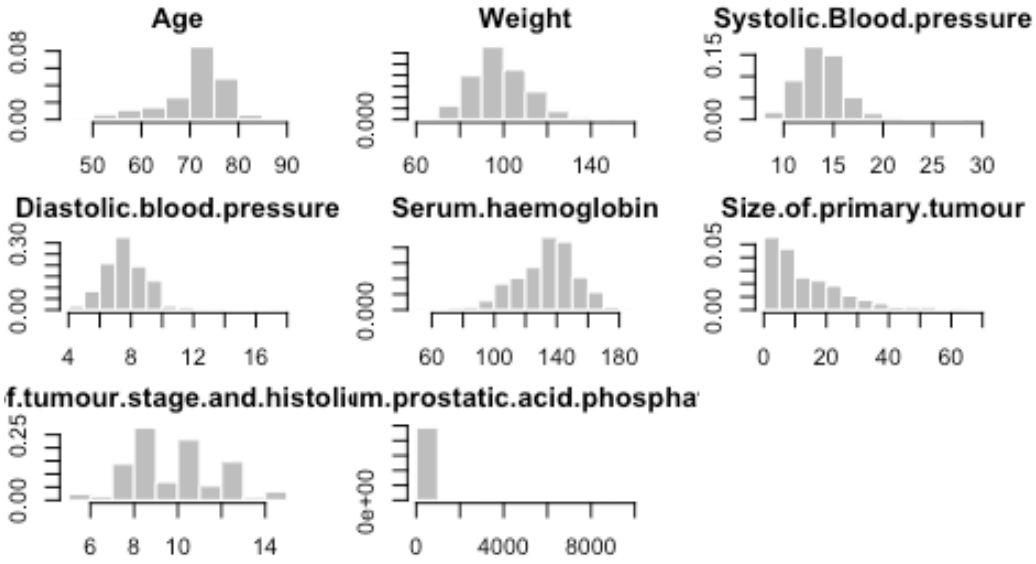


Figure 1 *Continuous Variables – Histograms*

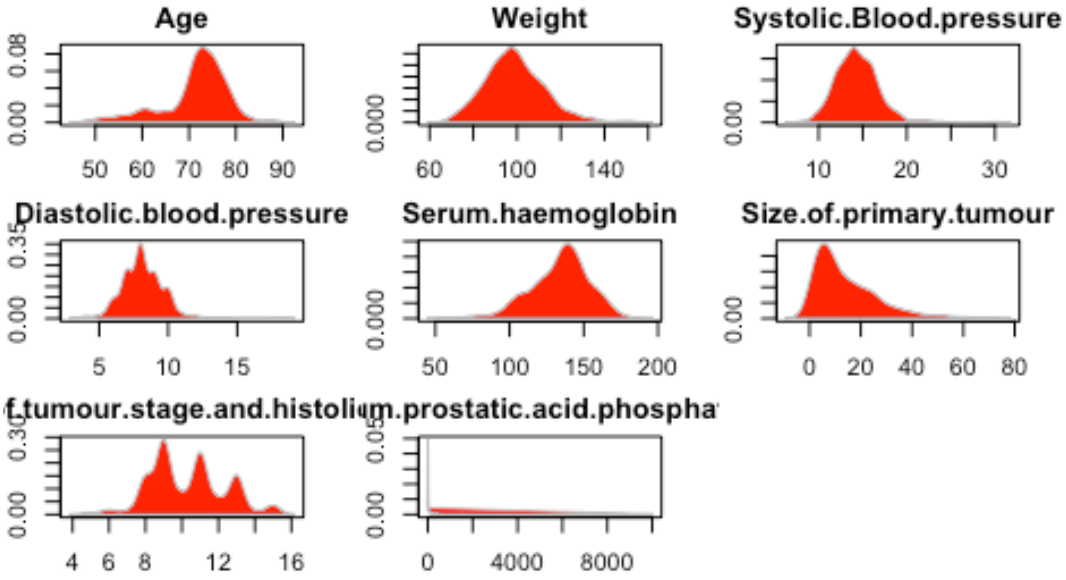


Figure 2 *Continuous Variables – Density Plots*

The QQ plots of Figure 3 below, which represent the pairing of two probability distributions, the one of the given sample and the other of the normal distribution are used to visually check the normality of the continuous data in the given dataset by plotting their quantiles against each other. It is known that, if the two distributions being compared are similar, the points in the Q-Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q-Q plot will approximately lie on a line, but not necessarily on the line y = x. If the distributions are not similar (demonstrating non-normality), then the points in the tails of the plot will deviate from the overall trend of the points.

The afore mentioned being taken into account, when examining each of the subplots in this Figure in detail, it is concluded that normality does not exist for any of these quantitative characteristics. Especially for the variables of *age*, *primary tumour size*, *tumour stage* and *histolic grade index* and *serum prostatic acid phosphatase*, the phenomenon of non-normality is more apparent as the points seem to form a curve that deviates markedly from the straight line of reference. In this Figure, a number of possible outliers is also observed since there are points at the end of each reference line which are distanced from the bulk of the remaining observations. These outliers are additionally calculated by using the method of percentiles (where any observation that lies outside the interval formed by the 0,01 and 99,0 percentiles is considered to be a potential outlier) and presented in Table 8.

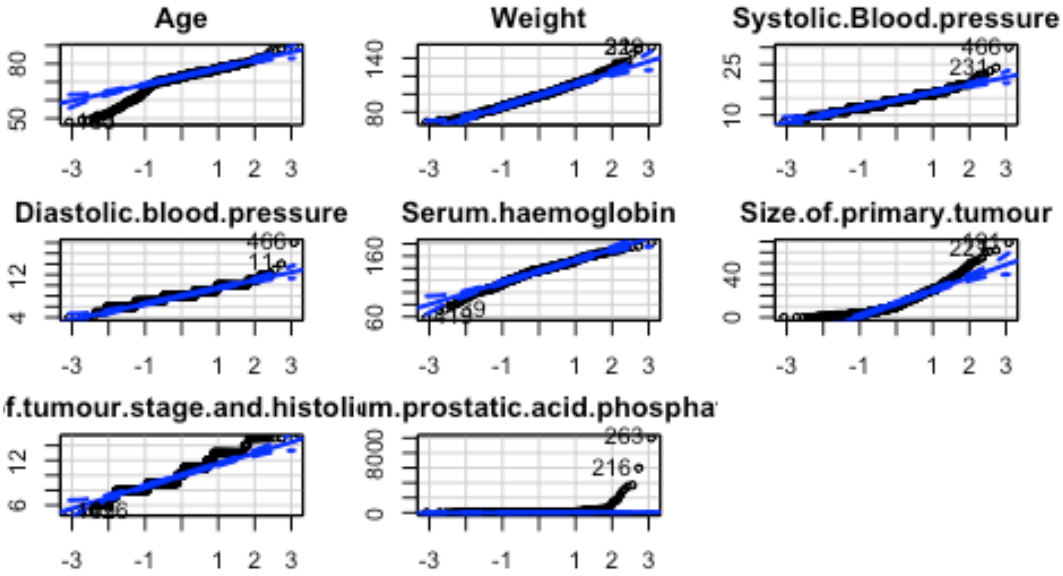


Figure 3 *Continuous Variables – QQ Plots*

When examining the results of Table 8, it is found that all continuous variables, except this of tumour stage and histolic grade index, appear to have outliers the count of which may vary from 3 to 6 in total. In general, this count represents an extremely small percentage (ranging from 0,63% to 1,26%) for the given dataset that consists of 475 observations. However, since outliers may skew the clusters substantially and thus significantly hinder the efficiency of the respective clustering algorithm, it should be further investigated whether the current ones are influential points. If that is the case, then it should be subsequently examined whether outlier detection methods are automatically used by the respective clustering algorithm to isolate outliers or manual removal of these outliers is required.

Variable Name	**Observation Sequence [Value]**	**Percentage (Outliers / Observations)**
Age	15 [87], 116 [85], 183 [48], 188 [87], 306 [89], 432 [88]	1,26%
Weight	31 [150], 140 [69], 229 [152], 257 [136], 299 [145]	1,05%
Systolic.Blood.pressure	64 [8], 178 [23], 231 [24], 466 [30]	0,84%
Diastolic.blood.pressure	11 [14], 72 [13], 466 [18]	0,63%
Serum.haemoglobin	3 [176], 94 [182], 319 [175], 324 [173], 332 [175], 419[59]	1,26%
Size.of.primary.tumour	107 [61], 191 [69], 217 [61], 221 [62], 387 [55]	1,05%
Index.of.tumour.stage.and.histolic.grade	-	-
Serum.prostatic.acid.phosphatase	30 [3.160], 137 [3.670], 216 [5.960],	1,05%

	243 [3.535], 263 [9.999]	

Table 8 *Continuous Variables - Outliers*

As previously highlighted, non-normality is recognized for the continuous data by using the visualization means of histogram and QQ plots and this is further quantified by applying the Shapiro-Wilk normality test. In this test, the p-value is compared to the selected significance level of 0,05 which indicates that the risk of concluding the data is 5% in case these data do not follow the specified distribution — when in actual fact the data do follow it.

- **P-value ≤ 0,05: The data do not follow the specified distribution (Reject H_0)**

If the p-value is less than or equal to the significance level of 0,05, the decision is to reject the null hypothesis and conclude that the data do not follow the specified distribution.

- **P-value > 0,05: Cannot conclude the data do not follow the specified distribution (Fail to reject H_0)**

If the p-value is larger than the significance level of 0,05, the decision is to fail to reject the null hypothesis because there is not enough evidence to conclude that the data do not follow the specified distribution. However, it cannot be concluded that the data do follow the specified distribution.

Null hypothesis	H_0: Data follow a normal distribution
Alternative hypothesis	H_1: Data do not follow a normal distribution

Table 9 *Shapiro-Wilk Normality Test – Hypotheses*

For each of the continuous variables, the p-value is calculated and presented in Table 10. It is concluded that since all the calculated p-values are less than the

significance level of 0,05, the decision is to reject the null hypothesis; the data do not follow a normal distribution.

Variable Name	P-value
Age	6,194e - 16
Weight	1,378e - 05
Systolic.Blood.pressure	1,384e - 12
Diastolic.blood.pressure	3,407e - 15
Serum.haemoglobin	6,971e - 06
Size.of.primary.tumour	< 2,2e - 16
Index.of.tumour.stage.and.histolic.grade	1,619e – 12
Serum.prostatic.acid.phosphatase	< 2,2e – 16

Table 10 *Continuous Variables - Shapiro-Wilk Normality Test*

Another aspect for the exploration of the continuous data deals with the phenomenon of collinearity. This aspect needs to be investigated since when the variables used in clustering are collinear, then they get a higher weight than others. If two variables are perfectly correlated, they effectively represent the same concept. But that concept is represented twice in the data and hence gets twice the weight of all the other variables. The final solution is likely to be skewed in the direction of that concept, which could be a problem if it is not anticipated. In the case of multiple variables and multicollinearity, the analysis is in effect being conducted on some unknown number of concepts that are a subset of the actual number of variables being used in the analysis.

In the following Figure 4, the correlation between each pair of variables in the given dataset is depicted. As it is visible, the size of the dot indicates the degree of correlation between the variables. It is observed that

- Diastolic and systolic blood pressure present a strong positive relationship with their correlation to be more than 0,5, an expected conclusion to be drawn,
- Weight and diastolic or systolic blood pressure are also positively correlated with this degree to be close to 0,2 and
- Size of primary tumour and index of tumour stage and holistic grade have the most negative association among all that is close to 0,5

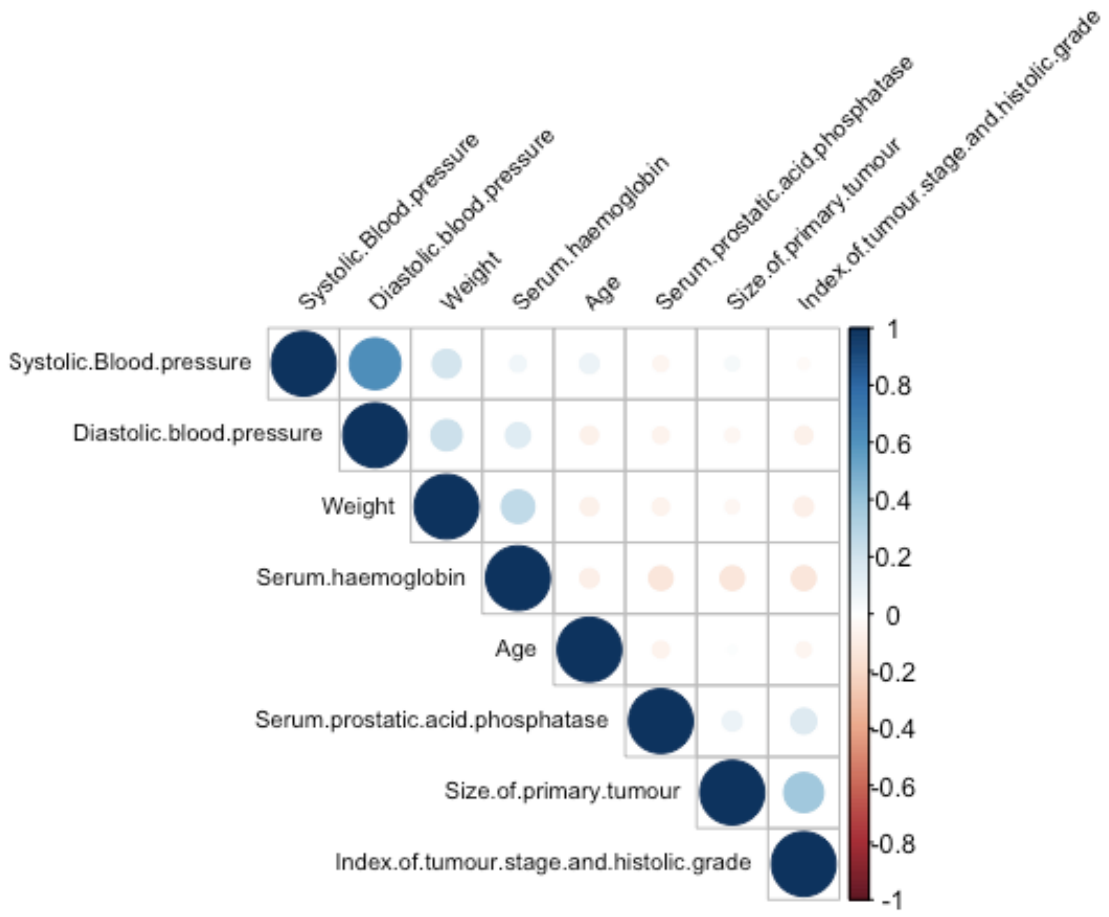


Figure 4 *Continuous Variables – Correlation*

Upon further analyzing data and going on to do so on the categorical variables, several remarkable conclusions are also drawn. Figure 9 below depicts the distribution of patients per survival status and performance rating. As shown, out of 475 people diagnosed with prostate cancer, a percentage of approximately 30% is still alive (*Level 0*), 25% have died due to cancer (*Level 1*) while 20% have died due to heart or vascular disease (*Level 3*). It is also of importance to notice that the overwhelming majority of the patients deceased due to heart/vascular disease (84 in total out of 93) or cancer (101 in total out of 121) had a normal activity (*Level 0*) (e.g., walking, and other aerobic exercise like outdoor work). This fact is of particular interest as it may imply that this type of cancer may not physically overwhelm the patients so much.

SurvStat \ Performance.rating	0	1	2	3	Totals
0	133	2	2		137
1	101	9	9	2	121
2	84	8	1		93
3	25	6			31
4	12	2			14
5	24				24
6	16				16
7	23	3	1		27
8	4	2			6
9	6				6
Totals	428	32	13	2	475

Table 11 *Number of Patients Per Survival Status & Performance Rating*

Figure 10 showcases the fact that the activity of the patients (e.g., normality, bed confinement, etc.) is barely affected by the cancer stage. More specifically, it is observed that irrespective of whether patients are at a stage 3 or 4, most of them continue to have a normal activity in their life (*Level 0*). Nevertheless, it is also noticeable that compared to stage 3 patients, at cancer stage 4, patients stay in bed more than 50% of the daytime (*Level 2*) while few of them are bedridden (*Level 3*), which verifies once again the afore mentioned conclusion regarding the physical impact of this potentially terminal disease on patients.

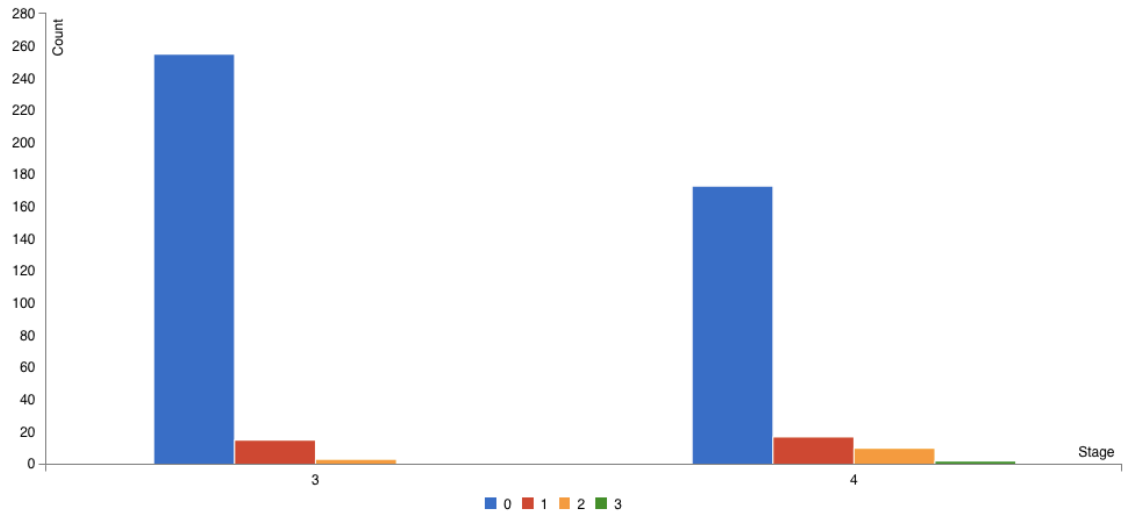


Figure 5 *Number of Patients Per Cancer Stage & Performance Rating*

It is a well-known fact worldwide that cardiovascular disease is the second most common cause of death in prostate cancer patients. This is clearly observed in Figure 6 as it is derived that the most important underlying cause of death in patients diagnosed with prostate cancer is either the cancer itself (*Level 1*) or the heart and blood vessels (*Level 2*). In the case of cancer stage 3, the majority of the patients who have died, have done so due to heart or vascular disease while in the case of cancer stage 4 this happens owing to prostate cancer. Based on these statistics, the other factors such as cerebrovascular accident (*Level 3*), pulmonary embolism (*Level 4*), respiratory disease (*Level 6*), etc. do not appear to significantly affect the mortality of prostate cancer patients. Of course, it is noteworthy that there is a - not negligible at all - survival rate in the prostate cancer patients considering that several of them at both stage 3 and 4 stay alive (around 140 in total representing a percentage of 30% of the patients).

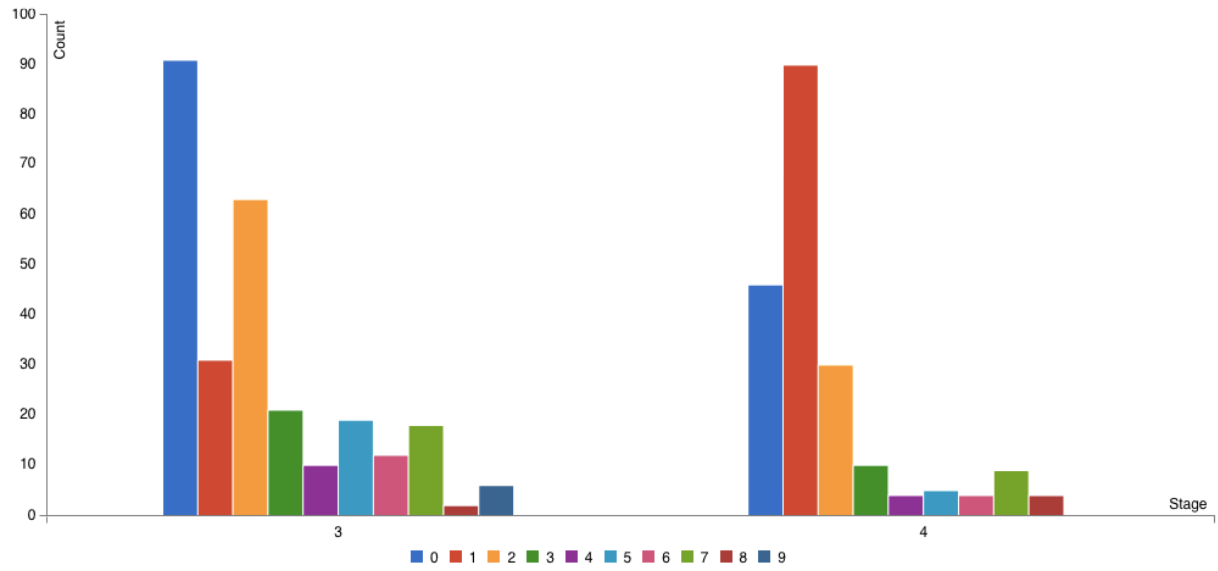


Figure 6 *Number of Patients Per Survival Status & Cancer Stage*

Going deeper into the cardiovascular disease of the under-study prostate cancer patients, it is observed that about half of them (207 out of 475) have heart problems (*Level 1*) with a history of heart strain (*Level 4*) to be present in the majority and a normal electrocardiogram (*Level 0*) or an old myocardial infarct (*Level 5*) to the rest of them. This group of people is likely to benefit from aggressive primary and secondary prevention therapies by improving the survival care in these people. In case the cardiovascular disease history is proved to affect the discrimination mechanism of

the clusters, then specific care processes could be proposed and applied to this patient population to improve patient specific cardiac and cancer related outcomes and identify at-risk individuals including standard cardiovascular risk factor assessment and modification or referral to a cardiology oncology clinic where available.

Cardiovascular.disease.history \ Electrocardiogram.code	0	1	2	3	4	5	6	Totals
0	113	15	28	15	72	25		268
1	48	8	22	10	73	45	1	207
Totals	161	23	50	25	145	70	1	475

Table 12 *Number of Patients Per Cardiovascular Disease History & Electrocardiogram Code*

As already analyzed, when prostate cancer spreads, the bones, such as the hip, spine, and pelvis ones, are typically the first area affected. It can be by a direct invasion or by traveling through the blood or lymphatic system. Bone metastases can weaken the patients' bones and lead to symptoms like bone pain, bladder and urinary troubles, soreness in the groin, leg swelling or unexplained weight loss. This kind of information is depicted in Figure 9 where it is observed that a number of about 76 patients with cancer stage 4 out of 202 in total have their cancer spread to the bones (*Level 1*). The factor of bone metastases could be proved determinant for the subsequent cluster analysis because this group of people is by default treated differently in terms of surveillance and treatment.

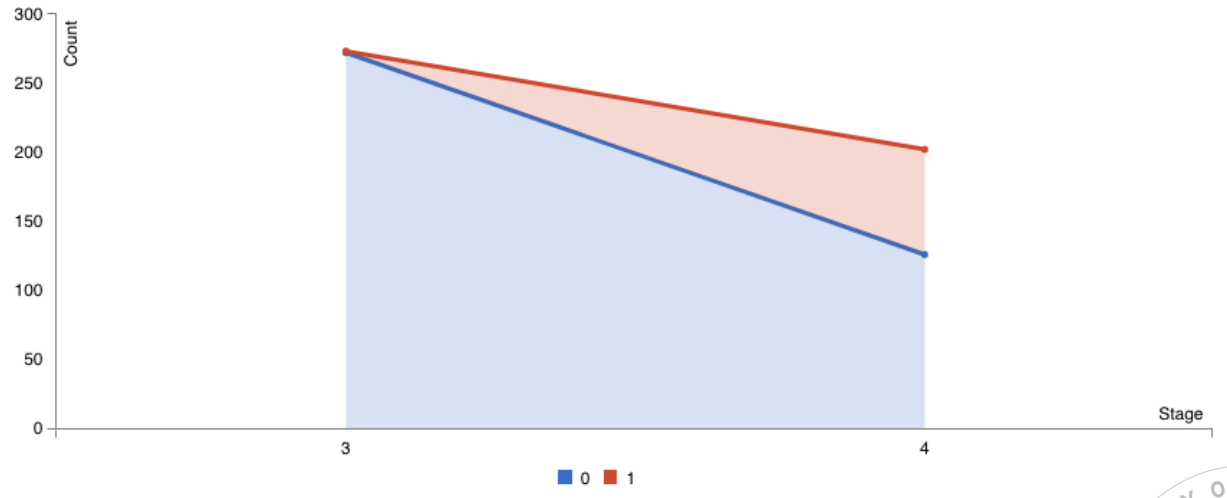


Figure 7 *Number of Patients with Bone Metastases Per Stage Cancer*

To summarize, it is of importance to draw special attention to the fact that the findings presented above undoubtedly shed light on the medical profile of the under-study patients for several key areas of their health status. The major challenges about to be approached when it comes to clustering of these patients diagnosed with prostate cancer are associated with the identification of their various groups, each one of which has different needs in terms of care, vigilance, and therapeutics. The primary goal of the current thesis is to determine the baseline variables which most efficiently discriminate the clusters of the patients and propose customized actionable items that best cater for the health management of these patients.

D. Data Manipulation & Transformation

As thoroughly analyzed in the previous subsection *Data Insights*, phenomena such as this of the asymmetric distribution, the usage of different units of measurement, and the outliers are present in the dataset. To reduce the negative effect of the first two phenomena in the subsequent clustering performed by the Kamila and K-Prototypes algorithms, the transformation steps below are applied:

- The initial understanding of the data has shown us that they contain attributes of different units. Units affect what clustering algorithms will discover. Due to this, the data standardization is used to make continuous variables in a common scale (i.e., -1 and 1) which practically means to have each attribute weighed properly. If non-normalized data existed, then this would lead to simply disregarding the attribute with the smaller range.
- The highly skewed continuous variable *Serum.prostatic.acid.phosphatase* is transformed by using the log. This is required to firstly create a smoother distribution, as the values of this variable are ranged over several orders of magnitude, and secondly to minimize as much as possible the time efficiency of the clustering algorithms which might hinder processing these wide data ranges.
- The variables *Performance.rating*, *Cardiovascular.disease.history*, *Electrocardiogram.code*, *Bone.metastases*, *Stage* and *SurvStat* are transformed to factors.

It is of importance to highlight that regarding the Latent Class Model, only the above steps of data standardization and log transformation are applied. Additionally, the categories of the nominal variable *Electrocardiogram.code* are reduced to 3 (instead of 7) by combining some categories (i.e., Levels 0,1 -> *Category 1*, Levels 2,3,4 -> *Category 2*, Levels 5,6,7 -> *Category 3*) since the Monte Carlo approximation used in Latent Class Model can be inefficient if there is a small number of observations in a particular category for a particular cluster. Moreover, regardless of the clustering algorithm used, additional steps are applied to drop the variables of *Age*, *Weight* and *Observation* which represents a random patient ID number of no statistical interest.

Finally, several potential problems of different nature for cluster analysis could emerge. These problems that could potentially arise are related to:

- The number of clusters to be defined: Identifying the number of clusters is a difficult task if the number of class labels is not known beforehand. A careful analysis of the number of clusters is necessary in order to produce correct results. Otherwise, it is found that heterogenous tuples may merge or similar types of tuples may be broken into many.
- The data structure: Real life data may not always contain clearly identifiable clusters. Also, the order in which the tuples are arranged may affect the results when an algorithm is executed if the distance measure used is not perfect. With structureless data, even identification of the appropriate number of clusters will not yield good results. For instance, if a record has all values missing, then this should be removed from the dataset. If an attribute has missing values in all tuples, then that attribute should also be removed.
- The identification of distance measure: For numerical attributes, distance measures able to be used are standard equations like Euclidean, Manhattan, etc. However, the identification of measure for categorical attributes is difficult and the respective clustering algorithms may not have the same efficiency in this area.
- The sensitivity of clustering algorithm to the outliers: Outliers may be of significant importance. In general, finding these outliers is highly non-trivial and removing them is not necessarily desirable.

IV. Clustering Application

In the current chapter, it is in-depth analyzed how the three different clustering algorithms (Kamila, K-Prototypes, Latent Class Model) presented in the section *Literature Review: Methodologies for clustering mixed mode data* are applied to the dataset of prostate cancer patients. The perspective which is distinctively investigated for each case of algorithm relates to the cluster characteristics; the key baseline variables which most suitably discriminate the clusters are identified. Moreover, regardless of the clustering method, since the true number of clusters is always unknown, this significant parameter is determined as a combination of a relevant objective criterion (e.g., BIC, Silhouette score, etc.) and the practitioner prior experience and knowledge.

The open-source tool of RStudio (*Version 1.3.959*) is used to develop the R code (*Version 4.0.2*) for the implementation of the clustering algorithms on the prostate cancer data.

Method 1: Kamila Clustering (Kamila R package)

Our research is initiated by running a Kamila clustering procedure on the data. It is of importance to note that several experiments are made regarding the ideal number of clusters that are meaningful in terms of interpretation and practical application. Moreover, what is of particular interest with this experimentation is that as the number of clusters is increased, their vagueness becomes more intense as it is quite difficult to identify the characteristics that differentiate the clusters among them, meaning that the clusters have the tendency to present similar characteristics. In any case, taking into account the possible patterns identified in the proposed clusters of each experiment, 3 clusters are eventually created with the Kamila clustering method by

- specifying the number of cross-validation runs to 10 and
- defining the threshold for determining the number of clusters to 50 and
- setting the weights of continuous and categorical variables equal to 1

As observed in Table 13, the prostate cancer patients are relatively evenly distributed in the clusters as opposed to the cases where more clusters existed (e.g., five, six, etc.) in which unexpectedly small groups of patients formed some of these clusters. As seen below, in the implemented solution most of the patients belong to Cluster 1 (approximately 40%), then to Cluster 2 (approximately 31%) and the rest of them to Cluster 3 (approximately 27%).

Cluster 1	194 patients
Cluster 2	150 patients
Cluster 3	131 patients

Table 13 *Kamila - Cluster Membership*

To accurately comprehend the differentiated features of the three clusters, statistical data on important variables are retrieved and visualizations are drawn, as presented below. To begin with, a trial is made to investigate whether the survival status is a decisive factor affecting the formulation of the clusters. Significant conclusions are drawn from the subsequent Table 14 when identifying that each survival status category (i.e., alive, or dead due to prostate cancer, or dead due to other factors) is particularly dominant in each cluster. More specifically, as indicated by the red colour of each category in this table, a high proportion of patients who are

- **Dead due to prostate cancer** constitute *Cluster 1* (89 out of 194)
- **Survivors or have died due to reasons not related with cancer** constitute *Cluster 2* (129 out of 150)
- **Dead due to other factors** such as heart or vascular disease, cerebrovascular accident, pulmonary embolism, other cancer, respiratory disease, etc. constitute *Cluster 3* (95 out of 131)

Cluster No.	Survival Status		
	Alive	Dead due to prostate cancer	Dead due to other factors
Cluster 1	44	**89**	61
Cluster 2	**68**	21	61

Cluster 3	25	11	**95**

Table 14 *Kamila - Patients' Survival Status Per Cluster*

Based on these figures, it is unquestionably suggested that the Kamila clustering depicts mortality from a disease-specific perspective. As it might be expected, the highly correlated relationship between the structure of the clusters and the outcome variable (i.e., survival status) is further approved by the Chi-Square Goodness of Fit Test as p-value ≤ 0,05.

```
        Pearson's Chi-squared test

data:  kamilaSurvTab
X-squared = 105.56, df = 4, p-value < 2.2e-16
```

When also looking at the results depicted in Figure 8, we see that Cluster 1 appears to have a prevalence of patients with high indices of tumour stage and holistic grade, and serum prostatic acid phosphatase (ranged mostly between 0,6 - 1 and 0,25 - 1 respectively), while Clusters 2 and 3 seem to primarily relate with patients with low levels of serum prostatic acid phosphatase and low indices of tumour stage and holistic grade (ranged mostly between 0 - 0,5 and 0 - 0,25 respectively).

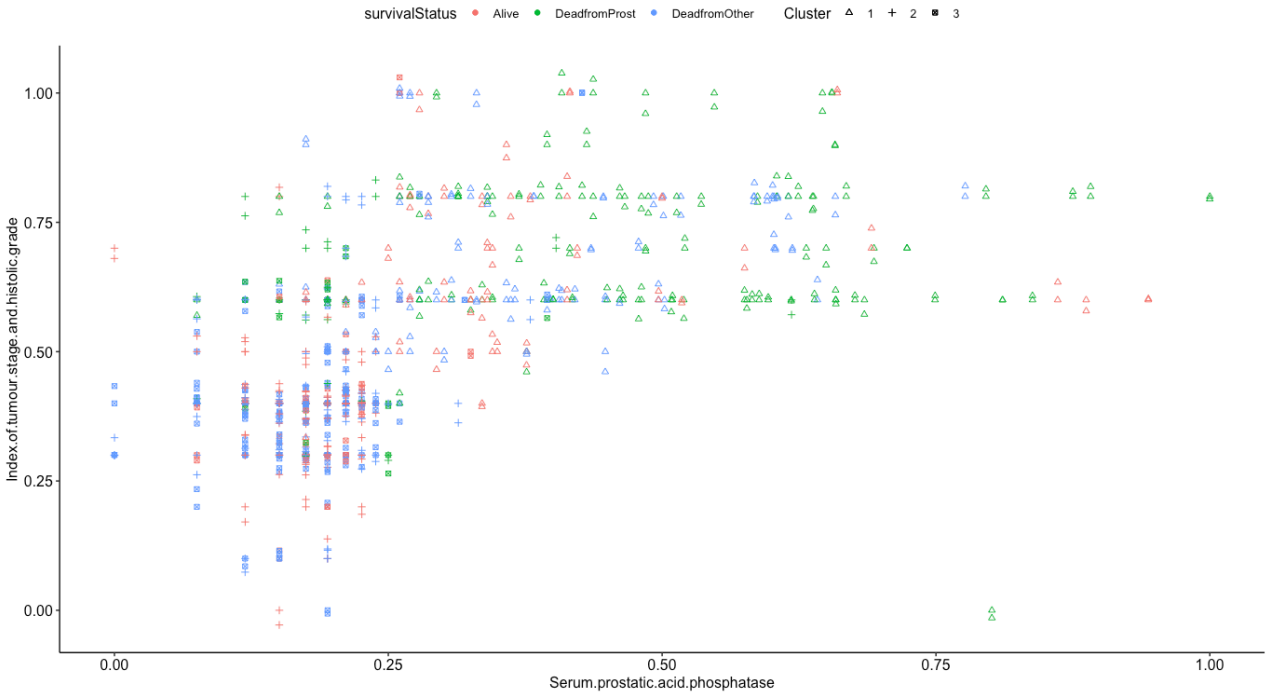


Figure 8 *Kamila - Index of Tumour Stage & Serum Prostatic Acid Phosphatase Levels Per Cluster*

To further analyse the profile of patients in each cluster, it is found in the results of Figure 9 that the vast majority of patients of Cluster 1 is in Stage 4 of prostate cancer with the size of their primary tumour to mostly variate at high levels. On the contrary, as observed in the same Figure, the patients of Cluster 2 and 3 have comparatively smaller sizes of primary tumour, a condition dominantly justified by the fact that they are numerically dominant in Stage 3.

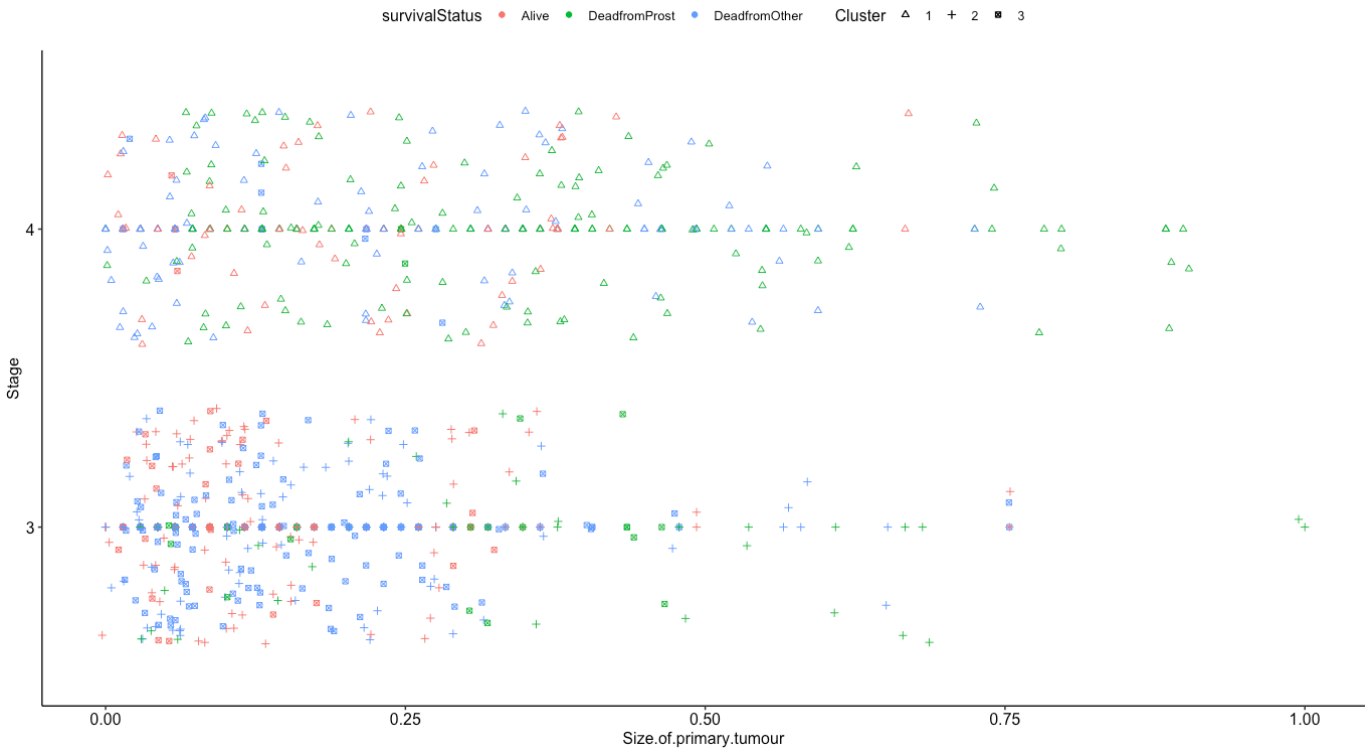


Figure 9 *Kamila - Size of Primary Tumour in Each Stage Per Cluster*

Additionally, Table 15 below that includes tumour stage and bone metastasis for each Kamila cluster membership seems to imply relationships with other variables, as well. More specifically, we can see that every single patient in Cluster 1 had a tumour of stage 4 with the significant percentage of 40% of the patients to have metastatic prostate cancer (76 out of 194). On the other hand, almost all individuals in clusters 2 and 3 have cancer tumours of stage 3. The exception to this are eight people who belong to Cluster 3 and have no spread of the prostate tumour and one person who belongs to Cluster 2 and has metastatic spread of the prostate tumour to the bone, as seen below. Therefore, these findings indicate that the bone metastasis

characteristic suggests an important feature for the individuals who compose Cluster 1 and in fact represent those patients that have mostly died due to prostate cancer.

Bone Metastasis	**Tumour Stage**	Cluster 1	Cluster 2	Cluster 3
No	Stage 3	0	149	123
	Stage 4	118	0	**8**
Yes	Stage 3	0	**1**	0
	Stage 4	76	0	0

Table 15 *Kamila - Bone Metastasis & Tumour Stage Per Cluster*

After having identified the various characteristics that uniquely determine each cluster, the prediction strength method offered in the Kamila package is also used to investigate in a more formal and accurate way whether the selected number of clusters (i.e., three) in this research is comparatively the most optimal choice to be made for the provided dataset. As observed in the below Figure 10, the prediction strength values are plotted against the number of clusters that range between 2 and 15, with the error bars to indicate plus or minus one standard error. The horizontal dotted line which counts at y = 0.8 denotes the default threshold for determining the number of clusters. According to the findings of this method, the ideal number of clusters to be selected is two (as this case presents the highest prediction strength value) followed by the number of three clusters. Apparently, it seems that when examining the case of a greater number of clusters (more than three) to be selected, this is not indicated by the results of the prediction strength method.

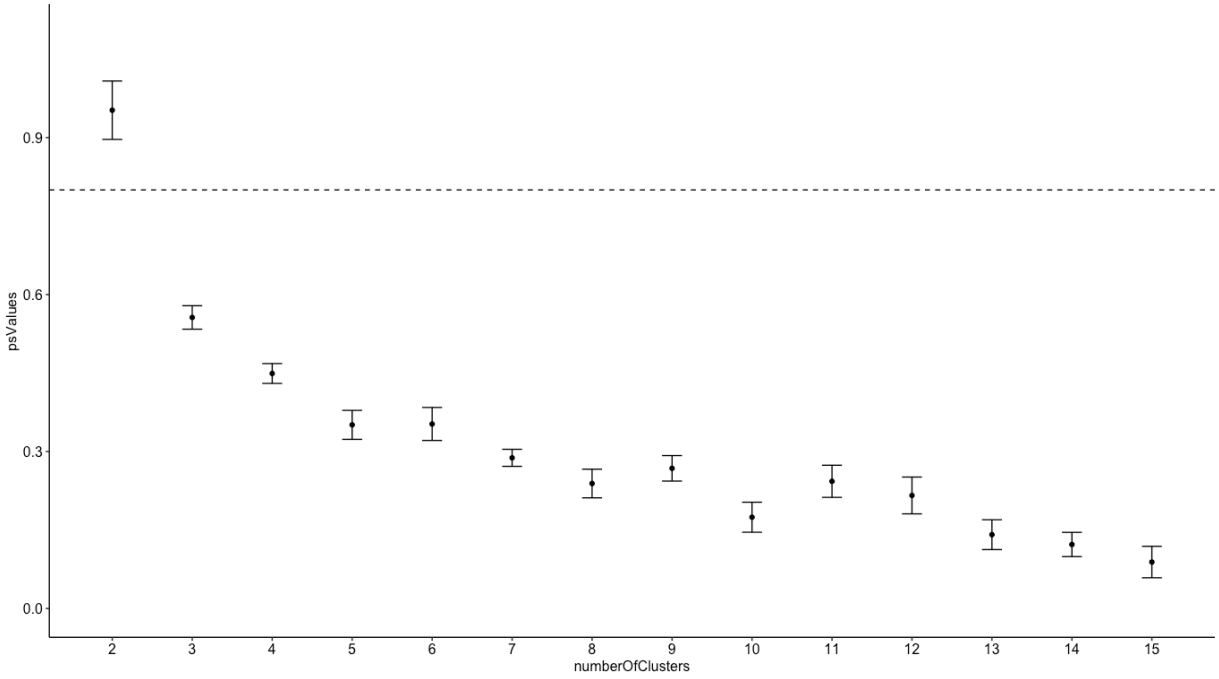


Figure 10 *Kamila - Prediction Strength Method (PS Values VS Number of Clusters)*

We run again the Kamila clustering on the prostate cancer data by using the prediction strength method that proposes two clusters. When selecting to compare how the patients are distributed in the three clusters of the implemented solution versus the two clusters of the prediction strength method, it is concluded that the two-cluster solution is quite similar to the three-cluster one. Apart from an extremely small number of nine patients, Clusters 2 and 3 with the lowest proportion of deaths due to prostate cancer appear seem to be somehow merged. Of course, this conclusion leads to an equally important conclusion that the initial selection of three clusters is well documented and on target.

Implemented Solution	**Kamila Predictive Method**	
	Cluster 1	Cluster 2
Cluster 1	194	0
Cluster 2	1	149
Cluster 3	8	123

Table 16 *Patients' Distribution: Implemented Solution VS Kamila Predictive Method*

Method 2: K-Prototypes Clustering (clustMix R package)

We continue by running a K-Prototypes clustering procedure on the data. The approach is based initially on the discovery of the optimal number of clusters by using measures such as these of the Elbow and Silhouette method.

The Elbow method plots the value of the cost function produced by different values of *k* (number of clusters). This method is helpful because it shows how increasing the number of the clusters contributes to separating the clusters in a meaningful, not marginal way. When *k* increases, the average distortion is decreased, each cluster has fewer constituent instances, and the instances are closer to their respective centroids. On the other hand, the improvements in average distortion are declined as *k* increases. The value of *k* at which improvement in distortion declines the most is called the *elbow*, at which the data should no longer be divided into further clusters. According to this method and as depicted on Figure 11 below, for the given dataset it is concluded that the optimal number of clusters is **5**.

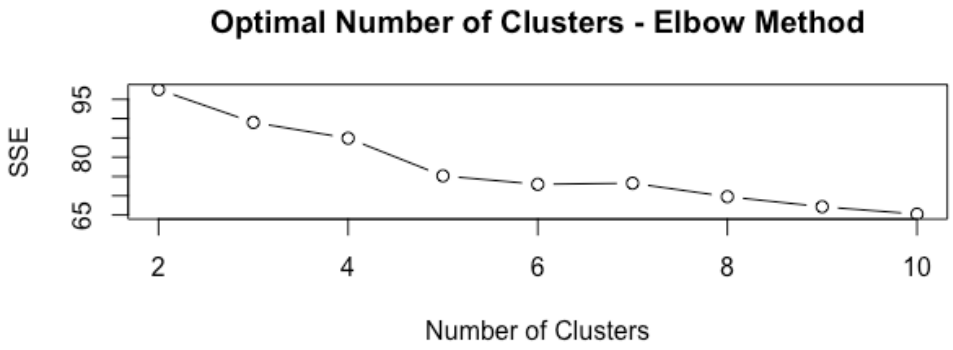


Figure 11 *K-Prototypes - Evaluation of Clusters' Number (Elbow Method)*

The Silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The value of the silhouette ranges between [1, -1], where a high value indicates that the object is well matched to its own cluster and poorly matched to neighbouring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters. So, the silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like the number of clusters visually. For the under-study case, the Silhouette method

(Figure 12, 13) suggests that the most suitable number of clusters is **2** as the "peak" characteristic is spotted there.

Figure 12 *K-Prototypes - Evaluation of Clusters' Number (Silhouette Method)*

```
Number of Clusters: 2
Cluster sizes: 277 198
Within cluster error: 49.81661 47.43263

Cluster prototypes:
  Performance.rating Cardiovascular.disease.history Systolic.Blood.pressure Diastolic.blood.pressure
1                  0                               1              0.2988185                0.3078907
2                  0                               0              0.2773186                0.2817460
  Electrocardiogram.code Serum.haemoglobin Size.of.primary.tumour
1                      0         0.6435091              0.1448229
2                      0         0.5663957              0.2941004
  Index.of.tumour.stage.and.histolic.grade Serum.prostatic.acid.phosphatase Bone.metastases Stage
1                                0.3949458                       0.1813458               0     3
2                                0.7191919                       0.4337131               0     4
  SurvStat
1        0
2        1
```

Figure 13 *Validating K-Prototypes (Method 'validation_kproto' - Silhouette index)*

Taking into account that the Elbow method uses only intra-cluster distances while the Silhouette method uses a combination of inter-cluster and intra-cluster distances in its scoring function, it is expected that these two methods end up with different results. However, it is decided to run a K-Prototypes clustering procedure on the data with the selected number of clusters to be 2, as indicated by the Silhouette method. This is justified by the fact that this method uses more evaluation criteria (both inter-cluster and intra-cluster distances) and from what we have already observed in the previous clustering method is that the differentiated characteristics of the clusters begin to disappear when their number is substantially increased. In the implemented K-Prototypes solution

- the threshold for determining the number of clusters is defined to 50 and
- the lambda parameter that specifies the tradeoff between the Euclidean distance of numeric variables and the simple matching coefficient between categorical variables is automatically calculated to 0,05076704 by using the K-Prototypes function *lambdaest*

In Table 17, the distribution of prostate cancer patients is shown. It is observed that most of the patients belong to Cluster 2 (approximately 59%) and all the others to Cluster 1 (approximately 41%).

Cluster 1	194 patients
Cluster 2	281 patients

Table 17 *K-Prototypes - Cluster Membership*

When also examining the prototypes of the two clusters (Table 18), several interesting conclusions are drawn regarding the characteristics of the patients belonging to each cluster. More specifically, it seems that on average the patients who belong to Cluster 1 have prostate cancer stage 3 with a proven record of cardiovascular disease history and symptoms of heart pressure. On the other hand, the patients of Cluster 2 are at a more advanced stage of cancer with their tumours to be almost double in size and their levels of serum prostatic acid phosphatase to rank much higher (0,43 as opposed to 0,18 of Cluster 1).

Variable	Cluster 1	Cluster 2
Performance rating	0	0
Cardiovascular disease history	1	0
Systolic blood pressure	0,2953737	0,2818650
Diastolic blood pressure	0,3050330	0,2853461
Electrocardiogram code	4	0
Serum haemoglobin	0,6418135	0,5672618
Size of primary tumour	0,1404405	0,3035261

Index of tumour stage and histolic grade	0,3996441	0,7190722
Serum prostatic acid phosphatase	0,1856953	0,4326166
Bone metastases	0	0
Stage	3	4
Survival status	0	1

Table 18 *K-Prototypes - Cluster Prototypes (2 Clusters)*

The above profiling is also verified when observing the visualizations of the K-Prototypes clustering result for cluster interpretation. As it can be seen, boxplots are generated for the numerical variables and bar plots for the factor variables of each cluster. The homogeneity of the clusters can be identified in characteristics like these of performance rating, systolic or diastolic pressure and serum haemoglobin. However, the clusters seem to differ significantly when it comes to factors such as these of *cardiovascular disease history*, *electrocardiogram code*, *size of primary tumour*, *index of tumour stage and histolic grade*, *serum prostatic acid phosphatase*, *bone metastasis*, *cancer stage* and *survival status*.

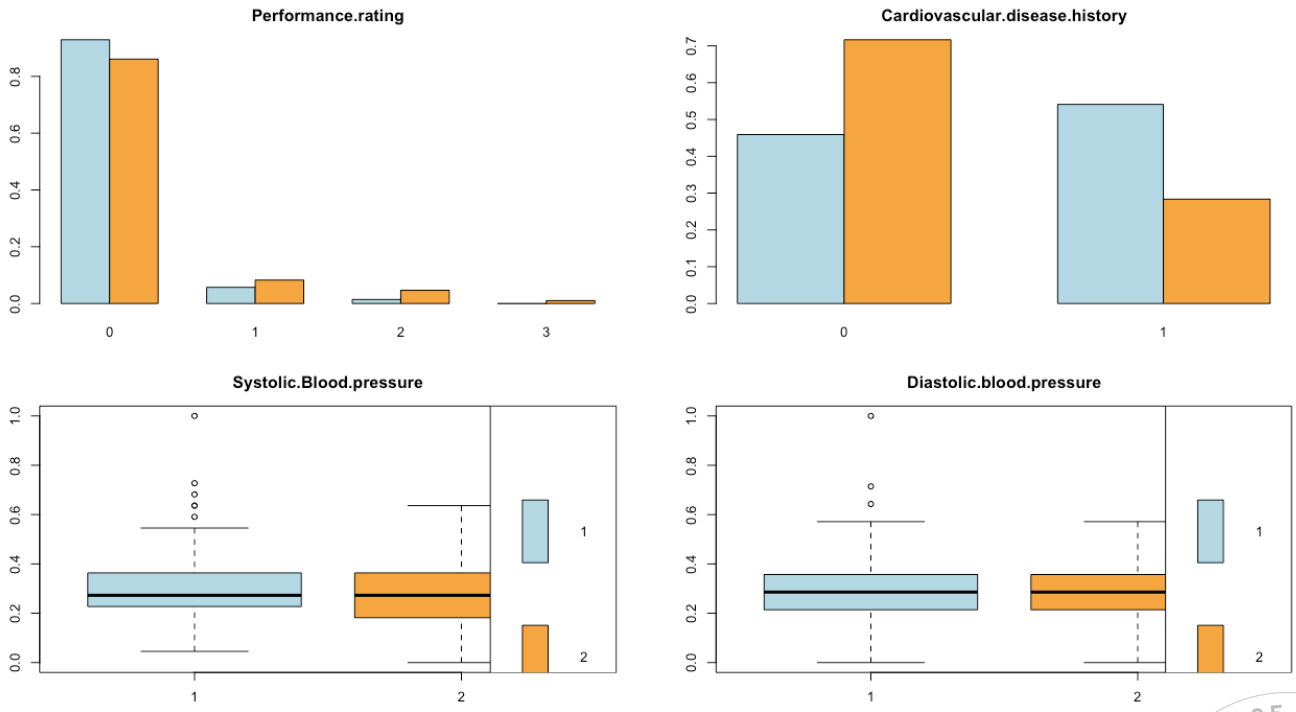


Figure 14 *K-Prototypes – Visualization Results (A)*

Regarding the survival status, it is reasonable to verify that the majority of patients who belong to Cluster 2 have died due to prostate cancer while the patients of Cluster 1 are mostly either alive or have died due to heart or vascular disease, as observed in Figure 16 below.

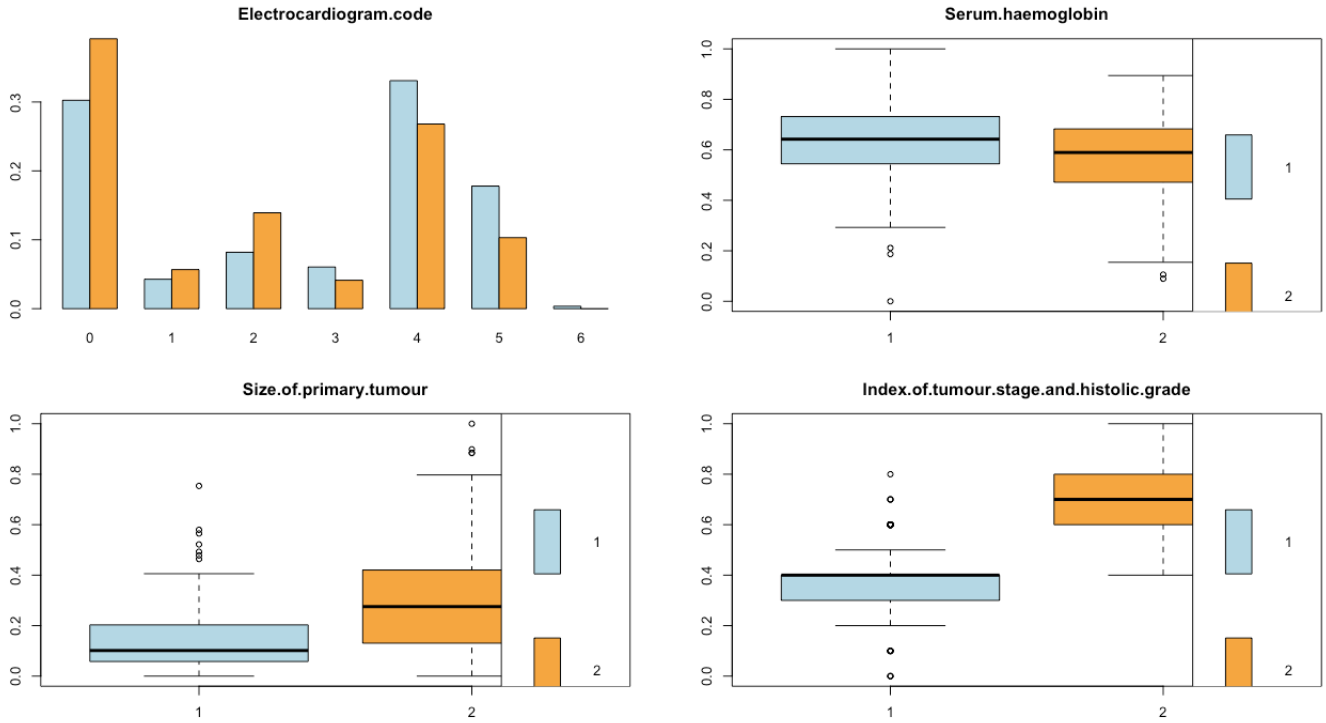


Figure 15 *K-Prototypes – Visualization Results (B)*

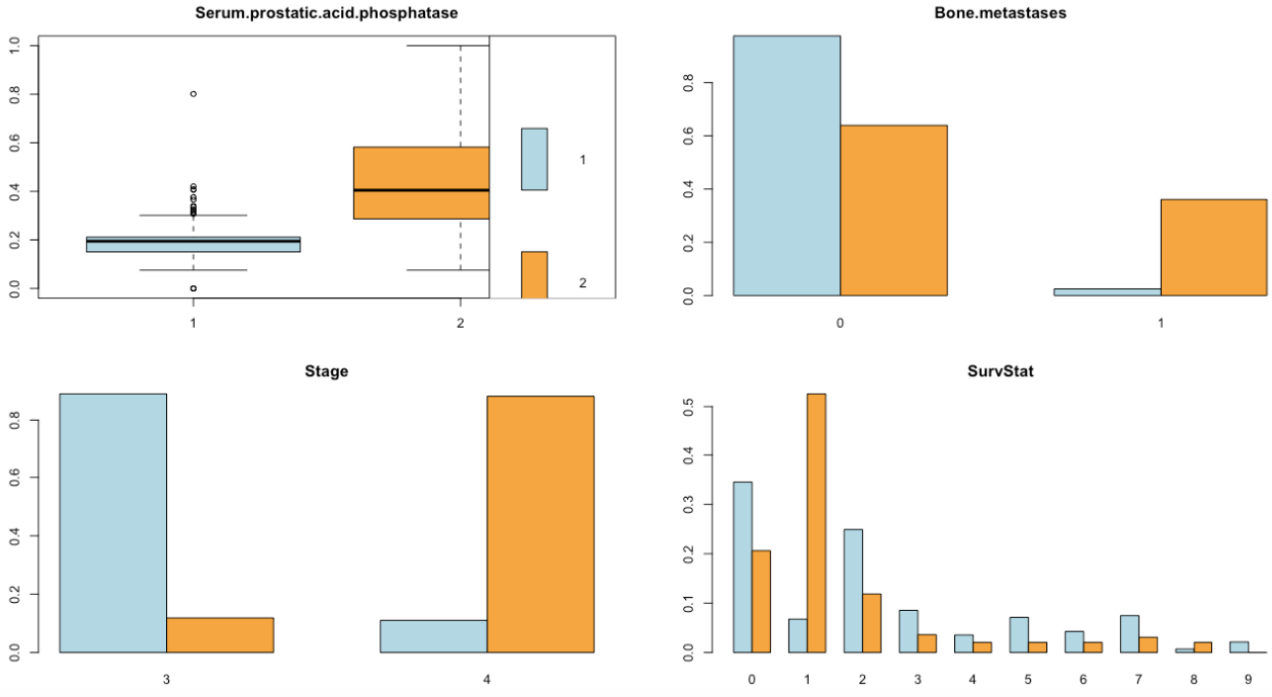


Figure 16 *K-Prototypes – Visualization Results (C)*

A K-Prototypes clustering procedure on the data is performed yet again. For this algorithm execution, the number of clusters is set to 3 in order to investigate if more than one sub-groups with different characteristics exist in the Cluster 1 presented above.

Variable	Cluster 1 (N = 146)	Cluster 2 (N = 173)	Cluster 3 (N = 156)
Performance rating	0	0	0
Cardiovascular disease history	1	0	0
Systolic blood pressure	0.3315691	0.2803468	0.2613636
Diastolic blood pressure	0.3302348	0.2824112	0.2820513
Electrocardiogram code	4	0	0
Serum haemoglobin	0.6308609	0.5574980	0.6528560
Size of primary tumour	0.1501886	0.3173327	0.1379599
Index of tumour stage and histolic grade	0.4164384	0.7323699	0.4121795
Serum prostatic acid phosphatase	0.1871842	0.4552140	0.1924813
Bone metastases	0	0	0
Stage	3	4	3
Survival status	2	1	0

Table 19 *K-Prototypes - Cluster Prototypes (3 Clusters)*

In this experimentation, it is observed that although Clusters 1 and 3 appear to be similar since on average patients with cancer of stage 3 belong to both clusters, the characteristic that differentiates the clusters between them is the fact that the patients in Cluster 1 seem to have a cardiovascular disease history and heart strain - features

that do not exist in patients of Cluster 3. The rest characteristics (e.g., systolic, and diastolic blood pressure, serum haemoglobin, performance rating, etc.) are more or less the same between Clusters 1 and 3. The value of this insight will be further investigated in the subsequent sub-section *Comparison of clustering methods*.

Method 3: Latent Variable Model (clustMD R package)

In the last clustering method, the optimal number of clusters for the provided dataset taking into account the characteristics of this algorithm is again investigated. Through this perspective, a suite of six different clustMD models was fitted to the set of prostate cancer patients with the number of clusters ranging from 1 to 4, as depicted in Figure 13. Each model represents a unique covariance structure offered by the algorithm. More specifically, *Model 1* matches to *EII*, *Model 2* to *VII*, *Model 3* to *EEI*, *Model 4* to *VEI*, *Model 5* to *EVI* and *Model 6* to *VVI*. The line plot below depicts the approximated BIC values for the models which vary from around -12.450 to -11.730. As it is observed the EII, VII, EEI and VEI. models perform worse in comparison to the EVI and VVI ones which are way ahead and competing with each other as to which of the two will be leading eventually with the EVI currently prevailing. Thus, it is concluded that in our case the appropriate model which maximizes the BIC criterion is a 3-cluster model, with the EVI covariance structure.

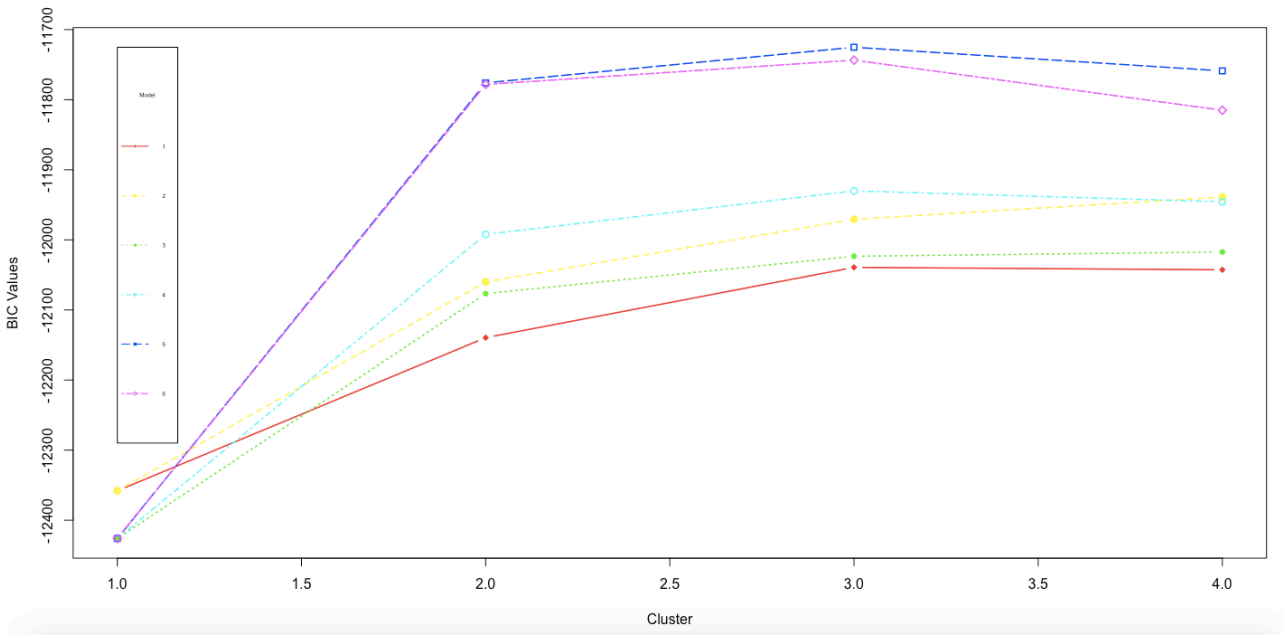


Figure 17 *Latent Class Model - Line Plot of BIC Values*

Based on this observation, we finally choose to implement a model of EVI covariance structure with 3 clusters. However, it is essential to highlight that, in comparison with the Kamila and K-Prototypes clustering algorithms, the computation time differed significantly for Latent Variable Model (for the same number of fifty maximum iterations) since for the latter method a two-fold clustering time is approximately required. A possible cause for this problem could be the Monte Carlo approximation used in the E-step of the model fitting algorithm, which is a simple and effective solution, but it does not come without issues. If the probability of observing a particular response on a nominal variable is very small for a particular cluster, then a large number of Monte Carlo samples may be required to observe a response in this category. This can slow the model fitting algorithm or even cause instability. Of course, having acknowledged that the specific dataset consists of only 475 observations, this problem scales significantly when dealing with larger datasets and undoubtedly raises concerns regarding the efficiency of the algorithm in such cases - this aspect should not be underestimated as datasets can quickly reach huge proportions in a big-data context. As McParland, D. and Gormley, I.C. already suggested, a more efficient way to approximate the intractable integrals could improve the model fitting efficiency.

Upon further analyzing the results of the implemented solution, the number of patients in each of these clusters is presented in Table 18. As we can see, around 42% of the patients belong to Cluster 3 while 26% of the patients belong to Cluster 2 and 30% of them belong to Cluster 1.

Cluster 1	145 patients
Cluster 2	127 patients
Cluster 3	203 patients

Table 20 *Latent Class Model - Cluster Membership*

When assessing the health status of the patients in each cluster with regards to the cancer stage (Table 19), it is easily observed that stage 4 patients are dominant in Cluster 3 while Clusters 1 and 2 seem quite similar since mostly patients diagnosed with stage 3 are included in them.

	Stage 3	Stage 4
Cluster 1	**139**	6
Cluster 2	**108**	19
Cluster 3	26	**177**

Table 21 *Latent Class Model – Distribution of Patients Per Stage & Cluster*

However, this similarity is questioned when comparing the estimated mean vectors for all three clusters (Figures 14, 15). It is surprising that patients in Cluster 1 have on average much higher levels of diastolic and systolic blood pressure; the metrics for these patients approximate around 0,360 and 0,396 respectively as compared to - 0,238 and - 0,351 of Cluster 2. Additionally, chances are higher for them to have a history of cardiovascular disease and their electrocardiogram index is more likely to point out a severe abnormality in their health. This suggests that a cardiovascular health issue differentiates patients in Cluster 1 from those in Cluster 2.

On the other hand, as also shown in Figures 14 and 15, it seems that the patients of Cluster 3 suffer the effects of the advanced prostate cancer on their health as they are more likely to have bone metastasis and on average the size of their primary tumour is enlarged significantly with the index of tumour stage, histolic grade and serum prostatic acid phosphatase levels to be extremely high.

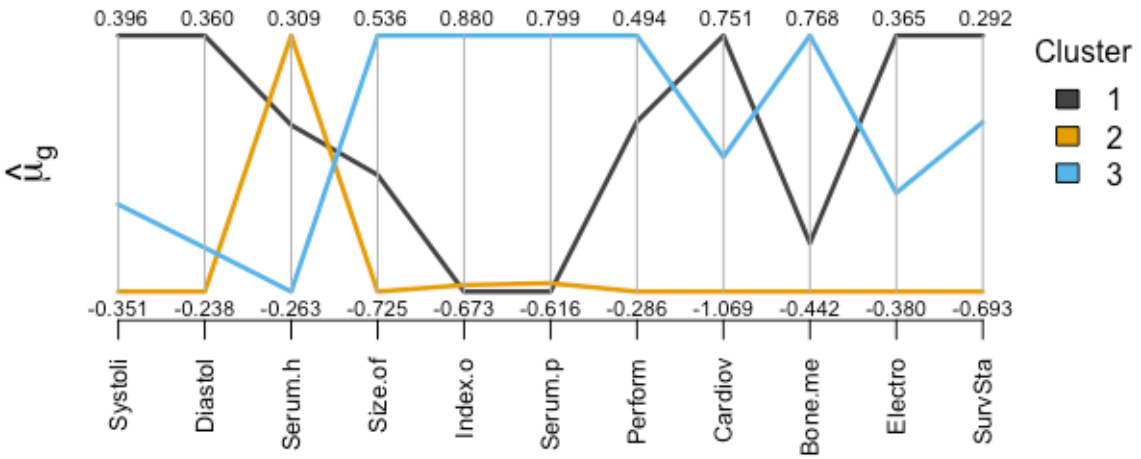


Figure 18 *Latent Class Model – Estimated Cluster Means (A)*

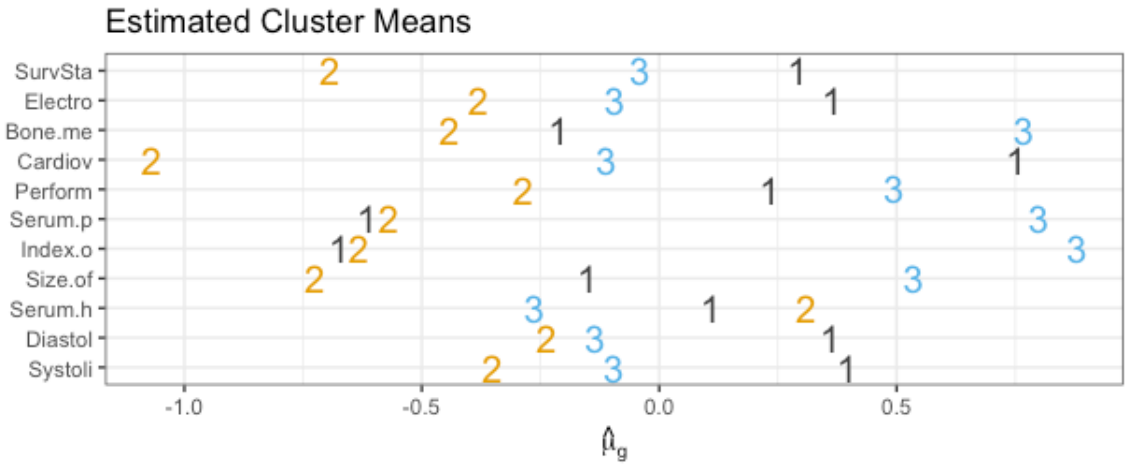


Figure 19 *Latent Class Model – Estimated Cluster Means (B)*

The above findings for the patients' profile are also verified by examining their post-trial survival status. Indicatively, it seems that 43% of Cluster 1 patients died due to heart or vascular disease or a cerebrovascular accident, 21% are alive and only 9% died due to prostate cancer. The remaining 27% died of other causes.

Moreover, it can be seen that 50% of patients in Cluster 3 died due to prostatic cancer (as compared to 9% in Cluster 1 and 6% in Cluster 2), 19% survived until the end of the trial (as compared to 21% in Cluster 1 and 62% in Cluster 2) and the rest 31% died of other reasons (as compared to 27% in Cluster 1 and 40% in Cluster 2).

Cluster No.	**Survival Status**		
	Alive	Died due to prostate cancer	Died due to heart or vascular disease or stroke
Cluster 1	30	13	**62**
Cluster 2	**68**	7	22
Cluster 3	39	**101**	40

Table 22 *Latent Class Model – Survival Status of Patients Per Cluster*

After having completed the profiling of the patients among clusters, the evaluation of the clustering performance follows by analysing the cluster variances and uncertainty, as depicted in Figures 16 and 17 below. As it is already known, the cluster variance is the coordinate-wise squared deviations from the mean of the cluster of all the observations belonging to that cluster. A small variance indicates that the

data points tend to be very close to the mean, and to each other within the cluster while a high variance indicates that the data points are very spread out from the mean, and from one another within the cluster.

In Figure 16, it is noticeable that the variances are relatively small (less than 2) for all three clusters. An exception to this are the variances for the variables of cardiovascular disease history, electrocardiogram code and survival status of Cluster 2 and the variable of cardiovascular disease of Cluster 1. For these cases, it seems that the variances range at higher levels indicating that some patients may have been wrongly included in these clusters as these clusters are not so clearly separated. In Figure 17, the misclassification rate is neither very high nor low as for around 250 patients (> 50%) the clustering uncertainty is rather low and for the other 225 patients, it presents a higher clustering uncertainty that ranges between 0,1 and 0,5 approximately.

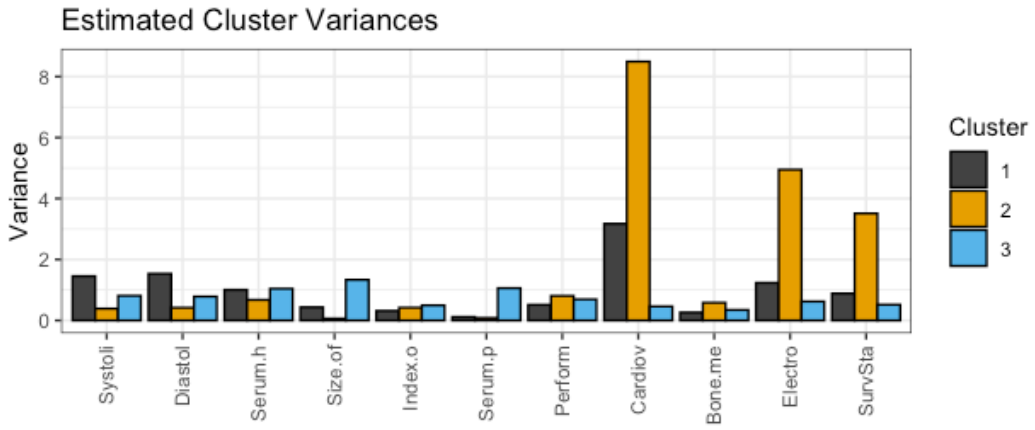


Figure 20 *Latent Class Model – Cluster Variances*

Figure 21 *Latent Class Model – Clustering Uncertainty*

Comparison of clustering methods

As known, cluster analysis can be considered as successful only if the partition makes sense for the practitioner. In the current section, the clustering results of all three methods are compared in terms of interpretation, performance and efficiency.

Interpretation. In the *Kamila clustering*, 3 clusters are created with the factor of the patients' survival status to stand out as the most significant factor in the formulation of clusters. In this method, one cluster (Cluster 1) consists primarily of patients whose mortality cause is prostate cancer, they are in cancer stage 4 with bone metastasis, have increased tumour sizes, high indices of tumour stage and holistic grade, and serum prostatic acid phosphatase. The other two clusters (Cluster 2 & 3) consist mainly of patients with less deteriorated health who have died due to other reasons different from prostate cancer (e.g., heart or vascular disease, cerebrovascular accident, etc.) or have survived from the trial and appear to have treatable tumours and low levels of serum prostatic acid phosphatase and indices of tumour stage and holistic grade.

In the *K-Prototypes clustering*, two different experiments are made involving clusters 2 and 3. For the former experiment, stage 3 patients on average are included in the first cluster (Cluster 1). In general, these people deal with cardiovascular problems, are diagnosed with heart pressure, and are still alive or dead because of heart or vascular diseases. The remaining health indicators for these patients (e.g., serum haemoglobin, tumour size, serum prostatic acid phosphatase, etc.) usually range within normal levels. In the second cluster (Cluster 2), patients with metastatic prostatic cancer belong to the group with oversized tumours, high levels of serum prostatic acid phosphatase and indices of tumour stage and holistic grade and for whom prostate cancer is the main cause of death. For the latter experiment, the patients who belong to Cluster 1 of the first experiment are practically divided into two different clusters (Cluster 1 & 3) with the distinction between them to be the severe cardiovascular problems of the patients in Cluster 1.

In the *Latent Variable Model clustering*, 3 clusters are formed. The first two clusters (Cluster 1 & 2) are comprised of patients with stage 3 who have survived or have died because of heart or vascular disease or stroke and for these patients, indices like primary tumour size, serum prostatic acid phosphatase, serum haemoglobin, etc.

are relatively low. The difference which is identified between these two clusters is that the patients of Cluster 1 cope with blood pressure and cardiovascular health issues. The third cluster (Cluster 3) consists of advanced prostate cancer patients who have mainly died due to cancer disease and present large tumours and high levels of serum prostatic acid phosphatase.

Based on the above findings, it is derived that the three algorithms move within an interpretative framework that is quite similar. All algorithms have separate clusters of stage 3 patients who are either in relatively good health conditions or suffer from various problems that are not related to cancer (e.g., cardiovascular ones) and stage 4 patients who face cancer-related issues (e.g., bone metastasis). Moreover, for any of the three clustering methods both categorical variables (e.g., stage, cardiovascular disease history, bone metastases, etc.) and continuous variables (e.g., tumour size, serum haemoglobin, serum prostatic acid phosphatase) play a significant role in the discrimination of the clusters with the most decisive factor for the generation of clusters to be most probably this of the patients' survival status (the outcome variable).

Additionally, at this point it is also of interest to emphasise that from the experience acquired in the context of the current research, it is observed that the K-Prototypes and Latent Variable Model are incorporated in "ready-to-use" R software packages which in some extend can be easily used by non-expert teams. This is because several built-in visualization functions are offered in these packages and any researcher who is not familiar with the relevant algorithm can effortlessly and directly interact with these functions to better interpret and fine-tune the clustering results as desired. On the other hand, even if Kamila algorithm is well suited for a big-data setting and recommends an advance over existing methods as it uses variables in their original measurement, ensures equitable impact of continuous and categorical variables and does not require the user to specify variable weights or coding schemes, it does not offer such provision for visualization to the potential researchers. The interested parties are required to invest more time and effort to extract the necessary data insights from the clustering of the Kamila algorithm.

Performance & Efficiency. To compare the partitions produced by any of the clustering methods, the *Adjusted Rand Index (ARI)* and the *Rand Index* (cluster_similarity function in clusteval R package) are used. The ARI and the Rand

measures provide an agreement score between a pair of partitions, ranging from 0 (complete disagreement) to 1 (complete agreement).

As depicted in Table 21 below, for some pairs of algorithms such as these of *Kamila – K-Prototypes (2 clusters)*, *K-Prototypes (2 clusters) – LVM* and *K-Prototypes (3 clusters) – LVM* where both the Rand Index and the ARI score are close to or above 50%, the clustering results can be described as quite satisfactory. This practically means that there is an evident similarity in the resulting partitions for these methods (*Kamila - K-Prototypes*, *K-Prototypes – LVM*). This is not, however, the case for the *Kamila - Latent Variable Model* methods since the performance measures for this specific clustering pair present the lowest values of 0,28 and 0,39 for the ARI and Rand Index respectively indicating the greatest dissimilarity among all possible combinations of methods.

Clustering Methods	Rand Index	Adjusted Rand Index
Kamila – K-Prototypes (2 clusters)	0,61	0,54
K-Prototypes (2 clusters) – LVM	0,55	0,50
LVM – Kamila	0,39	0,28
Kamila – K-Prototypes (3 clusters)	0,39	0,31
K-Prototypes (3 clusters) – LVM	0,51	0,50

Table 23 *Clustering Methods – RI & ARI Measures*

Additionally, as already indicated in the previous sub-section *Method 3: Latent Variable Model (clustMD R package)*, an increased computational complexity is shown for the Latent Variable Model when compared to the other algorithms since in the former case, almost double time is required for the clustering results of 475 observations to be generated. The computation time for this model-based method seems to heavily depend on the complexity of the selected model, the number of iterations and the additional features. When also judging the Kamila algorithm from this time-efficiency standpoint, this method appears to offer the best performance

when dealing with large datasets as it has been implemented to work in a Big-Data setting by taking advantage of the scalability of its algorithm.

Taking into consideration the aforementioned analysis with regards to the interpretation, performance, and efficiency pillars, it is considered that the selection of the optimal clustering algorithm lies between the K-Prototypes (2 clusters) and Kamila methods. A valid argument justifying this statement is that the greatest partition similarity is identified between these two algorithms (0,54 for ARI and 0,61 for RI). Moreover, from the interpretation perspective, they both have clusters consisting of

- stage 3 patients who have either survived or have died due to reasons different than cancer (e.g., cardiovascular diseases) and their health indicators such as tumour size, serum prostatic acid phosphatase, etc. are relatively normal and
- stage 4 patients who have oversized tumours, high indices of tumour stage and holistic grade, and serum prostatic acid phosphatase and die mostly due to the cancer disease

However, when evaluating the pros and cons of the algorithms themselves, we can state that the Kamila algorithm outweighs against K-Prototypes. It overcomes the challenges inherent in the various extant methods for clustering mixed continuous and categorical data (i.e., variables are used in their original measurement scale, the different types of variables are balanced by using the properties of Gaussian-multinomial mixture models, overly restrictive parametric assumptions are avoided for numeric features, etc). On the other hand, the K-Prototypes algorithm may cause uncertainty and inaccuracy due to the randomness in the initial clustering results and the use of the simple Hamming distance (0 or 1) to calculate the dissimilarity between the categorical data. It is therefore concluded that the Kamila clustering method might comprise a last resort selection for the clustering of the prostate cancer patients.

In any case, the above clustering results recommend a powerful asset of information likely leading to the generation of patient-specific and evidence-based advice and to the creation of reminders for preventive care, and alerts about potentially dangerous situations to aid clinical decision making by health care providers. While electronic health records and databases help physicians manage the rising tide of information, this kind of patient-specific recommendations could feed the clinical decision support systems by enhancing decision making and aiding the establishment of patient safety. More specifically, each category of cancer patients deriving from

the clustering process could belong to a different monitoring protocol in terms of both personalized diagnosis and treatment. In addition to this, significant pillars such as these of the health economics, the patients' financial costs, and the resources management in health care institutions could be augmented and more strategically organized based on the health needs and requirements of the respective cancer patient category. This patient-targeted and clinical knowledge unquestionably advances the quality and efficiency of medical services, reduces patient inconvenience, and probably results in lower costs through clinical interventions, decrease of inpatient length-of-stay, integrated systems suggesting cheaper medication alternatives, or reducing test duplication.

V. Conclusions & Future Work

This thesis aimed to effectively cluster prostate cancer patients into groups of similar characteristics by approaching the clustering problem of mixed mode data. Based on both the quantitative and the qualitative analyses of the prostate cancer data in response to the patients' health indicators, it can be derived that the mortality causes as well as the cancer stage are important factors to take into consideration when targeting on patients diagnosed with this disease. The data insights have also demonstrated that most patients have a normal activity in their life regardless of the cancer stage with the cardiovascular disease to be the second most common cause of death in these patients after cancer itself.

As thoroughly detailed, the focus of research in clustering data has recently moved from numeric data to categorical ones because almost all real data are categorical. And given the growing access to multiple data sources, it has become crucial to be able to manage all types of variables through recent advances in statistics and machine-learning. However, clustering categorical data is a slightly more challenging task than clustering numeric data because of the absence of any natural order, high dimensionality, and existence of subspace clustering. The earlier notions of statistics and geometry could not be applied to categorical data due to some limitations of the categorical data. As time passes by, researchers have proposed clustering methods that can directly be applied to categorical data with the most common approach to be this of their conversion into an equivalent numeric form easier to handle but with its own limitations. Over the years, various classic clustering algorithms have been proposed and are still used (PAM, CLARA, BIRCH, K-Modes, etc.). The new developments in this direction are either improvements or extensions of the old algorithms (Mixmox, Latent Class Model, Latent Class Analysis, etc.).

Within the scope of this thesis, the clustering problem of mixed mode data is modelled by applying three distance-based and model-based algorithms (Kamila, K-Prototypes, Latent Variable Model) which approached the various challenges derived from data of various types (continuous, nominal, ordinal, etc.) differently and uniquely. Several experimentations were performed for each algorithm in order to identify the ideal number of clusters which proved to range between two and three with the partition results to indicate a particular enlightening aspect for the prostate

cancer patients each time. In terms of clustering efficiency and performance, the Kamila and K-Prototypes methods ranked higher while all three methods appear to similarly illustrate how the prostate cancer patients could be grouped. In the end, after closely examining the outcome of each clustering method, the method chosen as the optimal one for the under-study case of prostate cancer patients was this of Kamila algorithm owing to its eminent scientific capabilities and advantages over the other methods.

According to the clustering results of the selected Kamila algorithm, the formulated clusters consisted of stage 3 patients who have either survived after trial and are in relatively good health or tackle with other major health diseases such as respiratory, cardiovascular, or cerebrovascular issues along with it. The results have also confirmed that these patients appear to have smaller tumours and moderately low levels of serum prostatic acid phosphatase and indices of tumour stage and holistic grade - a fact that justifies why the most frequent cause of their death is not cancer but other equally severe diseases. Another significant group that is also derived from the clustering process is this of stage 4 patients who are usually diagnosed with metastatic cancer, have heightened tumour sizes, elevated indices of tumour stage and holistic grade as well as serum prostatic acid phosphatase. It is the category of patients mostly dying of cancer. In practice, the described grouping of cancer patients could assist the clinical decision support systems used by health care providers in upgrading the profiling of patients and the recommended care and treatment actions performed for them.

Within the context of the current thesis, specific data manipulations and statistical approaches were adopted with a view to investigating the research problem of clustering. Nonetheless, there is clearly future work to be done on exploring this scientific area. The research represented in this thesis has addressed some of the fundamental problems with regards to clustering of mixed mode data and these hint at the direction for the further work as there are extensions and alternative ways of dealing with this problem. Indicatively, most close to this work would be to investigate the models with more combinations for the weights of continuous and categorical variables and the impact of the existing outliers in the proposed models as well. The feature of outliers was not used here due to lack of time so the impact of these characteristics would be worth probing into. In addition, scenarios with non-normal continuous variables would have yielded different results and could constitute the

object of further future studies. Future research could also fruitfully explore the issue further by potentially examining different number of initializations for the algorithms considering that initialization represents a trade-off between the robustness of the partition and computation time.

VI. References

Abbas, O.A. (2008). Comparisons Between Data Clustering Algorithms. International Arab Journal of Information Technology 5, Volume 5(Issue 3), pp.320–325.

Ahmad, A. and Khan, S.S. (2019). Survey of State-of-the-Art Mixed Data Clustering Algorithms. IEEE Access, Volume 7, pp.31883–31902.

Cancer.net. (2012). Prostate Cancer - Introduction. [online] Available at: https://www.cancer.net/cancer-types/prostate-cancer/introduction.

Cancer.net. (2018). Prostate Cancer - Risk Factors and Prevention. [online] Available at: https://www.cancer.net/cancer-types/prostate-cancer/risk-factors-and-prevention.

Cancer.net. (2018). Prostate Cancer - Types of Treatment. [online] Available at: https://www.cancer.net/cancer-types/prostate-cancer/types-treatment.

Carriere, K. and de Leon, A. (2013). Analysis of Mixed Data: Methods & Applications. [online] CRC/Chapman & Hall, pp.1–11. Available at: https://www.researchgate.net/publication/265601281_Analysis_of_mixed_data_An_overview [Accessed 20 May 2021].

Cross Validated. (n.d.). clustering - Latent Class Analysis vs. Cluster Analysis - differences in inferences? [online] Available at: https://stats.stackexchange.com/questions/122213/latent-class-analysis-vs-cluster-analysis-differences-in-inferences [Accessed 3 May 2021].

Foss, A., Markatou, M., Ray, B. and Heching, A. (2016). A semiparametric method for clustering mixed data. Machine Learning, [online] Volume 105(Issue 3), pp.419–458. Available at: https://link.springer.com/article/10.1007/s10994-016-5575-7 [Accessed 15 Nov. 2020].

Foss, A.H. and Markatou, M. (2018). kamila: Clustering Mixed-Type Data in R and Hadoop. Journal of Statistical Software, [online] Volume 83(Issue 13). Available at: https://www.jstatsoft.org/article/view/v083i13 [Accessed 14 Nov. 2020].

Hennig, C. (2015). What are the true clusters? Pattern Recognition Letters, Volume 64, pp.53–62.

Hendrickson, J.L. (2014). Methods for Clustering Mixed Data. [online] Available at: https://scholarcommons.sc.edu/etd/2590/ [Accessed 19 May 2021].

Huang, Z. (1998). Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery 2, [online] pp.283–304. Available at: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.15.4028&rep=rep1&type=pdf [Accessed 15 Dec. 2020].

Jia, Z. and Song, L. (2020). Weighted k-Prototypes Clustering Algorithm Based on the Hybrid Dissimilarity Coefficient. Mathematical Problems in Engineering, 2020, pp.1–13.

Leisch, F. (2004). FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression inR. Journal of Statistical Software, Volume 11(Issue 8).

Magidson, J. and Vermunt, J.K. (2002). Latent class models for clustering : A comparison with K-means. Canadian Journal of Marketing Research, [online] Volume 20. Available at: https://jeroenvermunt.nl/cjmr2002.pdf [Accessed 5 Jun. 2021].

Marbac, M., Mohammed, S. and Patin, T. (2019). Variable Selection for Mixed Data Clustering: Application in Human Population Genomics. Journal of Classification, [online] Volume 37(Issue 2). Available at: https://www.researchgate.net/publication/332097548_Variable_Selection_for_Mixed_Data_Clustering_Application_in_Human_Population_Genomics [Accessed 25 Nov. 2020].

Matthieu, M. and Mohammed, S. (2017). Variable selection for model-based clustering using the integrated complete-data likelihood. Statistics and Computing, [online] Volume 27(Issue 4), pp.1049–1063. Available at: https://www.researchgate.net/publication/271447963_Variable_selection_for_model-based_clustering_using_the_integrated_complete-data_likelihood [Accessed 19 Dec. 2020].

Mcparland, D. and Gormley, I.C. (2015). Model Based Clustering for Mixed Data: clustMD. Advances in Data Analysis and Classification, [online] Volume 10(Issue 2), pp.155–169. Available at: https://www.researchgate.net/publication/283531301_Model_Based_Clustering_for_Mixed_Data_clustMD [Accessed 10 Jun. 2021].

Parul, A., Afshar, A. and Ranjit, B. (2011). Issues, Challenges and Tools of Clustering Algorithms. International Journal of Computer Science Issues, Volume 8(Issue 3).

Preud'homme, G., Duarte, K., Dalleau, K., Lacomblez, C., Bresso, E., Smaïl-Tabbone, M., Couceiro, M., Devignes, M.-D., Kobayashi, M., Huttin, O., Ferreira, J.P., Zannad, F., Rossignol, P. and Girerd, N. (2021). Head-to-head comparison of clustering methods for heterogeneous data: a simulation-driven benchmark. Scientific Reports, Volume 11(Issue 1).

Prostate Cancer UK. (2019). Advanced prostate cancer. [online] Available at: https://prostatecanceruk.org/prostate-information/just-diagnosed/advanced-prostate-cancer.

Devos, Andy.K., van Huffel, S., Simonetti, A.W., van der Graaf, M., Heerschap, A. and Buydens, L.M.C. (2007). Classification of Brain Tumours by Pattern Recognition of Magnetic Resonance Imaging and Spectroscopic Data. Outcome Prediction in Cancer, pp.285–318.

Saxena, A. and Singh, M. (2016). Using Categorical Attributes for Clustering. International Journal of Scientific Engineering and Applied Science (IJSEAS), Volume 2(Issue 2), pp.324–329.

Szepannek, G. (2019). clustMixType: User-Friendly Clustering of Mixed-Type Data in R. The R Journal, Volume 10(Issue 2), p.200.

Wikipedia. (2021). Cluster analysis. [online] Available at: https://en.wikipedia.org/w/index.php?title=Cluster_analysis&oldid=1020699527 [Accessed 15 Jun. 2021].

WCRF International. (n.d.). Prostate cancer statistics | World Cancer Research Fund International. [online] Available at: https://www.wcrf.org/dietandcancer/prostate-cancer-statistics/.

Xu, D. and Tian, Y. (2015). A Comprehensive Survey of Clustering Algorithms. Annals of Data Science, Volume 2(Issue 2), pp.165–193.