CLUSTERING METHODS WITH VARIABLE SELECTION FOR DATA WITH MIXED VARIABLE TYPES OR LIMITS OF DETECTION

by

Shu Wang

MS, Columbia University, 2015

BS, Beijing Normal University, 2013

Submitted to the Graduate Faculty of

the Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2019

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Shu Wang

It was defended on

April 15th 2019

and approved by

Dissertation Advisor:

Jonathan G. Yabes, PhD, Assistant Professor, Department of Medicine and Department of Biostatistics, School of Medicine and Graduate School of Public Health, University of Pittsburgh

Dissertation Co-Advisor:

(Joyce) Chung-Chou H. Chang, PhD, Professor, Department of Medicine and Department of Biostatistics, School of Medicine and Graduate School of Public Health, University of Pittsburgh

Committee Member:

Stewart J. Anderson, PhD, Professor, Departmental of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Committee Member:

Qi Mi, PhD, Assistant Professor, Department of Sports Medicine and Nutrition, School of Health and Rehabilitation Sciences, University of Pittsburgh

Committee Member:

Christopher W. Seymour, MD, MSc, Associate Professor, Department of Critical Care Medicine and Emergency Medicine, School of Medicine, University of Pittsburgh

CLUSTERING METHODS WITH VARIABLE SELECTION FOR DATA WITH MIXED VARIABLE TYPES OR LIMITS OF DETECTION

Shu Wang, PhD

University of Pittsburgh, 2019

ABSTRACT

Clustering has emerged as one of the most essential and popular techniques for discovering patterns in data. However, challenges exist in application of clustering. First, many of the existing clustering methods are only useful for data with either all continuous or all categorical variables, despite the abundance of data with mixed variable types. Second, clustering algorithms typically require complete data. But measurements for clinical biomarkers are often subject to limits of detection (LOD). In addition, researchers are getting more interest in knowing variable importance due to the increasing number of variables that become available for clustering. To overcome aforementioned challenges, this dissertation proposes clustering methods for mixed data with the ability of variable selection and handling censored biomarker variables.

In the first section, we propose a hybrid density- and partition-based (HyDaP) algorithm for mixed data. The HyDaP algorithm involves two steps: variable selection step and clustering step. In the first step, variables that have much contribution to clustering will be selected; in the second step, a novel dissimilarity measure will be applied on those selected variables and obtain final results. Simulations and real data analysis were conducted to compare the performance of the HyDaP algorithm with other commonly used clustering algorithms.

In the second section, we propose a Bayesian finite mixture model to simultaneously conduct variable selection, account for biomarker LOD and obtain clustering results. We put a spike-and-slab type of prior on each variable to obtain variable importance. To account for LOD, we added one more step in Gibbs sampling that iteratively fills in censored biomarker values. The same simulation settings and real data were used to evaluate its clustering performance.

PUBLIC HEALTH SIGNIFICANCE: This dissertation proposes clustering algorithms that can be applied to any mixed data with or without censored biomarkers like electronic health record (EHR) data and other clinical data. The identified patient subgroups could provide medical experts more knowledge of patient heterogeneity and the selected important variables could let them better know where the heterogeneity comes from. Thus these information could help develop precision medicine for better patient care.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to thank my advisors Dr. Yabes and Dr. Chang for their tremendous help and guidance in my dissertation. I really appreciate the opportunity of working with them and really enjoy the time we work together. I also want to express my gratitude to Dr. Chang for her valuable help and support in my collaboration work. I also would like to thank my committee members: Dr. Anderson, Dr. Mi and Dr. Seymour for their suggestions and comments for my dissertation work. Finally, I would like to thank all my family and friends, especially my husband, for their constant support and encourage.

1.0 INTRODUCTION

In precision medicine, the prevention and treatment strategies are tailored according to individual characteristics. Such practice has been greatly improved by using information obtained from large databases (Council et al., 2011) including electronic health record (EHR) which contains patient information such as demographics, daily charts, medical history, lab results, medication use, billing information and others (Häyrinen et al., 2008). In order to efficiently process data and extract useful information, machine learning methods are often applied (Coorevits et al., 2013). Clustering is an important aspect of unsupervised machine learning methods which aims to uncover hidden patient subgroups that may have different diagnoses and treatment responses in EHR data. Further investigations on these subgroups together with current clinical guidelines could help design precision medicine strategies to further assist physicians in providing better patient care (Jensen et al., 2012).

The main challenge in applying clustering methods on clinical data is that most clinical data contain mixed types of variables (both continuous and categorical). However, most clustering methods can only handle single type of variables - either all continuous or all categorical. In addition, the increasing amount of collected clinical information motivates the need of distinguishing important variables, namely variable selection, for better interpretation. Therefore, the main objective of this dissertation is to develop clustering methods that can be used on mixed data and with the ability of variable selection.

In Chapter 2, we proposed a non-parametric method from a frequentist perspective; and in Chapter 3, we proposed a model-based method from a Bayesian perspective. Both are able to conduct variable selection and cluster mixed data. For the method proposed in Chapter 3, we took advantage of its parametric form and incorporated the ability of handling censored biomarkers, which is a common issue we may encounter in clinical data.

Chapter 2 introduces proposed hybrid density- and partition-based (HyDaP) clustering algorithm that is able to cluster mixed data as well as conduct variable selection. Simulations across different scenarios were conducted to compare its clustering performance with other existing methods. A real EHR data was also used to demonstrate the performance of the HyDaP algorithm. Chapter 3 introduces proposed Bayesian finite mixture model with variable selection that is able to cluster mixed data, conduct variable selection, and handle censored biomarker variables simultaneously. Simulation settings used in Chapter 2 were re-conducted to compare clustering performance of Bayesian finite mixture model with variable selection with other methods including the HyDaP algorithm. The EHR data was re-analyzed for performance evaluation. Chapter 4 concludes with discussions of proposed clustering methods and future work directions.

2.0 HYBRID DENSITY- AND PARTITION-BASED CLUSTERING ALGORITHM FOR DATA WITH MIXED-TYPE VARIABLES

2.1 INTRODUCTION

The basic concept of clustering is to divide individuals into a number of subgroups such that individuals within the same subgroup have more similar characteristics, as defined by a set of variables, than the individuals who belong to different subgroups. One of the main challenges in clustering is how to define "dissimilarity" between subjects with data of mixed variable types (continuous and categorical). If all variables are continuous, we can view the collection of information from an individual as a data point, or a vector of variables in a high-dimensional covariate space. The *distance* between the data points of two individuals is used to determine the *dissimilarity* between these two subjects so that a closer distance indicates lower dissimilarity. If all variables are categorical, *dissimilarity measures* (or *similarity measures*) were proposed to evaluate how often two individuals are in the same category among those variables. In this context we will use "distance" and "dissimilarity" interchangeably. Gower distance (Gower, 1971), distance defined in factorial analysis of mixed data (FAMD) (Pagès, 2014), and K-prototypes (Huang, 1998) are possible methods to address the above mentioned issue.

Gower distance was proposed to measure dissimilarity between subjects with mixed types of variables. The distance measure used in FAMD can be applied on mixed data as well, even though FAMD was not originally intended for clustering. Distance measure defined in K-prototypes is similar to Gower distance, but it incorporates a user-defined weight for each type of variables. Therefore, K-prototypes assumes that all categorical variables have the same weight, and that all continuous variables have the same weight. This design may not be practical if within the same variable type, some are clinically more important than others in terms of clustering.

Finite mixture model (FMM) (McCutcheon, 1987; Moustaki, 1996) is a model-based clustering method that bypasses the challenge of defining dissimilarity between subjects with mixed types of variables. It assumes that the data is a mixture of several parametric distributions. The unknown distributional parameters including cluster membership can be solved via maximizing likelihood using the expectation-maximization (EM) algorithm. Moreover, it is able to transfer the task of selecting the optimal number of clusters into model selection problem which is much more straightforward. However, its main drawback is that all the distributional assumptions are conditional on the unknown cluster, making those assumptions unverifiable from the data.

In order to identify cluster memberships, it is also important to know the underlying data structure. For example, whether distinct clusters exist in the feature space; or if no natural clusters exist but the data is heterogeneous enough to be partitioned. Such information is crucial in understanding data, selecting clustering methods, and interpreting clustering results. However, to our knowledge, none of the existing methods incorporates this data structure information into clustering.

To address the limitations of the existing methods, we propose a Hybrid Density- and Partition-based (HyDaP) algorithm to identify clusters for data with mixed types of variables and use this method to discover sepsis phenotypes using demographic and clinical data in EHR for sepsis patients at university affiliated hospitals.

In Section 2.2 we introduce the most commonly used dissimilarity measures and clustering algorithms; in Section 2.3 we define three data structures and propose a new clustering algorithm, HyDaP; in Section 2.4 we present performance comparisons among different methods under various simulation settings; in Section 2.5 we demonstrate the use of HyDaP algorithm to identify sepsis phenotypes; and Section 2.6 is discussion.

2.2 REVIEW OF DISSIMILARITY MEASURES AND CLUSTERING ALGORITHMS

In this section, we briefly review some existing dissimilarity measures and clustering algorithms. In addition, we discuss the pros and cons of each measure or algorithm.

2.2.1 Dissimilarity measures

Minkowski distance is a family of dissimilarity measures for numeric variables. Let $\mathbf{x}_i$ be a vector $(x_{i1}, x_{i2}, ..., x_{ip})^T$ representing p variables of subject i. For subjects i and i', Minkowski distance between the two is defined as follows:

$$d(\mathbf{x}_i, \mathbf{x}_{i'}) = \left(\sum_{j=1}^{p} |x_{ij} - x_{i'j}|^m \right)^{\frac{1}{m}}, m \geqslant 1$$

where m is related to the *shape of unit circle* which is a two-dimensional contour with every point on the contour at distance of 1 from the center $(0,0)$. Different choices of m lead to different distance measures. For example, $m = 2$ leads to the famous Euclidean distance which is intuitive and able to represent physical distances. When $m = 1$, we obtain Manhattan distance which is often used to detect hyperrectangular clusters. When $m \to \infty$, we obtain Chebyshev (maximum) distance which is the same as chess board distance since it is defined as the greatest value of the differences among all dimensions. A potential problem of using the Minkowski distance is that variables with larger variances tend to dominate the others (Xu and Wunsch, 2005; Shirkhorshidi et al., 2015), therefore, it is recommended to perform variable standardization (that is, rescale the variable by dividing by its standard deviation) before applying this measure.

Other dissimilarity measures for numeric variables include cosine similarity measure, Pearson correlation, Mahalanobis distance, to name a few. Cosine similarity measures the angle between two vectors regardless of vector magnitudes. It is usually applied if we are not interested in magnitudes, for example, for text mining as it captures text meanings instead of counting numbers (Xu and Wunsch, 2005; Han et al., 2011). Pearson correlation is usually used in clustering gene expression data (Xu and Wunsch, 2005), but it is sensitive to outliers. Mahalanobis distance is scale-invariant, and takes into account variable correlations.

When variables are all categorical, simple matching dissimilarity is usually used:

$$d(\mathbf{x}_i, \mathbf{x}_{i'}) = \sum_{j=1}^{p} \delta(x_{ij}, x_{i'j}),$$

where $\delta(x_{ij}, x_{i'j}) = I(x_{ij} \neq x_{i'j})$ indicating whether variable j are the same for individuals i and i'.

None of above-mentioned dissimilarity measures can be applied to mixed data. Gower distance was proposed to calculate the distance between subjects with mixed types of variables. Let $\mathbf{X}$ be a data matrix with $n \times p$ dimensions. Let the first h variables of $\mathbf{X}$ be continuous and the $(h+1)^{th}$ to p^{th} variables be multilevel categorical variables or symmetric binary variables. Let $\mathbf{X}_j$ be a vector $(x_{1j}, x_{2j}, ..., x_{nj})^T$ representing variable j. Gower distance between individuals i and i' is defined as:

$$d(\mathbf{x}_i, \mathbf{x}_{i'}) = \sum_{j=1}^{p} d_j(\mathbf{x}_i, \mathbf{x}_{i'}),$$

where

$$d_j(\mathbf{x}_i, \mathbf{x}_{i'}) = \begin{cases} \frac{|x_{ij} - x_{i'j}|}{max(\mathbf{X}_j) - min(\mathbf{X}_j)} & \text{if } j \in \{1, 2, ..., h\} \\ I(x_{ij} \neq x_{i'j}) & \text{if } j \in \{h+1, h+2, ..., p\}, \end{cases}$$

$$max(\mathbf{X}_j) = x_{i^\star j} \text{ if } x_{i^\star j} \geqslant x_{ij} \text{ for all } i,$$

$$min(\mathbf{X}_j) = x_{i^\star j} \text{ if } x_{i^\star j} \leqslant x_{ij} \text{ for all } i.$$

Gower distance for an *asymmetric* binary variable is calculated differently. Asymmetry occurs when similarity within one level is perceived to be higher compared to the other level. For example, breast cancer (yes/no) could be viewed as an asymmetric binary variable since individuals with breast cancer are much more similar than those without breast cancer (which could include men and women, adolescents and elder people). If variable j is an asymmetric binary variable, then the Gower distance between individuals i and i' with respect to this variable is defined as:

$$d_j(\mathbf{x}_i, \mathbf{x}_{i'}) = \begin{cases} 0 & \text{if } x_{ij} = x_{i'j} \text{ and they are the level with larger similarity} \\ 1 & \text{otherwise.} \end{cases}$$

In practice, there is one issue in applying Gower distance: as we will later show in simulations, Gower distance tends to give much larger weights to categorical variables than to continuous ones. This is because the distance due to a categorical variable is always 0 or 1, the minimum and the maximum of possible distance values, granting categorical variables more power in distinguishing subjects.

Another distance that could be used for mixed types of variables is the distance defined in FAMD:

$$d^2(\mathbf{x}_i, \mathbf{x}_{i'}) = \sum_{j=1}^{p} d_j^2(\mathbf{x}_i, \mathbf{x}_{i'}),$$

where

$$d_j^2(\mathbf{x}_i, \mathbf{x}_{i'}) = \begin{cases} (x_{ij} - x_{i'j})^2 & \text{if } j \in \{1, 2, ..., h\} \\ \sum_{l=1}^{C_j} \frac{1}{p_{jl}} (\frac{y_{ijl}}{p_{jl}} - \frac{y_{i'jl}}{p_{jl}})^2 & \text{if } j \in \{h+1, h+2, ..., p\}, \end{cases}$$

$$y_{ijl} = I(x_{ij} = L_{jl}), \sum_{l=1}^{C_j} y_{ijl} = 1; \; j \in \{h+1, h+2, ..., p\}$$

C_j is number of levels of categorical variable j; p_{jl} is proportion of l^{th} category of variable j; L_{jl} is l^{th} category of variable j.

2.2.2 K-means-based clustering algorithms

K-means (MacQueen et al., 1967) is the most well-known and applied clustering method in practice. The basic idea is to partition subjects with respect to minimizing the within-cluster sum of squares (WCSS). This algorithm is very efficient and has been the root of many later developed ones. It is usually used together with Euclidean distance. To cluster categorical data, K-modes (Huang, 1998) algorithm was developed by replacing Euclidean distance with simple matching dissimilarity measure, and replacing mean with mode to represent cluster centers.

To identify clusters with mixed types of variables, the partition around medoids (PAM) (Kaufman and Rousseeuw, 2009) has been proposed. PAM is a modification of K-means with a different definition of cluster centers. Unlike K-means which uses within-cluster mean to represent its centers, PAM uses *medoids* which are actual data points in the dataset. This makes defining centers of categorical variables possible. Moreover, medoids are analogous to medians and hence PAM is more robust to outliers. One drawback however is that PAM is computationally intensive and inefficient, making it less ideal for processing large data sets.

K-prototypes algorithm is another modified version of K-means with the ability of handling mixed types of variables. Its centers are called *prototypes*, which use within-cluster mean to represent continuous variables and mode for categorical variables. The distance between subjects i and i' is defined as:

$$d(\mathbf{x}_i, \mathbf{x}_{i'}) = \sum_{j=1}^{h} d_j(\mathbf{x}_i, \mathbf{x}_{i'}) + \gamma \sum_{j=h+1}^{p} d_j(\mathbf{x}_i, \mathbf{x}_{i'}),$$

where

$$d_j(\mathbf{x}_i, \mathbf{x}_{i'}) = \begin{cases} (x_{ij} - x_{i'j})^2 & \text{if } j \in \{1, 2, ..., h\} \\ I(x_{ij} \neq x_{i'j}) & \text{if } j \in \{h+1, h+2, ..., p\}, \end{cases}$$

and γ is a user-defined weight parameter for categorical variables. K-prototypes lacks flexibility in variables weights as it assumes equal importance for variables of the same type. Moreover, the tuning parameter γ is user-defined rather than data-driven.

2.2.3 Hierarchical clustering

Hierarchical clustering is another category of clustering methods. It first grows a dendrogram which is a tree-like diagram showing hierarchical structure of subjects and then cuts the dendrogram to obtain clusters. One advantage of hierarchical clustering is that the generated dendrogram is very informative and provides information of cluster structure besides cluster assignments. Its disadvantages include no global objective function, a greedy type of procedure, the sensitivity to outliers, and inefficient for large data sets.

2.2.4 Extended clustering framework

In many situations researchers are also often interested in variables' importance, not just cluster identification. Motivated by this interest, sparse clustering framework (Witten and Tibshirani, 2010) was proposed. It incorporates feature selection through a Lasso-type penalty, and adds variable weights to the objective function:

$$\text{Maximize} \sum_{j=1}^{p} w_j f_j(\mathbf{X}_j; \mathbf{\Theta})$$

with restriction $\mathbf{w}\|^2 \leqslant 1, \|\mathbf{w}\|_1 \leqslant s, w_j \geqslant 0 \ \forall j$. where n is number of subjects; p is number of features; $\mathbf{w} = (w_1, w_2, ..., w_p)^T$ is the weight vector; $\mathbf{\Theta}$ is a parameter vector restricted to lie in a set D; $f_j(\mathbf{X}_j; \mathbf{\Theta})$ is some function that involves feature j only; and s is a $L1$ norm restriction, which is a tuning parameter in the algorithm. We could plug in many algorithms like K-means, hierarchical clustering into this framework to obtain sparse version algorithms. One of the main attractions of sparse clustering is that it conducts data-driven variable selection and clustering simultaneously. However, the selection of tuning parameter s may not be straightforward.

Many partition-based algorithms require pre-specification of K, the optimal number of clusters, but how to choose it is another important question. The consensus clustering framework (Monti et al., 2003; Wilkerson and Hayes, 2010) can help determine number of clusters and obtain cluster memberships simultaneously. In addition, it can assess stability of discovered clusters. Consensus clustering incorporates results from multiple runs of an inner-loop clustering algorithm (e.g., K-means) on sub-sampled subjects. For each pair of subjects, a consensus index is obtained by calculating proportion of times the pair was assigned to the same cluster among times both pair members were sampled. The consensus index can then serve as a similarity measure and subjected to a hierarchical clustering algorithm to form final clusters. Choosing K is achieved by checking the consensus matrix heatmaps and cluster-consensus values. The number of clusters that yields the cleanest heatmap and highest cluster-consensus values is preferred.

2.2.5 Density-based clustering

Another important category of clustering methods is density-based clustering. All above-mentioned algorithms are distance-based methods which are more appropriate for detecting clusters that are convex shaped and with similar sizes and densities. If the underlying clusters have arbitrary shapes, density-based clustering algorithms may work better. Density-based spatial clustering of applications with noise (DBSCAN) (Ester et al., 1996) and ordering points to identify the clustering structure (OPTICS) (Ankerst et al., 1999) are two widely used density-based algorithms. DBSCAN does not need input of K, and it is robust to noise. However, it is not well suited for high dimensional data or for clusters with varying densities. OPTICS is an improved method which can detect clusters with varying densities while not being over sensitive to its user-specified tuning parameters.

2.2.6 Model-based clustering

FMM is a model-based clustering method assuming that the data is consists of K latent clusters. Its density function is defined as:

$$f(\mathbf{X}) = \sum_{k=1}^{K} \pi_k g_k(\mathbf{X}),$$

where π_k is the cluster mixture probability, $\sum_{k=1}^{K} \pi_k = 1$; and g_k is the conditional distribution given cluster k. For a sample of size n, the log-likelihood can be written as:

$$L = \sum_{i=1}^{n} \log f(\mathbf{x}_i) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k g_k(\mathbf{x}_i).$$

FMM assumes conditional independence given cluster k, that is, $g_k(\mathbf{x}_i) = \prod_{j=1}^{p} g_k(x_{ij})$, and the EM algorithm is usually used to obtain the MLE. The posterior probability of each subject belonging to each cluster can be calculated as:

$$\hat{p}(k|\mathbf{x}_i) = \frac{\hat{\pi_k}\hat{g_k}(\mathbf{x}_i)}{\sum_{k=1}^{K} \hat{\pi_k}\hat{g_k}(\mathbf{x}_i)}.$$

Subjects are then assigned to the cluster with which the posterior probability is the largest. These probabilities help discriminate core subjects (those with high probability of belonging to assigned cluster) and border subjects (those with low probability of belonging to assigned cluster) within each cluster. Given the parametric form of FMM, formal inference is possible. In addition, selecting the number of clusters becomes a model selection problem. The main drawback however is its unverifiable distributional assumptions; all the inferences are conducted conditional on unknown cluster assignments.

There are some other approaches to handle mixed types of variables. These include categorizing all continuous variables (Haripriya et al., 2015) or converting categorical variables into continuous or dummy variables and then treat the dummy variables as continuous (Hennig and Liao, 2013). However, both ideas will lead to information loss. Another common idea is to cluster continuous part of the data and categorical part separately. The final clusters are obtained by ensembling these two sets of clustering results (Reddy and Kavitha, 2012). This method impractically weigh continuous and categorical variables equally and ignore possible mutual influences between the two variable types.

2.3 PROPOSED HYBRID DENSITY- AND PARTITION-BASED CLUSTERING (HYDAP) ALGORITHM FOR MIXED DATA

To address the limitations of the existing clustering methods in handling data containing mixed types of variables, we propose a hybrid density- and partition-based clustering (HyDaP) algorithm which consists of a **pre-processing step** (step 1) and a **clustering step** (step 2). The pre-processing step identifies the data structure formed by continuous variables and recognizes the important variables for clustering. In the clustering step, our proposed dissimilarity measure is used to obtain a dissimilarity matrix, which can be fed into PAM to obtain the final results. We describe the HyDaP algorithm in detail below.

2.3.1 Pre-processing step (Step 1)

To help with variable selection and better understand the data set, we first define 3 data structures for the space spanned by the continuous variables as: *natural cluster structure* (data structure 1); *partitioned cluster structure* (data structure 2); and *homogeneous structure* (data structure 3). Once the data structure is known, we apply tailored variable selection procedures. At the end of the pre-processing step, a set of selected variables will proceed to the clustering step (step 2) (Figure 1).

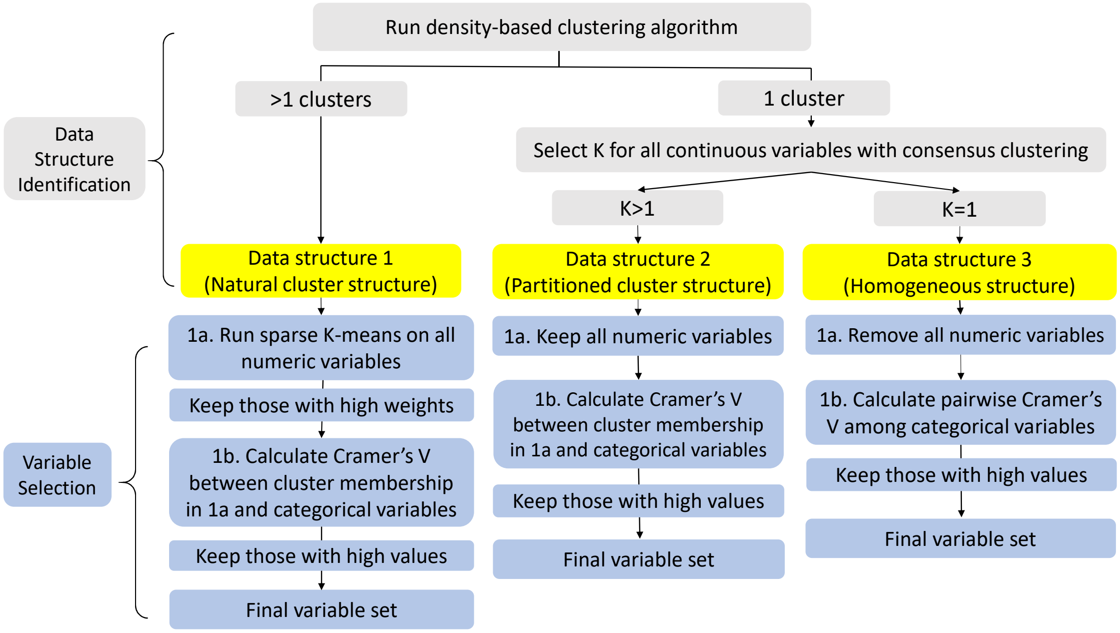

Run density-based clustering algorithm
>1 clusters
1 cluster
Data Structure Identification
Select K for all continuous variables with consensus clustering
K>1
K=1
Data structure 1 (Natural cluster structure)
Data structure 2 (Partitioned cluster structure)
Data structure 3 (Homogeneous structure)
1a. Run sparse K-means on all numeric variables
Keep those with high weights
1b. Calculate Cramer's V between cluster membership in 1a and categorical variables
Keep those with high values
Final variable set
1a. Keep all numeric variables
1b. Calculate Cramer's V between cluster membership in 1a and categorical variables
Keep those with high values
Final variable set
1a. Remove all numeric variables
1b. Calculate pairwise Cramer's V among categorical variables
Keep those with high values
Final variable set
Variable Selection

Figure 1: Flowchart of Step 1 of the HyDaP algorithm

2.3.1.1 Data structure identification Data spanned in the covariate space of continuous variables can be divided into two scenarios: with and without natural clusters. A hypothetical example of these two scenarios is depicted in Figure 2. We can observe that both Data 1 and Data 2 contain two variables, but natural clusters only exist in Data 1. Although this conclusion is straightforward for Data 1 and Data 2, when data is spanned in a high-dimensional space, it is impossible to visually examine existence of natural clusters. Therefore, we use a density-based clustering algorithm (e.g., OPTICS) and resulted reachability plot to help understand the spatial structure of the data. Reachability plot is a bar plot showing ordered reachability distances among subjects (Ankerst et al., 1999). A reachability plot provides an overall 2-dimensional spatial structure of a dataset regardless of its original dimensions. The horizontal axis of the plot is the processing order and the vertical axis is the reachability distance. Each trough on the reachability plot can be viewed as a single cluster. Edges between two side-by-side troughs represent the distance between two closest border points from the corresponding two clusters. Higher edges imply that the corresponding two clusters are farther apart while lower edges or unclear edges imply that clusters are not that distinct from each other.

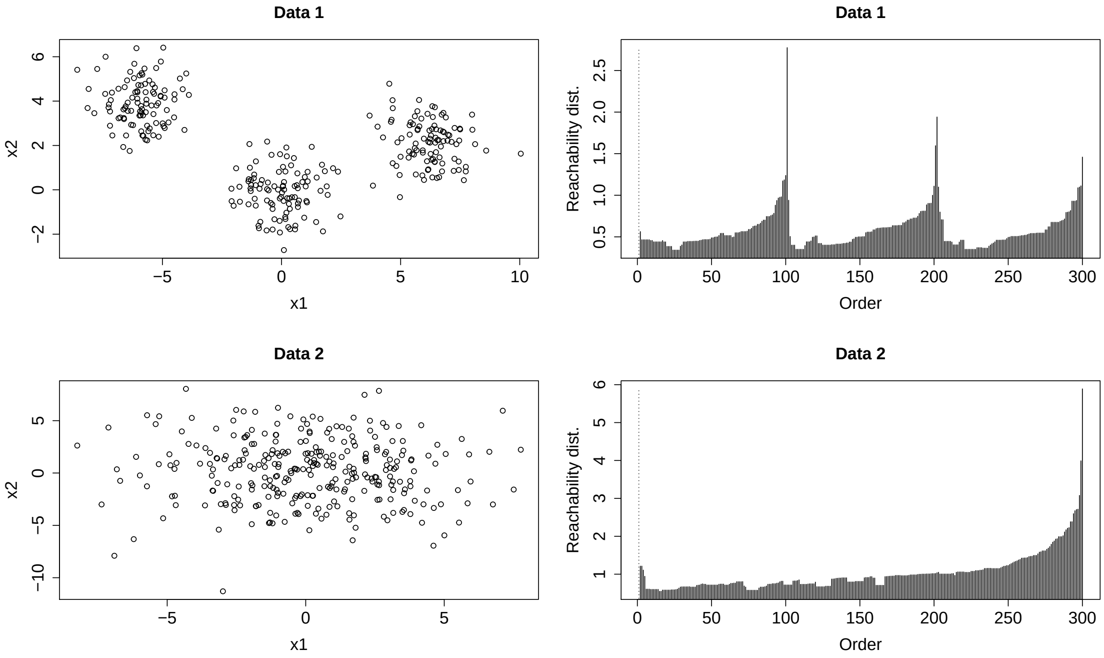


Figure 2: Illustration of different reachability plots

If we observe multiple troughs in a reachability plot, as illustrated in reachability plot of Data 1 in Figure 2, this indicates existence of distinct clusters, i.e., the corresponding dataset has natural clusters. We call this type of structure *natural cluster structure* (*data structure 1*) and aim to identify these distinct clusters. If we only observe one trough or no clear through in the reachability plot (e.g., reachability plot of Data 2 in Figure 2), this indicates that distinct clusters do not exist. Then we will investigate whether data points in the continuous covariate space are sufficiently heterogeneous to be further partitioned. We use consensus clustering framework for all continuous variables to access the possible heterogeneity by checking the selected optimal number of clusters. If we obtain ≥ 2 clusters in consensus clustering, this indicates that heterogeneity exists and we can obtain stable clusters through partitioning. We call this type of structure *partitioned cluster structure* (*data structure 2*). If the optimal number of clusters is one from the consensus clustering results, this indicates that continuous part of the data is highly homogeneous and cannot be further partitioned. We call this type of structure *homogeneous structure* (*data structure 3*).

2.3.1.2 Variable selection After identifying the data structure, we conduct data structure tailored variables selection.

Under the *natural cluster structure*, distinct clusters can be determined by continuous variables. Therefore, we would like to select those having high contributions. As shown in Figure 1, we apply sparse K-means on all continuous variables and keep those with high weights (suggestions of the weight threshold can be found in Section 2.3.3.2). Number of clusters under this structure can be determined by the number of troughs in the reachability plot. Next, we calculate Cramer's V between each categorical variable and the cluster membership obtained from sparse K-means. We will only select categorical variables with high Cramer's V values. Cramer's V has been used to measure the association between nominal variables. It ranges from 0 to 1. A larger number indicates a stronger association, vice versa. Unlike the p-value, Cramer's V is not affected by the sample size. Researchers suggested the use of 0.3 as the cutoff value, namely Cramer's V larger than 0.3 indicates a moderate to strong association.

Under the *partitioned cluster structure*, distinct clusters do not exist; however, covariate space of all continuous variables are sufficiently heterogeneous to be further partitioned. This structure indicates that all of the continuous variables together contribute to heterogeneity but none of them has the driving influence. Therefore, we keep all continuous variables and run consensus K-means to select the optimal number of clusters. Next, we calculate Cramer's V between each categorical variable and the cluster membership obtained from consensus K-means. We will only select categorical variables with high Cramer's V values.

Under the *homogeneous structure*, no distinct cluster exists and we are not able to further partition continuous covariate space into ≥ 2 homogeneous subgroups. Therefore, we dropped all continuous variables as they are non-distinguishable across clusters. Next, we calculate pairwise Cramer's V values among categorical variables and only select pairs with high Cramer's V values.

2.3.2 Clustering step (Step 2)

After variables with high contributions are selected, we proceed to the final clustering step. This step is the same across all data structures. We calculate the dissimilarities between subjects using our proposed dissimilarity measure, a modified version of the Gower distance. Assume that the first h variables are continuous and the rest are categorical. Our proposed dissimilarity between subjects i and i' is defined as:

$$d(\mathbf{x}_i, \mathbf{x}_{i'}) = \sum_{j=1}^{p} \frac{d_j(\mathbf{x}_i, \mathbf{x}_{i'})}{\sum_{o \neq o'} d_j(\mathbf{x}_o, \mathbf{x}_{o'})},$$

where

$$d_j(\mathbf{x}_i, \mathbf{x}_{i'}) = \begin{cases} \frac{|x_{ij} - x_{i'j}|}{max(\mathbf{X}_j) - min(\mathbf{X}_j)} & \text{if } j \in \{1, 2, \ldots, h\} \\ I(x_{ij} \neq x_{i'j}) & \text{if } j \in \{h+1, h+2, \ldots, p\} \end{cases}$$

$$max(\mathbf{X}_j) = x_{i^\star j} \text{ if } x_{i^\star j} \geqslant x_{ij} \text{ for all } i,$$

$$min(\mathbf{X}_j) = x_{i^\star j} \text{ if } x_{i^\star j} \leqslant x_{ij} \text{ for all } i.$$

Our modification is based on the idea of standardization to avoid variables with high variability be extremely influential to clustering results. It is motivated by the definition of Gower distance for categorical variables as they receive extreme dissimilarity values 0 or 1, which could exhibit high variability. This allows them to exert greater influence in the clustering results even if they are less informative than the continuous ones.

Below we show how our modification on dissimilarities is analogous to the standardization on continuous variables. Standardized squared Euclidean distance between subjects i and i' with respect to a continuous variable j is:

$$d_j^2(\mathbf{x}_i, \mathbf{x}_{i'}) = \left\{ \frac{x_{ij}}{sd(\mathbf{X}_j)} - \frac{x_{i'j}}{sd(\mathbf{X}_j)} \right\}^2$$

which can be re-written as:

$$d_j^2(\mathbf{x}_i, \mathbf{x}_{i'}) = \frac{(x_{ij} - x_{i'j})^2}{n^{-2}\sum_{o \neq o'}(x_{oj} - x_{o'j})^2},$$

where the numerator is the original squared Euclidean distance, the denominator is proportional to the sum of all pairwise distances. We adopt this idea to standardize the Gower distance, namely we divide the original Gower distance of variable j by sum of all pairwise Gower distance of variable j as shown above.

If after the pre-processing step all selected variables are continuous, we can just apply usual clustering methods to obtain the final clustering results.

2.3.3 Parameters selection

In this section we provide general suggestions on the selection of (1) the optimal number of clusters; (2) continuous variables under *natural cluster structure*.

2.3.3.1 Number of clusters Under the *natural cluster structure*, the number of clusters can be decided by the number of troughs in the reachability plot. Under the *partitioned cluster structure*, the number of clusters can be selected from the results of the consensus clustering. Under the *homogeneous structure*, we only select categorical variables in determining cluster membership. Hence we suggest constructing a dissimilarity matrix using our proposed dissimilarity measure and then plot the number of clusters against the corresponding within-cluster sum of dissimilarities. In this plot, we look for an *elbow* for the optimal number of clusters.

2.3.3.2 Selecting continuous variables under the natural cluster structure Selection of the continuous variables with high weights under the *natural cluster structure* could be subjective because of the choice of the weight threshold. We suggest applying sparse K-means for continuous part of each bootstrapping data set and then calculate the between-cluster sum of squares (BCSS). We then order these variables by their median BCSS from the smallest to the largest and plot the median (with 2.5^{th} quantile and 97.5^{th} quantile interval) of BCSS. Then we drop variables whose BCSS values are small or far away from the others. Our suggestion here is a heuristic one. Users can always incorporate other information and make their own judgements.

Table 1: Simulation settings

Variable	Cluster[a]	Sim 1(a)	Sim 1(b)	Sim 2(a)	Sim 2(b)	Sim 3
x_1	1	$\boldsymbol{N(-2,2)}^{b}$	$\boldsymbol{N(-2,2)}$	$\boldsymbol{N(-2,2)}$	$\boldsymbol{Beta(0.1,5)}$	$N(0,0.5)$
	2	$\boldsymbol{N(2,2)}$	$\boldsymbol{N(-1,2)}$	$\boldsymbol{N(2,2)}$	$\boldsymbol{Beta(0.1,5)+0.3}$	
	3	$\boldsymbol{N(6,2)}$	$\boldsymbol{N(0,2)}$	$\boldsymbol{N(6,2)}$	$\boldsymbol{Beta(0.1,5)+0.5}$	
x_2	1	$\boldsymbol{N(20,1)}$	$\boldsymbol{N(20,1)}$	$\boldsymbol{N(20,1)}$	$\boldsymbol{Beta(0.2,5)}$	$N(-3,1)$
	2	$\boldsymbol{N(25,1)}$	$\boldsymbol{N(24,1)}$	$\boldsymbol{N(25,1)}$	$\boldsymbol{Beta(0.1,5)+0.3}$	
	3	$\boldsymbol{N(18,1)}$	$\boldsymbol{N(21,1)}$	$\boldsymbol{N(18,1)}$	$\boldsymbol{Beta(0.1,5)+0.5}$	
x_3	1	$\boldsymbol{N(0,1)}$	$\boldsymbol{N(5,1)}$	$\boldsymbol{N(0,1)}$	$\boldsymbol{Beta(0.2,3)}$	$N(4,2)$
	2	$\boldsymbol{N(-7,1)}$	$\boldsymbol{N(8,1)}$	$\boldsymbol{N(-7,1)}$	$\boldsymbol{Beta(0.2,3)+0.3}$	
	3	$\boldsymbol{N(4,1)}$	$\boldsymbol{N(7,1)}$	$\boldsymbol{N(4,1)}$	$\boldsymbol{Beta(0.2,3)+0.5}$	
x_4	1	$N(0,1)$	$N(0,1)$	$N(0,1)$	$\boldsymbol{Beta(0.1,3)}$	$N(0,1)$
	2				$\boldsymbol{Beta(0.1,3)+0.3}$	
	3				$\boldsymbol{Beta(0.2,3)+0.5}$	
x_5	1	$\boldsymbol{M(0.1,0.1,0.8)}$	$N(40,1)$	$M(0.3,0.3,0.4)$	$N(0,0.01)$	$\boldsymbol{M(0.05,0.05,0.9)}$
	2	$\boldsymbol{M(0.1,0.8,0.1)}$		$M(0.3,0.3,0.4)$		$\boldsymbol{M(0.05,0.9,0.05)}$
	3	$\boldsymbol{M(0.8,0.1,0.1)}$		$M(0.4,0.3,0.3)$		$\boldsymbol{M(0.9,0.05,0.05)}$
x_6	1		$\boldsymbol{N(-1,1)}$		$M(0.3,0.3,0.4)$	$M(0.3,0.3,0.4)$
	2		$\boldsymbol{N(1,1)}$		$M(0.4,0.3,0.3)$	$M(0.4,0.3,0.3)$
	3		$\boldsymbol{N(-2,1)}$		$M(0.3,0.4,0.3)$	$M(0.3,0.4,0.3)$
x_7	1		$\boldsymbol{N(0,1)}$		$M(0.3,0.3,0.4)$	$\boldsymbol{M(0.9,0.05,0.05)}$
	2		$\boldsymbol{N(-1,1)}$			$\boldsymbol{M(0.05,0.9,0.05)}$
	3		$\boldsymbol{N(2,1)}$			$\boldsymbol{M(0.05,0.05,0.9)}$
x_8	1		$\boldsymbol{N(2,1)}$		$M(0.3,0.3,0.4)$	
	2		$\boldsymbol{N(1,1)}$			
	3		$\boldsymbol{N(0,1)}$			
$x_9 \sim x_{11}$	1,2,3		$N(0,1)$			
x_{12}	1		$M(0.3,0.3,0.4)$			
	2		$M(0.4,0.3,0.3)$			
	3		$M(0.3,0.4,0.3)$			
x_{13}	1		$\boldsymbol{M(0.9,0.05,0.05)}$			
	2		$\boldsymbol{M(0.05,0.9,0.05)}$			
	3		$\boldsymbol{M(0.05,0.05,0.9)}$			
x_{14}	1		$\boldsymbol{M(0.05,0.05,0.9)}$			
	2		$\boldsymbol{M(0.05,0.9,0.05)}$			
	3		$\boldsymbol{M(0.9,0.05,0.05)}$			

[a] Sample sizes for 3 clusters are 40, 40 and 120; [b] variables with bolded distributions are truly important in clustering

2.4 SIMULATION STUDIES

In this section we use simulations to evaluate the performance of the HyDaP algorithm relative to the existing approaches. Assuming that there are 3 underlying true clusters with cluster sizes of 40, 40, and 120. In terms of variable importance, we considered scenarios (1) both variable types contribute to clustering, (2) only continuous variables contribute to clustering, and (3) only categorical variables contribute to clustering. In terms of data structures, all 3 data structures were covered in simulations. Details of the distributions and parameters used in these simulation settings are shown in Table 1.

For each setting, 500 datasets were generated. Cluster analysis was performed on each dataset using the proposed HyDaP algorithm. We compared its performance with PAM with Gower distance, K-prototypes, FMM, and PAM with FAMD distance. Since we know the true cluster labels, the adjusted rand index (ARI) was calculated and used to evaluate the performances of different methods. ARI is used to measure the agreement between two nominal variables. Its largest value is 1 indicating perfect agreement and its smallest value is close to 0 indicating no agreement. For the purpose of evaluating clustering performance in simulations, higher ARI values indicate better agreement with true cluster labels and hence better performance. The reachability plot for each setting is illustrated in Figure 3. Table 2 summarizes the results of the pre-processing step of the HyDaP algorithm. The clustering performance with respect to ARI across all simulation settings is shown in Table 3. To examine the impact of conditional correlation on clustering performance, each simulation setting was imbued with a pairwise correlation of 0.4 conditional on true cluster labels. Results are shown in Table 5. Median along with the 2.5^{th} and 97.5^{th} percentiles were reported for all statistics.

2.4.1 Setting 1: Both types of variables contribute to clustering

2.4.1.1 Natural cluster structure In simulation 1(a), we simulated a total of 5 variables: 4 continuous and 1 categorical. All except one continuous variable truly contribute to clustering. The sole categorical variable also contributes to clustering.

In Step 1 of the HyDaP algorithm, the reachability plot (Figure 3a) indicated 3 clusters. Therefore, this setting has natural cluster structure as 3 distinct clusters exist. Table 2 shows the very low contribution of x_4 from the sparse K-means and the strong association between x_5 and the clusters identified by the sparse K-means. We dropped x_4 and kept all the others.

In Step 2, we applied PAM along with the proposed dissimilarity measure on the selected variables from Step 1: x_1, x_2, x_3, and x_5.

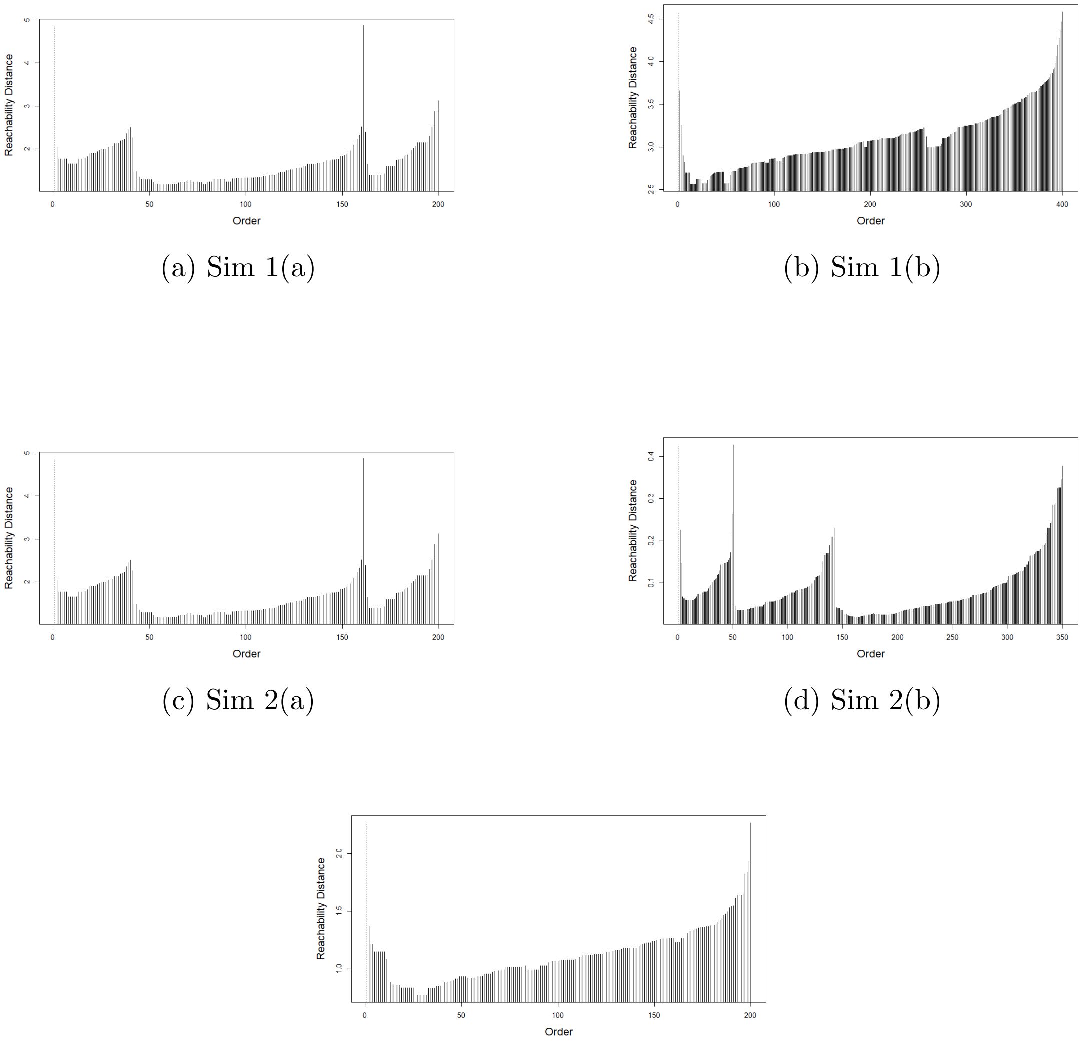
As shown in Table 3, HyDaP algorithm performed very well (ARI: 0.97 [0.92, 1.00]). Although K-prototypes (ARI: 1.00 [0.96, 1.00]) and FMM (ARI: 1.00 [0.98, 1.00]) both performed slightly better, our HyDaP algorithm was able to identify important variables. PAM with Gower distance (ARI: 0.70 [0.58, 0.80]) and PAM with FAMD distance (ARI: 0.78 [0.66, 0.89]) performed poorly. This is not surprising of the results using the Gower distance since it tends to downplay contributions of continuous variables, although in this setting continuous variables x_1 to x_3 all have large contributions to clustering.

2.4.1.2 Partitioned cluster structure

In simulation 1(b), we simulated a total of 14 variables: 11 continuous and 3 categorical. Six out of eleven continuous variables truly contribute to clustering; two out of three categorical variables contribute to clustering.

In Step 1 of the HyDaP algorithm, Figure 3b indicated that no natural clusters exists. After conducting consensus K-means, we chose 3 as the optimal number of clusters as its corresponding cluster-consensus values were the largest. Thus, a partitioned cluster structure was identified. All continuous variables were retained for the next step. Variable x_{12} was dropped because of its small Cramer's V with cluster assignments obtained in consensus K-means.

In Step 2, PAM with proposed dissimilarity measure was applied on x_1, x_2,..., x_{11}, x_{13}, and x_{14} to obtain final results.



(a) Sim 1(a)

(b) Sim 1(b)

(c) Sim 2(a)

(d) Sim 2(b)

(e) Sim 3

Figure 3: Reachability plots in different simulation settings

Table 2: Results from pre-processing step in different simulation settings

Sim 1(a)	Sim 1(b)	Sim 2(a)	Sim 2(b)	Sim 3
		Data Structure		
1	2	1	1	3
Weight	-	Weight	Weight	-
x_1: 0.49 (0.46, 0.52)		x_1: 0.49 (0.46, 0.52)	x_1: 0.54 (0.51, 0.57)	
x_2: 0.59 (0.58, 0.61)	Keep all continuous variables	x_2: 0.59 (0.58, 0.61)	x_2: 0.51 (0.48, 0.54)	Drop all continuous variables
x_3: 0.64 (0.62, 0.65)		x_3: 0.64 (0.62, 0.65)	x_3: 0.44 (0.38, 0.48)	
x_4: 0.00 (0.00, 0.02)		x_4: 0.00 (0.00, 0.02)	x_4: 0.50 (0.45, 0.54)	
			x_5: 0.00 (0.00, 0.02)	
Cramer's V	Cramer's V	Cramer's V	Cramer's V	Pairwise Cramer's V
x_5: 0.66 (0.57, 0.75)	x_{12}: 0.12 (0.05, 0.21)	x_5: 0.13 (0.06, 0.21)	x_6: 0.12 (0.05, 0.21)	x_5 x_6: 0.12 (0.04, 0.20)
	x_{13}: 0.77 (0.69, 0.85)		x_7: 0.09 (0.04, 0.17)	x_5 x_7: 0.69 (0.60, 0.77)
	x_{14}: 0.78 (0.69, 0.86)		x_8: 0.09 (0.04, 0.17)	x_6 x_7: 0.12 (0.04, 0.19)

[a]all weights and cramer's v values are presented in the form of median (2.5th percentile, 97.5th percentile).

Performance of the HyDaP algorithm is satisfactory (ARI 0.95 [0.87, 1.00]). Although it was unable to eliminate continuous variables that are purely noise, the HyDaP algorithm revealed that no continuous variable has driving effect but all of them together lead to heterogeneity in the feature space spanned by all of these continuous variables. In this setting, K-prototypes (ARI: 0.93 [0.79, 1.00]) and PAM with FAMD distance (ARI: 0.93 [0.84, 0.98]) also worked well while performance of FMM varied widely from sample to sample (ARI: 0.98 [0.44, 1.00]). PAM with Gower distance did not perform as well as others (ARI: 0.87 [0.76, 0.96]). This is because a noise categorical variable x_{12} was included and Gower distance tends to amplify its contribution.

2.4.2 Setting 2: Only continuous variables contribute to clustering

2.4.2.1 Natural cluster structure In simulation 2(a), we simulated a total of 5 variables: 4 continuous and 1 categorical. This setting is the same as simulation 1(a) except that the sole categorical variable does not contribute to clustering.

In Step 1 of the HyDaP algorithm, x_4 was dropped due to its low contribution in the sparse K-means. Table 2 shows a weak association between the categorical variable x_5 and clusters identified by the sparse K-means.

In Step 2, we applied the sparse K-means on x_1, x_2, and x_3 as they are all continuous variables.

In this setting, the HyDaP algorithm (ARI: 0.98 [0.94, 1.00]) and K-prototypes (ARI: 0.98 [0.92, 1.00]) both worked well. There were a few simulation runs the performance of FMM was not satisfactory (ARI: 1.00 [0.56, 1.00]). PAM with Gower distance (ARI: 0.01 [-0.01, 0.04]) and PAM with FAMD distance (ARI: 0.09 [-0.01, 0.44]) performed extremely poor. As mentioned in simulation 1(b), Gower distance tends to amplify the contributions of the categorical variables. Meanwhile, FAMD was not originally designed for clustering.

2.4.2.2 Natural cluster structure In simulation 2(b), we simulated a total of 8 variables: 5 continuous and 3 categorical. Four out of five continuous variables truly contribute to clustering and follow highly skewed distributions. None of the categorical variables contributes to clustering.

In Step 1 of the HyDaP algorithm, Figure 3d shows 3 distinct clusters and hence this setting was identified as natural cluster structure. We dropped x_5 because of its small contribution to clustering as shown in Table 2. All categorical variables were dropped as well given their weak associations with clusters obtained in the sparse K-means.

In Step 2, we applied the sparse K-means on x_1, x_2, and x_3 since they are all continuous variables.

In this setting, the HyDaP algorithm performed the best (ARI: 0.98 [0.92, 1.00]). PAM with Gower distance (ARI: 0.23 [0.00, 0.34]), K-prototypes (ARI: 0.58 [0.38, 0.99]), FMM (ARI: 0.41 [0.33, 0.58]), and PAM with FAMD distance (ARI: 0.34 [0.08, 0.39]) all performed poorly. This was expected for FMM because most of the continuous variables were not normally distributed conditional on the true cluster labels.

2.4.3 Setting 3: only categorical variables contribute to clustering

2.4.3.1 Homogeneous structure In simulation 3, we simulated a total of 7 variables: 4 continuous and 3 categorical. None of the continuous variables truly contributes to clustering. Two out of three categorical variables contribute to clustering.

Table 3: Performance comparison under different simulation settings

Clustering Method	**ARI, median (2.5th percentile, 97.5th percentile)**				
	Sim 1(a)	Sim 1(b)	Sim 2(a)	Sim 2(b)	Sim 3
HyDaP	0.97 (0.92, 1.00)	0.95 (0.87, 1.00)	0.98 (0.94, 1.00)	0.98 (0.92, 1.00)	0.75 (0.63, 0.85)
PAM + Gower distance	0.70 (0.58, 0.80)	0.87 (0.76, 0.96)	0.01 (-0.01, 0.04)	0.23 (0.00, 0.34)	0.71 (0.31, 0.84)
K-prototypes	1.00 (0.96, 1.00)	0.93 (0.79, 0.95)	0.98 (0.92, 1.00)	0.58 (0.38, 0.99)	0.17 (-0.01, 0.26)
Finite mixture model	1.00 (0.98, 1.00)	0.98 (0.44, 1.00)	1.00 (0.56, 1.00)	0.41 (0.33, 0.58)	0.72 (0.56, 0.85)
PAM + FAMD distance	0.78 (0.66, 0.89)	0.93 (0.84, 0.98)	0.09 (-0.01, 0.44)	0.34 (0.08, 0.39)	0.73 (0.22, 0.84)

In Step 1 of the HyDaP algorithm, Figure 3e indicates no natural clusters exist. After conducting consensus K-means, the optimal number of clusters chosen was 1 because cluster-consensus values were low for all numbers of clusters. Hence this was identified as homogeneous structure. All continuous variables were dropped but categorical variables x_5 and x_7 were kept due to their strong association with each other as shown in Table 2.

In Step 2, PAM with proposed dissimilarity measure was applied on x_5 and x_7.

In this setting, the HyDaP algorithm performed the best (ARI: 0.75 [0.63, 0.85]) and K-prototypes did the worst (ARI: 0.17 [-0.01, 0.26]). Performance of PAM with Gower distance (ARI: 0.71 [0.31, 0.84]), FMM (ARI: 0.72 [0.56, 0.85]) and PAM with FAMD distance (ARI: 0.73 [0.22, 0.84]) were similar.

2.4.4 Standardizing continuous variables

To assess if pre-processing the data could change the performance of the existing methods, we reanalyzed the datasets generated from all 5 simulation settings above. We first standardized all continuous variables to a have mean 0 and variance 1 prior to subjecting them to the clustering algorithms. There was minimal change in the clustering performance (Table 4) as compared to no standardization (Table 3). This was expected since the main challenge in clustering mixed data is to balance the contribution between continuous and categorical variables. Standardizing continuous variables alone was not expected to improve existing methods. Note HyDaP algorithm was not applied to the standardized datasets so the results in both tables were the same. In addition, the results for PAM with FAMD distance are the same in both tables standardization was performed in both as recommended in the literature (Pagès, 2014).

Table 4: Performance comparison with standardizing continuous variables

Clustering Method	**ARI, median (2.5th percentile, 97.5th percentile)**				
	Sim 1(a)	Sim 1(b)	Sim 2(a)	Sim 2(b)	Sim 3
HyDaP	0.97 (0.92, 1.00)	0.95 (0.87, 1.00)	0.98 (0.94, 1.00)	0.98 (0.92, 1.00)	0.75 (0.63, 0.85)
PAM + Gower distance	0.70 (0.58, 0.80)	0.87 (0.76, 0.96)	0.01 (-0.01, 0.04)	0.23 (0.00, 0.34)	0.73 (0.28, 0.84)
K-prototypes	0.98 (0.95, 1.00)	0.97 (0.92, 1.00)	0.97 (0.89, 1.00)	0.36 (0.33, 0.67)	0.19 (0.00, 0.31)
Finite mixture model	1.00 (0.56, 1.00)	0.98 (0.44, 1.00)	1.00 (0.98, 1.00)	0.41 (0.33, 0.58)	0.72 (0.56, 0.85)
PAM + FAMD distance	0.78 (0.66, 0.89)	0.93 (0.84, 0.98)	0.09 (-0.01, 0.44)	0.34 (0.08, 0.39)	0.73 (0.22, 0.84)

2.4.5 Variables are conditionally correlated

To assess the impact of within-cluster correlation, simulations for each of the 5 settings above was repeated with pairwise correlation of 0.4 for all continuous variables conditional on true cluster labels. As summarized in Table 5, within-cluster correlation had little to no impact on the performance of the HyDaP algorithm, PAM with Gower distance, K-prototypes, and PAM with FAMD distance. In some situations, it led to worse performance of FMM. This is expected since FMM assumes conditional independency, namely all variables are independent with each other conditional on clusters labels. However, we did observe that in simulation 3 when none of the continuous variables contributes to clustering, the optimal number of clusters selected by the consensus K-means was 2 instead of 3 (figures not shown here). This is understandable since all pairs of continuous variables are correlated given true cluster labels, therefore, they share a lot of common information. To some extent we can use only one of them without losing much information as all others as redundant. For any single continuous variable we can potentially divide it into 2 subgroups that have some differences. But this does not essentially mean these 2 subgroups can be viewed as 2 clusters. Therefore, if we observe that 2 is the optimal number of clusters in consensus clustering results and most pairs of continuous variables have high conditional correlations, we should be cautious. One suggestion is that we could try to look for continuous variables that have similar clinical meanings e.g., Aspartate Aminotransferase (AST) and Alanine Aminotransferase (ALT), since these variables are very likely to have high correlations within clusters. For these variables we can only keep one of them in clustering to avoid such situation.

Table 5: Performance comparison with existence of correlation

Clustering Method	**ARI, median (2.5th percentile, 97.5th percentile)**				
	Sim 1(a)	Sim 1(b)	Sim 2(a)	Sim 2(b)	Sim 3
HyDaP	0.97 (0.92, 1.00)	0.94 (0.33, 1.00)	1.00 (0.95, 1.00)	1.00 (0.97, 1.00)	0.74 (0.61, 0.83)
PAM + Gower distance	0.71 (0.59, 0.80)	0.87 (0.77, 0.95)	0.01 (-0.01, 0.04)	0.33 (0.21, 0.38)	0.73 (0.28, 0.83)
K-prototypes	0.98 (0.95, 1.00)	0.62 (0.32, 0.98)	0.97 (0.89, 1.00)	0.99 (0.95, 1.00)	0.13 (-0.01, 0.23)
Finite mixture model	1.00 (0.46, 1.00)	0.47 (0.36, 1.00)	1.00 (0.97, 1.00)	0.58 (0.46, 1.00)	0.29 (0.14, 0.82)
PAM + FAMD distance	0.79 (0.67, 0.88)	0.93 (0.34, 0.98)	0.34 (-0.01, 0.44)	0.35 (0.22, 0.41)	0.72 (0.22, 0.84)

2.4.6 Simulation summary

From the simulation studies, we found that our proposed HyDaP algorithm was consistently the top or one of the top performers across all simulation settings. Moreover, we found that (1) when categorical variables do not contribute much to clustering, PAM with Gower distance performed poorly; (2) when continuous variables follow arbitrary distributions, FMM may not perform well due to assumption violation; (3) when none of continuous variables contributes to clustering, K-prototypes may fail; (4) performance of PAM with FAMD distance was not stable across different scenarios as its distance measure is not specifically designed for clustering.

In terms of impact of standardizing continuous variables and conditional correlation, we observed that: (1) standardizing continuous variables before clustering will not affect clustering performance of any methods; (2) with existence of conditional correlation, most methods including our HyDaP algorithm received little or no impact on clustering performance. However, FMM did have worse results under many scenarios. This is expected since conditional correlation violates the assumption of conditional independency FMM assumes.

2.5 REAL DATA APPLICATION

We used the EHR data collected from the Sepsis ENdotyping in Emergency CAre (SENECA) project to demonstrate the use of our proposed HyDaP algorithm for identifying phenotypes in patients with sepsis. The SENECA data contains 20,189 sepsis encounters collected from 12 University of Pittsburgh Medical Center (UPMC) healthcare systems from year 2010 to 2012. The goal is to explore whether clinical sepsis phenotypes are identifiable for a patient that presents at the emergency department and whether the phenotypes are associated with various clinical endpoints. The objectives of the analysis are to select the most important variables among 30 demographic and clinical variables, and to identify homogeneous clusters (phenotypes). Twenty eight variables were continuous and 2 were categorical. Although we do not have much information about the optimal number of clusters for the data set, our clinician colleagues suggested that larger numbers of clusters are preferred.

Data structure identification: The reachability plot in Figure 4 indicates that there is no natural clusters in the SENECA data. Unlike the genetic data, we rarely observe natural clusters in data collected from clinical settings. We then performed the consensus K-means for all continuous variables, and the results are depicted in Figure 5 suggesting that the optimal number of clusters is 4, which indicates that the data structure of the SENECA data belongs to *partitioned cluster structure*.

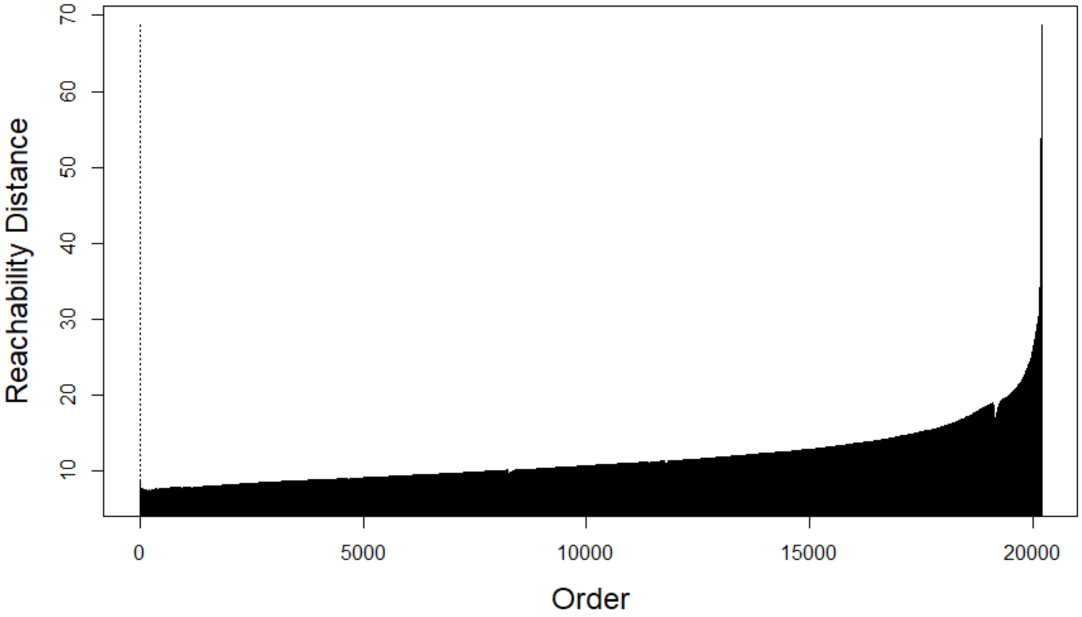


Figure 4: OPTICS reachability plot for the SENECA data

Variable selection: Under *partitioned cluster structure* we kept all continuous variables. For categorical variables, Cramer's V is 0.05 between gender and cluster membership from consensus clustering and it is also 0.05 between race and cluster membership. Therefore, we dropped gender and race before proceeding to the final clustering step.

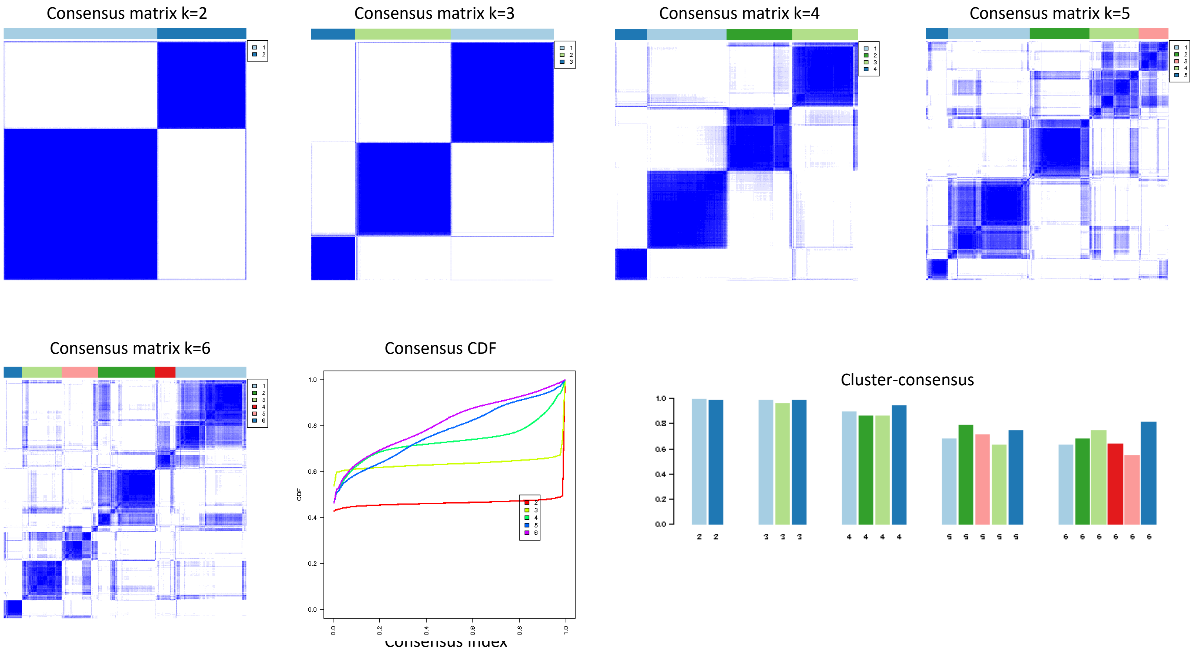


Figure 5: Consensus K-means results of SENECA data

Clustering step: All categorical variables were excluded after the pre-processing step, so we took the results from the consensus K-means as our final clustering results. In terms of variable importance, all continuous variables together had contributions to the obtained partitions but none of them showed dominant impact. Neither gender nor race were important clustering variables. We obtained 4 clusters with relatively balanced sample sizes: $6,625$, $5,512$, $5,385$, and $2,667$. Within each cluster, distributions of some important clinical endpoints are shown in top left plot of Figure 6. We can observe that Cluster 1 has the lowest proportion for all clinical endpoints while Cluster 2 has the second lowest ones. Cluster 4 has the highest proportions. With our clinician colleagues, we examined patient characteristics of the resulting clusters. We observed that sepsis patients in Cluster 1 had fewer other health issues; patients in Cluster 2 were those who were older, had multi morbidities, and renal dysfunctions; patients in Cluster 3 were those who had more inflammations and pulmonary dysfunctions; and patients in Cluster 4 were whose who had more acidosis, liver, and cardiovascular dysfunctions.

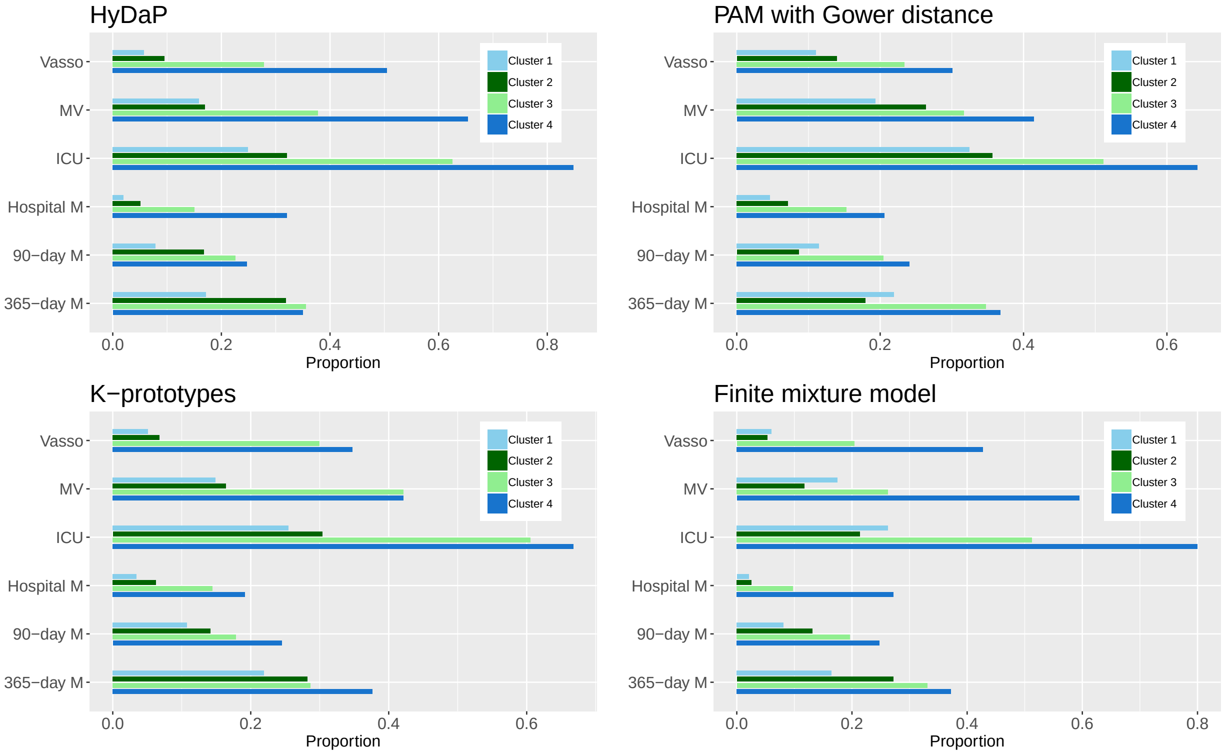


Figure 6: Clinical endpoints across 4 clusters identified by different methods

For comparison, we applied PAM with Gower distance, K-prototypes, and FMM to obtain cluster memberships assuming 4 clusters. The results are summarized in Figure 6. For PAM with Gower distance, we took a random sample of the whole SENECA data with size $5,000$ because the computation time of this algorithm was very long. After further exploration we found that gender dominated the clustering result as the proportion of male is 0.0% in Cluster 1, 2.7% in Cluster 4, 99.4% in Cluster 2, and 99.8% in Cluster 3. Note that gender was not relevant based on our proposed HyDaP algorithm. For the K-prototypes, we found that the 4 clusters obtained were not that distinct from each other in terms of the distribution of clinical endpoints. The 4 clusters obtained from the FMM appeared to be distinct from each other and similar to what we observed in HyDaP algorithm. However, Cluster 1 has larger proportion of patients admitted to ICU, use of mechanical ventilation and vasopressor compared to Cluster 2, but it has lower mortality rate.

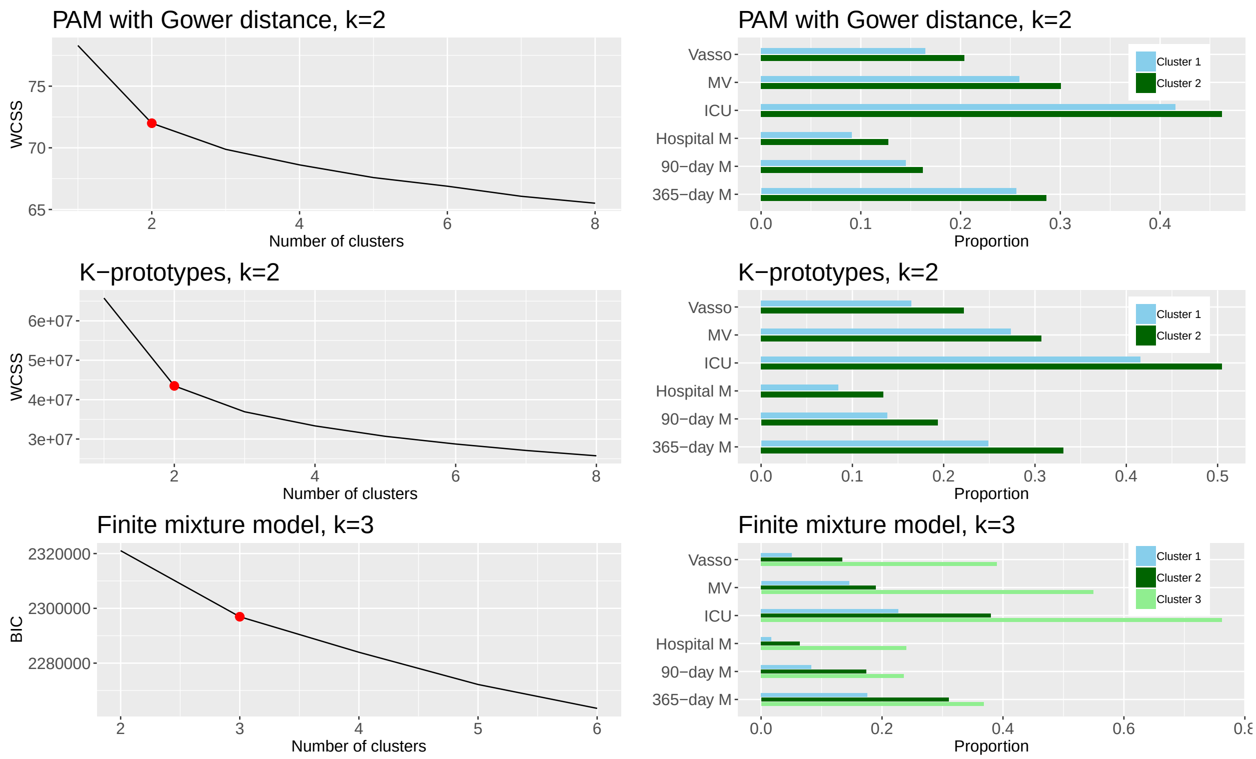


Figure 7: Clinical endpoints across clusters with the optimal number of clusters

Next, we re-applied the existing methods by first selecting the method-specific optimal number of clusters. The number of clusters versus the total WCSS or BIC values are shown in left column of Figure 7. We found that the optimal number of clusters was 2 for PAM with Gower distance and for K-prototypes, and 3 for FMM. We once again observed that the clustering results were dominated by gender when using PAM with Gower distance. The proportion of men was 1.2% in Cluster 1 and 98.8% in Cluster 2. The two clusters were quite similar in terms of clinical endpoints. Similarly, the 2 clusters identified by K-prototypes were not quite distinct in terms of clinical endpoints as well. The FMM identified 3 clusters with quite different distribution of clinical endpoints, but the HyDaP algorithm was able to identify one more cluster with distinct clinical features.

2.6 DISCUSSION

We proposed a hybrid density- and partition-based clustering (HyDaP) algorithm to conduct variable selection and identify clusters in data consisting of mixed types of variables. Our algorithm involves a pre-processing step to identify the data structure formed by continuous variables and to select important variables, and a clustering step to determine the cluster membership. In the clustering step, we proposed a dissimilarity measure that balances the contributions between continuous and categorical variables, which the existing clustering methods do not offer. Through simulation studies, we showed that the proposed HyDaP algorithm is robust to different data structures and outperforms or at par commonly used methods. We also defined 3 different data structures to help understand data and better interpret clustering results. Our method successfully identified four clinically meaningful sepsis phenotypes for data extracted from EHR of multiple health care systems. The resulting phenotypes are highly associated with several clinical endpoints.

Our HyDaP algorithm inherits the limitations of sparse K-means. For data under the *natural cluster structure*, if the continuous variables contain many outliers or excessive zeros (a.k.a. zero-inflated), the sparse K-means procedure cannot correctly identify variables with high contributions. Another situation that could affect the later steps in our method and lead to unsatisfactory results is that when data contains continuous variables of the same value for the majority of subjects while other few have different values. We also suggest checking variables that have similar clinical meanings or highly clinically related before clustering and only keep one of them to avoid existence of within-cluster correlations.

Clustering has emerged as one of the essential and popular techniques for discovering patterns in data or disease phenotypes. Although clustering methods keep evolving to cope with increasing complexity in data, certain features in data sets could limit the utilization of the existing approaches. Unlike genetics or genomics data, data collected from clinical settings often include various types. Our proposed algorithm overcomes the drawbacks of the commonly used clustering algorithms therefore the results from using our method may be more helpful to clinicians in making good medical decisions.

3.0 BAYESIAN FINITE MIXTURE MODEL WITH VARIABLE SELECTION

3.1 INTRODUCTION

The finite mixture model (FMM) (McCutcheon, 1987; Moustaki, 1996; Nylund et al., 2007) has been used to uncover the latent mixture probability distributions in a combined statistical distribution of a population (Deb et al., 2008) when the population is heterogeneous in characteristics and consisting with a combination of several more homogeneous subgroups. This method has a natural interpretation of heterogeneity through the mixture of finite components with a distributional assumption conditioning on components for each variable.

Unlike the commonly used nonparametric clustering algorithms (e.g., distance-based, density-based), the FMM technique is a model-based method that can easily handle variables with mixed variable types and can do variable selection while performing clustering. This is especially important when the number of variables involved is large.

The current method of FMM has to deal with several challenges. The first challenge is related to the use of the EM algorithm (Dempster et al., 1977), which is the mostly commonly used estimation procedure for the FMMs. In the EM algorithm the convergence rate could be very slow and the solution could be highly dependent on the choices of initial values, especially in the multivariate settings (Biernacki et al., 2003; Karlis and Xekalaki, 2003; McLachlan and Krishnan, 2007).

The second challenge of using FMM is how to handle censored probability distributions. Oftentimes, covariates collected from medical setting include biomarker data of patients. Biomarkers are characteristics that can be accurately and reproducibly measured and accessed as an indicator of various biological processes (Group et al., 2001; Strimbu and Tavel, 2010). Different biomarkers serve for different purposes. For example, temperature can be seen as a biomarker for fever; C-reactive protein (CRP) and Interleukin 6 (IL-6) are commonly used as biomarkers for sepsis (Pierrakos and Vincent, 2010). Therefore, biomarkers usually contain important diagnosis information about subjects. However, many biomarkers are subject to a limit of detection (LOD). LOD could be lower detection limit (i.e., values below this limit could not be measured), higher detection limit (i.e., values above this limit could not be measured), or both. When biomarkers are outcome variables in data analysis, biomarker values more extreme than the detection limit are usually represented by the detection limit value and an additional binary variable is included to indicating whether the corresponding value is actually undetected or measured. A semiparametric censored regression model can then be used (Tobin, 1958; Powell, 1984; Honoré, 1992). When biomarkers are predictors in analysis, multiple imputations are often used (Lubin et al., 2004; Lee et al., 2012; Bernhardt et al., 2015). If the objective of the analysis is to cluster the feature space when data containing censored biomarkers, the two aforementioned approaches cannot be applied. Therefore, conducting clustering for data with LODs is still not well-addressed.

The objective of this study is to develop a new clustering algorithm under the FMM framework to handle censored biomarker variables. Our method also incorporates a variable selection procedure to determine the importance of variables on clustering

The basic concepts of our development is to overcome the estimation limitation of the EM algorithm from a Bayesian perspective. We first adopt Gibbs sampling, which has been shown to be a valid and practical way in estimation of the FMMs (Geman and Geman, 1984), (Diebolt and Robert, 1994). Second, we propose the use of a spike-and-slab type prior for categorical variables for the purpose of variable selection. Together with the traditional spike-and-slab prior for continuous variables, we incorporate variable selection into our Bayesian framework to provide quantitative information about variable importance. Third, we introduce an additional sampling step into our framework so that it is able to handle censored biomarker variables that are often encountered in clinical data.

Section 3.2 contains our review of currently used variable selection methods for the finite mixture models. We describe our proposed method in detail in Section 3.3. Section 3.4 includes simulation studies that are used to assess the performance of our methods and compare the performance to existing methods. In Section 3.5, our proposed method is applied to identify sepsis phenotypes using demographic, clinical, and biomarker data collected from electronic health records. Our conclusions and the summary of this study are in Section 3.6.

3.2 EXISTING METHODS REVIEW

Overall, there are two categories of variable selection methods for supervised or unsupervised machine learning: filter methods and wrapper methods (Blum and Langley, 1997; Guyon and Elisseeff, 2003; Fop et al., 2018). Filter methods refer to those whose feature selection procedures are conducted separately with clustering procedures. On the contrary, wrapper methods refer to those whose feature selection is conducted simultaneously with clustering procedures, like "wrapped" around clustering procedures. For example, the step 1 of our proposed HyDaP algorithm can be viewed as a filter method as it is conducted separately with the actual clustering algorithm. Wrapper methods are relatively more popular since it is naturally incorporated in clustering algorithms. In this section we focus on wrapper methods for finite mixture models.

Liu et al. proposed to conduct principle component analysis (PCA) before fitting Gaussian finite mixture models (Liu et al., 2003). This method assumes that only the first K factors are relevant to clustering, where K is a random variable that has a prior distribution. However, factors having larger eigen values in PCA do not necessarily contain more important information for clustering (Chang, 1983).

Law et al. defined a binary indicator called feature saliency for each variable to reflect whether this variable is relevant to clustering or not. EM algorithm was used for estimation (Law et al., 2004). Let ϕ_m denote saliency for variable m. If $\phi_m = 1$ then variable m is relevant, otherwise variable m is not, namely its distribution is independent of cluster labels. Let $\boldsymbol{X}$ denote a data matrix with n subjects and M variables. Let x_{im} denote variable m of subject i. Let G denote number of clusters. Let $\boldsymbol{Z}$ denote cluster indicator matrix; z_{ig} denote indicator variable of subject i belonging to cluster g. Let $\boldsymbol{\beta}$ denote distributional parameters, $\boldsymbol{\beta}_{mg}$ denote parameters of variable m in cluster g, $\boldsymbol{\beta}_m$ denote marginal parameter of variable

m. Then L_m, the likelihood due to variable m, is defined as:

$$\begin{cases} L_m|\boldsymbol{X}, \boldsymbol{Z}, G, \boldsymbol{\beta} = \prod_{i=1}^{n} \prod_{g=1}^{G} [f(x_{im}|\boldsymbol{\beta}_{mg})]^{z_{ig}} & \text{if variable } m \text{ is relevant} \\ L_m|\boldsymbol{X}, \boldsymbol{\beta} = \prod_{i=1}^{n} f(x_{im}|\boldsymbol{\beta}_m) & \text{if variable } m \text{ is irrelevant} \end{cases}$$

Tadesse et al. and Li et al. used similar definitions of likelihood function and feature saliency. (Tadesse et al., 2005; Li et al., 2009). Different with the one proposed by Law et al., the method proposed by Tadesse et al. detects discriminating variables through reversible-jump MCMC instead of EM algorithm for high-dimensional Gaussian finite mixture models. Later, White et al. applied this idea on latent class analysis (White et al., 2016) for Bayesian variable selection. While the method proposed by Li et al. adopted Variational Learning of Bayesian approximation (VB) for inference. They claimed that although using EM algorithm and VB usually lead to identical results, VB could avoid the situation of getting infinite likelihood when there is singular cluster, which may encounter using EM. Later Sun et.al. used the same framework and extended Gaussian mixture model to Student's t mixture model to better handle outliers (Sun et al., 2018).

Another category of methods is penalization approach (Fop et al., 2018). The general idea is to maximize a penalized log-likelihood which is defined as:

$$l = \sum_{i=1}^{n} log\{\sum_{g=1}^{G} \tau_g f(\boldsymbol{x}_i|\boldsymbol{\beta}_g)\} - Q_\lambda(\boldsymbol{\beta})$$

where $\boldsymbol{x}_i$ is vector $(x_{i1}, x_{i2}, \ldots, x_{iM})^T$; $\boldsymbol{\beta}$ is vector $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_G)^T$ and each $\boldsymbol{\beta}_g$ represents distributional parameters for cluster g; Q is a function of distributional parameters; λ is penalty parameter. Different penalty functions were proposed including $L1$ penalty (Pan and Shen, 2007), sample size weighted $L1$ penalty (Bhattacharya and McNicholas, 2014), L_∞ penalty (Wang and Zhu, 2008) and other variations to achieve the goal of variable selection. These methods all assume that the differences of mixture components lie in mean parameters, which assumes similar variance within all clusters. This assumption was later relaxed by Xie et al. by adding two penalty terms so that the variance covariance structure is cluster-specific diagonal matrices (Xie et al., 2008). Later this method was further extended to unconstrained variance covariance matrices (Zhou et al., 2009).

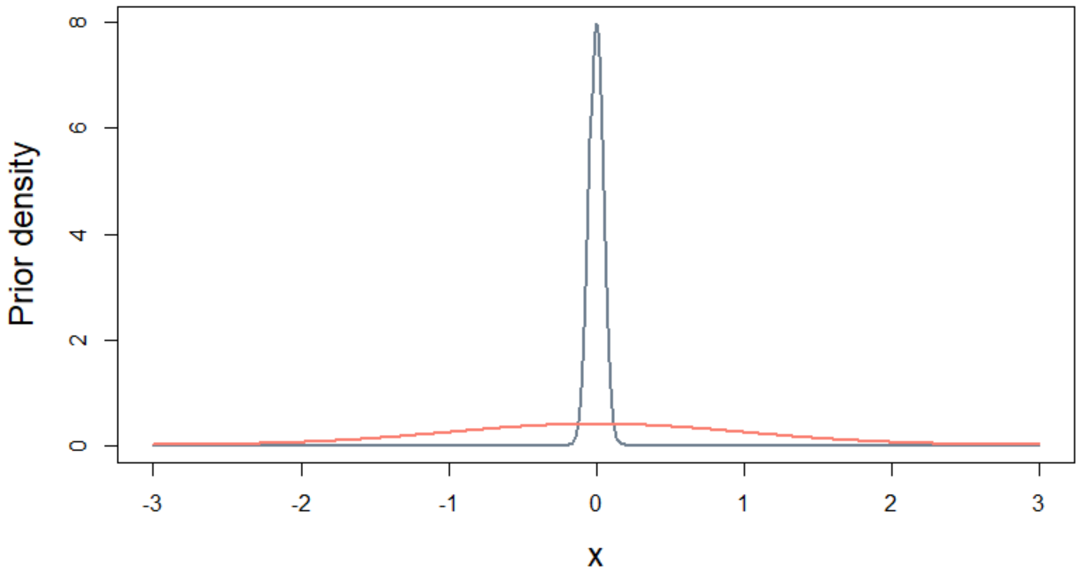


Figure 8: Demonstration of a spike and slab prior

Another category of methods that is commonly applied is spike and slab prior (Mitchell and Beauchamp, 1988; Madigan and Raftery, 1994; George and McCulloch, 1997; Ishwaran et al., 2005). It is demonstrated in Figure 8. Spike and slab prior is originally proposed for variable selection in linear regressions, but it can be naturally applied on Gaussian mixture models for unsupervised clustering. This method first chooses one cluster as reference, and obtain mean difference between other clusters with this reference cluster. Similar to feature saliency, a binary indicator representing variable importance is also defined. For other methods defined feature saliency, each variable has only one feature saliency indicating whether it is relevant to clustering or not. While under spike and slab prior, each variable first has a cluster-specific importance, for cluster $2, 3, \dots, G$ and then we aggregate these cluster-specific importance values into one overall importance or weight value. Let Δ_{mg} denote importance of variable m within cluster g, μ_{mg} the mean difference between cluster g and cluster 1 of variable m, where $g = 2, 3, \dots, G$. $\Delta_{mg} = 0$ indicates that for variable m, cluster g is not different with reference cluster, namely variable m within cluster g is not important. Then in next iteration, we assign μ_{mg} a “spike” prior (the grey density in Figure 8): a Gaussian prior centered at 0 with very small variance. In this way, a value close to 0 is very likely to be sampled as the updated value of μ_{mg}. Otherwise if $\Delta_{mg} = 1$, it indicates that for variable m, cluster g is different with reference cluster, namely variable m within cluster g is important. Then we assign μ_{mg} a "slab" prior (the red density in Figure 8): a Gaussian prior centered at 0 with very large variance. Therefore, any value is possible to be sampled as the updated value of μ_{mg}. Let Δ_m denote final weight of variable m. Within each iteration, as long as $\Delta_{mg} = 1$ for at least one cluster, we assign 1 to Δ_m for that iteration to represent that variable m is important. After all iterations, we calculate $p(\Delta_m = 1)$ to be the weight of variable m.

Although researchers have explored a lot of variable selection methods for finite mixture models, all above methods only focus on single type of variables. Not many methods incorporate both types of variables. The method proposed by Raftery and Dean is able to handle various types of variables through stepwise regression like procedures (Raftery and Dean, 2006). They used approximate Bayes factor as variable selection criteria and obtained final optimal variable sets through a greedy search algorithm. This method naturally takes advantage of the parametric form of finite mixture models, but it could be extremely slow if a data set contains a large number of variables. Besides, this method only identifies that whether a variable is important or not instead of providing quantitative values for importance. Later it was extended through combining LASSO-like procedures and model selection procedures (Celeux et al., 2018) to be more efficient. But similarly, this method does not provide real-valued variable weights.

3.3 BAYESIAN FINITE MIXTURE MODEL WITH VARIABLE SELECTION

3.3.1 Notation and proposed model

We will first define some notations that will be used to define our proposed Bayesian finite mixture model.

Let $\boldsymbol{X}$, $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iM})^T$ denote the vector of covariates for subject i and $\boldsymbol{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^T$ be an $n \times M$ data matrix, where n is the number of subjects and M is the number of variables. Let $\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iG})^T$ denote the vector of cluster-membership indicators for subject i, where $z_{ig} = 1$ if the subject i belongs to cluster g and 0 otherwise. $\sum_{g=1}^{G} z_{ig} = 1$, where G is the number of clusters. Without loss of generality, we let the first q variables to be continuous normally distributed and the rest $q + 1^{th}$ to M^{th} variables to be categorical.

Let $\boldsymbol{A}_1 = (A_{11}, A_{12}, \ldots, A_{1q})^T$ denote the vector of mean values of all continuous variables in cluster 1, where A_{1m} is the mean of variable m in cluster 1, $m = 1, 2, \ldots, q$. We also define $\boldsymbol{\sigma}^2 = (\sigma_1^2, \sigma_2^2, \ldots, \sigma_q^2)^T$ as the vector of variances for all continuous variables, where σ_m^2 is the variance of variable m, $m = 1, 2, \ldots, q$.

We further define $\boldsymbol{\mu}$ as a $q \times G$ matrix with its (m, g) element μ_{mg} representing the mean difference of variable m between cluster g and cluster 1, where $m = 1, 2, \ldots, q$. Note that cluster 1 is the reference cluster. Then, vector $\boldsymbol{\mu}_g = (\mu_{1g}, \mu_{2g}, \ldots, \mu_{qg})^T$ represents the mean difference between cluster g and cluster 1 of all normally distributed variables and vector $\boldsymbol{\mu}_m = (\mu_{m1}, \mu_{m2}, \ldots, \mu_{mG})^T$ represents the mean differences of variable m for all G clusters. For identifiability, $\boldsymbol{\mu}_1$ is set to be a vector with all elements being 1.

For categorical variables, let $\boldsymbol{\theta}$ be a $(M - q) \times G$ matrix with its (m, g) element $\boldsymbol{\theta}_{mg}$ representing the distributional parameters of variable m within cluster g, where $m = q + 1, q + 2, \ldots, M$. Vector $\boldsymbol{\theta}_g = (\boldsymbol{\theta}_{(q+1)g}, \boldsymbol{\theta}_{(q+2)g}, \ldots, \boldsymbol{\theta}_{Mg})^T$ represents the parameters of all categorical variables within cluster g and vector $\boldsymbol{\theta}_m = (\boldsymbol{\theta}_{m1}, \boldsymbol{\theta}_{m2}, \ldots, \boldsymbol{\theta}_{mG})^T$ represents the parameters of variable m for $m = q + 1, q + 2, \ldots, M$.

We define density function of the finite mixture model (FMM) with G clusters as

$$f(\mathbf{x}_i|\boldsymbol{\tau}, \boldsymbol{\theta}, \boldsymbol{A}_1, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) = \sum_{g=1}^{G} \tau_g f(\mathbf{x}_i|\boldsymbol{\theta}_g, \boldsymbol{A}_1, \boldsymbol{\mu}_g, \boldsymbol{\sigma}^2),$$

where $\sum_{g=1}^{G} \tau_g = 1$ and τ_g is the probability that a subject belongs to cluster g. We let $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_G)^T$ be the vector of cluster mixture probabilities.

For subject i, the corresponding density function for data ($\mathbf{x}_i$ given its cluster membership $\mathbf{z}_i$) could be written as:

$$f(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\theta}, \boldsymbol{A}_1, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) = \prod_{g=1}^{G} [f(\mathbf{x}_i|\boldsymbol{\theta}_g, \boldsymbol{A}_1, \boldsymbol{\mu}_g, \boldsymbol{\sigma}^2)]^{z_{ig}},$$

where $f(\mathbf{x}_i|\boldsymbol{\theta}_g, \boldsymbol{A}_1, \boldsymbol{\mu}_g, \boldsymbol{\sigma}^2) = \prod_{m=1}^{M} f(x_{im}|\theta_{mg}, A_{1m}, \mu_{mg}, \sigma_m^2)$ based on conditional independence, which assumes variables are independent with each other conditional on the cluster labels.

3.3.2 Priors

We define conjugate priors for all parameters specified in the above-defined FMM. Details of the prior distribution for each parameter is specified below.

For the cluster indicator matrix $\mathbf{Z}$, we specify a multinomial prior distribution with the form: $\mathbf{z}_i \sim Multinomial\,(G, \tau_1, \tau_2, \ldots \tau_G)$. We also let the probability of cluster membership $\boldsymbol{\tau}$ follow a Dirichlet distribution

$$\boldsymbol{\tau} \sim Dir\,(\delta_1, \delta_2, \ldots \delta_G)\,,$$

where $\delta_1, \delta_2, \ldots \delta_G$ are the hyper-parameters of $\boldsymbol{\tau}$.

Let parameter A_{1m} follows $N\,(\mu_A, \sigma_A^2)$, where μ_A and σ_A^2 are both hyper-parameters.

Let $\boldsymbol{\Delta}$ be a $G \times M$ matrix representing the collection of importance indicator for each variable, where Δ_{mg} is the importance indicator of variable m within cluster g. Vector $\boldsymbol{\Delta}_g$ is $(\Delta_{1g}, \Delta_{2g}, \ldots, \Delta_{Mg})^T$ representing the importance indicator of all variables within cluster g; and vector $\boldsymbol{\Delta}_m$ is $(\Delta_{m1}, \Delta_{m2}, \ldots, \Delta_{mG})^T$ representing the importance indicator of variable

m. For the purpose of variable selection, we apply a spike-and-slab priors for parameter μ_{mg} with the form

$$\begin{cases} \mu_{mg} \sim N\left(0, \sigma^2_{\Delta_0}\right) & \text{if } \Delta_{mg} = 0 \\ \mu_{mg} \sim N\left(0, \sigma^2_{\Delta_1}\right) & \text{if } \Delta_{mg} = 1, \end{cases}$$

where $\sigma^2_{\Delta_0}$ and $\sigma^2_{\Delta_1}$ are both hyper-parameters. For identifiability, we define $\mu_{mg} = 0$ when $g = 1$. When $\Delta_{mg} = 1$, namely cluster g is different from cluster 1, μ_{mg} should be away from 0. Therefore, $\sigma^2_{\Delta_1}$ should be a very large number, for increasing the variability of cluster mean difference, so that it is likely to be different with 0. When $\Delta_{mg} = 0$, which occurs if cluster g is not different from cluster 1, μ_{mg} should be close to 0. Therefore, $\sigma^2_{\Delta_0}$ should be a very small number so that the mean differences among cluster means would be close to zero. We used an Inverse-Gamma prior for parameter $\sigma^2_{\Delta_0}$:

$$\sigma^2_{\Delta_0} \sim Inv\Gamma\left(a_{\Delta_0}, b_{\Delta_0}\right),$$

where a_{Δ_0} and b_{Δ_0} are both hyper-parameters.

The precision parameter $\gamma_m = \sigma_m^{-2}$ is assigned a prior following a Gamma distribution,

$$\gamma_m \sim \Gamma\left(\tilde{a}, \tilde{b}\right),$$

where $\tilde{a}$ and $\tilde{b}$ are both hyper-parameters.

Inspired by the usual spike-and-slab priors for continuous variables, we propose similar priors for categorical variables. The challenge of building a spike-and-slab prior for a categorical variable include (1) we do not have an appropriate distribution for the probability difference; (2) a categorical variable may contain multiple levels so it is hard to compare all levels altogether between two clusters. Therefore, for a categorical variable, instead of comparing to a reference clustering group, we compare the distribution of the levels within each of the G clusters to the overall marginal distribution of this categorical variable. If $\Delta_{mg} = 0$, this indicates that the distribution of the categories within cluster g is identical to the marginal distribution. If $\Delta_{mg} = 1$, it implies that the distribution within cluster g is different from the marginal distribution.

We define a Dirichlet distribution for categorical variable parameter $\boldsymbol{\theta}_{mg}$:

$$\begin{cases} \boldsymbol{\theta}_{mg} \sim Dir\left(\boldsymbol{\alpha}_{m\Delta_0}\right) & \text{if } \Delta_{mg} = 0 \\ \boldsymbol{\theta}_{mg} \sim Dir\left(\boldsymbol{\alpha}_{\Delta_1}\right) & \text{if } \Delta_{mg} = 1 \end{cases}$$

where $\boldsymbol{\alpha}_{m\Delta_0}$ and $\boldsymbol{\alpha}_{\Delta_1}$ are hyper-parameters. Vector $\boldsymbol{\alpha}_{m\Delta_0} = (\alpha_{m\Delta_0 1}, \alpha_{m\Delta_0 2}, \ldots, \alpha_{m\Delta_0 L_m})^T$ is proportional to $\boldsymbol{\theta}_{mg}$ to make the corresponding prior center at marginal parameters of variable m, where L_m is the number of categories of variable m; its elements are relatively larger numbers so that the prior is "spike" at marginal parameters of variable m when $\Delta_{mg} = 0$. Elements of vector $\boldsymbol{\alpha}_{\Delta_1}$ are all 1 thus the pdf of this Dirichlet distribution is a constant. Therefore, this prior becomes a "slab" one when $\Delta_{mg} = 1$.

We used Bernoulli distribution for the importance indicator variable Δ_{mg}:

$$\begin{cases} \Delta_{mg} \sim Bern\left(p_{1m}\right) & \text{if } m \in \{1, 2, \ldots, q\} \\ \Delta_{mg} \sim Bern\left(p_{2m}\right) & \text{if } m \in \{q+1, q+2, \ldots, M\} \end{cases}$$

where p_{1m} and p_{2m} are hyper-parameters of Δ_{mg}.

Beta distribution for p_{1m} and p_{2m}:

$$p_{1m} \sim Beta(a_{p_1}, b_{p_1})$$

$$p_{2m} \sim Beta(a_{p_2}, b_{p_2})$$

where a_{p_1}, b_{p_1}, a_{p_2}, b_{p_2} are hyper-parameters.

A graphical depiction of our model is shown in Figure 9.

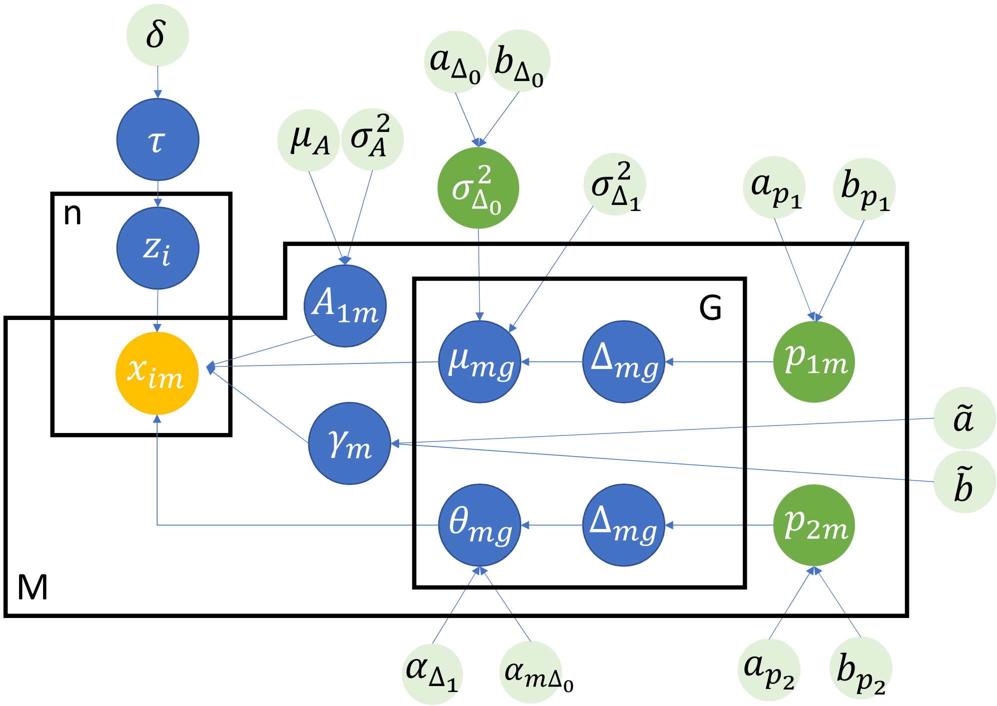


Figure 9: Graphical representation of the Bayesian framework of our proposed finite mixture model

3.3.3 Posteriors

With specified prior distributions of all parameters, the corresponding posterior distribution then can be obtained with the form:

$$\begin{aligned} f(Parameters|\mathbf{X}) \propto & \prod_{i=1}^{n}[\prod_{m=1}^{q} f(x_{im}|z_{im}, A_{1m}, \boldsymbol{\mu}_m, \gamma_m) \prod_{m=q+1}^{M} f(x_{im}|z_{im}, \boldsymbol{\theta}_m)] \\ & \times \prod_{i=1}^{n} f(\mathbf{z}_i|\boldsymbol{\tau}) f(\boldsymbol{\tau}) \prod_{m=1}^{q} N(A_{1m}; \mu_A, \sigma_A^2) \prod_{m=1}^{q} \Gamma(\gamma_m; \tilde{a}, \tilde{b}) \\ & \times \prod_{m=1}^{q} \prod_{g=2}^{G} [N(\mu_{mg}; 0, \sigma_{\Delta_1}^2)\Delta_{mg} + N(\mu_{mg}; 0, \sigma_{\Delta_0}^2)(1 - \Delta_{mg})] \\ & \times Inv\Gamma(\sigma_{\Delta_0}^2; a_{\Delta_0}, b_{\Delta_0}) \prod_{m=1}^{q} \prod_{g=2}^{G} Bern(\Delta_{mg}; p_{1m}) \prod_{m=1}^{q} Beta(p_{1m}; a_{p_1}, b_{p_1}) \\ & \times \prod_{m=q+1}^{M} \prod_{g=1}^{G} [Dir(\theta_{mg}; \boldsymbol{\alpha}_{\Delta_1})\Delta_{mg} + Dir(\theta_{mg}; \boldsymbol{\alpha}_{m\Delta_0})(1 - \Delta_{mg})] \\ & \times \prod_{m=q+1}^{M} \prod_{g=1}^{G} Bern(\Delta_{mg}; p_{2m}) \prod_{m=q+1}^{M} Beta(p_{2m}; a_{p_2}, b_{p_2}) \end{aligned}$$

We can derive posterior distribution for each parameter as below:
For cluster indicator matrix $\mathbf{Z}$:

$$\boldsymbol{z}_i|\mathbf{x}_i, \boldsymbol{\tau}, \boldsymbol{A_1}, \boldsymbol{\mu}, \boldsymbol{\gamma}, \boldsymbol{\theta} \sim Multinomial\left(\frac{\tau_1 f(\mathbf{x}_i|\boldsymbol{A_1}, \boldsymbol{\mu}_1, \boldsymbol{\gamma}_1, \boldsymbol{\theta}_1)}{\sum_{g=1}^{G} \tau_g f(\mathbf{x}_i|\boldsymbol{A_1}, \boldsymbol{\mu}_g, \boldsymbol{\gamma}_g, \boldsymbol{\theta}_g)}, \ldots, \frac{\tau_G f(\mathbf{x}_i|\boldsymbol{A_1}, \boldsymbol{\mu}_G, \boldsymbol{\gamma}_G, \boldsymbol{\theta}_G)}{\sum_{g=1}^{G} \tau_g f(\mathbf{x}_i|\boldsymbol{A_1}, \boldsymbol{\mu}_g, \boldsymbol{\gamma}_g, \boldsymbol{\theta}_g)}\right).$$

For cluster mixture proportion $\boldsymbol{\tau}$:

$$\boldsymbol{\tau}|\mathbf{Z}, \delta_1, \ldots, \delta_G \sim Dir\left(\delta_1 + \sum_{i=1}^{n} z_{i1}, \ldots, \delta_G + \sum_{i=1}^{n} z_{iG}\right).$$

For A_{1m} in non-censored normal distributed variable m within cluster g, where $m = 1, 2, \ldots, q$:

$$A_{1m}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\mu}_m, \gamma_m, \mu_A, \sigma_A^2 \sim N\left(\frac{\sigma_A^2(\sum_{i=1}^{n} x_{im} - \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\mu_{mg}) + (\mu_A/\gamma_m)}{n\sigma_A^2 + (1/\gamma_m)}, \frac{\sigma_A^2/\gamma_m}{n\sigma_A^2 + (1/\gamma_m)}\right).$$

For mean difference μ_{mg}, where $m = 1, 2, \ldots, q$:

when $\Delta_{mg} = 1$,

$$\mu_{mg}|\boldsymbol{X}, \boldsymbol{Z}, A_{1m}, \gamma_m, \sigma^2_{\Delta_1} \sim N\left(\frac{\sigma^2_{\Delta_1}(\sum_{i=1}^n z_{ig}x_{im} - \sum_{i=1}^n z_{ig}A_{1m})}{\sigma^2_{\Delta_1}\sum_{i=1}^n z_{ig} + (1/\gamma_m)}, \frac{\sigma^2_{\Delta_1}/\gamma_m}{\sigma^2_{\Delta_1}\sum_{i=1}^n z_{ig} + (1/\gamma_m)}\right),$$

when $\Delta_{mg} = 0$,

$$\mu_{mg}|\boldsymbol{X}, \boldsymbol{Z}, A_{1m}, \gamma_m, \sigma^2_{\Delta_0} \sim N\left(\frac{\sigma^2_{\Delta_0}(\sum_{i=1}^n z_{ig}x_{im} - \sum_{i=1}^n z_{ig}A_{1m})}{\sigma^2_{\Delta_0}\sum_{i=1}^n z_{ig} + (1/\gamma_m)}, \frac{\sigma^2_{\Delta_0}/\gamma_m}{\sigma^2_{\Delta_0}\sum_{i=1}^n z_{ig} + (1/\gamma_m)}\right).$$

For precision parameter γ_m, where $m = 1, 2, \ldots, q$:

$$\gamma_m|\mathbf{X}, \mathbf{Z}, A_{1m}, \boldsymbol{\mu}_m, \tilde{a}, \tilde{b} \sim \Gamma\left(\tilde{a} + \frac{1}{2}n, \tilde{b} + \frac{1}{2}\sum_{i=1}^n\sum_{g=1}^G z_{ig}(x_{im} - A_{1m} - \mu_{mg})^2\right).$$

For hyper-parameter $\sigma^2_{\Delta_0}$ of μ_{mg}:

$$\sigma^2_{\Delta_0}|\boldsymbol{\Delta}, \boldsymbol{\mu}, a_{\Delta_0}, b_{\Delta_0} \sim Inv\Gamma\left(a_{\Delta_0} + \frac{1}{2}\sum_{m=1}^q\sum_{g=2}^G(1 - \Delta_{mg}), b_{\Delta_0} + \frac{1}{2}\sum_{m=1}^q\sum_{g=2}^G(1 - \Delta_{mg})\mu^2_{mg}\right).$$

For parameter $\boldsymbol{\theta}_{mg}$, where $m = q+1, q+2, \ldots, M$:

When $\Delta_{mg} = 1$,

$$\boldsymbol{\theta}_{mg}|\mathbf{X}, \mathbf{Z}, \boldsymbol{\alpha}_{\Delta_1} \sim Dir\left(\alpha_{\Delta_1 1} + \sum_{i=1}^n x_{im1}z_{ig}, \ldots, \alpha_{\Delta_1 L_m} + \sum_{i=1}^n x_{imL_m}z_{ig}\right)$$

When $\Delta_{mg} = 0$,

$$\boldsymbol{\theta}_{mg}|\mathbf{X}, \mathbf{Z}, \boldsymbol{\alpha}_{m\Delta_0} \sim Dir\left(\alpha_{m\Delta_0 1} + \sum_{i=1}^n x_{im1}z_{ig}, \ldots, \alpha_{m\Delta_0 L_m} + \sum_{i=1}^n x_{imL_m}z_{ig}\right).$$

For importance indicator variable $\boldsymbol{\Delta}$:

When $m \in 1, 2, \ldots, q$:

$$\Delta_{mg}|\mu_{mg}, p_{1m}, \sigma^2_{\Delta_0}, \sigma^2_{\Delta_1} \sim Bernoulli\left(\frac{p_{1m}N(\mu_{mg}; 0, \sigma^2_{\Delta_1})}{p_{1m}N(\mu_{mg}; 0, \sigma^2_{\Delta_1}) + (1 - p_{1m})N(\mu_{mg}; 0, \sigma^2_{\Delta_0})}\right)$$

When $m \in q+1, q+2, \ldots, M$:

$$\Delta_{mg}|\mu_{mg}, p_{2m}, \boldsymbol{\alpha}_{m\Delta_0}, \boldsymbol{\alpha}_{\Delta_1} \sim Bernoulli\left(\frac{p_{2m}Dir(\theta_{mg}; \boldsymbol{\alpha}_{\Delta_1})}{p_{2m}Dir(\theta_{mg}; \boldsymbol{\alpha}_{\Delta_1}) + (1 - p_{2m})Dir(\theta_{mg}; \boldsymbol{\alpha}_{m\Delta_0})}\right).$$

For p_{1m} for continuous variables, where $m = 1, 2, \ldots, q$:

$$p_{1m}|\boldsymbol{\Delta}_m, a_{p_1}, b_{p_1} \sim Beta\left(a_{p_1} + \sum_{g=2}^{G} \Delta_{mg}, b_{p_1} + \sum_{g=2}^{G}(1 - \Delta_{mg})\right).$$

For p_{2m} for categorical variables, where $m = q + 1, q + 2, \ldots, M$:

$$p_{2m}|\boldsymbol{\Delta}_m, a_{p_2}, b_{p_2} \sim Beta\left(a_{p_2} + \sum_{g=1}^{G} \Delta_{mg}, b_{p_2} + \sum_{g=1}^{G}(1 - \Delta_{mg})\right).$$

We set different hyper-parameters for p_{1m} and p_{2m} to make variable selection more flexible. In this way, we can control the extent of shrinkage for continuous and categorical variables separately. In the next section, we will introduce how different choices of hyper-parameters could affect variable selection with more details.

For continuous variables with detection limit, we add one more sampling step: sample values below the lower detection limit and above the upper detection limit from the remainder of truncated normal distributions, respectively.

$$f(x_{i'm}|A_{1m}, \mu_{mg}, \gamma_m) = \left(\frac{\sqrt{\gamma_m} h(x_{i'm})}{\Phi[\sqrt{\gamma_m}(C_{Lm} - \mu_{mg} - A_{1m})]}\right)^{z_{ig}}$$

$$f(x_{i^*m}|A_{1m}, \mu_{mg}, \gamma_m) = \left(\frac{\sqrt{\gamma_m} h(x_{i^*m})}{1 - \Phi[\sqrt{\gamma_m}(C_{Um} - \mu_{mg} - A_{1m})]}\right)^{z_{ig}},$$

where C_{Um} is the upper limit and C_{Lm} is the lower limit for variable m; i' represents subjects whose real values of variable m are lower than C_{Lm}; i^* represents subjects whose real values of variable m are higher than C_{Um}; h is standard normal distribution; Φ is CDF of standard normal distribution.

3.3.4 Hyper-parameters

From the Bayesian framework of our proposed model, we will need to specify the values of the following hyper-parameters before starting the sampling procedures: $\boldsymbol{\delta}$ for $\boldsymbol{\tau}$; μ_A and σ_A^2 for $\boldsymbol{A_1}$; a_{Δ_0} and b_{Δ_0} for $\sigma_{\Delta_0}^2$; $\sigma_{\Delta_1}^2$ for $\boldsymbol{\mu}$; $\tilde{a}$ and $\tilde{b}$ for $\boldsymbol{\gamma}$; $\boldsymbol{\alpha}_{\Delta_1}$ and $\boldsymbol{\alpha}_{m\Delta_0}$ for $\boldsymbol{\theta}$; a_{p_1} and b_{p_1} for p_{1m}; and a_{p_2} and b_{p_2} for p_{2m}.

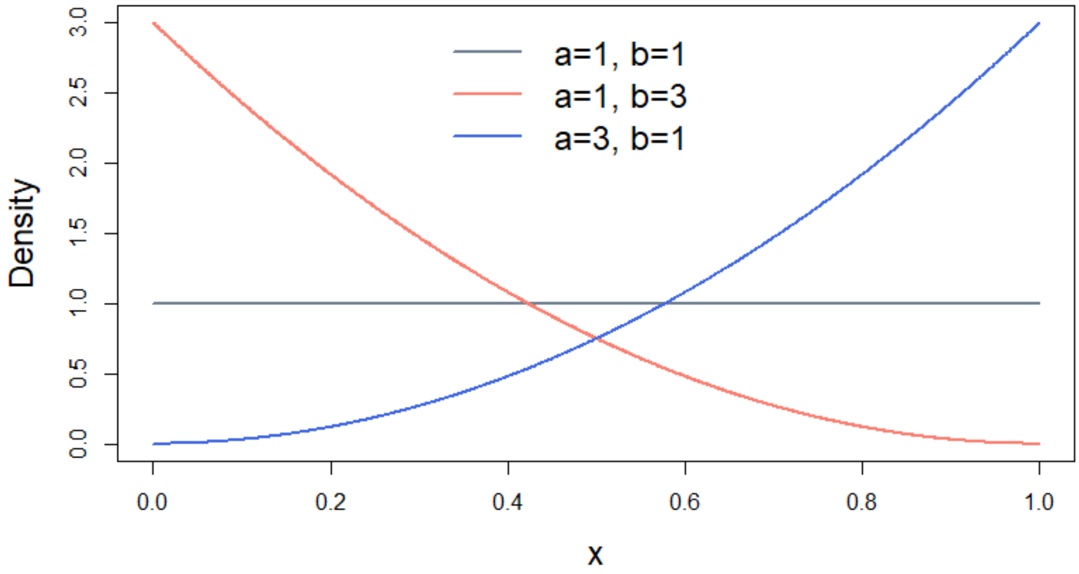


Figure 10: Beta priors

We can control the extent of shrinkage through controlling hyper-parameters a_{p_1} and b_{p_1} for continuous variables, and a_{p_2} and b_{p_2} for categorical variables. We can start with $a_{p_1} = b_{p_1} = a_{p_2} = b_{p_2} = 1$, namely $Beta(1,1)$ to be the prior distribution of p_{1m} and p_{2m}. If we prefer more shrinkage, then we could increase b_{p_1} and b_{p_2}; otherwise we could decrease b_{p_1} and b_{p_2}. In Figure 10 we show probability density functions of $Beta(1,1)$, $Beta(1,3)$ and $Beta(3,1)$ as an example. We can find that $Beta(1,1)$ (the grey line in Figure 10) is the same as $Unif(0,1)$, namely it is a non-informative flat prior. Under prior $Beta(1,3)$ (the red curve in Figure 10), it is more likely to sample smaller numbers for corresponding Δ_{mg} so that we can achieve more shrinkage. While under prior $Beta(3,1)$ (the blue curve in Figure 10), it is more likely to sample larger numbers for the corresponding Δ_{mg} so that we put less shrinkage. Assigning different hyper-parameters to p_{1m} and p_{2m} makes the shrinkage more flexible since we can control continuous and categorical variable separately. If we prefer the same shrinkage for all variables, we can simply set $a_{p_1} = a_{p_2}$ and $b_{p_1} = b_{p_2}$.

3.3.5 Estimation procedure

Since all parameters in the model have conjugate priors, we can apply Gibbs sampling to estimate unknown parameters and the procedure is summarized as follows.

Step 1: Set initial values for all parameters.

Step 2: For each variable in the dataset (e.g., age), update its distributional parameters (e.g., mean and standard deviation), namely $\boldsymbol{A_1}$, $\boldsymbol{\mu}$, and $\boldsymbol{\gamma}$ for a continuous variable or $\boldsymbol{\theta}$ for a categorical variable by sampling from the corresponding posterior distributions of parameters. If censored biomarker variables are encountered, update the censored variables before updating the distributional parameters.

Step 3: For each variable, update $\boldsymbol{\Delta}$ given the values of all other parameters.

Step 4: For each variable, update p_{1m} or p_{2m} given the values of all other parameters.

Step 5: Update $\sigma^2_{\Delta_0}$, $\mathbf{Z}$, and $\boldsymbol{\tau}$ given the values of all other parameters.

Step 6: Repeat Steps 2 to 5 for many iterations.

Step 7: Estimate each parameter using the mean value of its posterior distribution. Also, calculate the posterior probabilities of cluster membership for each subject and assign the cluster with the highest probability to that subject.

3.3.6 Label switching in Gibbs sampling

Label switching is a common problem in Gibbs sampling. It has been extensively studied and discussed in the literatures (Diebolt and Robert, 1994; Stephens, 2000; Frühwirth-Schnatter, 2001; Marin et al., 2005; Papastamoulis and Iliopoulos, 2010). Scrambling the cluster labels will not affect the likelihood function, the priors, and the posterior distributions. In Gibbs sampling of the cluster-specific parameters, there is a possibility of assigning wrong cluster labels to these parameters. This is called the label switching issue, which could cause biased estimation of the cluster-specific parameters. We adopt Stephen's method (Stephens, 2000; Papastamoulis, 2015) to resolve this issue.

Let $\boldsymbol{P}^{(t)}$ be a $n \times G$ matrix with element $p_{ig}^{(t)}$ representing posterior probabilities of a subject i belonging to cluster g at MCMC iteration t; $t = 1, 2, \ldots, T$, where n is total number of subjects, T is the total MCMC iteration times and G is the total number of clusters. The basic idea of Stephen's method is trying to permute the estimates from the MCMC iterations if needed so that those posterior probability matrix $\boldsymbol{P}^{(t)}$ agree each other for $t = 1, 2, \ldots, T$. Specifically, Stephen's Method minimizes the Kullback-Leibler divergence (also called *relative entropy*) between $\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{P}^{(t)}$ and $\boldsymbol{P}^{(t)}$ for each t.

The relabeling algorithm is summarized as follows:

1. Assign an initial cluster labeling for all the MCMC iterations. We let $\boldsymbol{U}$ be a $T \times G$ matrix whose tth row represents the cluster labels for the tth iteration. For example, if $t = 1$ and $G = 3$, the cluster labels for the first MCMC iteration could be $(1, 2, 3)$, $(1, 3, 2)$, $(2, 1, 3)$, $(2, 3, 1)$, $(3, 1, 2)$ or $(3, 2, 1)$. To assign an initial clustering labeling is equivalent to assigning an initial values of the matrix $\boldsymbol{U}$.

2. Based on the current cluster labels, we calculate the mean posterior probability of cluster membership for all subjects across all iterations by averaging $\boldsymbol{P}^{(1)}, \boldsymbol{P}^{(2)}, \ldots, \boldsymbol{P}^{(T)}$. That is, the mean posterior probability of the gth cluster membership for subject i has the form:

$$\bar{p}_{ig} = \frac{1}{T} \sum_{t=1}^{T} p_{ig}^{(t)},$$

for $i = 1, 2, ..., n$ and $g = 1, 2, ..., G$, where $p_{ig}^{(t)}$ is the (i, g)th element of the matrix $\boldsymbol{P}^{(t)}$.

3. Update each row of the matrix $\boldsymbol{U}$, the cluster labels of an iteration, by minimizing

$$\sum_{i=1}^{n} \sum_{g=1}^{G} p_{ig}^{(t)} \log(\frac{p_{ig}^{(t)}}{\bar{p}_{ig}}),$$

for $t = 1, 2, \ldots, T$.

4. Repeat Steps 2 and 3 until matrix $\boldsymbol{U}$ is unchanged. Let $\boldsymbol{\Psi}$ be the $T \times G$ matrix of a parameter that needs to be estimated, where T is the total number of MCMC iterations and G is the total number of clusters. We will obtain the final parameter estimates for each

iteration (each row of $\boldsymbol{\Psi}$) by permuting the cluster labels based on the values in matrix $\boldsymbol{U}$.

3.4 SIMULATION STUDIES

We examined the properties of our proposed method under simulation 1(a), 1(b), 2(a) and 3 from the first part of dissertation. Simulation 2(b) (continuous variables follow arbitrary distribution) was not examined since it violates the distributional assumptions. For simulation 1(b), we removed 4 variables that are not distinguishable across clusters for simplicity. To evaluate clustering performance of different methods in the presence of censored biomarker, we selected simulation 1(a) where most methods performed well under without censored variables. We examined combination of varying censoring proportions and varying number of censored variables.

For each setting, 500 datasets were generated. Our hyper-parameters used in simulations are as follows: $(0.33, 0.33, 0.33)^T$ was used for $\boldsymbol{\delta}$ assuming we do not have much informative on $\boldsymbol{\tau}$; 0 and 100 were used for μ_A and σ_A^2 assuming we do not have much informative on A_{1m}; 2 and 0.0001 were used for a_{Δ_0} and b_{Δ_0} as we believe $\sigma^2_{\Delta_0}$ should be a small number so that the corresponding prior for μ_{mg} is a "spike" one; 1000 was used for $\sigma^2_{\Delta_1}$ as we believe $\sigma^2_{\Delta_1}$ should be a large number so that the corresponding prior for μ_{mg} is a "slab" one; 2 and 1 were used for $\tilde{a}$ and $\tilde{b}$ assuming we do not have much informative on γ_m; 1 for $\boldsymbol{\alpha}_{\Delta_1}$ so that the corresponding prior for θ_{mg} is a "slab" one. We used $(10, 10, 10)^T$ for $\boldsymbol{\alpha}_{m\Delta_0}$, where $m = 1, 2, \ldots, M$, since every categorical variable has 3 categories with true marginal probabilities $(0.33, 0.33, 0.33)^T$. We used 1 and 2 for a_{p_1} and b_{p_1}, 1 and 2 for a_{p_2} and b_{p_2} as we would like some extent of shrinkage on both continuous and categorical variables.

3.4.1 Without existence of censored biomarker variables

Clustering performances in terms of ARI are shown in Table 6. The first row, Bayesian FMM, is our proposed Bayesian FMM with variable selection. All other rows are the same with Table 3 we presented in the first part of this dissertation. In addition, the median and 2.5^{th} to 97.5^{th} percentile interval of variable weights obtained from our proposed Bayesian FMM with variable selection across all simulated datasets are shown in Table 7.

In simulation 1(a), we simulated a total of 5 variables: 4 continuous and 1 categorical. All except one continuous variable truly contribute to clustering. The sole categorical variable also contributes to clustering. Our Bayesian FMM with variable selection performed the best (ARI 1.00 [0.98, 1.00]) compared with other methods. Table 7 shows that the median weights of variables x_1, x_2, x_3 and x_5 are 1.00 with 2.5^{th} and 97.5^{th} percentile interval $(1.00, 1.00)$, indicating that these variables were found to be important for clustering. Meanwhile, variable x_4 had low weights (0.01 [0.01, 0.06]).

In simulation 1(b), we simulated a total of 10 variables: 7 continuous and 3 categorical. Six out of seven continuous variables truly contribute to clustering; two out of three categorical variables contribute to clustering. Our Bayesian FMM with variable selection performed the best (ARI 0.99 [0.96, 1.00]) compared with other methods. In Table 7, variables x_1, x_4 and x_7 had weights 0.05 (0.04, 0.07), 0.04 (0.03, 0.04), 0.05 (0.04, 0.07) indicating none of them has contribution to clustering; variables x_3, x_5, x_6 and x_8 had weights 0.10 (0.05, 0.19), 0.12 (0.06, 0.16), 0.11 (0.06, 0.17) and 0.12 (0.08, 0.22) respectively indicating that they have small contributions to clustering; variable x_2 had weight 0.30 (0.11, 0.41) indicating that it had slightly larger importance than x_3, x_5, x_6 and x_8. Variables x_9 and x_{10} had weights 1.00 (1.00, 1.00) indicating they're dominant variables that are highly relevant to clustering. These weights correctly reflect the true setting.

Table 6: Performance comparison in different simulation settings

Clustering Method	**ARI, median (2.5th percentile, 97.5th percentile)**			
	Sim 1(a)	Sim 1(b)	Sim 2(a)	Sim 3
Bayesian FMM	1.00 (0.98, 1.00)	0.99 (0.96, 1.00)	1.00 (0.98, 1.00)	0.74 (0.65, 0.83)
HyDaP	0.97 (0.92, 1.00)	0.95 (0.87, 1.00)	0.98 (0.94, 1.00)	0.75 (0.63, 0.85)
PAM + Gower distance	0.70 (0.58, 0.80)	0.87 (0.76, 0.96)	0.01 (-0.01, 0.04)	0.71 (0.31, 0.84)
K-prototypes	1.00 (0.96, 1.00)	0.93 (0.79, 1.00)	0.98 (0.92, 1.00)	0.17 (-0.01, 0.26)
Finite mixture model	1.00 (0.98, 1.00)	0.98 (0.44, 1.00)	1.00 (0.56, 1.00)	0.72 (0.56, 0.85)
PAM + FAMD distance	0.78 (0.66, 0.89)	0.93 (0.84, 0.98)	0.09 (-0.01, 0.44)	0.73 (0.22, 0.84)

In simulation 2(a), we simulated a total of 5 variables: 4 continuous and 1 categorical. This setting is the same as setting 1(a) except that the sole categorical variable does not contribute to clustering. Our Bayesian FMM with variable selection performed the best (ARI 1.00 [0.98, 1.00]) compared with other methods. Variables x_1 to x_4 exhibited similar weights as those in simulation 1(a), but variable x_5 now had a low weight (0.12 [0.08, 0.23]), reflecting the underlying setting.

In simulation 3, we simulated a total of 7 variables: 4 continuous and 3 categorical. None of the continuous variables truly contribute to clustering. Two out of three categorical variables contribute to clustering. Our Bayesian FMM with variable selection performed similarly (ARI 0.74 [0.65, 0.83]) with the best performer: HyDaP algorithm (ARI 0.75 [0.63, 0.85]). Table 7 shows that variables x_1 to x_4 all had weights close to 0 as in the true setting none of them are distinguishable across clusters. While categorical variable x_5 had very low weight as it has very small differences across clusters in true setting. Meanwhile, variables x_6 and x_7 which truly are associated with clustering had high weights (1.00 [1.00, 1.00]).

Table 7: Obtained variable weights in different simulation settings

Variable	**Weight, median (2.5th percentile, 97.5th percentile)**			
	Sim 1(a)	Sim 1(b)	Sim 2(a)	Sim 3
x_1	1.00 (1.00, 1.00)	0.05 (0.04, 0.08)	1.00 (0.95, 1.00)	0.00 (0.00, 0.02)
x_2	1.00 (1.00, 1.00)	0.30 (0.11, 0.43)	1.00 (1.00, 1.00)	0.00 (0.00, 0.03)
x_3	1.00 (1.00, 1.00)	0.10 (0.05, 0.18)	1.00 (1.00, 1.00)	0.01 (0.00, 0.08)
x_4	0.01 (0.01, 0.06)	0.04 (0.03, 0.04)	0.01 (0.01, 0.06)	0.00 (0.00, 0.03)
x_5	1.00 (1.00, 1.00)	0.12 (0.06, 0.15)	0.12 (0.08, 0.24)	1.00 (1.00, 1.00)
x_6		0.11 (0.06, 0.16)		0.12 (0.09, 0.21)
x_7		0.05 (0.04, 0.07)		1.00 (1.00, 1.00)
x_8		0.12 (0.08, 0.22)		
x_9		1.00 (1.00, 1.00)		
x_{10}		1.00 (1.00, 1.00)		

In summary, the clustering performance of our Bayesian FMM is always the top performer across all simulation scenarios. In addition, our Bayesian FMM with variable selection is able to provide quantitative variable importance which is more informative than the dichotomous information we obtained using the HyDaP algorithm, especially for simulation 1(b).

Table 8: Clustering performance with existence of censored biomarker variables

Clustering Method	**ARI, median (2.5th percentile, 97.5th percentile)**			
	Censored variables: x_3, x_4		Censored variables: x_1 - x_4	
	20%^{a}	50%^{a}	20%^{a}	50%^{a}
Bayesian FMM	1.00 (0.98, 1.00)	0.99 (0.96, 1.00)	1.00 (0.97, 1.00)	0.92 (0.42, 0.98)
Naive Bayesian FMM	1.00 (0.98, 1.00)	0.99 (0.44, 1.00)	0.99 (0.65, 1.00)	0.63 (0.46, 0.68)
HyDaP	0.96 (0.92, 0.99)	0.93 (0.89, 0.98)	0.90 (0.84, 0.96)	0.85 (0.75, 0.94)
PAM + Gower distance	0.62 (0.53, 0.71)	0.69 (0.59, 0.78)	0.72 (0.57, 0.80)	0.74 (0.65, 0.82)
K-prototypes	0.99 (0.96, 1.00)	0.98 (0.93, 1.00)	0.90 (0.63, 0.97)	0.60 (0.54, 0.66)

^{a}censoring proportion of each censored variable

3.4.2 With existence of censored biomarker variables

We evaluated the clustering performance of our proposed Bayesian FMM with variable selection under simulation setting 1(a), where existing methods performed well. We also evaluated performances of (1) our Bayesian FMM together with the naive method which uses half of lower detection limit to fill in undetected biomarker values; (2) the HyDaP algorithm together with the naive method; (3) PAM with Gower distance together with the naive method; and (4) K-prototypes together with the naive method. We designed 4 scenarios: (1) only variable x_3 and x_4 are censored and both with 20% censoring; (2) only variable x_3 and x_4 are censored and both with 50% censoring; (1) all continuous variables are censored and each with 20% censoring; (2) all continuous variables are censored and each with 50% censoring. Clustering results are shown in Table 8 and obtained variable weights using proposed Bayesian FMM with variable selection are shown in Table 9.

In Table 8 we can observe that except the last column, our Bayesian FMM with variable selection performed the best. When all the continuous variables have very high censoring proportions, our method performs typically well but with higher variability from simulation to simulation (ARI 0.92 [0.42, 0.98]). Using Bayesian FMM with naive method to fill in censored values instead of using our embedded sampling approach, the performance is worse. Meanwhile, performance of the HyDaP algorithm was consistently good, though not the best, across all scenarios. When all the continuous variables have large censoring proportions, it can still yield high ARI with narrower percentile interval (ARI 0.85 [0.75, 0.94]). PAM with Gower distance performed poorly in all scenarios. Performance of K-prototypes is satisfactory when only a few continuous variables are censored, but becomes worse when all continuous variables are censored. In terms of variable weights, Table 9 indicates that our Bayesian FMM with embedded sampling approach yields weights that reflect the true setting for all scenarios.

Table 9: Obtained variable weights with existence of censored biomarker variables

Weight	**ARI, median (2.5th percentile, 97.5th percentile)**			
	Censored variables: x_3, x_4		Censored variables: x_1 - x_4	
	20%^{a}	50%^{a}	20%^{a}	50%^{a}
x_1	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)
x_2	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)
x_3	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)
x_4	0.03 (0.02, 0.14)	0.01 (0.01, 0.05)	0.02 (0.01, 0.10)	0.01 (0.01, 0.04)
x_5	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)

^{a}censoring proportion of each censored variable

In summary, when most continuous variables are censored with high probability, we suggest using the HyDaP algorithm since it would provide robust performance under this scenario. In all other scenarios, our proposed Bayesian FMM with variable selection is preferred as it would yield the best results.

3.5 APPLICATION

We reanalyzed the SENECA data that was used in Chapter 2 of this dissertation and assumed that the data contains three latent clusters, which was obtained from the optimal number of clusters under the FMM framework.

The simulations in Chapter 2 of this dissertation showed that when data contains high conditional correlations the clustering performance of the FMM could be affected because that the method assumes conditional independency among variables. For the subsequent analysis, we dropped variables heart rate, blood sodium level (Na), aspartate aminotransferase (AST), and blood urea nitrogen (BUN) because of their high correlations to other variables in the dataset. The goals of our analysis are to obtain quantitative weights for all variables in the analytic dataset and to assess whether the identified clusters are related to those identified by the HyDaP algorithm.

Our hyper-parameters used in the analysis are as follows: $\boldsymbol{\delta} = (0.33, 0.33, 0.33)^T$ was used assuming that we do not have much information on $\boldsymbol{\tau}$. We set $\mu_A = 0$ and $\sigma^2_A = 1,000$ assuming that we do not have much information on A_{1m}. To set $a_{\Delta_0} = 2$ and $b_{\Delta_0} = 0.005$ as we believe that $\sigma^2_{\Delta_0}$ should be a small number so that the corresponding prior for μ_{mg} is a *spike* one. We set $\sigma^2_{\Delta_1} = 100$ as we believe that $\sigma^2_{\Delta_1}$ should be a large number so that the corresponding prior for μ_{mg} is a *slab* one. We also let $\tilde{a} = 2$ and $\tilde{b} = 1$ assuming that we do not have much information on γ_m. $\boldsymbol{\alpha}_{\Delta_1} = 1$ so that the corresponding prior for θ_{mg} is a *slab* one. We set $\boldsymbol{\alpha}_{\Delta_0} = (10.2, 10)^T$ for gender and $\boldsymbol{\alpha}_{\Delta_0} = (15.5, 2.4, 2.1)^T$ for race based on their marginal distributions. We would like some shrinkage on both continuous and categorical variables, so we set $a_{p_1} = 1$, $b_{p_1} = 2$, $a_{p_2} = 1$, and $b_{p_2} = 2$.

Table 10: A comparison between clusters identified by the Bayesian finite mixture model and the HyDaP algorithm

		HyDaP			
		Cluster 1	Cluster 2	Cluster 3	Cluster 4
Bayesian FMM	Cluster 1	5083 (50.9%)	2614 (26.2%)	1729 (17.3%)	557 (5.6%)
	Cluster 2	1241 (16.7%)	2411 (32.4%)	3073 (41.3%)	722 (9.7%)
	Cluster 3	156 (5.7%)	238 (8.6%)	730 (26.5%)	1635 (59.3%)

By applying our proposed Bayesian FMM, we obtained 3 clusters with sample sizes $9,983$, $7,447$, and $2,759$, respectively. Table 10 shows these results cross tabulating with the 4 clusters that were identified using the HyDaP algorithm. Cluster 1 and Cluster 3 identified by the Bayesian FMM are similar to Cluster 1 and Cluster 4 identified by the HyDaP algorithm, respectively. Cluster 2 of the Bayesian FMM is similar to Clusters 2 and 3 altogether from the HyDaP algorithm. The results show that the 3 clusters identified by using the proposed Bayesian FMM are consistent with those identified from the HyDaP algorithm.

Table 11 summarizes the clustering weight of each variable obtained from a Bayesian FMM. None of the variables had driving influences on the final clustering results as their weights were low. This finding is consistent with what we found using the HyDaP algorithm. An advantage of using our proposed Bayesian FMM in ranking variable importance is its ability to provide a real-valued weight for each variable, not just dichotomizing certain variables as important or not. In SENECA, we found that ESR, troponin, lactate, sex, albumin, bicarbonate, GCS, and INR had relatively high weights, which mean that these variables contributed more in forming the 3 clusters than other variables did. Such information provides more details than the general importance obtained from the HyDaP algorithm. This is especially useful for data under the *partitioned cluster structure*.

Table 11: Clustering weights of the SENECA variables using a Bayesian finite mixture model

Variable	**Weight**	**Variable**	**Weight**	**Variable**	**Weight**
Age	0.05	GCS	0.12	CRP	0.10
Temperature	0.05	Elixhauser score	0.04	INR	0.12
Systolic blood pressure	0.11	White blood cell	0.06	Glucose	0.05
Respiration rate	0.08	Bands	0.08	Platelets	0.05
Albumin	0.13	Creatinine	0.05	SaO2	0.05
Cl	0.07	Bilirubin	0.09	PaO2	0.08
ESR	0.30	Troponin	0.27	Gender	0.20
Hemoglobin	0.06	Lactate	0.22	Race	0.06
Bicarbonate	0.12	ALT	0.11		

Abbreviation: ESR: Erythrocyte sedimentation rate; GCS: Glasgow coma scale; ALT: Alanine aminotransferase; CRP: C-reactive protein; INR: International normalized ratio; SaO2: Oxygen saturation; PaO2: Partial pressure of oxygen

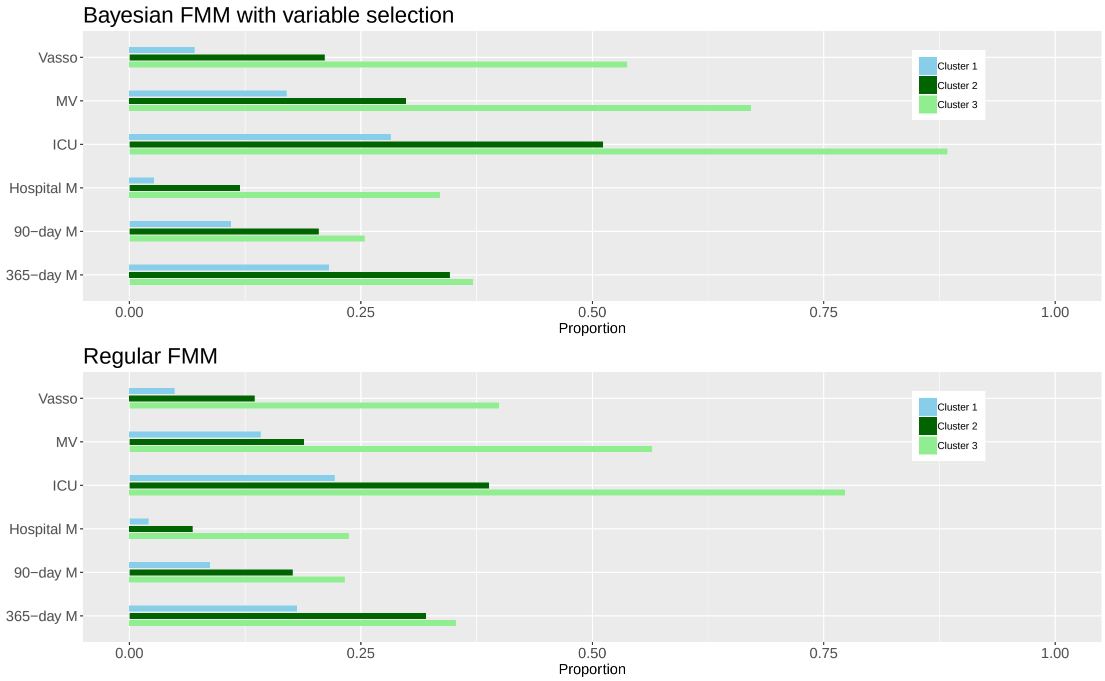


Figure 11: Distributions of selected clinical endpoints for the three clusters identified by various Bayesian finite mixture models

Table 12 and Figure 11 show the distributions of clinical endpoints across the 3 clusters obtained by Bayesian FMM with variable selection. We observe that these are similar to those identified by the traditional FMM but our method is able to provide weights of all involved variables so that we have better knowledge about their importance and better interpretation of clustering results.

Table 12: Distributions of selected clinical endpoints for the three clusters identified by a Bayesian finite mixture model

Clinical Endpoints	All Clusters	Cluster 1	Cluster 2	Cluster 3	p-value
	Bayesian FMM with variable selection				
Cluster size	20189	9984 (49.4%)	7447 (36.9%)	2759 (13.7%)	
Admitted to ICU	9063 (44.9%)	2817 (28.2%)	3809 (51.1%)	2437 (88.3%)	<0.001
Mechanical Ventilation	5773 (28.6%)	1694 (17.0%)	2227 (29.9%)	1852 (67.1%)	<0.001
Vasopressor	3755 (18.6%)	703 (7.0%)	1568 (21.1%)	1484 (53.8%)	<0.001
In-hospital mortality	2081 (10.3%)	267 (2.7%)	889 (11.9%)	926 (33.6%)	<0.001
90-day mortality[a] (exclude in hospital mortality)	2758 (14.2%)	1029 (11.0%)	1286 (20.4%)	443 (25.4%)	<0.001
365-day mortality[b] (exclude in hospital mortality)	5043 (27.9%)	2096 (21.6%)	2268 (34.6%)	679 (37.0%)	<0.001
	Traditional FMM				
Cluster size	20189	6410 (31.7%)	7581 (37.6%)	6198 (30.7%)	
Admitted to ICU	9063 (44.9%)	1455 (22.7%)	2882 (38.0%)	4726 (76.3%)	<0.001
Mechanical Ventilation	5773 (28.6%)	933 (14.6%)	1436 (18.9%)	3404 (54.9%)	<0.001
Vasopressor	3755 (18.6%)	324 (5.1%)	1016 (13.4%)	2415 (39.0%)	<0.001
In-hospital mortality	2081 (10.3%)	109 (1.7%)	486 (6.4%)	1487 (24.0%)	<0.001
90-day mortality[a] (exclude in hospital mortality)	2758 (14.2%)	504 (8.3%)	1192 (17.4%)	1061 (23.6%)	<0.001
365-day mortality[b] (exclude in hospital mortality)	5043 (27.9%)	1109 (17.6%)	2200 (31.0%)	1734 (36.8%)	<0.001

[a] Total number is 17,432 after excluding in hospital death and missing.

[b] Total number is 18,107 after excluding in hospital death.

3.6 DISCUSSION

We proposed a Bayesian finite mixture model (FMM) that can simultaneously cluster variables with mixed types, calculate variable weights, as well as handle censored biomarker variables. In this method we apply a Bayesian framework in order to bypass the limitations in the EM algorithm which is the standard estimation method for a FMM. In addition to identifying clusters, our model can provide real-valued variable weights which are more informative than a dichotomy of a variable being important vs. not in clustering. In addition, our proposed method is able to handle censored biomarker variables through recovering their underlying distributions. For a dataset of variables with mixed types but without censoring, our proposed Bayesian FMM performs better than the HyDaP algorithm and other existing methods across various simulation settings. If censored variables exist in the data, proposed Bayesian FMM performs better than other clustering algorithms using ad-hoc imputation methods. However, when variables have high censoring proportions, the proposed Bayesian FMM may not consistently outperform other approaches. In this case, the HyDaP algorithm with ad-hoc imputations provides better and robust results. Under all scenarios, our proposed Bayesian FMM with variable selection is able to provide reasonable weights of all variables.

Users of the proposed Bayesian FMM model need to be aware of certain limitations. Same as finite mixture models, the proposed model also has the disadvantage of unverifiable distributional assumptions. Besides, the model assumes conditional independence (i.e., variables are independent conditional on cluster membership) so it may not perform well when variables are subject to within-cluster correlations. In computations, because that we need to specify the values for all hyper-parameters before running the algorithm, the computation time for the proposed model is usually longer than that for the EM algorithm.

4.0 DISCUSSION AND FUTURE WORK

Clustering has received a lot of attention and been applied in various areas these days. However, clustering methods that can handle mixed types of variables (both continuous and categorical) are still limited. The goal of this dissertation is to develop methods that can cluster data with mixed types of variables and perform variable selection in clustering. The two methods we proposed achieve this goal from different perspectives: (1) the HyDaP algorithm is a nonparametric approach while the Bayesian FMM with variable selection is a model-based one; (2) variable selection procedure of the HyDaP algorithm is a filter method while the Bayesian FMM with variable selection is a wrapper method. Overall, our HyDaP algorithm and Bayesian FMM with variable selection perform better than other existing methods under the three data structures we specified in the dissertation. Other existing methods may fail under at least one type of data structures. Besides, both the HyDaP algorithm and the Bayesian FMM with variable selection provide the information on variable importance for clustering which is usually of interest for many researchers, particularly in interpreting clustering results.

Future work can be done on the top of the development of this dissertation. For the HyDaP algorithm, choice of the optimal number of clusters can be further explored. We can also extend the method to account for outliers or zero-inflated data. For the Bayesian FMM with variable selection, the framework can be extended to account for more distributions. Approaches of handling censored variables can also be further developed.

APPENDIX

VARIABLES USED IN SENECA DATA ANALYSIS

Age

Gender: categorical variable; 2 levels (male/female)

Race: categorical variables; 3 levels (white/black/hispanic)

Maximum temperature within 6 hours of ER presentation

Maximum heart rate within 6 hours of ER presentation

Minimum systolic blood pressure within 6 hours of ER presentation

Maximum respiration rate within 6 hours of ER presentation

Maximum albumin within 6 hours of ER presentation

Maximum Cl within 6 hours of ER presentation

Maximum erythrocyte sedimentation rate (ESR) within 6 hours of ER presentation

Maximum hemoglobin within 6 hours of ER presentation

Maximum bicarbonate within 6 hours of ER presentation

Maximum Sodium within 6 hours of ER presentation

Minimum Glasgow Coma Scale (GCS) within 6 hours of ER presentation

Elixhauser Score

Maximum white blood cell within 6 hours of ER presentation

Maximum bands within 6 hours of ER presentation

Maximum creatinine within 6 hours of ER presentation

Maximum bilirubin within 6 hours of ER presentation

Maximum troponin within 6 hours of ER presentation

Maximum lactate within 6 hours of ER presentation

Maximum alanine aminotransferase (ALT) within 6 hours of ER presentation

Maximum aspartate aminotransferase (AST) within 6 hours of ER presentation

Maximum C-reactive protein within 6 hours of ER presentation

Maximum international normalized ratio (INR) within 6 hours of ER presentation

Maximum glucose within 6 hours of ER presentation

Maximum Platelets within 6 hours of ER presentation

Maximum blood urea nitrogen (BUN) within 6 hours of ER presentation

Oxygen saturation (SaO2)

Minimum partial pressure of oxygen (PaO2) within 6 hours of ER presentation

BIBLIOGRAPHY

Ankerst, M., Breunig, M. M., Kriegel, H.-P., and Sander, J. (1999). Optics: ordering points to identify the clustering structure. In *ACM Sigmod record*, volume 28, pages 49–60. ACM.

Bernhardt, P. W., Wang, H. J., and Zhang, D. (2015). Statistical methods for generalized linear models with covariates subject to detection limits. *Statistics in biosciences*, 7(1):68–89.

Bhattacharya, S. and McNicholas, P. D. (2014). A lasso-penalized bic for mixture model selection. *Advances in Data Analysis and Classification*, 8(1):45–61.

Biernacki, C., Celeux, G., and Govaert, G. (2003). Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics & Data Analysis*, 41(3-4):561–575.

Blum, A. L. and Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial intelligence*, 97(1-2):245–271.

Celeux, G., Maugis-Rabusseau, C., and Sedki, M. (2018). Variable selection in model-based clustering and discriminant analysis with a regularization approach. *Advances in Data Analysis and Classification*, pages 1–20.

Chang, W.-C. (1983). On using principal components before separating a mixture of two multivariate normal distributions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 32(3):267–275.

Coorevits, P., Sundgren, M., Klein, G. O., Bahr, A., Claerhout, B., Daniel, C., Dugas, M., Dupont, D., Schmidt, A., Singleton, P., et al. (2013). Electronic health records: new opportunities for clinical research. *Journal of internal medicine*, 274(6):547–560.

Council, N. R. et al. (2011). *Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease.* National Academies Press.

Deb, P. et al. (2008). Finite mixture models. *Hunter College and the Graduate Center, CUNY NBER, FMM Slides*, 42.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 363–375.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231.

Fop, M., Murphy, T. B., et al. (2018). Variable selection methods for model-based clustering. *Statistics Surveys*, 12:18–65.

Frühwirth-Schnatter, S. (2001). Markov chain monte carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, 96(453):194–209.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741.

George, E. I. and McCulloch, R. E. (1997). Approaches for bayesian variable selection. *Statistica sinica*, pages 339–373.

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

Group, B. D. W., Atkinson Jr, A. J., Colburn, W. A., DeGruttola, V. G., DeMets, D. L., Downing, G. J., Hoth, D. F., Oates, J. A., Peck, C. C., Schooley, R. T., et al. (2001). Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical Pharmacology & Therapeutics*, 69(3):89–95.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182.

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Haripriya, H., Amrutha, S., Veena, R., and Nedungadi, P. (2015). Integrating apriori with paired k-means for cluster fixed mixed data. In *Proceedings of the Third International Symposium on Women in Computing and Informatics*, pages 10–16. ACM.

Häyrinen, K., Saranto, K., and Nykänen, P. (2008). Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *International journal of medical informatics*, 77(5):291–304.

Hennig, C. and Liao, T. F. (2013). How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(3):309–369.

Honoré, B. E. (1992). Trimmed lad and least squares estimation of truncated and censored regression models with fixed effects. *Econometrica: journal of the Econometric Society*, pages 533–565.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, 2(3):283–304.

Ishwaran, H., Rao, J. S., et al. (2005). Spike and slab variable selection: frequentist and bayesian strategies. *The Annals of Statistics*, 33(2):730–773.

Jensen, P. B., Jensen, L. J., and Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395.

Karlis, D. and Xekalaki, E. (2003). Choosing initial values for the em algorithm for finite mixtures. *Computational Statistics & Data Analysis*, 41(3-4):577–590.

Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons.

Law, M. H., Figueiredo, M. A., and Jain, A. K. (2004). Simultaneous feature selection and clustering using mixture models. *IEEE transactions on pattern analysis and machine intelligence*, 26(9):1154–1166.

Lee, M., Kong, L., and Weissfeld, L. (2012). Multiple imputation for left-censored biomarker data based on gibbs sampling method. *Statistics in medicine*, 31(17):1838–1848.

Li, Y., Dong, M., and Hua, J. (2009). Simultaneous localized feature selection and model detection for gaussian mixtures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):953–960.

Liu, J. S., Zhang, J. L., Palumbo, M. J., and Lawrence, C. E. (2003). Bayesian clustering with variable and transformation selections. *Bayesian statistics*, 7:249–275.

Lubin, J. H., Colt, J. S., Camann, D., Davis, S., Cerhan, J. R., Severson, R. K., Bernstein, L., and Hartge, P. (2004). Epidemiologic evaluation of measurement data in the presence of detection limits. *Environmental health perspectives*, 112(17):1691.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association*, 89(428):1535–1546.

Marin, J.-M., Mengersen, K., and Robert, C. P. (2005). Bayesian modelling and inference on mixtures of distributions. *Handbook of statistics*, 25:459–507.

McCutcheon, A. L. (1987). *Latent class analysis*. Number 64. Sage.

McLachlan, G. and Krishnan, T. (2007). *The EM algorithm and extensions*, volume 382. John Wiley & Sons.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032.

Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning*, 52(1-2):91–118.

Moustaki, I. (1996). A latent trait and a latent class model for mixed observed variables. *British journal of mathematical and statistical psychology*, 49(2):313–334.

Nylund, K. L., Asparouhov, T., and Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study. *Structural equation modeling*, 14(4):535–569.

Pagès, J. (2014). *Multiple factor analysis by example using R*. Chapman and Hall/CRC.

Pan, W. and Shen, X. (2007). Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research*, 8(May):1145–1164.

Papastamoulis, P. (2015). label. switching: An r package for dealing with the label switching problem in mcmc outputs. *arXiv preprint arXiv:1503.02271*.

Papastamoulis, P. and Iliopoulos, G. (2010). An artificial allocations based solution to the label switching problem in bayesian analysis of mixtures of distributions. *Journal of Computational and Graphical Statistics*, 19(2):313–331.

Pierrakos, C. and Vincent, J.-L. (2010). Sepsis biomarkers: a review. *Critical care*, 14(1):R15.

Powell, J. L. (1984). Least absolute deviations estimation for the censored regression model. *Journal of Econometrics*, 25(3):303–325.

Raftery, A. E. and Dean, N. (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473):168–178.

Reddy, M. J. and Kavitha, B. (2012). Clustering the mixed numerical and categorical dataset using similarity weight and filter method. *International Journal of Database Theory and Application*, 5(1):121–134.

Shirkhorshidi, A. S., Aghabozorgi, S., and Wah, T. Y. (2015). A comparison study on similarity and dissimilarity measures in clustering continuous data. *PloS one*, 10(12):e0144059.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809.

Strimbu, K. and Tavel, J. A. (2010). What are biomarkers? *Current Opinion in HIV and AIDS*, 5(6):463.

Sun, J., Zhou, A., Keates, S., and Liao, S. (2018). Simultaneous bayesian clustering and feature selection through students mixtures model. *IEEE transactions on neural networks and learning systems*, 29(4):1187–1199.

Tadesse, M. G., Sha, N., and Vannucci, M. (2005). Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, 100(470):602–617.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society*, pages 24–36.

Wang, S. and Zhu, J. (2008). Variable selection for model-based high-dimensional clustering and its application to microarray data. *Biometrics*, 64(2):440–448.

White, A., Wyse, J., and Murphy, T. B. (2016). Bayesian variable selection for latent class analysis using a collapsed gibbs sampler. *Statistics and Computing*, 26(1-2):511–527.

Wilkerson, M. D. and Hayes, D. N. (2010). Consensusclusterplus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics*, 26(12):1572–1573.

Witten, D. M. and Tibshirani, R. (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490):713–726.

Xie, B., Pan, W., and Shen, X. (2008). Penalized model-based clustering with cluster-specific diagonal covariance matrices and grouped variables. *Electronic journal of statistics*, 2:168.

Xu, R. and Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678.

Zhou, H., Pan, W., and Shen, X. (2009). Penalized model-based clustering with unconstrained covariance matrices. *Electronic journal of statistics*, 3:1473.