Journal of Data Science 19 (1), 15–36
January 2021
DOI: 10.6339/21-JDS996
Statistical Data Science

Hybrid Density- and Partition-Based Clustering Algorithm for Data With Mixed-Type Variables

Shu Wang[1,2], Jonathan G. Yabes[3,4], and Chung-Chou H. Chang[3,4,*]

[1]*Department of Biostatistics, College of Public Health and Health Professions, University of Florida*
[2]*University of Florida Health Cancer Center*
[3]*Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh*
[4]*Department of Medicine, School of Medicine, University of Pittsburgh*

Abstract

Clustering is an essential technique for discovering patterns in data. Many clustering algorithms have been developed to tackle the ever increasing quantity and complexity of data, yet algorithms that can cluster data with mixed variables (continuous and categorical) remain limited despite the abundance of mixed-type data. Of the existing clustering methods for mixed data types, some posit unverifiable distributional assumptions or rest on unbalanced contributions of different variable types. To address these issues, we propose a two-step hybrid density- and partition-based (HyDaP) algorithm to detect clusters after variable selection. The first step involves both density-based and partition-based algorithms to identify the data structure formed by continuous variables and determine important variables (both continuous and categorical) for clustering. The second step involves a partition-based algorithm together with our proposed novel dissimilarity measure to obtain clustering results. Simulations across various scenarios were conducted to compare the HyDaP algorithm with other commonly used methods. Our HyDaP algorithm was applied to identify sepsis phenotypes and yielded important results.

Keywords *mixed data; variable selection*

1 Introduction

Sepsis is a potentially life-threatening complication to infection with high mortality (Angus and Van der Poll, 2013; Liu et al., 2014; Seymour et al., 2016). Although several potential therapies for sepsis were promising in animal models, these therapies showed either no significant benefits or conflicting results in humans. One difficulty in developing therapy for sepsis is that sepsis is not a single disease; It comprises of endotypes with different responses to treatment (Scicluna et al., 2017; Seymour et al., 2019). The objective of this study is to identify homogeneous subgroups of sepsis patients so that we can better understand the heterogeneity of sepsis endotypes. Further investigations on these subgroups together with current clinical guidelines could help physicians design precision medicine strategies for better patient care (Jensen et al., 2012).

Our study takes advantage of the vast and variety of data in the electronic health records (EHR) to address the knowledge gap existing in precision medicine for sepsis. We studied 20 189 patients recruited in the Sepsis ENdotyping in Emergency CAre (SENECA) project and aimed to explore whether clinical sepsis phenotypes are identifiable for a patient in the emergency department. Among the thirty clinical variables involved in the SENECA data, twenty-eight of them are continuous, and the other two are categorical. One of the main challenges in clustering

*Corresponding author. Email: changj@pitt.edu.

Received September 13, 2020; Accepted October 15, 2020

SENECA data is how to define *dissimilarity* between subjects in the data set as mixed variable types (continuous and nominal categorical) are involved.

Several existing clustering methods claimed to be able to handle mixed types of variables. Gower distance (Gower, 1971) was proposed to measure dissimilarity between subjects with mixed types of variables, but the clustering results using Gower distance are dominated by categorical variables, as shown in our simulations. Note that in this context we will use "distance" and "dissimilarity" interchangeably. The distance measure used in Factor Analysis of Mixed Data (FAMD) can be applied on mixed data as well, even though FAMD was not originally intended for clustering (Pagès, 2014), therefore, its performance is not guaranteed. Distance measure defined in K-prototypes (Huang, 1998) involves user-defined variable weights, but it assumes that all categorical variables have the same weight, and that all continuous variables have the same weight, which is not practical in most clinical applications. Finite mixture model (FMM) (McCutcheon, 1987; Moustaki, 1996) is a model-based clustering method that can be used to cluster mixed data. Assuming that the data is a mixture of several parametric distributions, one can bypass the challenge of defining dissimilarity between subjects and transfer the task of selecting the optimal number of clusters into model selection, which is a much more straightforward approach in practice. By using this approach, we need to make parametric assumptions for each variable. Unfortunately, the distributional assumptions are conditional on the unknown cluster assignment, making them hard to verify from the data itself.

Motivated by the challenges encountered in clustering EHR data with mixed types using the aforementioned existing methods, we proposed a novel dissimilarity measure for mixed variables and developed a Hybrid Density- and Partition-based (HyDaP) algorithm for clusters identification which also can select the most important variables that drive clustering results.

The organization of the rest of the paper is as follows. In Section 2 we review the most commonly used dissimilarity measures and clustering algorithms; In Section 3 we define three data structures and propose a new clustering algorithm, HyDaP; In Section 4 we present performance comparisons among different methods under various simulation settings; In Section 5 we demonstrate the use of the HyDaP algorithm in the SENECA data to identify sepsis phenotypes; And Section 6 is discussion.

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.

2.1 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:

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

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 $m \to \infty$, Equation (1) becomes 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:

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

where

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

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 $d_j(\mathbf{x}_i, \mathbf{x}_{i'})$ is defined as 0 when $x_{ij} = x_{i'j}$ and they are the level with larger similarity, $d_j(\mathbf{x}_i, \mathbf{x}_{i'})$ is defined as 1 otherwise.

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 data 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 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 inefficiency, 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 variable weights as it assumes equal importance for variables of the same type. Moreover, the tuning parameter γ is user-defined rather than data-driven.

Sparse clustering (Witten and Tibshirani, 2010) is an advanced clustering framework proposed for cluster identification as well as variable selection. The capability in performing data-driven variable selection in the sparse clustering method is acquired through incorporating a lasso-type penalty and variable weights to its objective function:

$$\begin{aligned} &\underset{\mathbf{w};\, \Theta \in D}{\text{maximize}} && \sum_{j=1}^{p} w_j f_j(\mathbf{X}_j; \Theta) \\ &\text{subject to} && \|\mathbf{w}\|^2 \leqslant 1, \|\mathbf{w}\|_1 \leqslant s, w_j \geqslant 0\ \forall j, \end{aligned}$$

where n is number of subjects; p is number of features; $\mathbf{w} = (w_1, w_2, ..., w_p)^T$ is the vector of weights; Θ is a vector of parameters restricted to be in set D; $f_j(\mathbf{X}_j; \Theta)$ is a function involving feature j only; and s is the $L1$ norm restriction, which is a tuning parameter in the algorithm. If we use the objective function of sparse clustering to optimize the weighted WCSS, we will obtain sparse K-means. Although tuning the parameter s may be a complicated task, the sparse K-means clustering method distinguishes itself for being able to identify clusters and assess the relative importance of variables concurrently.

The optimal number of clusters, K, can be determined through applying the consensus clustering framework (Monti et al., 2003; Wilkerson and Hayes, 2010), and assessing the number of clusters, cluster memberships, and stability of clusters discovered from multiple runs of the K-means algorithm with different values of K. In each run of the K-means algorithm, a consensus index between each pair of subjects is calculated as the ratio of the times the pair was assigned to the same cluster over the times both members in the pair were sampled. These consensus indices can serve as measurements of similarity. The determination of 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 the optimal number of K.

2.3 Hierarchical Clustering

Hierarchical clustering (Ward Jr, 1963) 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. Hierarchical clustering can be inserted into sparse clustering framework as well.

2.4 Density-Based Clustering

Another important category of clustering methods is density-based clustering. Some researchers group density-based algorithms under the partition-based category, but most of the researchers including us list them as a separate category. 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.

Both algorithms involve two parameters: ε and *MinPts*. ε represents the radius of an object's neighborhood and *MinPts* represents number of objects within this neighborhood. The rough idea of DBSCAN is that every object in a cluster has at least *MinPts* objects including itself in its neighborhood defined by ε (Ester et al., 1996; Ankerst et al., 1999). 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 capable of detecting clusters with varying densities and suitable for high dimensional data. In OPTICS, a *core object* is one having at least *MinPts* objects including itself within the ε neighborhood. While *core distance* is the smallest value that makes a *core object* to be a *core object*.

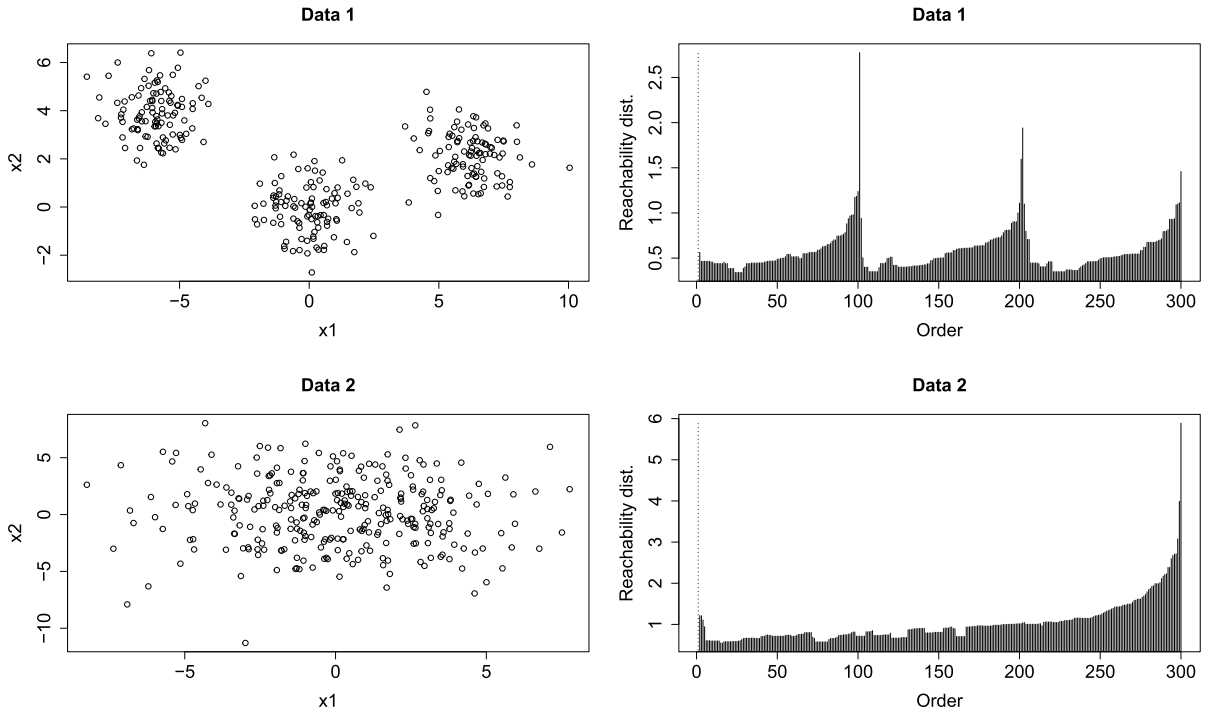

Figure 1: Illustration of different reachability plots.

If object A in the dataset is a *core object*, the algorithm will process all the objects in the ε neighborhood of object A from the nearest to the farthest. For each object processed, OPTICS also records its processing order within the dataset and calculates its *reachability distance* to the nearest *core object*. *Reachability distance* of object A with respect to a *core object* C, for example, is defined as the larger of the two values: The actual distance between objects A and C vs. object C's *core distance*. If A is not a *core object*, the algorithm will move on to the next object in the dataset that is not yet processed. After processing all the objects, OPTICS can generate a reachability plot for the dataset in which processing order and *reachability distance* are the horizontal axis and vertical axis, respectively.

Reachability plot provides an overall two-dimensional spatial structure of a dataset regardless of its original dimensions. Each trough on the reachability plot can be viewed as a single cluster. The edge between two side-by-side troughs represents the distance between two closest bordering objects from the corresponding two clusters. A higher edge implies that the corresponding two clusters are farther apart while a low or unclear edge implies that the two adjacent clusters are not much distinct from each other. A hypothetical example is depicted in Figure 1. From the left panel we can observe that data 1 contains 3 clusters whereas no natural cluster exists in data 2. On the right panel, we can observe 3 clear troughs in the reachability plot of data 1 and only 1 large trough in the reachability plot of data 2. This is consistent with what we observed from the scatter plots on the left panel. Although one can show clustering in data 1 and data 2 with scatter plots, the reachability plot is much easier to use for demonstrating the overall data structure when dealing with data in a high-dimensional space.

2.5 Model-Based Clustering

FMM is a model-based clustering method assuming that the data 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:

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

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 that the distributional assumptions are conditional on the unknown cluster assignment, making those assumptions hard to verify from the data.

2.6 Clustering Algorithms for Mixed Types of Variables

Except the algorithms mentioned in the introduction, 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 results (Reddy and Kavitha, 2012). This method impractically weighs continuous and categorical variables equally and ignores possible mutual influences between the two variable types.

3 Hybrid Density- and Partition-Based Clustering Algorithm

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 one) and a **clustering step** (step two). 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. The rationale that we use continuous variables only in pre-processing step to identify data structure is that we believe in general continuous variables provide richer information than categorical ones. The HyDaP algorithm is shown below.

Algorithm 1: HyDaP Algorithm.

```
Run a density-based algorithm on all continuous variables
if Number of clusters in the OPTICS reachability plot > 1 then
    It is defined as natural cluster structure
    Select continuous variables through Sparse K-means
    Select categorical variables through Cramer's V
else
    Run consensus K-means on all continuous variables
    if K> 1 from results of consensus K-means then
        It is defined as partitioned cluster structure
        Keep all continuous variables
        Select categorical variables through Cramer's V
    else
        It is defined as homogeneous structure
        Remove all continuous variables
        Select categorical variables through Cramer's V
    end
end
Run PAM on selected variables using proposed dissimilarity
```

3.1 Pre-Processing Step (Step One)

To help with variable selection and better understand the data set, we first define three data structures for the space spanned by the continuous variables as: *Natural cluster structure* (data structure one); *partitioned cluster structure* (data structure two); and *homogeneous structure* (data structure three). 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 two). A flow chart of step one is in online Supplementary Materials.

3.1.1 Definition of Three Data Structures

Data spanned in the covariate space of continuous variables can be divided into two scenarios: With and without 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. If we observe multiple troughs in a reachability plot, as illustrated in reachability plot of data 1 in Figure 1, 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 one) 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 1), 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 $\geqslant 2$ 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 two). If the optimal number of clusters is one from the consensus clustering results, this indicates that continuous part of the data is

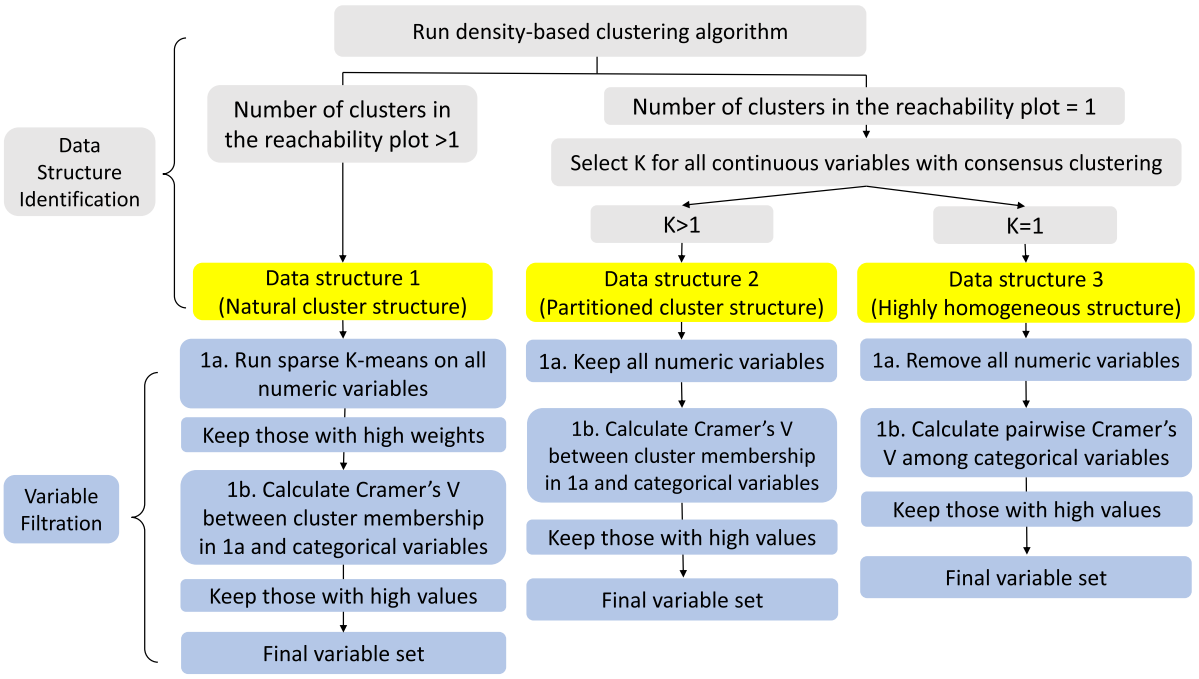

Figure 2: Flowchart of Step 1 of the HyDaP algorithm

highly homogeneous and cannot be further partitioned. We call this type of structure *homogeneous structure* (data structure three).

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 2, 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 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 is defined as:

$$\mathrm{V} = \sqrt{\frac{\chi^2/n}{\min(k-1, r-1)}},$$

where χ^2 is Pearson Chi-square test statistics, n is sample size, k and r are number of columns and rows of the contingency table, respectively. 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. Cramer's V exceeding 0.3 is defined as high values in this step indicating a moderate to strong association as 0.3 is a commonly used cutoff value. This is a general suggestion rather than a universal consensus cutoff. In fact, any number between 0.3 and 0.5 led to the same results in our simulation studies.

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 $\geqslant 2$ 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.

3.2 Clustering Step (Step Two)

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:

$$d(\mathbf{x}_i, \mathbf{x}_{i'}) = \sum_{j=1}^{p} \frac{d_j(\mathbf{x}_i, \mathbf{x}_{i'})}{\sum_{o \neq o'} d_j(\mathbf{x}_o, \mathbf{x}_{o'})}, \tag{3}$$

where $d_j(\mathbf{x}_i, \mathbf{x}_{i'})$ is the same as in Equation (2).

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.

3.3 Parameter 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*.

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.

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. When it is not that obvious to select variables based on weights, 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.

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 three 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. All three data structures were covered in simulations. Details of simulation settings are in online Supplementary Materials.

For each setting, 500 datasets were generated. Clustering was performed on each dataset using the proposed HyDaP algorithm. We compared its performance with PAM with Gower distance, PAM with FAMD distance, K-prototypes, and FMM assuming normal distribution for continuous variables and binomial/multinomial distribution for categorical variables. To compare our proposed dissimilarity with Gower distance, the HyDaP with Gower distance and PAM with proposed dissimilarity were also added in the simulation. To examine the impact of conditional correlation on clustering performance, additional simulations imbued with a pairwise correlation of 0.4 conditional on true cluster labels were conducted. Since we know the true cluster labels, the adjusted Rand index (ARI) (Rand, 1971; Hubert and Arabie, 1985) 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

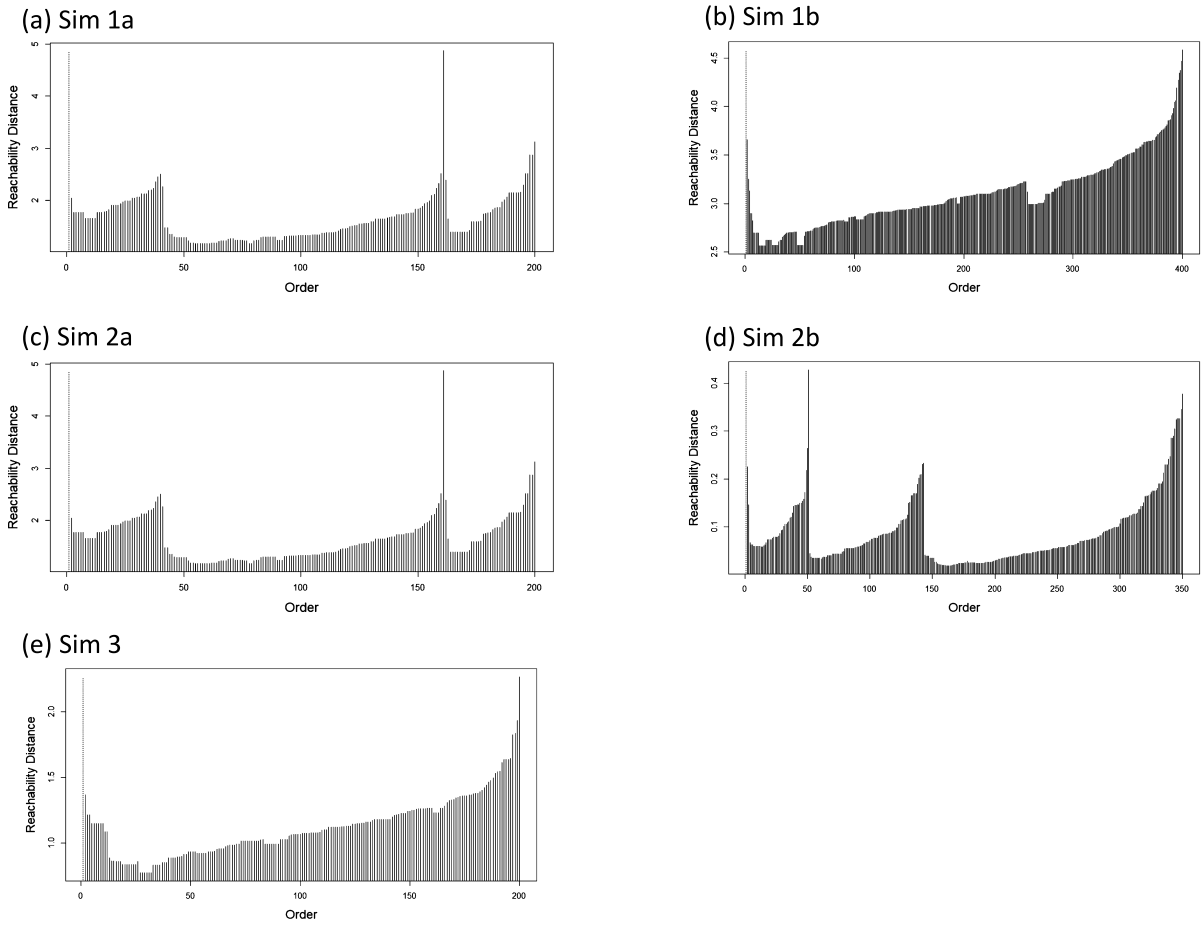

Figure 3: Reachability plots in different simulation settings.

better performance. Median along with the 2.5^{th} and 97.5^{th} percentiles were reported for all statistics.

4.1 Setting 1: Both Types of Variables Contribute to Clustering

4.1.1 Setting 1(a): Natural Cluster Structure

In simulation 1(a), we simulated a total of five variables: Four continuous and one categorical. All except one continuous variable truly contribute to clustering. The sole categorical variable also contributes to clustering.

In Step one of the HyDaP algorithm, the reachability plot (Figure 3a) indicated three clusters. This setting has natural cluster structure as three distinct clusters exist. Table 1 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 two, 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 2, performance of PAM with our proposed dissimilarity (ARI: 0.97 [0.90, 1.00]) is almost the same as the HyDaP algorithm (ARI: 0.97 [0.92, 1.00]). Similarly, performance of PAM with Gower distance (ARI: 0.70 [0.58, 0.80]) is the same as the HyDaP with Gower

Table 1: Results from pre-processing step in different simulation settings. All weights and Cramer's V values are presented in the form of median (2.5th percentile, 97.5th percentile).

Sim 1(a)	Sim 1(b)	Sim 2(a)	Sim 2(b)	Sim 3
		Data Structure		
Natural cluster	Partitioned cluster	Natural cluster	Natural cluster	Homogeneous structure
Weight	-	Weight	Weight	-
x_1: 0.49 (0.46, 0.52)		x_1: 0.58 (0.57, 0.58)	x_1: 0.54 (0.51, 0.57)	
x_2: 0.59 (0.58, 0.61)	Keep all	x_2: 0.56 (0.54, 0.57)	x_2: 0.51 (0.48, 0.54)	Drop all
x_3: 0.64 (0.62, 0.65)	continuous	x_3: 0.60 (0.59, 0.61)	x_3: 0.44 (0.38, 0.48)	continuous
x_4: 0.00 (0.00, 0.02)	variables	x_4: 0.00 (0.00, 0.02)	x_4: 0.50 (0.45, 0.54)	variables
			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: 0.09 (0.04, 0.17)	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)

distance (ARI: 0.70 [0.59, 0.80]). This is expected as HyDaP only removed one variable in this setting. In addition, these results indicate that our proposed dissimilarity is superior to Gower distance regardless of clustering algorithm as Gower distance tends to downplay contributions of continuous variables. 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 FAMD distance (ARI: 0.78 [0.66, 0.89]) performed poorly.

4.1.2 Setting 1(b): Partitioned Cluster Structure

In simulation 1(b), we simulated a total of fourteen variables: Eleven continuous and three categorical. Six out of eleven continuous variables truly contribute to clustering; Two out of three categorical variables contribute to clustering.

In Step one of the HyDaP algorithm, Figure 3b indicated that no natural clusters exists. After conducting consensus K-means, we chose three 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 two, PAM with proposed dissimilarity measure was applied on x_1, x_2,..., x_{11}, x_{13}, and x_{14} to obtain final results.

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. The HyDaP with Gower distance performed similarly well as noise categorical variable x_{12} was removed. On the other hand, PAM with Gower distance performed worse (ARI: 0.87 [0.76, 0.96]) as variable x_{12} was included and Gower distance tends to amplify its contribution. If Gower distance is replaced to our proposed dissimilarity, the performance becomes as good as the HyDaP (ARI: 0.93 [0.83, 0.99]). In this setting, K-prototypes (ARI: 0.93 [0.79, 0.95]) 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]).

Table 2: 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)	1.00 (1.00, 1.00)	0.98 (0.92, 1.00)	0.75 (0.63, 0.85)
HyDaP + Gower distance	0.70 (0.59, 0.80)	0.95 (0.88, 1.00)	1.00 (1.00, 1.00)	0.98 (0.92, 1.00)	0.71 (0.57, 0.82)
PAM + HyDaP dissimilarity	0.97 (0.90, 1.00)	0.93 (0.83, 0.99)	0.99 (0.90, 1.00)	0.34 (0.32, 0.98)	0.73 (0.28, 0.84)
PAM + Gower distance	0.70 (0.58, 0.80)	0.87 (0.76, 0.96)	0.00 (−0.01, 0.02)	0.23 (0.00, 0.34)	0.71 (0.31, 0.84)
PAM + FAMD distance	0.78 (0.66, 0.89)	0.93 (0.84, 0.98)	0.34 (−0.01, 0.42)	0.34 (0.08, 0.39)	0.73 (0.22, 0.84)
K-prototypes	1.00 (0.96, 1.00)	0.93 (0.79, 0.95)	1.00 (1.00, 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 (0.69, 1.00)	0.41 (0.33, 0.58)	0.72 (0.56, 0.85)

4.2 Setting 2: Only Continuous Variables Contribute to Clustering

4.2.1 Setting 2(a): Natural Cluster Structure

In simulation 2(a), we simulated a total of five variables: Four continuous and one categorical. This setting is the same as simulation 1(a) except that the sole categorical variable does not contribute to clustering.

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

In Step two, we applied the sparse K-means on x_1, x_2, and x_3 as they are all continuous variables. In this setting, the performance of the HyDaP algorithm (ARI: 1.00 [1.00, 1.00]) and the HyDaP with Gower distance (ARI: 1.00 [1.00, 1.00]) are the same as both of them became sparse K-means after all categorical variables were removed. However, PAM with Gower distance (ARI: 0.00 [−0.01, 0.02]) performed extremely poorly as clustering is driven by the single noise categorical variable. Replacing Gower distance with proposed dissimilarity provides satisfactory results (ARI: 0.99 [0.90, 1.00]). K-prototypes (ARI: 1.00 [1.00, 1.00]) also worked well. There were a few simulation runs of FMM whose performance was not satisfactory (ARI: 1.00 [0.69, 1.00]). PAM with FAMD distance (ARI: 0.34 [−0.01, 0.42]) performed poorly.

4.2.2 Setting 2(b): Natural Cluster Structure

In simulation 2(b), we simulated a total of eight variables: Five continuous and three 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 one of the HyDaP algorithm, Figure 3d shows three 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 1. All categorical variables were dropped as well given their weak associations with clusters obtained in the sparse K-means.

In Step two, we applied the sparse K-means on x_1, x_2, x_3, and x_4 since they are all continuous variables. In this setting, the performance of the HyDaP algorithm (ARI: 0.98 [0.92, 1.00]) and the HyDaP with Gower distance (ARI: 0.98 [0.92, 1.00]) are the same, similar to setting 2(a). PAM with proposed dissimilarity (ARI: 0.34 [0.32, 0.98]), 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.

4.3 Setting 3: Only Categorical Variables Contribute to Clustering

4.3.1 Setting 3: Homogeneous Structure

In simulation 3, we simulated a total of seven variables: Four continuous and three categorical. None of the continuous variables truly contributes to clustering. Two out of three categorical variables contribute to clustering.

In Step one of the HyDaP algorithm, Figure 3e indicates no natural clusters exist. After conducting consensus K-means, the optimal number of clusters chosen was one 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 1.

In Step two, 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 the HyDaP with Gower distance (ARI: 0.71 [0.57, 0.82]) is similar to the HyDaP as both of them only involved categorical variables in final clustering. Performance of PAM with proposed dissimilarity (ARI: 0.73 [0.28, 0.84]), 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]) are similar.

4.4 Variables Are Conditionally Correlated

To assess the impact of within-cluster correlation, simulations for each of the five settings above was repeated with pairwise correlation of 0.4 for all continuous variables conditional on true cluster labels. Results are summarized in online Supplementary Materials. Within-cluster correlation had little to no impact on all the methods except FMM. 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 two instead of three (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 two subgroups that have some differences. But this does not essentially mean these two subgroups can be viewed as two clusters. Therefore, if we observe that two is the optimal number of clusters in consensus clustering results and most pairs of continuous variables have high conditional correlations, we should be cautious. Our suggestion is 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.

4.5 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) Our proposed dissimilarity is superior to Gower distance as it better balances the contribution between continuous and categorical variables; (2) When categorical variables do not contribute much to clustering, PAM with Gower distance performed poorly; (3) When continuous variables follow arbitrary distributions, FMM may not perform well due to assumption violation; (4) When none of continuous variables contributes to clustering, K-prototypes may fail; (5) Performance of PAM with FAMD distance was not stable across different scenarios as its distance measure is not specifically designed for clustering.

4.6 Special Cases

In this section, we will discuss two special cases in which the proposed HyDaP algorithm may not perform well. Special case 1 occurs when categorical variables can further divide those clusters formed by continuous variables. Special case 2 occurs when continuous variables alone are not informative in clustering (i.e., continuous variables are homogeneous) but can detect clusters if used jointly with categorical variables.

4.6.1 Special Case 1

Consider two variables where x_1 is continuous and x_2 is categorical with three levels A, B, and C. If we use x_1 only, we will obtain two clusters with respect to the lower and the higher values of x_1 (bimodal), as shown in the top left panel of Figure 4. When x_2 is added, we can detect 4 clusters, with the lower values of x_1 divisible into categories A and B of x_2 and higher values of x_1 divisible into categories B and C of x_2, see top right panel of Figure 4. The HyDaP algorithm detects just two clusters because it can only use continuous variables x_1 to determine the optimal number of clusters. With this limitation in mind, we investigated the following: (1) Whether it is possible for other methods to select the correct number of clusters, and (2) how different clustering methods perform given the true number of clusters. In this simulation, we set 4 clusters of size 50 for a total sample of 200, also used 500 datasets to be consistent with the previous settings.

Under the assumption of unknown true number of clusters, we compared methods by finding the optimal number of clusters according to the corresponding criteria (WCSS, average silhouette, BIC, etc.). The results are that PAM with proposed HyDaP dissimilarity, K-prototypes, and PAM with FAMD distance chose 4 as the optimal number of clusters while PAM with Gower distance selected 3, and FMM selected 2 in most of the simulation runs. Note that the HyDaP and PAM with proposed HyDaP dissimilarity did the same in this setting because no variable was dropped in the HyDaP. Therefore, if we suspect that the categorical variables will further divide those clusters formed by the continuous variables, we can overcome the limitation of the HyDaP by trying different numbers of clusters and selecting the optimal one.

We further explored the clustering performance of all methods assuming that 4 is the optimal number of clusters. All methods performed similarly with a median ARI around 0.7 to 0.8 except FMM (ARI: 0.47 [0.39, 0.65]).

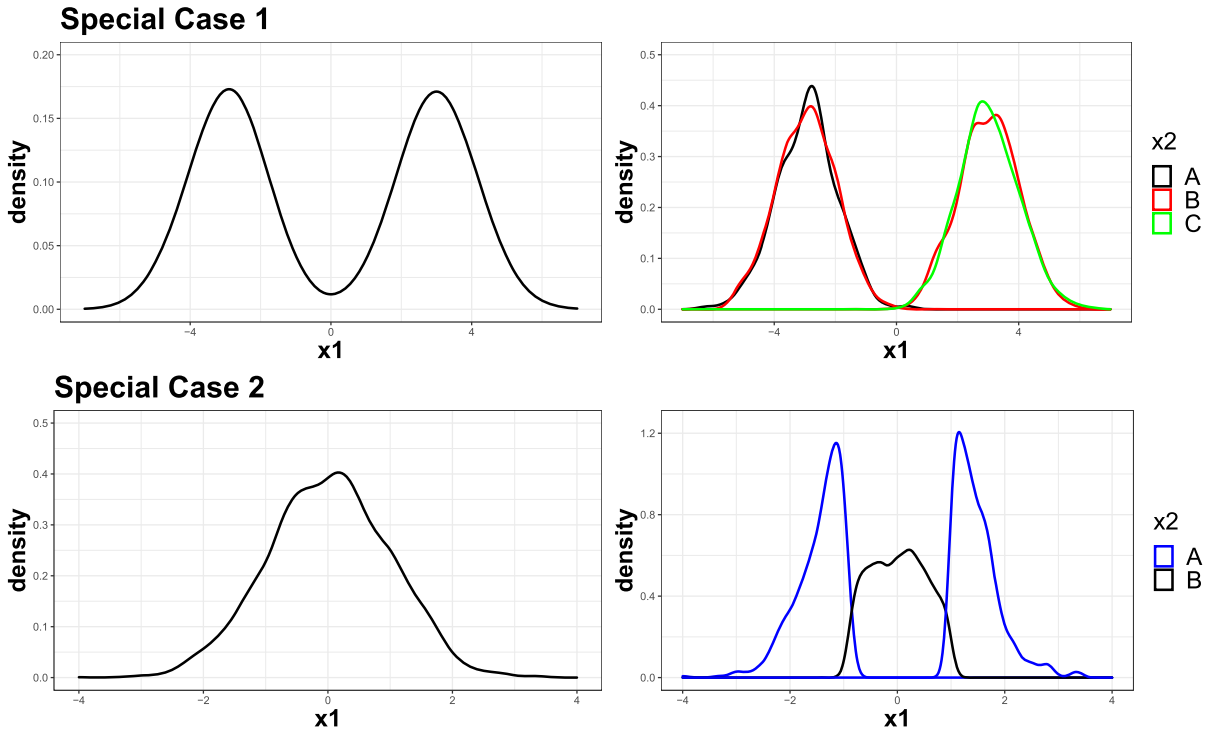

Figure 4: Simulation settings of two special cases.

In summary, the proposed HyDaP algorithm identifies fewer clusters in case 1 initially, but after trying different numbers of clusters, the HyDaP is able to select the optimal one and achieve satisfactory clustering performance. PAM with proposed HyDaP dissimilarity, K-prototypes, and PAM with FAMD distance can select the correct number of clusters and obtain good performance as well, while PAM with Gower distance and FMM cannot. FMM has poor clustering performance even when given the correct number of clusters.

4.6.2 Special Case 2

Consider two variables where x_1 is continuous and x_2 is categorical with two levels A and B. If we use variable x_1 only, we will detect one overall cluster. When x_2 is added, we can detect 3 clusters as shown in the bottom panel of Figure 4. Unfortunately, HyDaP will drop x_1 in step one so that eventually we can only obtain 2 clusters using x_2. In simulations, we set the true cluster sizes as 50, 200, and 50 for lower values of x_1 and x_2 =A, middle values of x_1 and x_2=B, and higher values of x_1 and x_2=A (corresponding to the bottom right panel of Figure 4), respectively. We also generated 500 datasets as before.

Similar to what has been done in special case 1, we identified the number of clusters using several other methods. The results are that PAM with proposed HyDaP dissimilarity, K-prototypes, and PAM with FAMD distance chose 3 as the optimal number of clusters. However, PAM with Gower distance was unable to select an optimal number, and FMM selected 4, 5, or 6 with about equal occurrences in simulation runs.

We also assessed the clustering performance of all methods assuming that 3 is the optimal number of clusters. All methods performed perfectly except FMM (ARI: 0.56 [0.44, 0.89]).

In summary, the proposed HyDaP algorithm identifies fewer clusters in this case, and it is unable to select the optimal number of clusters because of the criteria in step one. PAM with proposed HyDaP dissimilarity, K-prototypes, and PAM with FAMD distance can select the correct number of clusters and obtain good performance. However, PAM with Gower distance and FMM are unable to select the optimal number of clusters. Similar to special case 1, FMM has poor clustering performance even when given the correct number of clusters.

5 Real Data Application

We used the EHR data collected from the SENECA project to demonstrate the use of proposed HyDaP algorithm for identifying phenotypes in patients with sepsis. The SENECA data contains 20 189 sepsis encounters collected from 12 healthcare systems from year 2010 to 2012. We aimed to identify several heterogeneous subgroups (phenotypes) among sepsis patients using information collected at their emergency room presence and select the most important variables that drive clustering results. After obtaining phenotypes, we planned to check whether they are associated with different clinical endpoints. The list of the thirty variables that were used for identifying sepsis phenotypes is shown in Supplementary Materials. 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 were preferred.

Data structure identification: Natural clusters are rarely observed in data collected from clinical settings. This is also true for the SENECA data. Its reachability plot (provided in online Supplementary Materials) shows no natural cluster. After performing the consensus K-means for all continuous variables, we obtained the results in Figure 5 suggesting that the optimal number of clusters is four, and found that the SENECA data belongs to *partitioned cluster structure*.

Variable selection step: For *partitioned cluster structure*, we decided to keep all the continuous variables. To determine if categorical variables gender and race were to be used for clustering, we checked the Cramer's V between each categorical variable and the cluster membership from the consensus clustering and found that Cramer's V for both were 0.05, a very small value. Therefore, we dropped these categorical variables before proceeding to the final clustering step.

Clustering step: We took the consensus K-means results as our final clustering results. In terms of variable importance, under *partitioned cluster structure*, all continuous variables together contributed to the obtained partitions, yet no single variable showed dominant impact.

We obtained four clusters with relatively balanced sample sizes: 6 625, 5 512, 5 385, and 2 667. After examining the characteristics of these clusters, our clinician colleagues found 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 those who had more acidosis, liver, and cardiovascular dysfunctions. These findings indicate that the 4 clusters identified by the HyDaP algorithm are clinically heterogeneous. We then examined the distribution of some important clinical endpoints across 4 clusters; Results are shown in the top left plot of Figure 6. We found that Cluster 1 has the lowest proportions in adverse events defined by those clinical endpoints while Cluster 2 has the second lowest ones. Cluster 4 has the highest proportions. These clinical endpoints were not included in the clustering algorithm, so the distinct distributions across clusters can serve as a "validation" that the clusters we identified are clinically meaningful, as they show different clinical outcomes. The information obtained will be useful in making prognosis for sepsis patients of different phenotypes.

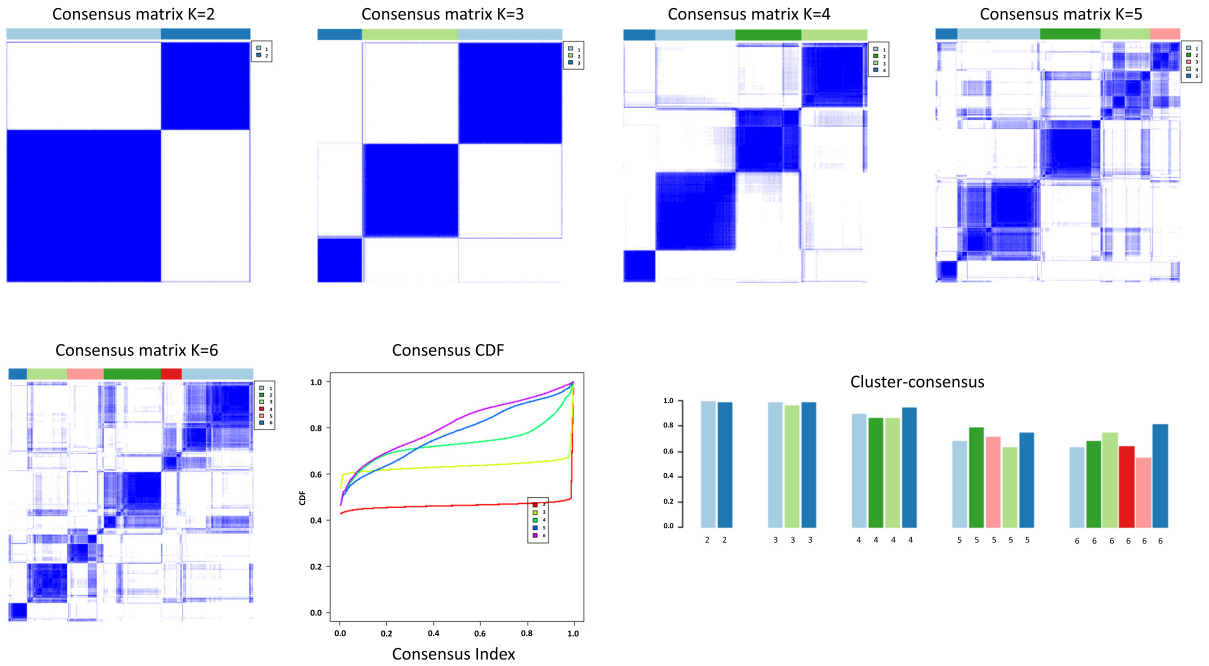

Figure 5: Consensus K-means optimal number of clusters selection.

We examined some other commonly used algorithms for clustering: PAM with Gower distance, K-prototypes, and FMM, assuming that four clusters were to be generated from the same SENECA dataset by each algorithm. The resulting clusters were assessed by distinct clinical characteristics and clinical endpoints. 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 in our proposed HyDaP algorithm gender was not relevant. For the K-prototypes, we found that the 4 clusters obtained were not that distinct from each other in terms of the distributions of the clinical endpoints. The 4 clusters obtained from the FMM appeared to be distinct from each other and were similar to that obtained from using the HyDaP algorithm. However, Cluster 1 has larger proportion of patients admitted to ICU or used mechanical ventilation and vasopressor as compared with Cluster 2 but has lower mortality rate, which is difficult to explain. Moreover, FMM does not provide information about variable importance.

We compared the above-mentioned three methods for application-specific optimal number of clusters. We found that the optimal number of clusters was two for PAM with Gower distance, two for K-prototypes, and three for FMM. We once again observed that the clustering results from PAM with Gower distance were dominated by gender and the two clusters had quite similar distributions of clinical endpoints. Similarly, the two clusters identified by K-prototypes were not distinct in terms of clinical endpoints. The FMM identified three clusters with quite different distributions of clinical endpoints whereas the HyDaP algorithm was able to identify one more cluster with distinct clinical features. The results are provided in the online supplement.

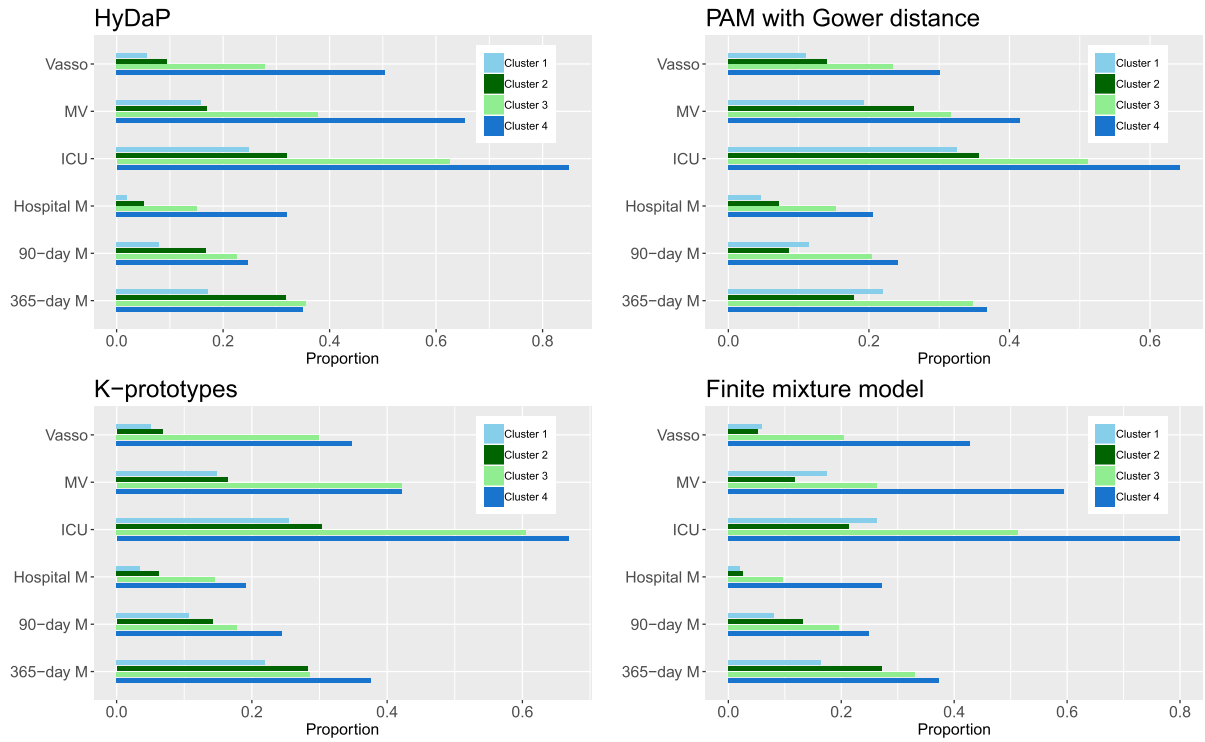

Figure 6: Clinical endpoints across 4 clusters identified by different methods.

6 Discussion

We proposed a novel clustering algorithm (HyDaP) to identify clusters and variable importance in data of mixed variable types. By applying the method to the SENECA data for sepsis patients in the emergency department, we identified four clinically meaningful disease phenotypes that are highly associated with a number of clinical endpoints. All selected continuous demographic and clinical covariates contributed significantly to the determination of the cluster membership with no single feature dominated the process. These covariates also provided clinicians directions of further research in treatments of sepsis. Note that other existing methods mentioned in the paper identified none or fewer clusters. To identify disease phenotypes using the HyDaP algorithm, we first analyzed and determined the data structure which was important in understanding the data and interpreting the clustering results. We then found the cluster membership via our proposed dissimilarity measure which can balance the contribution between continuous and categorical variables. Through simulation studies, we showed that our proposed HyDaP algorithm is robust to different data structures, and can outperform or be on a par with the commonly used methods. If multiple variables that are clinically similar or related exist, we suggest that only one is kept for clustering in order to avoid within-cluster correlations.

Our HyDaP algorithm has a few limitations. First, it inherits the limitations of the sparse K-means algorithm; i.e., for data under the *natural cluster structure*, the sparse K-means procedure cannot correctly identify variables of high contributions if there is a continuous variable containing many outliers or excessive zeros (a.k.a. zero-inflated). The limitation does not affect the results of this study because SENECA data contains no such variable. Second, since the HyDaP algorithm only uses continuous variables in identifying data structure and selecting number

of clusters, it may detect fewer clusters if those clusters are divisible by categorical variables. We have discussed this in detail in Section 4.6. This limitation is not a concern for SENECA data because the two categorical variables in SENECA have low Cramer's V values.

Clustering has emerged as an essential and popular technique for discovering patterns in data. In dealing with the complexity of clinical data, we proposed the HyDaP algorithm to address some of the issues found in the commonly used clustering algorithms, and successfully applied it to identify sepsis phenotypes with distinct demographics, biomarkers, or clinical conditions. The approach will help clinicians gain insight into different sepsis types and treatments thereof and fine-tune precision medicine to reduce the mortality rate associated with sepsis.

Supplementary Material

The R codes and a brief tutorial of implementing the HyDaP are available at GitHub: https://github.com/gmailw1264648156/HyDaP.

References

Angus DC, Van der Poll T (2013). Severe sepsis and septic shock. *New England Journal of Medicine*, 369: 840–851.

Ankerst M, Breunig MM, Kriegel HP, Sander J (1999). OPTICS: Ordering points to identify the clustering structure. In: *ACM Sigmod Record*, volume 28, 49–60. ACM.

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

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

Han J, Pei J, Kamber M (2011). *Data Mining: Concepts and Techniques*. Elsevier.

Haripriya H, Amrutha S, Veena R, 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*, 10–16. ACM.

Hennig C, Liao TF (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.

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.

Hubert L, Arabie P (1985). Comparing partitions. *Journal of Classification*, 2(1): 193–218.

Jensen PB, Jensen LJ, Brunak S (2012). Mining electronic health records: Towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6): 395–405.

Kaufman L, Rousseeuw PJ (2009). *Finding Groups in Data: An Introduction to Cluster Analysis*, volume 344. John Wiley & Sons.

Liu V, Escobar GJ, Greene JD, Soule J, Whippy A, Angus DC, et al. (2014). Hospital deaths in patients with sepsis from 2 independent cohorts. *Journal of the American Medical Association*, 312(1): 90–92.

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, 281–297. Oakland, CA, USA.

McCutcheon AL (1987). *Latent Class Analysis*. 64. Sage.

Monti S, Tamayo P, Mesirov J, 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.

Pagès J (2014). *Multiple Factor Analysis by Example Using R*. CRC Press.

Rand WM (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336): 846–850.

Reddy MJ, 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.

Scicluna BP, Van Vught LA, Zwinderman AH, Wiewel MA, Davenport EE, Burnham KL, et al. (2017). Classification of patients with sepsis according to blood genomic endotype: A prospective cohort study. *The Lancet Respiratory Medicine*, 5(10): 816–826.

Seymour CW, Kennedy JN, Wang S, Chang CCH, Elliott CF, Xu Z, et al. (2019). Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. *Journal of the American Medical Association*, 321(20): 2003–2017.

Seymour CW, Liu VX, Iwashyna TJ, Brunkhorst FM, Rea TD, Scherag A, et al. (2016). Assessment of clinical criteria for sepsis: For the third international consensus definitions for sepsis and septic shock (sepsis-3). *Journal of the American Medical Association*, 315(8): 762–774.

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

Ward Jr JH (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301): 236–244.

Wilkerson MD, Hayes DN (2010). ConsensusClusterPlus: A class discovery tool with confidence assessments and item tracking. *Bioinformatics*, 26(12): 1572–1573.

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

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