Clustering of mixed data and continuous data

: Comparison between k-prototypes, average agglomerative clustering,

k-means, partitioning-around-medoids, mixture of multivariate models, and density-based clustering

Jung-a Kim

Ray Chen

Department of Mathematics and Statistics,

San José State University, California

Abstract

There have been many clustering approaches towards mixed data with distance-based clustering and density-based clustering techniques. K-prototypes is one of the classic clustering methods for the mixed data.With the extended version of weight parameters, it produces more robust results. DBSCAN can identify non-spherical clusters in clearly separated settings. Mixture model-based clustering outperforms partitioning methods by identifying overlapped clusters in the simulated datasets and the real data examples used in this paper.

1. Introduction

In this paper, several clustering methods were tested under different situations considering the separation between homogeneous groups, types of variables, correlation among variables, and within variance. Depending on the type of data, different method are compared to each other.

There have been many researches to develop a clustering method for mixed data, and one of the classic ones is k-prototype presented by Huang in 1997 [2] which is the combination of k-means and k-modes. Since its appearance, there have also been extensions of k-prototypes to enhance robustness such as fuzzy k-prototype clustering [9]. K-prototypes has the same iteration steps as k-means such that it is sensitive to initial random prototypes and assumes the spherical distribution of clusters.

A density-based approaches have been developed to handle the nonspherical continuous data. The traditional method is density-based spatial clustering of applications with noise(DBSCAN) proposed by Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu in 1996 [3]. This method works similarly to aggolmerative clustering except it has parameters which are the minimum number of members in each cluster and the maximum distance(noted as ε) between members. This has great advantage over paritioning methods as it does not require the number of clusters to be created and can classify strength of each point's membership. The downside is that DBSCAN is very sensitive to the parameter setting and works well when data has clusters with similar density. To overcome these limitations, recent studies have modified DBSCAN for efficient parameter specification and they have been extended for mixed data as well [14].

Both paritioning and density-based methods depend heavily on the distances between data points so overlapped clusters are often considered as one. On the other hand, a mixture model uses likelihood to determine the number of clusters and distribution of each cluster so overlapped case can be considered [4-7]. It is a robust model as it involves iteration steps and considers various shapes of the distribution. But it cannot be used for mixed data. In this paper, we used skew-t mixture model [6].

In high-dimensional real dataset, we will use dimensional reduction techniques, principal component analysis(PCA)[11] and factor analysis of mixed data(FAMD)[12] to interpret the cluster solutions, compare the homogeneity of each cluster in the graph, and class agreement with each other and labels from the dataset.

2. Methods

K-prototypes

K-prototypes is created to complement the confined usage of k-means to numeric data. The dissimilarity measure of k-prototypes is the combination of Euclidean distance and probability of modes across all the variables. The cost function for k-prototypes used in this paper is

$$E_l = \sum_{i=1}^{n} y_{il} \sum_{j=1}^{m_r} \lambda_j \left(x_{ij}^r - q_{lj}^r\right)^2 + \sum_{j=m_r+1}^{m} \lambda_j n_l (1 - p(q_{lj}^c \in C_j | l)) = E_l^r + E_l^c$$

where m_r is the number of numeric variables and m is the total number of variables

r represents the numerical data and c represents the categorical data. n is the number of observations. y_{il} is the membership of X_i to the cluster l. Thus, $y_{il} = 1$, if X_i belongs to the cluster l. $y_{il} = 0$, otherwise. m_r is the number of numerical variables. $m - m_r$ is the number of categorical variables. n_l is the number of members in cluster l. λ_j is the weight parameter which determines the degree of the variable's influence on the cost. Originally in the paper proposed by Huang in 1997, λ_j is only applied to categorical variables as one unison real-value λ. The above cost function is an extended version that accomodates each variable's variability which was proposed

by Szepannek in 2018 [13]. In this paper we will call $[\lambda_1, \dots, \lambda_m]$a lambda vector. A lamda vector has inverse values of each variable's variance. In order to minimize the cost function, E_l, E_l^r and E_l^c are minimized independently. E_l^r is minimized if $q_{lj}^r = \frac{1}{n_l}\sum_{i=1}^{n} y_{il}x_{ij}$ for j = 1, … , m. Thus, q_{lj}^r is the centroid of cluster l. E_l^c is minimized if and only if $p\left(q_{lj}^c \in C_j \middle| l\right) \geq p(c_j \in C_j|l)$ for $q_{lj}^c \neq c_j$ for all categorical variables where C_j is the set of all unique values of categorical variable j and c_j is an arbitrary value of the categorical variable j. Thus, the minimizer q_{lj}^c has to be the mode of the categorical variable j.

The iteration steps are equivalent to the k-means method. First, k random initial prototypes are selected from the data. Then, each object is assigned to a cluster where its cost is the minimum. Next, the prototypes are modified. The similarity between the objects and its prototypes are re-measured to check if the object is in the right cluster. If not, the objects are re-allocated and prototypes are updated again. This iteration stops when no object changes its membership.

DBSCAN

DBSCAN detects dense and sparse regions of data. The degree of density is solely dependent on the parameter ε. This ε is also determined by the minimum sample size which is another parameter. The common minimum sample size is $2 * \text{number of dimensions}$. ε should be small enough exclude noise from a sample, but large enough to form as few clusters as possible. Thus, the minimum sample size must be fixed to determine the optimal ε which is the maximum boundary. This boundary point is the elbow in a k-nearest-neighbor distance plot which is the plot of each data point's average distance to the minimum sample size -1 neighbors in ascending order. `kNNdist` function from the R-package `dbscan` [20] can be used for the plot. The optimal ε is where curvature of this graph is maximized. With minimum sample size and ε, DBSCAN can classify each point's strength of membership. If a point has k(minimum sample size) or more neighbors, the point is classified as a core point. A border point is one that has less than k neighbors, but has a core point in the neighborhood. If a point is neither a core point nor a border point, it is a noise point. The following graph shows the k-nearest-neighbor distance plot of Travel Review Ratings data used as a real data example.

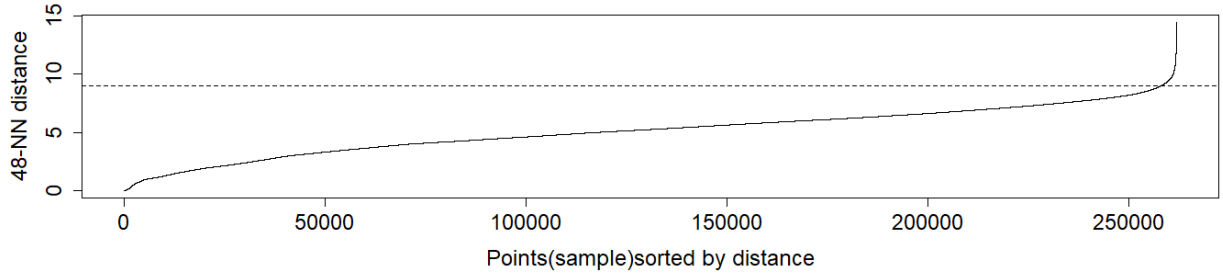

Figure 1. 48-nearest-neighbor distance plot of Travel Review Ratings data

Mixture of skew-t distributions

Skew-t mixture model is a special case of generalized hyperbolic distribution model which considers skewness in a t-distribution. It uses EM algorithm to find the maximum likelihood estimates of its parameters, mean, covariance, proportion, degrees of freedom, and skewness which are symbolized as $\mu, \Sigma, \upsilon, \pi$ and δ. The mixture model uses complete likelihood function where each data point's membership variable z_{ig} and gamma variable u_{ig} are involved, so EM algorithm starts with initial random parameters to find the expected z_{ig} and u_{ig}. Using the expected values, parameters are estimated and the data points are re-allocated. These two steps are repeated until the membership converges or the maximum number of iterations set by the user is reached.

Other methods

In addition to the methods mentioned above, Gaussian mixture model [4], mixture of t distributions [5], mixture of skew-Gaussian distribution model [7], partitioning-around-medoids(PAM) [8], k-means [9], and agglomerative clustering [10] are used to compare.

3. Simulation setup

12 different scenarios were set up based on separation between each sample distribution, the number of different distributions, their types, correlations, and sizes. For each scenario, 10 different datasets were randomly created using `rmnorm` function(from R-package `lmf` [21]), `rmvt`(from `mvtnorm` [22]), `rdmsn`(from `EMMIXskew` [23]), and `rdmst`(from `EMMIXskew`). The dimension of each distribution was fixed as 5.

Simulation 1 Clearly separated with 2 clusters. Distributions: (Gaussian, Gaussian) Correlations: (none, none) Dimensions: (5, 5) Sample sizes: (336, 615)	Simulation 2 Clearly separated with 4 clusters. Distributions: (Gaussian, Gaussian, Gaussian, t) Correlations: (none, none, none, none, none) Dimensions: (5, 5, 5, 5) Sample sizes: (336, 615, 400, 500)	Simulation 3 Clearly separated with 2 clusters. Distributions: (Gaussian, Gaussian) Correlations: (none, moderately positive) Dimensions: (5, 5) Sample sizes: (336, 615)
Simulation 4 Clearly separated with 4 clusters. Distributions: (Gaussian, Gaussian, Gaussian, t) Correlations: (none, moderately positive, moderately positive, moderately positive) Dimensions: (5, 5, 5, 5) Sample sizes: (336, 615, 400, 500)	Simulation 5 Clearly Separated with 2 clusters. Distributions: (Gaussian, skew-Gaussian) Correlations: (none, moderately positive) Dimensions: (5, 5) Sample sizes: (336, 615)	Simulation 6 Clearly Separated with 4 clusters. Distributions: (Gaussian, skew-Gaussian, Gaussian, t) Correlations: (none, moderately positive, moderately positive, moderately positive) Dimensions: (5, 5, 5, 5) Sample sizes: (336, 615, 400, 500)
Simulation 7 Overlapped 2 clusters. Distributions: (Gaussian, Gaussian) Correlations: (none, none) Dimensions: (5, 5) Sample sizes: (336, 615)	Simulation 8 Overlapped 4 clusters. Distributions: (Gaussian, Gaussian, Gaussian, skew-t) Correlations: (none, none, none, none) Dimensions: (5, 5, 5, 5) Sample sizes: (336, 615, 400, 500)	Simulation 9 Overlapped 2 clusters. Distributions: (Gaussian, Gaussian) Correlations: (none, moderately positive) Dimensions: (5, 5) Sample sizes: (336, 615)

Simulation 10 Overlapped with 4 clusters. Distributions: (Gaussian, Gaussian, Gaussian, skew-t) Correlations: (none, moderately positive, moderately positive, moderately positive) Dimensions: (5, 5, 5, 5) Sample sizes: (336, 615, 400, 500)	Simulation 11 Overlapped with 2 clusters Distributions: (Gaussian, skew-Gaussian) Correlations: (none, moderately positive) Dimensions: (5, 5) Sample sizes: (336, 615)	Simulation 12 Overlapped with 4 clusters Distributions: (Gaussian, skew-Gaussian, Gaussian, skew-t) Correlations: (none, moderately positive, moderately positive, moderately positive) Dimensions: (5, 5, 5, 5) Sample sizes: (336, 615, 400, 500)

4. Simulation results

K-means, PAM, DBSCAN, and skew-t mixture models were tested in each simulated dataset. Each k-means solution was generated starting with 20 different sets of initial points and 100 maximum number of iterations using `kmeans` function from `stats` package [24]. PAM solution was generated based on Euclidean distance using `pam` function from `cluster` package [25]. The criterion for selecting the number of clusters was average silhouette width for both k-means and PAM.

The minimum sample size for DBSCAN was fixed as 10(=2*number of dimensions in each dataset) for all 120 datasets. For estimating optimal ε in each dataset, we used average slopes to find the ε that maximized the curvature since the distance function was discrete. We used 50 average slopes which were considered as enough number of slopes to estimate the optimal ε. DBSCAN results were generated using `dbscan` function from `dbscan` package.

Skew-t mixture solution was generated using `EmSkew` function from `EMMIXskew` package and BIC was the criterion for selecting the number of clusters.

For each simulated dataset, ARI between each solution and true labels were recorded. Table 1 shows the average ARI per method in each scenario.

	k-means	PAM	DBSCAN	Skew-t
Simulation 1	1.0000	1.0000	0.9866	1.0000
Simulation 2	0.9880	0.9968	0.9580	0.9998
Simulation 3	1.0000	1.0000	0.9877	1.0000
Simulation 4	0.9911	0.9911	0.9688	0.9997
Simulation 5	0.6379	0.6495	0.9320	1.0000
Simulation 6	0.8587	0.8458	0.9582	0.9999
Simulation 7	0.0136	0.0122	0.0006	0.0000
Simulation 8	0.0297	0.1232	0.0113	0.3327
Simulation 9	0.0149	0.0120	0.0086	0.6157
Simulation 10	0.0289	0.1231	0.0089	0.8600
Simulation 11	0.0801	0.1712	-0.0204	0.7971
Simulation 12	0.0860	0.1877	0.0012	0.9170

Table 1. Average ARI of k-means, PAM, DBSCAN, and Skew-t

Overall, skew-t mixture model clearly showed the best results maintaining much higher ARI throughout all the scenarios than the other methods whose ARI dropped dramatically in the datasets with overlapped clusters. In cases of overlapped clusters, ARI with four true clusters was higher than ARI with two true clusters in general.

K-means and PAM showed almost perfect ARI when non-skewed clusters were clearly separated regardless of whether they were distributed with t or Gaussian(simulation 1-4) and whether the moderate correlation existed in some of the clusters or not. But both methods were weak in detecting skewed clusters. When one of two Gaussian distributions(simulation 3) became skewed(simulation 5), ARI of K-means and PAM dropped from 1 to about 0.65. This indicated that both methods were more appropriate for non-skew clusters. But in simulation 6 where only one out of four distributions was skewed with Gaussian distribution, both methods' ARI(0.85) were not much affected by its skewness. This indicated that if most of the distributions are non-skewed, k-means and PAM are robust. It was also clear that both methods performed poorly when clusters were overlapped. The average ARI in this situation for k-means was about 0.05 and that for PAM was about 0.1 whereas the average ARI in clearly separated clusters for both methods were about 0.91. PAM worked slightly better than k-means in overlapped clusters when there were more than two clusters.

DBSCAN had such dramatic performance gap between clearly separated clusters and overlapped clusters that it was clear DBSCAN cannot detect overlapped clusters. For all the situations where clusters were overlapped regardless of the number of clusters and their distributions, DBSCAN worked poorly as if it chose random clusters. But it was robust in skewed clusters(simulation1-4) since it maintained as high ARI's as those in non-skewed clusters(simulation 5-6).

Skew-t mixture models could detect the true clusters in all the situations where clusters were clearly separated. In overlapped clusters, it showed varing degrees of class agreement according to the situations. In simulation 7 where two spherical Gaussian distributions were overlapped, skew-t mixture model considered it as one cluster. Thus, the ARI resulted in zero. On the other hand, when one of the two Gaussian had moderately positive correlation between variables(simulation 9), it could detect different clusters to some high degree observed by ARI jumping from zero to 0.62. Skew-t mixture model worked much better when overlapped clusters had different shapes. When analogously shaped clusters overlapped, there was less misclassification rate with higher number of clusters than less number comparing ARI's from simulation 7 and 8 results. In scenarios with overlapped clusters, ARI increased noticeably as each cluster became variant with distribution, skewness, and correlations.

5. Real data example

Four different real datasets were collected from different sources. 'Species' from Tetragonula Bee Species and 'General Type' from Schools dataset were excluded from the analysis procedure as they were used to identify true clusters in the data.

- Tetragonula Bee Species [15]
  - Genetic data for 236 Tetragonula bees from Australia and Southeast Asia
  - 14 categorical variables: L1 – L13 are strings of six digits which encode a pair of alleles with no numeric information. Species represent the species out of 9 categories labeled from 1 to 9.
  - 2 numerical variables: C1 and C2 are coordinates of locations of individual bees. C1 is is latitude (negative values are South). C2 is longitude (negative values are West).

- Schools [16]
  - Features of 445 public and private schools for infant, Pre-K, and K-14 students in San Francisco, California
  - 5 variables were selected out of 16 variables.
  - 3 categorical variables:
    - CCSF Entity: City College of San Francisco entities
      - Private
      - SFCCD = San Francisco Community College District
      - SFUSD = San Francisco Unified School District
    - General Type: Broad category of schools
      - CC = Community College
      - CDC = Child Development Center
      - IND = Independent / Private
      - PS = Public School
    - Supervisor District: City and County Supervisor District number
      - 1-9 (9 levels)
  - 2 numerical variables:
    - Lower Age: Lower bound of generic age of the education program
    - Upper Age: Upper bound of generic age of the education program

- Aircraft Noise Complaint [17]
  - Counts of aircraft noise complaints by community and by month which were collected via Complaint hotline, Online Complaint Form, emails, letters and telephone calls to the Aircraft Noise Abatement Office in San Francisco, California
  - 4305 counts in each combination of community, year, and month
  - 2 categorical variables:
    - Month: The month of the aircraft noise complaint disturbance.
    - Community: The name of the community where the aircraft noise disturbance occurred.
  - 3 numerical variables:
    - Year: The year of the aircraft noise complaint disturbance.
    - Total Complaints: The number of monthly complaints associated by community.
    - Total Number of Callers: The number of Complainants associated by community.

- Travel Review Ratings [18]
  - 5456 Google ratings on attractions from 24 categories across Europe. The rating ranges from 1 to 5.
  - 24 numerical variables: Average user rating per category

Travel Review Ratings data is a continuous dataset whereas the others are mixed data. Thus, we implemented different methods for these two types of datasets.

For mixed data, we used FAMD for dimensional reduction for the purpose of visualizing the clusters and discuss the relations between the components and clustering results using `FAMD` function from `FactoMineR` R-package [26] and functions from `factoextra` package [27]. The methods used for mixed data were k-prototypes, agglomerative clustering, and PAM. The gower dissimilarity was used for agglomerative clustering and selecting the number of clusters for PAM and k-prototypes.

For continuous data, we used PCA for the same purpose as FAMD using `princomp` and `biplot` functions from `stats` package. The methods for continuous data were k-means, agglomerative clustering, PAM, DBSCAN, and four types of mixture models(Skew-t, Skew-normal, t, and Gaussian).

Tetragonula Bee Species data

Using 'Species' as the true labels from the original dataset, the homogeneous groups were clearly separated from each other except the two clusters overlapping in the left(Figure 2).

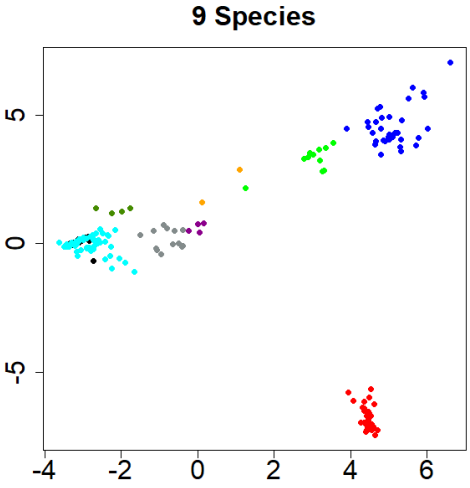

Figure 2. 2-d plot of Tetragonula Bee Species data

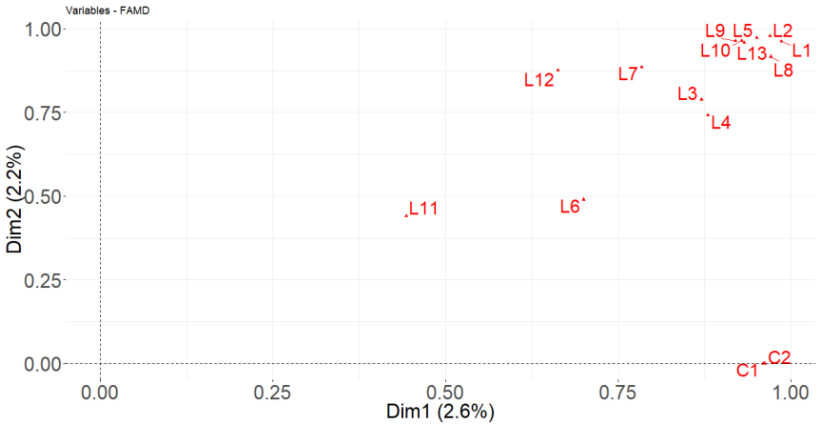

Figure 3. Correlation plot of Tetragonula Bee Species data

But figure 2 did not represent the true positions of data since the two components had very low variability as 4.8% in total as we can see in the following correlation plot(Figure 3) displayed by `fviz_pca_var` function from `factoextra` package. The top five components could only explain around 2% each. C1, C2, L1, L2, L5, L8, L9, L10, L13 were explained by Dim1 by more than 90%. L1, L2, L5, L8, L9, L10, L13, L7, and L12 were mostly explained by Dim2. The numerical variables were well explained by Dim1, but not by Dim2.

Conducting k-prototypes, lambda vector was determined by inverse of each variable's variances. As the criterion for selecting the number of clusters, average silhouette width using gower dissimilarity was computed. The solution with 3 clusters had the highest average silhouette width compared to the others with the number of clusters varying from 2 to 10.

For agglomerative clustering, we used gower dissimliarity and compared the four different methods of clustering: Average, Ward, Single Linkage, and Complete Linkage. Average agglomerative clustering had the highest cophenetic correlation with cophenetic matrix. After looking at the dendrogram, we chose 4 clusters.

For PAM, gower dissimilarity measure was used and the number of clusters was selected based on the average silhouette width.

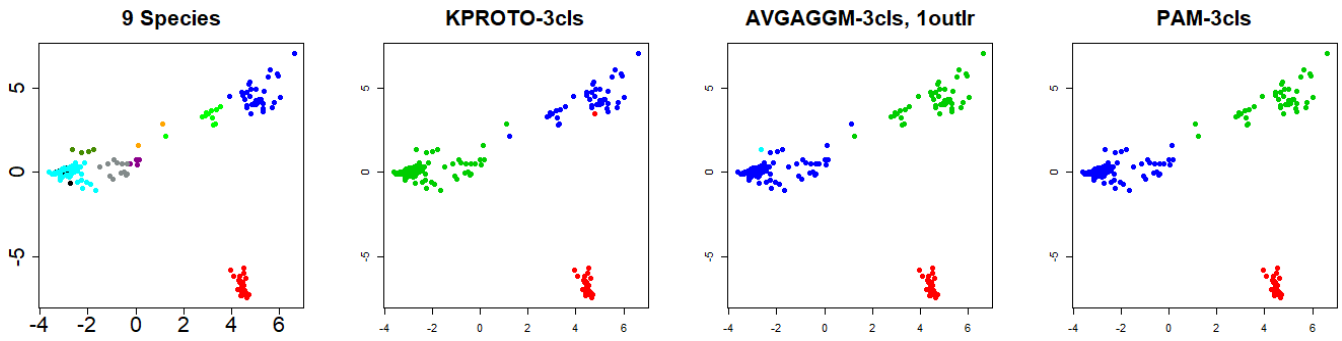

Figure 4. 2-d plots of 9 Species, k-prototypes, average agglomerative clustering, and PAM in Tetragonula data

None of the three methods could identify different clusters in the overlapping region in the left, which lowered the class agreement by large amount. All of these methods resulted in three clusters and each of their classification agreements with species was about the same as 45%. For k-prototypes, most lambda values of categorical variables for k-prototypes were around 1 and the average lambda values of numerical variables was almost close to zero which meant vertical distance was mainly used for partitioning. The red-colored dot placed in the region of the blue cluster was similar to the red cluster in terms of L4 and L11 whereas it was close to the blue cluster in terms of L2, L5, and L10. Although the lambda for L11 was more than twice higher(2.63) than the other variables, the vertical axis was not highly correlated with L11(r=0.5) as shown in figure 3. Thus, it was positioned closer to the blue cluster. If we placed the same weight for all the variables, this dot would have belonged to the blue cluster which was the true cluster it was supposed to belong to. Thus, placing different variable weights based only on its variance may not be the remedy for some cases.

Schools data

Since this data was mixed-type as well, the same procedure as did in Tetragonula Bee Species dataset was conducted. The correlation plot below(Figure 5) showed the total variability of the two components was 24.8% and Dim1 was highly correlated with lower and upper age of schools and CCSF entity and general type of schools were moderately correlated with both components. Supervisor district had moderate correlation with Dim2, but no correlation with Dim1.

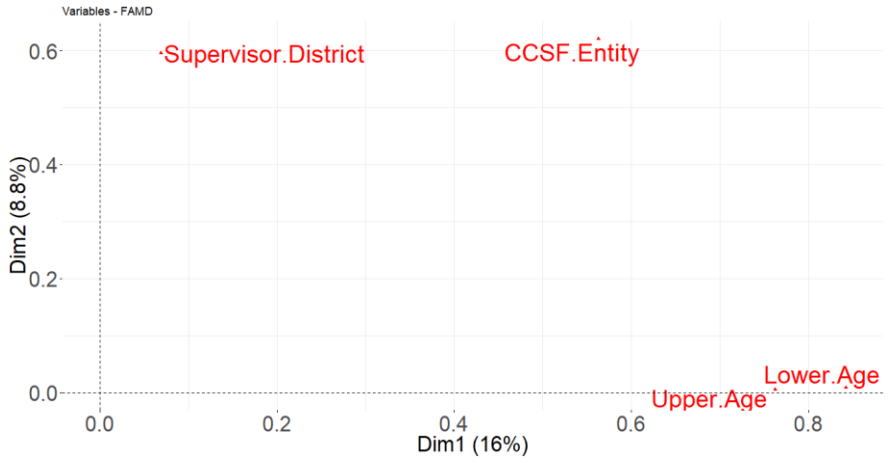

Figure 5. Correlation plot of Schools data

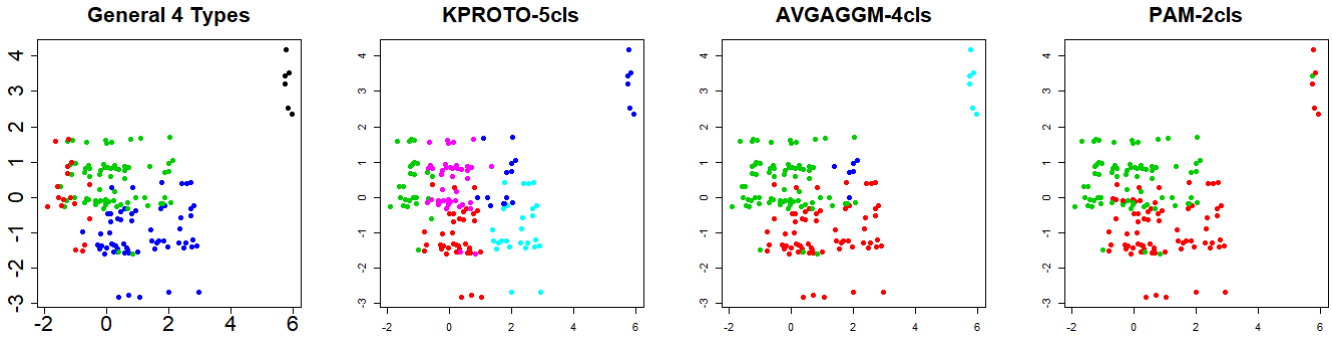

Figure 6. 2-d plots of General Types, k-prototypes, average agglomerative clustering, and PAM in Schools data

K-prototypes showed high class agreement as 71% with general types of school. ARI for PAM was 60% and ARI for average agglomerative clustering was 53%. For k-prototypes the lambda vector was *[CCSF entity=2.24, supervisor district=1.1, lower age=0.07, upper age=0.04]*. Thus, the misclassified blue dots that are spread vertical wise in the lower left group could be the result from not having enough influence of 'age' to be considered as a different cluster from the upper right group of blue dots since those two groups are different horizontal-wise. The green dot in the red cluster in the plot for PAM was an outlier in terms of age since both lower age and upper age were the highest number possible(18 and 19) which accounted for 2% each in either of the two clusters. But it was decided as being closer to the green cluster than the red cluster because its supervisor district was the mode of the green cluster. This showed us an example of misclassification resulting from putting the same weight for all the variables in contrast to the previous case of misclassification in k-prototypes from Tetragonula Bee Species data.

Aircraft Noise Complaint data

The same procedure as the previous analysis was done to produce the results. The following correlation plot shows that the variability of Dim1 and Dim2 was very low as 2.4%. Thus, the 2-d plot could not incorporate the true sparsity of the data. The horizontal axis was moderately correlated with total complaints and community. Month and the total number of callers were not correlated with either of the two dimensions. Year was weakly correlated with both dimensions.

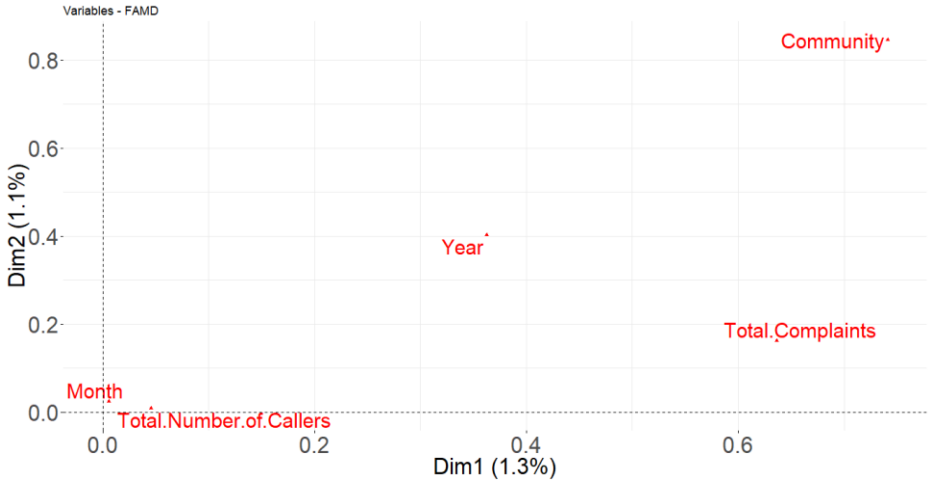

Figure 7. Correlation plot of Aircraft Noise Complaint data

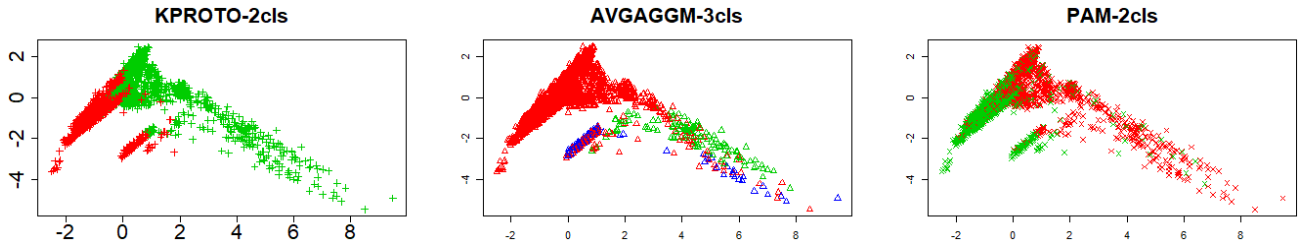

Figure 8. 2-d plots of k-prototypes, average agglomerative clustering, and PAM in Aircraft Noise Complaint data

Since there was no labeled variable that could be representative of homogenous groups, we could only compare the class agreement with each other's results. The class agreement between PAM and k-prototypes was 0.57 whereas the other two combinations had negative ARI. Since agglomerative clustering is more sensitive to outliers than PAM, this may be the cause of having an additional cluster(blue dots) in agglomerative clustering results. The lambda vector in k-prototypes was *[year=0.06, month=1.10, community=1.02, total complaints= 0, Total number of callers = 0]* which indicated that the clusters were determined majorly based on month and community. Thus, the dots colored in blue from the agglomerative clustering plot were not identified as a separate cluster in k-prototypes results although they had relatively higher total complaints and total number of callers (within top 20%, each) than the leftmost red cluster.

Travel Review Ratings data

Since this data has only continuous variables, we used different methods from the previous analysis. The mixture models of four types, average agglomerative clustering, PAM, k-means, and DBSCAN were compared and Euclidean distance was used for all these methods. For reducing the dimensions, PCA was used to display the 2-d plot and to examine the variability of the two components. From the following correlation plot, we could verify that the first two components had variability about 34.4%. Except for category 2 and category 16, all the other variables had moderate correlation with both or either of the components.

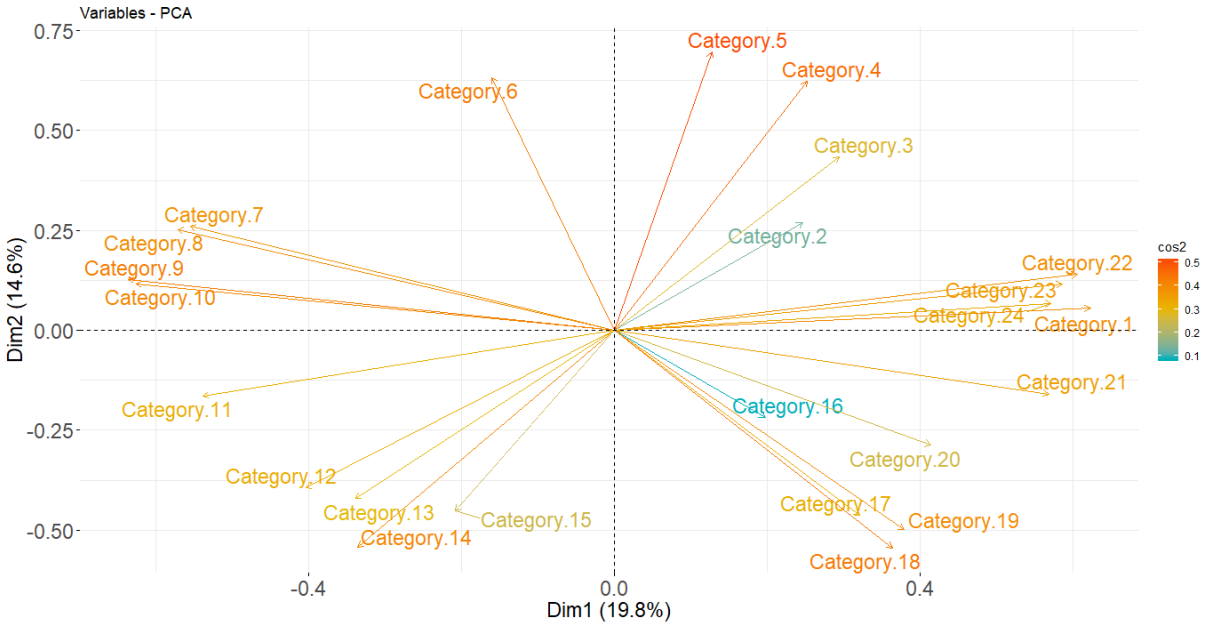

Figure 9. Correlation plot of Travel Review Ratings data

First, we compared the four mixture models: Gaussian, t, skew-Gaussian, and skew-t. For each model, BIC was the criterion to select the final number of clusters. 2-dimensinoal plot was not helpful in comparing the models since distinguishing homogeneous groups in each plot was difficult due to overlaps. Thus, we compared the class agreement to select the least complicated mixture model. For skew-normal mixture model, BIC kept increasing as the number of clusters increased. We assumed that the data had many outliers which forced the skew-normal model to keep detecting new clusters. On the other hand, BIC of skew-t mixture models stopped increasing when the number of clusters reached 8. The class agreement between skew-normal and skew-t was low as 22%. Thus, we chose skew-t over skew-normal model. In the next step, we compared skew-t model and t mixture model to see if the clusters were generally skewed. The class agreement between them was also low as 37.5%, so we chose skew-t over t mixture model.

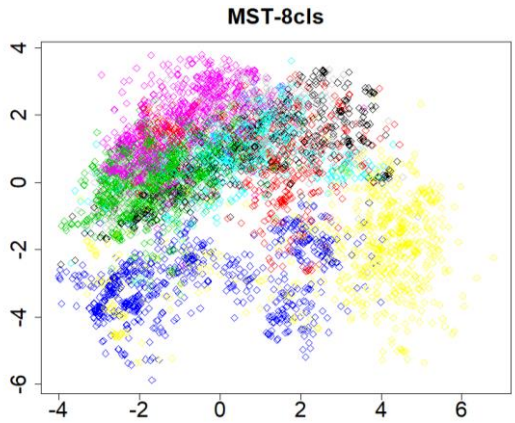

Figure 10. Mixture model of Skew-t in Travel Review Ratings data

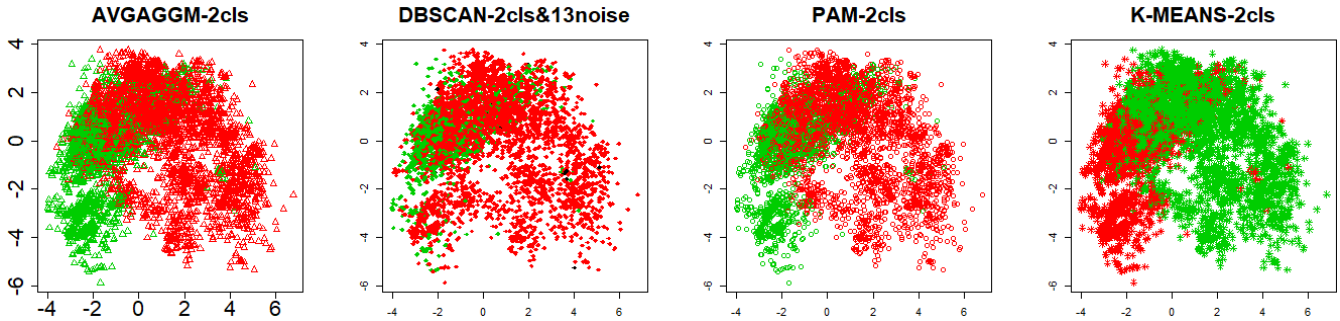

Figure 10. 2-d plots of Average agglomerative clustering, DBSCAN, PAM, and k-means in Travel Review Ratings data

Considering the number of categorical variables(24) and the dataset size(5456 observations), there could have been many shapes of homogeneous groups and we could verify this from figure 9 which showed that the sparsity in the data was variant across most regions. Based on the class agreements of each combination, mixture model of skew-t had close-to-zero agreement with the other methods. Agglomerative clustering, PAM, and k-means had very similar results. DBSCAN had 50% agreement with these three methods. This meant either mixture of skew-t performed very poorly or superior to the other methods. Referring back to the simulation results, the ARI table proved that the mixture model of skew-t was robust in overlapping clusters especially when the homogenous groups had different shapes. In the same environment, k-means, DBSCAN, and PAM were as bad as random clusters. It was evident that the mixture of skew-t had advantage in predicting the true clusters in this dataset which satisfied both conditions. Thus, we could conclude that the mixture model of skew-t was likely to explain the homogeneous groups with much more accuracy than the other methods.

6. Conclusion

In this paper, we tested several methods according to different types of datasets in terms of variable types, sample size, shapes, and distributions. From the simulated datasets, mixture model of skew-t showed consistently superior performance in all the scenarios compared to the other distance-based methods especially when the data were overlapped. Although we should take into account that the simulated datasets were in favor of mixture models since they were generated from random mixture models with fixed parameters, the performance was incomparable. The downside of the mixture model was that when clusters were both overlapped and analogously shaped, it performed as poorly as the other methods. For the continuous real dataset, we could once again verify the robustness of mixture model when dataset was large with many variables and different shapes.

When the clusters were clearly separated, DBSCAN performed well in identifying non-spherical shapes. Although we had not discussed the influence of DBSCAN's parameter ε in this paper, there have been extended algorithms to improve the measure of local density and relative distances [14].

From the mixed real dataset examples, k-prototypes assigned different weights for cost of each variable inversely proportional to its variance, which made the method more robust than agglomerative clustering and PAM which used gower dissimilarity. All these three methods assumed spherical shapes in the data.

Future studies for finding appropriate dissimilarity measures of mixed data for model-based clustering and density-based clustering could be helpful for further applications.

References:

[1] Anderlucci, L. and Hennig, C. (2014). Clustering of categorical data: A comparison of a model-based and a distance-based approach. Comm. Statist. Theory Methods 43 704 – 721.

[2] Huang, Z. “Clustering Large Data Sets with Mixed Numeric and Categorical Values”, In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD’97), Singapore, 22–23 February 1997; pp. 21–34. Available: https://pdfs.semanticscholar.org/d42b/b5ad2d03be6d8fefa63d25d02c0711d19728.pdf

[3] Ester M., Kriegel H., Sander J., Xu X. “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”, *KDD'96 Proceedings of the Second Int. Conf. on Knowledge Discovery and Data Mining*, Portland, Oregon, 1996, pp.226-231. [Online.] Available: https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf

[4] Fraley C., Raftery A. “Model-Based Clustering, Discriminant Analysis, and Density Estimation,” *Journal of the American Statistical Association*, vol. 97, no. 458, pp. 611-631, Jun 2002. [Online]. doi: 10.1198/016214502760047131

[5] Andrews, J. L. & McNicholas, P. D. (2012). Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions. Statistics and Computing, 22(5), 1021-1029.

[6] Theodossiou, P. “Financial Data and the Skewed Generalized T Distribution,” *Management Science*, vol.44, no.12, pp.1650-1661, Dec 1998.

[7] O'Hagan, A. and Leonard, T. (1976). Bayes estimation subject to uncertainty about parameter constraints. Biometrika, 63, 201-202.

[8] Kaufman, L. and Rousseeuw, P.J. (1990) Partitioning around Medoids (Program PAM). In: Kaufman, L. and Rousseeuw, P.J., Eds., Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley & Sons, Inc., Hoboken, 68-125.

[9] MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. 1. University of California Press. pp. 281–297.

[10] Sokal R, Michener C (1958). "A statistical method for evaluating systematic relationships". University of Kansas Science Bulletin. 38: 1409–1438.

[11] Jollife, I.T. 2002. Principal Component Analysis. 2nd ed. New York: Springer-Verlag. https://goo.gl/SB86SR.

[12] Pagès, J. 2004. “Analyse Factorielle de Donnees Mixtes.” Revue Statistique Appliquee 4: 93–111.

[13] Szepannek, G. “clustMixType: User-Friendly Clustering of Mixed-Type Data in R,” *The R Journal*, vol. 10/2, pp.3-4, Dec. 2018. Available: https://journal.r-project.org/archive/2018/RJ-2018-048/RJ-2018-048.pdf

[14] Duan, B., Han, L., Gou Z., Yang Y., Chen S. “Clustering Mixed Data Based on Density Peaks and Stacked Denoising Autoencoders,” *Symmetry*, vol. 11, no. 163, pp.3-4, Feb. 2019. [Online]. doi: 10.3390/sym11020163

[15] Franck, P., E. Cameron, G. Good, J.-Y. Rasplus, and B. P. Oldroyd (2004) Nest architecture and genetic differentiation in a species complex of Australian stingless bees. Mol. Ecol._ 13, 2317-2331.

[16] DataSF, “Schools,” July 11, 2016. [Online]. Available: https://data.sfgov.org/Economy-and-Community/Schools/tpp3-epx2

[17] DataSF, “Aircraft Noise Complaint Data,” Feb. 13, 2019. [Online]. Available: https://data.sfgov.org/Transportation/Aircraft-Noise-Complaint-Data/q3xd-hfi8

[18] Renjith, Shini, A. Sreekumar, and M. Jathavedan. 2018. â€œEvaluation of Partitioning Clustering Algorithms for Processing Social Media Data in Tourism Domainâ€ • . In 2018 IEEE Recent Advances in Intelligent Computational Systems (RAICS), 127â€“31. IEEE.

[19] Everitt, B. S., S. Landau, M. Leese, and D. Stahl (2011, January). Cluster Analysis (5th ed.), Volume 14. John Wiley & Sons, Ltd.

[20] Michael Hahsler and Matthew Piekenbrock (2018). dbscan: Density Based Clustering of Applications with Noise (DBSCAN) and Related Algorithms. R package version 1.1-3. https://CRAN.R-project.org/package=dbscan

[21] Thomas Kvalnes (2013). lmf: Functions for estimation and inference of selection in age-structured populations. R package version 1.2. https://CRAN.R-project.org/package=lmf

[22] Alan Genz, Frank Bretz, Tetsuhisa Miwa, Xuefei Mi, Friedrich Leisch, Fabian Scheipl, Torsten Hothorn (2019). mvtnorm: Multivariate Normal and t Distributions. R package version 1.0-10. URL http://CRAN.R-project.org/package=mvtnorm

[23] Kui Wang, Angus Ng and Geoff McLachlan. (2018). EMMIXskew: The EM Algorithm and Skew Mixture Distribution. R package version 1.0.3. https://CRAN.R-project.org/package=EMMIXskew

[24] R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[25] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.(2018). cluster: Cluster Analysis Basics and Extensions. R package version 2.0.7-1.

[26] Sebastien Le, Julie Josse, Francois Husson (2008). FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software, 25(1), 1-18. 10.18637/jss.v025.i01

[27] Alboukadel Kassambara and Fabian Mundt (2017). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.5. https://CRAN.R-project.org/package=factoextra

[28] Ryan P. Browne, Aisha ElSherbiny and Paul D. McNicholas (2018). mixture: Mixture Models for Clustering and Classification. R package version 1.5.

[29] Ryan P. Browne and Paul D. McNicholas (2014). Estimating Common Principal Components in High Dimensions. Advances in Data Analysis and Classification 8(2), 217-226.

[30] Gilles Celeux and Gerard Govaert (1995). Gaussian Parsimonious Clustering Models. Pattern Recognition 28(5), 781-793.

[31] Andrews JL, Wickins JR, Boers NM, McNicholas PD (2018). “teigen: An R Package for Model-Based Clustering and Classification via the Multivariate t Distribution.” _Journal of Statistical Software_, *83*(7), 1-32. doi: 10.18637/jss.v083.i07 (URL: http://doi.org/10.18637/jss.v083.i07).

[32] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel and Friedrich Leisch (2019). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-0.1. https://CRAN.R-project.org/package=e1071

[33] Alan Genz, Frank Bretz (2009), Computation of Multivariate Normal and t Probabilities. Lecture Notes in Statistics, Vol. 195., Springer-Verlag, Heidelberg. ISBN 978-3-642-01688-2