K-Means Clustering on Multiple Correspondence Analysis Coordinates

Le Phan, Hongzhe Liu and Cristina Tortora

Abstract On April 18, 2017, the International Federation of Classification Societies (IFCS) issued a challenge to its members and the classification community to analyze a data set of 928 low back pain patients. In this paper, we present our contribution in terms of a cluster analysis of this data set. We will discuss our data cleaning process, which we view as a two-pronged approach: inferring values that are missing not at random and imputing values that are missing at random. We will also discuss the challenges in clustering mixed data types and the required data transformation prior to applying a clustering algorithm. We call our proposed data transformation process split-then-join. Finally, we offer our interpretation of the clustering results with respect to validation variables and we present some thoughts on selecting important variables to classify new observations.

Le Phan
San José State University
🖂 lephan1205@gmail.com

Hongzhe Liu
San José State University
🖂 hongzheliu11@gmail.com

Cristina Tortora
San José State University
🖂 cristina.tortora@sjsu.edu

ARCHIVES OF DATA SCIENCE, SERIES B
(ONLINE FIRST)
KIT SCIENTIFIC PUBLISHING
Vol. 1, No. 1, 2019

DOI 10.5445/KSP/1000085952/05
ISSN 2510-0564

1 Introduction

The 2017 International Federation of Classification Societies (IFCS) cluster analysis challenge concerns the analysis of data from 928 low back pain patients. For each patient, 121 variables have been measured (see van Mechelen and Vach (2018) for details). There are several challenging aspects in the data set. First, the data come from self-reported questionnaires and examinations and contain several missing values. The first task is to distinguish between values missing at random and not missing at random. Specifically, we use a two-pronged approach: inferring values missing not at random and imputing values missing at random.

A second challenging aspect is the presence of variables of mixed type; the data set contains binary, categorical, ordinal, and continuous variables. Most of the classical clustering techniques only work on one kind of variables. Moreover, the data set is characterized by a high number of variables. We propose a data transformation process called split-then-join. We divide the data set in binary or categorical and ordinal or continuous data and we transform the categorical data using multiple correspondence analysis (MCA) (Greenacre and Blasius, 2006). MCA is an extension of correspondence analysis (CA) (Greenacre, 1984) for multivariate data sets. It projects the observations in a lower dimensional subspace producing two major effects: It reduces the dimensionality of the data set, and it projects the observations on a continuous space. Specifically, the transformed data set contains only seven numerical dimensions derived from 73 categorical variables. The resulting data set can now "join" the continuous variables and create a new reduced data set of continuous variables.

Several clustering techniques can now be applied. K-means clustering (Hartigan and Wong, 1979) is one of the most well-known cluster analysis techniques thanks to its simplicity and speed. After a random initialization, k-means clustering finds the cluster centers that minimize the within-cluster variance over all variables. K-means works well for detecting spherical clusters; some research on the initialization procedure has been done to improve the performance of the method; a review can be found in (Steinley and Brusco, 2007); in our analysis we will use multiple starting points. Despite this improvement, one of the main drawbacks of k-means is a lack of robustness. Fuzzy k-means (FKM) (Ruspini, 1969) and Partition Around Medoids (PAM) (Kaufman and Rousseeuw, 1990) are both more robust; FKM is a soft clustering version of k-means, while PAM minimizes the dissimilarities between the points and the cluster centers. Unlike k-means, PAM uses data points as cluster centers and the Manhattan norm instead of the Euclidean distance. Another robust algorithm is Probabilistic Distance Clustering (PDClust) (Ben-Israel and Iyigun, 2008). It is based on the assumption that the probability of any point belonging to a cluster is inversely proportional to its distance from the center of that cluster.

Model-based clustering or mixture modeling, instead, assumes a density that is a convex combination of a finite number of component density functions; accordingly, it is very well suited to clustering problems. The Gaussian distribution (Titterington et al, 1985) has been one of the most widely used component distributions until recently. The recent literature has seen the use of different component distributions. Among others, mixture generalized hyperbolic distributions (MGHD) (Browne and McNicholas, 2015) is remarkable for its flexibility. We compared the results of several clustering techniques and we chose the best using the validation variables.

To further describe the clusters we performed a principal component analysis (Hotelling, 1933) on the baseline variables. The clusters are almost separable on the first two components. The remainder of this paper is structured as follows: Section 2 describes the data cleaning process. Section 3 describes the clustering methods and the interpretation of the results. Section 4 contains some concluding thoughts.

2 Data Cleaning Process

The original data set contains patient self-reported questionnaires and examinations recorded by clinicians, resulting in 112 baseline variables and three outcome variables for different time periods (two weeks, three months, and 12 months) following the initial clinical consultation. These variables covered various domains from pain history, activity limitation, work-related questions, validated questionnaires, fear avoidance, etc. There are many missing values in the data, as some patients and clinicians did not fill out all the questions. To choose the right imputation strategy, missing values need to be divided into missing at random and not missing at random. The values that are not missing at random can be deduced based on their relationship with other variables. For example, many questions inquire whether activities at work impact the patients' pain level, or whether their pain limits their activity at work. Obviously, patients who do not have jobs (i.e., students, unemployed, and pensioners) will not fill out these questions as they are not applicable. Using patients' employment status, we can infer that some of the missing values are not missing at random (i.e., they can be substituted with a new value indicating that the patient did not complete the questionnaires for legitimate reasons). Table 1 shows which missing values can be inferred. Our goal is to fill in as many missing values as possible by inference prior to imputing values that are assumed to be missing at random. We also removed 13 observations with more than 30 % missing values. At this point, we assume that the remaining missing values are missing at random and thus can be imputed.

Table 1: Treatment of values missing not at random.

Variables with missing values	**Reasons for missing**	**Inference**
fabq60 – fabq 140.	Questions involving pain level with respect to work condition; only to be answered if patient is working.	Replace NAs with new category (-1) if patient's employment situation, barb0, indicates not working.
facetextrot, facetsit, facetwalk, parasping_debut.	Questions only to be asked if patients answer yes to having dominating back pain.	Replace NAs with new category (-1) if patient does not have dominating back pain (i.e., domin_bp is 1).
musclegroup_palp	Question involving pain caused by different muscle groups.	Replace NAs with new category (-1) if patient has no pain referred from triggerpoint (i.e., triggerpoint is 0) and no replication of pain during palpation (i.e., musclepalp is 0).
musclepalp	Highly correlated with musclegroup_palp	Use musclegroup_palp to update NAs in musclepalp.

Our next task is to determine the appropriate techniques to impute the remaining missing values. Since the data set consists of mixed data types (i.e., binary, continuous, and categorical), the selected techniques must be appropriate for each type. Binary data can be forecasted with logistic regression while continuous data can be forecasted with linear regression. Categorical data can be forecasted with a multinomial method. For this reason, we favor the Multiple Imputation by Chained Equations (MICE) method (van Buuren and Groothuis-Oudshoorn, 2011). In R (R Core Team, 2016), MICE is implemented by a function with a similar name, mice(), which handles both data missing at random (MAR) and missing not at random (MNAR). This gives us an extra layer of comfort in the event that the inference process did not completely remove MNAR items. The MICE algorithm can be summarized as follows. An incomplete column is imputed using the default imputation methods: Predictive mean matching (numeric data), logistic regression imputation (binary data), polytomous regression (unordered categorical data with more than two levels), proportional odds model (ordered with more than two levels). Each incomplete column is then predicted based on all other columns in the data. For incomplete predictors, the most recently generated imputations are used prior to imputing the target column.

A cycling through all variables is considered one iteration. At the end of each iteration, the missing values are all replaced by predicted values. MICE converges when the variance between sequences is smaller than the variance with each individual sequence. van Buuren and Groothuis-Oudshoorn (2011) suggest 10-20 cycles. In addition to specifying the number of iterations, users can also set the number of imputations, m, for each missing value resulting in m data sets. We impute our data set using 10 iterations; each missing value was further imputed five times, which produced five imputed data sets. We then took the average (for numeric variables) and the mode (for categorical variables) of the five imputed values to obtain a complete data set. Some statisticians may consider taking the average of multiple imputed data sets an improper use of multiple imputation as this ignores the variability across the imputed datasets. Multiple imputation is usually considered for parameter estimations. Our ultimate goal is more complex and a subsequent combination of the different end results is not easy as the final output of our analyses is a partition of the units rather than a simple parameter. An improvement of our technique can be obtained by applying the selected method (i.e. MCA of the categorical variables and k-means on the combined data) on all the data sets obtained with multiple imputation and compare the clustering partition using the ARI to measure the variability of the results.

3 Clustering Methods

Clustering mixed data types presents some challenges. Popular algorithms such as k-means, fuzzy k-means, probabilistic distance clustering, and mixture models work well with numeric data but not with categorical data. Some clustering methods for categorical data have been proposed (e.g. Hwang et al (2006); van Buuren and Heiser (1989); D'Enza and Palumbo (2013)); however, they don't work on continuous data. Therefore, a data transformation is needed to obtain a final data set with a single data type, while preserving the relationships between the variables, before applying a clustering algorithm. To transform the data, we split our data set into two subsets: one purely categorical and the other purely numeric. Ordinal data were treated as numeric. We then applied multiple correspondence analysis (MCA) to the categorical subset and examined their principal coordinates. With MCA, we were able to reduce the dimension of the categorical subset from 73 to just seven. This was decided based on an eigenvalue contribution analysis which suggests that 95 percent of the total variation can be explained by seven dimensions. Since the principal coordinates were numeric linear combinations of categorical data, we appended these seven numeric columns to the purely numeric subset of 38 numeric variables, resulting in one numeric data set. Figure 1 illustrates this idea. It is worth to consider that when the number of numerical variables is much higher than the number of selected coordinates this approach may lead to underweighting the categorical variables. At this stage, our data is completely cleaned and ready to be clustered. We will refer to this as the **transformed data set** going forward.

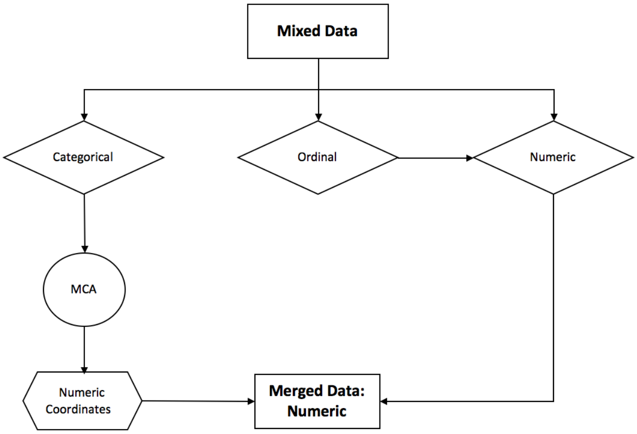


Figure 1: Data Transformation Method.

3.1 Determine the Number of Clusters

We subjected the transformed data set to the following seven clustering methods:

- K-means,
- Partition Around Medoids (PAM),
- Fuzzy K-means (FKM),
- Probabilistic Distance Clustering (PDClust),
- Mixture of Multivariate Normal Distributions (MVN),
- Mixture of Skewed-t Distributions (MST),
- Mixture Generalized Hyperbolic Distributions (MGHD).

In order to apply a clustering algorithm, we needed to determine the number of clusters in our data. Where appropriate, we used cluster comparison metrics such as the Caliński-Harabasz criterion (Caliński and Harabasz, 1974) and the Bayesian Information criterion (BIC) (Schwarz et al, 1978) to analyze the preliminary clustering results in order to determine the number of clusters.

Clustering solutions with a higher Caliński-Harabasz value are preferred over those with a lower Caliński-Harabasz value. In contrast, solutions with a smaller BIC are preferred over solutions with a larger BIC. The results suggested that three to four clusters exist within the data. We then used these values to evaluate the performance of the algorithms.

The mentioned clustering methods and other analytical tools are available in the following R packages:

- Clustering methods
  - cluster::PAM (Kaufman and Rousseeuw, 1990)
  - EMMIXskew::Emskew::mst (Wang et al, 2013)
  - EMMIXskew::Emskew::mvt (Wang et al, 2013)
  - fclust::FKM (Giordani et al, 2015)
  - FPDclustering::PDclust (Tortora and McNicholas, 2017)
  - MixGHD::ARI and MixGHD::MGHD (Tortora et al, 2017)
  - stats::kmeans (Hartigan and Wong, 1979)
- Other analytical tools
  - factoextra::fviz_cluster (Kassambara, 2017)
  - FactoMiner::MCA and FactoMiner::PCA (Lê et al, 2008)
  - MASS::mca (Venables and Ripley, 2002)
  - mice::mice (van Buuren and Groothuis-Oudshoorn, 2010)

For interested readers, the R codes will be provided as supplementary materials for reproducibility.

3.2 Clustering Algorithm Selection

We compared the obtained clustering results using the Adjusted Rand Index (ARI) (Rand, 1971). For the fuzzy techniques we used the hard clustering partition derived from the fuzzy output. ARI measures the similarity of two data partitions and ranges from zero to one. An ARI-value of zero indicates a match not different from a random match, while an ARI-value of one indicates a perfect match between two clustering results. Using the algorithm comparisons in table 2, we noticed that the k-means algorithm on average produced the highest pairwise ARI among the non-model-based clustering techniques (i.e. k-means, FKM, PAM, and PDClust). Therefore, we considered the k-means solution for further analysis. However, this partition is different from the partitions obtained using model-based clustering techniques, as indicated by lower ARI-values. This means that the partitions obtained using non-model-based and model-based techniques are different and need further study. Multivariate skew-t and GHD produce, on average, high ARI among the model-based techniques. Thus, we chose their solutions, too, as potential candidates for further analysis.

We then analyzed the three selected clustering solutions with respect to the validation variables and found that the k-means algorithm produces the best cluster separation (see further next subsection). Additionally, the projection of the transformed data onto the first two dimensions of a principal component analysis (PCA) of the transformed 112 baseline variables reveals that the three clusters produced by k-means are almost separable (figure 2). It is true that there are no gaps between the clusters but there is also very little overlap. Therefore, we could use the positions of the clusters on the principal components to understand the differences among them.

Table 2: Algorithm Comparison Using ARI.

Algorithm Comparisons	Adjusted Rand Index (ARI)
k-means vs. PAM	0.8060
k-means vs. FKM	0.8735
k-means vs. PDClust	0.5293
k-means vs. MVN	0.0359
k-means vs. MST	0.0359
PAM vs. FKM	0.7077
PAM vs. PDClust	0.5856
PAM vs. MVN	0.0098
PAM vs. MST	0.0521
FKM vs. PDClust	0.5404
FKM vs. MVN	0.0173
FKM vs. MST	0.0465
MVN vs. MST	0.7129
MVN vs. GHD	0.5147
MST vs. GHD	0.6837

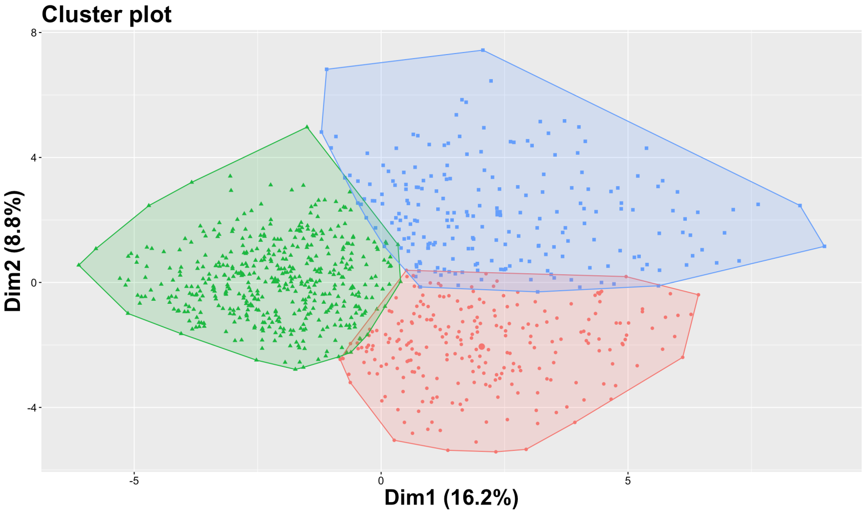


Figure 2: Three-Cluster Solution Resulting from K-Means Analysis Plotted in Projection of Transformed 112 Baseline Variables on their first two Principal Components.

3.3 Validation Variables

The k-means clustering algorithm separates the clusters reasonably well, as evidenced by distinctive patterns for all three clusters with respect to the validation variables. Figure 3 shows that patients in all three clusters experienced a decline in both LBP intensity and Roland-Morris scores at 12 months after the initial consultation. Patients in cluster 3 experienced the highest LBP intensity and Roland-Morris scores at the three time points, but they had the greatest perceived improvements compared to the other two clusters.

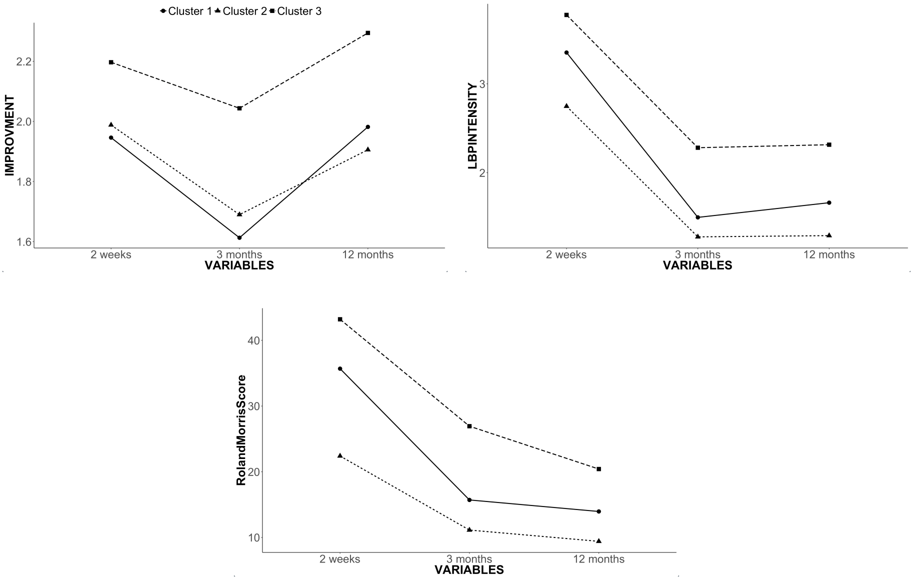


Figure 3: Validation Variable – Top left: Improvement, Top right: LBP Intensity, Bottom: Roland-Morris Score.

We also noticed an interesting pattern in the validation variables: While all three clusters exhibited a decline in LBP intensity and Roland-Morris score from the period of two weeks to three months, their perceived improvement actually declined in this period. One would think that when the physical condition improves (i.e., declining LBP intensity and Roland-Morris score), patients should experience an increase rather than a decline in perceived improvement.

We suspect that there is a time lag between when patients feel better versus when their conditions improve. This is also evidenced by the flattening out of LBP intensity while the perceived improvements rose from the three-month to the 12-month time point.

4 Concluding Thoughts

Up to this point, we have been performing all analyses based on the full set of 112 baseline variables. Plotting the first seven principal components derived from the transformed variables shows that the clusters are separable using the first two principal components (figure 4). Consequently, we believe it should be possible to describe the full data set at baseline using fewer variables.

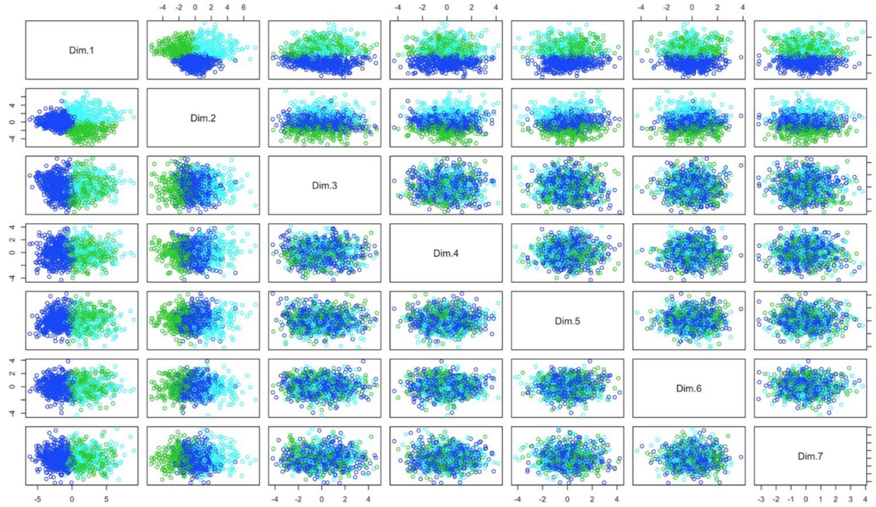


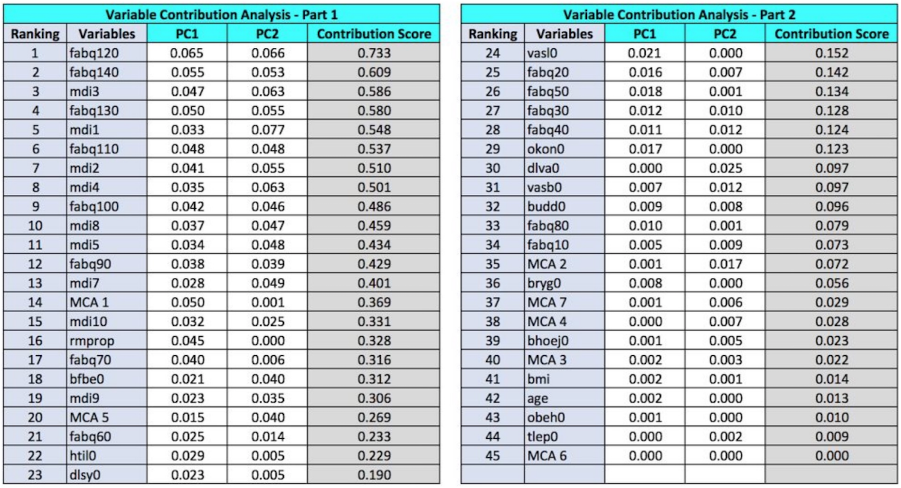
Figure 4: Pair-Plots of First 7 Principal Components Using Transformed Baseline Data Set.

To identify the important variables in the first two principal components, we analyzed their contributions to each of these components. We found that the data can be reduced to 27 numeric variables and 2 MCA dimensions. For this reduction we analyzed the contribution of each variable (see table 3) and ranked their contribution score (from highest to lowest). Variables with a contribution score lower than 0.1 are excluded. We will refer to the resulting smaller set of selected variables as the **reduced data set**. Table 3 and figure 5 below summarize these variables.

To validate this reduction, we reran the k-means algorithm using only the selected variables and compared the clustering result with the result for the full set of the transformed 112 baseline variables using the Adjusted Rand Index (ARI). This comparison produced a high ARI of 0.92, which indicates that the two clustering results are very similar. Figure 6 displays the three-cluster solution resulting from the k-means analysis of the reduced baseline data plotted in a projection of these data onto their first two principal components. This figure confirms that the selected variables perform just as well as the larger set with respect to separating the clusters and, therefore, are sufficient to describe the original dataset.

Figure 7 shows how the three clusters differ with respect to the numeric variables from the reduced data set. Note that these variables have been scaled to eliminate the effect of different ranges in their values. Cluster 1 is described by high scores on the Fear-Avoidance Beliefs Questionnaire (fabqs) and the Roland-Morris summary score (rmprop). Similarly, cluster 3 is described by high scores on items of the self-reported mood questionnaire, the Major Depression Inventory (mdi-variables). High mdi scores indicate poor mood or mental health. In contrast, patients in cluster 2 have the lowest scores on these items. In words, cluster 1 is characterized by above average back pain and limited functional activities. Patients in cluster 3 experienced higher levels of depression characterized by having low spirit, sadness, loss of appetite, and inability to sleep at night, and are also considered to be a higher risk group. Cluster 2 patients can be seen as the average patients as they exhibit an average score on the variables of all three categories mentioned.

Table 3: Contributions of Transformed Baseline Variables to first two Principal Components Derived from these Data.



Variable Contribution Analysis - Part 1				
Ranking	Variables	PC1	PC2	Contribution Score
1	fabq120	0.065	0.066	0.733
2	fabq140	0.055	0.053	0.609
3	mdi3	0.047	0.063	0.586
4	fabq130	0.050	0.055	0.580
5	mdi1	0.033	0.077	0.548
6	fabq110	0.048	0.048	0.537
7	mdi2	0.041	0.055	0.510
8	mdi4	0.035	0.063	0.501
9	fabq100	0.042	0.046	0.486
10	mdi8	0.037	0.047	0.459
11	mdi5	0.034	0.048	0.434
12	fabq90	0.038	0.039	0.429
13	mdi7	0.028	0.049	0.401
14	MCA 1	0.050	0.001	0.369
15	mdi10	0.032	0.025	0.331
16	rmprop	0.045	0.000	0.328
17	fabq70	0.040	0.006	0.316
18	bfbe0	0.021	0.040	0.312
19	mdi9	0.023	0.035	0.306
20	MCA 5	0.015	0.040	0.269
21	fabq60	0.025	0.014	0.233
22	htil0	0.029	0.005	0.229
23	dlsy0	0.023	0.005	0.190

Variable Contribution Analysis - Part 2				
Ranking	Variables	PC1	PC2	Contribution Score
24	vasl0	0.021	0.000	0.152
25	fabq20	0.016	0.007	0.142
26	fabq50	0.018	0.001	0.134
27	fabq30	0.012	0.010	0.128
28	fabq40	0.011	0.012	0.124
29	okon0	0.017	0.000	0.123
30	dlva0	0.000	0.025	0.097
31	vasb0	0.007	0.012	0.097
32	budd0	0.009	0.008	0.096
33	fabq80	0.010	0.001	0.079
34	fabq10	0.005	0.009	0.073
35	MCA 2	0.001	0.017	0.072
36	bryg0	0.008	0.000	0.056
37	MCA 7	0.001	0.006	0.029
38	MCA 4	0.000	0.007	0.028
39	bhoej0	0.001	0.005	0.023
40	MCA 3	0.002	0.003	0.022
41	bmi	0.002	0.001	0.014
42	age	0.002	0.000	0.013
43	obeh0	0.001	0.000	0.010
44	tlep0	0.000	0.002	0.009
45	MCA 6	0.000	0.000	0.000

- Variable X_k's contribution to principal component Y_i is defined as: $\frac{r^2_{Y_i,X_k}}{\sum_{i=1}^{p} r^2_{Y_i,X_k}}$ where p is the number of variables, $r^2_{Y_i,X_k}$ is the squared correlation between variable X_k, and principal component Y_i and $\sum_{i=1}^{p} r^2_{Y_i,X_k}$ is the total sum of the squared correlation coefficients between variable X_k and each of the components Y_i. A larger variable contribution implies a greater influence on the component than a smaller variable contribution.

- The contribution score for each variable is defined as: $\sum_{i=1}^{p} C_{ki}\lambda_i$ where λ_i is the i^{th}-eigenvalue and C_{ki} is the contribution of the k^{th}-variable to the i^{th} component.

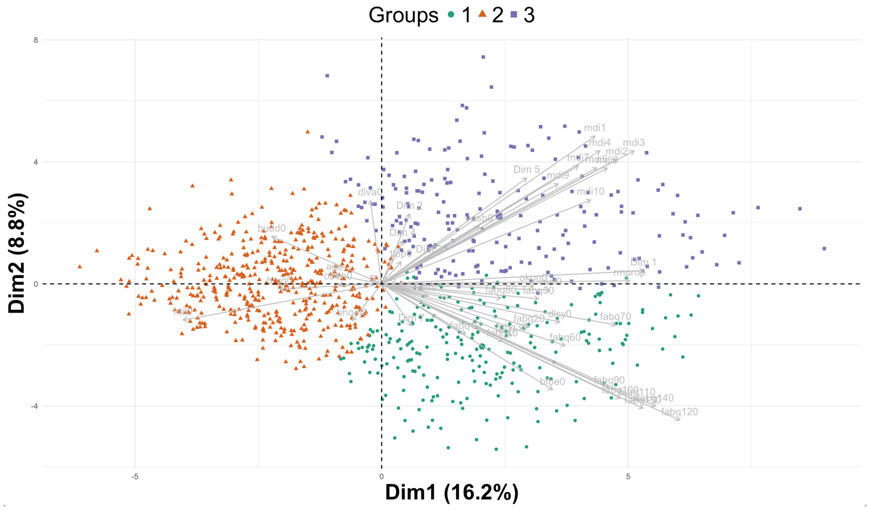


Figure 5: Biplot of first two Principal Components of Transformed Baseline Data with Patients Color Coded on the Basis of K-Means Cluster Membership.

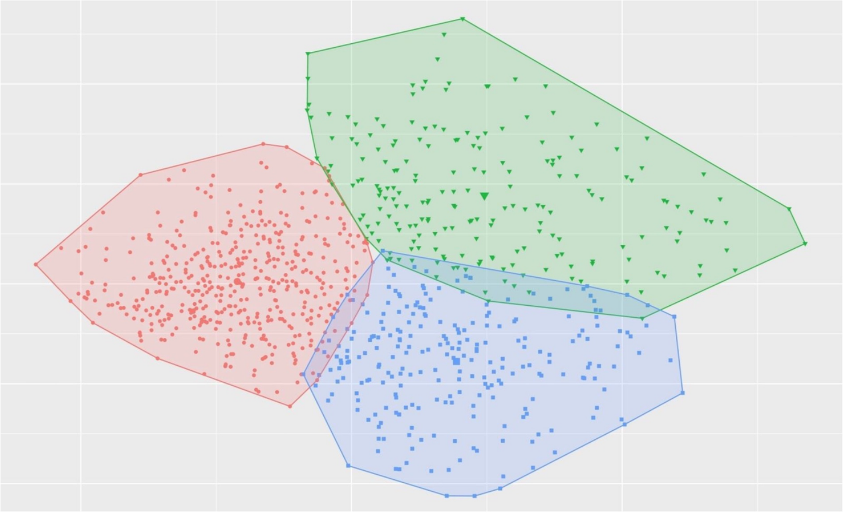


Figure 6: Three-Cluster Solution Resulting from K-Means Analysis of Reduced Baseline Data Plotted in Projection of these Data onto their first two Principal Components.

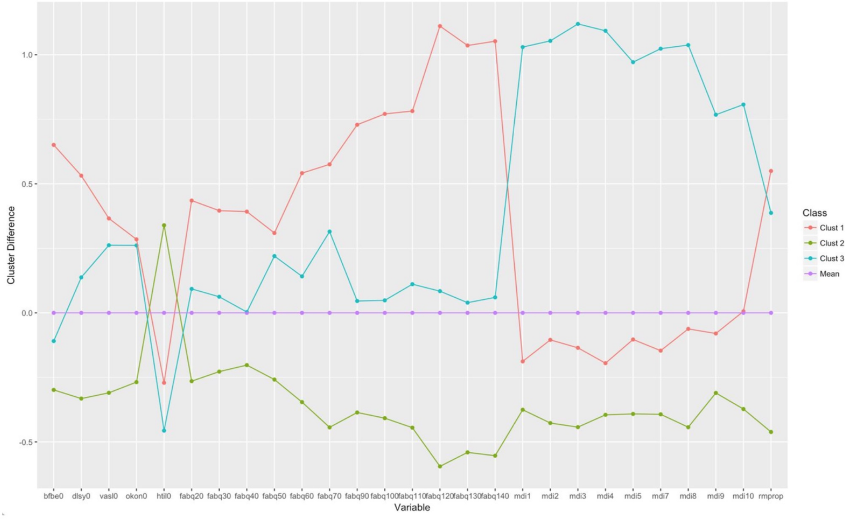


Figure 7: Average Scores of three K-Means Clusters on 27 Scaled Numerical Variables from Reduced Data Set.

Acknowledgements This work was supported by the just in time funds of the College of Science of San Jose Sate University.

References

Ben-Israel A, Iyigun C (2008) Probabilistic D-clustering. Journal of Classification 25(1):5–26, Springer, DOI 10.1007/s00357-008-9002-z

Browne RP, McNicholas PD (2015) A mixture of generalized hyperbolic distributions. Canadian Journal of Statistics 43(2):176–198, DOI 10.1002/cjs.11246

van Buuren S, Groothuis-Oudshoorn K (2010) MICE: Multivariate Imputation by Chained Equations. Journal of Statistical Software, URL `https://cran.r-project.org/web/packages/mice/`, R Package version 2.30

van Buuren S, Groothuis-Oudshoorn K (2011) MICE: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software 45(3):1–68, DOI 10.18637/jss.v045.i03

van Buuren S, Heiser WJ (1989) Clustering n objects into k groups under optimal scaling of variables. Psychometrika 54(4):699–706, DOI 10.1007/BF02296404

Caliński T, Harabasz J (1974) A dendrite method for cluster analysis. Communications in Statistics, Theory and Methods 3(1):1–27, Taylor & Francis, DOI 10.1080/03610927408827101

D'Enza AI, Palumbo F (2013) Iterative factor clustering of binary data. Computational Statistics 28(2):789–807, Springer, DOI 10.1007/s00180-012-0329-x

Giordani P, Brigida Ferraro M, Serafini A (2015) URL `https://CRAN.R-project.org/package=fclust`, R package version 1.1.2

Greenacre MJ (1984) Theory and application of correspondence analysis. Academic Press, London. ISBN: 978-0-122990-50-2

Greenacre MJ, Blasius J (2006) Multiple correspondence analysis and related methods. Chapman & Hall/CRC. ISBN: 978-1-584886-28-0, URL `https://www.crcpress.com/Multiple-Correspondence-Analysis-and-Related-Methods/Greenacre-Blasius/p/book/9781584886280`

Hartigan JA, Wong MA (1979) A K-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics) 28:100–108, Wiley, DOI 10.2307/2346830

Hotelling H (1933) Analysis of a Complex of Statistical Variables into Components. Journal of Educational Psychology 24(6):417–441

Hwang H, Dillon WR, Takane Y (2006) An extension of multiple correspondence analysis for identifying heterogenous subgroups of respondents. Psychometrika 71(1):161–171, DOI 10.1007/s11336-004-1173-x

Kassambara A (2017) Factoextra: Print method for an object of class factoextra, . URL `https://cran.r-project.org/web/packages/factoextra/`, R package version 1.04

Kaufman L, Rousseeuw PJ (1990) Partitioning around medoids (program PAM). Finding groups in data: an introduction to cluster analysis, p. 68–125, John Wiley & Sons, Hoboken (USA), DOI 10.1002/9780470316801

Lê S, Josse J, Husson F, et al (2008) FactoMineR: An R package for multivariate analysis. Journal of Statistical Software 25(1):1–18, DOI 10.18637/jss.v025.i01

van Mechelen I, Vach W (2018) Cluster analyses of a target data set in the IFCS cluster benchmark data repository: Introduction to the special issue. Archives of Data Science, Series B 1(1):1–12, DOI 10.5445/KSP/1000085952/01

R Core Team (2016) R: A Language and Environment for Statistical Computing, . R Foundation for Statistical Computing, Vienna (Austria)

Rand WM (1971) Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association 66(336):846–850, Taylor & Francis Group, DOI 10.2307/2284239

Ruspini EH (1969) A new approach to clustering. Information and Control 15(1):22–32, DOI 10.1016/S0019-9958(69)90591-9

Schwarz G, et al (1978) Estimating the dimension of a model. The Annals of Statistics 6(2):461–464, Institute of Mathematical Statistics

Steinley D, Brusco MJ (2007) Initializing k-means batch clustering: A critical evaluation of several techniques. Journal of Classification 24(1):99–121, Springer, DOI 10.1007/s00357-007-0003-0

Titterington DM, Smith AFM, Makov UE (1985) Statistical analysis of finite mixture distributions. John Wiley & Sons, Chichester. ISBN: 978-0-471907-63-3

Tortora C, McNicholas PD (2017) FPDclustering: PD-clustering and factor PD-clustering. URL `https://cran.r-project.org/web/packages/FPDclustering/index.html`, R package version 1.2

Tortora C, El-Sherbiny A, Browne RP, Franczak BC, McNicholas PD (2017) MixGHD: Model based clustering and classification using the mixture of generalized hyperbolic distributions. URL `https://cran.r-project.org/web/packages/MixGHD/index.html`, R package version 2.1

Venables WN, Ripley BD (2002) Modern applied statistics with S, 4th edn. Springer, New York (USA), DOI 10.1007/978-0-387-21706-2

Wang K, Ng A, McLachlan GJ (2013) EMMIXskew: The EM algorithm and skew mixture distribution. DOI 10.1109/DICTA.2009.88, URL `https://cran.r-project.org/web/packages/EMMIXskew/index.html`, R package version 1.0.1