

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 of cluster analysis for this data set. We will discuss our data cleaning process, which we view as a two-pronged approach: inferring values missing not at random and imputing values 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 this data transformation process split-then-join. Finally, we offer our interpretation of the clustering results with respect to validation variables and final 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

1 Introduction

The 2017 International Federation of Classification Societies (IFCS) data challenge concerns the analysis of 928 low back pain patients. For each patient, 121 variables have been measured (see cite(INTRODUCTION) for details). There are several challenging aspects in the data set; first, the data come from self-reported questionnaires and examinations, therefore, there are 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 variable; 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 dimensions of the data set, and it projects the observations on a continuous space. Specifically, the transformed data set contains only seven numerical factors out of 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 detecting spherical clusters; some research on the starting criterion has been done to improve the performance of the method, a review can be found in (Steinley et Brusco, 2007), in the following we will use multiple starting points. Despite this improvement, one of the main drawbacks is the 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. Differently from k-means, PAM uses data points as cluster centers and the

Manhattan norm instead of 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 the distance from the center of that cluster. Model-based clustering or mixture models, instead, assumes that a population is a convex combination of a finite number of density functions; accordingly, both are very well suited to clustering problems. The Gaussian distribution (Titterington et al, 1985) have been some of the most used until recently. The recent literature has seen the use of different 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 principal component analysis (Hotelling, 1933). The clusters are well separated on the first two components. The paper is structured as follows: Section 2 describes the data cleaning process. Sections 3 describes the clustering methods and the result interpretation. Section 4 contains some conclusive thoughts.

2 Data Cleaning Process

The original data set contains patient self-reported questionnaires and clinicians recorded examinations, 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 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 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 missing at random. We also removed 13 observations with more than 30% missing values. At this point, we assume 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, paraspinq_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 have 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 robust 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 with a function by 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. MICE algorithm can be generalized as followed. An incomplete column is imputed using the default imputation methods: predictive mean matching (numeric data), logistic regression imputation (binary data), polyto-

mous 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.

The cycling through each variable 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 imputed five times, which produced five imputed data sets. We then took the average (for numeric data) and the mode (for categorical) of the five data sets 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 that ignores the variability across multiple imputed datasets. Multiple imputation is usually considered for parameter estimations. Our data set is more complex and a subsequent combination of the results is not easy, the output of our algorithm is a partition of the units not a simple parameter. An improvement of our technique can be obtained 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 categorical data. Some clustering methods for categorical data have been proposed (e.g. Hwang et al (2006); Van Buuren and Heiser (1989); Iodice D’Enza and Palumbo (2013)); however, they don’t work on continuous data. Therefore, data transformation is needed to obtain the final data set with consistent 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 cate-

gorical 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.

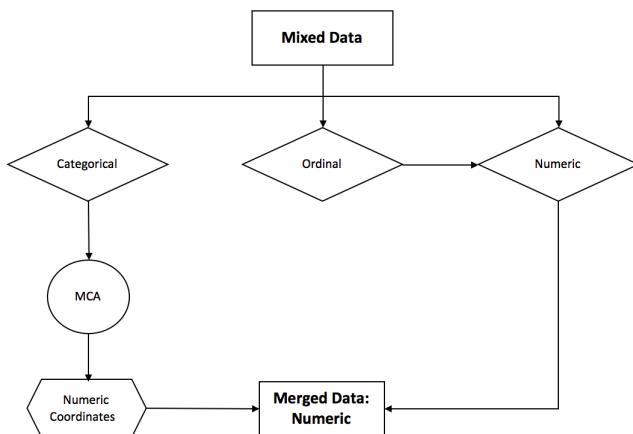


Fig. 1 Data Transformation Method

3.1 Determine the Number of Clusters

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 Calinski-Harabasz criterion (Calinski and Harabasz, 1974) and Bayesian Information criterion (BIC) (Schwarz et al, 1978) to analyze the preliminary clustering results for the following clustering algorithms while varying the number of clusters:

- 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).

Clustering solutions with higher Calinski-Harabasz criterion are preferred over those with lower Calinski-Harabasz criterion. In contrast, solutions with smaller BIC are preferred over large BIC. The initial run suggested that three to four clusters exist within the data. We then used these known values to re-run these algorithms and evaluate their performance. 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 (Wang et al, 2013)
 - fclust::FKM (Giordani and Ferraro, 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, 2000-2010)

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. ARI of zero indicates a match not different from a random match, while ARI of one indicates a perfect match between two clustering results. Using the algorithm comparisons in Table 2, we noticed that, among the non model based clustering technique, i.e. k-means, FKM, PAM, and PDClust, k-means algorithm on average, produced the highest pairwise ARI, therefore, we considered k-means solution for further analysis. However, this partition is different form the partitions obtained using model based clustering techniques, the ARI are lower. 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 also produce, on average, high ARI among the model based techniques. Thus, we chose these three algorithms as potential candidates for further analysis.

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

We then analyze those three clustering solutions with respect to the validation variables and found that k-means produces the best cluster separation. Additionally, the projection of the transformed data onto the first two dimensions of principal component analysis (PCA) reveals that the three clusters

produced by k-means are well separated (Figure 2). Our goal here is to use Figure 2 to help with the interpretation of the clusters. It is true that there is no separation but there is very little overlapping. Therefore, we use the position of the clusters with respect to the variables on the PCA factors to understand the differences among the clusters.

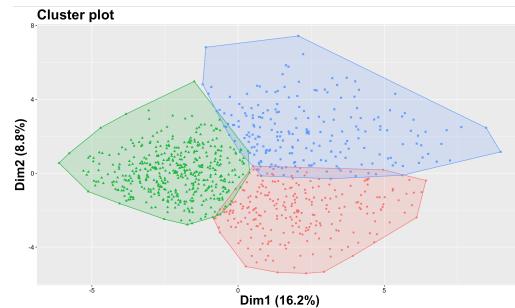


Fig. 2 Data Projection onto Dimension 1 and 2 for 112 Variables

3.3 Validation Variables

K-means clustering algorithm separates the clusters quite well, as evident 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, but they had the greatest perceived improvements compared to the other two clusters.

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 as physical condition improves (i.e., declining LBP intensity and Roland Morris score), patients should experience an increase rather than decline in perceived improvement. We suspect there is a time lag between when patients feel better versus when their conditions improved. This is evident by the flattening out of LBP inten-

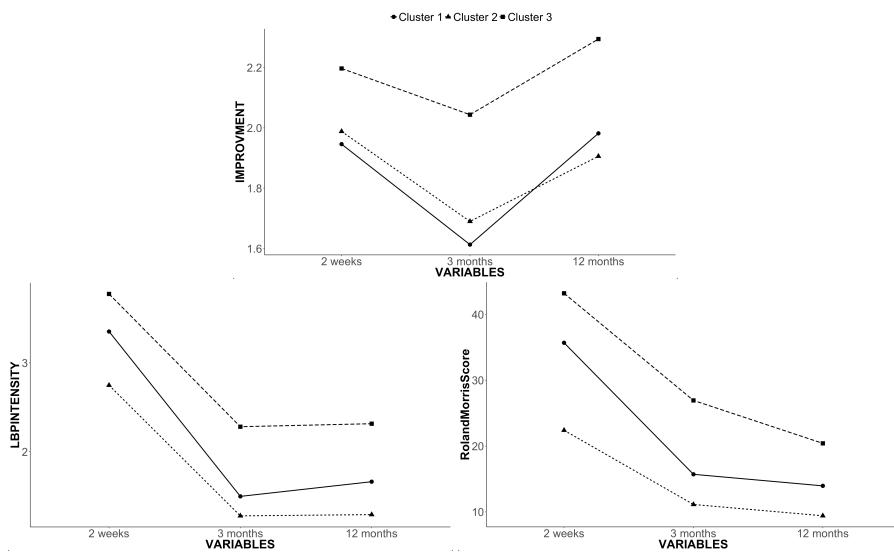


Fig. 3 Validation Variable—Top: Improvement, Bottom left: LBP Intensity, Bottom right: Roland Morris Score

sity while the perceived improvements rose from three-month to the 12-month period.

4 Final Thoughts

Up to this point, we have been performing all analyses based on the full set of 112 variables. Plotting the first seven principal components shows that the clusters are well separated using the first two principal components (Fig. 4). Consequently, we believe it is possible to describe the full data set using fewer variables.

To identify the important variables in the first two principal components, we analyzed their contributions to each dimension as well as examined their biplot. We found that the data can be reduced to 27 numeric and 28 categorical variables. This reduction is derived from a posteriori analysis. We analyzed the contribution of each variable (see Table 3) and ranked their contribution score (from highest to lowest). Variables with lower than 0.1 contribution score are

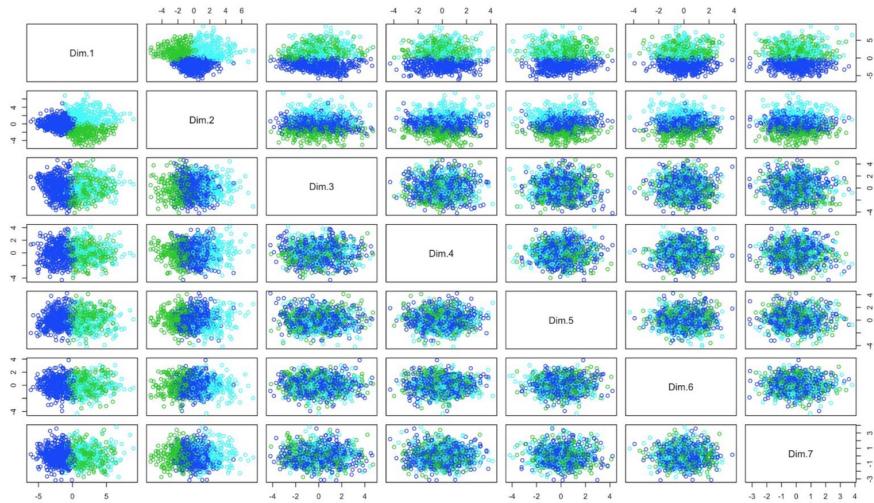


Fig. 4 Pair-Plots of First 7 Principal Components Using Transformed Data set

excluded. We will refer to this smaller set of selected variables as the **reduced data set**. Table 3 and Figure 5 below summarize these important variables.

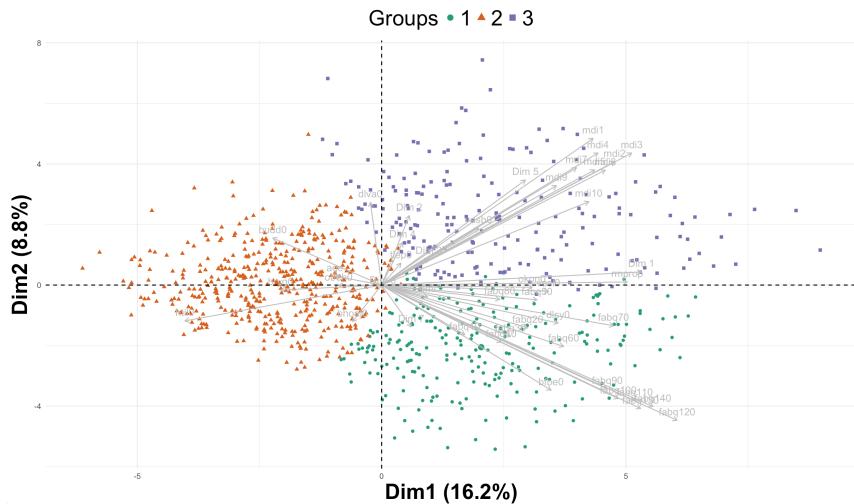


Fig. 5 Important Variables on Dimension 1 and 2

Table 3 Variable Contribution Analysis

Variable Contribution Analysis - Part 1					Variable Contribution Analysis - Part 2				
Ranking	Variables	PC1	PC2	Contribution Score	Ranking	Variables	PC1	PC2	Contribution Score
1	fabq120	0.065	0.066	0.733	24	vasl0	0.021	0.000	0.152
2	fabq140	0.055	0.053	0.609	25	fabq20	0.016	0.007	0.142
3	mdi3	0.047	0.063	0.586	26	fabq50	0.018	0.001	0.134
4	fabq130	0.050	0.055	0.580	27	fabq30	0.012	0.010	0.128
5	mdi1	0.033	0.077	0.548	28	fabq40	0.011	0.012	0.124
6	fabq110	0.048	0.048	0.537	29	okon0	0.017	0.000	0.123
7	mdi2	0.041	0.055	0.510	30	diva0	0.000	0.025	0.097
8	mdi4	0.035	0.063	0.501	31	vasb0	0.007	0.012	0.097
9	fabq100	0.042	0.046	0.486	32	budd0	0.009	0.008	0.096
10	mdi8	0.037	0.047	0.459	33	fabq80	0.010	0.001	0.079
11	mdi5	0.034	0.048	0.434	34	fabq10	0.005	0.009	0.073
12	fabq90	0.038	0.039	0.429	35	MCA 2	0.001	0.017	0.072
13	mdi7	0.028	0.049	0.401	36	bryg0	0.008	0.000	0.056
14	MCA 1	0.050	0.001	0.369	37	MCA 7	0.001	0.006	0.029
15	mdi10	0.032	0.025	0.331	38	MCA 4	0.000	0.007	0.028
16	rmprop	0.045	0.000	0.328	39	bhoej0	0.001	0.005	0.023
17	fabq70	0.040	0.006	0.316	40	MCA 3	0.002	0.003	0.022
18	bfbe0	0.021	0.040	0.312	41	bml	0.002	0.001	0.014
19	mdi9	0.023	0.035	0.306	42	age	0.002	0.000	0.013
20	MCA 5	0.015	0.040	0.269	43	obeh0	0.001	0.000	0.010
21	fabq60	0.025	0.014	0.233	44	tiep0	0.000	0.002	0.009
22	htil0	0.029	0.005	0.229	45	MCA 6	0.000	0.000	0.000
23	dlsy0	0.023	0.005	0.190					

- Variable X_k 's contribution to principal component Y_i is defined as: $\frac{P_{Y_i, X_k}^2}{\sum_{i=1}^p P_{Y_i, X_k}^2}$ where p is the number of variables, P_{Y_i, X_k}^2 is the square correlation between variable X_k , and principal component Y_i and $\sum_{i=1}^p P_{Y_i, X_k}^2$ is the total sum square of the correlation coefficients of the components Y_i . Larger variable contribution has greater influence on a component than smaller variable contribution.
- 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 k^{th} -variable contribution of the i^{th} component.

To validate this, we reran the k-means algorithm using only the selected variables and compared its clustering result against the result from the full set of 112 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. The projection of this reduced data onto the first two principal components (Figure 6) confirms that the selected variables perform just as well as the larger set with respect to separating the clusters and therefore are sufficient enough to describe the original dataset.

Fig. 7 shows how the three clusters differ with respect to the numeric variables from the transformed data set. Note 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 Roland Morris Summary (rmprop). Similarly, cluster 3 is described by high scores on self-reported mood questionnaires Major Depression Inventory (mdi's). High mdi scores indicates poor mood or mental health. In contrast, patients in clus-

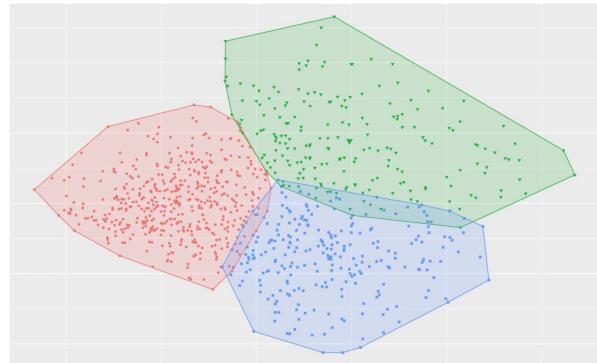


Fig. 6 Projection of Reduced Data onto the First Two Principal Components

ter 2 have the lowest score in these questionnaires. In words, cluster 1 is characterized by patients who experienced 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 the average score in all four categories mentioned.

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

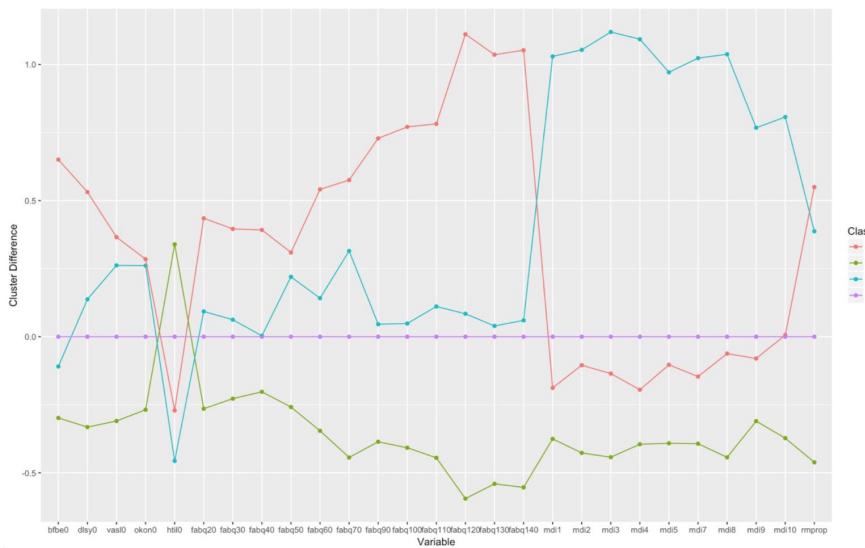


Fig. 7 Cluster Difference Using Scaled Variables

References

- Ben-Israel A, Iyigun C (2008) Probabilistic d-clustering. *Journal of Classification* 25(1):5–26
- Browne RP, McNicholas PD (2015) A mixture of generalized hyperbolic distributions. *Canadian Journal of Statistics* 43(2):176–198
- Browne RP, El-Sherbiny A, McNicholas PD (2015) Mixture: Mixture models for clustering and classification. R package version 1.4
- Caliński T, Harabasz J (1974) A dendrite method for cluster analysis. *Communications in Statistics, theory and Methods* 3(1):1–27
- Giordani P, Ferraro MB (2015) Fclust: Fuzzy k-means. R package version 1.1.2
- Greenacre MJ (1984) Theory and application of correspondence analysis. Academic Press, London
- Greenacre MJ, Blasius J (2006) Multiple correspondence analysis and related methods. Chapman & Hall/CRC
- Hartigan JA, Wong MA (1979) A k-means clustering algorithm. *Applied Statistics* 28:100–108

- Hotelling H (1933) Analysis of a complex of statistical variables into components. *Journal of Educational Psychology* 24:417–441
- Hwang H, Dillon WR, Takane Y (2006) An extension of multiple correspondence analysis for identifying heterogenous subgroups of respondents. *Psychometrika* 71:161–171
- Iodice D'Enza A, Palumbo F (2013) Iterative factor clustering of binary data. *Computational Statistics* 28(2):789–807
- Kassambara A (2017) Factoextra: Print method for an object of class 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 pp 68–125
- Lê S, Josse J, Husson F, et al (2008) FactoMineR: an R package for multivariate analysis. *Journal of statistical software* 25(1):1–18
- 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
- Ruspini EH (1969) A new approach to clustering. *Information and Control* 15(1):22 – 32
- Schwarz G, et al (1978) Estimating the dimension of a model. *The Annals of Statistics* 6(2):461–464
- Steinley D and Michael J. Brusco (2007) Initializing k-means batch clustering: A critical evaluation of several techniques." *Journal of Classification* 24.1(2007) Local optima in k-means clustering: What you don't know may hurt you. *Journal of Classification* 24(1):99–121.
- Titterington DM, Smith AFM, Makov UE (1985) Statistical analysis of finite mixture distributions. John Wiley & Sons, Chichester
- Tortora C, McNicholas PD (2017) FPDclustering: PD-clustering and factor PD-clustering. 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. R package version 2.1
- van Buuren S, Groothuis-Oudshoorn K (2000-2010) Mice: Multivariate imputation by chained equations. 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)

- Van Buuren S, Heiser WJ (1989) Clustering n objects into k groups under optimal scaling of variables. *Psychometrika* 54:699–706
- Venables WN, Ripley BD (2002) Modern applied statistics with S. Fourth Edition. Springer, New York
- Wang K, Ng A, McLachlan GJ (2013) EMMIXskew: The EM algorithm and skew mixture distribution. R package version 1.0.1