K-means Clustering on Multiple Correspondence Analysis Coordinates

Le Phan, Hongzhe Liu
Advisor: Cristina Tortora

Methodology

- Data cleaning
  - Inferring missing values
  - Imputing missing values & techniques
- Clustering methods
  - Data transformation
  - Choosing optimal number of clusters
  - Select appropriate clustering algorithm
- Validation & key findings

Data Cleaning

1. Inferring missing values
   a. What is it?
      i. Data missing not at random
      ii. Fill in missing values using known relationship with other variables
   b. Goal: fill in as many missing values as possible by inference before imputing

Table1: Treatment of Values Missing Not at Random

Variables with missing values	Reason for missing not at random	Inference
fabq60, fabq70, fabq80, fabq90, fab100, fabq110, fab120, fab130, fabq140	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, paraspin_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 (domin_bp = 1 for no)
musclegroup_palp	Question involving pain caused by different muscle groups.	Replace NAs with new category - 1 if patient have no pain referred from triggerpoint(triggerpoint = 0 for no) and no replication of pain during muscle palpation(musclepalp = 0)
musclepalp	Highly correlated with musclegroup_palp	Use musclegroup_palp to update NAs in musclepalp

Data Cleaning (cont.)

2. Imputing missing values
    a. Fill in values missing at random
    b. Method:
        i. Multiple Imputation by Chained Equations (MICE)
        ii. R-package and function: mice::mice
    c. More information on MICE: Stef van Buuren and Karin Groothuis-Oudshoorn, *"mice: Multivariate Imputation by Chained Equations in R", 2011*

MICE algorithm:

1. Initial imputation — random draw from the data as placeholders.
2. Placeholder for first variable is set back to missing.
3. The appropriate model (i.e., logistic regression, linear regression, multinominal, etc.) is used to predict the missing values.
4. The missing values is replaced with the predicted values.
5. Step 2-4 is repeated for the next variable.

Clustering Method – Data Transformation

- **The clustering problem:** mixed data types (i.e numeric and categorical)
  - Traditional cluster algorithms (e.g., k-means, fuzzy k-means, etc.) work with numeric but not categorical data
- **Solution to mixed data:** split-then-join
  - Split into two subsets
    - Pure categorical
    - Pure numeric
  - Transform categorical data with MCA
  - Reduce dimension
  - Append MCA's numeric coordinate to numeric data subset

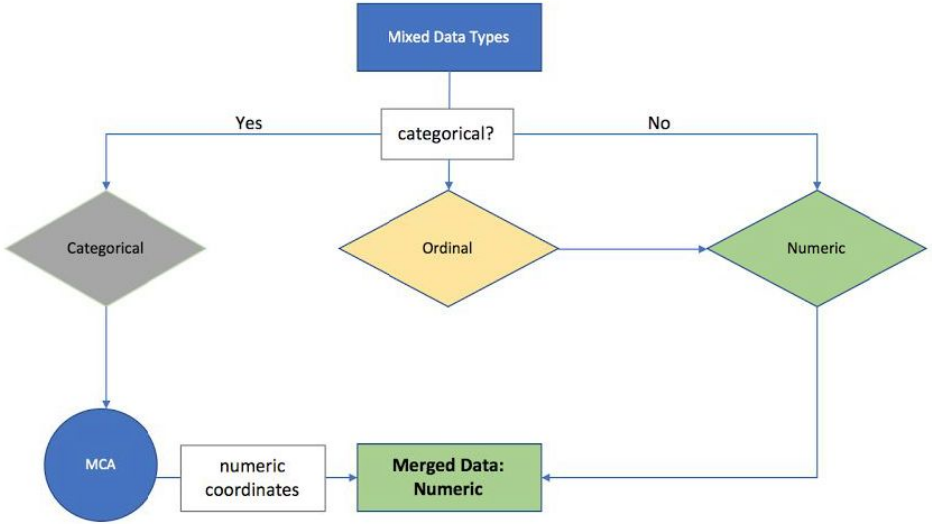

Figure 1: Data Transformation Method

Clustering Methods – Determine the Number of Clusters

- What is the optimal number of clusters
  - K-means
  - Partition Around Medoids (PAM)
  - Fuzzy K-means (FKM)
  - Probabilistic Distance Clustering (PDClust)
  - Gaussian Parsimonious Clustering Models
  - Mixture Generalized Hyperbolic Distributions (MGHD)
- Used Calinski-Hasbaraz criterion and Bayesian Information criterion
- These clustering algorithms suggest 3 or 4

Clustering Methods – Algorithm Selection

Algorithm Comparison	Adjusted Rand Index (ARI)
K-means vs. Partition around medoids (PAM)	0.8060
K-means vs. Fuzzy k-means (FKM)	0.8735
K-means vs. Probabilistic distance clustering (PDClust)	0.5293
K-means vs. multivariate normal distribution	0.0359
PAM vs. FKM	0.7077
PAM vs. PDClust	0.5856
FKM vs. PDClust	0.5404
Multivariate normal vs. multivariate skew-t	0.7129
Multivariate normal vs. GHD	0.5147
Multivariate skew-t vs. GHD	0.6837

- Goal: narrow down to two or three algorithm for comparison
- Compare clustering solutions
  - Clusters well separated?
- Final decision: K-means with 3 clusters
  - Clusters are well separated with respect to validation variables
  - To confirm our decision, we performed PCA and plot first two dimensions

Figure 2: Data Projection onto Dimension 1 and 2 for 112 Variables

Interpreting Clustering Result with Validation Variables

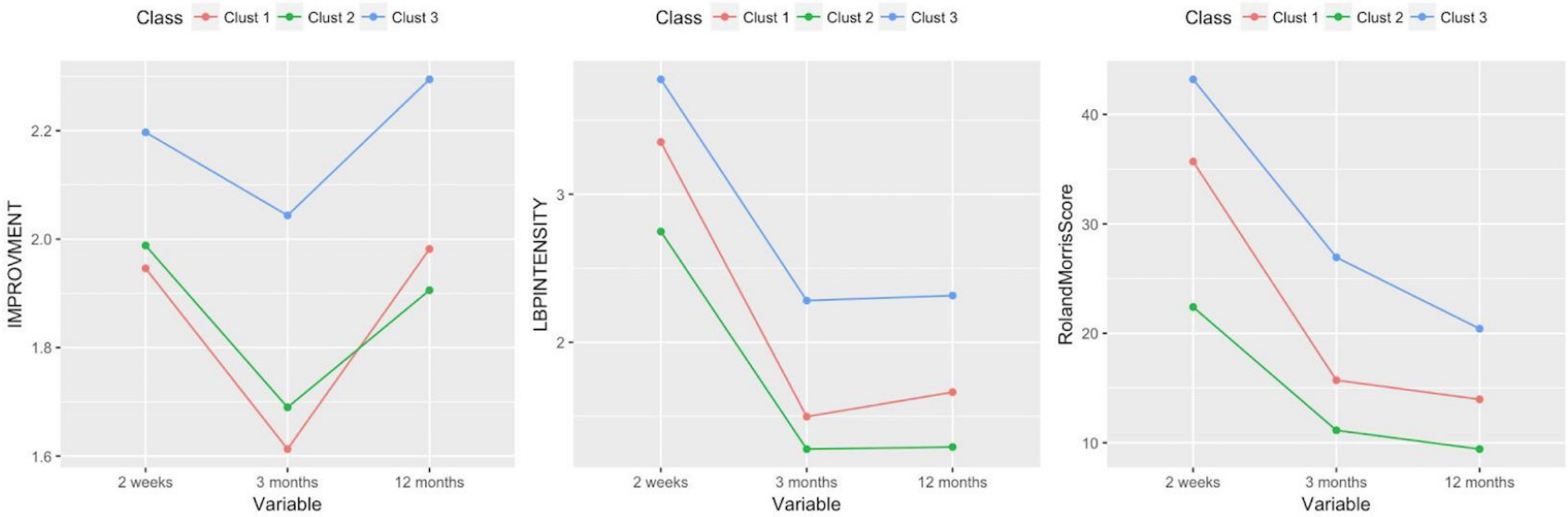

Figure 6: Validation Variables

Interpreting Clustering Result (cont.)

- Clusters are well separated
- All 3 groups experienced decline in LBP intensity and Roland Morris (RM) score—twelve months after initial consultation
- Cluster 3:
  - Highest LBP intensity and RM score but greatest perceived improvements
  - Considered at “high risk” for having back pain disability
- Interesting pattern:
  - All 3 groups experienced declined in LBP intensity and RM score (2-week to 3-month) while their perceived improvements actually declined during this period.
  - Is there a “time lag”?

Variables Selection

- Possible to describe dataset using fewer variables
- Reduced dataset
  - 27 numeric
  - 28 categorical

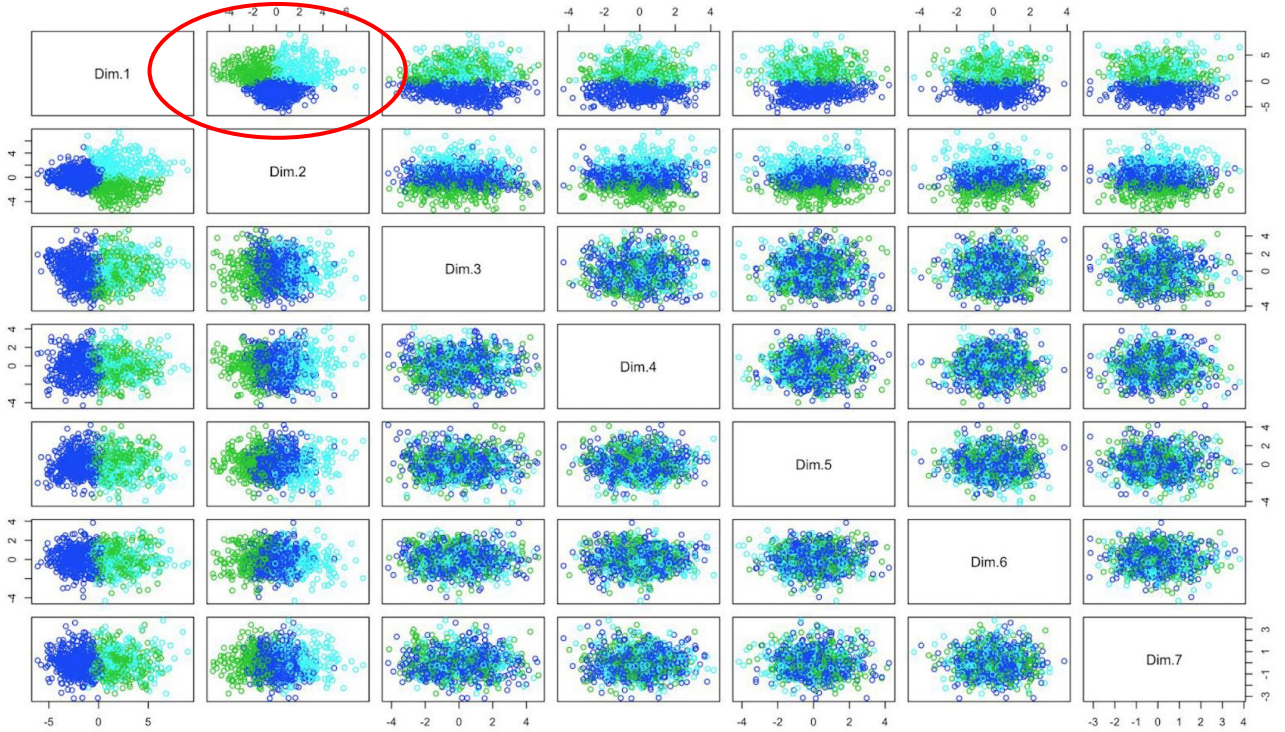

Figure 8: Pair-plots of First 7 Principal Components Using Transformed dataset

Identifying Important Variables

Figure 9: Important Variables in Each Cluster

Table 3: Variable Contribution Analysis

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: $\frac{P^2_{Y_i,X_k}}{\sum_{i=1}^{p} P^2_{Y_i,X_k}}$ where $P^2_{Y_i,X_k}$ is the square correlation between variable X_k and principal component Y_i and $\sum_{i=1}^{p} P^2_{Y_i,X_k}$ 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 to the i^{th} component.

Re-run K-mean with Reduced Dataset

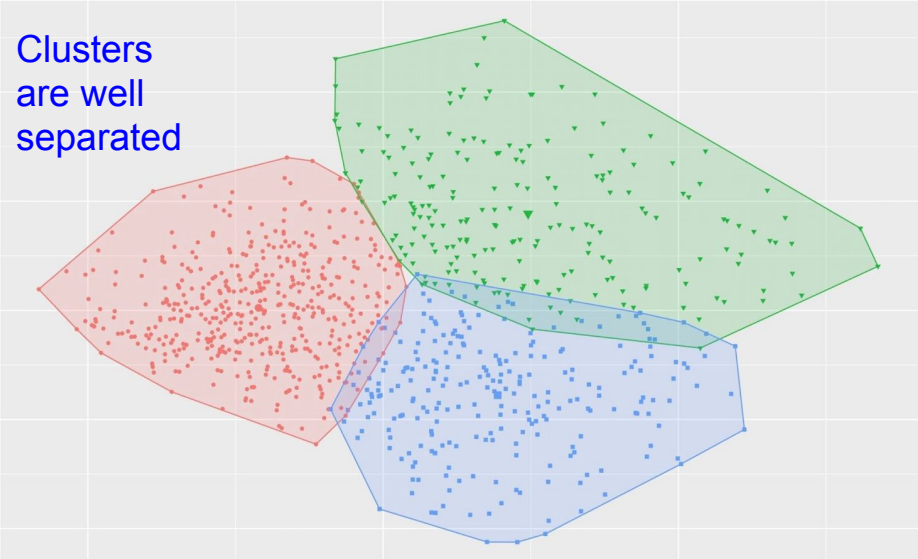

Figure 10: Projection of Reduced Data onto the First Two Principal Components

List of Important Variables

Continuous & Ordinal Variables			Categorical Variables		
fabq120	mdi8	fabq60	start_risk	rm50	rm90
fabq140	mdi5	htil0	rm30	start90	rm110
mdi3	fabq90	dlsy0	rm100	rm120	start80
fabq130	mdi7	vasl0	rm180	start30	rm80
mdi1	mdi10	fabq20	rm220	rm160	rm150
fabq110	rmprop	fabq50	rm140	rm200	romflex
mdi2	fabq70	fabq30	rm70	start40	start60
mdi4	bfbe0	fabq40	rm60	rm40	rm170
fabq100	mdi9	okon0	rm130	rm10	start50
				romrotl	

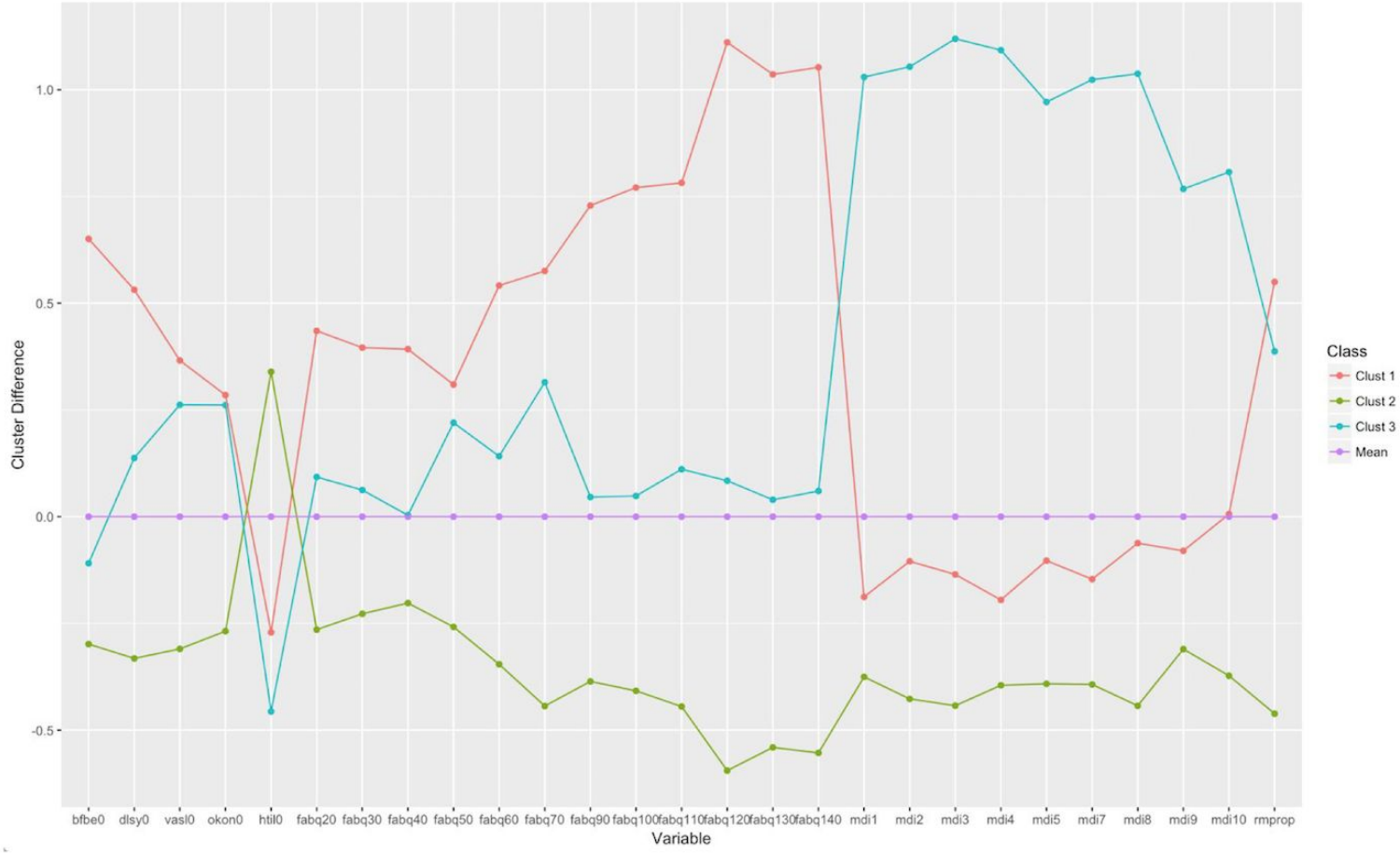

Figure 11: Cluster Difference Using Scaled Variables

New Observation Classification Guide

Table 4: New Observation Classification Guide

Fear-Avoidance Beliefs Questionaires (fabq's)	Major Depression Inventory (mdi's)	Roland Morris Summary Score (rmprop)	Back Pain Disability Risk Profile (start_risk)	Cluster Membership
Above Average		Above Average	Low - Medium	1
Below Average	Below Average	Below Average	Low - Medium	2
Above Average	Above Average		Medium - High	3

References

1. van Buuren, Stef, & Karin Groothuis-Oudshoorn. "mice: Multivariate Imputation by Chained Equations in R." Journal of Statistical Software [Online], 45.3 (2011): 1 - 67. Web. 6 Aug. 2017 https://www.jstatsoft.org/article/view/v045i03
2. Calinski, T., and J. Harabasz. "Communications in Statistics A dendrite method for cluster analysis." 27 Jun. 1974 https://www.researchgate.net/profile/Tadeusz_Calinski/publication/233096619_A_Dendrite_Method_for_Cluster_Analysis/links/555213e108aeaaff3befe29b/A-Dendrite-Method-for-Cluster-Analysis.pdf
3. G. Schwarz. "Estimating the Dimension of a Model - ResearchGate. https://www.researchgate.net/publication/38358303_Estimating_the_Dimension_of_a_Model
4. W. Rand. "Objective Criteria for the Evaluation of Clustering Methods - jstor." https://www.jstor.org/stable/2284239

R Packages

- mice::mice
- MASS::mca
- MixGHD::ARI and MixGHD::MGHD
- EMMIXskew::Emskew
- mixture::gpcm
- FactoMiner::MCA and FactoMiner::PCA
- factoextra::fviz_cluster
- fclust::FKM
- FPDclustering::PDclust