Statistical Learning 2020 Coursework 2 – Cluster Analysis
CFAS 420 – 19/20

May 12, 2020

Athanasios Kostaras/ ID: 35286525 Submission date: 2020-05-12 (Tue)

Contents

Introduction

For somebody using healthcare services, to be handled as an individual is an essential part of their whole experience during a tough period. Each patient undergoes healthcare uniquely and individually. To be identified and treated as an individual remains fundamental to a person when he/she becomes a patient. In this project, some qualities of life measures were used to identify distinct groups of respondents in a hospital. Clustering is a technique that can help us deal with these kind of problems. By applying four different methods; two distance-based clusterings (k-means and PAM), and two model-based clusterings(Gaussian mixture models and Latent Class models); the patients were classified in different groups. The results of this process can give to hospital information about the best way to treat each sufferer according to her/his needs.

1 Data Pre-Processing

First and foremost, missing values were identified. In this data-set, missing values were represented with -9. These prices were detected and deleted. Although the relationship variable was not used in the first steps of this project, rows that contained NA's about this column were deleted too, as later on this project this variable used. Hence, we did not want to have any bias by comparing models with different number of rows. Table 2, that can be found in the Appendix contains a description of the data and depicts that there are no variables where missing values were too many, compared with other variables. When the missing values were deleted the data-set contained 292 rows from 377. In this part of the project, the variables were treated as continuous. As we wanted to apply Principal Component Analysis, it was recommended to standardise the variables although the scales for this data-set was similar. A new data-set was constructed just for this procedure. After applying PCA, the plots of Figure 1 and 2 were extracted. It was evident that with 10 PC briefly 80 per cent of the variance was explained. With two of them, 50 per cent of the variance was explained. The biplot illustrates the vectors of the variables for the first two principal components. It was evident that all the vectors for the quality life measures had the same direction; on the left of the graph. Some of them were bigger according to the values of the coordinates. Moreover, some of them looking on the upper left and some on the down left. Consequently, we can assume that observations on the left part of the graph had higher values about life measures. Generally, the further to the left was a remark the bigger its life quality measure. Hence, observations on the right side of the graph corresponded to patients with low life measures and the observations on the centre of the graph, matched with those with average values. Moreover, it is observable that the biggest proportion of individuals was connected with average values of life measures, while few of them had high values.

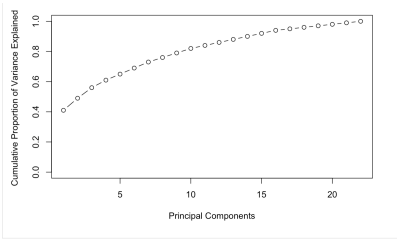

Figure 1: PC Cummulative Graph

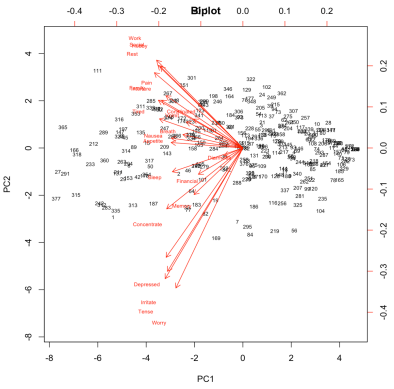

Figure 2: PCA Biplot

2 Distance Based Clustering

2.1 K-Means Clustering

For the two different distance based clustering models, we worked with the unscaled data of 292 rows and 22 columns. K-means clustering is the most frequently used unsupervised machine learning algorithm for partitioning a data set into a set of k groups, where k expresses the number of groups. The objective is that elements within the same cluster are as related as possible, whereas objects from diverse clusters are as different as possible. In K-Means clustering, each cluster is described by its centre. The basic concept of this algorithm is to define clusters in such a way that the total within-cluster variation to be minimised. We implemented one of many K-means algorithms, which uses the total within-cluster variation[1]. This measure was represented by the Euclidean distances between items and their centroids. Each observation is allocated to a cluster with the objective that the distance of the observation from its assigned cluster centres to be minimum. The total within-cluster variation was used to evaluate the optimum number of clusters for this process. This measure estimates the goodness of fit of the clustering. We want this metric to be as small as possible.

Before applying the K-Means clustering, let us explain some parameters that have been used(there is clear explanation for every model parameters in the R-Source Code). "iter. max" which was the number of times the algorithm would be reproduced the cluster designation and moving of centroids was defined to 300. "nstart" which was the number of times the initial starting points are re-sampled was set to 100. Furthermore, the procedure was implemented for a different number of groups.

Figure 3 illustrates the vision of clusters for a different number of groups. It was evident that 3 clusters seemed to be the best choice in order to separate observations with the best way. One thing that should be noticed in these graphs is the numbers in x and y-axis. When the "fviz_cluster" command identifies more than two dimensions (variables), it performs

automatically (PCA) and depicts the observations according to the first two principal components that describe the majority of the variance. Although this visualisation can give us a good intuition about the delineations of the data, it does not identify the optimal number of clusters. As to estimate this number, the elbow method was implemented. The aim was to minimise the total within-cluster sum of squares. Firstly the clustering algorithm was computed for a different number of clusters. Then the TWSS was computed, and the curve was plotted. The location of a bend (knee) in the curve is considered as an indicator of the proper amount of clusters. Figure 4, shows that the optimum number of clusters is 3, as the TWSS after this number diminishes very slowly.

Commenting on the 4 clusters model the below information was extracted. In this kind of partition, the population for the clusters observations was: 41, 96, 60 and 95. Hence, the second cluster was the group with the biggest amount of patients, while the first was this with the smaller one. The fourth and third cluster had 95 and 60 accordingly. Looking at the cluster means table, which can be found in the Appendix, it is observable that the fourth group of observations, on the right corner of the third graph of Figure 3, belongs to people with small prices for all the life measures. The number of cluster means for this group were the smaller, which meant that no one of the 22 life measures had high values for this group. Hence, the patients who belonged to this group were the healthier overall. This result agrees with the direction of vectors from the PCA graph, which also indicates something like that. Referring to the second cluster(centre and right), there were patients with mean values of life qualities. No variable has high values for this category; hence the most of the patients have values between 1 and 2 in the evaluation of life measures. In the third cluster(centre and left), some patients had measures bigger than the average but not the biggest. These patients should be treated with care as they looked less healthy than the previous ones. The first group(on the left) of people contained patients with biggest life qualities overall. We could assume that these were the patients that need to be cared the most overall, as they seemed the least healthy. All the variables had the biggest mean in this group. Again PCA had already informed us about these results. Figure 5 expresses precisely the same idea. Every line represents the mean for every variable in each cluster. The higher the values of y the less healthy the patient. Moreover, Figure 6 represents the population for each cluster.

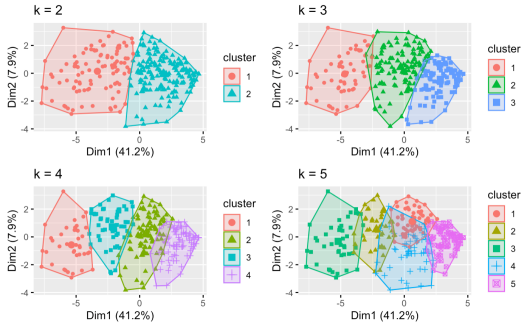

Figure 3: K-Means Clustering with different number of centres

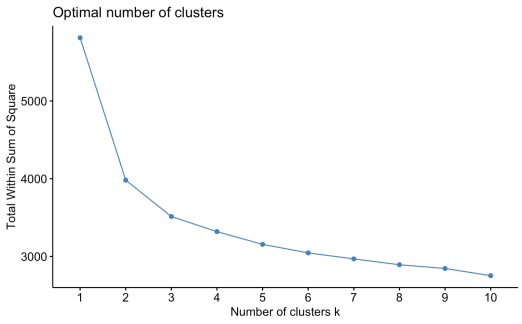

Figure 4: Elbow method for K-Means Clustering

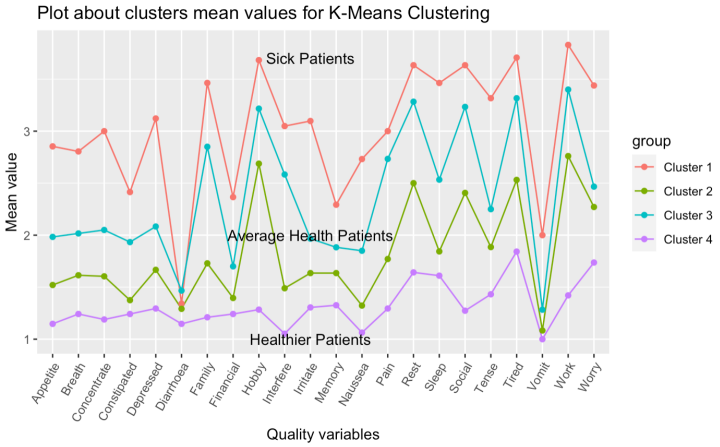

Figure 5: Plot about clusters mean values for K-Means Clustering

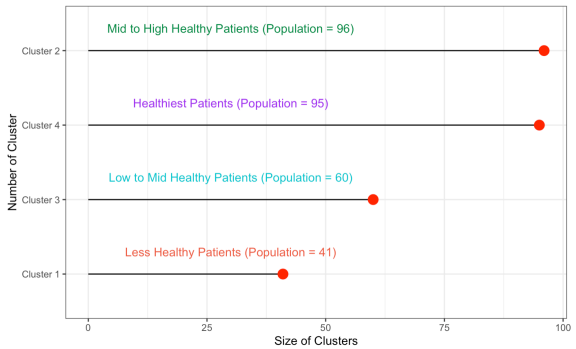

Figure 6: K-means Clusters Population

2.2 Partitioning Around Medoids (PAM) technique

A k-medoids algorithm is a clustering approach, where each cluster is represented by one data point in the cluster, called cluster medoids. A medoid is a point within a cluster for which average dissimilarity between it and all the other members of the cluster is minimal. It corresponds to the most central point in the cluster[2]. This way of clustering is more robust than k-means, where the mean value of all data points, is calculated as the centre of the cluster. In this project, the PAM technique applied. PAM, is a classic algorithm for k-medoids clustering. The goal in this way of clustering is to find representative points that reduce the sum of the differences of the observations to their closest centre. Two different measures of distance were used. The one was the "Euclidean distances", which are the root sum-of-squares of differences. The other was the "Manhattan distances", which are the sum of absolute distances. We chose to compare the algorithm with the euclidean as this distance measure was also applied for the k-means clusters.

The distribution of observations with this way was 51, 105, 85 and 51 from the first to fourth cluster respectively. Looking at the values of the means(Figure 14 in the Appendix) and Figure 8, we conclude that with PAM method, fewer observations(51) categorised in the cluster with the lowest means (on the right corner of Figure 7) compared to the k-means implementation. More observations classified in the two central clusters. The cluster that contains people with mean quality life scores, near the one with the lowest values, consisted of 85 measurements, instead of 96 for the k-means. The cluster with the patients with mid and high volumes had 105 instead of 60 for the k-means, while this one with the highest values had 51 compared to 41 for the k-means. Although the main idea was the same, the separation with this process was more ambiguous than k-means. This is because, in this instance, k-means dealt with bigger flexibility; since this approach estimated a centre according to the mean value instead of a mean point that medoids did. Although k-medoids is a more robust technique than k-means in this data-set, the values were actually discrete and not continuous; hence, the estimation about the centre was not the perfect one. As four clusters seemed not to be the ideal partition for this method, the optimum cluster value was estimated. To get this number, the Silhouette method was constructed. The purpose is to calculate the PAM algorithm using various values of clusters. The average silhouette estimates the quality of a clustering. The higher the average silhouette, the better the cluster fit. The optimal number of clusters is this one which maximises the average silhouette over a range of possible values. Figure 9 illustrates that the best choice is 2 clusters. Figure 8 represents the population for each cluster.

Taking into account both distance based clustering methods, we could assume that the best number of clusters was 3. This is a more reasonable decision as the vision of clusters looked significantly better, and the elbow method suggested this. However, the most significant was the measure values of patients. It seemed that if they would be separated into three categories, that is, those who had low, medium and high life quality rates, the distribution was better. The fourth category seemed redundant.

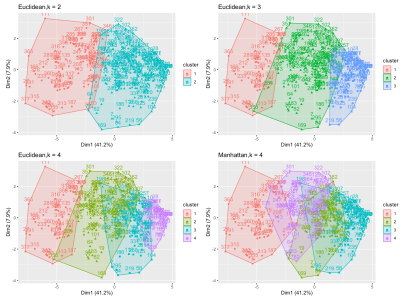

Figure 7: K-Medoids clustering with different number of clusters

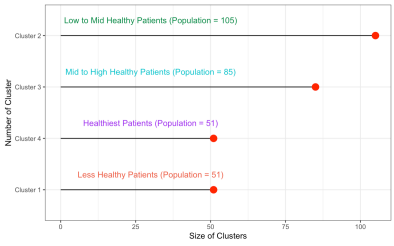

Figure 8: PAM Clusters Population

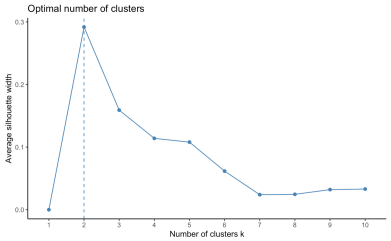

Figure 9: Silhouette method for K-Medoids clustering

3 Model Based Clustering

3.1 Gaussian mixture model (GMM)

Some more sophisticated techniques about clustering are model-based clustering. Using these techniques, and more specific mixture models, we tried to fit statistical models called mixture models to the data, to look for different distributions in our parameter space. Then, based on these distributions, the data is assigned to clusters and describe the joint distribution. Gaussian mixture models are a probabilistic model for describing normally distributed subpopulations within an overall group of observations.As there were several possible models, a model-selection concept was used to estimate the appropriate number of clusters. In this occasion, we take into account the BIC of the model and penalize concerning the number of model components. Again the continuous unscaled data-set was used. Figure 10 illustrates the BIC value in the y-axis. The number of clusters that the algorithm was tried to identify was designed in the x-axis. The BIC value is a measurement of how much variability is explained in the data. Thus the higher this value, the better the model is performing. Trying to find the model with the best BIC, it was apparent that the best model was the (VEE,2) model. This model had a BIC value of -14.518.08 and defined by two clusters. Looking at the mean values for each variable in every cluster (R-Source code) and Figure 11 and 12, it was evident that the first cluster in this procedure contained the 194 cases with higher values; hence the less healthy individuals. On the other side, the second cluster consisted of 98 more healthy patients. The VEE abbreviation refers to a model that has a variable volume, ellipsoidal distribution, equal shape and orientation. Looking at the covariance structure over the groups (in the source code as it was big enough due to many variables), it seems that this method does not work well with this data-set. The GMM has some assumptions about the form of the data. If those principles do not match, the performance might diminish significantly. GMM assumes an underlying Gaussian generative distribution. However, this data-set does not seem to satisfy this assumption, as the values were not really continuous and did not come from a Normal Distribution.

3.2 Latent Class Model

Latent Class Analysis (LCA) is a statistical model in which individual data points are classified into mutually exclusive (and exhaustive) types based on a set of categorical variables. LCA is a latent class model, similar to the GMM, except in this case the latent variable is discrete[3]. After importing the rest three variables(covariates) and encoding the 24 variables as factors (the age variable was not a factor but continuous), the procedure to build the best model based on several criteria has begun. In this part of Latent Class Model, the 3 covariates was not used yet. As in this process, the correct number of classes was not specified; different models were implemented and compared. For all the models, the maximum number of iterations

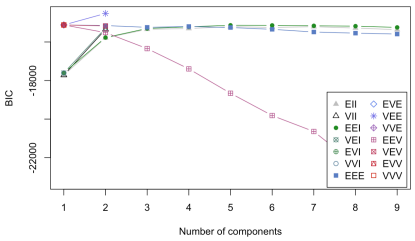

Figure 10: BIC graph

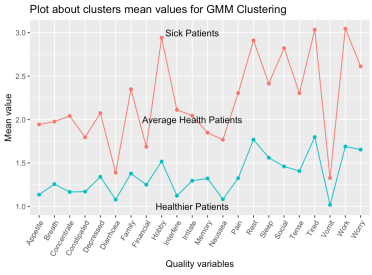

Figure 11: GMM Clustering mean values

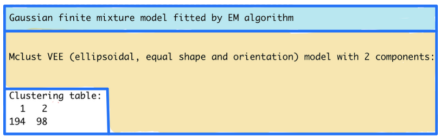

Figure 12: GMM fitted by EM algorithm

through which the estimation algorithm will cycle was set to 5000. The number of times to estimate the model, using different values of starting probabilities was set to 20. When setting "nrep>1," the search automates for the global, rather than just a local maximum of the log-likelihood function[4]. Firstly the model with two different classes was constructed. This two-class model can be interpreted like these. Patients in the first-class generally have low qualities of life variables and the inverse about the second class. Looking at Figure 15, it is evident that individuals from the first class, have a significant proportion of low values (1 and 2) in most of the categories; hence they were more healthy. On the other hand, people from the second category have bigger values, therefore this class represents the less healthier cases. The BIC value for this method was 13716.96. The population was 181(61.99%) for the first class and 111(38.01%) for the second; hence there were more patients with low-quality life measures in this partition. As it was wise to check the fit of more alternatives, the 3 class model was implemented. The interpretation here could be similar to the first one; however, we had another class with patients that had medium qualities of life variables. Something that should be noticed here was that looking at Figure 16; we observed that with this method, the classes were unordered. Therefore, we should be careful in the analysis of the model. As we understand that commenting on a Latent model with more than 3 unordered classes could be confusing for the reader, the results about the next models are summarised in Table 1 to be more understandable. During this procedure, class 1 represented the patients with low rates, class 2 symbolised the individuals with high values, while class 3 denotes the people with average rates. Individuals from the medium rate class were the most (127 - 43.50%). People from the second class were 74(25.34%), while the people in the first class with the lowest measures were 91(31.16%). The BIC value was 13494.85 , which was smaller than the previous model. Commonly, we compare the BIC value between the models to pick out the best one. Hence the second model was chosen. Lastly, as there was the need to check for a higher class model, the model with four classes was checked. The results of Table 3 illustrates that the BIC value for this model was bigger; therefore, the third class model was the best overall.

The next step was to try improving the results of the clusters by using some extra variables. As life quality rates were investigated for the patients without having some extra information, inserting some data about age, relationship and sex might give a better intuition on the data. Clustering was performed conditional on these variables. In this case, they work as covariates in the model, and we focus on constructing a clustered conditional distribution. When regressing onto covariates, the changes in the previous third-class model were insignificant. The BIC for this model was 13538.15, which is slightly bigger than the simple LCA model. Again attention should be paid with the interpretation of the classes as they were unordered. The first class during this procedure consisted of 72 patients(24.66%) and was the group of people with high life measures(less healthy). The second class contained cases with low rates(the healthiest patients). The quantity of this group was 91(31.16%). The third class, which was the class with the biggest proportion of cases (44.18%- 129 people), included the individuals with average quality life measures(average healthy cases). The results were similar to the previous 3 class model. Attention should be paid in the analysis of the coefficients presented in Table 4. In the first line of this table presented the coefficients according to the second and first class. The results showed that the three covariates; age, sex and relationship did not seem to to have a severe effect on people with different rates(low vs high). The p-values of all the covariates were insignificant, for this case. The same conclusion can be drawn comparing the second line of this table. In this procedure classes, 3 and 1 compared. Again it seemed there was no relation between covariates and the patients from high and average life measures. Hence, there was no evidence to reject the Null Hypothesis for both cases; that there was no relationship between the covariates and the classes separation.

Model	Healthy Patients	Average Healthy Patients	Less Healthy Patients
3-Class	91 (31.16%)	127 (43.50%)	74 (25.34%)
3-Class with Covs.	91 (31.16%)	129 (44.18%)	72 (24.66%)

Table 1: Separation of Classes with different Latent Models

4 Conclusion

After applying four different clustering techniques, we conclude that two of them(k-means and Latent Class model) worked reasonably good. The best specification of clusters it seemed to be three clusters, while four appeared to be redundant. The other two processes(PAM and GMM) did not work appropriately with this data-set, as PAM works right with really continuous data, and GMM mainly works with data that came from the normal distribution. Taking all the above into consideration the separation of patients to three groups will be a perfect choice; those with low, average and high values of life qualities.

5 Appendix

Life Measures	1	2	3	4	NA
Work - work limitations	58	82	69	83	15
Hobby - hobby limitations	77	74	65	76	20
Breath - short of breath	163	69	32	28	24
Pain - any pain	113	97	55	27	13
Rest - needed to rest	41	106	91	54	15
Sleep - trouble sleeping	93	108	49	42	15
Appetite - lacked appetite	163	79	30	20	17
Naussea - nausea	193	60	18	21	15
Vomit - vomited	249	27	9	7	17
Constipated - constipated	192	46	35	19	17
Diarrhoea - diarrhoea	233	42	9	8	19
Tired - tired	37	102	84	69	18
Interfere - pain interfering with daily life	160	62	41	29	22
Concentrate - difficulty concentrating	149	85	38	20	19
Tense - felt tense	97	121	47	27	16
Worry - been worried	60	131	54	47	18
Irritate - felt irritable	136	97	40	19	16
Depressed - felt depressed	122	117	32	21	18
Memory - memory difficulties	151	97	31	13	14
Family - Illness or medication interfered with family life	121	82	47	42	21
Social - Illness or medication interfered with social activities	84	76	69	63	17
Financial - Illness or medication caused financial difficulties	200	46	25	21	16

Table 2: ***Description of the data.***

```
K-means clustering with 4 clusters of sizes 41, 96, 60, 95

Cluster means:
       Work    Hobby   Breath     Pain     Rest    Sleep Appetite   Nausea    Vomit Constipated
1 3.829268 3.682927 2.804878 3.000000 3.634146 3.463415 2.853659 2.731707 2.000000    2.414634
2 2.760417 2.687500 1.614583 1.770833 2.500000 1.843750 1.520833 1.322917 1.083333    1.375000
3 3.400000 3.216667 2.016667 2.733333 3.283333 2.533333 1.983333 1.850000 1.283333    1.933333
4 1.421053 1.284211 1.242105 1.294737 1.642105 1.610526 1.147368 1.063158 1.000000    1.242105
  Diarrhoea     Tired Interfere Concentrate    Tense    Worry Irritate Depressed   Memory
1  1.341463 3.707317  3.048780    3.000000 3.317073 3.439024 3.097561  3.121951 2.292683
2  1.291667 2.531250  1.489583    1.604167 1.885417 2.270833 1.635417  1.666667 1.635417
3  1.466667 3.316667  2.583333    2.050000 2.250000 2.466667 1.966667  2.083333 1.883333
4  1.147368 1.842105  1.052632    1.189474 1.431579 1.736842 1.305263  1.294737 1.326316
    Family   Social Financial
1 3.463415 3.634146  2.365854
2 1.729167 2.406250  1.395833
3 2.850000 3.233333  1.700000
4 1.210526 1.273684  1.242105
```

Figure 13: Metrics about 4 clusters K-Means

```
Medoids:
     ID Work Hobby Breath Pain Rest Sleep Appetite Nausea Vomit Constipated Diarrhoea Tired
212 168    4     4      2    4    4     3         2      3     3            3         1     4
132 106    3     3      2    1    3     2         1      1     1            1         1     3
350 273    2     2      1    2    2     1         1      1     1            1         1     2
357 278    1     1      1    1    1     1         1      1     1            1         1     1
    Interfere Concentrate Tense Worry Irritate Depressed Memory Family Social Financial
212         4           2     3     4        3         4      2      4      4         2
132         1           2     2     2        2         2      2      2      3         2
350         1           1     2     2        2         2      1      2      2         1
357         1           1     1     1        1         1      1      1      1         1
```

Figure 14: Means about 4 clusters K-Medoids

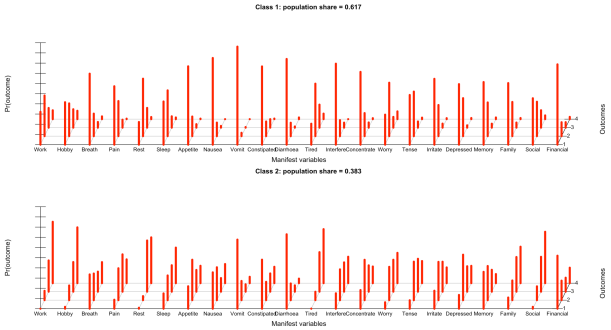

Figure 15: Latent 2 Class Model

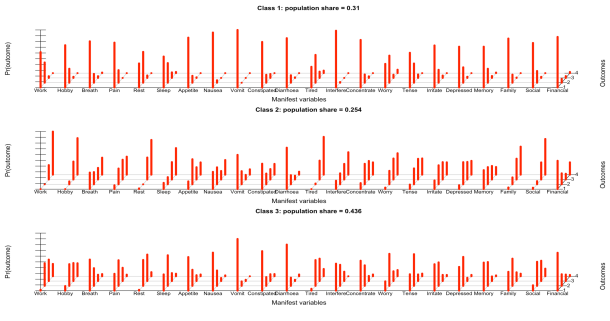

Figure 16: Latent 3 Class Model

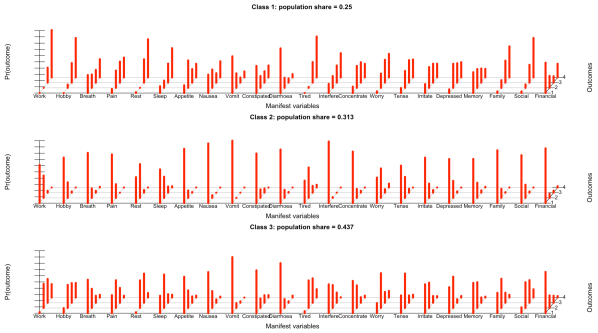

Figure 17: 3 Class Latent Model with Covariates

Model	k=1	k=2	k=3	k=4	k=3 + covariates
BIC	15142.89	13716.96	13494.85	13633.95	13538.15

Table 3: BIC values for Latent Models

	Intercept	Sex 2	Age	Relationship 2	Relationship 3	Relationship 4
Class 2/1	0.679	0.768	0.827	0.652	0.929	0.848
Class 3/1	0.628	0.589	0.999	0.565	0.214	0.643

Table 4: p-values about Latent Model with covariates

References

[1] Brad Boehmke and Brandon M Greenwell. *Hands-On Machine Learning with R*. CRC Press, 2019.

[2] Alboukadel Kassambara. *Practical guide to cluster analysis in R: Unsupervised machine learning*. Vol. 1. STHDA, 2017.

[3] Alex Gibberd. *Statistical Learning CFAS420*. Vol. (Part II) Clustering and Unsupervised Learning. Lancaster University, 2020.

[4] *https://www.rdocumentation.org/packages/poLCA/versions/1.4.1/topics/poLCA*.