Analysing Patient Data using Clustering Methods

CFAS420 Statistical Learning II

Harry Baines (35315878)

1 Introduction and Background

Clustering methods can be used to facilitate the understanding of a dataset. In this report we apply both distance-based and model-based clustering algorithms to the data to aid in understanding of the dataset, such that we can identify distinct groups of observations to gain insight into the issues facing hospital patients based on a set of quality of life variables. The data used in this report consists of 22 quality of life variables, measured on a 5-point likert scale, including an additional 3 variables collected from 377 hospital patients (see R code for explanations of these variables). The primary objective of this report is to extract insights from this data using cluster analysis to determine if distinct groups of respondents can be found. An interpretation of the clusters formed will be given to aid understanding of the issues facing the hospitals patients. We hypothesise that there exists at least 2 distinct sets of respondents, with one set representing those patients who experience more severe factors affecting their qualities of life compared to those experiencing more milder factors.

Extensive research has been undertaken in the area of analysing patient data using various clustering techniques, such as partitioning around medoids to compare symptom cluster phenotypes in patients with cancer and various kidney diseases [1] and using k-means clustering to cluster healthcare data [2]. We utilise both of these techniques in this report, as well as model-based clustering techniques.

2 Data Preprocessing and Exploratory Data Analysis

Firstly we identify any observations containing missing values. In this dataset, a value of -9 indicates the presence of a missing value for a given variable. 85 observations were found to have missing values, and following the removal of these, the resulting dataset consisted of 292 observations. One could impute the missing values which would enable a cluster analysis to be carried out on a larger set of data. Following removal of the missing values, the 22 quality of life variables were extracted into a separate data frame. Following Exploratory Data Analysis (EDA), the pairs of variables Interfere and Pain, Tired and Rest, and Hobby and Work were identified as strongly correlated variable pairs. The variables Interfere and Pain are similar in nature and could be combined or one of them could be removed. In our analysis however we keep these and apply clustering analysis to all quality of life variables.

3 Distance-Based Clustering

In this section we carry out a distance-based clustering analysis for the quality of life variables. Clustering methods are inherently unsupervised in contradistinction to supervised techniques, in which we are provided with a set of 'ground truth' labels. These clustering algorithms are based on the idea of separating observations into a set of clusters using commonly known distance metrics to quantify dissimilarity between them, such that they display different statistical characteristics, which facilitates the identification of patterns in the data.

3.1 K-Means Clustering

K-means is an iterative clustering algorithm and is commonly used due to its simplicity. It uses a cost function based on the squared Euclidean distance metric (ℓ_2 norm) which minimises the total within cluster variation for a set of K clusters $C_1, ..., C_K$. Initially, K data points are selected as the starting centroid locations. Data points are assigned to one of the K centroids based on the Euclidean distance (cluster assignment). The centroid locations are recalculated based on the mean of the data points assigned to that centroid's cluster (centroid update). The cluster assignment and centroid update steps are repeated until convergence is achieved (i.e. the centroid locations no longer change) and the sum of distances from data points to centroids is minimal.

We begin our analysis with $K = 4$ and apply the algorithm to the quality of life variables treating each as continuous. Due to the random initial assignment of the cluster centroids, it is recommended to repeat the k-means procedure multiple times, as results obtained on successive runs may be different and sub-optimal. The procedure was run 1000 times with 1000 random starting values to ensure a strong solution was obtained from different starting configurations of the centroids. When $K = 4$, the best total within-cluster sums of squares obtained was 3319.53 (2.d.p.).

Given our dataset is 22-dimensional, we utilise principal component analysis to reduce dimensionality of the data following k-means to facilitate visualisation of the clustered observations (see Appendix A). Cluster 1 showed the most separation from all other clusters. Cluster 2 showed a slight overlap with cluster 3, with the latter overlapping with cluster 4. Despite a reasonable separation of observations into 4 clusters, the overlap could suggest this number can be reduced to obtain a more optimal separation with less cluster overlap. Due to the interpretability constraints of the previous plot, we provide a further visualisation to help interpret the distribution of the quality of life variables with respect to the means computed for each of the generated clusters:

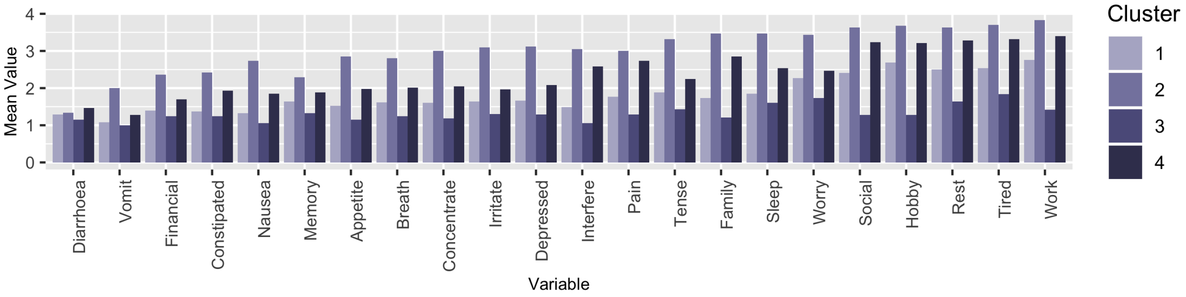


Figure 1: Bar plot of mean trajectories for each of the 22 quality of life variables.

Figure 1 was constructed by computing the mean values for all variables, grouped by their allocated cluster number following k-means. We notice nearly all quality of life variables have relatively low and stable mean values for cluster 3. This highlights a potential group of patients whose qualities of life are fairly good (i.e. near consistent scores of 1 on the likert scale across the variables indicate they don't have these qualities at all). Cluster's 2 and 4 showed a high variance in the mean values across the variables, with cluster 1 also showing some variation across the variables. Patients at the upper end of the spectrum (e.g. variables such as tired, have work limitations or needed to rest) in both cluster's 2 and 4 belong in groups displaying different statistical characteristics to those belonging in the lower spectrum (e.g. variables such as diarrhoea, vomiting and financial). This is because there exists a notable difference in their mean values, with a larger mean value on the upper end compared to the lower end. In essence, we notice that the factors affecting the quality of life of an individual is dependent on the severity of that factor with regards to the patients quality of life (i.e. there exists a group of patients who have more severe medical conditions compared to those who are busy or tired). Despite the clear separation of observations into distinct groups, we cannot conclude however that 4 clusters is the optimal value of K to partition this data.

3.2 PAM Clustering

Unlike in k-means where cluster centroids are summarised by their mean, partitioning around medoids (PAM) uses actual data points (medoids). PAM is based on the Manhattan distance metric (ℓ_1 norm), and works by iteratively assigning data points to their closest medoid to construct clusters. The algorithm aims to minimise the sum of dissimilarities between the data points in a cluster. PAM is a more robust alternative to k-means, and is less sensitive to noise and outliers as it uses medoids as cluster centers instead of mean values.

Similarly to our analysis for k-means, we can produce a 2-dimensional plot to visualise the generated clusters using PCA (see Appendix A). We find that PAM produces a result similar to k-means, with more cluster overlap observed with PAM compared to k-means. This strongly suggests the number of clusters K can be reduced to obtain a more optimal cluster arrangement. For both k-means and PAM clustering methods, the value of K had to be specified manually. We can use standard metrics to assess cluster fit to the data. Increasing the value of K results in a more complex model as a new centroid has been introduced, hence we want to lower the objective function whilst ensuring our model is not too complex as we would risk overfitting. In our analysis we utilise 3 common methods to assess cluster fit using PAM for values of K ranging from 1 to 10:

1. The elbow method: utilises the total within cluster variation score. The optimal value of K is located at the elbow in the plot.

2. The silhouette method: based on the silhouette value, is a measure of how similar a data point is to its own cluster compared to the other clusters. We take an average silhouette value over all data points and attempt to maximise this value as a function of K. The value of K which yields the largest average silhouette width will represent the optimal number of clusters to use.

3. The Gap statistic: a more sophisticated technique which utilises re-sampling, and compares the total intra-cluster variation for different values of K with their expected values under the null reference distribution of the data (i.e. a distribution with no obvious clustering).

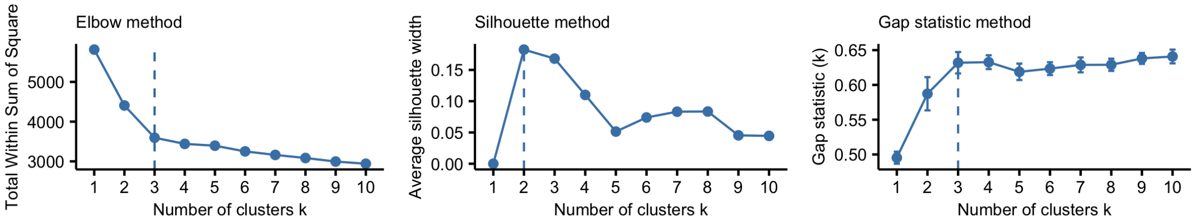


Figure 2: Methods for obtaining optimal value of K following PAM.

In Figure 2 we obtain $K = 3$ for the elbow method (the elbow in the first plot), $K = 2$ for the silhouette method (which yields the largest average silhouette width in the second plot) and $K = 3$ for the gap statistic method (which yields the largest gap statistic in the third plot). From these plots $K = 3$ should be chosen. Hence clustering the quality of life variables into 4 clusters, despite showing reasonable separation, is not the optimal number of clusters to partition the data into using PAM (see Appendix A for PAM clustering results using the optimal value of $K = 3$).

4 Model-Based Clustering

In this section we carry out a model-based clustering analysis for the quality of life variables. The clustering techniques studied in this section have a probabilistic foundation, that is, each observation is assigned a probability of belonging to a certain cluster. This is commonly known as soft clustering, which differs from hard clustering in which observations belong to a cluster completely or not. Mixture models can also account for clusters with varying shapes and can incorporate uncertainty in the estimates, hence have more flexibility than distance-based clustering methods.

4.1 Gaussian Mixture Models

A Gaussian Mixture Model (GMM) is based on a mixture of statistical distributions in which we try to model each sub-population of the data as its own distribution (known as a mixture component). In our analysis we fitted a GMM where each component was based on a Gaussian/Normal distribution. This distribution is parameterized by a mean and a variance (with the latter allowing us to capture more information compared to just the mean like in k-means). However this now involves computing means and variances for each cluster. The algorithm works in a similar iterative fashion to k-means, in which we assign random means representing the initial cluster centers, however we assign points to clusters based on the log-likelihood. Cluster centers are re-computed using the mean of observations in that cluster and observations are re-clustered to obtain a new arrangement.

We continue with our hypothesis that the data can be partitioned into 4 clusters, hence we specify 4 mixture components (clusters). We can visualise the fitted model in 2 dimensions using the top 2 principal components explaining the largest variance following PCA (see Appendix B).

Geometric features of each cluster are determined by the covariance matrix which has different parameterizations. To obtain the optimal number of mixture components we carry out model selection on the GMM using the BIC metric, and penalize w.r.t. the number of model components (and component complexity). This metric, based on a penalised form of the log-likelihood, can help to identify the optimal number of components to use. As the likelihood increases with more components, a penalty for the number of estimated parameters is subtracted from the log-likelihood. We compute a matrix of BIC values for all covariance structures up to 9 mixture components (see *?mclustModelNames* in R for all model names). Each identifier corresponds to the volume, shape and orientation respectively in that order, which allows different shapes of clusters to be recognised. Plotting the BIC against the number of components for different model types, 2 components yield the maximum BIC and hence represents the optimal number of clusters to use, with the best model being VEE (see Appendix B for both the BIC plot and the optimal cluster arrangements for the data with this model). This model captures ellipsoidal shapes with varying volume and equal orientations and shapes. Some model types could not be estimated however, and contained NA values in the BIC matrix. This arises due to a singularity present in the covariance matrix, and can be avoided using Bayesian regularisation [3]. Given we used the Gaussian distribution for each of the mixture components which is inherently continuous in nature, and given the quality of life variables are ordinal variables and not continuous, the values for the variables are not normally distributed, hence a GMM may not be the most appropriate statistical distribution to model dataset.

4.2 Latent Class Models

Latent Class Analysis (LCA) is a powerful method to extract insights from a dataset, in which data points are clustered into mutually exlusive regions based on a set of categorical variables. LCA models are similar to the GMM, however here the latent variables are discrete instead of continuous. For our analysis we fit a multicategory latent class model to the data in which we treat each of the 22 quality of life variables as categorical. Latent class models are known to work well with likert scales, and they aid in interpretation of differences between the classes.

Following model selection, the value of K which yielded the smallest BIC of 13505.76 (2.d.p.) was 3, hence 3 clusters represents the optimal number of classes here.

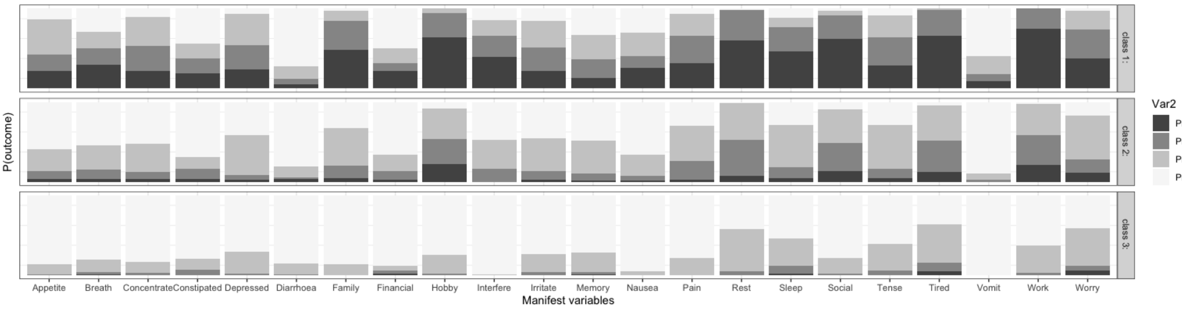


Figure 3: Estimated categorical distribution for the three classes.

We can infer from he distribution in Figure 3 that there are 3 groups of patients, each of which have different effects of the variables on their qualities of life. For example, patients whose quality of life is affected significantly (i.e. value of 4 on the scale) is observed in class 1, notably for variables such as Hobby, Rest, Sleep, Social, Tired and Work. Class 3 exhibits higher probabilities of patients who have more severe factors affecting their qualities of life, with the highest probabilities belonging to variables Vomit, Nausea and Interfere. This suggests a group of patients with potentially severe underlying medical conditions which affect their qualities of life and hence this group should be prioritised more in terms of medical treatment.

We can extend our latent class model to include further covariates we excluded from our clustering analysis: Age, Relationship and Sex (see Appendix C for plots). One can interpret the output of the fitted latent class model in relation to the multinomial distribution. The computed t-value enables us to assess the chance that a parameter of this size is obtained by chance. A small p-value would allow us to reject the null hypothesis, that is, the parameter is statistically significant at the given significance level. For 3 latent classes, the Sex covariate gave a p-value of 0.534, the Relationship covariate gave a p-value of 0.552, and the Age covariate gave a p-value of 0.999. Hence including these variables as covariates are not statistically significant.

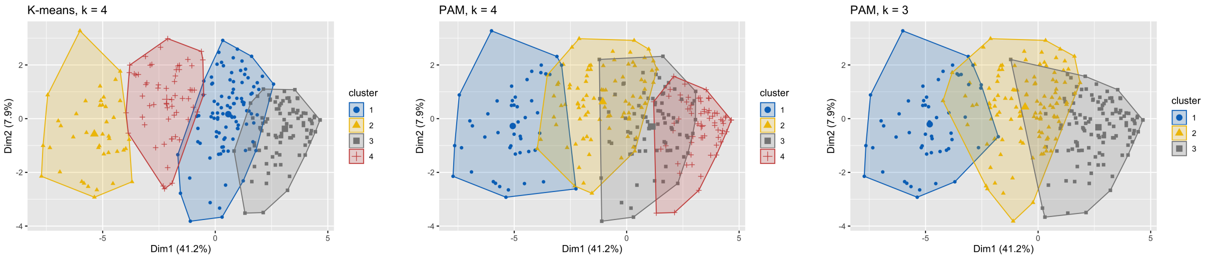
References

[1] M. Jhamb, K. Abdel-Kader, J. Yabes, Y. Wang, S. D. Weisbord, M. Unruh, and J. L. Steel, "Comparison of fatigue, pain, and depression in patients with advanced kidney disease and cancer-symptom burden and clusters," Mar 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30552961

[2] J. Samriya, "Efficient k-means clustering for healthcare data," pp. 2393–8390, 04 2016.

[3] L. Scrucca, M. Fop, T. B. Murphy, and A. E. Raftery, "mclust 5: Clustering, classification and density estimation using gaussian finite mixture models," Aug 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5096736/

A Distance-Based Clustering Visualisations (using PCA)

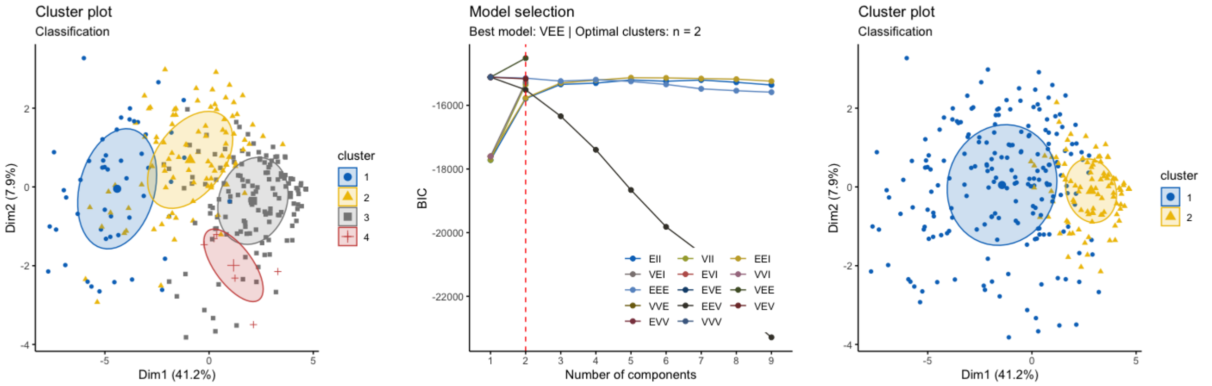


(a) K-means cluster allocations with $K = 4$.

(b) PAM cluster allocations with $K = 4$.

(c) PAM cluster allocations with optimal value of $K = 3$.

B Gaussian Mixture Model Visualisations

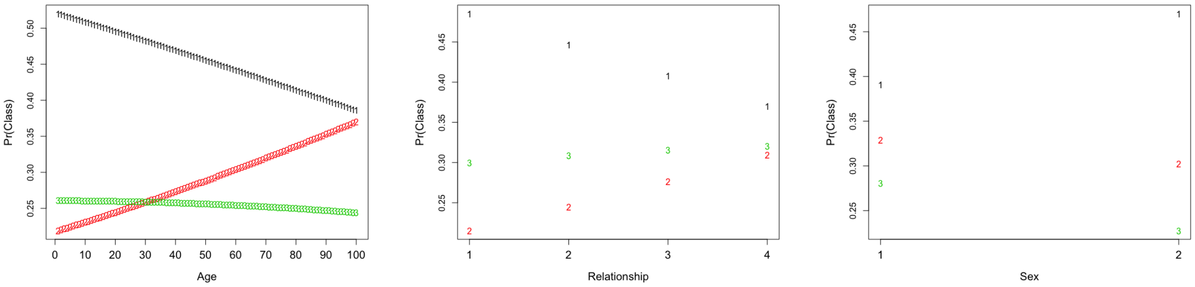


(a) GMM cluster allocations on 2 principal components with 4 mixture components.

(b) BIC values against number of components for different model types.

(c) GMM cluster allocations on 2 principal components with 2 mixture components.

C Latent Class Model Visualisations



(a) Class probs against age.

(b) Class probs against relationship.

(c) Class probs against sex.