Pairwise Overlap and Misclassification in Cluster Analysis

by

Birzhan Akynkozhayev

Supervisor: Igor Melnykov Second reader: Ayman Alzaatreh

Nazarbayev University

2015

Contents

1. Introduction

Separation of data into distinct groups is one of the most important tools of learning and means of obtaining valuable information from data. Cluster analysis studies the ways of distributing objects into groups with similar characteristics. Real-world examples of such applications are age separation of a population, loyalty grouping of customers, classification of living organisms into kingdoms, etc. In particular, cluster analysis is an important objective of data mining, which focuses on studying ways of extracting key information from data and converting it into some more understandable form. There is no single best algorithm for producing data partitions in cluster analysis, but many that perform well in various circumstances (Jain, 2008). Many popular clustering algorithms are based on an iterative partitioning method, where single items are moved step-by-step from one cluster to another based on optimization of some parameter. One of such algorithms, which will be mentioned in this paper is K-means algorithm, where data points are partitioned based on optimization of sum of squared distances within clusters (MacQueen, 1967). Another large class of algorithms are based on finite mixture model clustering methods. For example, stochastic emEMclustering method, which will also be covered in this article, is based on maximum likelihood estimation of statistical model parameters (Melnykov & Maitra).

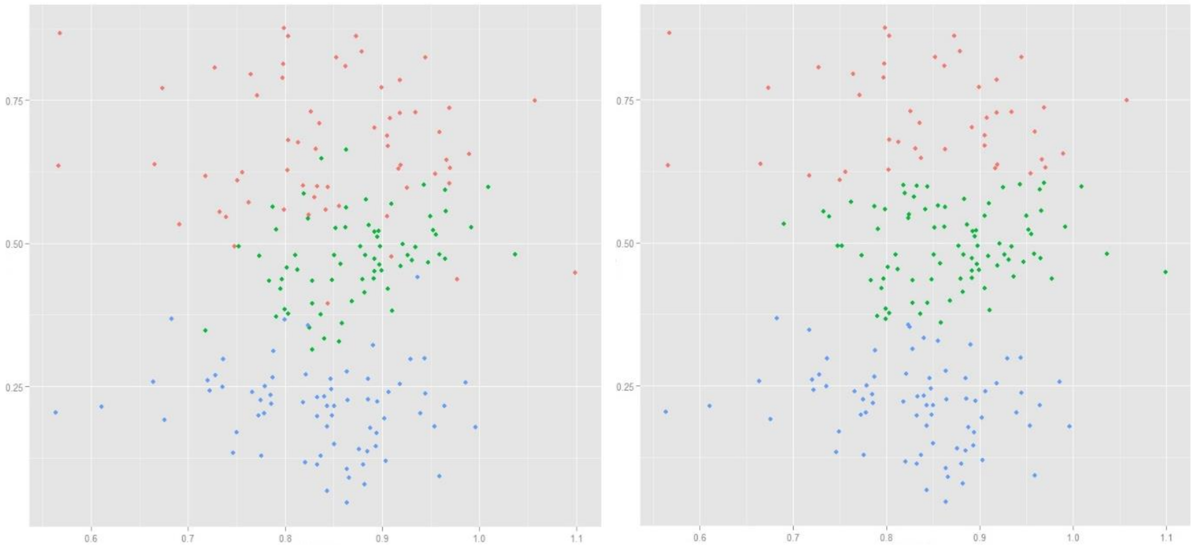


Figure 1. *a) True partition* *b) Partition obtained by K-means algorithm*

Misclassification of data is not a rare situation in cluster analysis. For instance, we can observe that several points have been misclassified on the previous figure (Figure 1) of true partition (a) versus the solution found by the K-means algorithm (b). Various factors lead to misclassification in clustering algorithms. The main goal of this paper is to analyze the effect of pairwise overlap, number of dimensions of data, and number of clusters on misclassification. The simplest case where misclassification can occur is when there are two clusters. The overlap is exact in this case, thus, we proceeded to use one of the simplest algorithms – K-means. At the higher number of clusters, when overlap is estimated, we considered more complex emEM algorithm.

2. The case of K = 2 clusters

Firstly, we began our investigation with the most simple scenario in which misclassification can occur, a case of two clusters. Most methods provide similar solutions in this case. Thus, we decided to use one of the simplest and fastest algorithms, which is K-means. Although K-means algorithm was published in 1950s and is 60 years old (Lloyd, 1957), it is still one of the most widely used and popular algorithms today. This algorithm aims at minimizing the following objective function:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

Here, $\left\| x_i^{(j)} - c_j \right\|$is the distance between cluster center c_j and data point $x_i^{(j)}$.

The K-means algorithm distributes N points in p dimensional space into K clusters, based on the minimization of sum of squared distances within clusters (MacQueen, 1967)

The algorithms consists of the following steps:

1) Pick K random points from the data set, these points will represent initial cluster centers.
2) Assign each point to the cluster with the closest center.
3) Calculate clusters' geometrical centers and assign them to be new centers.
4) Repeat Steps 2 and 3 until the centers are stabilized.

K-means algorithm is sensitive to the choice of initial centers. In some cases algorithm may not converge or one or more clusters can get dissolved. In that case, the algorithm is repeated with a different set of initial centers. The solution that produces the lowest value of objective function *J* is recorded as the best.

2.1 Mixture model

Other partition optimization algorithms rely on parametric methods, such as finite mixture model techniques. A mixture model is a statistical model, which specifies presence of subclasses in a data set, without identification to which subclass individual points belong (McLaughlan & Peel, 2000).

For independent identically distributed p-dimensional observations X_1, $X_2,\ldots,X_n$, the probability density function for mixture model with *K* components is

$$f(x;\pi) = \sum_{k=1}^{K} \pi_k f_k(x)$$

where f_k is the *k*th component and π_k is the probability that observation belongs to *k*th component ($\pi_k \geq 0, \sum_{k=1}^{K} \pi_k = 1$). Commonly, f_k is a normal (Gaussian) density $\varphi_k(x|\mu_k, \sum_k)$, where μ_k is the mean and $\sum_k$ is the covariance matrix.

$$\varphi_k(x|\mu_k, \textstyle\sum_k) = \frac{\exp\{-\frac{1}{2}(x_i - \mu_k)' \sum_k^{-1} (x_i - \mu_k)\}}{(2\pi)^{-\frac{p}{2}} |\sum_k|^{-\frac{1}{2}}}$$

Pairwise overlap is the measure of how much clusters penetrate each other. Pairwise overlap is the sum of misclassification probabilities $\omega_{i|j}$ and $\omega_{j|i}$ (Melnykov & Maitra, 2009)

$$\omega_{j|i} = \Pr[\pi_i \varphi(x;\mu_i,\Sigma_i) < \pi_j \varphi(x;\mu_j,\Sigma_j) \mid x \sim N_p(\mu_i,\Sigma_i)]$$

To analyze the degree of misclassification, we tried to fit different regression models to observe the behavior of misclassification probability relevant to overlap and number of dimensions of cluster data. Initially, we expected that misclassification would be higher for higher number of dimensions of data, since points can be close to each other in one dimension and be greatly separated in another.

2.2 Simulations

To generate data we used MixSim R package, which provides ways to generate multi-dimensional and multi-component Gaussian mixtures and specifies mean and maximum overlap between clusters in mixtures (Melnykov, Chen & Maitra, 2013). Datasets of sample size 1000 were generated using finite mixture model with Gaussian components for pre-specified level of maximum overlap between clusters. Covariance matrix structure was set to be spherical and the value of smallest mixing proportion was set to imply equal proportions. 1000 simulations of such datasets were used to obtain the median values of misclassification probabilities for each variation of overlap (ω) and number of dimensions (p). The following results (Table 1) were obtained for median values of misclassification proportions:

w/p	p=2	p=3	p=4	p=5	p=7	p=10
w=0.01	0,009	0.009	0.009	0.01	0.01	0.01
w=0.05	0.0360	0.0320	0.0335	0.0395	0.0345	0.0385
w=0.1	0.0600	0.0650	0.0660	0.0620	0.0650	0.0620
w=0.15	0.0945	0.0905	0.0940	0.0955	0.0935	0.0945
w=0.20	0.1175	0.1190	0.1195	0.1170	0.1210	0.1275
w=0.25	0.1405	0.1445	0.1430	0.1455	0.1385	0.1470
w=0.30	0.1665	0.1690	0.1720	0.1735	0.1760	0.1770

Table 1. Median misclassification probabilities for K=2.

Misclassification was measured by two parameters, adjusted Rand (AR) index and percentage of misclassification. AR index describes the agreement between two data partitions and attains value 1 when they agree perfectly. This index has the expected value equal to 0 when the data points are allocated to clusters completely at random (Hubert & Arabie, 1985). Since, labeling of the points as correctly or incorrectly classified is easy in the case of two clusters, we decided to use misclassification probability in that case. However, when the number of clusters is larger, the assignment of labels to the points is not trivial and calls for a more advanced measure, such as the adjusted Rand index.

3. Modeling misclassification probabilities

3.1 Logistic regression model

Initially, we tried to fit a logistic function to define the behavior of misclassification probability subject to overlap (ω) and number of dimensions (p). One of the reasons to choose the logistic model was the value of misclassification probability that ranges between 0 and 1, i.e. its value approaches the horizontal asymptotes at 0 and 1 (value 0 when there is no misclassification, value 1 when every point is misclassified). Assuming the failure or success of correctly identifying the case that a point belongs to a certain cluster as response variable $Y_{i,}$ one may consider Y as a Bernoulli random variable with parameter $\mathrm{E}\{Y\} = \tau$. $Y_{i,}$ can take value 0 with probability 1 - τ, which is a case of misclassifying a point, and value 1 with probability τ, i.e. correctly classifying that point. The expected value of Y_i is as follows:

$$E\{Y_i\} = \frac{e^{\beta_0+\beta_1\omega+\beta_2 p}}{1+e^{\beta_0+\beta_1\omega+\beta_2 p}}$$

The results show that misclassification probability of K-means algorithm depends on the overlap between clusters and that the number of dimensions p is not a significant predictor in the case of two clusters. Table 2).

```
Call:
glm(formula = misclass ~ w + p, family = "binomial")

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-0.137157  -0.051194   0.003566   0.049659   0.075608

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.622416   1.736253  -2.086   0.0369 *
w            7.211372   6.170064   1.169   0.2425
p            0.006436   0.203679   0.032   0.9748
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1.72902  on 41  degrees of freedom
Residual deviance: 0.18209  on 39  degrees of freedom
AIC: 14.191

Number of Fisher Scoring iterations: 6
```

Table 2. Summary of regression of misclassification percentage on degree of overlap (ω) and number of dimensions (p).

Therefore, we excluded the number of dimensions from the regression model and obtained the following result (Table 3) for the logistic model.

```
Call:
glm(formula = misclass ~ w, family = "binomial")

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-0.138254  -0.044873   0.004165   0.048136   0.073543

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.589      1.374  -2.613  0.00898 **
w              7.211      6.170   1.169  0.24250
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1.72902  on 41  degrees of freedom
Residual deviance: 0.18308  on 40  degrees of freedom
AIC: 12.191

Number of Fisher Scoring iterations: 6
```

Table 3. Summary of regression of misclassification percentage on degree of overlap (ω)

Thus, the analytic expression of the regression function is as follows:

$$\tau = \frac{e^{-3.589+7.211*\omega}}{1+e^{-3.589+7.211*\omega}}$$

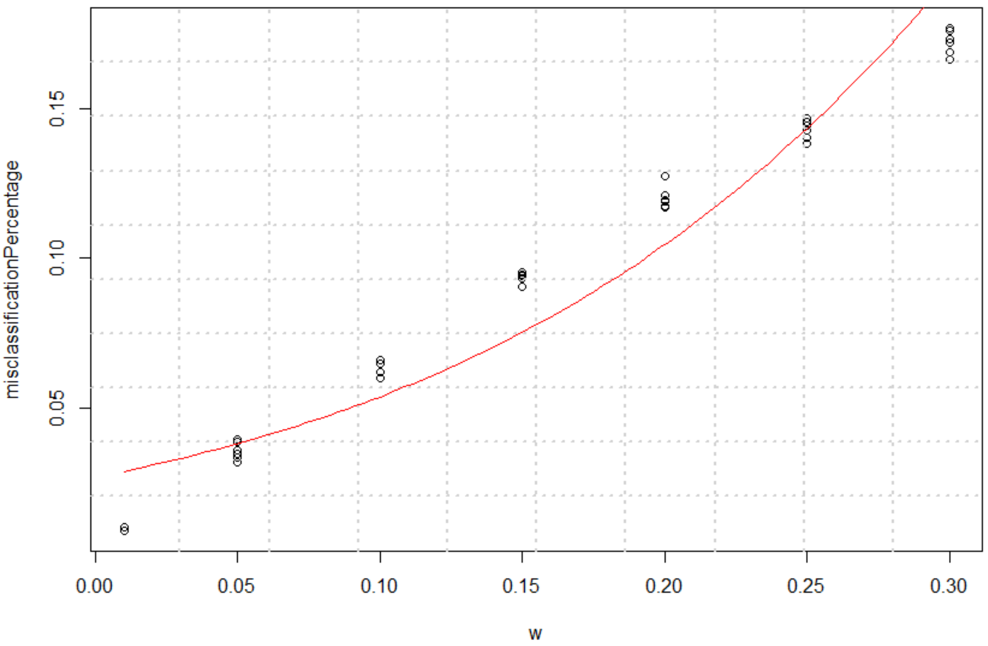


Figure 2. Observed values of misclassification probabilities and the graph of the fitted logistic model.

However, it can be observed from the graph in Figure 2 that the fit is quite poor; in addition, the chi-square goodness-of-fit test suggests that it is very unlikely (p-value < 0.001) that observed data comes from our logistic regression model. Thus, we had to search for a better fitting model.

3.2 Linear regression model

After denying logistic model, we considered the use of a linear model. Similarly, we tried to fit both predictor variables ω and p. The results are as follows:

```
Call:
lm(formula = misclass ~ w + p)

Residuals:
       Min         1Q     Median         3Q        Max 
-0.0082533 -0.0021750 -0.0002074  0.0020101  0.0069298 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.0045013  0.0013668   3.293  0.00211 ** 
w           0.5545801  0.0052282 106.075  < 2e-16 ***
p           0.0005153  0.0001916   2.690  0.01046 *  
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.003317 on 39 degrees of freedom
Multiple R-squared:  0.9965,   Adjusted R-squared:  0.9964 
F-statistic:  5630 on 2 and 39 DF,  p-value: < 2.2e-16
```

Table 4. Summary of linear regression of misclassification percentage on pairwise overlap and number of dimensions.

Again, number of dimensions parameter is not significant, thus we excluded it from further consideration:

$$\tau = 0.007 + 0.555 * \omega$$

```
Call:
lm(formula = misclass ~ w)

Residuals:
       Min         1Q     Median         3Q        Max 
-0.0073086 -0.0027094 -0.0003506  0.0028059  0.0094204 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.007164   0.001014   7.068 1.51e-08 ***
w           0.554580   0.005621  98.663  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.003566 on 40 degrees of freedom
Multiple R-squared:  0.9959,   Adjusted R-squared:  0.9958 
F-statistic:  9734 on 1 and 40 DF,  p-value: < 2.2e-16
```

Table 5. Summary of linear regression of misclassification percentage on pairwise overlap.

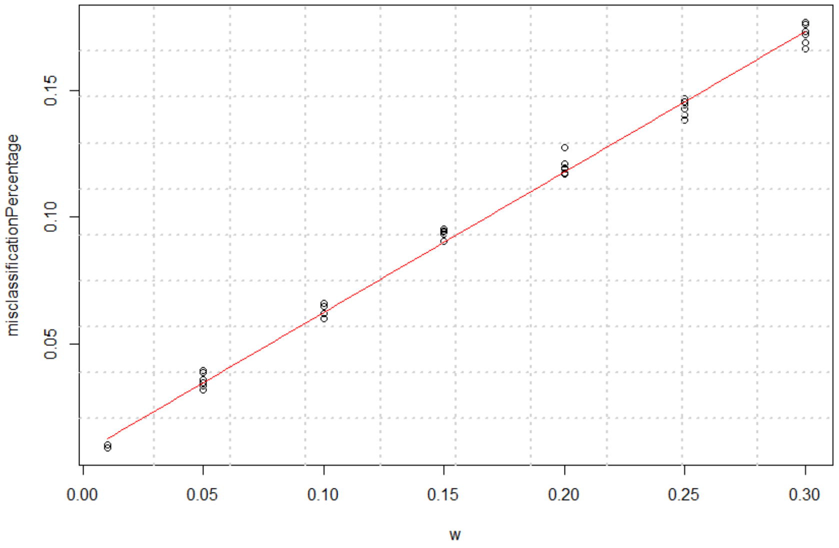


Figure 3. Observed values of misclassification probabilities and the graph of the fitted linear model.

We evaluated the assumptions of the model and in particular tested the residuals for normality.

Shapiro-Wilk test of normality for residuals yielded p-value = 0.7137.

Thus, the test of normality does not contradict the assumption that errors follow a normal distribution. Although the results for linear regression seemed to be reasonable enough, the following plot of residuals versus fitted values (Figure 4) is not evenly distributed and shows quadratic tendency. Therefore, we looked for a better model, and proceeded to the analysis of a quadratic model in ω.

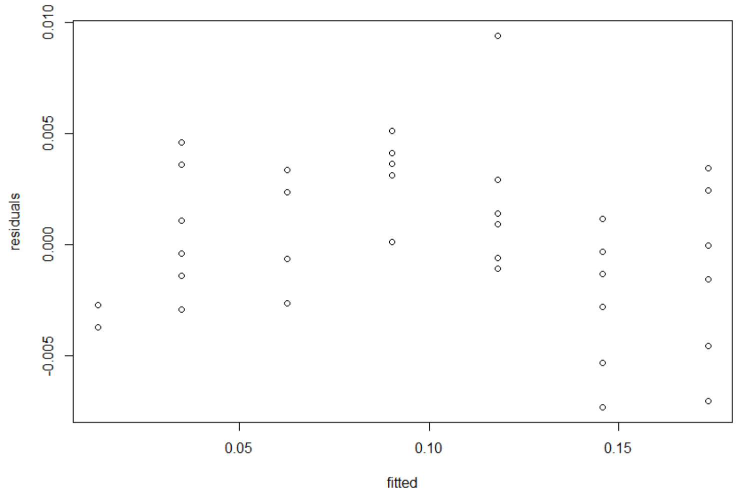


Figure 4. Residuals of linear model in ω versus the fitted values

3.3 Quadratic regression model

To obtain a better regression model, we further used a quadratic model to estimate the behavior of misclassification probability. The results of regression with both predictor variables ω and p:

```
Call:
lm(formula = misclass ~ w + I(w^2) + p + I(p^2))

Residuals:
       Min         1Q     Median         3Q        Max 
-0.0084293 -0.0012760 -0.0002441  0.0019262  0.0053670 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  9.742e-04  2.400e-03   0.406 0.687171    
w            6.232e-01  1.717e-02  36.304  < 2e-16 ***
I(w^2)      -2.228e-01  5.384e-02  -4.138 0.000194 ***
p            6.759e-04  8.424e-04   0.802 0.427484    
```

```
I(p^2)       -1.328e-05  6.833e-05  -0.194 0.847004    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.002815 on 37 degrees of freedom
Multiple R-squared:  0.9976,	Adjusted R-squared:  0.9974 
F-statistic:  3913 on 4 and 37 DF,  p-value: < 2.2e-16
```

Table 6. Summary of quadratic regression of misclassification percentage on pairwise overlap and number of dimensions.

Again, results show the insignificance of number of dimensions parameter. Therefore, it was removed from the model:

$$\tau = 0.004017 + 0.623199\omega - 0.222804\omega^2$$

```
Call:
lm(formula = misclass ~ w + I(w^2))

Residuals:
       Min         1Q     Median         3Q        Max 
-0.0073915 -0.0019689 -0.0002268  0.0014822  0.0077553 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.004017   0.001213   3.312 0.002002 ** 
w            0.623199   0.018862  33.040  < 2e-16 ***
I(w^2)      -0.222804   0.059164  -3.766 0.000548 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.003093 on 39 degrees of freedom
Multiple R-squared:  0.997,	Adjusted R-squared:  0.9968 
F-statistic:  6478 on 2 and 39 DF,  p-value: < 2.2e-16
```

Table 7. Summary of quadratic regression of misclassification percentage on pairwise overlap.

Shapiro-Wilk test of normality for residuals resulted in p-value = 0.8071.

The test does not contradict that residuals have a Gaussian distribution. Unlike the residuals in a linear model, the following plot (Figure 5) represents that residuals in quadratic model are more evenly distributed.

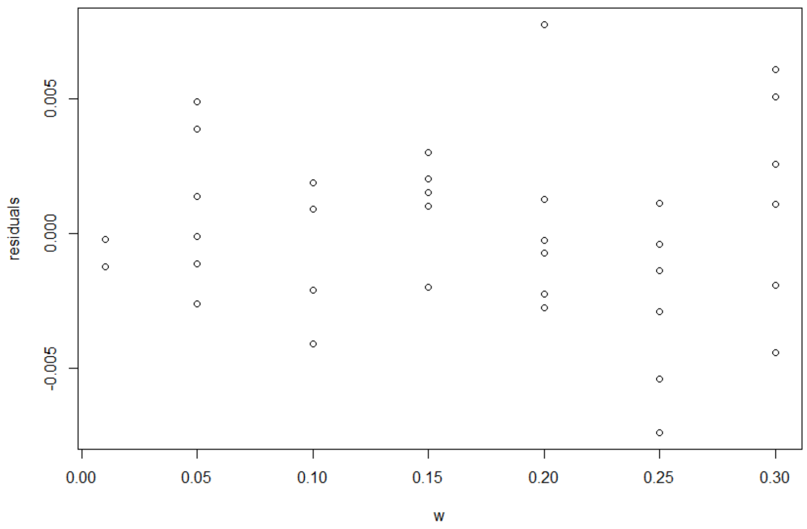


Figure 5. Residuals of quadratic model in ω versus the fitted values

We were satisfied with the results of quadratic regression, thus we refrained from considering other models and indicated quadratic model as our most successful fit. The following plot (Figure 6) represents observed values and fitted graph of quadratic model:

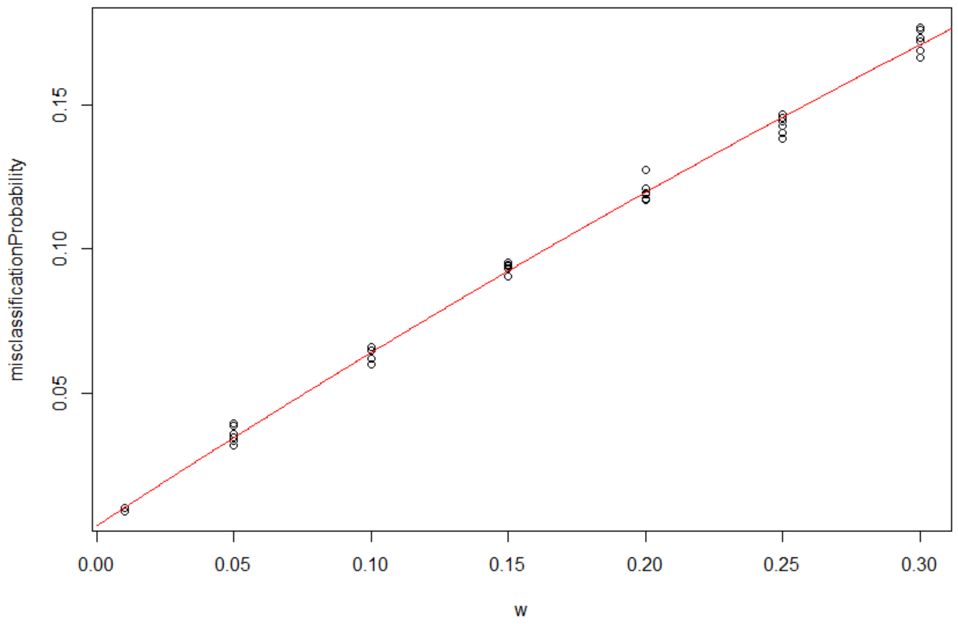


Figure 6. Observed values of misclassification probabilities and the graph of the fitted quadratic model.

4. The case of multiple clusters

4.1 Introduction to emEM Method

After, studying the misclassification for number of clusters K=2, we wanted to proceed further and study how misclassification occurs with a larger number of clusters. Though previous results of misclassification percentages, which were obtained for number of clusters K=2, seemed to be decent, with the increase in K, K-means algorithm rapidly deteriorate. For example, the simulation of 1000 runs of K-means algorithm with K=2 and maximum overlap ω=0.1 gives median adjusted Rand (AR) index of 0.7601456, however, the simulation with the same parameters except for K=5, shows median AR index of 0.350072. Thus, we sought for a better algorithm to obtain good solution for multiple clusters. Another popular clustering algorithm that we made an analysis on was Expectation-Maximization (EM) algorithm. The EM algorithm is a general statistical method of maximum likelihood estimation and in particular it can be used to perform clustering (Ordonez & Omiecinski, 2002).

The EM algorithm step by step improves a starting clustering model to better fit the data set and stops at a solution which is locally ideal or a saddle point (Bradley, Fayyad & Reina, 1998).

In analysis of EM algorithm, we continued to use Gaussian mixture as our choice for mixture model. The mixture density of which is the following:

$$f(x;\vartheta) = \sum_{k=1}^{K} \pi_k \varphi(x;\mu_k,\Sigma_k)$$

where μ_k is mean and $\sum_k$ is covariance matrix for k-th component.

The normal (Gaussian) density $\varphi_k(x|\mu_k,\sum_k)$ is:

$$\varphi_k(x|\mu_k,\textstyle\sum_k) = \frac{\exp\{-\frac{1}{2}(x_i-\mu_k)'\sum_k^{-1}(x_i-\mu_k)\}}{(2\pi)^{-\frac{p}{2}}|\sum_k|^{-\frac{1}{2}}}$$

According to V. Melnykov and R. Maitra (2009) the EM algorithm is carried out by assuming that there are missing data points, which together with the observed data compose “complete” data. The appropriate likelihood function is commonly easier to operate. Two steps, the expectation (E) and the maximization (M), compose the algorithm.

The M-step of s-th iteration aims to maximize the conditional loglikelihood function, called Q-function, with respect to parameter vector $\vartheta : Q(\vartheta; \vartheta^{(s-1)}, x_1, x_2, ..., x_n)$

The corresponding Q-function is given by:

$$Q(\vartheta; \vartheta^{(s-1)}, x_1, x_2, ..., x_n) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\pi_{ik}\{\log|\Sigma_k| + (x_i - \mu_k)' \Sigma_k{}^{-1}(x_i - \mu_k)\}$$
$$+\sum_{i=1}^{n}\sum_{k=1}^{K}\pi_{ik}\log\pi_k - \frac{pn}{2}\log 2\pi.$$

Given the current parameter ϑ^{s-1} estimates, E-step focuses on calculation of the following conditional probabilities:

$$\pi_{ik}^{(s)} = \text{Prob}\{X_i \in k^{th} cluster \mid X_i; \vartheta^{(s-1)}\} = \frac{\pi_k^{(s-1)} f_k(x_i; \vartheta_k^{(s-1)})}{\sum_{k'=1}^{K}\pi_{k'}^{(s-1)} f_k(x_i; \vartheta_{k'}^{(s-1)})}$$

Considering covariance matrix $\sum_k$ as a general unstructured dispersion matrix, the M-step gives the following solutions:

$$\pi_k^{(s)} = \frac{1}{n}\sum_{i=1}^{n}\pi_{ik}^{(s)}, \qquad \mu_k^{(s)} = \frac{\sum_{i=1}^{n}\pi_{ik}^{(s)} x_i}{\sum_{i=1}^{n}\pi_{ik}^{(s)}}$$

$$\Sigma_k^{(s)} = \frac{\sum_{i=1}^{n} \pi_{ik}^{(s)} (x_i - \mu_k^{(s)})(x_i - \mu_k^{(s)})'}{\sum_{i=1}^{n} \pi_{ik}^{(s)}}.$$

When the respective increase in likelihood function is not considerable, the algorithm is stopped.

4.2 Initialization of the algorithm

The likelihood function commonly can have multiple local maxima, thus the algorithm is very sensitive to the choice of starting points and good initialization is critical. There exist numerous initialization algorithms (Melnykov & Maitra, 2009). Since we chose the stochastic *em*EM method proposed by V. Melnykov and R. Maitra for clustering, the initialization was carried out as a part of the method. The *em*EM algorithm consists of two EM stages. First one, called short EM, runs EM algorithm with set of initial points chosen randomly. The solution, which produces the best log likelihood, is used afterwards as an initializer for the second long EM stage. The long EM runs until convergence criterion is met and final solution is obtained.

4.3 Simulations

The generation of datasets was similar to the simulations done for K-means algorithm. We used MixSim package, mentioned before and EMCluster R package, which provides ways for execution of EM algorithm. Datasets were again of size 1000 and generated using Gaussian finite mixture model with pre-specified levels of maximum overlap between clusters. The value of the smallest mixing proportion was set to imply equal proportions and covariance matrix structure was set to be non-spherical. Due to the complexity of algorithm, unlike for K-means we did 100 simulations of such datasets for obtaining values of AR index for each value of overlap (ω) and number of dimensions (p). Unlike the case of K=2 number of clusters, where we used misclassification percentage as a measure of misclassification, here we used AR index. The reason for the change was the difficulty in calculating misclassification percentages; thus, the following results (Table 6) were obtained for the median AR index:

w/p	p=2	p=4	p=5	p=7	p=10
w=0,001	0.8973324	0.8923708	0.888494	0.8823739	0.8775253
w=0,005	0.8812297	0.8804144	0.8643471	0.8446646	0.8191256
w=0,01	0.8700241	0.8507911	0.8190255	0.8005175	0.738382
w=0,03	0.7199319	0.6639654	0.6512065	0.5963805	0.5772604
w=0,04	0.6619233	0.6022783	0.590545	0.5137298	0.4294371
w=0,05	0.601245	0.5523285	0.5272748	0.4543913	0.3461704
w=0,07	0.512055	0.4411077	0.4112382	0.3361598	0.2577662
w=0,1	0.4084238	0.337837	0.3018632	0.2372028	0.1625142
w=0,15	0.3065079	0.2291579	0.1901517	0.1377378	0.0914074

Table 6.Medians of AR index values for K=5 number of clusters

5. Modeling AR Index

5.1 Logistic regression model

The value of AR index ranges between 0 and 1, thus first of all we tried to fit a logistic model. The summary of logistic regression (Table 8) shows that number of dimensions is not significant parameter.

$$AR = \frac{e^{2.1510-24.1801*\omega-0.1140*p}}{1+e^{2.1510-24.1801*\omega-0.1140*p}}$$

```
Call:
glm(formula = ARIndex ~ w + p, family = "binomial")

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.21221  -0.13066   0.01292   0.14064   0.38259

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.1510     0.9442   2.278  0.02272 *
w           -24.1801     8.7604  -2.760  0.00578 **
p            -0.1140     0.1263  -0.903  0.36651
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 12.1746  on 44  degrees of freedom
Residual deviance:  1.1413  on 42  degrees of freedom
AIC: 37.541

Number of Fisher Scoring iterations: 4
```

Table 8. Summary of regression of AR index on degree of overlap (ω) and number of dimensions (p).

Therefore, we excluded the number of dimensions from the regression model. However, it should be noted that p-value is larger in this case, thus the number of dimensions is less significant than in the case of K=2 clusters. The summary of regression (Table 9) is the following:

```
Call:
glm(formula = ARIndex ~ w, family = "binomial")

Deviance Residuals: 
     Min        1Q    Median        3Q       Max  
-0.45781  -0.08996   0.05744   0.18694   0.52454  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)   
(Intercept)   1.4820     0.5366   2.762  0.00575 **
w           -23.6985     8.6287  -2.746  0.00602 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 12.1746  on 44  degrees of freedom
Residual deviance:  1.9713  on 43  degrees of freedom
AIC: 38.199

Number of Fisher Scoring iterations: 4
```

Table 9. Summary of logistic regression of AR index on degree of overlap (ω)

The result of regression:

$$AR = \frac{e^{1.4820-23.6985*\omega}}{1+e^{1.4820-23.6985*\omega}}$$

The chi-square goodness of fit shows p-value < 0.001, suggesting that logistic model is a poor fit for observed data. Hence, we proceeded to consideration of a linear model.

5.2 Linear regression model

As the logistic model was unsatisfactory, a linear model in ω and p was tried.

The result for linear model was the following:

$$AR = 0.94101 - 4.90510 * \omega - 0.02181 * p$$

```
Call:
lm(formula = AR ~ w + p)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.13148 -0.05618 -0.02542  0.06124  0.15953 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.941005   0.029374  32.035  < 2e-16 ***
w           -4.905102   0.249856 -19.632  < 2e-16 ***
p           -0.021810   0.004255  -5.125 7.08e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07786 on 42 degrees of freedom
Multiple R-squared:  0.9074,   Adjusted R-squared:  0.903 
F-statistic: 205.8 on 2 and 42 DF,  p-value: < 2.2e-16
```

Table 10. Summary of linear regression of median AR index on pairwise overlap and number of dimensions

Due to the problems with curvature that we experienced before, the residuals were examined by means of the plot of residuals versus pairwise overlap (Figure 7). Once again it was clear that a quadratic term in ω needs to be added. We fitted a model with both quadratic terms (in ω and p) as well as the interaction term.

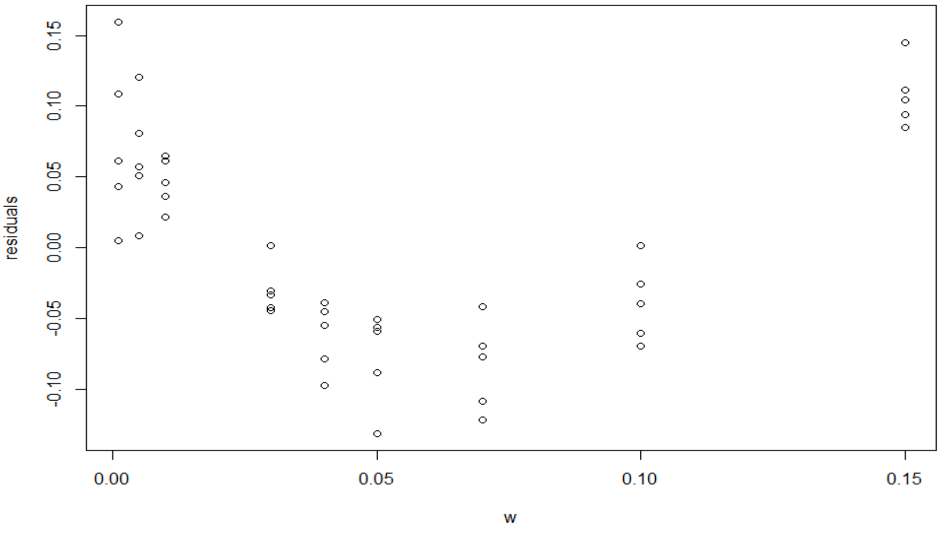


Figure 7. Residuals of linear model in ω versus the fitted values.

5.2 Second order regression model

Due to potential problems with high correlation among predictors and high order terms, the centering of predictor variables was performed. Those variables are ω_c and p_c, where

$\omega_{c_i} = \omega_i - \overline{\omega}$ and $\mathrm{p}_{c_i} = \mathrm{p}_i - \overline{\mathrm{p}}$.

The results of regression (Table 11):

```
Call:
lm(formula = AR ~ wc + I(wc^2) + pc + I(pc^2) + wc:pc)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.066103 -0.018914  0.004224  0.014751  0.052868 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.4984362  0.0070101  71.103  < 2e-16 ***
wc          -6.2489450  0.1145204 -54.566  < 2e-16 ***
I(wc^2)     32.1279955  1.8444322  17.419  < 2e-16 ***
pc          -0.0221540  0.0015478 -14.313  < 2e-16 ***
I(pc^2)      0.0003459  0.0005677   0.609    0.546    
wc:pc       -0.1496286  0.0310286  -4.822 2.19e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02637 on 39 degrees of freedom
Multiple R-squared:  0.9901,	Adjusted R-squared:  0.9889 
F-statistic:   783 on 5 and 39 DF,  p-value: < 2.2e-16
```

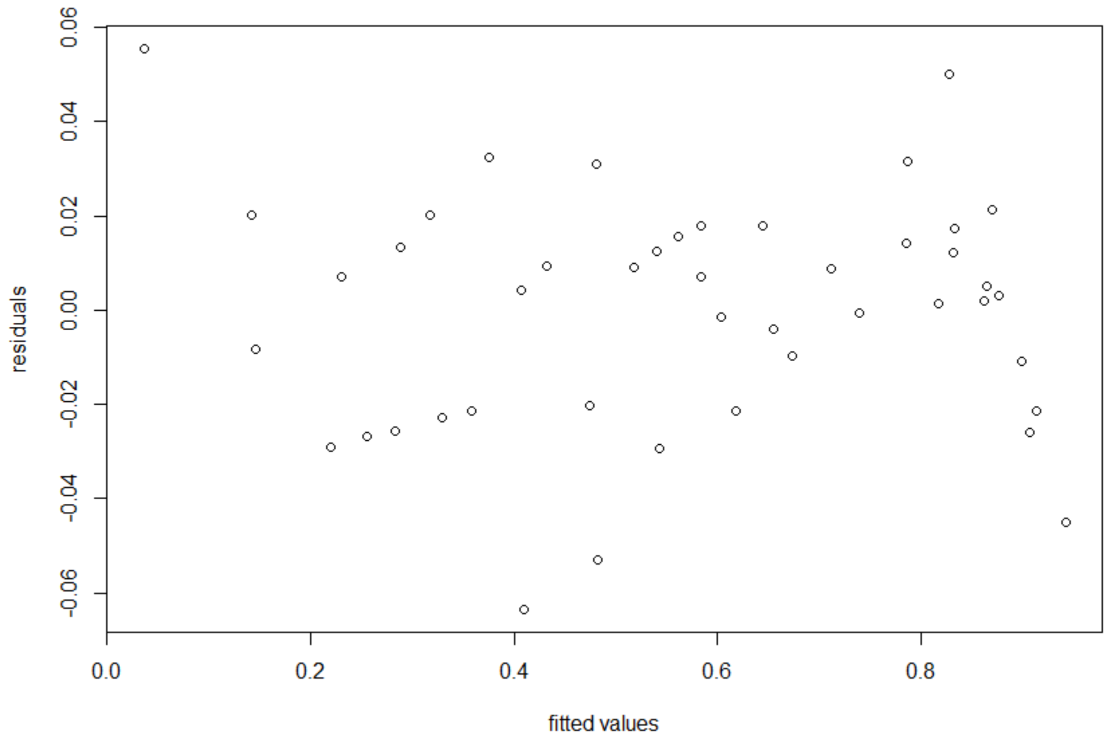
Table 11. Summary of second order regression of AR index on pairwise overlap and number of dimensions.

All terms except the quadratic term for the number of dimensions were significant. Thus, only that term had to be dropped and the model was refit:

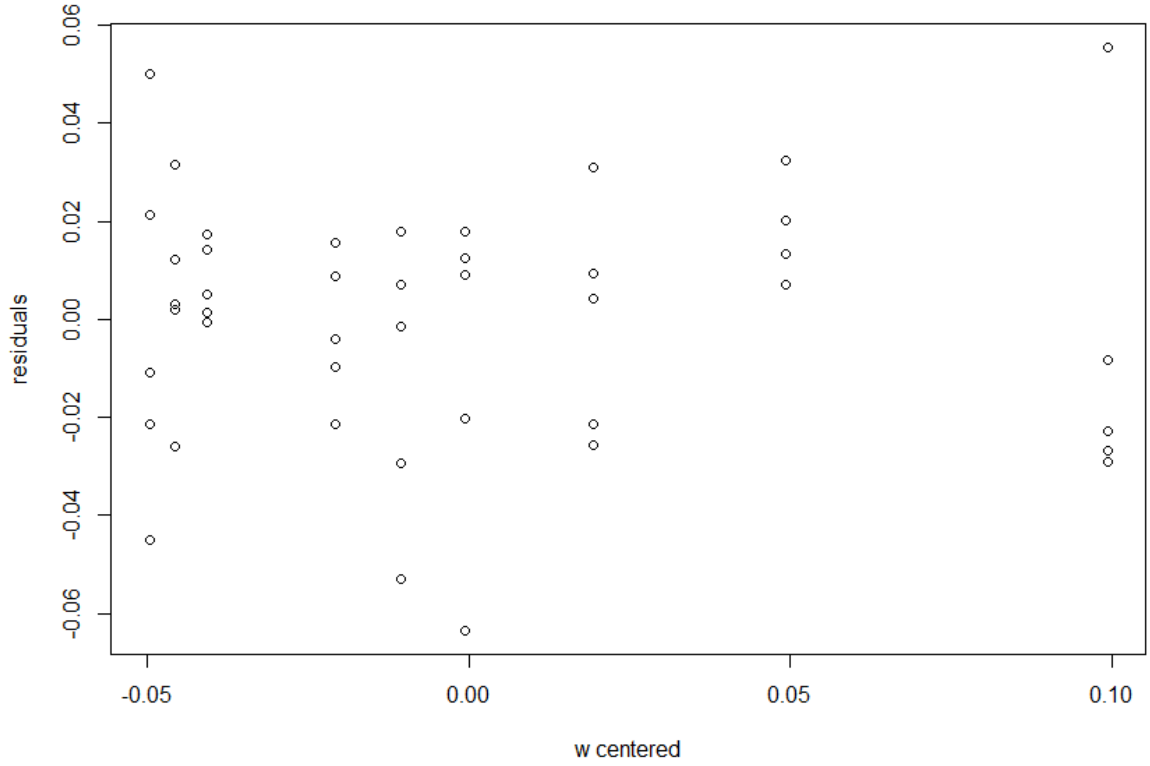
```
Call:
lm(formula = AR ~ wc + I(wc^2) + pc + wc:pc)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.063493 -0.021262  0.004211  0.015743  0.055479 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.501010   0.005551  90.263  < 2e-16 ***
wc          -6.248945   0.113617 -55.000  < 2e-16 ***
I(wc^2)     32.127996   1.829877  17.557  < 2e-16 ***
pc          -0.021810   0.001430 -15.252  < 2e-16 ***
wc:pc       -0.149629   0.030784  -4.861 1.84e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02617 on 40 degrees of freedom
Multiple R-squared:   0.99,	Adjusted R-squared:  0.989 
F-statistic: 994.3 on 4 and 40 DF,  p-value: < 2.2e-16
```

Table 12. The summary of the final second order regression of AR index on pairwise overlap and number of dimensions.

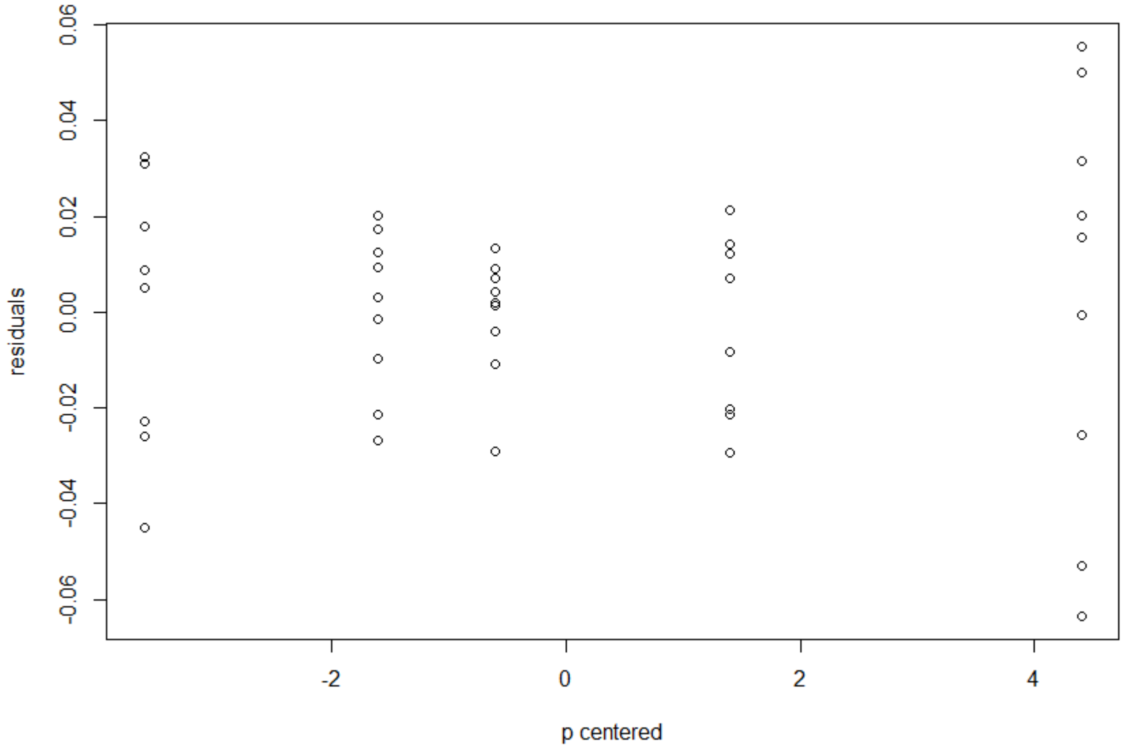
The plots of residuals versus both predictor variables as well as fitted values (Figure 8 a, b, c) were examined and did not reveal further problems.



a)



b)



c)

Figure 8. Residuals of second order model plotted against a) fitted values, b) ω_c *, c)* p_c

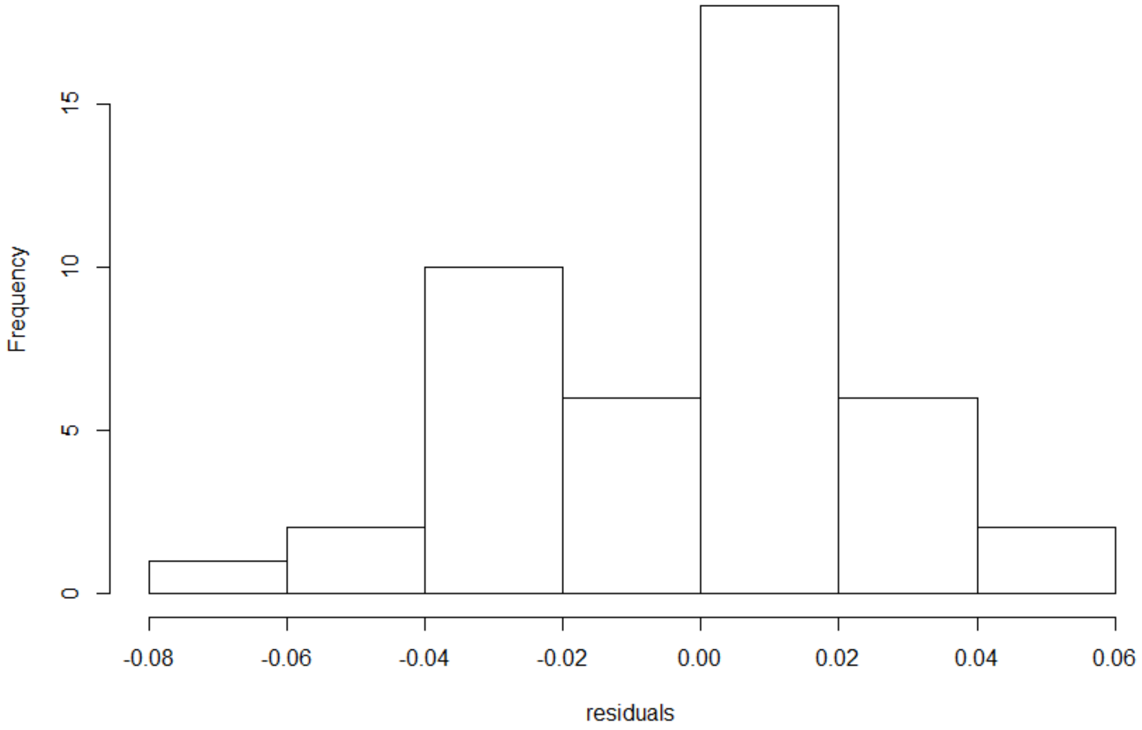


Figure 9. The histogram of residuals of the final model.

Shapiro-Wilk test of normality for residuals resulted in p-value = 0.4401. The assumption of a normal distribution of errors is upheld.

Rewriting the model in terms of non-centered variables we obtain the following equation:

$$AR = 0.979782 - 8.66668*\omega + 32.128 * \omega^2 - 0.0142288 * p - 0.149629*\omega*p$$

Outside of the scope of our model, an extrapolation would lead to inaccurate results. Thus, the model should be used for values of p between 2 and 10, ω between 0.001 and 0.15 and number of clusters K=5.

We can observe that both predictors negatively affect the values of AR index and in addition the interaction term between ω and p has a negative coefficient meaning that the adverse effect of the increase in number of dimensions and pairwise overlap on AR index is reinforced when both of them occur at the same time.

We stopped further model selection and chose the second order model to be our best fit to the observed data.

6. Discussion

This study analyzed the impact of several parameters on misclassification of data in cluster analysis. In particular, pairwise overlap and number of dimensions were studied as such predictors. Different algorithms were considered for different number of clusters: K-means algorithm in the case of two clusters and emEM method in the case when the number of clusters is greater than two. The results that were obtained showed that pairwise overlap was a significant factor in both cases. However, the number of dimensions was not a significant factor in the case of two clusters, but it was a considerable factor in the case of multiple clusters. In general, a higher number of dimensions or higher pairwise overlap mean that misclassifications will occur more frequently. Logistic, linear, and second order regression models were tried as possible approximations. Due to poor fitting characteristics, both the logistic and linear model were rejected. In the case of multiple clusters, regression shows that both overlap and number of dimensions are significant as well as interaction between them. In both cases, a second order regression model provided the best results.

References

Bradley, P.S., Fayyad U., Reina C. (1998). Scaling EM (Expectation-Maximization) Clustering to Large Databases. (Report No. MSR-TR-98-35). Redmond: Microsoft Research.

Hubert, L. & Arabie, P. (1985). Comparing partitions. *Journal of Classification.* 193–218. doi: 10.1007/BF01908075

Jain, A.K. (2008). Data Clustering: 50 Years Beyond K-Means. *Pattern Recognition Letters*, Vol. 31, No. 8, pp. 651-666, 2010.

Lloyd, S.P. (1982). Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*. Vol. it-28, no.2, 129-137

MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability . University of California Press. pp. 281–297.

McLachlan, G. & Peel, D. (2000). Finite mixture models. New York: Wiley.

Melnykov, V. & Maitra, R. (2009). Finite mixture models and model-based clustering. *Statistics Surveys.* Vol. 4 (2010) 80–116 ISSN: 1935-7516. DOI: 10.1214/09-SS053

Melnykov, V., Maitra, R., Chen, W. (2013). Simulating Data to Study Performance of Clustering Algorithms. CRAN Repository.

Ordonez, C. & Omiecinski, E. (2002). FREM: fast and robust EM clustering for large data sets. Proceedings of the eleventh international conference on Information and knowledge management. (CIKM '02). ACM, New York, NY, USA, 590-599. DOI=10.1145/584792.584889

Appendix

Code for simulation of K-means method:

```
v<-vector()
m<-vector()
for (k in 1:10) {
for (o in 1:1000) {
A<-MixSim(MaxOmega=0.001*k, p=2, K=5, PiLow=1,sph=TRUE)
B<-simdataset(1000, Pi = A$Pi, Mu=A$Mu, S=A$S)
trueID<-B$id
x<-B$X
Q<-kmeans(x,5,nstart = 10000,iter.max=200,algorithm="Lloyd")
estID<-Q$cluster
t<-table(trueID,estID)
nom<-0
for (i in 1:5) {
column<-t[,i]
maxPos<-which.max(column)
nom<-nom+sum(column)-t[maxPos,i]
for(j in 1:5) {
if(j!=maxPos) {
t[j,i]<-0
}
}
}
for (i in 1:5) {
row<-t[i,]
maxPos<-which.max(row)
nom<-nom+sum(row)-t[i,maxPos]
}
v[o]<-nom
}
m[k]<-median(v)
}
------------------------------------------------------------------------------------------------------------
```

Code for simulation of emEM method:

```
--------------------------------------------------------------------------------------------------------------
library(EMCluster)
library(MixSim)
library(e1071)

n<-1000
K<-10
p<-2
v<-c()

for (w in 1:100) {
   A<-MixSim(BarOmega=0.01, p=2, K=10, PiLow=1,sph=FALSE)
   B<-simdataset(n, Pi = A$Pi, Mu=A$Mu, S=A$S)
   trueID<-B$id
   x<-B$X
   dim(x)<-dim(B$X)

#### Clustering after initialization with the true cluster centers #####
    gamma <- matrix(rep(0, n*K), ncol = K)
    for (i in 1:n){
        gamma[i,trueID[i]] <- 1
    }
    init <- m.step(x, Gamma = gamma)
    Qtarget <- emcluster(x, pi = init$pi, Mu = init$Mu, LTSigma=init$LTSigma, assign.class =
    TRUE)
    targetID<-Qtarget$class

    targetAR<-classAgreement(table(trueID,targetID))$crand
```

```
 #### Clustering after random initialization #####

    bestBIC <- Inf

    .EMC$short.iter <- n / 5
    .EMC$short.eps <- 0.01
    EMruns <- 100

    for (i in 1:100) {
        Q <- em.EM(x, nclass = K)
        estID<-Q$class
        m <- K - 1 + K * p + K * p * (p + 1) / 2
        BIC <- -2 * Q$llhdval + m * log(n)

        AR <- classAgreement(table(trueID, estID))$crand

         if (BIC < bestBIC){
                   bestBIC <- BIC
                   bestAR <- AR
          }
    }
    cat("Run =", w, " :   Target AR =", targetAR, "  Found AR =", AR, "\n")
    plot(x, col=estID)
v<-c(v, AR)
} # end of loop in w
------------------------------------------------------------------------------------------------------------------------
```