Model-based clustering with non-parametric initialization

Kyungduk Lim

Korea University, Seoul, South Korea

Abstract

Model-based clustering including k-means clustering is widely used method for unsupervised classification. In model-based clustering, data are assumed to be generated by a mixture model, and clustering will be done with parameters of the model. Though we regard all the parameters of the model, the result of model-based clustering may vary by initial values of parameters. However, if too strong initialization is imposed, the clustering cannot get over from the initial status. A suggestion for this issues is using non-parametric initialization such as centroid of grid cell and directions of points in the cell. The non-parametric initialization can induce the clustering algorithm to stable solution.

Key words

model-based clustering; Mahalanobis distance; non-parametric initialization;

1. Introduction

On model-based clustering, observations are assumed to be generated by a mixture model. Then clustering will be done based on parameters in the mixture model. The simplest model-based clustering method is K-means clustering which is widely used for unsupervised classification. However, the clustering uses only Euclidean distances of observations from each centroid to allocate each observation to proper cluster

disregarding the other parameters of model. In this paper, general model-based clustering will be introduced, specifically for gaussian mixtures of multivariate normal distribution. In section 2, a brief overview for multivariate normal distribution based clustering using Mahalanobis distance will be given with general clustering algorithm.

It is well-known fact that results of model-based clustering depend on initial values of parameters such as cluster mean and variance. I will show that initialization affects the result of clustering significantly. In section 3, to solve this initialization problem, non-parametric initialization method using count, grid cell, and direction will be given for general p-dimensional case. To help understading the algorithm, each step will be explained with figures with Sepal.Length and Sepal.Width from iris data. Example of the initialization and clustering with whole iris data is going to be shown in section 4. Moreover, visualization of results of the clustering by principal variables will be suggested in section 5.

2. Multivariate Normal-Based Clustering

In model-based clustering, the observations are assumed to be generated by a mixture model. In this paper, the model is restricted to multivariate normal distribution. Then the density of observation $x = (x_1, \cdots, x_n)$ is given as follows.

$$f(x) = \prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k \, f_k(x_i|\theta_k)$$

where G is the number of clusters, τ_k is the probability of belonging to the k-th cluster, and $f_k(x;\mu,\Sigma)$ is multivariate normal distribution with mean μ_k and covariance Σ_k, i.e.

$$f_k(x;\mu_k,\Sigma_k) = (2\pi)^{-p/2}|\Sigma_k|^{-p/2}\exp(-(x-\mu_k)^{\top}\Sigma_k^{-1}(x-\mu_k)/2).$$

Therefore, for each cluster, 2 parameters of μ_k and Σ_k is given and then, there are totally $2\times G$ parameters in the model.

On the mixture model, Euclidean distance

$$d_E(x;\mu) = \sqrt{(x-\mu)^\top (x-\mu)}$$

cannot explain covariance Σ_k which is set of scales or correlations of x. Therefore, Mahalanobis distance

$$d_M(x;\mu,\Sigma) = \sqrt{(x-\mu)^\top \Sigma^{-1}(x-\mu)}$$

should be applied in model-based clustering.

Let μ_k^j and Σ_k^j be the mean and covariance of k-th cluster on the beginning of j-th step respectively. Also, define $d_M^{i,k}$ be Mahalanobis distance of x_i from centroid of k-th cluster, and $g^{i,j}$ be allocated group of x_i after j-th step. Then model-based clustering is given as follows.

(Multivariate Nomal) Model-based algorithm

---

1. Initialize : Set μ_k^0, Σ_k^0, $j=1$ where $k=1, \cdots, G$.
2. For each x_i, allocate x_i to a group $g^{i,j}$, where $g^{i,j} = \arg\max_k (d_M^{i,k})$
3. Update group mean μ_k^j and group covariance Σ_k^j of j-th repeat with

$$\mu_k^{j+1} = \frac{\sum_{g^{i,j}=k} x_i}{\sum I[g^{i,j}=k]}, \quad \Sigma_k^{j+1} = \frac{\sum_{g^{i,j}=k} (x_i-\mu_k^{j+1})(x_i-\mu_k^{j+1})^\top}{\left(\sum I[g^{i,j}=k]\right)-1}.$$

   Also, replace j with $j+1$.
4. Repeat step 2 and step 3 until μ_k^j and Σ_k^j converges.

---

3. Non-parametric Initialization

Model-based clustering algorithm is sensitive to the selection of the initialization. With inappropriate initializations, the clustering algorithm may converge to solutions of local minimums of the criterion functions. Therefore proper initialization should be suggested

for improved results. Besides, with appropriate initialization, efficiency of the algorithm is expected to be increased.

On model-based clustering of multivariate normal distribution, common methods used for initialization is random points for cluster centroid, μ_i, and I_p for cluster covariance, Σ_i. Though the initialization can avoid interventions, for example, of strong assumption of parametric distributions, insights and striking features of data are also ignored.

To adjust intervention and disregard, nonparametric initialization could be suggested. Without any model assumption, each initial values of cluster centroids and covariances will be suggested by centroids of grid cells and directions of points in the cell respectively.

Grid generating algorithm
1. For each x_i, calculate $R_i = x_i^{U95} - x_i^{L95}$.
2. Devide variables x_A and x_B respectively to k equidistance intervals where x_A and x_B have the largest and the second largest range. Then, the grid of k row and k column is generated. (k equals the number of clusters)

Let x_i^{U95} be the maximum of x_i not greater than upper 95% quantile of x_i, and x_i^{L95} be the minimum of x_i not less than lower 95% quantile of x_i. Define a quantile range of x_i as $R_i = x_i^{U95} - x_i^{L95}$. Suppose $A = \text{argmax}\{R_1, R_2, \cdots, R_p\}$ and $B = \text{argmax}\{\{R_1, R_2, \cdots, R_p\} - \{R_A\}\}$, then two variables x_A and x_B have the largest and the second largest range. Agha, M.E. and Ashour, W.M. (2012) used maximum and minimum of each variables, but both maximum and minimum can be exaggerated by a outlier. Therefore excepting the upper and lower 5% observations could be robust and proper.

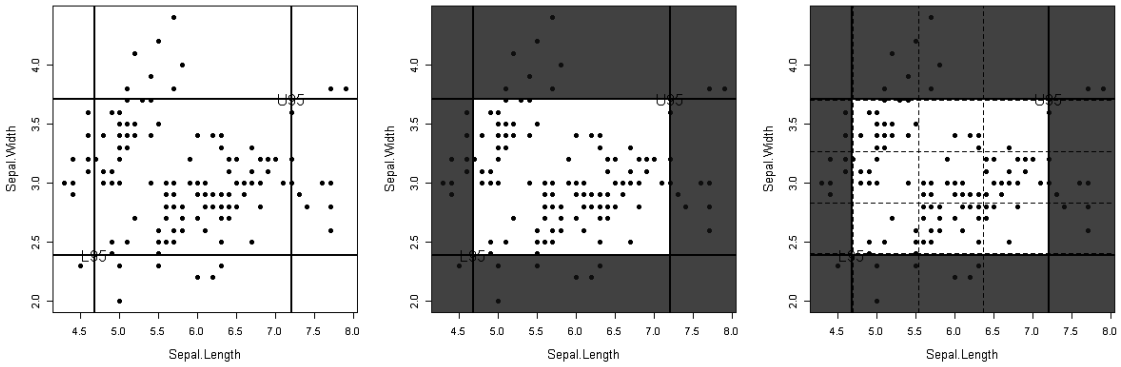

Figure 1. Grid generating step for Sepal.Length and Sepal.Width from iris data. First of all, observations outside R_is are neglected(middle), and then divide each R_i by $k=3$ intervals(right).

After generating the grid, $k \times k$ cells are given and by counting the number of observations in each cell, a table $T=(t_{(l,m)})$ which consists of $t_{(l,m)}$, counts of (l,m) -th cell also can be calculated. By using the table, initial cells will be chosen and, from the information withing the cells, parameter initialization will be done.

Let a location indicator $(a,\ b)$ is adjacent to $(c,\ d)$ if $(\ a=c\pm1,\ b=d\ )$ or $(\ a=c,\ b=d\pm1\ )$.

Cell selection algorithm

1. Calculate $T_1=(t_{(l,m)}),\quad l,\ m=1,\ \cdots,k$

   where $t_{(l,m)}$ = the number of obs in (l,m)th cell.

2. Let $L_j=\arg\max_{(l)} t_{(l,\,m)}$ and $M_j=\arg\max_{(m)} t_{(l,\,m)}$.

3. Update the table of counts T_j as $T_{j+1}=(t'_{(l,m)})$,

   where $t'_{(l,\,m)}=\begin{cases}0 & if\ \ (l,m)\ \ is\ \ adjacent\ \ to\ \ (L_j,\ M_j)\\ t_{(l,\ m)} & o.w.\end{cases}$

4. Repeat k times to get coordinates of k-cells $(L_j,\ M_j),\ j=1,\ \cdots,\ k$

   ※If there are same counts, then do selection randomly.

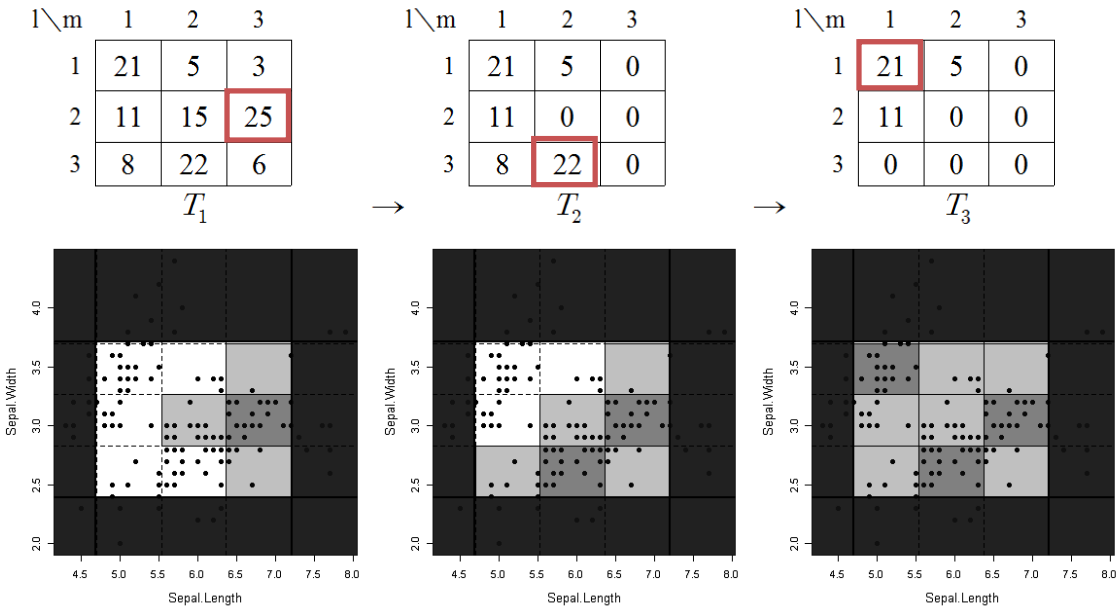

Figure 2. Steps of cell selection for Sepal.Length and Sepal.Width. $t_{(2,3)}$ is selected by T_1, and then $t_{(3,2)}$ and $t_{(1,1)}$ are selected by T_2 and T_3 in sequence. Each step is visualization(the darker cells are selected cells and lightly shaded cells are the adjacent cells of selected cells

Within the selected k cells, cell centroids will be proposed with m_j, means of observations in each cell. Then, m_j is given as

$$m_j = \frac{\sum_{x_i \in cell(j)} x_i}{n_j} = (m_{1j},\ m_{2j},\ \cdots,\ m_{pj})$$

where $cell(j)$ is a cell with coordinates $(L_j,\ M_j),\ j = 1,\ \cdots,\ k$ and $n_j = \sum_{i=1}^{n} I[x_i \in cell(j)]$, the number of observations within $cell(j)$. Then $m_j,\ j = 1,\ \cdots,\ k$ are initial centroids for model based-clustering.

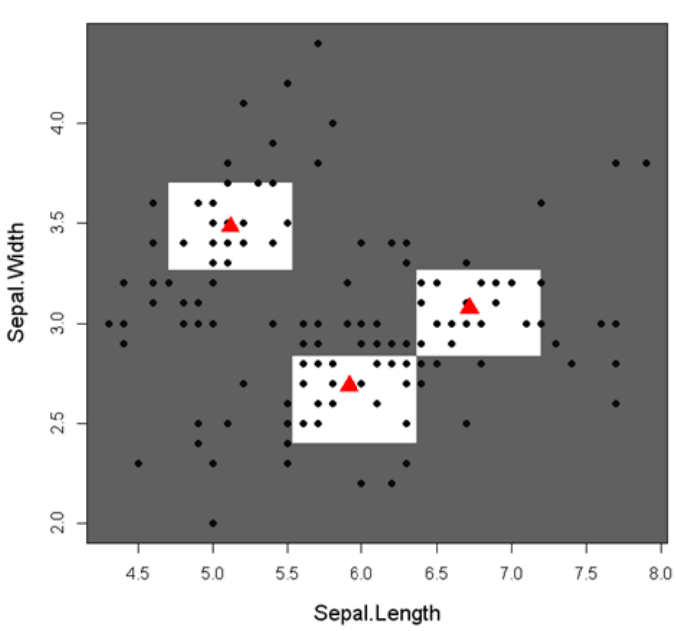

Figure 3. m_j are marked as triangles. Namely, triangles are mean points of each cell. Observations in shaded region are not concerned in initialization.

In sequence, direction method will be suggested for initialization for covariance matrices of each $cell(j)$. Because covariance matrix is a composition of covariances, covariance of each pair of variables should be suggested. However, for some convenience, variances of variables are fixed to 1, so I will focus on correlation matrix rather than covariance matrix.

Let $x_h^j = (x_{1h}^j,\ x_{2h}^j,\ \cdots,\ x_{ph}^j)$ be h-th observation in $cell(j)$, $h = 1,\ \cdots,\ n_j$, and $(x_{uh}^j,\ x_{vh}^j)$ be uth and vth variables of x_h^j. Then the direction of $(x_{uh}^j,\ x_{vh}^j)$ from the cell centroid $(m_{uj},\ m_{vj})$, can be defined as $\phi_h^j(u,v)$, where

$$\phi_h^j(u,v) = \left(\arccos\left[\frac{(x_{uh}^j)}{\sqrt{(x_{uh}^j)^2 + (x_{vh}^j)^2}}\right] - \pi\right) \times sign\left(\arccos\left[\frac{(x_{vh}^j)}{\sqrt{(x_{uh}^j)^2 + (x_{vh}^j)^2}}\right]\right) + \pi.$$

Then $0 \le \phi_h^j(u,v) \le 2\pi$, and directions of each $(x_{uh}^j,\ x_{vh}^j)$ are measured by the angle $\phi_h^j(u,v)$.

Correlation of two variable will be close to +1 if the two variable have same sign

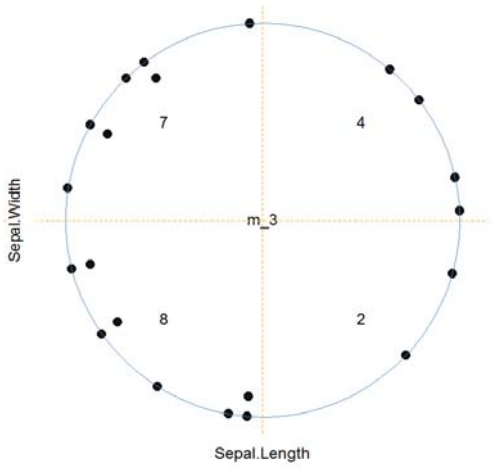

Figure 4. Direction diagram for cell $(1, 1)$. 12 observations are on 1st and 3rd quadrant. Therefore, with $n_3 = 21$, $\rho^{3}_{(S.L,\ S.W)} = 2 \times \frac{12}{21} - 1 = 0.14$.

from the mean, and close to -1 if the two variable have opposite signs. Therefore, for correlation of $(x_u^j,\ x_v^j)$ variable x_u and x_v in $cell(j)$, $\rho^{j}_{(u,\,v)}$ could be suggested as follows.

$$\rho^{j}_{(u,\,v)} = 2\left[\frac{\sum_{h=1}^{n_j} I\left[0 \le \phi_h^j \le \pi/2 \ \text{ or } \ \pi \le \phi_h^j \le 3\pi/2\right]}{n_j}\right] - 1$$

If all the $({x_u}^j_h,\ {x_v}^j_h)$ are in the 1st and 3rd quadrant with origin $(m_{uj},\ m_{vj})$, then $\rho^{j}_{(u,\,v)}$ equals to +1, otherwise, all the $({x_u}^j_h,\ {x_v}^j_h)$ are in the 2nd and 4th quadrant, then $\rho^{j}_{(u,\,v)}$ equals to -1. $\rho^{j}_{(u,\,v)}$ could be relatively strong for initialization, for proper constant c, $c \times \rho^{j}_{(u,\,v)}$ would be better. (Although all the observations are in the 1st and 3rd quadrant, the correlation can not be +1 unless they are in a line.)

In sequence,

$$P_j = (c \times \rho^{j}_{(u,\,v)}),\ \ u, v = 1,\ \cdots,\ p$$

will be suggested for initial value of covariance matrix of $cell(j)$.

In the end, $(m_j,\ P_j),\ j = 1,\ \cdots,\ k$ is a non-parametric initialization for model-based clustering

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1.00			
Sepal.Width	-0.12	1.00		
Petal.Length	0.87	-0.43	1.00	
Petal.Width	0.82	-0.37	0.96	1.00

Table 1. Correlation matrix of 4 variables from iris data

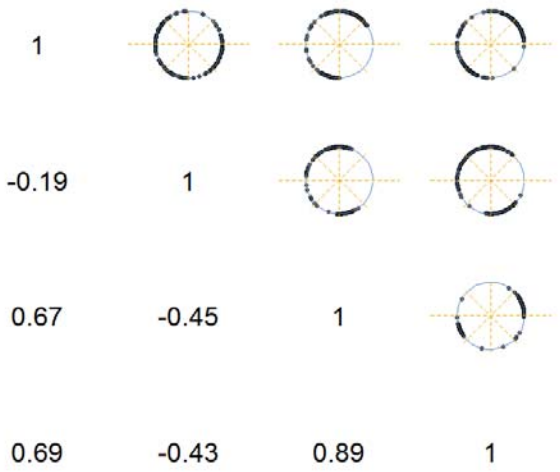

Figure 5. Direction diagrams and $\rho_{(u,v)}$ for whole iris data with $c=1$. Compared with actual correlation of 4 variables in iris data (given above), a non-parametric initialization P_{iris} seem to be similar with the correlation.

4. Example of Model-based clustering with non-parametric Initialization

Iris data will be used for an example of non-parametric Initialization.

4.1 Clustering Iris data without "Species"

There are 4 variables of Sepal.Length(x_1), Sepal.Width(x_2), Petal.Length(x_3), and Petal.Width(x_4) in the iris data. x_3 and x_1 are the first two variables of large range. Suppose $k=3$. Then, by cell selection algorithm, $(1,\ 3)$, $(2,\ 2)$, and $(3,\ 1)$th cells are selected where $n_1=33,\ n_2=16,\ n_3=30$.

For the centroid, $m_1=(5.06,\ 3.46,\ 1.53,\ 0.25)$, $m_2=(5.82,\ 2.73,\ 4.04,\ 1.23)$, and $m_3=(6.70,\ 3.03,\ 5.33,\ 1.93)$ are given as centroid. For the initialization of covariance matrices, $P_1=\begin{bmatrix} 1 & & & \\ 0.52 & 1 & & \\ -0.15 & -0.15 & 1 & \\ 0.21 & 0.21 & 0.27 & 1 \end{bmatrix}$, $P_2=\begin{bmatrix} 1 & & & \\ 0.12 & 1 & & \\ 0 & 0.38 & 1 & \\ 0.25 & 0.62 & 0.5 & 1 \end{bmatrix}$, and $P_3=\begin{bmatrix} 1 & & & \\ 0.27 & 1 & & \\ 0.2 & 0 & 1 & \\ 0.07 & 0.13 & 0.33 & 1 \end{bmatrix}$ are suggeted.

With just 10 iteration, $\widehat{\mu_j}$ and $\widehat{\Sigma_j}$ are converged. ($n_1=50,\ n_2=41,\ n_3=59$,

$\widehat{\mu_1}=(5.01,\ 3.43,\ 1.46,\ 0.25)$, $\widehat{\mu_2}=(5.74,\ 2.73,\ 4.09,\ 1.27)$, $\widehat{\mu_3}=(6.57,\ 2.95,\ 5.39,\ 1.91)$,

$\widehat{\Sigma_1}=\begin{bmatrix} .124 & .099 & .016 & .010 \\ & .144 & .012 & .009 \\ & & .030 & .006 \\ & & & .011 \end{bmatrix}$, $\widehat{\Sigma_2}=\begin{bmatrix} .169 & .072 & .121 & .038 \\ & .100 & .086 & .046 \\ & & .172 & .057 \\ & & & .033 \end{bmatrix}$, $\widehat{\Sigma_3}=\begin{bmatrix} .346 & .084 & .250 & .044 \\ & .100 & .070 & .048 \\ & & .356 & .114 \\ & & & .113 \end{bmatrix}$)

In existing random initialization, initial μ_js are randomly selected from k observation, and initial Σ_ks are set I_p. Table 2. shows the number of observations in each cluster from each 10 clustering.

(Because order of clusters could be different in each simulation, index of cluster are standardized by increasing order of Sepal.Length)

Initial obs	Cluster 1	Cluster 2	Cluster 3
25, 109, 104	80	87	13
105, 129, 38	50	45	55
7, 77, 21	24	26	100
126, 81, 88	21	53	76
78,117, 51	55	50	45
24, 68, 97	50	15	85
97, 98, 10	82	50	18
52, 8, 110	50	45	55
43, 17, 143	24	26	100

Table 2. Result of 10 clustering with random initialization

Table 2. says, by initial values of clustering, the outcome could be significantly different while non-parametric method give only one solution for a clustering. Though we cannot say which method is better by the number of observations in each cluster, non-parametric method is much stable than random one.

5. Visualization by Principal Variables

For p-dimensional data, usual approach for visualization is dimension reduction by principal component analysis or factor analysis. By the methods, observations can be expressed by two main axes which are combinations of variables. These combination explains much of variance of variables, however, explanation of themselves would be uncertain. Principal variable is a method to deal with this uncertainty. Rather than making a few combinations, choosing some variables which have more impact than the others would be better. Though explanation power could be reduced, principal variable gives clear impression.

In this clustering problem, it is desirable to select principal variables which can tell the differences between clusters. Because visualization space is restricted to 2-dimensional space, generally, two principal variables will be enough to be visualized.

Let $\mu_j = (\mu_{j1},\ \mu_{j2},\ \cdots,\ \mu_{jp})$ be the final mean and $\Sigma_j = (\sigma_{j,lm})$ be final covariance of j-th cluster. ($\sigma_{j,lm}$ is covariance of x_l and x_m in j-th cluster.) Because only two variables are to be chosen, mean and covariance also reduced to 2-dimension. Suppose x_a and x_b are the chosen variable, and define $\mu_j^{a,b} = (\mu_{ja},\ \mu_{jb})$ and $\Sigma_j^{a,b} = \begin{bmatrix} \sigma_{j,aa} & \sigma_{j,ab} \\ \sigma_{j,ba} & \sigma_{j,bb} \end{bmatrix}$, reduced mean and covariance with two variables x_a and x_b.

To select principal variables, a proper criterion should be introduced.

Define

$$d^{a,b}_{inter} = \sum_{i=1}^{k} \sum_{j \neq i} \left[(\mu_j^{a,b} - \mu_i^{a,b})^\top (\Sigma_i^{a,b})^{-1} (\mu_j^{a,b} - \mu_i^{a,b}) \right].$$

Then, $d_{inter}^{a,b}$ equals sum of inter-Mahalanobis distances of $\mu_j^{a,b}$s, each mean of k clusters of principal variables x_a and x_b.

For all possible $p(p-1)/2$ combination of variables of size 2, i.e. for all possible combinations of $(a,\ b)$, calculate $d_{inter}^{a,b}$ and select a combination $(a',\ b')$ where

$$(a',\ b') = \mathrm{argmax}_{(a,\ b)}\ d_{inter}^{a,b}.$$

Namely, two variables which maximize variance between groups will be chosen. The next thing is to plot $(a',\ b')$ with $\mu_j^{a',b'} = (\mu_{ja'},\ \mu_{jb'})$ and $\Sigma_j^{a',b'} = \begin{bmatrix} \sigma_{j,a'a'} & \sigma_{j,a'b'} \\ \sigma_{j,b'a'} & \sigma_{j,b'b'} \end{bmatrix}$.

The visualization is shown below with the example of clustering from 4.1.

$(a,\ b)$	(1, 2)	(1, 3)	(1, 4)	(2, 3)	(2, 4)	(3, 4)
$d_{inter}^{a,b}$	138.75	880.58	428.97	974.86	525.70	973.27

Table 3. $d_{inter}^{a,b}$s are calculated for the clustering of iris data from 4.1. Shown above, variable set (2, 3) are selected for principal variables.

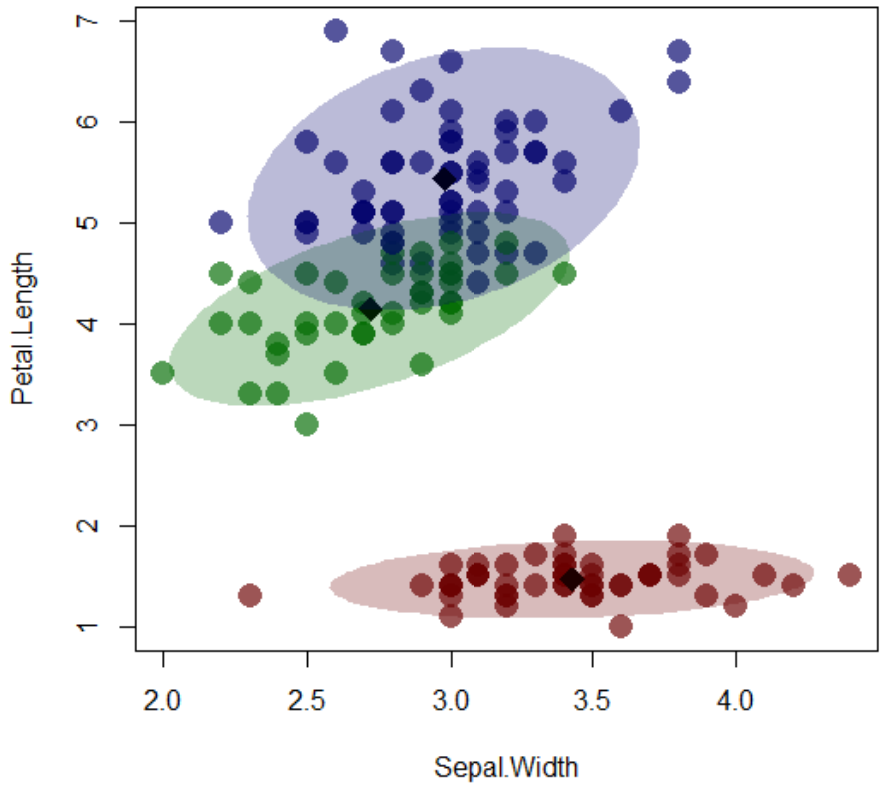

Figure 6. Visualization with two principal variables Sepal.Width and Petal.Length from clutering of iris data. The color of points means the final cluster each observation is belonging to. The ellipse is drawn based on the covariance of each cluster by eigenvalue decomposition, but sizes of ellipses do not have meaning. However, we can find out different structure of clusters.

6. Discussions about initialization

6.1 Determining proper constant c of $P_j = (c \times \rho_{(u,v)}^{j})$

As mentioned introduction above, a constant c have to be set. If c is close to 1, then initiation of covariance matrix is too strong. But if c is close to 0, it's same as assuming there's no correlation. i.e. $P_j = I_p$. Therefore we should set c properly. However, c may be vary by data, and no algorithm for setting c is suggested yet.

6.2 Determining proper number of segments on making grid

Suppose k is proper number of clusters for a certain data. Then, by grid generating algorithm, $k \times k$ cells are set. In the $k \times k$ cells, only k cells are chosen for initial centroid. Then the ratio of the number of selected cells from total number of cells equals to $\frac{1}{k}$. It means for large enough k, too many cells are generated. Also for small k, generating more cells could be better in initialization. For example, for $k = 2$, the ratio equals to 0.5. 2 cells of 4 cells are chosen. But if $k = 4$, the ratio drops to 0.25, the half of 0.5. Therefore, with further analysis, upper and lower bound for the ratio should be suggested for better initialization.

7. Conclusion

Initialization is a issue on the model-based clustering. With proper initial values of parameters, the number of iterations can be reduced and also clustering algorithm converges to better estimations. But if the initialization is too strong, then the result of clustering could be pulled toward initial values, so that it can be ruined. Therefore, one should set proper but not heavy initial values. Existing random initialization avoid intervening into clustering steps, but also at the same time, miss some inspiration that data shows. In this situation, non-parametric initialization gives some implications. Non-parametric initial values do not ask heavy calculation. For example, to calculate covariance, we need multiple of n^2 calculations. But direction methods ask multiple of n calculations giving good suggestion for initial covariance matrix. Thus, non-parametric initialization is expected to be effective also for massive data.

Moreover, by the non-parametric initialization gives a solution of clustering whereas random initialization gives totally different clustering by selection of initial values. Therefore non-parametric initialization is more stable in comparison with random methods.

8. Reference

- Fraley, C. and Raftery, A.E. (2007) “Bayesian Regularization for Normal Mixture Estimation and Model-Based Clustering”, *Journal of Classification*, 24. 155-181.
- Agha, M.E. and Ashour, W.M. (2012) “Efficient and Fast Initialization Algorithm for K-means Clustering”, *I.J. Interlligent Systems and Applicatoins*, 1. 21-31.
- Huh, M.H. (2011). *Exploratory Multivariate Data Anaylsis*. Jayu Academy. 95-104.