Model Based Clustering

1.Model Selection

1.1 Basic Case

If the number of clusters K is fixed, we only need to measure the probability of $f(X \mid \Theta_k)$,given M_K(model k), where k is the index of model ,then using EM to do clustering.

1.2 Select K and Model together

Idea: *To measure the integrated likelihood* $f(\Theta_k \mid X)$ *instead of just* $f(X \mid \Theta_k)$ *for a specific model.*

For comparing two models, M_i and M_j , we have posterior comparing:

$$\frac{f(M_i \mid X)}{f(M_j \mid X)} = \frac{f(X \mid M_i) f(M_i) / \sum_{k=1}^{K} f(X \mid M_k) f(M_k)}{f(X \mid M_j) f(M_j) / \sum_{k=1}^{K} f(X \mid M_k) f(M_k)}$$

$$= \frac{f(X \mid M_i) f(M_i)}{f(X \mid M_j) f(M_j)}$$

Bayesian Factor:

$$B_{ij} = f(X \mid M_i) / f(X \mid M_j).$$

Approache 1: **Bayesian Information Criterion(BIC)**

$$2 \log f(X \mid M_k) \approx 2 \log f(X \mid \hat{\Theta}_k, M_k) - v_k \log(n) = BIC$$

For comparing two models, M_i and M_j , we have

$$\log(B_{ij}) = \log f(X \mid M_i) - \log f(X \mid M_j) = \tfrac{1}{2}(BIC_i - BIC_j)$$

If $BIC_i > BIC_j$ then $\log(B_{ij}) > 0$, ie. $B_{ij} > 1$. Model i is better than model j.

Approach 2: **AIC** penalizes the number of parameters less strongly than does the BIC, and AIC can be derived in the same Bayesian framework as BIC, just by using a different prior.

$$\mathrm{AIC} = 2k - 2\ln(\hat{L})$$

where

- $\hat{L}$ = the maximized value of the likelihood function of the model, i.e. $\hat{L} = p(x|\hat{\theta}, M)$, where $\hat{\theta}$ are the parameter values that maximize the likelihood function;
- x = the observed data;
- k = the number of free parameters to be estimated.

2. Clustering Procesure

2.1 Clustering by EM

steps:
1. Set a maximum for the number of clusters, which is usually less that 10. A set of mixture models, say, a subset of 8 covariance structures is considered.

2. Perform bottom up clustering to approximately maximize the classification likelihood for each model, and obtain the corresponding classifications for up to M groups.

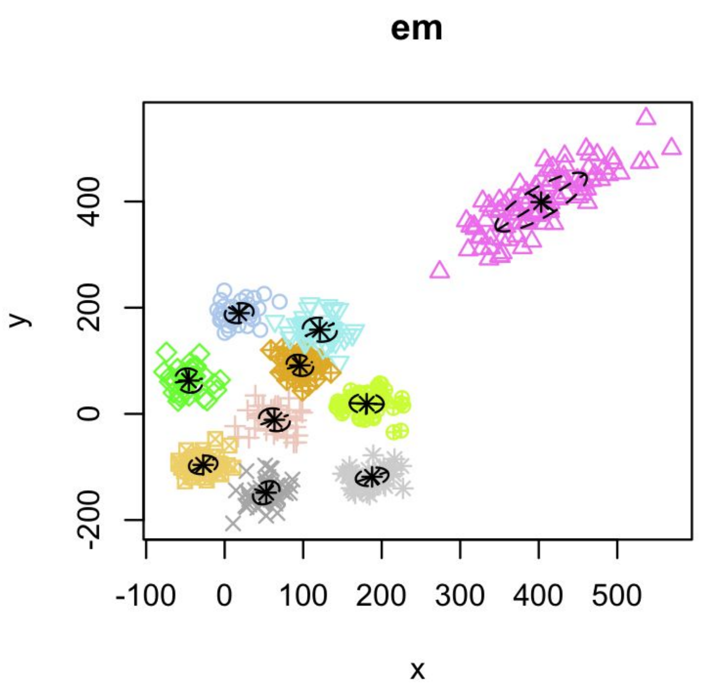
3. Apply EM algorithm for each model and each number of clusters 2,…,M, starting with the classification result from hierarchical agglomeration(bottom up clustering).

4. Compute BIC for the one-cluster case for each model and for the mixture model with the optimal parameters from EM for 2, …, M clusters.

5. Choose the model corresponding to the largest BIC.

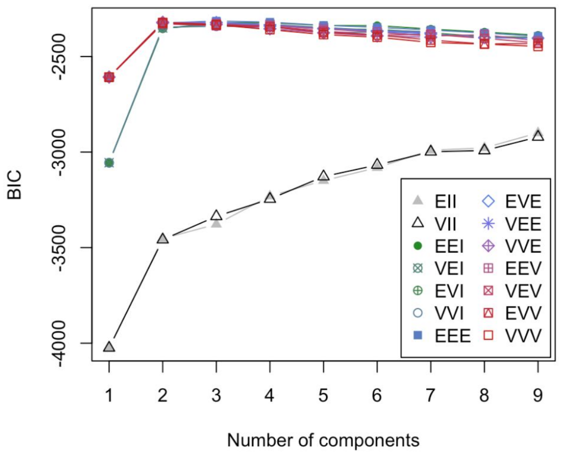
3. Software and Demo(unsupervised clustering)

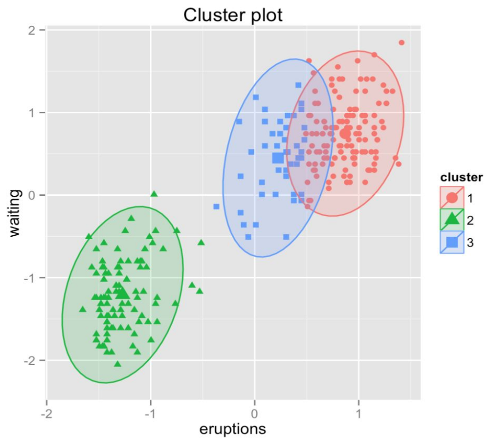
3.1 “EMCLUST” using EM




```
[> library(EMCluster, quiet = TRUE)
[> set.seed(1234)
[> x <- da1$da
[> TC <- da1$class
[> n <- nrow(x)
[> p <- ncol(x)
[> et.em <- init.EM(x, nclass = 10, method = "em.EM")
[> ret.em <- init.EM(x, nclass = 10, method = "em.EM")
[> plotem(ret.em, x, main = "em")
```

3.2 MClust: *provides the optimal mixture model estimation according to BIC*






```
library(mclust)
# Model-based-clustering
mc <- Mclust(faithful)
# Print a summary
summary(mc)
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust EEE (ellipsoidal, equal volume, shape and orientation) model with 3 components:
##
##  log.likelihood   n df       BIC       ICL
##       -1126.361 272 11 -2314.386 -2360.865
##
```

```
# Optimal number of cluster
mc$G
```

```
# Probality for an observation to be in a given cluster
head(mc$z)
```

```
##              [,1]         [,2]         [,3]
## 1 2.181744e-02 1.130837e-08 9.781825e-01
## 2 2.475031e-21 1.000000e+00 3.320864e-13
## 3 2.521625e-03 2.051823e-05 9.974579e-01
## 4 6.553336e-14 9.999998e-01 1.664978e-07
## 5 9.838967e-01 7.642900e-20 1.610327e-02
## 6 2.104355e-07 9.975388e-01 2.461029e-03
```

```
# BIC values used for choosing the number of clusters
plot(mc, "BIC")
```