# Penalized Model-Based Clustering
## with Application to Variable Selection

by Wooyoul Na, Jaehyeong Ahn

Department of Education, Konkuk University
Department of Applied Statistics, Konkuk University

*jayahn0104@gmail.com*

# Model-based Clustering

**Data**

$$X = (X_1, \cdots, X_p) = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

$$Z = (z_1, \cdots, z_k)^T, \quad z_k \in 0, 1$$

$X$: Observed data of $nxp$ dimension matrix

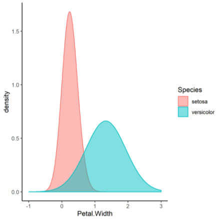
$z_k$: Binary indicator of whether $x_i$ is from component $k$ (unobserved data)

# Model-based Clustering

## Data

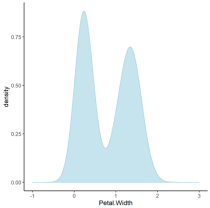
When $z_k$ is observed

| Length | Width | Species |
|---|---|---|
| 1.4 | 0.2 | setosa |
| 1.4 | 0.2 | setosa |
| 1.3 | 0.2 | setosa |
| 1.5 | 0.2 | setosa |
| 1.4 | 0.2 | setosa |
| 4.7 | 1.4 | versicolor |
| 4.5 | 1.5 | versicolor |
| 4.9 | 1.5 | versicolor |
| 4.0 | 1.3 | versicolor |
| 4.6 | 1.5 | versicolor |




When $z_k$ is not observed

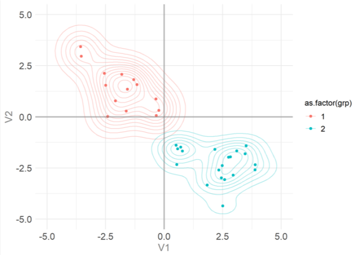
| Length | Width |
|---|---|
| 1.4 | 0.2 |
| 1.4 | 0.2 |
| 1.3 | 0.2 |
| 1.5 | 0.2 |
| 1.4 | 0.2 |
| 4.7 | 1.4 |
| 4.5 | 1.5 |
| 4.9 | 1.5 |
| 4.0 | 1.3 |
| 4.6 | 1.5 |



# Mixture Model

**Example of mixture model with clusters**

$$X \sim f(x;\Theta) = \pi_1 f_1(;\theta_2) + \pi_2 f_2(x;\theta_2) = \sum_{k=1}^{K} \pi_k f_k(x;\theta_k)$$




**Under Assumption** $f_1$ and $f_2$ are normal density

$$f_1(x;\theta_1) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma_1|^{\frac{1}{2}}} \exp(-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1))$$

$$f_2(x;\theta_2) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma_2|^{\frac{1}{2}}} \exp(-\frac{1}{2}(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2))$$

# Model-based Clustering

## Model

Suppose that $P(z_k = 1) = \pi_k$. $\pi_k$ is satisfying $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$

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

$f_k$: k'th component from specific distribution

$\theta_k$: parameter of k'th component from specific distribution

(i) Likelihood for incomplete data

$$\log L(X|\Theta) = \sum_{i=1}^{n} \log[\sum_{k=1}^{K} \pi_k f_k(x_i|\theta_k)]$$

(ii) Likelihood for complete data

$$logL_c(X|\Theta) = \sum_{k} \sum_{i} z_{ki}[log\pi_k + log f_k(x_i|\theta_k)]$$

$z_k \sim Ber(\tau_{ki})$

$$\tau_{ki} = P(z_k = 1|x) = \frac{\pi_k f_k(x_i|\theta_k)}{\sum_{k=1}^{K} \pi_k f_k(x_i|\theta_k)}$$

# Model-based Clustering

**EM algorithm**

E-step (Expectation)

$$\tau_{ki}^{(m)} = \frac{\pi_i^{(m)} f_k(x_i|\theta_k^{(m)})}{\sum_{k=1}^{K} \pi_k^{(m)} f_k(x_i|\theta_k^{(m)})}$$

M-step (Maximization)

$$argmax_{(\Theta)}[E(logL_c(X|\Theta)] \qquad E(logL_c(X|\Theta)) = \sum_k \sum_l \tau_{ki}[log\pi_k + logf_k(x_i|\theta_k)$$

# EM algorithm for Normal density

## Assumption

$$f_k(x;\theta_k) = N(\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma_k|^{\frac{1}{2}}} exp(-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k))$$

Take initial values about $\hat{\mu}_1, \cdots, \hat{\mu}_k, \hat{\Sigma}_1, \cdots, \hat{\Sigma}_k, \hat{\pi}_1, \cdots, \hat{\pi}_k$

E-step(Expectation)

$$\tau_{ki}^{(m)} = \frac{\pi_i^{(m)} f_k(x_i|\theta_k^{(m)})}{\sum_{k=1}^{K} \pi_k^{(m)} f_k(x_i|\theta_k^{(m)})}$$

M-step(Maximazation)

$$\hat{\pi}_k^{(m+1)} = \sum_{i=1}^{n} \tau_{ki}^{(m)}/n$$

$$\hat{\mu}_k^{(m+1)} = \frac{\sum_{i=1}^{n} \tau_{ki}^{(m+1)} \hat{\mu}_k^{(m)}}{\sum_{i=1}^{n} \tau_{ki}^{(m+1)}}$$

$$\hat{\Sigma}_k^{(m+1)} = \frac{\sum_{i=1}^{n} \tau_{ki}^{(m)}(x_i-\hat{\mu}_k^{(m)})(x_i-\hat{\mu}_k^{(m)})^T}{\sum_{i=1}^{n} \tau_{ki}^{(m)}}$$

Iterate E-step and M-step until converge

# Penalized Model-based Clustering

## Penalized likelihood

(i) Penalized log-likelihood

$$\log L_P(X|\Theta) = \Sigma_{i=1}^{n} \log[\Sigma_{k=1}^{K} \pi_k f_k(x_i|\theta_k)] - h_\lambda(\Theta) \qquad h_\lambda(\Theta) = \sum_{k=1}^{K} \sum_{p=1}^{P} |\mu_{kp}|$$

(ii) Penalized log-likelihood for complete data

$$\log L_{c,P}(X|\Theta) = \Sigma_k \Sigma_i z_{ki} [\log \pi_k + \log f_k(x_i|\theta_k)] - h_\lambda(\Theta)$$

# EM algorithm for penalized model-based clustering

**E-step**

$$\tau_{ki} = P(z_k = 1|x) = \frac{\pi_k f_k(x_i|\theta_k)}{\sum_{k=1}^{K} \pi_k f_k(x_i|\theta_k)}$$

**M-step**

$$argmax_{(\Theta)}[E(logL_c(X|\Theta) - h_\lambda(\Theta)]$$

# EM algorithm for penalized model-based clustering

## Assumption

$$f_k(x;\theta_k) = N(\mu_k, V) = \frac{1}{(2\pi)^{\frac{p}{2}}|V|^{\frac{1}{2}}} \exp(-\frac{1}{2}(x-\mu_k)^T V^{-1}(x-\mu_k))$$

$$\text{where} \quad \Sigma_1 = \cdots = \Sigma_k = V = \begin{pmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_p^2 \end{pmatrix}$$

$$h_\lambda(\Theta) = \sum_{k=1}^{K}\sum_{p=1}^{P}|\mu_{kp}|$$

# Formula: Estimation for penalized EM-algorithm

## E-step

$$\tau_{ki}^{(m)} = \frac{\pi_i^{(m)} f_k(x_i|\theta_k^{(m)})}{\sum_{k=1}^{K} \pi_k^{(m)} f_k(x_i|\theta_k^{(m)})}$$

## M-step

$$\hat{\sigma}_p^{2/(m+1)} = \Sigma_{k=1}^{K}\Sigma_{i=1}^{n}\left(\frac{\tau_{ki}^{(m)}(x_{ip}-\mu_{kp}^{(m)})^2}{n}\right) \quad \text{where } n = \Sigma_{k=1}^{K}\Sigma_i^{m}\tau_{ki} \qquad \hat{\pi}_k^{(m+1)} = \sum_{i=1}^{n}\tau_{ki}^{(m)}/n$$

$$\hat{\mu}_k^{(m+1)} = sign(\tilde{\mu}_k^{(m+1)})(|\tilde{\mu}_k^{(m+1)}| - \frac{\lambda}{\Sigma_i \tau_{ki}^{(m+1)}} V^{(m+1)}\mathbf{1})_+ \qquad \text{where } \tilde{\mu}_k^{(m+1)} = \frac{\sum_{i=1}^{n}\tau_{ki}^{(m+1)}x_i}{\sum_{i=1}^{n}\tau_{ki}^{(m+1)}}$$

$$\text{If } \lambda \leq \left|\frac{\Sigma_{i=1}^{n}\tau_{ki}^{(m+1)}x_i}{\sigma_p^{2}(m+1)}\right|, \hat{\mu}_{kp}^{m+1)} = |\tilde{\mu}_{kp}^{(m+1)}| - \frac{\lambda\sigma_p^{2(m+1)}}{\Sigma_{i=1}^{n}\tau_{ki}^{(m+1)}} \qquad \text{If } \lambda > \left|\frac{\Sigma_{i=1}^{n}\tau_{ki}^{(m+1)}}{\sigma_p^{2(m+1)}}\right|, \hat{\mu}_{kp}^{(m+1)} = 0$$