Complete Membership Model for High Dimensional Clustering

1 Introduction

Literature review:

Finite mixture model clustering.

Difficulty in high dimensional clustering.

The difficulty in Determinantal process clustering, repulsive clustering.

Classification EM.

2 Complete Membership Model

2.1 General Framework

Let y_i be the observed data following the distribution F with parameter θ_i. Consider a membership model with κ possible membership:

$$\begin{aligned} \pi(y_i) &\overset{indep}{\sim} F(y_i|\theta_i) \\ \pi(\theta_i) &= \sum_{k=1}^{\kappa} z_{k,i}\delta_{\theta_k^*}(\theta_i) \end{aligned} \tag{1}$$

for each i, $\{z_{\cdot,i}\}$ denotes the membership with only one 1 and $(k-1)$ many 0's. Note $\{z_{\cdot,i}\}$ is not random for each i, although collectively $\{\sum_i z_{1,i}, \sum_i z_{2,i}, \ldots \sum_i z_{\kappa,i}\}$ follows a multinomial distribution $Multinomial(n, \{w_1, w_2, \ldots, w_\kappa\})$.

2.2 Estimation

$$\log[\{y_i\}_i, \{z_{k,i}\}_{k,i}] = \sum_i \sum_{k=1}^{\kappa} z_{k,i} \log\{w_k F(y_i \mid \theta_k)\} \tag{2}$$

This allows us to utilize Expectation-Maximization algorithm for parameter estimation:

Algorithm 1 Estimation algorithm

while $||\theta^{(s)} - \theta^{(s+1)}|| > \epsilon$ **do**
  Maximize over $\{z_{i,k}\}_k$ by $\underset{\{z_{i,k}\}_k}{\text{argmax}} \{z_{i,k} w_k F(y_i|\theta_k^*)\}$ for all i;
  Update $\theta_k^* = \text{argmax} \sum_{i=1}^n z_{k,i} \log F(y_i|\theta_k^*)$ for all k
  Update $w_k = \sum_i z_{k,i} / \sum_i \sum_k z_{k,i}$

2.3 High Dimensional Estimation

Alternative Least Square with regularization

3 Theory

3.1 Increased Separation among the Centers

Consider the finite mixture model with the center estimates as $\mu_k = \frac{\sum_i \mathbb{E} z_{k,i} y_i}{\sum_i \mathbb{E} z_{k,i}}$, and complete membership model with the center estimates as $\mu_k^* = \frac{\sum_i z_{k,i} y_i}{\sum_i z_{k,i}}$. We are interested in comparing the pairwise distance among the centers from the two models.

Let the pairwise distance be $||\mu_1 - \mu_2||$ between two centers in the finite mixture. As the $||\mu_1 - \mu_2|| = \sqrt{\sum_{l=1}^p ||\mu_{1,l} - \mu_{2,l}||^2}$, we focus on one sub-dimension $\mu_{1,l} - \mu_{2,l}$, without loss of generality, we assume $\mu_{1,l} > \mu_{2,l}$.

For any $y_{j,l} \geq \frac{\sum_{i \neq j} \mathbb{E} z_{1,i} y_{i,l}}{\sum_{i \neq j} E z_{1,i}} \geq \frac{\sum_{i \neq j} \mathbb{E} z_{2,i} y_{i,l}}{\sum_{i \neq j} E z_{2,i}}$ and $\mathbb{E} z_{1,j} \geq \mathbb{E} z_{2,j}$,

$$\begin{aligned}
\mu_{1,l} - \mu_{2,l} =& \frac{\sum_{i \neq j} \mathbb{E} z_{1,i} y_{i,l} + \mathbb{E} z_{1,j} y_{j,l}}{\sum_{i \neq j} \mathbb{E} z_{1,i} + \mathbb{E} z_{1,j}} - \frac{\sum_{i \neq j} \mathbb{E} z_{2,i} y_{i,l} + \mathbb{E} z_{2,j} y_{j,l}}{\sum_{i \neq j} \mathbb{E} z_{2,i} + \mathbb{E} z_{2,j}} \\
\leq& \frac{\sum_{i \neq j} \mathbb{E} z_{1,i} y_{i,l} + y_{j,l}}{\sum_{i \neq j} \mathbb{E} z_{1,i} + 1} - \frac{\sum_{i \neq j} \mathbb{E} z_{2,i} y_{i,l}}{\sum_{i \neq j} \mathbb{E} z_{2,i}}
\end{aligned}$$

For any $y_{j,l} \leq \frac{\sum_{i\neq j} \mathbb{E}z_{2,i}y_{i,l}}{\sum_{i\neq j} Ez_{2,i}} \leq \frac{\sum_{i\neq j} \mathbb{E}z_{1,i}y_{i,l}}{\sum_{i\neq j} Ez_{1,i}}$ and $\mathbb{E}z_{1,j} \leq \mathbb{E}z_{2,j}$,

$$\mu_{1,l} - \mu_{2,l} \leq \frac{\sum_{i\neq j} \mathbb{E}z_{1,i}y_{i,l}}{\sum_{i\neq j} \mathbb{E}z_{1,i}} - \frac{\sum_{i\neq j} \mathbb{E}z_{2,i}y_{i,l} + y_{j,l}}{\sum_{i\neq j} \mathbb{E}z_{2,i} + 1}$$

By induction, this converts all the $Ez_{k,i}$ to $z_{k,i}$ hence $\mu_{1,l} - \mu_{2,l} \leq \mu^*_{1,l} - \mu^*_{2,l}$.

3.2 Convex Relaxation

Something similar to:

Agarwal, Alekh, Sahand Negahban, and Martin J. Wainwright. "Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions." The Annals of Statistics (2012): 1171-1197.

4 Simulation

Note: preliminary

RMSE:

Model	$n = 100, p = 100, p^* = 5$	$n = 100, p = 100, p^* = 100$
k-means	0.40	0.069
CM	0.43	0.069
GMM	0.46	0.069
PCA+ k-means	0.17	0.030
PCA + CM	0.07	0.030
PCA + GMM	0.10	0.030
new model		
new model		
new model		

5 Application

6 Discussion

References