Modified EM algorithms for High-Dimensional Clustering

Beomseok Seo

Penn State University

Nov. 29. 2018

Motivation

- For **unsupervised learning** techniques, model-based approach is one of the most popular methods.
- Model-based approach exploits latent variable Gaussian mixture model (**GMM**) and estimates the model through expectation and maximization (**EM**) algorithm.

$$
\begin{aligned}
\text{E-step}: \quad Q(\theta|\theta^{(t)}) &= \hat{E}_{Z|X,\theta^{(t)}} l_c(\theta; X, Z) \\
&= \sum_{i}^{n} \sum_{k}^{K} \underbrace{\hat{E}_{\theta^{(t)}}[z_{ki}|x_i]}_{\tau_{ki}^{(t)}} \{\log \pi_k + \log f_k(x_i; \theta_k)\} \\
\text{M-step}: \quad \theta^{(t+1)} &= \arg\max_{\theta} Q(\theta|\theta^{(t)})
\end{aligned}
$$

- However, when data is **high dimensional**, EM algorithm is confronted with identifiability, stability and computational efficiency problems.

Motivation

- Literature approached this issue by imposing sparsity through modification of either E-step or M-step in EM algorithm.

  $\Rightarrow$ **Modifying E-step**

  $$Q(\theta|\theta^{(t)}) = \hat{E}_{Z|X,\theta^{(t)}}\{l_c(\theta; X, Z) + p(\theta)\}$$

  $\Rightarrow$ **Modifying M-step**

  $$\theta^{(t+0.5)} = \arg\max_{\theta} Q(\theta|\theta^{(t)})$$
  $$\theta^{(t+1)} = s(\theta^{(t+0.5)})$$

- Different modifications have different performances in stability and computational efficiency.

Comparison between different modifications

- Where the modification is applied is not critical.
- Different modifications result in different update rule.

E.g. [Modification of M-step]: truncation step (Wang et al. 2014)
The update rule changes as follows.

$$\tau_{ki}^{(t)} = \frac{\pi_k^{(t)} f_k(x_i; \theta_k^{(t)})}{\sum_k^K \pi_k^{(t)} f_k(x_i; \theta_k^{(t)})}, \quad \pi_k^{(t+1)} = \frac{1}{n} \sum_i^n \tau_{ki}^{(t)} \quad \mu_k^{(t+0.5)} = \frac{\sum_i^n \tau_{ki}^{(t)} x_i}{\sum_i^n \tau_{ki}^{(t)}},$$

And additional step impose sparcity.

$$\begin{aligned} \hat{\mathcal{S}}^{(t+0.5)} &= \text{set of index } j\text{'s of the top } s \text{ largest } |\mu_{kj}^{(t+0.5)}| \\ \hat{\mu}_k^{(t+1)} &= \begin{cases} \mu_k^{(t+0.5)} & j \in \hat{\mathcal{S}}^{(t+0.5)} \\ 0 & j \notin \hat{\mathcal{S}}^{(t+0.5)} \end{cases} \end{aligned}$$

Comparison between different modifications

E.g. [Modification of E-step]: L_1 penalty (Pan and Shen 2007)

$$Q_p(\theta;\theta^{(t)}) = \sum_i^n \sum_k^K \tau_{ki}^{(t)} \{\log \pi_k + \log f_k(x_i;\theta_k)\} - \lambda \sum_k^K \sum_j^p |\mu_{kj}|$$

The update rule is same as previous example except that the additional step changes as follows.

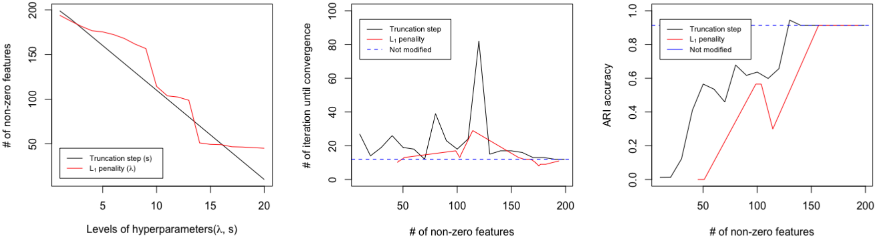
$$\hat{\mu}_k^{(t+1)} = sign(\mu_k^{(t+0.5)}) \left(|\mu_k^{(t+0.5)}| - \frac{\lambda}{\sum_i \tau_{ki}^{(t+1)}} V 1_p \right)_+ ,$$

For truncation step

$$\hat{\mu}_k^{(t+1)} \quad = \quad sign(\mu_k^{(t+0.5)}) \left(|\mu_k^{(t+0.5)}| - \mu_{k(n-s)}^{(t+0.5)} \right)_+ .$$

Simulation study

- $N = 100$, $p = 200$, $K = 3$(# of clusters), K-means initialization.

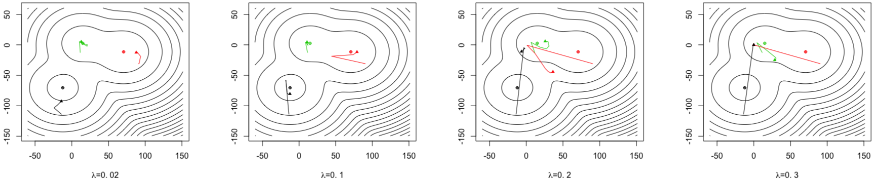


- Number of iteration till convergence is not proportional to dimension size.
- L1 penalized EM converges more quickly than EM with truncation step.
- Accuracy could not be improved by dim reduction.
- Any type of EM did not work well on random initialization.

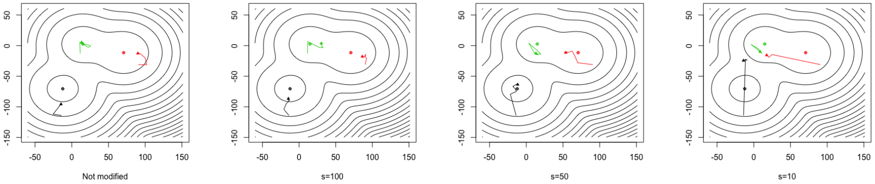
Simulation study

- Geometric interpretation on 2D principal components space.

L_1 penalty with different λ

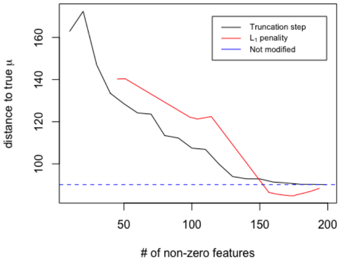


Truncation step with different thresholding s



Simulation study

- Geometric interpretation.



- Bias of estimated μ decreases at some level of hyperparameter and then increases as more dimensions are reduced.
- $\Rightarrow$ Bias corrected penalty such as SCAD, MCP may work better.