One-Hot Membership Model

1 Introduction

hard clustering: several overly strong assumptions in k-means

soft clustering: large variance induced by way too many non-zeros in the probability simplex.

2 One-Hot 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:

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

where $\{z_{.,i}\}$ denotes the membership with only one 1 and $(k-1)$ many 0's.

To allow borrowing of strength among $\{z_{.,i}\}$, we assume $z_{.,i}$ follows:

$$\begin{aligned} &\pi(z_{.,i}) \propto \prod_{k=1}^{\kappa} (u_{k,i}v_k)^{z_{k,i}} \\ &v_k \sim Dir_\kappa(1, 1, \dots 1) \\ &u_{.,i} \overset{iid}{\sim} Dir_\kappa(\epsilon, \epsilon, \dots \epsilon) \end{aligned} \tag{2}$$

where ϵ is a small number close to 0, which causes $u_{k,i}$ to have only one value close to 1 and the rest close to 0. Due to the $z_{.,i}$ will take one-of-κ memberships with probability almost 1, we refer this as one-hot membership model.

To compare with other methods, note when $\{z_{.,i}\}$ is assumed to be independent for each i, (1) is exactly the same as the k-means model; when $u_{k,i}$ is fixed to $1/\kappa$ for all k, i, the membership $z_{.,i}$ can be integrated out to obtain general finite mixture model.

The key difference is in one-hot membership model, the multinomial distribution of $z_{.,i}$ has its probability weight concentrated to one vertex, so that the estimate of $z_{.,i}$ is obtained via maximization, instead of expectation or sampling as in the general mixture model.

2.2 Estimation

$$\log[\{y_i\}_i, \{z_{k,i}\}_{k,i}] = \sum_i \sum_{k=1}^{\kappa} z_{k,i} \log\{\delta_{\theta_k}(y_i) v_k\} \tag{3}$$

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

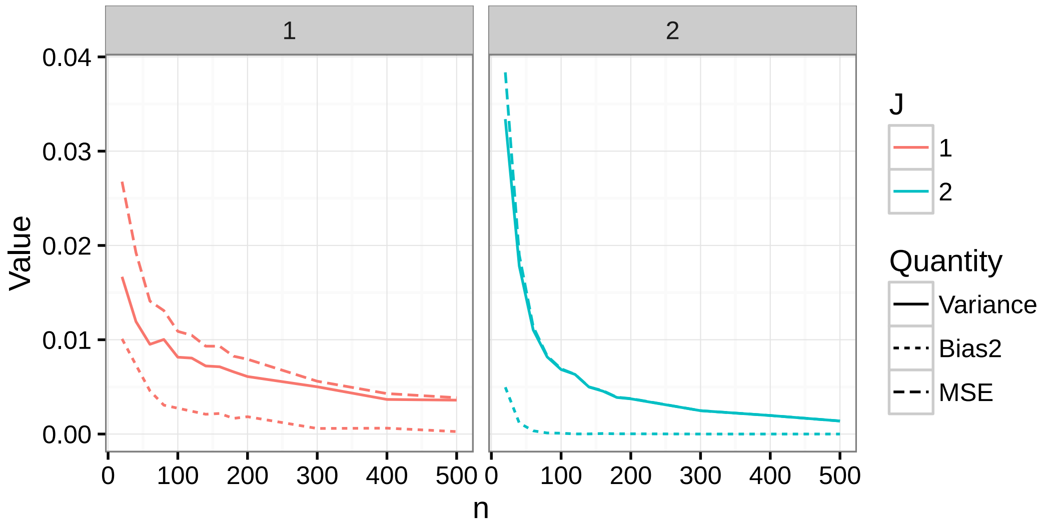
Algorithm 1 Estimation algorithm

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

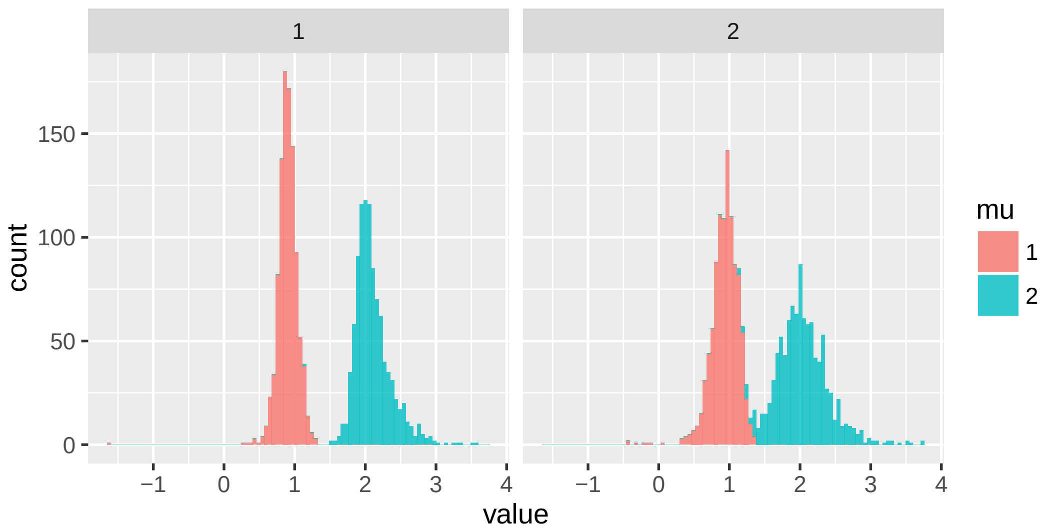
3 Theory

3.1 Empirical Result: Bias-Variance Trade-off

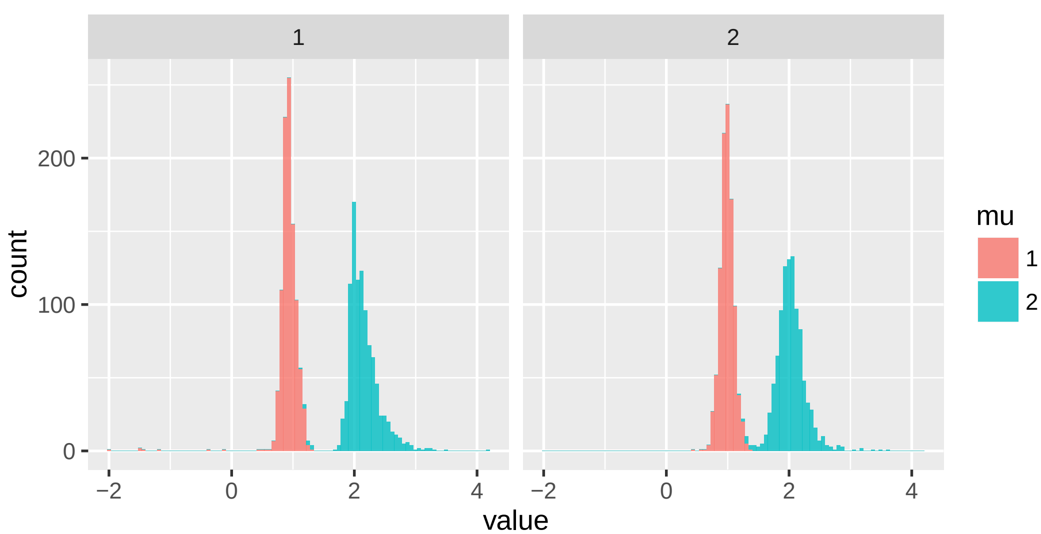
We generate data y_i from $N(\mu_i, 0.5^2)$, using 75% with $\mu_i = 1$ and 25% with $\mu_i = 2$.



(a) The bias-variance trade-off.



(b) The distribution of MLE for μ_1 and μ_2 over 1,000 simulation data sets, with total 40 data points.



(c) The distribution of MLE for μ_1 and μ_2 over 1,000 simulation data sets, with total 150 data points.

3.2 Variance Reduction in Small Sample Size

Let superscript denote the variable to be integrated over. Using variance decomposition, the variance of the parameter estimator $\hat{\theta}$ has:

$$\mathrm{Var}^y(\hat{\theta}) = \mathrm{E}^z\mathrm{Var}^y(\hat{\theta}|z) + \mathrm{Var}^z\mathrm{E}^y(\hat{\theta}|z)$$

Using subscript FMM and OH to denote the different models. As each z_i is maximized to a fixed value in each dataset in one-hot model, It can be immediately seen that $\mathrm{Var}^z_{OH}\mathrm{E}^y(\hat{\theta}|z) = 0$ and $\mathrm{E}^z_{OH}\mathrm{Var}^y(\hat{\theta}|z) = \mathrm{Var}^y_{OH}(\hat{\theta}|z)$; whereas each z_i is random in finite mixture model, $\mathrm{Var}^z_{FMM}\mathrm{E}^y(\hat{\theta}|z) \geq 0$. Therefore, it suffices to have $\mathrm{Var}^y_{OH}(\hat{\theta}|z) - \mathrm{E}^z\mathrm{Var}^y_{FMM}(\hat{\theta}|z) < \mathrm{Var}^z_{FMM}\mathrm{E}^y(\hat{\theta}|z)$ for a reduction in variance.

4 Simulation

5 Application

6 Discussion

References