GAUSSIAN MIXTURE MODEL FOR CLUSTERED OBSERVATIONS

ANIKO SZABO

1. Mixture model

Let Y_{ij}, $i = 1, \dots, \mathcal{I}$, $j = 1, \dots, n_i$ denote a continuous random variable measured in subject j of cluster i. We want to model the distribution of Y as a mixture of K latent normal distributions:

$$[Y_{ij} \mid \xi_{ij} = k] \sim N(\mu_k, \sigma_k), \tag{1}$$

where ξ_{ij} is the latent class indicator for subject j of cluster i taking values from 1 to K.

Due to the clustering we cannot assume that ξ_{ij} are independent, however we will assume that they follow and exchangeable multinomial distribution $\mathcal{EM}(K, \boldsymbol{q})$.

Denoting $\phi(y; \mu, \sigma)$ the normal pdf with mean μ and standard deviation σ, the likelihood is

$$\begin{aligned} L =& \prod_{i=1}^{\mathcal{I}} \sum_{\boldsymbol{x} \in \mathbb{Z}_K^{n_i}} Pr(\boldsymbol{\xi}_i = \boldsymbol{x}) \times \prod_{j=1}^{n_i} \phi(y_{ij}; \mu_{x_j}, \sigma_{x_j}) && (2) \\ =& \prod_{i=1}^{\mathcal{I}} \sum_{\boldsymbol{x} \in \mathbb{Z}_K^{n_i}} \frac{q_{\dot{\boldsymbol{x}}|n_i}}{\binom{n_i}{\dot{\boldsymbol{x}}}} \times \prod_{j=1}^{n_i} \phi(y_{ij}; \mu_{x_j}, \sigma_{x_j}), && (3) \end{aligned}$$

where $\boldsymbol{x} = (x_1, \dots, x_{n_i})$, $\boldsymbol{\xi}_i = (\xi_{i1}, \dots, \xi_{in_i})$, $\dot{\boldsymbol{x}} = (\sum_j I(x_j = 1), \dots, \sum_j I(x_j = K))$ is the vector of latent category frequencies, and

$$q_{\boldsymbol{r}|n_i} = \sum_{\boldsymbol{t} \in \mathbb{Z}_N : \boldsymbol{t}'\boldsymbol{1} = N} h(\boldsymbol{r}, \boldsymbol{t}, N, n_i) q_{\boldsymbol{t}|L} \tag{4}$$

are the exchangeable and marginally reproducible outcome probabilities of the exchangeable multinomial distribution with maximal cluster size N with $h(\boldsymbol{r}, \boldsymbol{t}, N, n_i)$ defined as multivariate hypergeometric probabilities.

2. EM algorithm

We can consider ξ_{ij} missing data. Then the complete data likelihood is

$$L_c(\mu, \sigma, q | \xi_{ij}, y_{ij}) = \sum_{i=1}^{\mathcal{I}} \frac{q_{\dot{\boldsymbol{\xi}}|n_i}}{\binom{n_i}{\dot{\boldsymbol{\xi}}}} \prod_{j=1}^{n_i} \log \phi(y_{ij}; \mu_{\xi_{ij}}, \sigma_{\xi_{ij}}) = \sum_{i=1}^{\mathcal{I}} \prod_{\boldsymbol{z} \in \mathbb{Z}_K^{n_i}} \big[\frac{q_{\dot{\boldsymbol{z}}|n_i}}{\binom{n_i}{\dot{\boldsymbol{z}}}} \prod_{j=1}^{n_i} \log \phi(y_{ij}; \mu_{z_j}, \sigma_{z_j}) \big]^{I(\boldsymbol{\xi}_i = \boldsymbol{z})}. \tag{5}$$

Its logarithm without the binomial term that does not contain unknown parameters is

$$\log L_c(\mu, \sigma, q | \xi_{ij}, y_{ij}) = \sum_{i=1}^{\mathcal{I}} \sum_{\boldsymbol{z} \in \mathbb{Z}_K^{n_i}} I(\boldsymbol{\xi}_i = \boldsymbol{z}) \big[\log q_{\dot{\boldsymbol{z}}|n_i} + \sum_{j=1}^{n_i} \log \phi(y_{ij}; \mu_{z_j}, \sigma_{z_j}) \big] \tag{6}$$

The expected complete data log-likelihood, given previous estimates $\mu^{(m)}, \sigma^{(m)}, q^{(m)}$ is

$$\begin{aligned} Q(\mu, \sigma, q) =& E\big[\log L_c(\mu, \sigma, q \mid \mu^{(m)}, \sigma^{(m)}, q^{(m)}, y_{ij}) \big] = \\ & \sum_{i=1}^{\mathcal{I}} \sum_{\boldsymbol{z} \in \mathbb{Z}_K^{n_i}} Pr(\boldsymbol{\xi}_i = \boldsymbol{z} \mid \mu^{(m)}, \sigma^{(m)}, q^{(m)}, y_{ij}) [\log q_{\dot{\boldsymbol{z}}|n_i} + \sum_{j=1}^{n_i} \log \phi(y_{ij}; \mu_{x_j}, \sigma_{x_j})], && (7) \end{aligned}$$

Date: November 27, 2019.

2.1. **E-step.** Using the Bayes theorem

$$e_{i\boldsymbol{z}}^{(m)} = Pr(\boldsymbol{\xi}_i = \boldsymbol{z} \mid \mu^{(m)}, \sigma^{(m)}, q^{(m)}, y_{ij}) = \tag{8}$$

$$\frac{Pr(\boldsymbol{y}_i \mid \boldsymbol{\xi}_i = \boldsymbol{z}, \mu^{(m)}, \sigma^{(m)}, q^{(m)}) Pr(\boldsymbol{\xi}_i = \boldsymbol{z} \mid \mu^{(m)}, \sigma^{(m)}, q^{(m)})}{\sum_{\boldsymbol{w}} Pr(\boldsymbol{y}_i \mid \boldsymbol{\xi}_i = \boldsymbol{w}, \mu^{(m)}, \sigma^{(m)}, q^{(m)}) Pr(\boldsymbol{\xi}_i = \boldsymbol{w} \mid \mu^{(m)}, \sigma^{(m)}, q^{(m)})} = \frac{\prod_{j=1}^{n_i} \phi(y_{ij}; \mu_{z_j}^{(m)}, \sigma_{z_j}^{(m)}) q_{\dot{\boldsymbol{z}}|n_i}^{(m)} / \binom{n_i}{\dot{\boldsymbol{z}}}}{\sum_{\boldsymbol{w}} \prod_{j=1}^{n_i} \phi(y_{ij}; \mu_{w_j}^{(m)}, \sigma_{w_j}^{(m)}) q_{\dot{\boldsymbol{w}}|n_i}^{(m)} / \binom{n_i}{\dot{\boldsymbol{w}}}}. \tag{9}$$

2.2. **M-step.**

$$Q(\mu, \sigma, q) = \sum_{i=1}^{\mathcal{I}} \sum_{\boldsymbol{z} \in \mathbb{Z}_K^{n_i}} e_{i\boldsymbol{z}}^{(m)} [\log q_{\dot{\boldsymbol{z}}|n_i} + \sum_{j=1}^{n_i} \log \phi(y_{ij}; \mu_{z_j}, \sigma_{z_j})] = \tag{10}$$

$$\sum_{i=1}^{\mathcal{I}} \sum_{\boldsymbol{z} \in \mathbb{Z}_K^{n_i}} e_{i\boldsymbol{z}}^{(m)} \log q_{\dot{\boldsymbol{z}}|n_i} + \sum_{i=1}^{\mathcal{I}} \sum_{\boldsymbol{z} \in \mathbb{Z}_K^{n_i}} e_{i\boldsymbol{z}}^{(m)} \sum_{j=1}^{n_i} \log \phi(y_{ij}; \mu_{z_j}, \sigma_{z_j}) = \tag{11}$$

$$\sum_{i=1}^{\mathcal{I}} \sum_{\boldsymbol{z} \in \mathbb{Z}_K^{n_i}} e_{i\boldsymbol{z}}^{(m)} \log q_{\dot{\boldsymbol{z}}|n_i} + \sum_{k=1}^{K} \sum_{i=1}^{\mathcal{I}} \sum_{j=1}^{n_i} \sum_{\boldsymbol{z}: z_j = k} e_{i\boldsymbol{z}}^{(m)} \log \phi(y_{ij}; \mu_k, \sigma_k), \tag{12}$$

where the two components can be maximized separately, and within the last component the terms corresponding to different k's can be maximized separately as well.

2.2.1. *Update for q.* In terms of q, the log-likelihood can be viewed as the $\mathcal{EM}$ log-likelihood for an extended data set: for each cluster i and for each possible frequency vector of responses $\boldsymbol{r} \in \mathbb{Z}_{n_i}$, $\boldsymbol{r}'\mathbf{1} = n_i$ we compute the 'observed' frequency as $a_{i\boldsymbol{r}}^{(m)} = \sum_{\boldsymbol{z}: \dot{\boldsymbol{z}} = \boldsymbol{r}} e_{i\boldsymbol{z}}^{(m)}$. Then $q_{\boldsymbol{t}|N}^{(m+1)}$ can be obtained from the EM algorithm for fitting the $\mathcal{EM}$ model to these frequencies.

2.2.2. *Update for μ and σ.* In terms of the normal distribution parameters, we have a weighted normal log-likelihood for each k. The weight corresponding to observation j in cluster i is $b_{ijk}^{(m)} = \sum_{\boldsymbol{z}: z_j = k} e_{i\boldsymbol{z}}^{(m)}$. Then

$$\mu_k^{(m+1)} = \sum_{i=1}^{\mathcal{I}} \sum_{j=1}^{n_i} b_{ijk}^{(m)} y_{ij} \tag{13}$$

$$[\sigma_k^{(m+1)}]^2 = \sum_{i=1}^{\mathcal{I}} \sum_{j=1}^{n_i} b_{ijk}^{(m)} (y_{ij} - \mu_k^{(m+1)})^2 \tag{14}$$