Computational Statistics Midterm: EM for mixture distribution

Pei-Shan Yen, Jieqi Tu, Jun Lu, Hajwa Kim, Yanli Gao

Nov 25 2020

1 Introduction

The random variable Y, finite mixture distribution, defined as a random drawn from a collection of underlying individual random variables $X_j, j = 1, 2, ..., k$ with the corresponding probability density function f_j. The individual random variables that are combined to form the mixture distribution are called the mixture components, and the probabilities (or weights) associated with each component are called the mixture weights. Assume the chance that the mixture Y following X_j is the mixture weight p_j. The sum of the mixture weights should be 1, i.e., $\sum_{j=1}^{k} p_j = 1$.

The probability density function of the mixture Y is

$$f(y) = \sum_{j=1}^{k} p_j f_j \tag{1}$$

If a mixture dataset with sample size n is observed, and the ith observation is from the jth component, the indicator variable δ can be defined as

$$\delta_{ij} = 1 \ when \ y_i \sim f_j \tag{2}$$

Obviously, $\underset{\sim}{\delta_i} = (\delta_{i1}, \delta_{i2}, ..., \delta_{ik})$ follows a multinomial distribution

$$\underset{\sim}{\delta_i} \sim MN(size = 1; p_1, p_2, .., p_k) \tag{3}$$

2 EM algorithm for mixture Poisson

2.1 Point Estimate for mixture Poisson

Assume the underlying individual distribution is Poisson distribution with rate parameter λ_j. We have

$$Y_i|\underset{\sim}{\delta_i} \sim Poi(\lambda_j) \tag{4}$$

The joint distribution of (Y, δ) is

$$f(\underset{\sim}{Y}, \underset{\sim}{\delta}) = \prod_{i=1}^{n} f(\underset{\sim}{Y_i}|\underset{\sim}{\delta_i}) f(\underset{\sim}{\delta_i}) = \prod_{i=1}^{n} \prod_{j=1}^{k} \left[(e^{-\lambda_j} \frac{\lambda_j^{y_i}}{y_i!}) p_j \right]^{\delta_{ij}} \tag{5}$$

The log-likelihood of (Y, δ) is

$$logf(\underset{\sim}{Y}, \underset{\sim}{\delta}) = \sum_{i=1}^{n} \sum_{j=1}^{k} \delta_{ij} \Big[-\lambda_j + y_i log\lambda_j - y_i! + logp_j \Big] \tag{6}$$

The Expectation-Maximization algorithm (EM) can be used to estimate the unknown parameter $\theta = (p_1, p_2, ..., p_{k-1}, \lambda_1, \lambda_2, ..., \lambda_k)$ for the mixture distribution. Given the initial value $\theta^{(0)}$, the objective function Q is defined as

$$Q(\theta|\theta^{(0)}, \underset{\sim}{Y}) = E_{\delta_{ij}} \Big[logf(\underset{\sim}{Y}, \underset{\sim}{\delta}) \Big] = \sum_{i=1}^{n} \sum_{j=1}^{k} E\Big[\delta_{ij}\Big] \Big[-\lambda_j + y_i log\lambda_j - y_i! + logp_j \Big] \tag{7}$$

From equation (7), the conditional expectation δ_{ij} given the observation Y_i at iteration t of the algorithm is

$$E[\delta_{ij}|Y_i] = P(\delta_{ij} = 1|Y_i) = \frac{P(Y_i|\delta_{ij} = 1)P(\delta_{ij} = 1)}{\sum_{j=1}^{k} P(Y_i|\delta_{ij} = 1)P(\delta_{ij} = 1)} = \frac{f_j^{(t)} p_j^{(t)}}{\sum_{m=1}^{k} f_m^{(t)} p_m^{(t)}} = \frac{\left[e^{-\lambda_j^{(t)}} \frac{\lambda_j^{(t) y_i}}{y_i!}\right] p_j^{(t)}}{\sum_{m=1}^{k} \left[e^{-\lambda_m^{(t)}} \frac{\lambda_m^{(t) y_i}}{y_i!}\right] p_m^{(t)}} = w_{ij}^{(t)} \quad (8)$$

Hence, in the E-step, the objective function Q can be simplified to

$$Q(\theta|\theta^{(0)}, \underset{\sim}{Y}) = Q(p_j, \lambda_j|p_j^{(t)}, \lambda_j^{(t)}, \underset{\sim}{Y}) = \sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij}^{(t)}(-\lambda_j + y_i ln\lambda_j - y_i! + logp_j) \quad (9)$$

In the M-step, we maximize the objective function Q with respect to θ.

For the Poisson rate parameter λ_j, we have

$$\frac{dQ}{d\lambda_j} = \frac{dE_{\delta_{ij}}\left[logf(\underset{\sim}{Y}, \underset{\sim}{\delta})\right]}{d\lambda_j} = w_{ij}^{(t)}\left(\frac{y_i}{\lambda_j} - 1\right) \quad (10)$$

The maximum likelihood estimate of $\lambda_j^{(t+1)}$ is

$$\lambda_j^{(t+1)} = \frac{\sum_{i=1}^{n} w_{ij}^{(t)} y_i}{\sum_{i=1}^{n} w_{ij}^{(t)}} \quad (11)$$

For the mixture weight parameter p_j, we have

$$\frac{dQ}{dp_j} = \frac{dE_{\delta_{ij}}\left[logf(\underset{\sim}{Y}, \underset{\sim}{\delta})\right]}{dp_j} = \sum_{i=1}^{n}\left[\frac{w_{ij}^{(t)}}{p_j} - \frac{w_{ik}^{(t)}}{p_k}\right] \quad (12)$$

The maximum likelihood estimate of $p_j^{(t+1)}$ is

$$p_j^{(t+1)} = \frac{\sum_{i=1}^{n} w_{ij}^{(t)}}{n} \quad (13)$$

2.2 Variance Estimate for mixture Poisson

The variance estimation of the parameters can be derived from Louis formula. The information matrix $I(\theta) = I(p_1, p_2, ...p_{k-1}, \lambda_1, \lambda_2, ...\lambda_k)$ is

$$I(\theta) = -Var\left[\frac{d}{d\theta} logf(\underset{\sim}{Y}, \underset{\sim}{\delta})\right] + E\left[-\frac{d^2}{d\theta^2} logf(\underset{\sim}{Y}, \underset{\sim}{\delta})\right] = -\Psi + \Gamma \tag{14}$$

2.2.1 The matrix Ψ

The matrix Ψ with dimension of $(2K-1) \times (2k-1)$ can be partitioned into 4 sub-matrix, including the Ψ_{11}, Ψ_{12}, Ψ_{21}, Ψ_{22}.

$$\Psi_{11} = \begin{pmatrix} Var[\frac{d}{dp_1} logf] & Cov[\frac{d}{dp_1} logf, \frac{d}{dp_2} logf] & ... & ... & Cov[\frac{d}{dp_1} logf, \frac{d}{dp_{(k-1)}} logf] \\ Cov[\frac{d}{dp_2} logf, \frac{d}{dp_1} logf] & Var[\frac{d}{dp_2} logf] & ... & ... & Cov[\frac{d}{dp_2} logf, \frac{d}{dp_{(k-1)}} logf] \\ ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... \\ Cov[\frac{d}{dp_{(k-1)}} logf, \frac{d}{dp_1} logf] & ... & ... & ... & Var[\frac{d}{dp_{(k-1)}} logf] \end{pmatrix}_{(k-1)\times(k-1)} \tag{15}$$

$$\Psi_{22} = \begin{pmatrix} Var[\frac{d}{d\lambda_1} logf] & Cov[\frac{d}{d\lambda_1} logf, \frac{d}{d\lambda_2} logf] & ... & ... & Cov[\frac{d}{d\lambda_1} logf, \frac{d}{d\lambda_k} logf] \\ Cov[\frac{d}{d\lambda_2} logf, \frac{d}{d\lambda_1} logf] & Var[\frac{d}{d\lambda_2} logf] & ... & ... & Cov[\frac{d}{d\lambda_2} logf, \frac{d}{d\lambda_k} logf] \\ ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... \\ Cov[\frac{d}{d\lambda_k} logf, \frac{d}{d\lambda_1} logf] & ... & ... & ... & Var[\frac{d}{d\lambda_k} logf] \end{pmatrix}_{(k)\times(k)} \tag{16}$$

$$\Psi_{12} = \begin{pmatrix} Cov[\frac{d}{dp_1} logf, \frac{d}{d\lambda_1} logf] & Cov[\frac{d}{dp_1} logf, \frac{d}{d\lambda_2} logf] & ... & ... & Cov[\frac{d}{dp_1} logf, \frac{d}{d\lambda_k} logf] \\ Cov[\frac{d}{dp_2} logf, \frac{d}{d\lambda_1} logf] & Cov[\frac{d}{dp_2} logf, \frac{d}{d\lambda_2} logf] & ... & ... & Cov[\frac{d}{dp_2} logf, \frac{d}{d\lambda_k} logf] \\ ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... \\ Cov[\frac{d}{dp_{(k-1)}} logf, \frac{d}{d\lambda_1} logf] & ... & ... & ... & Cov[\frac{d}{dp_{(k-1)}} logf, \frac{d}{d\lambda_k} logf] \end{pmatrix}_{(k-1)\times(k)} \tag{17}$$

$$\Psi_{12} = \Psi_{21}^t \tag{18}$$

2.2.2 The sub-matrix Ψ_{11}

The matrix Ψ_{11} with dimension $(k-1) \times (k-1)$ represents the covariance between the first derivative log-likelihood with respect to p_l and the first derivative log-likelihood with respect to p_s. The diagonal entry of the matrix Ψ_{11} is

$$\psi_{ll} = Var\left[\frac{d}{dp_l} logf\right] = \sum_{i=1}^{n} Var\left[\left(\frac{\delta_{il}}{p_l} - \frac{\delta_{ik}}{p_k}\right)\right], \quad for\ 1 \le l \le k-1 \tag{19}$$

When $1 \le j, j' \le k-1$, from the properties of the multinomial distribution, we have

$$E[\delta_{ij}|\underset{\sim}{Y}] = \delta_{ij}^* \tag{20}$$

$$Var[\delta_{ij}|\underset{\sim}{Y}] = \delta_{ij}^*(1 - \delta_{ij}^*) \tag{21}$$

$$Cov[\delta_{ij}, \delta_{ij'}] = -\delta_{ij}^* \delta_{ij'}^* \tag{22}$$

$$Cov[\delta_{ij}, \delta_{i'j}] = 0 \tag{23}$$

$$Cov[\delta_{ij}, \delta_{ik}|\underset{\sim}{Y}] = Cov[\delta_{ij}, 1-(\delta_{i1}+\delta_{i2}+...+\delta_{i(k-1)})] = -\sum_{l=1}^{k-1} Cov[\delta_{ij}, \delta_{il}] = -\left[\sum_{l\neq j}(-\delta^*_{ij}\delta^*_{il})+\delta^*_{ij}(1-\delta^*_{ij})\right] = -\delta^*_{ij}\delta^*_{ik} \tag{24}$$

$$Var[\delta_{ik}|\underset{\sim}{Y}] = Cov[1-(\delta_{i1}+\delta_{i2}+...+\delta_{i(k-1)}), \delta_{ik}] = \sum_{l\neq k} -\delta^*_{il}\delta^*_{ik} = \delta^*_{ik}(1-\delta^*_{ik}) \tag{25}$$

From equation (20)-(25), the diagonal entry of matrix Ψ_{11} expressed in equation (19) is rewritten as

$$\psi_{ll} = \sum_{i=1}^{n}\left[\frac{\delta_{il}}{p_l} - \frac{\delta_{ik}}{p_k}\right] = \sum_{i=1}^{n}\left[\frac{\delta^*_{il}(1-\delta^*_{il})}{p_l^2} + \frac{\delta^*_{ik}(1-\delta^*_{ik})}{p_k^2} + 2\frac{\delta^*_{il}\delta^*_{ik}}{p_l p_k}\right] \tag{26}$$

, and the off-diagonal entry of the matrix Ψ_{11} is

$$\psi_{ll'} = Cov\left[\frac{d}{dp_l}logf, \frac{d}{dp_{l'}}logf\right] = \sum_{i=1}^{n} Cov\left[(\frac{\delta_{il}}{p_l}-\frac{\delta_{ik}}{p_k}), (\frac{\delta_{il'}}{p_{l'}}-\frac{\delta_{ik}}{p_k})\right] = \sum_{i=1}^{n}\left[\frac{-\delta^*_{il}\delta^*_{il'}}{p_l p_{l'}} + \frac{\delta^*_{il}\delta^*_{ik}}{p_l p_k} + \frac{\delta^*_{ik}\delta^*_{il'}}{p_k p_{l'}} + \frac{\delta^*_{ik}(1-\delta^*_{ik})}{p_k^2}\right] \tag{27}$$

2.2.3 The sub-matrix Ψ_{22}

The matrix Ψ_{22} with dimension $(k)\times(k)$ represents the covariance between the first derivative log-likelihood with respect to λ_l and the first derivative log-likelihood with respect to λ_s. The diagonal entry of the matrix Ψ_{22} is

$$\psi_{ll} = Var\left[\frac{d}{d\lambda_l}logf\right] = \sum_{i=1}^{n} Var\left[\delta_{il}\left(\frac{y_i}{\lambda_l}-1\right)\right] = \sum_{i=1}^{n}\left(\frac{y_i}{\lambda_l}-1\right)^2 Var[\delta_{il}] = \sum_{i=1}^{n}\left(\frac{y_i}{\lambda_l}-1\right)^2 \delta^*_{il}(1-\delta^*_{il}) \tag{28}$$

, and the off-diagonal entry of the matrix Ψ_{22} is

$$\psi_{ll'} = \sum_{i=1}^{n} Cov\left[\delta_{il}\left(\frac{y_i}{\lambda_l}-1\right), \delta_{il'}\left(\frac{y_i}{\lambda_{l'}}-1\right)\right] = \sum_{i=1}^{n}\left(\frac{y_i}{\lambda_l}-1\right)\left(\frac{y_i}{\lambda_{l'}}-1\right)\left(-\delta^*_{il}\delta^*_{il'}\right) \tag{29}$$

2.2.4 The sub-matrix Ψ_{12}

The matrix Ψ_{12} with dimension $(k-1)\times(k)$ represents the covariance between the first derivative log-likelihood with respect to p_l and the first derivative log-likelihood with respect to λ_s. The diagonal entry of the matrix Ψ_{12} is

$$\psi_{ll} = \sum_{i=1}^{n} Cov\left[\left(\frac{\delta_{il}}{p_l}-\frac{\delta_{ik}}{p_k}\right), \delta_{il}\left(\frac{y_i}{\lambda_l}-1\right)\right] = \sum_{i=1}^{n}\left(\frac{y_i}{\lambda_l}-1\right)\left[\frac{\delta^*_{il}(1-\delta^*_{il})}{p_l} + \frac{\delta^*_{il}\delta^*_{ik}}{p_k}\right] \tag{30}$$

, and the off-diagonal entry of the matrix Ψ_{12} is

$$\psi_{ll'} = \sum_{i=1}^{n} Cov\left[\left(\frac{\delta_{il}}{p_l}-\frac{\delta_{ik}}{p_k}\right), \delta_{il'}\left(\frac{y_i}{\lambda_{l'}}-1\right)\right] = \sum_{i=1}^{n}\left(\frac{y_i}{\lambda'_l}-1\right)\left[-\frac{\delta^*_{il}\delta^*_{il'}}{p_l} + \frac{\delta^*_{ik}\delta^*_{il'}}{p_k}\right] \tag{31}$$

2.2.5 The matrix Γ

The matrix Γ with dimension of $(2K-1)\times(2k-1)$ can be partitioned into 4 sub-matrix, including the Γ_{11}, Γ_{12}, Γ_{21}, Γ_{22}

2.2.6 The sub-matrix Γ_{11}

The matrix Γ_{11} with dimension $(k-1)\times(k-1)$ represents the expectation between the second derivative log-likelihood with respect to p_l and the second derivative log-likelihood with respect to p_s. The diagonal entry of the matrix Γ_{11} is

$$\gamma_{ll} = E\Big[-\frac{d^2}{dp_l^2}logf\Big] = -\sum_{i=1}^{n} E\Big[-\frac{\delta_{il}}{p_l^2} - \delta_{ik}\Big(-\frac{1}{p_k^2}\Big)\frac{d1-(p_1+p_2+...+p_{(k-1)})}{dp_l}\Big] = \sum_{i=1}^{n}\Big[\frac{\delta_{il}^*}{p_l^2}+\frac{\delta_{ik}^*}{p_k^2}\Big] \quad (32)$$

, and the off-diagonal entry of the matrix Γ_{11} is

$$\gamma_{ll'} = E\Big[-\frac{d^2}{dp_l dp_l'}logf\Big] = \sum_{i=1}^{n} E\Big[\frac{\delta_{ik}}{p_k^2}\Big] = \sum_{i=1}^{n}\frac{\delta_{ik}^*}{p_k^2} \quad (33)$$

2.2.7 The sub-matrix Γ_{22}

The matrix Γ_{22} with dimension $(k)\times(k)$ represents the expectation between the second derivative log-likelihood with respect to λ_l and the second derivative log-likelihood with respect to λ_s. The diagonal entry of the matrix Γ_{22} is

$$\gamma_{ll} = E\Big[-\frac{d^2}{d\lambda_l^2}logf\Big] = \sum_{i=1}^{n} E\Big[\frac{y_i}{\lambda_l^2}\delta_{il}\Big] = \sum_{i=1}^{n}\frac{y_i}{\lambda_l^2}\delta_{il}^* \quad (34)$$

, and the off-diagonal entry of the matrix Γ_{11} is

$$\gamma_{ll'} = E\Big[-\frac{d^2}{d\lambda_l d\lambda_l'}logf\Big] = E\Big[0\Big] = 0 \quad (35)$$

From equation (34)-(35), Γ_{22} is a diagnoal matrix.

2.2.8 The sub-matrix Γ_{12}

The matrix Γ_{12} with dimension $(k-1)\times(k)$ represents the expectation between the second derivative log-likelihood with respect to p_l and the second derivative log-likelihood with respect to λ_s. Obviously, Γ_{12} is a matrix all of whose entries are zero.

3 EM algorithm for mixture Exponential

3.1 Point Estimate

Assume the underlying individual distribution is Exponential distribution with rate parameter λ_j. We have

$$Y_i|\underset{\sim}{\delta}_i \sim Exp(\lambda_j) \tag{36}$$

The joint distribution of (Y, δ) is

$$f(\underset{\sim}{Y}, \underset{\sim}{\delta}) = \prod_{i=1}^{n} f(\underset{\sim}{Y}_i|\underset{\sim}{\delta}_i) f(\underset{\sim}{\delta}_i) = \prod_{i=1}^{n}\prod_{j=1}^{k} \Big[\big(\lambda_j e^{-\lambda_j y_i}\big)p_j\Big]^{\delta_{ij}} \tag{37}$$

The log-likelihood of (Y, δ) is

$$logf(\underset{\sim}{Y}, \underset{\sim}{\delta}) = \sum_{i=1}^{n}\sum_{j=1}^{k} \delta_{ij}\Big[log\lambda_j - y_i\lambda_j + logp_j\Big] \tag{38}$$

In E-step, the objective function Q is

$$Q(p_j, \lambda_j|p_j^{(t)}, \lambda_j^{(t)}, \underset{\sim}{Y}) = \sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij}^{(t)}(log\lambda_j - y_i\lambda_j + logp_j) \tag{39}$$

where

$$w_{ij}^{(t)} = \frac{f_j^{(t)}p_j^{(t)}}{\sum_{m=1}^{k} f_m^{(t)}p_m^{(t)}} = \frac{\Big[\lambda_j^{(t)}e^{-\lambda_j^{(t)}y_i}\Big]p_j^{(t)}}{\sum_{m=1}^{k}\Big[\lambda_m e^{-\lambda_m^{(t)}y_i}\Big]p_m^{(t)}} \tag{40}$$

In M-step, the maximum likelihood estimate of $p_j^{(t+1)}$ is identical to equations (13). The maximum likelihood estimate of $\lambda_j^{(t+1)}$ is $\frac{\sum_{i=1}^{n} w_{ij}^{(t)}}{\sum_{i=1}^{n} w_{ij}^{(t)} y_i}$.

3.2 Variance Estimate

4 EM algorithm for mixture Rayleigh

4.1 Point Estimate

Assume the underlying individual distribution is Rayleigh distribution with parameter σ_j. We have

$$Y_i|\underset{\sim}{\delta_i} \sim Rayleigh(\sigma_j) \tag{41}$$

The joint distribution of (Y, δ) is

$$f(\underset{\sim}{Y}, \underset{\sim}{\delta}) = \prod_{i=1}^{n} f(\underset{\sim}{Y_i}|\underset{\sim}{\delta_i}) f(\underset{\sim}{\delta_i}) = \prod_{i=1}^{n}\prod_{j=1}^{k}\left[(\frac{y_i}{\sigma_j^2} e^{\frac{-y_i^2}{2\sigma_j^2}})p_j\right]^{\delta_{ij}} \tag{42}$$

The log-likelihood of (Y, δ) is

$$logf(\underset{\sim}{Y}, \underset{\sim}{\delta}) = \sum_{i=1}^{n}\sum_{j=1}^{k}\delta_{ij}\Big[-\frac{y_i^2}{2\sigma_j^2} + logy_i - 2log\sigma_j + logp_j\Big] \tag{43}$$

In E-step, the objective function Q is

$$Q(p_j, \lambda_j|p_j^{(t)}, \lambda_j^{(t)}, \underset{\sim}{Y}) = \sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij}^{(t)}(-\frac{y_i^2}{2\sigma_j^2} + logy_i - 2log\sigma_j + logp_j) \tag{44}$$

where

$$w_{ij}^{(t)} = \frac{f_j^{(t)} p_j^{(t)}}{\sum_{m=1}^{k} f_m^{(t)} p_m^{(t)}} = \frac{\left[\frac{y_i}{\sigma_j^2(t)} e^{\frac{-y_i^2}{2\sigma_j^2(t)}}\right] p_j^{(t)}}{\left[\sum_{m=1}^{k} \frac{y_i}{\sigma_m^2(t)} e^{\frac{-y_i^2}{2\sigma_m^2(t)}}\right] p_m^{(t)}} \tag{45}$$

In M-step, the maximum likelihood estimate of $p_j^{(t+1)}$ is identical to equations (13). The maximum likelihood estimate of $\sigma_j^{(t+1)}$ is $\sqrt{\frac{\sum_{i=1}^{n} w_{ij}^{(t)} y_i^2}{2\sum_{i=1}^{n} w_{ij}^{(t)}}}$.

4.2 Variance Estimate

The detailed point estimation of EM algorithm for each underlying distribution are included in Appendix.