Computational Statistics Final Project: EM for finite mixture models

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

Dec 31 2020

Abstract

Finite mixture models are frequently seen in real-world data applications. To identify appropriate mixture models, we need to isolate the true cluster of the data, tackling the density estimation. EM algorithm is commonly used to obtain the parameter estimates of mixture model. Most of the published R packages, such as **mixture**, **mclust**, and **EMCluster**, were developed primarily for handling mixture Gaussian cases. For cases outside the scope of mixture Gaussian, the relevant R packages are quit limited. To cope with this problem, we develop a new R package **MixPoiRayExp** in this project. **MixPoiRayExp** can provide the results of the point estimation with its precision in mixture Poisson, mixture exponential, and mixture Rayleigh. To help users obtain reliable results, two additional functions, which provides suggestions of the number of mixture components and initialization strategies for the EM algorithm, are also included in **MixPoiRayExp**.

keywords: finite mixture model, EM, mixture Poisson, mixture exponential, mixture Rayleigh

1 Definition of finite mixture

Let the data $y = (y_1, y_2, ...y_n)$ be a sample of n independent, identically distributed observations. If the distribution of the observation y_i follows

$$f(y_i; \Theta) = \sum_{j=1}^{k} p_j f_j(y_i; \theta_j), \tag{1}$$

then the density of y is called a finite mixture model with k components, where $\Theta = (p_1, p_2, ..., p_k, \theta_1, \theta_2, ..., \theta_k)$ are the parameters of the mixture model. $f_j(y_i; \theta_j)$ is the density of the jth component X_j for the observation y_i. The mixture components $(X_1, X_2, ..., X_k)$ have their corresponding densities $(f_1, f_2, ..., f_k)$. The mixture weights are $(p_1, p_2, ..., p_k)$. The sum of the mixture weights should be one, i.e., $\sum_{j=1}^{k} p_j = 1$.

Let the indicator variable be δ, where $\delta_{ij} = 1$ given $y_i \sim f_j$. Obviously, $\underset{\sim}{\delta_i} = (\delta_{i1}, \delta_{i2}, ..., \delta_{ik})$ follows a multinomial distribution with sample size 1 and its proportion parameter are mixture weights. This is given as

$$\underset{\sim}{\delta_i} \sim MN(1; p_1, p_2, .., p_k) \tag{2}$$

2 EM algorithm for finite mixture distribution

2.1 mixture Poisson

2.1.1 Point estimate for mixture Poisson

Let the underlying individual distribution follows Poisson distribution with rate parameter λ_j, namely,

$$Y_i | \underset{\sim}{\delta_i} \sim Poi(\lambda_j) \tag{3}$$

The joint distribution of (Y, δ) is

$$f(\underset{\sim}{Y}, \underset{\sim}{\delta}) = \prod_{i=1}^{n} f(Y_i | \underset{\sim}{\delta_i}) f(\underset{\sim}{\delta_i}) = \prod_{i=1}^{n} \prod_{j=1}^{k} \left[(e^{-\lambda_j} \frac{\lambda_j^{y_i}}{y_i!}) p_j \right]^{\delta_{ij}} \tag{4}$$

and the log-likelihood of (Y, δ) is

$$logf(\underset{\sim}{Y}, \underset{\sim}{\delta}) = \sum_{i=1}^{n}\sum_{j=1}^{k}\delta_{ij}\Big[-\lambda_j + y_i log\lambda_j - y_i! + logp_j\Big] \tag{5}$$

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

$$Q(\theta|\theta^{(0)}, \underset{\sim}{Y}) = E_{\delta_{ij}}\Big[logf(\underset{\sim}{Y}, \underset{\sim}{\delta})\Big] = \sum_{i=1}^{n}\sum_{j=1}^{k}E\Big[\delta_{ij}\Big]\Big[-\lambda_j + y_i log\lambda_j - y_i! + logp_j\Big] \tag{6}$$

Equation (7) shows that, with the observation Y_i, the conditional expectation δ_{ij} at iteration t is

$$E[\delta_{ij}|Y_i] = P(\delta_{ij} = 1|Y_i) = \frac{P(Y_i|\delta_{ij} = 1)P(\delta_{ij} = 1)}{\sum_{j=1}^{k} P(Y_i|\delta_{ij} = 1)P(\delta_{ij} = 1)} = \frac{f_j^{(t)}p_j^{(t)}}{\sum_{m=1}^{k} f_m^{(t)}p_m^{(t)}} = \frac{\Big[e^{-\lambda_j^{(t)}}\frac{\lambda_j^{(t)y_i}}{y_i!}\Big]p_j^{(t)}}{\sum_{m=1}^{k}\Big[e^{-\lambda_m^{(t)}}\frac{\lambda_m^{(t)y_i}}{y_i!}\Big]p_m^{(t)}} = w_{ij}^{(t)} \tag{7}$$

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

$$Q(\theta|\theta^{(0)}, \underset{\sim}{Y}) = Q(p_j, \lambda_j|p_j^{(t)}, \lambda_j^{(t)}, \underset{\sim}{Y}) = \sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij}^{(t)}(-\lambda_j + y_i log\lambda_j - y_i! + logp_j) \tag{8}$$

Furthermore, we maximize Q with respect to θ in the M-step.

For the Poisson rate parameter λ_j, we have

$$\frac{dQ}{d\lambda_j} = \frac{dE_{\delta_{ij}}\Big[logf(\underset{\sim}{Y}, \underset{\sim}{\delta})\Big]}{d\lambda_j} = \sum_{i=1}^{n} w_{ij}^{(t)}\left(\frac{y_i}{\lambda_j} - 1\right) \tag{9}$$

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

$$\lambda_j^{(t+1)} = \frac{\sum_{i=1}^{n} w_{ij}^{(t)} y_i}{\sum_{i=1}^{n} w_{ij}^{(t)}} \tag{10}$$

For the mixture weight parameter p_j, we have

$$\frac{dQ}{dp_j} = \frac{dE_{\delta_{ij}}\Big[logf(\underset{\sim}{Y}, \underset{\sim}{\delta})\Big]}{dp_j} = \sum_{i=1}^{n}\left[\frac{w_{ij}^{(t)}}{p_j} - \frac{w_{ik}^{(t)}}{p_k}\right] \tag{11}$$

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

$$p_j^{(t+1)} = \frac{\sum_{i=1}^{n} w_{ij}^{(t)}}{n} \tag{12}$$

2.1.2 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

$$I(\theta) = -Var\left[\frac{d}{d\theta} logf(\underset{\sim}{Y}, \underset{\sim}{\delta})\right] + E\left[-\frac{d^2}{d\theta^2} logf(\underset{\sim}{Y}, \underset{\sim}{\delta})\right] = -\Psi + \Gamma \tag{13}$$

2.1.2.1 Matrix Ψ

The matrix Ψ, which has dimensions $[(2K-1)\times(2k-1)]$, can be partitioned into 4 sub-matrices, Ψ_{11}, Ψ_{12}, Ψ_{21}, and Ψ_{22}. They are give as follows.

(I) sub-matrix Ψ_{11}

The matrix Ψ_{11}, which has dimensions $[(k-1)\times(k-1)]$, is

$$\Psi_{11} = \begin{pmatrix} Var[\frac{d}{dp_1}logf] & Cov[\frac{d}{dp_1}logf, \frac{d}{dp_2}logf] & ... & ... & Cov[\frac{d}{dp_1}logf, \frac{d}{dp_{(k-1)}}logf] \\ Cov[\frac{d}{dp_2}logf, \frac{d}{dp_1}logf] & Var[\frac{d}{dp_2}logf] & ... & ... & Cov[\frac{d}{dp_2}logf, \frac{d}{dp_{(k-1)}}logf] \\ ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... \\ Cov[\frac{d}{dp_{(k-1)}}logf, \frac{d}{dp_1}logf] & ... & ... & ... & Var[\frac{d}{dp_{(k-1)}}logf] \end{pmatrix}_{(k-1)\times(k-1)} \tag{14}$$

Ψ_{11} 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 entries of Ψ_{11} are

$$\psi_{ll} = Var\left[\frac{d}{dp_l}logf\right] = \sum_{i=1}^{n} Var\left[\left(\frac{\delta_{il}}{p_l} - \frac{\delta_{ik}}{p_k}\right)\right], \quad for\ 1 \le l \le k-1 \tag{15}$$

Given $1 \le j, j^{'} \le k-1$, we have that from the properties of the multinomial distribution

$$E[\delta_{ij}|\underset{\sim}{Y}] = \delta^*_{ij} \tag{16}$$

$$Var[\delta_{ij}|\underset{\sim}{Y}] = \delta^*_{ij}(1-\delta^*_{ij}) \tag{17}$$

$$Cov[\delta_{ij}, \delta_{ij'}] = -\delta^*_{ij}\delta^*_{ij'} \tag{18}$$

$$Cov[\delta_{ij}, \delta_{i'j}] = 0 \tag{19}$$

$$Cov[\delta_{ij}, \delta_{ik}|\underset{\sim}{Y}] = Cov[\delta_{ij}, 1-(\delta_{i1}+\delta_{i2}+...+\delta_{i(k-1)})] = -\sum_{l=1}^{k-1} Cov[\delta_{ij}, \delta_{il}] = -\left[\sum_{l\neq j}(-\delta^*_{ij}\delta^*_{il}) + \delta^*_{ij}(1-\delta^*_{ij})\right] = -\delta^*_{ij}\delta^*_{ik} \tag{20}$$

$$Var[\delta_{ik}|\underset{\sim}{Y}] = Cov[1-(\delta_{i1}+\delta_{i2}+...+\delta_{i(k-1)}), \delta_{ik}] = \sum_{l\neq k} -\delta^*_{il}\delta^*_{ik} = \delta^*_{ik}(1-\delta^*_{ik}) \tag{21}$$

Integrating all the results, the diagonal entries of Ψ_{11} defined in equation (16) can be re-cast as

$$\psi_{ll} = \sum_{i=1}^{n} Var\left[\frac{\delta_{il}}{p_l} - \frac{\delta_{ik}}{p_k}\right] = \sum_{i=1}^{n}\left[\frac{\delta^*_{il}(1-\delta^*_{il})}{p_l^2} + \frac{\delta^*_{ik}(1-\delta^*_{ik})}{p_k^2} + 2\frac{\delta^*_{il}\delta^*_{ik}}{p_l p_k}\right] \tag{22}$$

and the off-diagonal entries of Ψ_{11} are

$$\psi_{ll'} = Cov\left[\frac{d}{dp_l}logf, \frac{d}{dp_{l'}}logf\right] = \sum_{i=1}^{n} Cov\left[(\frac{\delta_{il}}{p_l} - \frac{\delta_{ik}}{p_k}), (\frac{\delta_{il'}}{p_{l'}} - \frac{\delta_{ik}}{p_k})\right] = \sum_{i=1}^{n}\left[\frac{-\delta^*_{il}\delta^*_{il'}}{p_l p_{l'}} + \frac{\delta^*_{il}\delta^*_{ik}}{p_l p_k} + \frac{\delta^*_{ik}\delta^*_{il'}}{p_k p_{l'}} + \frac{\delta^*_{ik}(1-\delta^*_{ik})}{p_k^2}\right] \tag{23}$$

(II) sub-matrix Ψ_{22}

The matrix Ψ_{22}, which has dimensions $(k \times k)$, is

$$\Psi_{22} = \begin{pmatrix} Var[\frac{d}{d\lambda_1}logf] & Cov[\frac{d}{d\lambda_1}logf, \frac{d}{d\lambda_2}logf] & ... & ... & Cov[\frac{d}{d\lambda_1}logf, \frac{d}{d\lambda_k}logf] \\ Cov[\frac{d}{d\lambda_2}logf, \frac{d}{d\lambda_1}logf] & Var[\frac{d}{d\lambda_2}logf] & ... & ... & Cov[\frac{d}{d\lambda_2}logf, \frac{d}{d\lambda_k}logf] \\ ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... \\ Cov[\frac{d}{d\lambda_k}logf, \frac{d}{d\lambda_1}logf] & ... & ... & ... & Var[\frac{d}{d\lambda_k}logf] \end{pmatrix}_{k \times k} \tag{24}$$

Ψ_{22} 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 entries of Ψ_{22} are

$$\psi_{ll} = Var\left[\frac{d}{d\lambda_l}logf\right] = \sum_{i=1}^{n} Var\left[\delta_{il}\left(\frac{y_i}{\lambda_l} - 1\right)\right] = \sum_{i=1}^{n}\left(\frac{y_i}{\lambda_l} - 1\right)^2 Var[\delta_{il}] = \sum_{i=1}^{n}\left(\frac{y_i}{\lambda_l} - 1\right)^2 \delta^*_{il}(1 - \delta^*_{il}) \tag{25}$$

and the off-diagonal entries of Ψ_{22} are

$$\psi_{ll'} = \sum_{i=1}^{n} Cov\left[\delta_{il}\left(\frac{y_i}{\lambda_l} - 1\right), \delta_{il'}\left(\frac{y_i}{\lambda_{l'}} - 1\right)\right] = \sum_{i=1}^{n}\left(\frac{y_i}{\lambda_l} - 1\right)\left(\frac{y_i}{\lambda_{l'}} - 1\right)\left(-\delta^*_{il}\delta^*_{il'}\right) \tag{26}$$

(III) sub-matrix Ψ_{12}

The matrix Ψ_{12}, which has dimensions $[(k-1) \times k]$, is

$$\Psi_{12} = \begin{pmatrix} Cov[\frac{d}{dp_1}logf, \frac{d}{d\lambda_1}logf] & Cov[\frac{d}{dp_1}logf, \frac{d}{d\lambda_2}logf] & ... & ... & Cov[\frac{d}{dp_1}logf, \frac{d}{d\lambda_k}logf] \\ Cov[\frac{d}{dp_2}logf, \frac{d}{d\lambda_1}logf] & Cov[\frac{d}{dp_2}logf, \frac{d}{d\lambda_2}logf] & ... & ... & Cov[\frac{d}{dp_2}logf, \frac{d}{d\lambda_k}logf] \\ ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... \\ Cov[\frac{d}{dp_{(k-1)}}logf, \frac{d}{d\lambda_1}logf] & ... & ... & ... & Cov[\frac{d}{dp_{(k-1)}}logf, \frac{d}{d\lambda_k}logf] \end{pmatrix}_{(k-1) \times k} \tag{27}$$

Ψ_{12} 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 elements of Ψ_{12} are

$$\psi_{ll} = \sum_{i=1}^{n} Cov\left[\left(\frac{\delta_{il}}{p_l} - \frac{\delta_{ik}}{p_k}\right), \delta_{il}\left(\frac{y_i}{\lambda_l} - 1\right)\right] = \sum_{i=1}^{n}\left(\frac{y_i}{\lambda_l} - 1\right)\left[\frac{\delta^*_{il}(1 - \delta^*_{il})}{p_l} + \frac{\delta^*_{il}\delta^*_{ik}}{p_k}\right] \tag{28}$$

and the off-diagonal elements of Ψ_{12} are

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

(IV) sub-matrix Ψ_{21}

Lastly, the matrix Ψ_{21} is the the transpose of the matrix Ψ_{12}, namely,

$$\Psi_{21} = \Psi^t_{12} \tag{30}$$

2.1.2.2 Matrix Γ

Likewise, the matrix Γ, which has dimensions $[(2K-1)\times(2k-1)]$, can be partitioned into 4 sub-matrices, Γ_{11}, Γ_{12}, Γ_{21}, and Γ_{22}.

(I) sub-matrix Γ_{11}

The matrix Γ_{11}, which has dimensions $[(k-1)\times(k-1)]$, 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 entries of Γ_{11} are

$$\gamma_{ll} = E\bigg[-\frac{d^2}{dp_l^2}logf\bigg] = \sum_{i=1}^{n}\bigg[\frac{\delta_{il}^*}{p_l^2}+\frac{\delta_{ik}^*}{p_k^2}\bigg] \tag{31}$$

and the off-diagonal entries of Γ_{11} are

$$\gamma_{ll'} = E\bigg[-\frac{d^2}{dp_l dp_l^{'}}logf\bigg] = \sum_{i=1}^{n}E\bigg[\frac{\delta_{ik}}{p_k^2}\bigg] = \sum_{i=1}^{n}\frac{\delta_{ik}^*}{p_k^2} \tag{32}$$

(2) sub-matrix Γ_{22}

The matrix Γ_{22}, which has dimensions $(k\times k)$, 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 entries of Γ_{22} are

$$\gamma_{ll} = E\bigg[-\frac{d^2}{d\lambda_l^2}logf\bigg] = \sum_{i=1}^{n}E\bigg[\frac{y_i}{\lambda_l^2}\delta_{il}\bigg] = \sum_{i=1}^{n}\frac{y_i}{\lambda_l^2}\delta_{il}^* \tag{33}$$

and the off-diagonal entries of Γ_{11} are

$$\gamma_{ll'} = E\bigg[-\frac{d^2}{d\lambda_l d\lambda_l^{'}}logf\bigg] = E\bigg[0\bigg] = 0 \tag{34}$$

which clearly shows that Γ_{22} is a diagonal matrix.

(3) sub-matrix Γ_{12}

The matrix Γ_{12}, which has dimensions $(k-1)\times k$, 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 whose entries are all zero.

2.1.2.3 The covariance related to p_k

The covariance estimate for mixture distribution discussed above does not include the parameter p_k due to the constraint of the mixture weight. Our R package **MixPoiRayExp** includes the estimated results related to p_k based on the following calculation to help users have the whole covariance estimation.

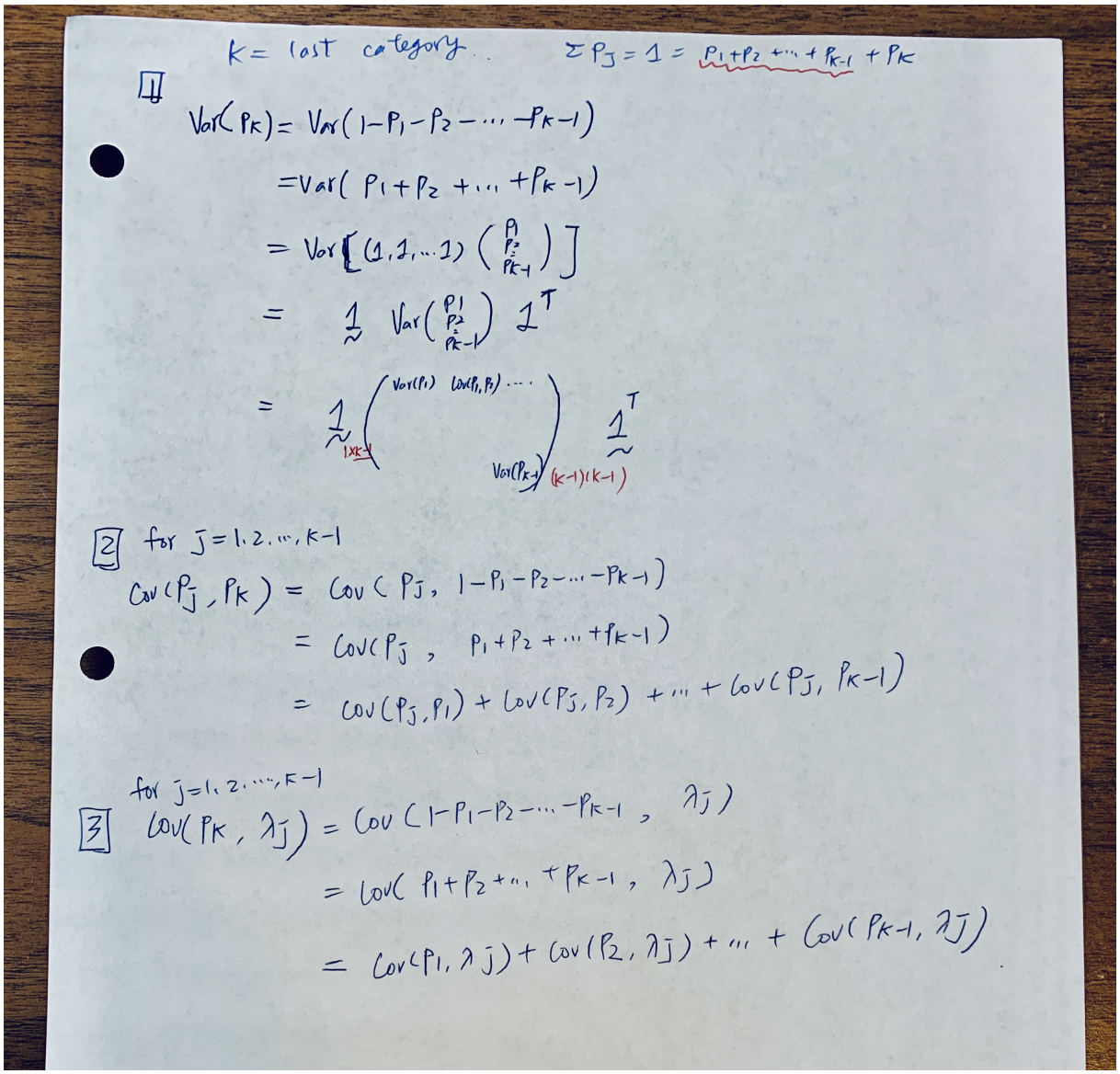


Figure 1: covariance for p_k

2.2 EM algorithm for mixture exponential

2.2.1 Point Estimate for mixture exponential

Assume that the underlying individual distribution follows an exponential distribution with rate parameter λ_j. That is,

$$Y_i|\underset{\sim}{\delta}_i \sim Exp(\lambda_j) \tag{35}$$

The joint distribution of (Y, δ) is

$$f(\underset{\sim}{Y}, \underset{\sim}{\delta}) = \prod_{i=1}^{n} f(\underset{\sim}{Y}_i|\underset{\sim}{\delta}_i) f(\underset{\sim}{\delta}_i) = \prod_{i=1}^{n}\prod_{j=1}^{k} \left[\left(\lambda_j e^{-\lambda_j y_i}\right) p_j \right]^{\delta_{ij}} \tag{36}$$

and the log-likelihood of (Y, δ) is

$$logf(\underset{\sim}{Y}, \underset{\sim}{\delta}) = \sum_{i=1}^{n}\sum_{j=1}^{k} \delta_{ij} \left[log\lambda_j - y_i\lambda_j + logp_j \right] \tag{37}$$

In E-step, the objective function Q is

$$Q(p_j, \lambda_j|p_j^{(t)}, \lambda_j^{(t)}, \underset{\sim}{Y}) = \sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij}^{(t)} (log\lambda_j - y_i\lambda_j + logp_j) \tag{38}$$

where

$$w_{ij}^{(t)} = \frac{f_j^{(t)} p_j^{(t)}}{\sum_{m=1}^{k} f_m^{(t)} p_m^{(t)}} = \frac{\left[\lambda_j^{(t)} e^{-\lambda_j^{(t)} y_i}\right] p_j^{(t)}}{\sum_{m=1}^{k} \left[\lambda_m e^{-\lambda_m^{(t)} y_i}\right] p_m^{(t)}} \tag{39}$$

In M-step, the maximum likelihood estimate of $p_j^{(t+1)}$ is equivalent to the one defined in equations (12). The maximum likelihood estimate of $\lambda_j^{(t+1)}$ is $\frac{\sum_{i=1}^{n} w_{ij}^{(t)}}{\sum_{i=1}^{n} w_{ij}^{(t)} y_i}$.

2.2.2 Variance estimate for mixture exponential

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

$$I(\theta) = -Var\left[\frac{d}{d\theta} logf(\underset{\sim}{Y}, \underset{\sim}{\delta})\right] + E\left[-\frac{d^2}{d\theta^2} logf(\underset{\sim}{Y}, \underset{\sim}{\delta})\right] = -\Psi + \Gamma \tag{40}$$

2.2.2.1 Matrix Ψ

The matrix Ψ, which has dimensions $[(2k-1) \times (2k-1)]$, can be partitioned into 4 sub-matrices, Ψ_{11}, Ψ_{12}, Ψ_{21}, and Ψ_{22}. They are illustrated below.

(I) sub-matrix Ψ_{11}

The matrix Ψ_{11}, which has dimensions $[(k-1) \times (k-1)]$, is

$$\Psi_{11} = \begin{pmatrix} Var[\frac{d}{dp_1} logf] & Cov[\frac{d}{dp_1} logf, \frac{d}{dp_2} logf] & ... & ... & Cov[\frac{d}{dp_1} logf, \frac{d}{dp_{(k-1)}} logf] \\ Cov[\frac{d}{dp_2} logf, \frac{d}{dp_1} logf] & Var[\frac{d}{dp_2} logf] & ... & ... & Cov[\frac{d}{dp_2} logf, \frac{d}{dp_{(k-1)}} logf] \\ ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... \\ Cov[\frac{d}{dp_{(k-1)}} logf, \frac{d}{dp_1} logf] & ... & ... & ... & Var[\frac{d}{dp_{(k-1)}} logf] \end{pmatrix}_{(k-1)\times(k-1)} \tag{41}$$

Ψ_{11} 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 entries of Ψ_{11} are

$$\psi_{ll} = Var\left[\frac{d}{dp_l} logf\right] = \sum_{i=1}^{n} Var\left[\left(\frac{\delta_{il}}{p_l} - \frac{\delta_{ik}}{p_k}\right)\right], \quad for\ 1 \le l \le k-1 \tag{42}$$

Given $1 \le j, j^{'} \le k-1$ and employing the properties of the multinomial distribution, the following identities hold

$$E[\delta_{ij}|\underset{\sim}{Y}] = \delta^*_{ij} \tag{43}$$

$$Var[\delta_{ij}|\underset{\sim}{Y}] = \delta^*_{ij}(1-\delta^*_{ij}) \tag{44}$$

$$Cov[\delta_{ij}, \delta_{ij'}] = -\delta^*_{ij}\delta^*_{ij'} \tag{45}$$

$$Cov[\delta_{ij}, \delta_{i'j}] = 0 \tag{46}$$

$$Cov[\delta_{ij}, \delta_{ik}|\underset{\sim}{Y}] = Cov[\delta_{ij}, 1-(\delta_{i1}+\delta_{i2}+...+\delta_{i(k-1)})] = -\sum_{l=1}^{k-1} Cov[\delta_{ij}, \delta_{il}] = -\left[\sum_{l\neq j}(-\delta^*_{ij}\delta^*_{il}) + \delta^*_{ij}(1-\delta^*_{ij})\right] = -\delta^*_{ij}\delta^*_{ik} \tag{47}$$

$$Var[\delta_{ik}|\underset{\sim}{Y}] = Cov[1-(\delta_{i1}+\delta_{i2}+...+\delta_{i(k-1)}), \delta_{ik}] = \sum_{l\neq k} -\delta^*_{il}\delta^*_{ik} = \delta^*_{ik}(1-\delta^*_{ik}) \tag{48}$$

Combining all the results, the diagonal entries of Ψ_{11} defined in equation (43) can be rewritten as

$$\psi_{ll} = \sum_{i=1}^{n} Var\left[\frac{\delta_{il}}{p_l} - \frac{\delta_{ik}}{p_k}\right] = \sum_{i=1}^{n}\left[\frac{\delta^*_{il}(1-\delta^*_{il})}{p_l^2} + \frac{\delta^*_{ik}(1-\delta^*_{ik})}{p_k^2} + 2\frac{\delta^*_{il}\delta^*_{ik}}{p_l p_k}\right] \tag{49}$$

and and the off-diagonal entries of Ψ_{11} are

$$\psi_{ll'} = Cov\left[\frac{d}{dp_l} logf, \frac{d}{dp_{l'}} logf\right] = \sum_{i=1}^{n} Cov\left[(\frac{\delta_{il}}{p_l} - \frac{\delta_{ik}}{p_k}), (\frac{\delta_{il'}}{p_{l'}} - \frac{\delta_{ik}}{p_k})\right] = \sum_{i=1}^{n}\left[\frac{-\delta^*_{il}\delta^*_{il'}}{p_l p_{l'}} + \frac{\delta^*_{il}\delta^*_{ik}}{p_l p_k} + \frac{\delta^*_{ik}\delta^*_{il'}}{p_k p_{l'}} + \frac{\delta^*_{ik}(1-\delta^*_{ik})}{p_k^2}\right] \tag{50}$$

(II) sub-matrix Ψ_{22}

The matrix Ψ_{22}, which has dimensions $(k \times k)$, is

$$\Psi_{22} = \begin{pmatrix} Var[\frac{d}{d\lambda_1}logf] & Cov[\frac{d}{d\lambda_1}logf, \frac{d}{d\lambda_2}logf] & ... & ... & Cov[\frac{d}{d\lambda_1}logf, \frac{d}{d\lambda_k}logf] \\ Cov[\frac{d}{d\lambda_2}logf, \frac{d}{d\lambda_1}logf] & Var[\frac{d}{d\lambda_2}logf] & ... & ... & Cov[\frac{d}{d\lambda_2}logf, \frac{d}{d\lambda_k}logf] \\ ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... \\ Cov[\frac{d}{d\lambda_k}logf, \frac{d}{d\lambda_1}logf] & ... & ... & ... & Var[\frac{d}{d\lambda_k}logf] \end{pmatrix}_{k \times k} \tag{51}$$

Ψ_{22} 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 entries of Ψ_{22} are

$$\psi_{ll} = Var\left[\frac{d}{d\lambda_l}logf\right] = \sum_{i=1}^{n} Var\left[\delta_{il}\left(\frac{1}{\lambda_l} - y_i\right)\right] = \sum_{i=1}^{n}\left(\frac{1}{\lambda_l} - y_i\right)^2 Var[\delta_{il}] = \sum_{i=1}^{n}\left(\frac{1}{\lambda_l} - y_i\right)^2 \delta^*_{il}(1 - \delta^*_{il}) \tag{52}$$

and the off-diagonal entries of Ψ_{22} are

$$\psi_{ll'} = \sum_{i=1}^{n} Cov\left[\delta_{il}\left(\frac{1}{\lambda_l} - y_i\right), \delta_{il'}\left(\frac{1}{\lambda_{l'}} - y_i\right)\right] = \sum_{i=1}^{n}\left(\frac{1}{\lambda_l} - y_i\right)\left(\frac{1}{\lambda_{l'}} - y_i\right)\left(-\delta^*_{il}\delta^*_{il'}\right) \tag{53}$$

(III) sub-matrix Ψ_{12}

The matrix Ψ_{12}, which has dimensions $[(k-1) \times k]$ is,

$$\Psi_{12} = \begin{pmatrix} Cov[\frac{d}{dp_1}logf, \frac{d}{d\lambda_1}logf] & Cov[\frac{d}{dp_1}logf, \frac{d}{d\lambda_2}logf] & ... & ... & Cov[\frac{d}{dp_1}logf, \frac{d}{d\lambda_k}logf] \\ Cov[\frac{d}{dp_2}logf, \frac{d}{d\lambda_1}logf] & Cov[\frac{d}{dp_2}logf, \frac{d}{d\lambda_2}logf] & ... & ... & Cov[\frac{d}{dp_2}logf, \frac{d}{d\lambda_k}logf] \\ ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... \\ Cov[\frac{d}{dp_{(k-1)}}logf, \frac{d}{d\lambda_1}logf] & ... & ... & ... & Cov[\frac{d}{dp_{(k-1)}}logf, \frac{d}{d\lambda_k}logf] \end{pmatrix}_{(k-1)\times k} \tag{54}$$

Ψ_{12} 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 entries of Ψ_{12} are

$$\psi_{ll} = \sum_{i=1}^{n} Cov\left[\left(\frac{\delta_{il}}{p_l} - \frac{\delta_{ik}}{p_k}\right), \delta_{il}\left(\frac{1}{\lambda_l} - y_i\right)\right] = \sum_{i=1}^{n}\left(\frac{1}{\lambda_l} - y_i\right)\left[\frac{\delta^*_{il}(1 - \delta^*_{il})}{p_l} + \frac{\delta^*_{il}\delta^*_{ik}}{p_k}\right] \tag{55}$$

and the off-diagonal entries of Ψ_{12} are

$$\psi_{ll'} = \sum_{i=1}^{n} Cov\left[\left(\frac{\delta_{il}}{p_l} - \frac{\delta_{ik}}{p_k}\right), \delta_{il'}\left(\frac{1}{\lambda_{l'}} - y_i\right)\right] = \sum_{i=1}^{n}\left(\frac{1}{\lambda'_l} - y_i\right)\left[-\frac{\delta^*_{il}\delta^*_{il'}}{p_l} + \frac{\delta^*_{ik}\delta^*_{il'}}{p_k}\right] \tag{56}$$

(IV) sub-matrix Ψ_{21}

Lastly, the matrix Ψ_{21} is the transpose of the matrix Ψ_{21}, i.e.,

$$\Psi_{21} = \Psi^t_{12} \tag{57}$$

2.2.2.2 Matrix Γ

The matrix Γ, which has dimensions $[(2K-1)\times(2k-1)]$, can be decomposed into 4 sub-matrices, Γ_{11}, Γ_{12}, Γ_{21}, and Γ_{22}. They are given as follows

(I) sub-matrix Γ_{11}

The matrix Γ_{11}, which has dimensions $[(k-1)\times(k-1)]$, 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 entries of Γ_{11} are

$$\gamma_{ll} = E\bigg[-\frac{d^2}{dp_l^2}logf\bigg] = \sum_{i=1}^{n}\bigg[\frac{\delta_{il}^*}{p_l^2}+\frac{\delta_{ik}^*}{p_k^2}\bigg] \tag{58}$$

and the off-diagonal entries of Γ_{11} are

$$\gamma_{ll'} = E\bigg[-\frac{d^2}{dp_l dp_l^{'}}logf\bigg] = \sum_{i=1}^{n}E\bigg[\frac{\delta_{ik}}{p_k^2}\bigg] = \sum_{i=1}^{n}\frac{\delta_{ik}^*}{p_k^2} \tag{59}$$

(II) sub-matrix Γ_{22}

The matrix Γ_{22}, which has dimensions $(k\times k)$, 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 elements of Γ_{22} are

$$\gamma_{ll} = E\bigg[-\frac{d^2}{d\lambda_l^2}logf\bigg] = \sum_{i=1}^{n}E\bigg[\frac{1}{\lambda_l^2}\delta_{il}\bigg] = \sum_{i=1}^{n}\frac{1}{\lambda_l^2}\delta_{il}^* \tag{60}$$

and the off-diagonal elements of Γ_{11} are

$$\gamma_{ll'} = E\bigg[-\frac{d^2}{d\lambda_l d\lambda_l^{'}}logf\bigg] = E\bigg[0\bigg] = 0 \tag{61}$$

which indicates that Γ_{22} is a diagonal matrix.

(III) sub-matrix Γ_{12}

The matrix Γ_{12}, which has dimensions $[(k-1)\times k]$, 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 whose all entries are zero.

2.3 EM algorithm for mixture Rayleigh

2.3.1 Point estimate for mixture Rayleigh

Assume the underlying individual distribution is Rayleigh distribution with parameter σ_j, which is given as

$$Y_i|\underset{\sim}{\delta_i} \sim Rayleigh(\sigma_j) \tag{62}$$

and the joint distribution of (Y, δ) is

$$f(\underset{\sim}{Y}, \underset{\sim}{\delta}) = \prod_{i=1}^{n} f(\underset{\sim}{Y_i}|\underset{\sim}{\delta_i}) f(\underset{\sim}{\delta_i}) = \prod_{i=1}^{n}\prod_{j=1}^{k} \left[(\frac{y_i}{\sigma_j^2} e^{\frac{-y_i^2}{2\sigma_j^2}}) p_j \right]^{\delta_{ij}} \tag{63}$$

The log-likelihood of (Y, δ) is

$$logf(\underset{\sim}{Y}, \underset{\sim}{\delta}) = \sum_{i=1}^{n}\sum_{j=1}^{k} \delta_{ij} \left[-\frac{y_i^2}{2\sigma_j^2} + logy_i - 2log\sigma_j + logp_j \right] \tag{64}$$

In E-step, the objective function Q is

$$Q(p_j, \lambda_j | p_j^{(t)}, \lambda_j^{(t)}, \underset{\sim}{Y}) = \sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij}^{(t)} (-\frac{y_i^2}{2\sigma_j^2} + logy_i - 2log\sigma_j + logp_j) \tag{65}$$

where

$$w_{ij}^{(t)} = \frac{f_j^{(t)} p_j^{(t)}}{\sum_{m=1}^{k} f_m^{(t)} p_m^{(t)}} = \frac{\left[\frac{y_i}{\sigma_j^2(t)} e^{\frac{-y_i^2}{2\sigma_j^2(t)}} \right] p_j^{(t)}}{\left[\sum_{m=1}^{k} \frac{y_i}{\sigma_m^2(t)} e^{\frac{-y_i^2}{2\sigma_m^2(t)}} \right] p_m^{(t)}} \tag{66}$$

In M-step, the maximum likelihood estimate of $p_j^{(t+1)}$ is identical to the one given in equation (12). The maximum likelihood estimate of $\sigma_j^{(t+1)}$ is $\sqrt{\frac{\sum_{i=1}^{n} w_{ij}^{(t)} y_i^2}{2\sum_{i=1}^{n} w_{ij}^{(t)}}}$.

2.3.2 Variance estimate for mixture Rayleigh

The variance estimation of the parameters can be derived from Louis formula. The information matrix, $I(\theta) = I(p_1, p_2, ...p_{k-1}, \sigma_1, \sigma_2, ...\sigma_k)$, is

$$I(\theta) = -Var\left[\frac{d}{d\theta} logf(\underset{\sim}{Y}, \underset{\sim}{\delta})\right] + E\left[-\frac{d^2}{d\theta^2} logf(\underset{\sim}{Y}, \underset{\sim}{\delta})\right] = -\Psi + \Gamma \tag{67}$$

2.3.2.1 Matrix Ψ

The matrix Ψ, which has dimensions $[(2K-1)\times(2k-1)]$, can be decomposed into 4 sub-matrices, Ψ_{11}, Ψ_{12}, Ψ_{21}, Ψ_{22}.

(I) sub-matrix Ψ_{11}

The matrix Ψ_{11}, which has dimensions $[(k-1)\times(k-1)]$, is

$$\Psi_{11} = \begin{pmatrix} Var[\frac{d}{dp_1} logf] & Cov[\frac{d}{dp_1} logf, \frac{d}{dp_2} logf] & ... & ... & Cov[\frac{d}{dp_1} logf, \frac{d}{dp_{(k-1)}} logf] \\ Cov[\frac{d}{dp_2} logf, \frac{d}{dp_1} logf] & Var[\frac{d}{dp_2} logf] & ... & ... & Cov[\frac{d}{dp_2} logf, \frac{d}{dp_{(k-1)}} logf] \\ ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... \\ Cov[\frac{d}{dp_{(k-1)}} logf, \frac{d}{dp_1} logf] & ... & ... & ... & Var[\frac{d}{dp_{(k-1)}} logf] \end{pmatrix}_{(k-1)\times(k-1)} \tag{68}$$

Ψ_{11} 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 entries of Ψ_{11} are

$$\psi_{ll} = Var\left[\frac{d}{dp_l} logf\right] = \sum_{i=1}^{n} Var\left[\left(\frac{\delta_{il}}{p_l} - \frac{\delta_{ik}}{p_k}\right)\right], \quad for\ 1 \le l \le k-1 \tag{69}$$

Suppose $1 \le j, j^{'} \le k-1$, we then have from the properties of the multinomial distribution

$$E[\delta_{ij}|\underset{\sim}{Y}] = \delta^*_{ij} \tag{70}$$

$$Var[\delta_{ij}|\underset{\sim}{Y}] = \delta^*_{ij}(1-\delta^*_{ij}) \tag{71}$$

$$Cov[\delta_{ij}, \delta_{ij'}] = -\delta^*_{ij}\delta^*_{ij'} \tag{72}$$

$$Cov[\delta_{ij}, \delta_{i'j}] = 0 \tag{73}$$

$$Cov[\delta_{ij}, \delta_{ik}|\underset{\sim}{Y}] = Cov[\delta_{ij}, 1-(\delta_{i1}+\delta_{i2}+...+\delta_{i(k-1)})] = -\sum_{l=1}^{k-1} Cov[\delta_{ij}, \delta_{il}] = -\left[\sum_{l\neq j}(-\delta^*_{ij}\delta^*_{il}) + \delta^*_{ij}(1-\delta^*_{ij})\right] = -\delta^*_{ij}\delta^*_{ik} \tag{74}$$

$$Var[\delta_{ik}|\underset{\sim}{Y}] = Cov[1-(\delta_{i1}+\delta_{i2}+...+\delta_{i(k-1)}), \delta_{ik}] = \sum_{l\neq k} -\delta^*_{il}\delta^*_{ik} = \delta^*_{ik}(1-\delta^*_{ik}) \tag{75}$$

Integrating all the results, the diagonal entries of Ψ_{11} defined in equation (70) can be re-cast as

$$\psi_{ll} = \sum_{i=1}^{n} Var\left[\frac{\delta_{il}}{p_l} - \frac{\delta_{ik}}{p_k}\right] = \sum_{i=1}^{n}\left[\frac{\delta^*_{il}(1-\delta^*_{il})}{p_l^2} + \frac{\delta^*_{ik}(1-\delta^*_{ik})}{p_k^2} + 2\frac{\delta^*_{il}\delta^*_{ik}}{p_l p_k}\right] \tag{76}$$

and the off-diagonal entries of Ψ_{11} are

$$\psi_{ll'} = Cov\left[\frac{d}{dp_l} logf, \frac{d}{dp_{l'}} logf\right] = \sum_{i=1}^{n} Cov\left[(\frac{\delta_{il}}{p_l} - \frac{\delta_{ik}}{p_k}), (\frac{\delta_{il'}}{p_{l'}} - \frac{\delta_{ik}}{p_k})\right] = \sum_{i=1}^{n}\left[\frac{-\delta^*_{il}\delta^*_{il'}}{p_l p_{l'}} + \frac{\delta^*_{il}\delta^*_{ik}}{p_l p_k} + \frac{\delta^*_{ik}\delta^*_{il'}}{p_k p_{l'}} + \frac{\delta^*_{ik}(1-\delta^*_{ik})}{p_k^2}\right] \tag{77}$$

(II) sub-matrix Ψ_{22}

The matrix Ψ_{22}, which has dimensions $(k \times k)$, is

$$\Psi_{22} = \begin{pmatrix} Var[\frac{d}{d\sigma_1}logf] & Cov[\frac{d}{d\sigma_1}logf, \frac{d}{d\sigma_2}logf] & ... & ... & Cov[\frac{d}{d\sigma_1}logf, \frac{d}{d\sigma_k}logf] \\ Cov[\frac{d}{d\sigma_2}logf, \frac{d}{d\sigma_1}logf] & Var[\frac{d}{d\sigma_2}logf] & ... & ... & Cov[\frac{d}{d\sigma_2}logf, \frac{d}{d\sigma_k}logf] \\ ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... \\ Cov[\frac{d}{d\sigma_k}logf, \frac{d}{d\sigma_1}logf] & ... & ... & ... & Var[\frac{d}{d\sigma_k}logf] \end{pmatrix}_{k \times k} \tag{78}$$

Ψ_{22} 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 entries of Ψ_{22} are

$$\psi_{ll} = Var\left[\frac{d}{d\sigma_l}logf\right] = \sum_{i=1}^{n} Var\left[\delta_{il}\left(-\frac{2}{\sigma_l} + \frac{y_i^2}{\sigma_l^3}\right)\right] = \sum_{i=1}^{n}\left(-\frac{2}{\sigma_l} + \frac{y_i^2}{\sigma_l^3}\right)^2 Var[\delta_{il}] = \sum_{i=1}^{n}\left(-\frac{2}{\sigma_l} + \frac{y_i^2}{\sigma_l^3}\right)^2 \delta_{il}^*(1-\delta_{il}^*) \tag{79}$$

and the off-diagonal entries of Ψ_{22} are

$$\psi_{ll'} = \sum_{i=1}^{n} Cov\left[\delta_{il}\left(-\frac{2}{\sigma_l} + \frac{y_i^2}{\sigma_l^3}\right), \delta_{il'}\left(-\frac{2}{\sigma_{l'}} + \frac{y_i^2}{\sigma_{l'}^3}\right)\right] = \sum_{i=1}^{n}\left(-\frac{2}{\sigma_l} + \frac{y_i^2}{\sigma_l^3}\right)\left(-\frac{2}{\sigma_{l'}} + \frac{y_i^2}{\sigma_{l'}^3}\right)\left(-\delta_{il}^*\delta_{il'}^*\right) \tag{80}$$

(III) sub-matrix Ψ_{12}

The matrix Ψ_{12}, which has dimensions $[(k-1) \times k]$, is

$$\Psi_{12} = \begin{pmatrix} Cov[\frac{d}{dp_1}logf, \frac{d}{d\sigma_1}logf] & Cov[\frac{d}{dp_1}logf, \frac{d}{d\sigma_2}logf] & ... & ... & Cov[\frac{d}{dp_1}logf, \frac{d}{d\sigma_k}logf] \\ Cov[\frac{d}{dp_2}logf, \frac{d}{d\sigma_1}logf] & Cov[\frac{d}{dp_2}logf, \frac{d}{d\sigma_2}logf] & ... & ... & Cov[\frac{d}{dp_2}logf, \frac{d}{d\sigma_k}logf] \\ ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... \\ Cov[\frac{d}{dp_{(k-1)}}logf, \frac{d}{d\sigma_1}logf] & ... & ... & ... & Cov[\frac{d}{dp_{(k-1)}}logf, \frac{d}{d\sigma_k}logf] \end{pmatrix}_{(k-1)\times k} \tag{81}$$

Ψ_{12} 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 elements of Ψ_{12} are

$$\psi_{ll} = \sum_{i=1}^{n} Cov\left[\left(\frac{\delta_{il}}{p_l} - \frac{\delta_{ik}}{p_k}\right), \delta_{il}\left(-\frac{2}{\sigma_l} + \frac{y_i^2}{\sigma_l^3}\right)\right] = \sum_{i=1}^{n}\left(-\frac{2}{\sigma_l} + \frac{y_i^2}{\sigma_l^3}\right)\left[\frac{\delta_{il}^*(1-\delta_{il}^*)}{p_l} + \frac{\delta_{il}^*\delta_{ik}^*}{p_k}\right] \tag{82}$$

and the off-diagonal elements of Ψ_{12} are

$$\psi_{ll'} = \sum_{i=1}^{n} Cov\left[\left(\frac{\delta_{il}}{p_l} - \frac{\delta_{ik}}{p_k}\right), \delta_{il'}\left(-\frac{2}{\sigma_{l'}} + \frac{y_i^2}{\sigma_{l'}^3}\right)\right] = \sum_{i=1}^{n}\left(-\frac{2}{\sigma_{l'}} + \frac{y_i^2}{\sigma_{l'}^3}\right)\left[-\frac{\delta_{il}^*\delta_{il'}^*}{p_l} + \frac{\delta_{ik}^*\delta_{il'}^*}{p_k}\right] \tag{83}$$

(IV) sub-matrix Ψ_{21}

The matrix Ψ_{21} is the transpose of the matrix Ψ_{12}, i.e.,

$$\Psi_{21} = \Psi_{12}^t \tag{84}$$

2.3.2.2 The matrix Γ

The matrix Γ, which has dimensions $[(2k-1)\times(2k-1)]$, can be decomposed into 4 sub-matrices, Γ_{11}, Γ_{12}, Γ_{21}, Γ_{22}.

(I) sub-matrix Γ_{11}

The matrix Γ_{11}, which has dimensions $[(k-1)\times(k-1)]$, 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 entries of Γ_{11} are

$$\gamma_{ll} = E\bigg[-\frac{d^2}{dp_l^2}logf\bigg] = \sum_{i=1}^{n}\bigg[\frac{\delta_{il}^*}{p_l^2}+\frac{\delta_{ik}^*}{p_k^2}\bigg] \tag{85}$$

and the off-diagonal entries of Γ_{11} are

$$\gamma_{ll'} = E\bigg[-\frac{d^2}{dp_l dp_l^{'}}logf\bigg] = \sum_{i=1}^{n}E\bigg[\frac{\delta_{ik}}{p_k^2}\bigg] = \sum_{i=1}^{n}\frac{\delta_{ik}^*}{p_k^2} \tag{86}$$

(II) sub-matrix Γ_{22}

The matrix Γ_{22}, which has dimensions $(k\times k)$, 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 entries of Γ_{22} are

$$\gamma_{ll} = E\bigg[-\frac{d^2}{d\sigma_l^2}logf\bigg] = \sum_{i=1}^{n}E\bigg[\bigg(-\frac{2}{\sigma_l^2}+\frac{3y_i^2}{\sigma_l^4}\bigg)\delta_{il}\bigg] = \sum_{i=1}^{n}\bigg(-\frac{2}{\sigma_l^2}+\frac{3y_i^2}{\sigma_l^4}\bigg)\delta_{il}^* \tag{87}$$

and the off-diagonal entries of Γ_{11} are

$$\gamma_{ll'} = E\bigg[-\frac{d^2}{d\sigma_l d\sigma_l^{'}}logf\bigg] = E\bigg[0\bigg] = 0 \tag{88}$$

which shows that Γ_{22} is a diagonal matrix.

(III) sub-matrix Γ_{12}

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

3 The choice of starting value

EM algorithm has some drawbacks, such as the sensitivity to initial values and the possibility of being trapped in local optima. Nevertheless, due to its appealing properties, EM plays an important role in estimating the parameters of mixture models. Finite mixture modeling can suffer from locally optimal solutions, and the final parameter estimates are dependent on the initial starting values of the EM algorithm (Shireman et al., 2017). Good choice of initial values can help to reach the global maximum in fewer iterations and converge faster. In this section, some initialization strategies for selecting the set of starting values for mixture models will be discussed. These initialization strategies are very commonly implemented in software (R packages "mclust" and "mixture", the statistical computing software LatentGOLD and Mplus).

3.1 Random starting values

To initialize a mixture model, each observation is randomly classified to one of k clusters and each cluster contains an equal number of observations. The first E-step of EM algorithm is to estimate parameters (both the group-specific mean vectors and covariance matrices) from this initial cluster. Due to random assignment, it's recommended to start many times to generalize. Several criteria have been proposed and the effect of stopping early has been examined (Seidel et al., 2000).EM algorithm stops ierating when the value of the chosen criterion becomes smaller than a specified constant. The smaller this constant, the more severe the criterion.

Some alternative ways of generating random starting values including randomly picking starting parameter values from a uniform distribution with bounds in each cluster; using only one observation per cluster, etc. However, it's found that randomly drawing parameters from respective distributions had poor performance (Karlis and Xekalaki, 2003). Although random starting values methods are comparatively easy to apply, the fact that many random starts must be implemented to settle on a solution is its main weakness, and not easy to decide when the number of initialization will be sufficient.

3.2 Iteratively constrained EM

To construct the search space for global optimum, several constrained EM algorithms initialized with different random starting values will be run iteratively. Among those, the quickest increase in likelihood will be considered as better initial values in EM algorithm (Lubke and Muthén, 2007). Some drawbacks of this methods including the choice of number of iterations in initializing EM, the value chosen could be sufficient for some likelihood functions, but perhaps not others (Biernacki et al., 2003). Additionally, run number of initial-stage EM algorithms based on different starting values will greatly increase the amount of time required in computation.

3.3 K-means clustering

The k-means clustering is a technique relies on the center of cluster. This is often represented by the average of cluster. The clustering measure the similarity of the group by iterating the measurement distance between each point and the center of each cluster using Euclidean distance measuring. The results from the k-means algorithm will be served as initial values for the EM algorithm, which can provide close to accurate parameters.This method is implemented in R package "mixture" and has been recommended by (McLachlan and Peel, 2004) as variable initialization strategy. However, when the clusters are highly heterogeneous and non-spherical, the results from a k-means clustering may not be accurate and thus may not provide adequate starting values. In addition, it's subject to local optima under certain conditions as well. Also the number of initialization in k-means is quite a subjective determination depending on researcher's choice.

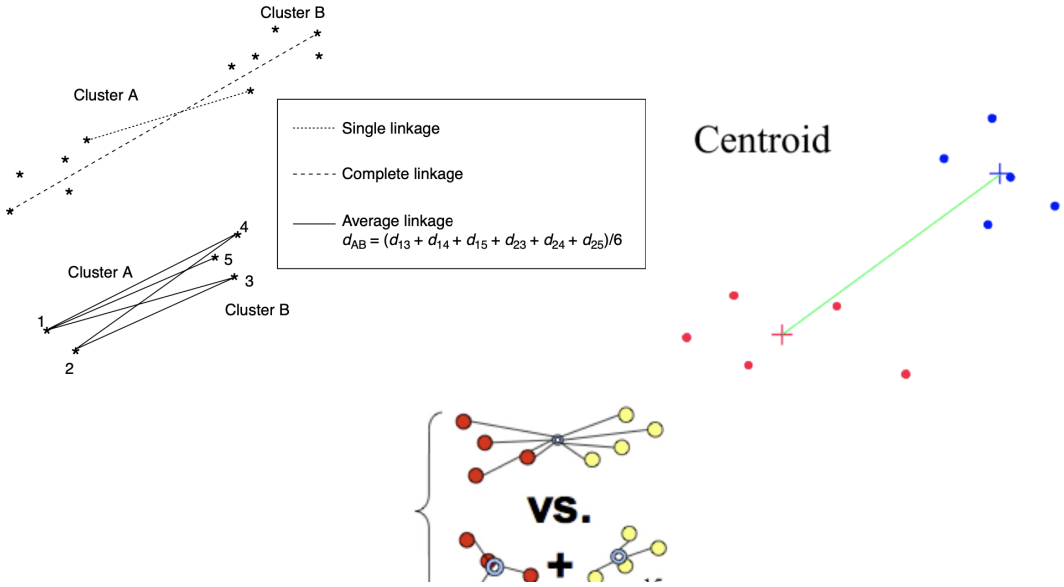
3.4 Agglomerative hierarchical clustering

In a hierarchical classification the data are not partitioned into a particular number of classes or clusters at a single step. Instead it consists of a series of partitions. Agglomerative methods is a hierarchical clustering technique, proceeded by a series of successive fusions of the n individuals into groups. The idea is to ensure nearby points end up in the same cluster. Start with a collection C of n singleton clusters with each cluster contains one data point. At each stage the method find a pair of clusters that is closest, merge the clusters c_i, c_j into a new cluster $c_(i+j)$ and then remove c_i, c_j from the collection C, add $c_(i+j)$. Repeat until only one cluster is left. A hierarchical cluster analysis is performed and a dendrogram will be produced to show the hierarchical tree of clusters.

There are different measures for cluster distance. Single link is the distance between closet elements in clusters and it produces long chains. Complete link is the distance between farthest elements in clusters and forces 'spherical'

clusters with consistent 'diameter'. Average link is the average of all pairwise distances so it will be less affected by outliers compared to single and complete link. Centroids is the distance between centroids (means) of two clusters. Ward's method considers joining two clusters and new centroids will be the mean of two joined clusters together. Total deviation is defined as the total distance from centroid of each point in the cluster. Then by checking the amount of increase of total deviation in combined clusters vs. total deviation in seperate clusters, the combination with smallest total deviation increase indicate best cluster combination. See below for visual illustration of these measures.

Figure 2: Cluster distance measures



There is R package "mclust" utilizing this hierarchical clustering for computing starting values in the first E-step of mixture model estimation (Fraley and Raftery, 2006).Hierarchical clustering maybe an accurate way to describe data organized hierarchically but may not perform well in other data structures.

3.5 Sum scores

Sum score as a representation of the data, will be used in initializing class memberships. For example, in a cognitive ability assessment study, the sum score for each individual would be the sum of each item score. Based on the sum score calculated, the data are split into k equally sized ordered groups, which are used for the M-step of the EM algorithm. This method is simple and quick to implement, but sum score may not be an adequate representation of the data, thus clusters made will not be nearly accurate. It's recommended that any method advanced by researchers for scoring scales needs evidence to support its use, and considering sum scores as a factor model demands such evidence (McNeish and Wolf, 2020). However, this technique is fairly fast and takes a very short time to calculate. Also due to only one possible set of starting values, there is no need for several iterations for initialization.

4 The choice of mixture components

The choice of the number of components can be a difficult problem in cluster analysis. Considering the finite mixture models we assumed, there is a variety of model selection methods for us to compare models with different components, including Bayesian information criterion (BIC), Akaike information criterion (AIC), likelihood-ratio test, entropy criterion, and etc. Furthermore, Richardson (Richardson and Green, 1997)proposed that fully Bayesian mixture modelling can be appropriate way to estimates the number of components. Here, we discussed four methods to choose the number of components for our EM algorithms.

4.1 Akaike information criterion

Bozdogan and Sclove (1984) (Bozdogan and Sclove, 1984)studied the Akaike information criterion (Akaike 1974) (Akaike, 1974)in the mixture context. AIC takes the form

$$AIC(K) = -2L(K) + 2\upsilon(K)$$

where $\upsilon(K)$ is the number of free parameters in the mixture model with K components and $L(K)$ is the maximized value of the likelihood function of the model with K components. However, some authors (e.g. Koehler and Murphree 1988 (Koehler and Murphree, 1988)) found that AIC has a problem of order inconsistent and tends to overestimate the correct number of components in the mixture context.

4.2 Bayesian information criterion

Schwarz (1978) (Schwarz et al., 1978)proposed the Bayesian information which is an approximation to the exact Bayes solution in model selection problem. BIC takes the form

$$BIC(K) = -2L(K) + \upsilon(K)ln(n)$$

where $\upsilon(K)$ is the number of free parameters in the mixture model with K components; $L(K)$ is the maximized value of the likelihood function of the model with K components; n is the number of the observations. Compared to AIC, it has been proved that BIC is order consistent in suitable conditions and it is a better criterion for our problem. However, using BIC tends underestimate the number of components.

4.3 Likelihood-ratio test

The likelihood ratio test statistic λ can be used to find the smallest model with the smallest number of components which is also consistent with the data. Unfortunately, with mixture, the test statistic $-2log\lambda$ does not follow chi-squared distribution asymptotically and it's also hard to derive. As an alternative, McLachlan (1987)(McLachlan, 1987) suggested the bootstrap could be employed to find the distribution of $-2log\lambda$ under the null hypothesis.

4.4 Entropy criterion

Celeux (1996) (Celeux and Soromenho, 1996) proposed an entropy criterion for selecting the number of components in a mixture model. He derived this criterion from the maximum likelihood and the classification likelihood of a mixture. Entropy criterion shows a favorable results compared AIC and BIC. Entropy criterion takes the form

$$t_{ik} = \frac{p_k f(\boldsymbol{y_i}, \boldsymbol{\theta_k})}{\sum_{j=1}^{K} p_j f(\boldsymbol{y_i}, \boldsymbol{\theta_k}))}$$

$$E(K) = -\sum_{k=1}^{K}\sum_{i=1}^{n} t_{ik} lnt_{ik}$$

$$NEC(K) = \frac{E(K)}{L(K) - L(1)}$$

where t_{ik} is the conditional probability that y_i is from the k th cluster; the E(K) is the entropy term which measures the overlap of mixture components; L(K) is log likelihood of component K; NEC(K) is the entropy criterion.

5 R package performance

The package **MixPoiRayExp**, which was developed based on EM algorithm, is used to provide parameter estimation of mixture distribution. In this package, the three functions `mixture_poisson`, `mixture_rayleigh`, and `mixture_exponential` are employed to produce the results of point estimate and variance estimate for the parameters. Users can also use the function `find_start_value` of **MixPoiRayExp** to choose the starting values pertaining to the k-means clustering, agglomerative hierarchical clustering, and the method of sum scores. If users are somewhat unfamiliar with the numbers of mixture components, the function `decide_component`, which was developed based on BIC criteria, can be used to help them make final decisions.

To evaluate the performance of **MixPoiRayExp**, we simulated 1,000 datasets with sample size 1,000 for mixture distribution with 3 mixture components. The accuracy and precision of the simulation replications are accessed. Tables 1-3 show the point estimate, the accuracy, and the precision of the parameters.

In these three tables, the true values (abbreviated as TV) are the true parameters. The average estimates (or AE) and the standard deviations of the estimates (or SD) are calculated across these 1,000 replications. Let the true parameter be θ, and the estimated value be $\hat{\theta}$. For checking the accuracy, the relative bias (RB), defined as $E[\frac{\hat{\theta}-\theta}{\theta}] \times 100\%$, and the standardized bias (SB), defined as $E\left[\frac{|\hat{\theta}-\theta|}{SD(\hat{\theta})}\right] \times 100\%$, are used. Furthermore, to have hybrid measure of accuracy and precision, the root mean square error (RMSE) of θ, defined as $\sqrt{E[\hat{\theta}-\theta]^2}$, is also employed.

As indicated by the simulation results, the parameter is well-estimated for mixture exponential and mixture Rayleigh. However, the performance in mixture Poisson is not stable because the data displays two peaks. If the true mixture weight is equally distributed, using the method of sum scores yields the the result at the fastest speed. This is followed by k-means clustering, and agglomerative hierarchical clustering has the slowest performance.

Table 1: Poisson distribution with 3 mixture components (sample size = 1,000)

Initial value	Efficiency	Time	Iteration	Parameter	TV	AE	SD	RB	SB	RMSE
k-means	Simulations	1000		weight	0.3333	0.4610	0.1627	38.3014	78.4576	0.2068
	Min	0.1400	24		0.3333	0.2717	0.0983	-18.4897	62.6719	0.1160
	Mean	22.7452	3159		0.3333	0.2673	0.1040	-19.8115	63.4967	0.1232
	SD	157.7525	19282	parameter	0.5000	1.3628	1.0966	172.5676	78.6851	1.3949
	Median	0.3000	35		5.0000	21.4969	21.0511	329.9375	78.3658	26.7367
	Max	3081.7200	374671		50.0000	50.8673	3.0870	1.7346	28.0950	3.2051
hierarchical	Simulations	1000		weight	0.3333	0.6523	0.0672	95.6930	474.6005	0.3260
	Min	0.1000	11		0.3333	0.1914	0.1177	-42.5811	120.5666	0.1844
	Mean	72.5805	9913		0.3333	0.1563	0.1203	-53.1116	147.2041	0.2140
	SD	533.0929	64523	parameter	0.5000	2.6560	0.4593	431.2075	469.3998	2.2044
	Median	3.7900	541		5.0000	46.2282	9.3998	824.5646	438.6071	42.2852
	Max	13795.3600	1720132		50.0000	52.6566	5.4060	5.3132	49.1414	6.0211
sum scores	Simulations	1000		weight	0.3333	0.3335	0.0178	0.0372	0.6964	0.0178
	Min	0.0900	15		0.3333	0.3323	0.0179	-0.3073	5.7168	0.0179
	Mean	0.2141	28		0.3333	0.3342	0.0151	0.2701	5.9559	0.0151
	SD	0.0567	4	parameter	0.5000	0.5001	0.0585	0.0185	0.1578	0.0585
	Median	0.2000	28		5.0000	5.0052	0.1478	0.1032	3.4912	0.1478
	Max	0.5500	48		50.0000	50.0081	0.3834	0.0162	2.1135	0.3833

Table 2: Exponential distribution with 3 mixture components (sample size = 1,000)

Initial Value	Efficiency	Time	Iteration	Parameter	TV	AE	SD	RB	SB	RMSE
k-means	Simulations	1000		weight	0.3333	0.3319	0.0244	-0.4292	5.8735	0.0244
	Min	0.4300	89		0.3333	0.3352	0.0287	0.5713	6.6448	0.0287
	Mean	0.8673	155		0.3333	0.3329	0.0268	-0.1421	1.7641	0.0268
	SD	0.2532	35	parameter	0.5000	0.5005	0.0363	0.1016	1.3984	0.0363
	Median	0.8300	150		5.0000	5.0352	0.6991	0.7047	5.0400	0.6997
	Max	2.5300	382		50.0000	50.8718	5.6256	1.7437	15.4976	5.6900
hierarchical	Simulations	1000		weight	0.3333	0.3319	0.0244	-0.4291	5.8731	0.0244
	Min	0.5300	105		0.3333	0.3352	0.0287	0.5713	6.6448	0.0287
	Mean	0.9965	175		0.3333	0.3329	0.0268	-0.1421	1.7641	0.0268
	SD	0.2919	42	parameter	0.5000	0.5005	0.0363	0.1016	1.3984	0.0363
	Median	0.9400	167		5.0000	5.0352	0.6991	0.7047	5.0399	0.6997
	Max	2.7400	555		50.0000	50.8718	5.6256	1.7437	15.4976	5.6900
sum scores	Simulations	1000		weight	0.3333	0.3319	0.0244	-0.4286	5.8665	0.0244
	Min	0.2000	41		0.3333	0.3352	0.0287	0.5716	6.6479	0.0287
	Mean	0.5889	105		0.3333	0.3329	0.0268	-0.1429	1.7742	0.0268
	SD	0.1917	27	parameter	0.5000	0.5005	0.0363	0.1019	1.4025	0.0363
	Median	0.5500	100		5.0000	5.0353	0.6991	0.7065	5.0528	0.6997
	Max	1.6500	275		50.0000	50.8724	5.6257	1.7448	15.5077	5.6902

6 Discussion

We encountered some negative variance estimates produced although it was low percentage of all simulation cases. According to our literature review the negative variance may occur when improper initial values for variance components are used, data characteristics such as low redundancy or extreme outliers, violation of linearity, actual value of variance close to zero or improper stochastic model (El Leithy et al., 2016) (Wang and Wu, 2020). While some researchers choose to replace the negative variance with zero, there are modified variance component estimates(VCE) approaches suggesting to add some restrictions through EM procedures such as Modified REML (MREML) (Thompson et al., 1962), Modified Minimum Norm Quadratic Unbiased Estimation (MMINQUE) (Subramani, 2012) (Rao, 1972) and Modified Iterative Almost Unbiased Estimation IAUE (MIAUE) (Rao, 1970) (El Leithy et al., 2016). Still some of the methods produce negative variances although the chance is critically reduced, and this methodology area has not been studied enough with widely accepted solutions.

Since this may become another big research area that requires substantial amount of theoretical development, in our package building project we have decided that, when the negative variance issue occurs to present the warning message and suggest better starting values. Based on our simulation results the Sum of Scores for the initial value finding approach seems to work mostly well. This is because we thought negative variance may occur by the nature of EM limitation to generate correct estimates when the initial value is not proper.

Table 3: Rayleigh distribution with 3 mixture components (sample size = 1,000)

Initial Value	Efficiency	Time	Iteration	Parameter	TV	AE	SD	RB	SB	RMSE
k-means	Simulations	1000		weight	0.3333	0.3410	0.0487	2.2903	15.6920	0.0492
	Min	0.2700	44		0.3333	0.3323	0.0163	-0.3097	6.3434	0.0163
	Mean	10.0765	1250		0.3333	0.3267	0.0488	-1.9811	13.5452	0.0492
	SD	174.7226	22961	parameter	0.5000	0.5635	0.3981	12.7052	15.9589	0.4029
	Median	1.2400	140		5.0000	6.0545	6.6461	21.0891	15.8657	6.7260
	Max	5450.7200	718995		50.0000	50.1951	3.9954	0.3902	4.8831	3.9982
hierarchical	Simulations	1000		weight	0.3333	0.3413	0.0495	2.3828	16.0512	0.0501
	Min	0.1700	26		0.3333	0.3328	0.0168	-0.1506	2.9849	0.0168
	Mean	24.0865	2723		0.3333	0.3259	0.0528	-2.2329	14.0876	0.0533
	SD	383.6347	41985	parameter	0.5000	0.5659	0.4046	13.1733	16.2787	0.4097
	Median	1.8600	207		5.0000	6.0964	6.7697	21.9279	16.1956	6.8545
	Max	11309.7000	1228588		50.0000	50.1919	3.9971	0.3839	4.8018	3.9997
sum scores	Simulations	1000		weight	0.3333	0.3333	0.0158	0.0034	0.0720	0.0158
	Min	0.0900	16		0.3333	0.3325	0.0163	-0.2435	4.9927	0.0163
	Mean	0.1877	19		0.3333	0.3341	0.0159	0.2395	5.0170	0.0159
	SD	0.0541	2	parameter	0.5000	0.4999	0.0164	-0.0142	0.4352	0.0164
	Median	0.1800	19		5.0000	4.9912	0.1667	-0.1750	5.2499	0.1668
	Max	0.5300	27		50.0000	50.0055	1.3642	0.0110	0.4042	1.3635

References

Emilie Shireman, Douglas Steinley, and Michael J Brusco. Examining the effect of initialization strategies on the performance of gaussian mixture modeling. *Behavior research methods*, 49(1):282–293, 2017.

Wilfried Seidel, Karl Mosler, and Manfred Alker. A cautionary note on likelihood ratio tests in mixture models. *Annals of the Institute of Statistical Mathematics*, 52(3):481–487, 2000.

Dimitris Karlis and Evdokia Xekalaki. Choosing initial values for the em algorithm for finite mixtures. *Computational Statistics & Data Analysis*, 41(3-4):577–590, 2003.

Gitta Lubke and Bengt O Muthén. Performance of factor mixture models as a function of model size, covariate effects, and class-specific parameters. *Structural Equation Modeling*, 14(1):26–47, 2007.

Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics & Data Analysis*, 41(3-4): 561–575, 2003.

Geoffrey J McLachlan and David Peel. *Finite mixture models*. John Wiley & Sons, 2004.

Chris Fraley and Adrian E Raftery. Mclust version 3: an r package for normal mixture modeling and model-based clustering. Technical report, WASHINGTON UNIV SEATTLE DEPT OF STATISTICS, 2006.

Daniel McNeish and Melissa Gordon Wolf. Thinking twice about sum scores. *Behavior Research Methods*, pages 1–19, 2020.

Sylvia Richardson and Peter J Green. On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)*, 59(4):731–792, 1997.

Hamparsum Bozdogan and Stanley L Sclove. Multi-sample cluster analysis using akaike's information criterion. *Annals of the Institute of Statistical Mathematics*, 36(1):163–180, 1984.

Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6): 716–723, 1974.

Anne B Koehler and Emily S Murphree. A comparison of the akaike and schwarz criteria for selecting model order. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 37(2):187–195, 1988.

Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

Geoffrey J McLachlan. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 36(3):318–324, 1987.

Gilles Celeux and Gilda Soromenho. An entropy criterion for assessing the number of clusters in a mixture model. *Journal of classification*, 13(2):195–212, 1996.

Heba A El Leithy, Zakaria A Abdel Wahed, and Mohamed S Abdallah. On non-negative estimation of variance components in mixed linear models. *Journal of advanced research*, 7(1):59–68, 2016.

Leyang Wang and Qiwen Wu. Non-negative variance component estimation for the partial eiv model by the expectation maximization algorithm. *Geomatics, Natural Hazards and Risk*, 11(1):1278–1298, 2020.

William A Thompson et al. The problem of negative estimates of variance components. *The Annals of Mathematical Statistics*, 33(1):273–289, 1962.

J Subramani. On modified minimum variance quadratic unbiased estimation (mivque) of variance components in mixed linear models. *Model assisted statistics and applications*, 7(3):179–200, 2012.

C Radhakrishna Rao. Estimation of variance and covariance components in linear models. *Journal of the American Statistical Association*, 67(337):112–115, 1972.

C Radhakrishna Rao. Estimation of heteroscedastic variances in linear models. *Journal of the American Statistical Association*, 65(329):161–172, 1970.