Statistical Computing Homework 4

110024516 邱繼賢

Problem 1.

Assume that *waiting time* (Y_i) follows the mixture normal distribution as below

$$Y_i \ \sim \ \tau \ N\left(\mu_1 \ , \ \sigma_1^2\right) \ + \ (1-\tau) \ N\left(\mu_2 \ , \ \sigma_2^2\right)$$

can be represented as

$$Y_{i1} \ \sim \ N\left(\mu_1 \ , \ \sigma_1^2\right) \ ; \ Y_{i2} \ \sim \ N\left(\mu_2 \ , \ \sigma_2^2\right) \ ,$$

$$\Rightarrow \ Y_i \ = \ \gamma_i \ Y_{i1} \ + \ (1-\gamma_i) \ Y_{i2} \ , \ \gamma_i \ \overset{iid}{\sim} \ Ber(\tau)$$

thus we have to estimate the parameters : $\theta \ = \ \left(\tau \ , \ \mu_1 \ , \ \sigma_1^2 \ , \ \mu_2 \ , \ \sigma_2^2\right)$

Log-likelihood based on the complete data :

$$\log L\left(\theta \mid Y, \gamma\right) \ = \ \sum_{i=1}^{n} \left\{\gamma_i \ \log f\left(Y_i \ ; \ \mu_1, \sigma_1^2\right) \ + \ (1-\gamma_i) \ \log f\left(Y_i \ ; \ \mu_2, \sigma_2^2\right)\right\} + \sum_{i=1}^{n} \left\{\gamma_i \ \log(\tau) \ + \ (1-\gamma_i) \ \log(1-\tau)\right\}$$

<u>*EM Algorithm*</u> :

(1) compute

$$\hat{\gamma}_i^{(t)} \ = \ E_{\hat{\theta}^{(t)}}\left[\gamma_i \mid Y_i\right] \ = \ \frac{\hat{\tau}^{(t)} \ f\left(Y_i \ ; \ \hat{\mu}_1^{(t)} \ , \ \hat{\sigma}_1^{2(t)}\right)}{\hat{\tau}^{(t)} \ f\left(Y_i \ ; \ \hat{\mu}_1^{(t)} \ , \ \hat{\sigma}_1^{2(t)}\right) \ + \ \left(1-\hat{\tau}^{(t)}\right) \ f\left(Y_i \ ; \ \hat{\mu}_2^{(t)} \ , \ \hat{\sigma}_2^{2(t)}\right)}$$

(2) E-step :

$$Q\left(\theta \mid \hat{\theta}^{(t)}\right) \ = \ E_{\hat{\theta}^{(t)}}\left[\log L\left(\theta \mid Y \ , \ \hat{\gamma}^{(t)}\right) \ \mid Y\right]$$

$$= \sum_{i=1}^{n} \left\{\hat{\gamma}_i^{(t)} \ \log f\left(Y_i \ ; \ \mu_1, \sigma_1^2\right) \ + \ \left(1-\hat{\gamma}_i^{(t)}\right) \log f\left(Y_i \ ; \ \mu_2, \sigma_2^2\right)\right\} \ + \ \sum_{i=1}^{n} \left\{\hat{\gamma}_i^{(t)} \ \log(\tau) \ + \ \left(1-\hat{\gamma}_i^{(t)}\right) \log(1-\tau)\right\}$$

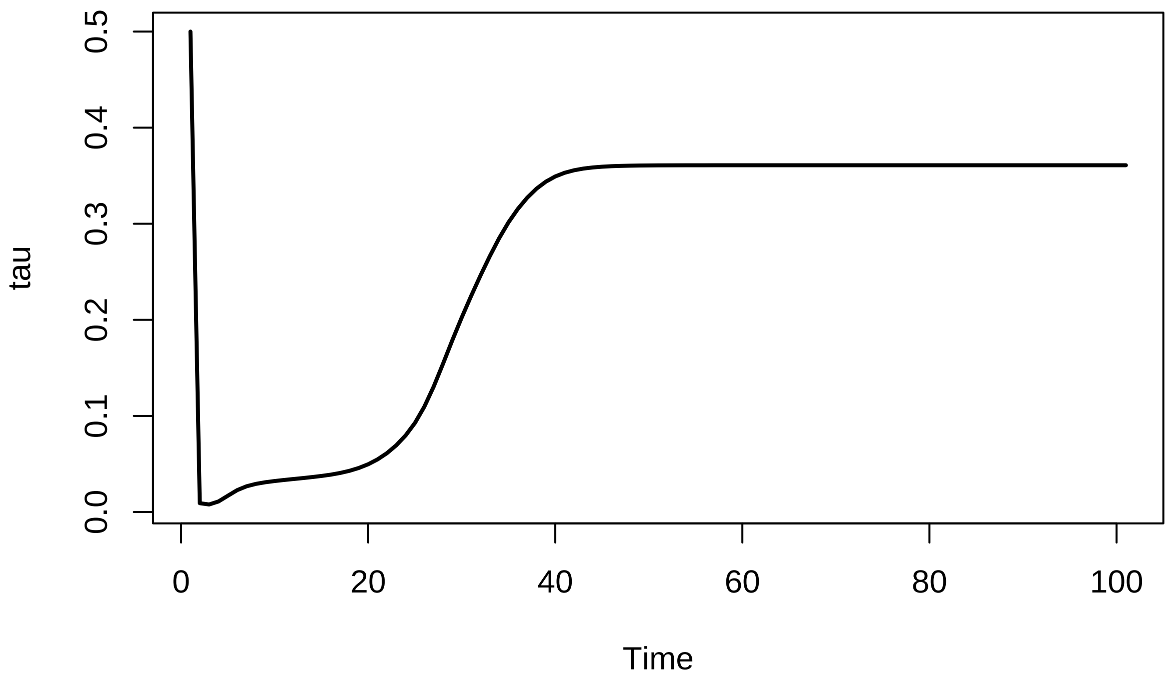
(3) M-step : $\hat{\theta}^{(t+1)} \ = \ \arg \ max_\theta \ Q\left(\theta \mid \hat{\theta}^{(t)}\right)$

$$\Rightarrow \begin{cases} \hat{\mu}_1^{(t+1)} \ = \ \dfrac{\sum_{i=1}^n \hat{\gamma}_i^{(t)} Y_i}{\sum_{i=1}^n \hat{\gamma}_i^{(t)}} \ , \ \hat{\sigma}_1^{2(t+1)} \ = \ \dfrac{\sum_{i=1}^n \hat{\gamma}_i^{(t)} \left(Y_i - \hat{\mu}_1^{(t+1)}\right)^2}{\sum_{i=1}^n \hat{\gamma}_i^{(t)}} \\ \hat{\mu}_2^{(t+1)} \ = \ \dfrac{\sum_{i=1}^n \left(1-\hat{\gamma}_i^{(t)}\right) Y_i}{\sum_{i=1}^n \left(1-\hat{\gamma}_i^{(t)}\right)} \ , \ \hat{\sigma}_2^{2(t+1)} \ = \ \dfrac{\sum_{i=1}^n \left(1-\hat{\gamma}_i^{(t)}\right)\left(Y_i - \hat{\mu}_2^{(t+1)}\right)^2}{\sum_{i=1}^n \left(1-\hat{\gamma}_i^{(t)}\right)} \\ \hat{\tau}^{(t+1)} \ = \ \dfrac{\sum_{i=1}^n \hat{\gamma}_i^{(t)}}{n} \end{cases}$$

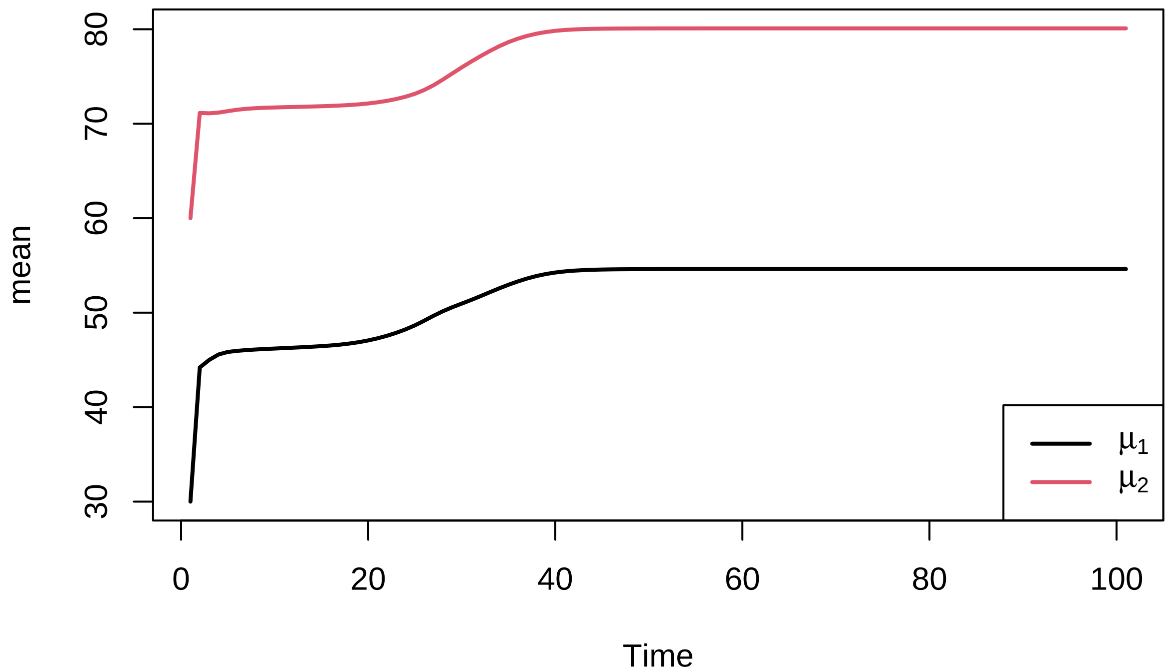
(4) Check convergence $||\hat{\theta}^{(t+1)} - \hat{\theta}^{(t)}|| \rightarrow 0$

Take initial parameter $\hat{\theta}^{(1)} = (0.5\ ,\ 30\ ,\ 1\ ,\ 60\ ,\ 1)$ for example, and iterate 100 times

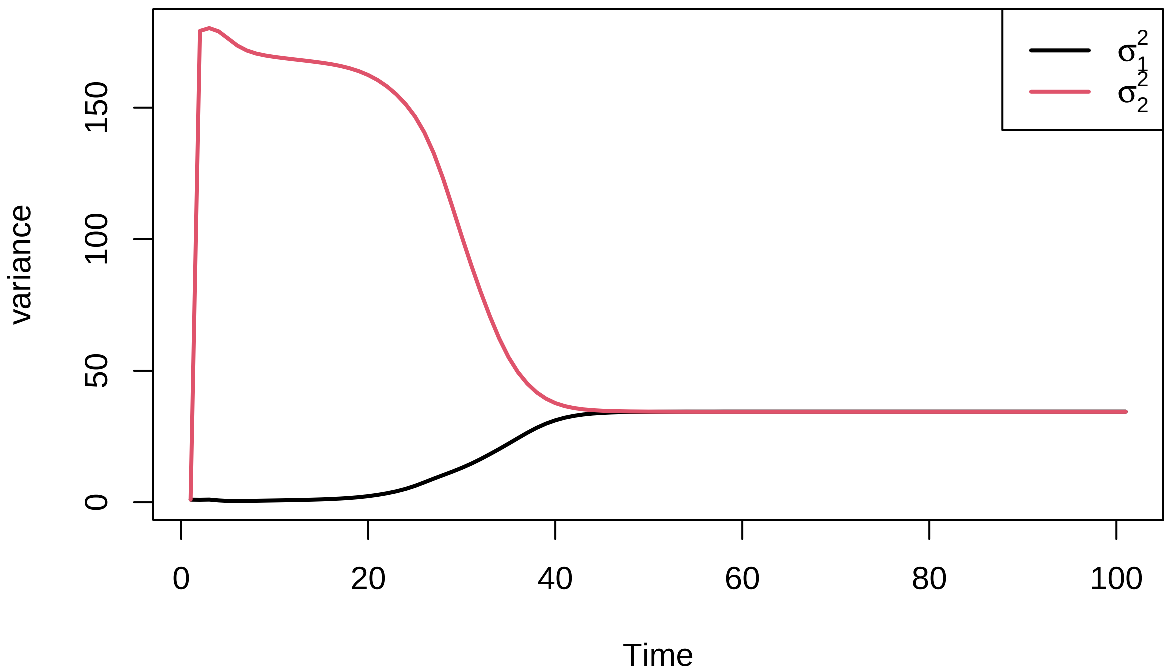
Check the convergence of the five parameters :



$\hat{\tau}$ converges to a stable value after about 60th iteration



$(\hat{\mu}_1\ ,\ \hat{\mu}_2)$ converge to two stable values after about 60th iteration

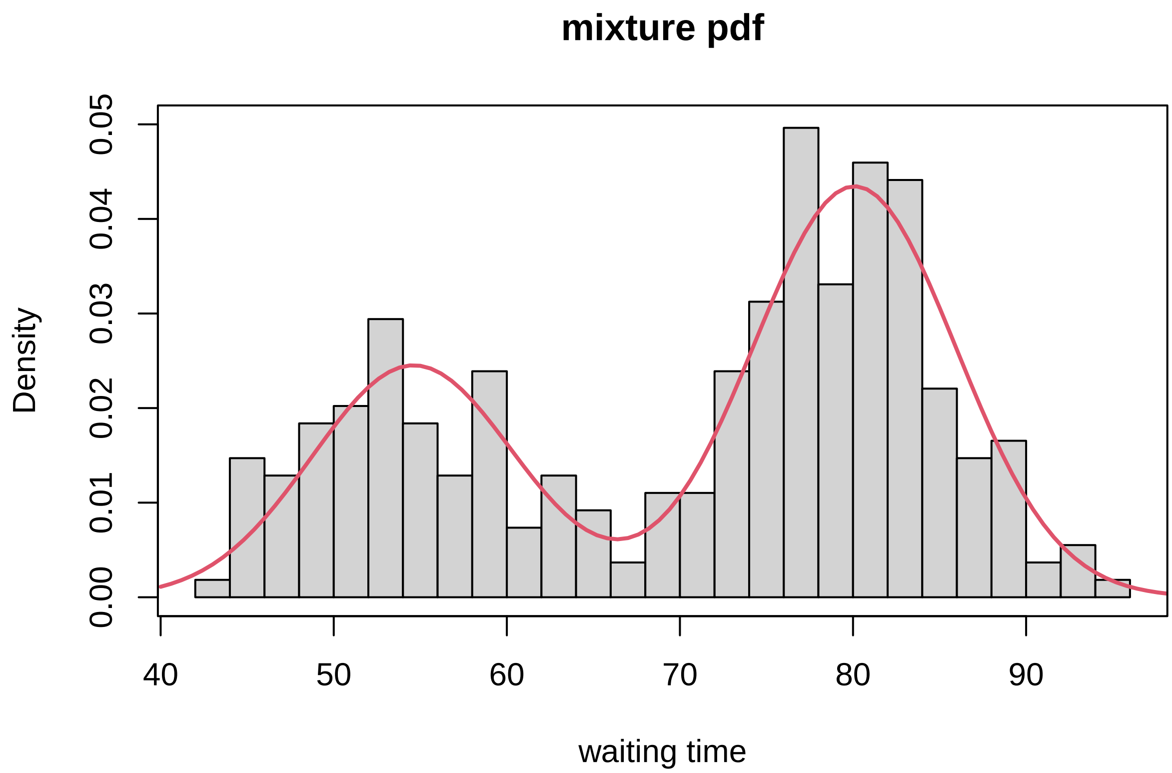


$(\hat{\sigma}_1^2\ ,\ \hat{\sigma}_2^2)$ converge to two stable values after about 60th iteration

Let's see the final estimation of the parameters

	$\hat{\tau}$	$\hat{\mu}_1$	$\hat{\sigma}_1^2$	$\hat{\mu}_2$	$\hat{\sigma}_2^2$
101	0.361	54.615	34.471	80.091	34.43

Sketch the mixture pdf curve by the above parameters and compare to the histogram of *waiting time*



They fit pretty well!