
1: Parallel Computing for EM Alogorithm (40%)

The EM algorithm in the question:

Given initial guess: $\pi_1^{(0)}, \pi_2^{(0)}, \mu_1^{(0)}, \mu_2^{(0)}, \mu_3^{(0)}, \sigma_1^{(0)}, \sigma_2^{(0)}, \sigma_3^{(0)}$, for $t \geq 0$ and $t \in \mathbb{Z}$:

E – step: Calculate $E(Z_i^{(t)} | \Theta^{(t)})$, where $\Theta^{(t)} = \pi_1^{(t)}, \pi_2^{(t)}, \mu_1^{(t)}, \mu_2^{(t)}, \mu_3^{(t)}, \sigma_1^{(t)}, \sigma_2^{(t)}, \sigma_3^{(t)}$.

$$\begin{aligned} \widehat{Z}_{ik}^{(t)} &= E(Z_i = k | \Theta^{(t)}) = E(Z_i^{(t)} | \pi_1^{(t)}, \pi_2^{(t)}, \mu_1^{(t)}, \mu_2^{(t)}, \mu_3^{(t)}, \sigma_1^{(t)}, \sigma_2^{(t)}, \sigma_3^{(t)}) \\ &= \frac{\pi_k^{(t)} \frac{1}{\sqrt{2\pi}\sigma_k^{(t)}} e^{-\frac{(y_i - \mu_k^{(t)})^2}{2\sigma_k^{(t)2}}}}{\pi_1^{(t)} \frac{1}{\sqrt{2\pi}\sigma_1^{(t)}} e^{-\frac{(y_i - \mu_1^{(t)})^2}{2\sigma_1^{(t)2}}} + \pi_2^{(t)} \frac{1}{\sqrt{2\pi}\sigma_2^{(t)}} e^{-\frac{(y_i - \mu_2^{(t)})^2}{2\sigma_2^{(t)2}}} + (1 - \pi_1^{(t)} - \pi_2^{(t)}) \frac{1}{\sqrt{2\pi}\sigma_3^{(t)}} e^{-\frac{(y_i - \mu_3^{(t)})^2}{2\sigma_3^{(t)2}}}} \end{aligned}$$

M – step: Update $\Theta^{(t+1)}$ by equations (1) to (8) listed at the next page.

Stopping criterion: $|L(\Theta^{(t)} | \mathbf{Y})) - L(\Theta^{(T+1)} | \mathbf{Y}))| < \text{tolerance}$.

Iterative scheme:

$$\pi_1^{(t+1)} = \frac{\sum_{i=1}^n \widehat{Z}_{i1}^{(t)}}{\sum_{i=1}^n \widehat{Z}_{i1}^{(t)} + \sum_{i=1}^n \widehat{Z}_{i2}^{(t)} + \sum_{i=1}^n \widehat{Z}_{i3}^{(t)}} \quad (1)$$

$$\pi_2^{(t+1)} = \frac{\sum_{i=1}^n \widehat{Z}_{i2}^{(t)}}{\sum_{i=1}^n \widehat{Z}_{i1}^{(t)} + \sum_{i=1}^n \widehat{Z}_{i2}^{(t)} + \sum_{i=1}^n \widehat{Z}_{i3}^{(t)}} \quad (2)$$

$$\mu_1^{(t+1)} = \frac{\sum_{i=1}^n \widehat{Z}_{i1}^{(t)} y_i}{\sum_{i=1}^n \widehat{Z}_{i1}^{(t)}} \quad (3)$$

$$\mu_2^{(t+1)} = \frac{\sum_{i=1}^n \widehat{Z}_{i2}^{(t)} y_i}{\sum_{i=1}^n \widehat{Z}_{i2}^{(t)}} \quad (4)$$

$$\mu_3^{(t+1)} = \frac{\sum_{i=1}^n \widehat{Z}_{i3}^{(t)} y_i}{\sum_{i=1}^n \widehat{Z}_{i3}^{(t)}} \quad (5)$$

$$\sigma_1^{2(t+1)} = \frac{\sum_{i=1}^n \widehat{Z}_{i1}^{(t)} (y_i - \mu_1^{(t)})^2}{\sum_{i=1}^n \widehat{Z}_{i1}^{(t)}} \quad (6)$$

$$\sigma_2^{2(t+1)} = \frac{\sum_{i=1}^n \widehat{Z}_{i2}^{(t)} (y_i - \mu_2^{(t)})^2}{\sum_{i=1}^n \widehat{Z}_{i2}^{(t)}} \quad (7)$$

$$\sigma_3^{2(t+1)} = \frac{\sum_{i=1}^n \widehat{Z}_{i3}^{(t)} (y_i - \mu_3^{(t)})^2}{\sum_{i=1}^n \widehat{Z}_{i3}^{(t)}} \quad (8)$$

where $\widehat{Z}_{i1}^{(t)} + \sum_{i=1}^n \widehat{Z}_{i2}^{(t)} + \sum_{i=1}^n \widehat{Z}_{i3}^{(t)} = n$ (the sample size).

In both E-step and M-step, the iterative schemes for each parameter and Z_i , $i = 1, 2, 3$ are independent. Therefore, we can apply parallel computing in updating all the parameters and Z 's.

2: Database Access from R (30%)

SQL in the pictures following highlighted in blue in the double quotes.

(a) The 'Book' Table:

```
> dbGetQuery(con,"SELECT * FROM Book;")
  BookNumber Classification
1          1 Natural Science
2          2 Natural Science
3          3 Natural Science
4          4      History
5          5      History
6          6    Philosophy
7          7    Philosophy
8          8    Philosophy
9          9    Philosophy
```

Figure 1: 'Book' Table

(b)

```
> dbGetQuery(con, "SELECT Student.StudentID, EntryYear FROM Student, Record, Book
+   where Book.Classification = 'Natural Science'
+   And Record.BookNumber = Book.BookNumber
+   And Record.StudentID = Student.StudentID;")
  StudentID EntryYear
1          1      2018
2          3      2019
3          8      2018
```

Figure 2: Students who borrowed natural science books

(c)

```
> dbGetQuery(con, "SELECT Student.StudentID, Major FROM Student, Record
+   where Record.BookNumber = '8'
+   And Record.StudentID = Student.StudentID
+   And TIMESTAMPDIFF(day, BorrowingTime, ReturnTime)>30;")
  StudentID Major
1          6    Art
```

Figure 3: Students who borrowed book 8 for more than 30 days

3: Parse HTML (30%)

(a) The result is stored in variable 'comp' in R code.

(b)