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 $t \in \mathbb{Z}$:

$\mathbf{E-step}$: Calculate $E(Z_i^{(t)}|\Theta^{(t)})$, where $\Theta^{(t)} = \pi_1^{(t)}, \pi_2^{(t)}, \mu_1^{(t)}, \mu_2^{(t)}, \mu_3^{(t)}, \sigma_1^{2(t)}, \sigma_2^{2(t)}, \sigma_3^{2(t)}$.

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

$\mathbf{M-step}$: Update $\Theta^{(t+1)}$ by equations (1) to (8).

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

Iterative scheme:

$$\pi_1^{(t+1)} = \frac{\sum_{i=1}^n \widehat{Z_{i1}}^{(t)}}{n} \tag{1}$$

$$\pi_2^{(t+1)} = \frac{\sum_{i=1}^n \widehat{Z_{i2}}^{(t)}}{n} \tag{2}$$

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

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

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

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

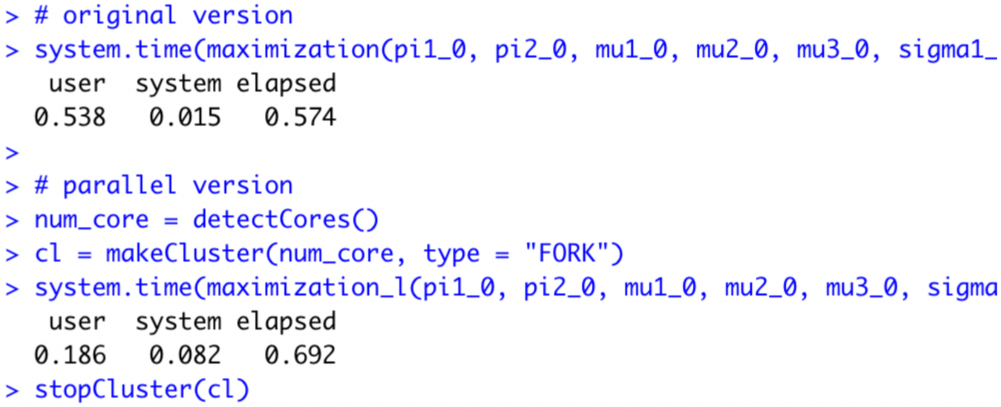
$$\sigma_2^{2(t+1)} = \frac{\sum_{i=1}^n \widehat{Z_{i2}}^{(t)} (y_i - \mu_2^{(t)})^2}{\sum_{i=1}^n \widehat{Z_{i2}}^{(t)}} \tag{7}$$

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

In both E-step and M-step, the iterative schemes for each parameter and missing data are independent. Therefore, we can apply parallel computing in updating all the parameters and Z's.

The detailed process is as follows: given $\Theta^{(t)}$,

1. E-step: Compute conditional expectation values to be stored in an $n \times 3$ matrix, in which the computation tasks are distributed in rows in parallel.
2. M-step:
When computing the above matrix, to pre-compute some intermediate parameters such as $\widehat{Z_{i1}}^{(t)} y_i$ and $\widehat{Z_{i1}}^{(t)} y_i^2$, which are collected in other $n \times 3$ matrices.
The master gather all values computed before and then distribute tasks about updating 8 parameters in parallel.
The master gather the updated parameters $\Theta^{(t+1)}$ and then go to the next loop.



```
> # original version
> system.time(maximization(pi1_0, pi2_0, mu1_0, mu2_0, mu3_0, sigma1_
   user  system elapsed 
  0.538   0.015   0.574 
> 
> # parallel version
> num_core = detectCores()
> cl = makeCluster(num_core, type = "FORK")
> system.time(maximization_l(pi1_0, pi2_0, mu1_0, mu2_0, mu3_0, sigma
   user  system elapsed 
  0.186   0.082   0.692 
> stopCluster(cl)
```

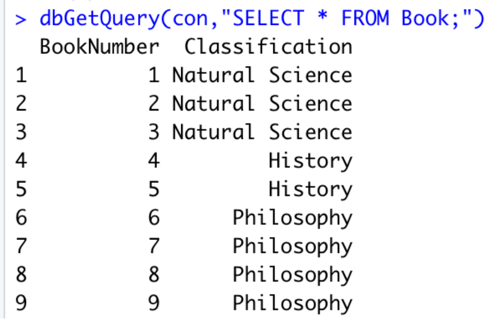
Figure 1: Original VS Parallel Computing Time

From the computation, we ca know that parallel computing is much faster than the original version.

2: Database Access from R (30%)

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

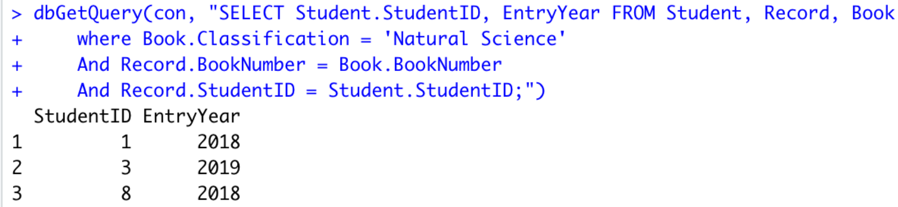
(a) The 'Book' Table:



```
> 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 2: 'Book' Tbale

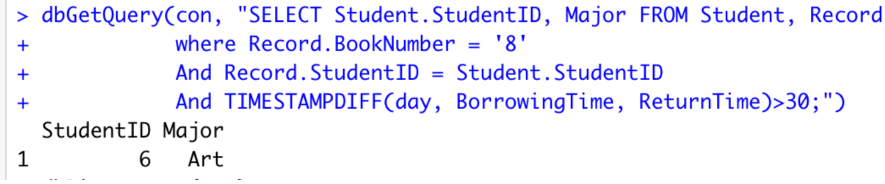
(b)



```
> 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 3: 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
```

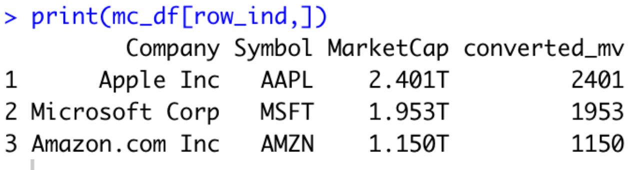
Figure 4: Students who borrowed book 8 for more than 30 days

3: Parse HTML (30%)

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

(b) In Figure.6 (Table of company, ticker symbol, market cap, price to book value, and dividend yield) on page 4.

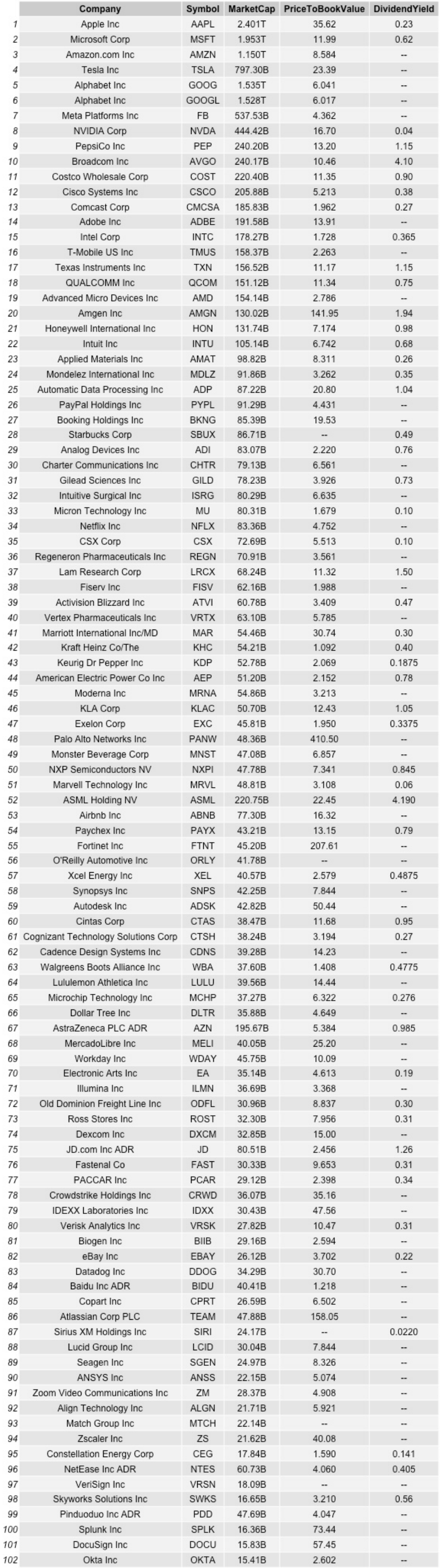
(c)



```
> print(mc_df[row_ind,])
         Company Symbol MarketCap converted_mv
1      Apple Inc   AAPL    2.401T         2401
2 Microsoft Corp   MSFT    1.953T         1953
3 Amazon.com Inc   AMZN    1.150T         1150
```

Figure 5: Top 3 companies with highest Market Cap

Top 3 companies with highest Market Cap are Apple, Microsoft, and Amazon with 2.201T, 1.953T, and 1.150T Market Cap respectively.



	Company	Symbol	MarketCap	PriceToBookValue	DividendYield
1	Apple Inc	AAPL	2.401T	35.62	0.23
2	Microsoft Corp	MSFT	1.953T	11.99	0.62
3	Amazon.com Inc	AMZN	1.150T	8.584	--
4	Tesla Inc	TSLA	797.30B	23.39	--
5	Alphabet Inc	GOOG	1.535T	6.041	--
6	Alphabet Inc	GOOGL	1.528T	6.017	--
7	Meta Platforms Inc	FB	537.53B	4.362	--
8	NVIDIA Corp	NVDA	444.42B	16.70	0.04
9	PepsiCo Inc	PEP	240.20B	13.20	1.15
10	Broadcom Inc	AVGO	240.17B	10.46	4.10
11	Costco Wholesale Corp	COST	220.40B	11.35	0.90
12	Cisco Systems Inc	CSCO	205.88B	5.213	0.38
13	Comcast Corp	CMCSA	185.83B	1.962	0.27
14	Adobe Inc	ADBE	191.58B	13.91	--
15	Intel Corp	INTC	178.27B	1.728	0.365
16	T-Mobile US Inc	TMUS	158.37B	2.263	--
17	Texas Instruments Inc	TXN	156.52B	11.17	1.15
18	QUALCOMM Inc	QCOM	151.12B	11.34	0.75
19	Advanced Micro Devices Inc	AMD	154.14B	2.786	--
20	Amgen Inc	AMGN	130.02B	141.95	1.94
21	Honeywell International Inc	HON	131.74B	7.174	0.98
22	Intuit Inc	INTU	105.14B	6.742	0.68
23	Applied Materials Inc	AMAT	98.82B	8.311	0.26
24	Mondelez International Inc	MDLZ	91.86B	3.262	0.35
25	Automatic Data Processing Inc	ADP	87.22B	20.80	1.04
26	PayPal Holdings Inc	PYPL	91.29B	4.431	--
27	Booking Holdings Inc	BKNG	85.39B	19.53	--
28	Starbucks Corp	SBUX	86.71B	--	0.49
29	Analog Devices Inc	ADI	83.07B	2.220	0.76
30	Charter Communications Inc	CHTR	79.13B	6.561	--
31	Gilead Sciences Inc	GILD	78.23B	3.926	0.73
32	Intuitive Surgical Inc	ISRG	80.29B	6.635	--
33	Micron Technology Inc	MU	80.31B	1.679	0.10
34	Netflix Inc	NFLX	83.36B	4.752	--
35	CSX Corp	CSX	72.69B	5.513	0.10
36	Regeneron Pharmaceuticals Inc	REGN	70.91B	3.561	--
37	Lam Research Corp	LRCX	68.24B	11.32	1.50
38	Fiserv Inc	FISV	62.16B	1.988	--
39	Activision Blizzard Inc	ATVI	60.78B	3.409	0.47
40	Vertex Pharmaceuticals Inc	VRTX	63.10B	5.785	--
41	Marriott International Inc/MD	MAR	54.46B	30.74	0.30
42	Kraft Heinz Co/The	KHC	54.21B	1.092	0.40
43	Keurig Dr Pepper Inc	KDP	52.78B	2.069	0.1875
44	American Electric Power Co Inc	AEP	51.20B	2.152	0.78
45	Moderna Inc	MRNA	54.86B	3.213	--
46	KLA Corp	KLAC	50.70B	12.43	1.05
47	Exelon Corp	EXC	45.81B	1.950	0.3375
48	Palo Alto Networks Inc	PANW	48.36B	410.50	--
49	Monster Beverage Corp	MNST	47.08B	6.857	--
50	NXP Semiconductors NV	NXPI	47.78B	7.341	0.845
51	Marvell Technology Inc	MRVL	48.81B	3.108	0.06
52	ASML Holding NV	ASML	220.75B	22.45	4.190
53	Airbnb Inc	ABNB	77.30B	16.32	--
54	Paychex Inc	PAYX	43.21B	13.15	0.79
55	Fortinet Inc	FTNT	45.20B	207.61	--
56	O'Reilly Automotive Inc	ORLY	41.78B	--	--
57	Xcel Energy Inc	XEL	40.57B	2.579	0.4875
58	Synopsys Inc	SNPS	42.25B	7.844	--
59	Autodesk Inc	ADSK	42.82B	50.44	--
60	Cintas Corp	CTAS	38.47B	11.68	0.95
61	Cognizant Technology Solutions Corp	CTSH	38.24B	3.194	0.27
62	Cadence Design Systems Inc	CDNS	39.28B	14.23	--
63	Walgreens Boots Alliance Inc	WBA	37.60B	1.408	0.4775
64	Lululemon Athletica Inc	LULU	39.56B	14.44	--
65	Microchip Technology Inc	MCHP	37.27B	6.322	0.276
66	Dollar Tree Inc	DLTR	35.88B	4.649	--
67	AstraZeneca PLC ADR	AZN	195.67B	5.384	0.985
68	MercadoLibre Inc	MELI	40.05B	25.20	--
69	Workday Inc	WDAY	45.75B	10.09	--
70	Electronic Arts Inc	EA	35.14B	4.613	0.19
71	Illumina Inc	ILMN	36.69B	3.368	--
72	Old Dominion Freight Line Inc	ODFL	30.96B	8.837	0.30
73	Ross Stores Inc	ROST	32.30B	7.956	0.31
74	Dexcom Inc	DXCM	32.85B	15.00	--
75	JD.com Inc ADR	JD	80.51B	2.456	1.26
76	Fastenal Co	FAST	30.33B	9.653	0.31
77	PACCAR Inc	PCAR	29.12B	2.398	0.34
78	Crowdstrike Holdings Inc	CRWD	36.07B	35.16	--
79	IDEXX Laboratories Inc	IDXX	30.43B	47.56	--
80	Verisk Analytics Inc	VRSK	27.82B	10.47	0.31
81	Biogen Inc	BIIB	29.16B	2.594	--
82	eBay Inc	EBAY	26.12B	3.702	0.22
83	Datadog Inc	DDOG	34.29B	30.70	--
84	Baidu Inc ADR	BIDU	40.41B	1.218	--
85	Copart Inc	CPRT	26.59B	6.502	--
86	Atlassian Corp PLC	TEAM	47.88B	158.05	--
87	Sirius XM Holdings Inc	SIRI	24.17B	--	0.0220
88	Lucid Group Inc	LCID	30.04B	7.844	--
89	Seagen Inc	SGEN	24.97B	8.326	--
90	ANSYS Inc	ANSS	22.15B	5.074	--
91	Zoom Video Communications Inc	ZM	28.37B	4.908	--
92	Align Technology Inc	ALGN	21.71B	5.921	--
93	Match Group Inc	MTCH	22.14B	--	--
94	Zscaler Inc	ZS	21.62B	40.08	--
95	Constellation Energy Corp	CEG	17.84B	1.590	0.141
96	NetEase Inc ADR	NTES	60.73B	4.060	0.405
97	VeriSign Inc	VRSN	18.09B	--	--
98	Skyworks Solutions Inc	SWKS	16.65B	3.210	0.56
99	Pinduoduo Inc ADR	PDD	47.69B	4.047	--
100	Splunk Inc	SPLK	16.36B	73.44	--
101	DocuSign Inc	DOCU	15.83B	57.45	--
102	Okta Inc	OKTA	15.41B	2.602	--

Figure 6: Table of company, ticker symbol, market cap, price to book value, and dividend yield