HW2-Exercise 2

Zining Fan(zf2234), Mutian Wang(mw3386), Siyuan Wang(sw3418)

2.1

Read Data

The Student Score Data read looks like this,

```
scores<-read.table("scores.txt")
```

We could see there are some missing values "NA" in this data.

1) complete case analysis

At this step, we dropped all the observations with missing data.

```
##        x1     x2     x3     x4     x5
## x1 216.30  -7.50  45.05  77.65  94.50
## x2  -7.50 221.50 117.50  77.00 226.75
## x3  45.05 117.50 157.30  85.90 242.00
## x4  77.65  77.00  85.90  75.20 132.25
## x5  94.50 226.75 242.00 132.25 422.00
```

```
## [1] 767.3178
```

The covariance matrix is shown above, we could see that the largest eigenvalue. $\hat{\lambda_1}$ is 767.3178.

2) available case analysis

In this step we use pairwise complete observation,

```
##            x1         x2        x3        x4       x5
## x1 121.363636   4.563636  35.79091  42.12727  94.5000
## x2   4.563636 179.134199 112.26840 114.60173 172.5000
## x3  35.790909 112.268398 151.48918 125.96537 182.3727
## x4  42.127273 114.601732 125.96537 153.56061 142.8636
## x5  94.500000 172.500000 182.37273 142.86364 294.5636
```

```
## [1] 655.6862
```

The covariance matrix is shown above, we could see that the largest eigenvalue $\hat{\lambda_1}$ is 655.6862.

3) mean imputation

In this step, we replace the missing values by the mean of each columns.

```
##          x1        x2        x3        x4        x5
## x1 57.79221   2.17316  17.04329  20.06061  21.50138
## x2  2.17316 179.13420 112.26840 114.60173  82.14286
## x3 17.04329 112.26840 151.48918 125.96537  86.84416
## x4 20.06061 114.60173 125.96537 153.56061  68.03030
## x5 21.50138  82.14286  86.84416  68.03030 140.26840
```

```
## [1] 458.2323
```

The covariance matrix and the largest eigenvalue $\hat{\lambda_1}$ are shown above.

4) mean imputation with the bootstrap

In this step, we used bootstrap. At each trial, we randomly selected 8 rows from the data. Below is average covariance matrix and its largest eigenvalue.

```
##           x1         x2        x3        x4        x5
## x1 42.982153  -1.887119  13.91275  14.45334  14.78145
## x2 -1.887119 184.562857 116.17500 120.19339  78.38760
## x3 13.912752 116.175000 151.25839 128.23536  85.02864
## x4 14.453344 120.193393 128.23536 160.20964  67.43138
## x5 14.781453  78.387595  85.02864  67.43138 131.67243
```

```
## [1] 463.2556
```

We could find that the largest eigenvalue is samller than the one from mean imputation.

5) EM-algorithm

```
## [1] 791.5246
```

The covariance matrix and the largest eigenvalue $\hat{\lambda_1}$ is shown above, which is 791.52. We could find that the largest eigenvalue λ_1 differs a lot among different mathods of imputation.

Comment

We could look at the largest eigenvalue of the covariance matrix , which is larger when we use complete case analysis or available case analysis and is smaller when we use mean imputation. This makes sense, since mean imputation will decerease the variance.

2.2

To compute the 95% confidence interval, $CI : (\hat{\lambda_1} \pm \frac{2}{\sqrt{n}} Z_{1-\frac{\alpha}{2}})$.

1) complete case analysis

```
## [1] 222.1407
```

```
## [1] 2650.467
```

2) available case analysis

```
## [1] 363.1191
```

```
## [1] 1183.976
```

3) mean imputation

```
## [1] 253.7691
```

```
## [1] 827.4325
```

4) mean imputation with the bootstrap

Now, we have 100 observations of λ_1,

```
## [1] 351.1095
```

```
## [1] 611.2217
```

5) EM-algorithm

```
## [1] 438.3464
```

```
## [1] 1429.26
```

2.3

In this question, we used the whole data.

```
##            mechanics   vectors   algebra  analysis statistics
## mechanics   305.7680 127.22257 101.57941 106.27273  117.40491
## vectors     127.2226 172.84222  85.15726  94.67294   99.01202
## algebra     101.5794  85.15726 112.88597 112.11338  121.87056
## analysis    106.2727  94.67294 112.11338 220.38036  155.53553
## statistics  117.4049  99.01202 121.87056 155.53553  297.75536
```

```
## [1] 686.9898
```

We could see that the largest eigenvalue $\hat{\lambda}_1$is 686.9, which is quite similar to the one by available case analysis.

Using Bootstrap:

```
##            mechanics   vectors   algebra  analysis statistics
## mechanics   313.0353 130.69197 106.94328 110.09514   126.2638
## vectors     130.6920 172.25102  87.34809  95.31332   102.5963
## algebra     106.9433  87.34809 114.99679 110.32840   124.3145
## analysis    110.0951  95.31332 110.32840 215.60570   155.7736
## statistics  126.2638 102.59631 124.31449 155.77364   295.9167
```

```
## [1] 700.0244
```

```
## [1] 675.5742
```

```
## [1] 743.1351
```

Based on the bootstrap, the first number largest eigenvalue, the following two numbers are the lower and upper bound for confidence interval. We could see that the confidence interval based on the original data is relatively larger than the one based on the mean imputed data. This makes sense, since mean imputation will decrease the variance.

2.4

$$\log f(\mathbf{X}_{io}, \mathbf{X}_{im}|\mu, \Sigma) = \log f(\mathbf{X}_{io}|\mu, \Sigma) + \log f(\mathbf{X}_{im}|\mathbf{X}_{io}, \mu, \Sigma)$$

E-step: use $E[\log f(\mathbf{X}_{im}|\mathbf{X}_{io}, \mu, \Sigma)|\mathbf{X}_{io} = \mathbf{x}_{io}, \mu', \Sigma']$ to replace $\log f(\mathbf{X}_{im}|\mathbf{X}_{io}, \mu, \Sigma)$.

$$\begin{aligned}
&\log f(\mathbf{X}_{io}|\mu, \Sigma) = \sum -\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_{ioo}| - \frac{1}{2}\left(\mathbf{X}_{io} - \mu_{io}\right)^T \Sigma_{ioo}^{-1}\left(\mathbf{X}_{io} - \mu_{io}\right) \\
&E[\log f(\mathbf{X}_{im}|\mathbf{X}_{io}, \mu, \Sigma)|\mathbf{X}_{io} = \mathbf{x}_{io}, \mu', \Sigma'] \\
= &\sum E[-\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_{\mathbf{X_{im}}|\mathbf{X_{io}}}| - \frac{1}{2}\left(\mathbf{X}_{im}|\mathbf{X}_{io} - \mu_{\mathbf{X_{im}}|\mathbf{X_{io}}}\right)^T \Sigma_{\mathbf{X_{im}}|\mathbf{X_{io}}}^{-1}\left(\mathbf{X}_{im}|\mathbf{X}_{io} - \mu_{\mathbf{X_{im}}|\mathbf{X_{io}}}\right)]
\end{aligned}$$

M-step: Now we want to solve for μ and Σ to maximize $\log f(\mathbf{X}_{io}|\mu, \Sigma) + E[\log f(\mathbf{X}_{im}|\mathbf{X}_{io}, \mu, \Sigma)]$. Then we plug in the new μ and Σ and iterate until convergence.

Therefore we can derive

$$\begin{aligned}
\mu^{k+1} &: \sum_{i=1}^{n}(\hat{X}_i - \mu) = 0 \\
\Sigma^{k+1} &: \sum_{i=1}^{n}\left(\Sigma - (\hat{X}_i - \mu)(\hat{X}_i - \mu)^T - \mathbf{C}_i^{(k)}\right) = 0
\end{aligned}$$

$\hat{X}_{io} = X_{io}$ and

$$\hat{X}_{im} = E[X_{im}|X_{io}] = \mu_{X_{im}|X_{io}} = \mu_{im}^{(k)} + \Sigma_{imo}^{(k)}(\Sigma_{ioo}^{(k)})^{-1}(X_{io} - \mu_{io}^{(k)})$$

and $C_{ijk}^{(k)} = 0$ if X_{ij} or X_{ik} are observed and

$$\mathbf{C}_{imm}^{(k)} = \Sigma_{X_{im}|X_{io}} = \Sigma_{imm}^{(k)} - \Sigma_{imo}^{(k)}(\Sigma_{ioo}^{(k)})^{-1}\Sigma_{iom}^{(k)}$$