UPPSALA UNIVERSITY
Department of Mathematics
Shaobo Jin

Exam in Mathematical Statistics
Multivariate Methods, 1MS003
2024-01-04

Time: 8.00-13.00. Limits for the credits 3, 4, 5 are 18, 25 and 32 points, respectively. The solutions should be well motivated.

Permitted aids: A sheet of your own notes (A4 paper, two-sided). Pocket calculator. Dictionary. No electronic device with internet connection is allowed.

1. (4p) Suppose that

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}, \begin{bmatrix} 2 & 1 & a \\ 1 & 3 & 1 \\ a & 1 & 2 \end{bmatrix} \right),$$

   (a) (1p) Find the distribution of $\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \mid X_3$.

   Solution: $\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \mid X_3$ remains normal with mean

$$\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} a \\ 1 \end{bmatrix} 2^{-1} (X_3 - \mu_3)$$

   and covariance matrix

$$\begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix} - \begin{bmatrix} a \\ 1 \end{bmatrix} 2^{-1} \begin{bmatrix} 1 & a \end{bmatrix}.$$

   (b) (2p) Find the joint distribution of $X_1 + X_3$ and $X_2 + X_3$.
   Solution: Note that

$$\begin{bmatrix} X_1 + X_3 \\ X_2 + X_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}.$$

   Hence, the joint distribution is normal with mean

$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} \mu_1 + \mu_3 \\ \mu_2 + \mu_3 \end{bmatrix}$$

   and covariance matrix

$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & a \\ 1 & 3 & 1 \\ a & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}.$$

(c) (1p) Which value a should take in order for X_1 being independent of X_3?
Solution: $a = 1$.

2. (9p) Suppose that

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \sim N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad \boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} \sim N_3(\boldsymbol{\kappa}, \boldsymbol{\Sigma}).$$

(a) (1p) What assumption is needed in order for the joint distribution of $\boldsymbol{X}$ and $\boldsymbol{Y}$ to be normal?
Solution: We need $\boldsymbol{X}$ and $\boldsymbol{Y}$ to be independent.

(b) (2p) Hereafter, suppose that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent. For each brand, 100 samples are measured, i.e., $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, ..., $\boldsymbol{X}_{100}$ and $\boldsymbol{Y}_1$, $\boldsymbol{Y}_2$, ..., $\boldsymbol{Y}_{100}$. Let $\bar{\boldsymbol{X}}$ and $\bar{\boldsymbol{Y}}$ be the sample mean of two brands, respectively. Find the distribution of $\bar{\boldsymbol{X}} - 2\bar{\boldsymbol{Y}}$.
Solution: Note that $\bar{\boldsymbol{X}} \sim N_3\left(\boldsymbol{\mu}, \frac{1}{100}\boldsymbol{\Sigma}\right)$ and $\bar{\boldsymbol{Y}} \sim N_3\left(\boldsymbol{\kappa}, \frac{1}{100}\boldsymbol{\Sigma}\right)$. Since they are also independent, then $\bar{\boldsymbol{X}} - 2\bar{\boldsymbol{Y}} = \begin{bmatrix} 1 & -2 \end{bmatrix} \begin{bmatrix} \bar{\boldsymbol{X}} \\ \bar{\boldsymbol{Y}} \end{bmatrix} \sim N_3\left(\boldsymbol{\mu} - 2\boldsymbol{\kappa}, \frac{5}{100}\boldsymbol{\Sigma}\right)$.

(c) (2p) Find the distribution of

$$\boldsymbol{S} = \frac{1}{198}\left[\sum_{i=1}^{100}\left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right)\left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right)^T + \sum_{j=1}^{100}\left(\boldsymbol{Y}_j - \bar{\boldsymbol{Y}}\right)\left(\boldsymbol{Y}_j - \bar{\boldsymbol{Y}}\right)^T\right].$$

Solution: Under the normality assumption,

$$\begin{aligned} \sum_{i=1}^{100}\left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right)\left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right)^T &\sim W_3(\boldsymbol{\Sigma}, 99), \\ \sum_{j=1}^{100}\left(\boldsymbol{Y}_j - \bar{\boldsymbol{Y}}\right)\left(\boldsymbol{Y}_j - \bar{\boldsymbol{Y}}\right)^T &\sim W_3(\boldsymbol{\Sigma}, 99). \end{aligned}$$

Hence,

$$198\boldsymbol{S} \sim W_3(\boldsymbol{\Sigma}, 198).$$

(d) (2p) Suppose that $\bar{\boldsymbol{X}} - 2\bar{\boldsymbol{Y}}$ is independent of $\boldsymbol{S}$. Suppose also that $\boldsymbol{\mu} = 2\boldsymbol{\kappa}$. Find the distribution of $20\left(\bar{\boldsymbol{X}} - 2\bar{\boldsymbol{Y}}\right)^T \boldsymbol{S}^{-1}\left(\bar{\boldsymbol{X}} - 2\bar{\boldsymbol{Y}}\right)$.
Solution: Note that $\sqrt{\frac{100}{5}}\left(\bar{\boldsymbol{X}} - 2\bar{\boldsymbol{Y}}\right) \sim N_3(\boldsymbol{0}, \boldsymbol{\Sigma})$ and $198\boldsymbol{S} \sim W_3(\boldsymbol{\Sigma}, 198)$. Then,

$$\begin{aligned} 198\left[\sqrt{\frac{100}{5}}\left(\bar{\boldsymbol{X}} - 2\bar{\boldsymbol{Y}}\right)\right]^T (198\boldsymbol{S})^{-1}\left[\sqrt{\frac{100}{2}}\left(\bar{\boldsymbol{X}} - 2\bar{\boldsymbol{Y}}\right)\right] &= \frac{100}{5}\left(\bar{\boldsymbol{X}} - 2\bar{\boldsymbol{Y}}\right)^T \boldsymbol{S}^{-1}\left(\bar{\boldsymbol{X}} - 2\bar{\boldsymbol{Y}}\right) \\ &\sim T^2(3, 198). \end{aligned}$$

(e) (2p) Suppose that $\boldsymbol{\mu} = \mathbf{0}$ is known. Find the maximum likelihood estimator of $\boldsymbol{\Sigma}$ using $\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_{100}$.
Solution: The log-likelihood function is

$$\begin{aligned} \ell(\boldsymbol{\Sigma}) &= -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log\det(\boldsymbol{\Sigma}) - \frac{1}{2}\sum_{j=1}^{n}\boldsymbol{X}_j^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_j \\ &= -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log\det(\boldsymbol{\Sigma}) - \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}\sum_{j=1}^{n}\boldsymbol{X}_j\boldsymbol{X}_j^T\right\}. \end{aligned}$$

Recall that $g(\boldsymbol{\Sigma}) = -q\log\det(\boldsymbol{\Sigma}) - \mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\right\}$ is maximized at $\boldsymbol{\Sigma} = q^{-1}\boldsymbol{A}$. Hence, the MLE is

$$\hat{\boldsymbol{\mu}} = n^{-1}\sum_{j=1}^{n}\boldsymbol{X}_j\boldsymbol{X}_j^T,$$

where $n = 100$ here.

3. (7p) Suppose that we have measured the weight of water, the weight of fat, and the weight of protein of one type of meat produced by five brands. We want to test whether different brands have the same weights. For each brand, 100 samples are measured.

(a) (1p) Formulate the corresponding MANOVA model. Note: No need to specify the identification restrictions.
Solution: The MANOVA model is

$$\boldsymbol{y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\tau}_i + \boldsymbol{e}_{ij},$$

where $i = 1, 2, ..., 5$ and $j = 1, 2, ..., 100$.

(b) (2p) What are the assumptions in order to carry out MANOVA?
Solution: The assumptions include (1) the sample $\boldsymbol{X}_{11}$, $\boldsymbol{X}_{12}$, ..., $\boldsymbol{X}_{1n_1}$ is a random sample of size n_1 from a p-variate population with mean vector $\boldsymbol{\mu}_1$ and covariance matrix $\boldsymbol{\Sigma}_1$, (2) the sample $\boldsymbol{X}_{21}$, $\boldsymbol{X}_{22}$, ..., $\boldsymbol{X}_{2n_2}$ is a random sample of size n_2 from a p-variate population with mean vector $\boldsymbol{\mu}_2$ and covariance matrix $\boldsymbol{\Sigma}_2$ (3) $\boldsymbol{X}_{11}, \boldsymbol{X}_{12}, ..., \boldsymbol{X}_{1n_1}$ are independent of $\boldsymbol{X}_{21}$, $\boldsymbol{X}_{22}$, ..., $\boldsymbol{X}_{2n_2}$. (4) Both populations are multivariate normal, (5) $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$.

(c) (1p) A statistician has performed the following MANOVA analysis in R.

```
##            Df   Wilks approx F num Df den Df Pr(>F)
## Brand       4 0.97101   1.2158     12 1304.6 0.2661
## Residuals 495
```

What conclusion can be drawn?
Solution: We cannot reject the null hypothesis that different brands have the same expected weight.

(d) (1p) Before drawing any conclusion, another statistician has done the following analysis to the weight of water, the weight of fat, and the weight of protein of the first brand.

```
## $multivariateNormality
##      Test        H      p value
## 1 Royston 34.90605 3.578046e-09
##
## $univariateNormality
##               Test  Variable Statistic   p value
## 1 Anderson-Darling   Water      6.5294  <0.001
## 2 Anderson-Darling    Fat       6.6318  <0.001
## 3 Anderson-Darling  Protein     4.2136  <0.001
```

How would you interpret the outputs?
Solution: We can reject the null hypothesis that data are normally distributed.

(e) (2p) Another statistician has done a regression analysis. The results are shown below.

```
## Response Fat :
##
## Call:
## lm(formula = Fat ~ Water, data = Brand1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7975 -0.5036  0.3221  1.0868  5.8743
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 96.89448    1.55776   62.20   <2e-16 ***
## Water       -1.24781    0.02342  -53.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.139 on 98 degrees of freedom
## Multiple R-squared:  0.9666,Adjusted R-squared:  0.9663
## F-statistic:  2839 on 1 and 98 DF,  p-value: < 2.2e-16
##
##
## Response Protein :
##
## Call:
## lm(formula = Protein ~ Water, data = Brand1)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -7.0067 -0.6194  0.2473  0.9392  4.6420
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.46526    1.34771   1.829   0.0704 .
## Water        0.24041    0.02026  11.866   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.851 on 98 degrees of freedom
## Multiple R-squared:  0.5896,Adjusted R-squared:  0.5854
## F-statistic: 140.8 on 1 and 98 DF,  p-value: < 2.2e-16
```

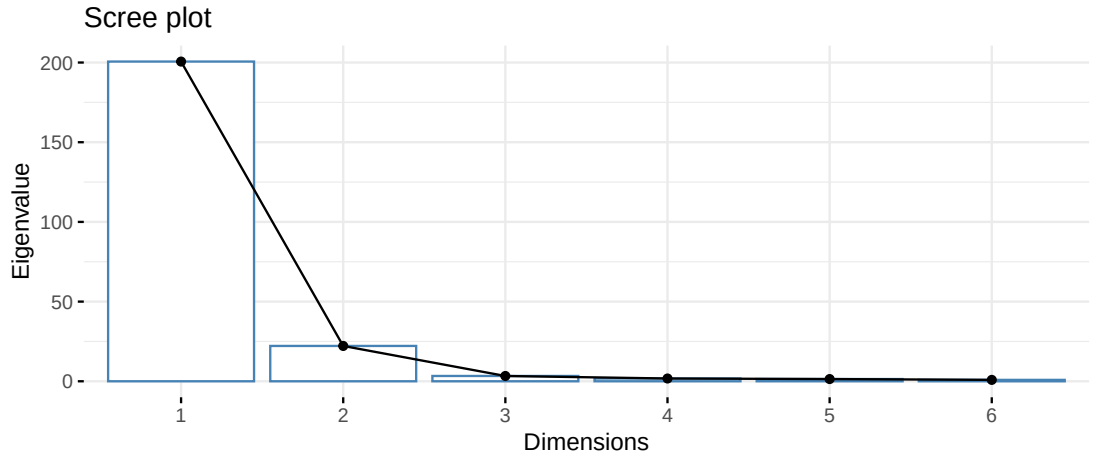
What is the fitted regression model?

Solution: $\begin{bmatrix} \hat{\text{Fat}} \\ \hat{\text{Protein}} \end{bmatrix} = \begin{bmatrix} 96.89448 & -1.24781 \\ 2.46526 & 0.24041 \end{bmatrix} \begin{bmatrix} \text{Intercept} \\ \text{Water} \end{bmatrix}$.

4. (10p) Suppose that we have observed a data set with six variables. Its sample covariance matrix that is shown below.

```
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]   10    9    6    2    2   14
## [2,]    9   11    6    2    1   12
## [3,]    6    6    5    1    1    8
## [4,]    2    2    1    6    5   25
## [5,]    2    1    1    5    5   24
## [6,]   14   12    8   25   24  179
```

(a) (1p) Statistician A performs a PCA to the above covariance matrix and obtains the following scree plot.



How many principal components would you like to choose? Motive your choice.
Solution: 2 components is a reasonable choice as it is where the elbow occurs.

(b) (1p) Statistician B chooses to perform PCA to the correlation matrix. The following eigenvalues-eigenvectors of the correlation matrix are obtained from R.

```
## eigen() decomposition
## $values
## [1] 3.4 1.9 0.2 0.2 0.2 0.1
##
## $vectors
##            [,1]       [,2]        [,3]        [,4]        [,5]
## [1,] -0.4396010 -0.3582039  0.09202419  0.35607223  0.35365600
## [2,] -0.4155137 -0.4143898  0.16378417  0.03297737  0.31619954
## [3,] -0.4130217 -0.4052381 -0.25698945 -0.32778826 -0.69515631
## [4,] -0.3972141  0.4094615 -0.70842435 -0.21235984  0.35662305
## [5,] -0.3817888  0.4528828  0.07572984  0.67199699 -0.40431609
## [6,] -0.3999634  0.4037222  0.62535003 -0.51769032  0.03242668
##             [,6]
## [1,]  0.64661618
## [2,] -0.72645343
## [3,]  0.09200066
## [4,] -0.02050572
## [5,] -0.16836879
## [6,]  0.13007792
```

How much variation has been explained by the first two principal components?
Solution: The proportion is $(3.4+1.9)/6 \approx 0.88$.

(c) (1p) Based on the analysis of Statistician B, how would you calculate the first principal component?
Solution: The first component is $-0.440x_1 - 0.416x_2 - 0.413x_3 - 0.397x_4 - 0.382x_5 - 0.400x_6$.

(d) (1p) Based on the analysis of Statistician B, what is the covariance matrix of the principal components?
Solution: The covariance matrix is

$$\begin{bmatrix} 3.4 & & & & & \\ & 1.9 & & & & \\ & & 0.2 & & & \\ & & & 0.2 & & \\ & & & & 0.2 & \\ & & & & & 0.1 \end{bmatrix}$$

(e) (1p) Are the results by Statistician A equivalent to the results by Statistician B?
Solution: No, PCA depends on the scale.

(f) (2p) Statistician C chooses to perform a factor analysis to the data set. Formulate the factor analysis model as well as its assumptions.
Solution: The factor model is $\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{LF} + \boldsymbol{e}$. Our assumptions are

$$\begin{aligned} \mathbb{E}(\boldsymbol{F}) &= \boldsymbol{0}, \\ \operatorname{cov}(\boldsymbol{F}) &= \boldsymbol{I}, \\ \mathbb{E}(\boldsymbol{e}) &= \boldsymbol{0}, \\ \operatorname{cov}(\boldsymbol{e}) &= \boldsymbol{\Psi} \text{ (diagonal)}, \\ \operatorname{cov}(\boldsymbol{F}, \boldsymbol{e}) &= \boldsymbol{0}. \end{aligned}$$

(g) (1p) What is orthogonal rotation in factor analysis?
Solution: For any orthogonal matrix $\boldsymbol{T}$, The covariance matrices of $\boldsymbol{\mu} + \boldsymbol{LF} + \boldsymbol{e}$ and $\boldsymbol{\mu} + (\boldsymbol{LT})\left(\boldsymbol{T}^T \boldsymbol{F}\right) + \boldsymbol{e}$ are the same. Hence, we can multiply any orthogonal matrix to the loading matrix.

(h) (2p) Show that factor analysis is scale invariant (i.e., the analysis to the covariance matrix and the analysis to the correlation matrix are equivalent).
Solution: The covariance matrix of $\boldsymbol{X}$ is $\boldsymbol{LL}^T + \boldsymbol{\Psi}$. If we have a symmetric matrix $\boldsymbol{C}$, then the covariance matrix of $\boldsymbol{CX}$ is

$$(\boldsymbol{CL})(\boldsymbol{CL})^T + \boldsymbol{C\Psi C},$$

which correspond to the model

$$\boldsymbol{CLF} + \boldsymbol{Ce}.$$

5. (6p) Consider the following two populations. The first population has the density

$$f_1(x) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left\{-\frac{x^2}{2\sigma_1^2}\right\}, \quad -\infty < x < \infty.$$

The second population has the density

$$f_2(x) = \frac{1}{\sigma_2} \exp\left\{-\frac{x}{\sigma_2} - \exp\left(-\frac{x}{\sigma_2}\right)\right\}, \quad -\infty < x < \infty.$$

(a) (2p) Suppose that both σ_1 and σ_2 are known. Determine the classifier that minimizes the ECM when the prior probabilities are uniform and the cost of misclassifying an object to population 1 and population 2 are 4 and 1, respectively.
Solution: We classify the object to class 1 if

$$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \frac{c(1 \mid 2)}{c(2 \mid 1)} \frac{p_2}{p_1} = 4,$$

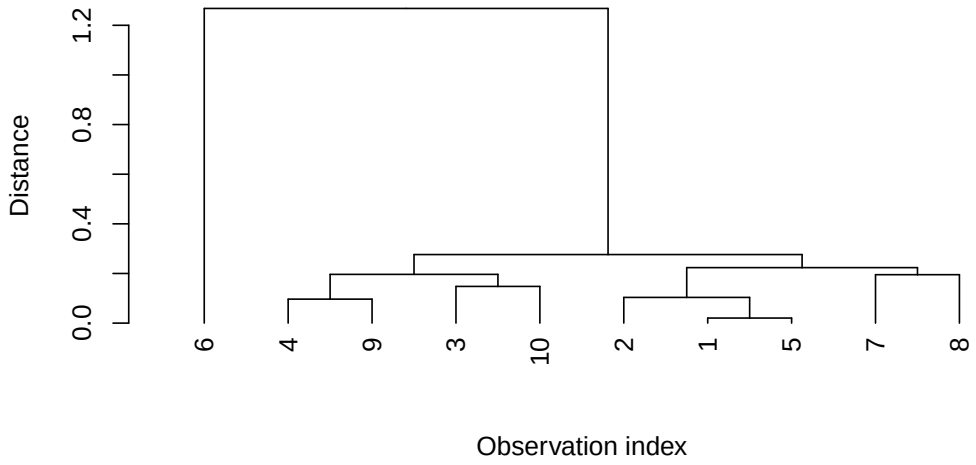
where

$$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} = \frac{\frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left\{-\frac{x^2}{2\sigma_1^2}\right\}}{\frac{1}{\sigma_2} \exp\left\{-\frac{x}{\sigma_2} - \exp\left(-\frac{x}{\sigma_2}\right)\right\}}.$$

(b) (2p) Suppose that both σ_1 and σ_2 are known. Derive the classifier that assigns an object to the class with the highest posterior probability, where the prior probability of population 1 is 0.4 and the prior probability of population 2 is 0.6.
Solution: Classifying based on the posterior probability is the same as setting the costs to be the same. Hence, an object belongs to the first class if

$$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \frac{0.6}{0.4} = 1.5.$$

(c) (1p) A sample of size 10 has been observed. But we have lost the information on which population our observations come from. Hence, a cluster analysis has been done. The results are shown below.



Which two objects are grouped into the same cluster at the first step?
Solution: 1 and 5, since the distance is the smallest.

(d) (1p) What is the (approximate) cluster distance when the cluster containing object 3 and the cluster containing object 9 are combined into one cluster?
Solution: It is approximately 0.2, just the distance axis when they merge into one class.

6. (4pt) Suppose that we have observed a random sample $x_1, x_2, ..., x_n$ from two populations, but we do not observe which population each observation comes from. We assume that $X|Z=k$ follows an exponential distribution with mean θ_k, $k=1,2$. Let $p_k = P(Z=k)$. Explain how the EM algorithm is used to find the estimator of p_1, θ_1 and θ_2.
Solution: In the E-step, we compute the conditional expectation given by

$$\begin{aligned} Q\left(\boldsymbol{\theta}, p_1 \mid \hat{\boldsymbol{\theta}}^{(t)}, \hat{p}^{(t)}\right) &= \sum_{j=1}^{N}\sum_{k=1}^{2}\left[\log p_k + \log f\left(\boldsymbol{x}_j \mid Z_j = k; \boldsymbol{\theta}\right)\right] P_{(t)}\left(Z_j = k \mid \boldsymbol{x}_j\right) \\ &= \sum_{j=1}^{N}\sum_{k=1}^{2}\left[\log p_k - \log\theta_k - \frac{x_j}{\theta_k}\right] P_{(t)}\left(Z_j = k \mid \boldsymbol{x}_j\right). \end{aligned}$$

The M-step maximizes $Q\left(\boldsymbol{\theta}, p_1 \mid \hat{\boldsymbol{\theta}}^{(t)}, \hat{p}^{(t)}\right)$. Note that

$$\frac{\partial Q}{\partial p} = \sum_{j=1}^{N} \frac{1}{p_1} P_{(t)}\left(Z_j = 1 \mid \boldsymbol{x}_j\right) - \sum_{j=1}^{N} \frac{1}{1-p_1} P_{(t)}\left(Z_j = 2 \mid \boldsymbol{x}_j\right).$$

Hence

$$\hat{p}_1 = \frac{1}{N} \sum_{j=1}^{N} P_{(t)}\left(Z_j = 1 \mid \boldsymbol{x}_j\right).$$

Note also that

$$\frac{\partial Q}{\partial \theta_k} = \sum_{j=1}^{N} \frac{x_j - \theta_k}{\theta_k^2} P_{(t)}\left(Z_j = k \mid \boldsymbol{x}_j\right).$$

Hence,

$$\hat{\theta}_k = \frac{\sum_{j=1}^{N} x_j P_{(t)}\left(Z_j = k \mid \boldsymbol{x}_j\right)}{\sum_{j=1}^{N} P_{(t)}\left(Z_j = k \mid \boldsymbol{x}_j\right)}.$$

It is enough to get full points until here.
To be more complete, by Bayes rule,

$$\begin{aligned} P_{(t)}\left(Z_j = k \mid x_j\right) &= \frac{P_{(t)}\left(x_j \mid Z_j = k\right) P_{(t)}\left(Z_j = k\right)}{\sum_{k=1}^{2} P_{(t)}\left(x_j \mid Z_j = k\right) P_{(t)}\left(Z_j = k\right)} \\ &= \frac{\left(\theta_k^{(t)}\right)^{-1} \exp\left(-x_j/\theta_k^{(t)}\right) \hat{p}^{(t)}}{\left(\theta_1^{(t)}\right)^{-1} \exp\left(-x_j/\theta_1^{(t)}\right) \hat{p}^{(t)} + \left(\theta_2^{(k)}\right)^{-1} \exp\left(-x_j/\theta_2^{(t)}\right) \left(1-\hat{p}^{(t)}\right)}. \end{aligned}$$