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 $\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \mid X_3$.

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

   (c) (1p) Which value a should take in order for X_1 being independent of X_3?

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?

   (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 $\bar{\boldsymbol{X}} - 2\bar{\boldsymbol{Y}}$.

   (c) (2p) Find the distribution of

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

   (d) (2p) Suppose that $\bar{\boldsymbol{X}} - 2\bar{\boldsymbol{Y}}$ is independent of $\boldsymbol{S}$. Find the distribution of $20\left(\bar{\boldsymbol{X}} - 2\bar{\boldsymbol{Y}}\right)^T \boldsymbol{S}^{-1}\left(\bar{\boldsymbol{X}} - 2\bar{\boldsymbol{Y}}\right)$ when $\boldsymbol{\mu} = 2\boldsymbol{\kappa}$.

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

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.

   (b) (2p) What are the assumptions in order to carry out MANOVA?

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

```
MANOVA <- manova(cbind(Water, Fat, Protein) ~ Brand, data = Data)
summary(MANOVA, test = "Wilks")
```

```
##             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?

   (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.

```
mvn(Data[Data$Brand == "1", 1 : 3], mvnTest = "royston")
```

```
## $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
```

   What conclusion can be drawn?

   (e) (2p) Another statistician has done a regression analysis using the data from the first brand. 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
```

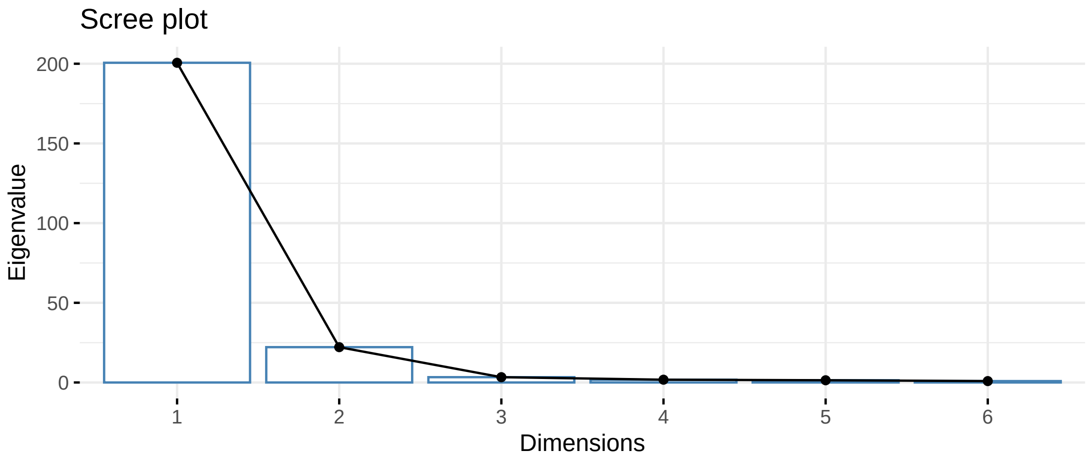
What is the fitted regression model?

4. (10p) Suppose that we have observed a data set with six variables (X_1, X_2, ..., X_6). Its sample covariance matrix is shown below.

```
cov(X)


##      [,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.

(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]   [,6]
## [1,] -0.440 -0.358  0.092  0.356  0.354  0.647
## [2,] -0.416 -0.414  0.164  0.033  0.316 -0.726
## [3,] -0.413 -0.405 -0.257 -0.328 -0.695  0.092
## [4,] -0.397  0.409 -0.708 -0.212  0.357 -0.021
## [5,] -0.382  0.453  0.076  0.672 -0.404 -0.168
## [6,] -0.400  0.404  0.625 -0.518  0.032  0.130
```

How much variation has been explained by the first two principal components?

(c) (1p) Based on the analysis of Statistician B, how would you calculate the first principal component?

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

(e) (1p) Are the results by Statistician A equivalent to the results by Statistician B?

(f) (2p) Statistician C chooses to perform a factor analysis to the data set. Formulate the factor analysis model as well as its assumptions.

(g) (1p) What is orthogonal rotation in factor analysis?

(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).

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

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

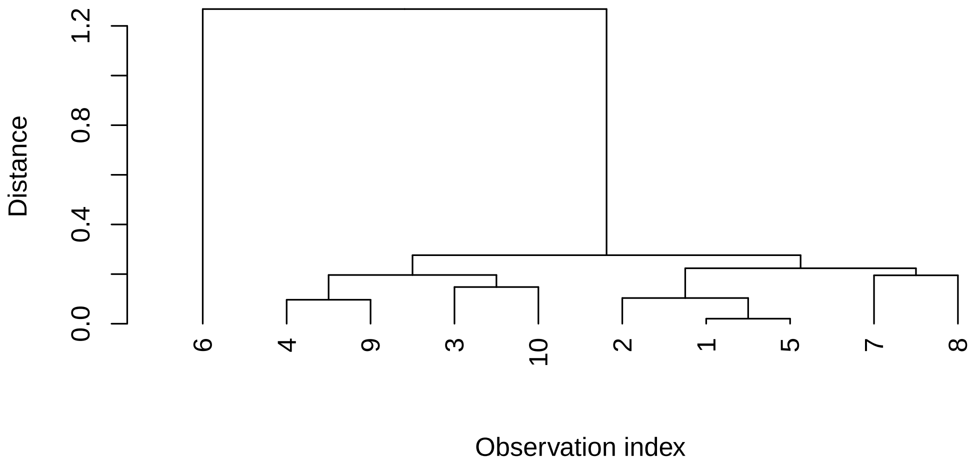
The second population has the density

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

(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 to an object to population 1 and population 2 are 4 and 1, respectively.

(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.

(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?

(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?

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

$$f(X|Z=k) \quad = \quad \frac{1}{\theta_k}\exp\left\{-\frac{x}{\theta_k}\right\}, \quad x>0,\ k=1,2.$$

Let $p_k = P(Z=k)$. Explain how the EM algorithm can be used to find the estimator of p_1, θ_1 and θ_2.