Training Exam, Multivariate Analysis

Shaobo Jin

1. (4pt) Let

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \boldsymbol{X}_3 \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \\ \boldsymbol{\mu}_3 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{13} \\ \boldsymbol{\Sigma}_{12}^T & \boldsymbol{\Sigma}_{22} & \boldsymbol{\Sigma}_{23} \\ \boldsymbol{\Sigma}_{13}^T & \boldsymbol{\Sigma}_{23}^T & \boldsymbol{\Sigma}_{33} \end{bmatrix} \right),$$

and

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{I}_{q\times q} & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ \boldsymbol{0} & \boldsymbol{I}_{r\times r} \end{bmatrix}.$$

   (a) Find the joint distribution of $\boldsymbol{A}\left(\begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix} - \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \right)$.

   (b) Find the conditional distribution of $\begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_3 \end{bmatrix} \mid \boldsymbol{X}_2 = \boldsymbol{a}$.

2. (4p) Let X_{jk} be the response to the kth treatment on the jth unit, where $k = 1, 2, .., q$ and $j = 1, 2, ..., n$. Let

$$\boldsymbol{X}_j = \begin{bmatrix} X_{j1} \\ X_{j2} \\ \vdots \\ X_{jq} \end{bmatrix}.$$

Suppose that $\boldsymbol{X}_1$, ..., $\boldsymbol{X}_n$ is a random sample from an $N_q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ population. We want to test a linear combination of $\boldsymbol{\mu}$ as

$$H_0 : \boldsymbol{C\mu} = \boldsymbol{0} \quad \text{versus} \quad H_1 : \boldsymbol{C\mu} \neq \boldsymbol{0},$$

where $\boldsymbol{C}$ is a matrix of constants.

   (a) Let

$$\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{X}_j,$$
$$\boldsymbol{S} = \frac{1}{n-1}\sum_{j=1}^{n} \left(\boldsymbol{X}_j - \bar{\boldsymbol{X}}\right)\left(\boldsymbol{X}_j - \bar{\boldsymbol{X}}\right)^T.$$

   Find the distribution of $\boldsymbol{C}\bar{\boldsymbol{X}}$ and $(n-1)\boldsymbol{CSC}^T$.

(b) Find the distribution of $T^2 = n\left(\boldsymbol{C}\bar{\boldsymbol{X}} - \boldsymbol{C}\boldsymbol{\mu}\right)^T \left(\boldsymbol{C}\boldsymbol{S}\boldsymbol{C}^T\right)^{-1} \left(\boldsymbol{C}\bar{\boldsymbol{X}} - \boldsymbol{C}\boldsymbol{\mu}\right)$.

3. (4p) Consider the OLS estimator of the multivariate regression coefficient $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{Y}$. The predicted value is $\hat{\boldsymbol{Y}} = \boldsymbol{Z}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{Y}$, and the residual matrix is $\hat{\boldsymbol{E}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$.

   (a) Show that the predicted value is $\hat{\boldsymbol{Y}}_{(i)}$ is perpendicular to the residual $\hat{\boldsymbol{E}}_{(k)}$, where $\hat{\boldsymbol{Y}}_{(i)}$ is the ith column of $\hat{\boldsymbol{Y}}$ and $\hat{\boldsymbol{E}}_{(k)}$ is the kth column of $\hat{\boldsymbol{E}}$.

   (b) Show that $\hat{\boldsymbol{E}}_{(k)}$ is perpendicular to the columns of $\boldsymbol{Z}$.

4. (15pt) Suppose that we have a dataset with the following variables *Station*, *PM25*, *SO2*, *NO2*, and *CO*. There are 12 stations in total.

   (a) Suppose that we have perform the following test in R

```
HotellingsT2Test(x = Station1, mu = c(60, 20, 60, 1000), test = "f")

##
##  Hotelling's one sample T2-test
##
## data:  Station1
## T.2 = 19.992, df1 = 4, df2 = 199, p-value = 7.472e-14
## alternative hypothesis: true location is not equal to c(60,20,60,1000)
```

   Which test problem has been considered here?

   (b) Given the degrees of freedoms of the test produced by the function HotellingsT2Test(), how many observations do you have in the first station?

   (c) A statistician has computed some confidence intervals for the mean values of these variables of the first station. The R code of computing those confidence interval are given below.

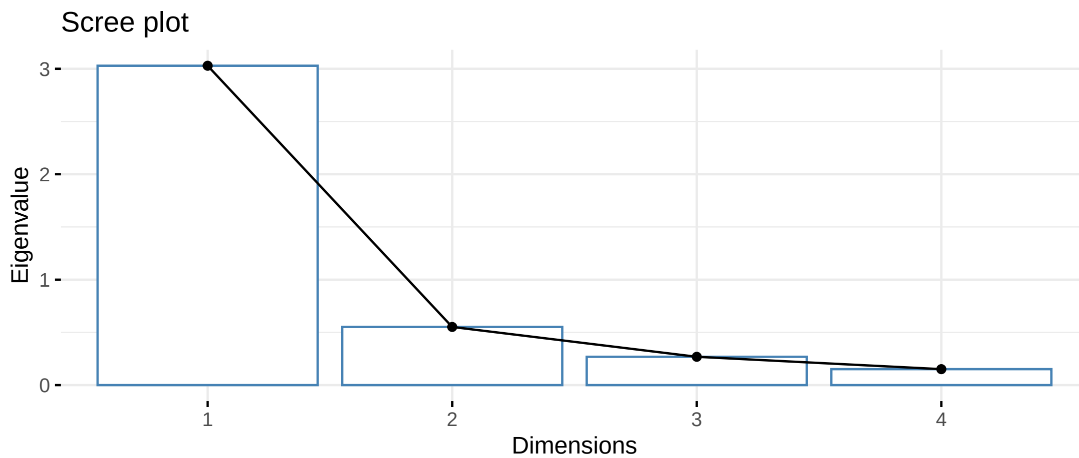
```
Xbar <- colMeans(Station1)
S <- cov(Station1)
c(Xbar[1] - sqrt(S[1, 1] / n1) * qt(0.975, n1 - 1),
  Xbar[1] + sqrt(S[1, 1] / n1) * qt(0.975, n1 - 1))
c(Xbar[2] - sqrt(S[2, 2] / n1) * qt(0.975, n1 - 1),
  Xbar[2] + sqrt(S[2, 2] / n1) * qt(0.975, n1 - 1))
c(Xbar[3] - sqrt(S[3, 3] / n1) * qt(0.975, n1 - 1),
  Xbar[3] + sqrt(S[3, 3] / n1) * qt(0.975, n1 - 1))
c(Xbar[4] - sqrt(S[4, 4] / n1) * qt(0.975, n1 - 1),
  Xbar[4] + sqrt(S[4, 4] / n1) * qt(0.975, n1 - 1))
```

   Let n1 be the sample size of the first station. Can these confidence intervals give you the correct confidence level? State also the reason.

(d) Write down the MANOVA model equation that tests whether different statitions have the same *PM25*, *SO2*, *NO2*, and *CO*. What is the null hypothesis in terms of your MANOVA model?

(e) Specify the assumptions in order to apply MANOVA.

(f) A statistician has applied PCA to the correlation matrix of the variables *PM25*, *SO2*, *NO2*, and *CO*. The following scree plot is obtained.



How many principal components would you like to choose? State also the reason.

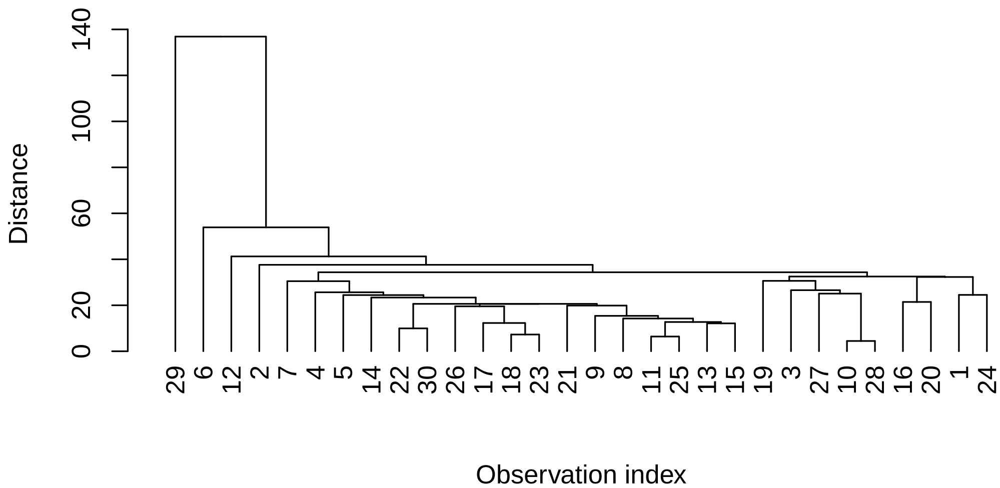
(g) Let the eigenvalues of the correlation matrix be 3, 0.6,0.3, and 0.1. How much total variation is explained by the first principal component?

(h) The statistician performed orthogonal factor rotation. Are the communalities after rotation the same as the communalities before rotation?

(i) Formulate what types of research questions principal component analysis and factor analysis can answer from this data set, respectively.

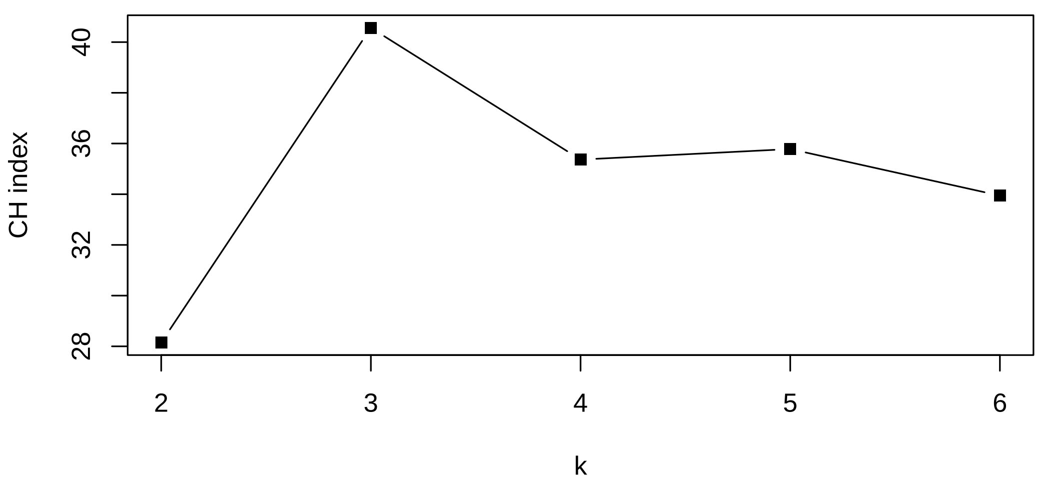
5. (5pt) Suppose that we have a dataset with the following variables. The response variable is *heart disease* with two classes: with heart disease (1) or without heart disease (0). The covariates are *age* (A), *resting blood pressure* (B) , and *cholesterol level* (C), and *maximum heart rate* (H).

   (a) Suppose that the response variable *heart disease* is actually missing. A hierarchical clustering analysis with single linkage has been performed. The resulting dendrogram is shown below.



If three clusters are desired, how would you cluster the observations based on the dendrogram?

(b) We also want to perform a k-means clustering analysis. To determine the number of clusters, we have obtained the following plot of the CH index.



How many clusters would like to choose?

(c) A k-means clustering has been done with the following R code. Can you identify any potential issues? Name at least two, and motivate your answer.

```r
kmeans(Data, centers = 3, nstart = 1)
```

6. (4pt) Let $Z \sim \text{Bernoulli}(\phi)$ be a random variable that indicates the class, where $Z = 1$ or 0:

$$\boldsymbol{X} \mid Z = 1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}), \quad \boldsymbol{X} \mid Z = 0 \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}).$$

   Here $\boldsymbol{\Sigma}$ is known, but ϕ, $\boldsymbol{\mu}_1$, and $\boldsymbol{\mu}_2$ are unknown. Derive a naive Bayes classifier.

7. (4pt) Suppose that we have observed a univariate random sample x_1, x_2, ..., x_n from K populations, but we do not observe which population each observation comes from. We assume that

$$\begin{aligned} X \mid Z = k &\sim N_q\left(\mu_k, \sigma_k^2\right), \\ P(Z = k) &= p_k, \end{aligned}$$

   where σ_k^2 are known. Find the expression of p_k and μ_k that maximizes the conditional expectation computed from the E-step of an EM algorithm.