Departamento de Matemática

Multivariate Analysis

Master in Eng. and Data Science & Master in Mathematics and Applications

1^{st} Test - Part II | 1^{st} Semester – 2020/2021

Duration: 45 minutes | 04/02/2021 – 15:30

Please justify conveniently your answers

If the first letter of your second name is between "A" and "L" solve **Group II - Version A**, otherwise solve **Group II - Version B**.

Any wrong choice of Group II Version will not be classified.

Group II - Version A **10.0 points**

1. Let U be a random variable with standard normal distribution and $\boldsymbol{a} = (a_1, a_2, a_3)^t$ a vector of positive constant (i.e. $a_i > 0,\ i = 1, 2, 3$). Suppose $\boldsymbol{X} = (a_1U, a_2U, a_3U)^t$. What are the principal components based on the correlation matrix of $\boldsymbol{X}$? (2.5)

2. The number of deaths in the 50 USA states were reported according to 7 death causes: accident, cardiovascular, cancer, pulmonary, pneumonia flu, diabetes, and liver. A principal component analysis of the data was carried out, based on the original variables and the eigenvalues of the sample covariance matrix are: {8069.40, a, 76.03, 25.21, 10.45, 5.76, 3.47}.

   (a) Knowing that the trace of the sample covariance matrix is 8379.54 determine a. (1.0)

   (b) Decide how many sample principal components to retain. (2.0)

   (c) Let (3.0)

   $$\boldsymbol{\gamma}_1 = (-0.06, 0.94, 0.34, 0.03, 0.02, 0.03, 0.01)^t \quad \text{and} \quad \boldsymbol{\gamma}_2 = (-0.34, -0.34, 0.86, 0.01, -0.11, 0.09, 0.11)^t$$

   be the two first eigenvectors of the sample covariance matrix. Define the first two sample principal components and interpret them.

   (d) Comment on the interest and limitations of this analysis. (1.5)

If the first letter of your second name is between "A" and "L" solve **Group II - Version A**, otherwise solve **Group II - Version B**.

Any wrong choice of Group II Version will no be classified.

Group II - Version B **10.0 points**

1. Let U be a random variable with uniform distribution in $(0, 1)$ and $\boldsymbol{a} = (a_1, a_2, a_3)^t$ a vector of positive constant (i.e. $a_i > 0,\ i = 1, 2, 3$). Suppose $\boldsymbol{X} = (a_1U, a_2U, a_3U)^t$. What are the principal components based on the correlation matrix of $\boldsymbol{X}$? (2.5)

Reminder: If $X \sim Uniform(0, 1)$ then $\mathrm{E}(X) = 1/2$ and $\mathrm{Var}(X) = 1/12$.

2. The number of deaths in the 50 USA states were reported according to 7 death causes: accident, cardiovascular, cancer, pulmonary, pneumonia flu, diabetes, and liver. A principal component analysis of the data was carried out, based on the standardized variables and the eigenvalues of the sample correlation matrix are: $\{3.40, 1.23, 1.06, 0.61, 0.43, 0.22, 0.05\}$.

   (a) Decide how many sample principal components to retain. (2.0)

   (b) Let

   $$\boldsymbol{\gamma}_1 = (a, -0.50, -0.52, -0.30, -0.27, -0.40, -0.18)^t \quad \text{and} \quad \boldsymbol{\gamma}_2 = (0.28, 0.13, -0.10, 0.38, 0.70, -0.37, -0.37)^t$$

   be the two first eigenvectors of the sample correlation matrix.

   i. Determine a. (1.0)

   ii. Define the first two sample principal components and interpret them, where the vector of sample means is $\bar{\boldsymbol{x}} = (44.31, 398.53, 178.41, 26.49, 21.04, 14.84, 10.55)^t$. (3.0)

   Reminder: In case you have not solve the previous question consider $a = 0.40$.

   (c) Comment on the interest and limitations of this analysis. (1.5)