Multivariate data analysis: Assigment 1

1 Around multivariate normal distributions [5 marks]

Let $A_1, \dots, A_n$ be $n < p$ matrices of dimension $p \times p$ such that:

- $A_k A_j = 0$ for $j \neq k$.
- $A_j^2 = A_j$;
- A_j is symmetrical of rank N_j
- $\sum_j N_j \leq p$.

1. Let X, Y be two random normal vectors with values in $\mathbb{R}^p$ and $\mathbb{R}^q$. We write their means μ_X and μ_Y and their covariances Σ_X and Σ_Y. Assuming that $\text{cov}(X, Y) = 0$, show that X and Y are independent.

   Clue: We can study the characteristic function of (X, Y).

   Let $X \sim \mathcal{N}_p(0, I_p)$.

2. show that $\mathbb{E}[A_j X] = 0$;
3. show that $\text{cov}(A_j X) = A_j$;
4. show that the joint law of $(A_1 X, \dots, A_n X)$ is normal and that $(A_j X)_j$ are mutually independent.
5. show that we can write: $\|A_j X\|^2 = Y^\top \Lambda Y$ where $\Lambda = \text{daig}(\lambda_1^j, \dots, \lambda_p^j)$ is the diagonal form of A_j, and Y is a random vector whose law is to be defined.
6. show that the eigenvalues of A_j are in $\{0, 1\}$. Deduce that $N_j = Card(i, \lambda_i^j = 1)$, and that $\|A_j X\|^2 \sim \chi^2_{N_j}$.
7. deduce that $\|A_1 X\|^2, \dots, \|A_n X\|^2$ are mutually independent.

We will now apply this result.

2 Test of normal mean equality [5 marks]

In your garden, you have k potato plants, when the harvest comes, you measure the mass of the tuber given by each plant. We write n_i the number of tuber given by the ith plant, and we assume each tuber mass measured to be independent with distribution $\mathcal{N}(m_i, \sigma^2)$ (that is, each tuber collected in the same plant has the same distribution, between two plants only the mean changes).

The goal is to know whether all the potato plant follow the same distribution.

Overall the statistical model writes:

$$\left(\mathcal{N}(m_1, \sigma^2)^{\otimes n_1} \otimes \cdots \otimes \mathcal{N}(m_k, \sigma^2)^{\otimes n_k} \mid m_1, \ldots, m_k \in \mathbb{R}^+, \sigma > 0\right).$$

Note that although it is possible under this model to observe negative masses, we neglect this.

1. Let $n = n_1, \ldots, n_k$, we write the previous sample as a Gaussian vector X which is the concatenation of all the observations in $\mathbb{R}^n$. Write its mean μ and its covariance.

   The test problem is to know whether $m_1 = \cdots = m_k$. This is hypothesis H_0.

2. write the hypothesis H_1.

   With a geometric interpretation, the test problem writes, with μ the mean of the total Gaussian vector and $V = Span((1, \ldots, 1)^\top)$:

   $$H_0 : \ \mu \in V \text{ v.s. } H_1 : \ \mu \notin V.$$

   In the following we write u_F the projection of u onto a vectorial subspace F. A natural rejection area for the test is of the form $\{u \in \mathbb{R}^n \mid \|u_E - u_F\| \geq s\}$. This corresponds to the fact that H_0 is rejected if the projections of the observations onto E and V are significatively different.

   Now we have to compute the distribution of $\|X_E - X_V\|$ under H_0.

3. By the previous exercise, show that $\|X_E - X_V\| \sim \sigma^2 \chi^2_{k-1}$.

   If σ was known, it would be finished, in our case further adaptations needs to be made.

4. by using the previous exercise, show that $X - X_E = \varepsilon - \varepsilon_E$ is independent from $X_E - X_v = \varepsilon_E - \varepsilon_V$, where $X = \mu + \varepsilon$ and ε is a random vector to define.

5. For $u \in \mathbb{R}^n$, we write:

   $$F(u) = \frac{\|u_E - u_V\|^2/k - 1}{\|u - u_E\|^2/n - k}.$$

6. Show that under H_0, $F(X)$ follows a known distribution that does not depend on σ. We might want to check for known properties of the previous distributions.

7. Write the rejection area for the test.