Devoir maison

Exercise 1 : Gaussian mixture

First consider a 2-d Gaussian vector $Z_1 = (X, Y)$ centered in the origin and with a covariance matrix $\text{Diag}(\sigma_1^2, \sigma_2^2)$, with $\sigma_1 > \sigma_2$. This distribution is subject to a rotation of angle θ from x axis and a translation of range L along this new axis. Corresponding variable is denoted $Z_2 = (V, W)$.

1 **Show that equi-probability curves of Z_1 have cartesian equations $\frac{x^2}{\sigma_1^2} + \frac{y^2}{\sigma_2^2} = C$.**

2 **This linear transformation can be formalized as $Z_2 = AZ + T$. What are A and T in this case ?**

3 **Is Z_2 a Gaussian vector ? (Justify)**

4 **What are eigenvectors associated with the Principal Component Analysis of Z_2 ? What is the contribution of the first PCA factor, depending on σ_1 and σ_2 ?**

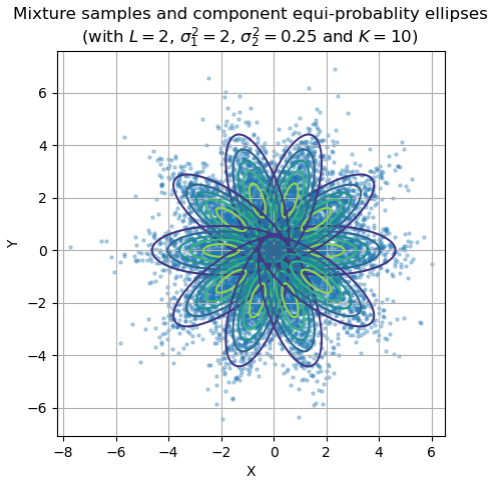
5 **Express equi-probability curves of Z_2 depending on A,T,σ_1 and σ_2.**

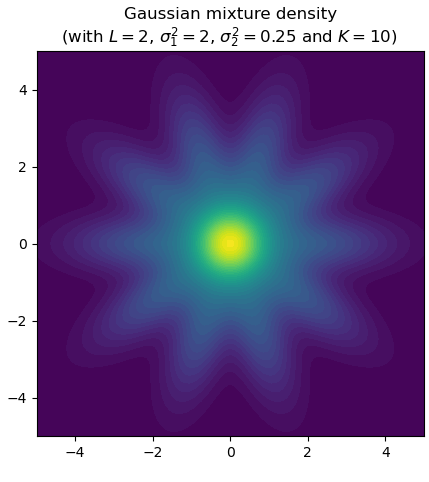
6 **Write, in Python or pseudo-code, a function *is_in_ellipse(samples, A, mu_1, mu_2 ,sig_1, sig_2, C)* that returns a boolean vector indicating whether each sample is inside Z_2 equi-probability ellipse of order C.** [1]

7 **Prove in general that $\text{Cov}(\text{AZ} + \text{T}) = \text{ACov}(\text{Z})\text{A}^{\text{T}}$.**

Now consider $Z_k \sim \mathcal{N}(\mu_k, \Sigma_k)$ with previously defined transformation from Z with angle θ_k, so each Z_k is associated with a rotation matrix A_k and a translation μ_k. From these variables Z_k, a mixture variable $M = (M_1, M_2)$ with $K > 2$ components is constructed, meaning that there is discrete variable $G \in \{1, \dots, K\}$ (independent from Z_k) such that $M = Z_k$ if $G = k$. This means $\mathbb{P}(M/G = k)$ is the probability of a Gaussian distribution of parameters (μ_k, Σ_k). Moreover the probability over mixture groups is considered uniform : $\mathbb{P}(G = k) = p_k = \frac{1}{K}$ and angles correspond to unit circle fractions : $\theta_k = \frac{2k\pi}{K}$.

In this context, we can show that the mixture density is the average of mixture components densities : $f_M = \sum_{k=1}^{K} p_k f_{Z_k}$ and so as the expected value of M : $\mu_M = \mathbb{E}[M] = \sum_{k=1}^{K} p_k \mu_k$.





[1] "sig_1, sig_2" refer to original σ_1, σ_2 and "mu_1, mu_2" refer to translation T.

8 Calculate μ_k and Σ_k in function of θ_k.

9 Show that $\mathrm{Cov(M)} = \sum_{\mathrm{k=1}}^{\mathrm{K}} \mathrm{p_k Cov(Z_k)} + \sum_{\mathrm{k=1}}^{\mathrm{K}} \mathrm{p_k} \mu_\mathrm{k} \mu_\mathrm{k}^\mathrm{T} + \mu_\mathrm{M} \mu_\mathrm{M}^\mathrm{T}$

Reminder :

- $\sum_{k=1}^{K} \cos(\frac{2k\pi}{K})^2 = \sum_{k=1}^{K} \sin(\frac{2k\pi}{K})^2 = \frac{K}{2}$
- $\sum_{k=1}^{K} \cos(\frac{2k\pi}{K}) \sin(\frac{2k\pi}{K}) = 0$

10 Show that, in the previously defined context : $\mu_M = (0,0)$, $\sum_{k=1}^{K} p_k \mu_k \mu_k^T = \mathrm{Diag(L,L)}$ **and** $\sum_{k=1}^{K} p_k \mathrm{Cov(Z_k)} = \mathrm{Diag}(\frac{\sigma_1^2+\sigma_2^2}{2}, \frac{\sigma_1^2+\sigma_2^2}{2})$

11 What would be then the result of a Principal Component Analysis on variable M ? Explain why is it the worst case scenario in terms of dimension reduction.

Consider now a closed shape $\mathcal{A}$, denote $\pi_{in} = \mathbb{P}(M \in \mathcal{A})$ and $\pi_{out} = 1 - \pi_{in}$. Then consider n i.i.d samples $(M^{(1)}, \dots, M^{(n)})$ drawn from the same distribution of M and $Q^{(i)} = \mathbb{1}_{\{M^{(i)} \in \mathcal{A}\}}$. Denote $b \in]0, \min(\pi_{in}, \pi_{out})[$ and $I =]\pi_{in} - b, \pi_{in} + b[$.

12 Show that $\widehat{\pi} = \frac{1}{n} \sum_{k=1}^{n} Q^{(i)}$ is the Maximum Likelihood Estimator of π_{in}.

13 What is the exact probability $\mathbb{P}(\widehat{\pi} \in I)$? What is the asymptotic probability of this interval ? (Depending on n)

Assume $\mathcal{A}_C$ is the union of K equi-probability ellipses of same parameter C, associated with Z_k variables. Denote $\pi_{in} = g(C)$ and $C_{0.95}$ the minimum value of C such that $\mathbb{P}(M \in \mathcal{A}_C) > 0.95$. (All other parameters K, L, σ_1, σ_2 are fixed).

14 Justify that g is a strictly increasing function.

15 Write, in Python or pseudo-code, a function *is_in_K_ellipses(samples, rotations, means,sig_1, sig_2, C)* that returns a boolean vector indicating whether each sample is inside $\mathcal{A}_C$. [2]

16 Using is_in_K_ellipses, write a code in Python or pseudo-code to determine empirically the value of $C_{0.95}$ and another code to determine the area of shape $\mathcal{A}_{C_{0.95}}$.

17 Depending on n and g^{-1}, determine an asymptotic confidence interval for this estimation of $C_{0.95}$.

[2] "rotations" refers to an array with rotation matrices A_k and "means" refers to an array with translations μ_k.

Exercise 2 : Video game data analysis

Age of empire is a strategical video game where two players choose their civilization to develop by gathering resources, building settlements, unlocking technologies and creating an army to destroy opponent's base.

A database with contextual informations about thousands of games is available online [3]. First column indicates the map on which the game is played, second column is the duration of the game, "elo" refers to the average ranking of the two players, "p1 civ" and "p2 civ" are the civilizations names chosen by the two players and last column indicates who won.

In this exercise, $X \in \{1, \dots, p\}$ denotes the row variable and $Y \in \{1, \dots, q\}$ the column variable and we consider data are drawn from n i.i.d variables $(X_1, \dots, X_n)$ and $(Y_1, \dots, Y_n)$ following same distribution as X and Y.

Then we denote :

- $C_{ij} = \sum_{k=1}^{n} \mathbb{1}_{\{X_k, Y_k = i,j\}}$
- $C_{i\cdot} = \sum_{k=1}^{n} \mathbb{1}_{\{X_k = i\}}$, $C_{\cdot j} = \sum_{k=1}^{n} \mathbb{1}_{\{Y_k = j\}}$
- $E_{ij} = \frac{C_{i\cdot} C_{\cdot j}}{n}$
- $p_{ij} = \frac{C_{ij}}{n}$
- $p_{i\cdot} = \sum_{j=1}^{q} p_{ij}$ $\left(\text{ resp. } p_{\cdot j} = \sum_{i=1}^{p} p_{ij}\right)$

map	duration	elo	p1 civ	p2 civ	winner
Arabia	3445	1104	Vikings	Mayans	0
Arena	2932	884.5	Britons	Goths	0
Four Lakes	2712	994.5	Khmers	Huns	1
...	...	...	...	...	...

Table 1: Example of extract of AoE games database

In order to study these data, Factorial Correspondence Analysis is applied on several couples of variables. Consider that all data are stored in a DataFrame named "df".

1 Which of the defined variables corresponds to the "contingency table" ?

2 What is the residuals matrix ?

3 What is $d^2_{\chi^2}$ the chi-squared distance (using previous notations) and what does it quantify ?

4 Show that : $\sum_{i=1}^{p} \sum_{j=1}^{q} \frac{\left(p_{ij} - p_{i\cdot} p_{\cdot j}\right)^2}{p_{i\cdot} p_{\cdot j}} = \frac{1}{n} d^2_{\chi^2}$

5 Explain what is done by the following Python code.

```
In[]:
contingency_table1 = pd.crosstab(df['p1_civ'][df["elo"] > 1800], df['map'][df["elo"] > 1800])

contingency_table2 = pd.crosstab(df['p1_civ'][df["elo"] < 800], df['map'][df["elo"] < 800])

deg1 = (contingency_table1.shape[0]-1)*(contingency_table1.shape[1]-1)
deg2 = (contingency_table2.shape[0]-1)*(contingency_table2.shape[1]-1)

Out[]:
Chi-squared Table :
[1595.33338286 1668.81956351 1690.05365201 1776.30548921]
[1405.33338956 1474.37234405 1494.34592441 1575.58483135]
```

Assume their is a function $chi\_square\_dist(C)$ that computes the chi-squared distance associated with contingency table C.

[3] https://www.kaggle.com/datasets/nicoelbert/aoe-matchups

```
In[]:
print("Chi-squared distance :")
print(chi_square_dist(contingency_table1)
print(chi_square_dist(contingency_table2)

Out[]:
Chi-squared distance :
2716.639822149627
1241.8086170719955
```

6 Are "deg1" and "deg2" equal ? How can we explain that ?

7 How do you interpret these Chi-squared distance results ?

8 What can you assume about map and civilization picks in this game, based on these observations ?

We would like to check if one aspect of the gameplay, the chosen civilization type, impacts the game duration. In other world, how much the game is likely to reach a certain duration depending on played civilizations. The following figures present histograms of game duration (between minimum and maximum duration in the dataset).

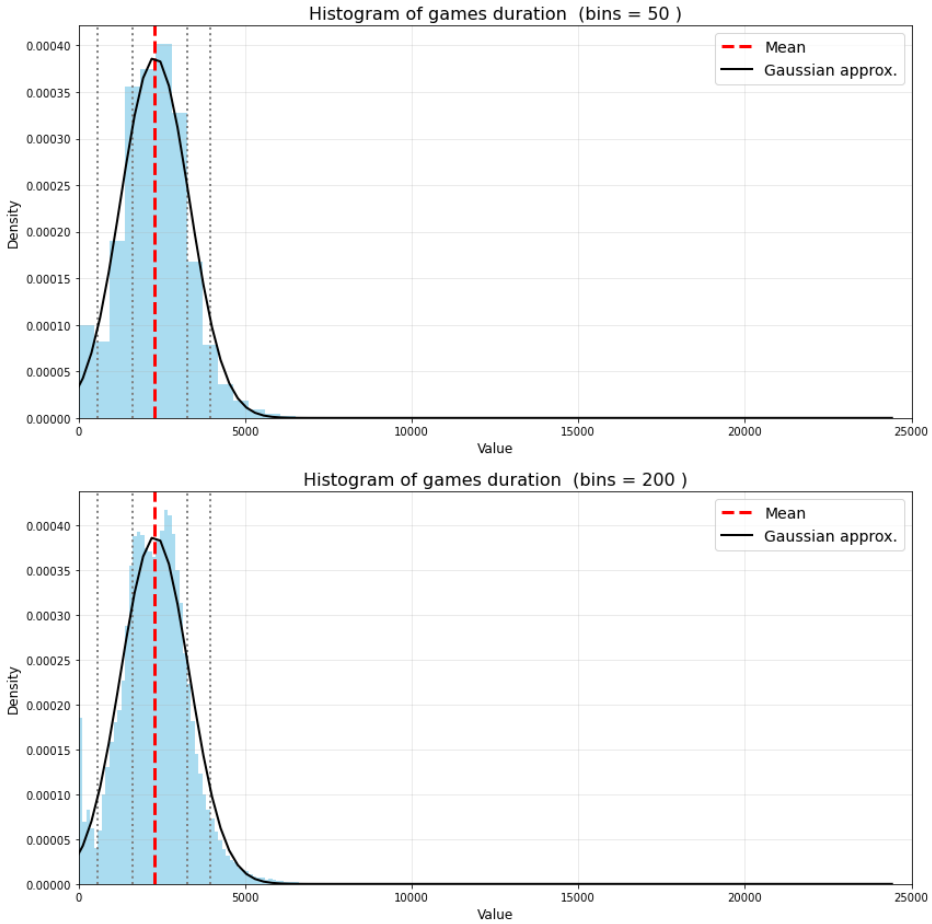


Figure 1: AoE game duration distribution

9 What can we observe about extreme values of game duration ? How can we explain that ? What do you suggest to deal with these aberrant data ?

10 Formulate an assumption about the game explaining local minimum around the mean.

In first approximation these data are modeled by a Gaussian distribution (as represented in the Figure). As game durations represent quantitative data, they are sorted in 6 groups depending on quantiles and empirical mean. Denoting $\widehat{m}$ the empirical mean and q_α quantile of order α : $G_0 = [0, q_{0.05}], G_1 = [q_{0.05}, q_{0.25}], G_2 = [q_{0.25}, \widehat{m}], G_3 = [\widehat{m}, q_{0.85}], G_4 = [q_{0.85}, q_{0.95}], G_5 = [q_{0.95}, +\infty[$.

11 Why the median of a Gaussian distribution is equal to its mean ? Find an example of distribution where the mean is outside $[q_{0.25}, q_{0.85}]$.

12 Explain why FCA is hardly applicable on quantitative data.

Factorial Correspondence Analysis is performed with rows representing civilizations and columns representing duration groups. To obtain a classification of civilization based on game duration a k-means clustering is applied on the first two factors projection, initializing cluster centers with column profiles coordinates. Each row profile dot color represent a cluster and red lines represent frontiers between clustering areas.

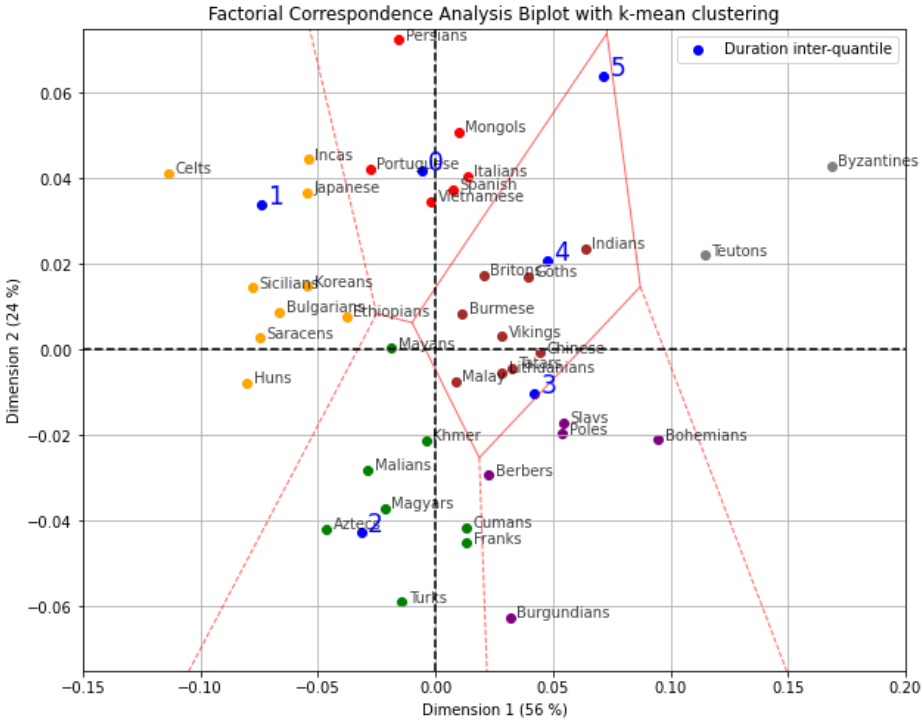


Figure 2: First two FCA axes projection with k-means clustering

13 **Which civilizations should be chosen for a very short game ? For a very long one ?**

14 **What is clustering area frontiers property regarding clustering centers ?**

15 **What does k-means update throughout iterations ?**

16 **In Figure 2, we observe that several column profiles are in the same clustering area. How do you explain that ?**