Final SY09 Printemps 2020

— **Répondre de manière manuscrite sur des feuilles au format A4.**
— **Chaque feuille doit comporter vos nom, prénom et signature.**
— **La qualité de la présentation sera prise en compte dans la notation.**

Exercice 1 (3 points)

On dispose d'un ensemble de trois individus mesurés par deux variables quantitatives. En effectuant l'ACP sur ces données, on obtient les axes factoriels

$$\mathbf{u}_1 = (0.816, -0.578)^T \quad \text{et} \quad \mathbf{u}_2 = (-0.578, -0.816)^T,$$

et les composantes principales

$$\mathbf{c}_1 = (-3.03, -1.97, 5)^T \quad \text{et } \mathbf{c}_2 = (0.916, -1.06, 0.139)^T.$$

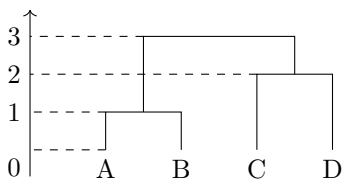
1. Reconstituer le tableau X_c des données centrées.

2. Quelles sont les valeurs propres de la matrice de variance-covariance associée à cette ACP, et les pourcentages d'inertie expliquée par chacun des axes ?

3. Sachant que les moyennes des variables initiales X_1 et X_2 étaient respectivement 5 et 4, déterminer le tableau de données initial.

Exercice 2 (2 points)

À la suite d'une analyse ascendante hiérarchique, on obtient la hiérarchie indicée suivante.



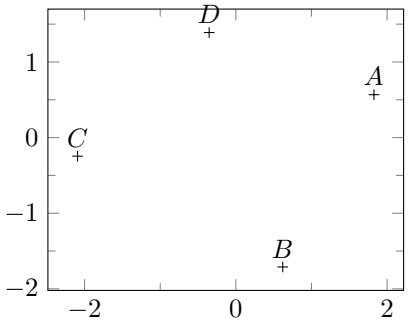
1. Placer les 4 points sur l'axe ainsi que leur inter-distance de telle sorte qu'une analyse ascendante hiérarchique avec le critére d'agrégation minimum donne la hiérarchie indicée ci-dessus.

2. On souhaite à présent faire une CAH avec le critère d'agrégation maximum. Quel est l'indice de l'ensemble $\{A, D\}$? Justifier.

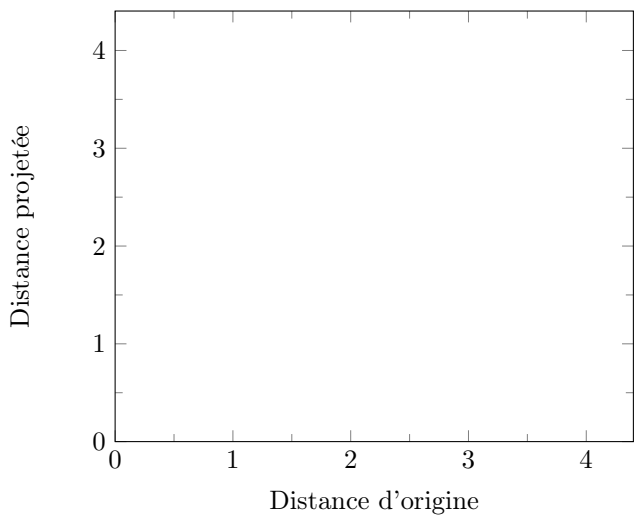
Exercice 3 (2 points)

On dispose d'une matrice de distance (ci-dessous, à gauche) portant sur 4 éléments, A, B, C et D. À la suite d'une AFTD projetée en dimension 2, on obtient la représentation suivante (ci-dessous, à droite).

	A	B	C	D
A	0.00	1.80	4.00	0.39
B	1.80	0.00	2.58	3.22
C	4.00	2.58	0.00	1.06
D	0.39	3.22	1.06	0.00

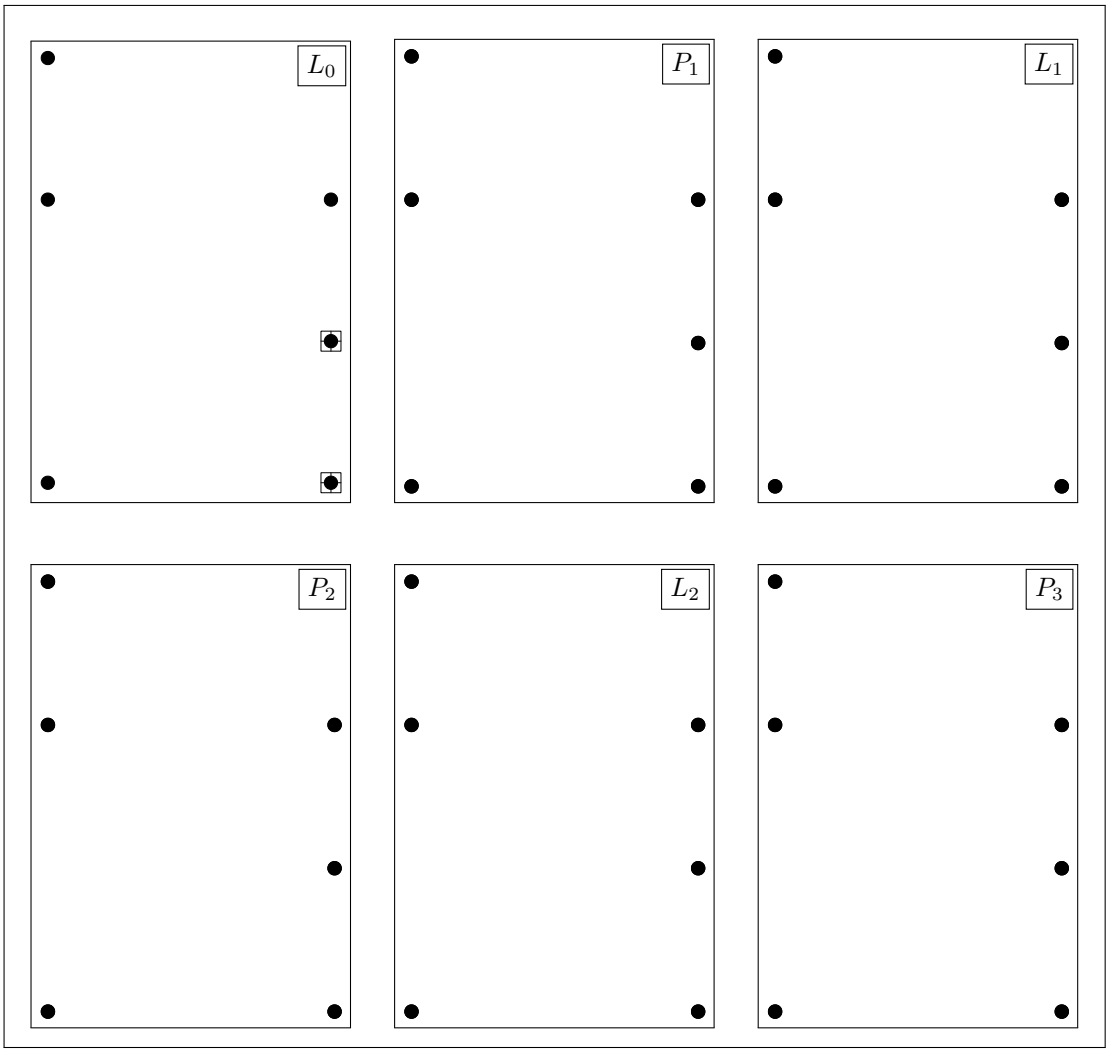


1. Tracer le diagramme de Shepard ; quel est le couple de points le moins bien représenté lors de la projection ?



Exercice 4 (3 points)

1. On considère cette fois 6 points situés dans le plan. On applique l'algorithme des centres mobiles avec 2 centres initiaux repérés dans l'étape L_0 par 2 carrés. Donner la suite des autres étapes en entourant à chaque fois les points constituant chaque groupement et en dénotant par un carré les 2 centres courants.



Exercice 5 (5 points)

On considère un problème de discrimination à $g = 2$ classes gaussiennes caractérisées par les paramètres suivants :

$$\boldsymbol{X}|\omega_k \sim \mathcal{N}(\mu_k, \Sigma_k), \quad \mu_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}, \quad \mu_2 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix};$$

1. Donner l'expression de la règle de décision optimale au sens de Neyman-Pearson. Quelle est la forme de la frontière de décision ? On l'exprimera en fonction de x_1, x_2 et on calculera la valeur exacte du seuil critique.

2. On suppose que les classes sont présentes en proportions π_1 et $\pi_2 = 1 - \pi_1$, connues. Calculer la règle de décision optimale au sens de Bayes, et exprimer la frontière en fonction de x_1 et x_2.

3. Calculer la probabilité d'erreur de Bayes $\varepsilon^\star$, puis de la règle de Neyman-Pearson ε_{NP}, pour $\pi_1 = \pi_2 = 1/2$.

Exercice 6 (4 points)

On considère un problème de discrimination à deux classes (avec ω_1 la classe « triangle », et ω_2 la classe « carré ») ; la figure 1 représente la partition associée à un arbre binaire de classification en cours de construction.

1. Donner l'arborescence décrivant l'arbre complet obtenu en utilisant l'algorithme CART. On représentera les nœuds *en les numérotant dans l'ordre dans lequel ils sont créés par l'algorithme.*

2. Quelle est la complexité $\xi(\mathcal{A})$ de l'arbre $\mathcal{A}$ ainsi obtenu ?

3. Dans l'arbre complet, calculer le gain (en termes d'indice de Gini) pour chacune des trois divisions terminales ; quelle est la moins informative ?

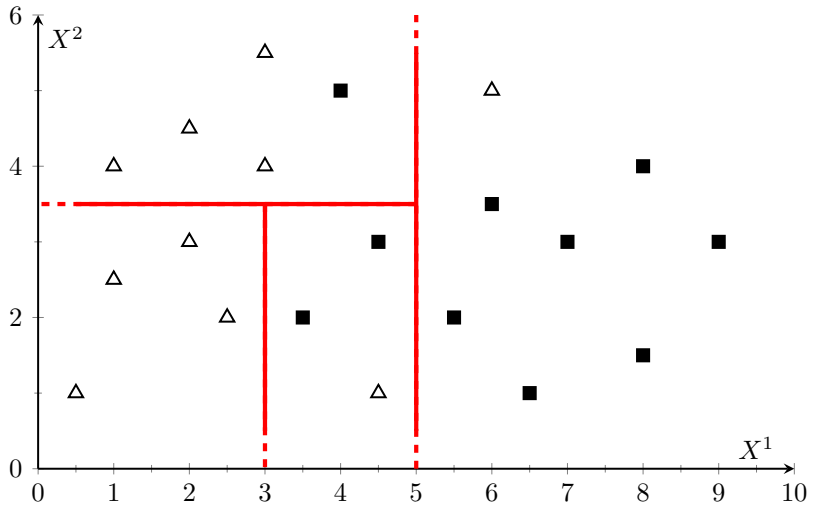


FIGURE 1 – Partition induite par un arbre en cours de construction.

Exercice 7 (7 points)

Dans le cadre d'un contrôle de qualité, on cherche à prédire si des briques de lait ont été correctement pasteurisées ou non : pour ce faire, on mesure le nombre de spores X de *clostridium botulinum* présentes dans une quantité de liquide donnée (fixe) après pasteurisation. On suppose que X suit une loi de Poisson dépendant du fait que la pasteurisation est correcte ($Z = 1$) ou non ($Z = 0$) :

$$p_k(x) = \mathbb{P}(X = x | Z = k) = \exp(-\lambda_k)\frac{(\lambda_k)^x}{x!}, \text{ pour } k = 1, 0.$$

On note également $\pi = \mathbb{P}(Z = 1)$ la probabilité que la pasteurisation est a priori correcte.
On cherche à mettre en œuvre une stratégie permettant de détecter les erreurs de pasteurisation basée sur des mesures de X sur des échantillons pris au hasard. On veut pour cela apprendre les paramètres λ_1, λ_0 et π permettant de discriminer les deux classes. On dispose pour cela d'un ensemble d'apprentissage $(x_1, z_1), \ldots, (x_n, z_n)$, où $x_i \in \mathbb{N}$ correspond à la mesure du nombre de spores effectuée sur le i^{e} tube à essai, et $z_i \in \{1, 0\}$ indique la qualité de pasteurisation.

1. Rappeler le principe de la stratégie de décision minimisant la probabilité d'erreur de classement.

2. Calculer l'expression de la probabilité $\mathbb{P}(X = x, Z = z)$, en fonction de $p_1(x)$ et $p_0(x)$, π et z.

3. En déduire $\mathbb{P}(X = x)$ puis $\mathbb{P}(Z = z | X = x)$.

4. Calculer la log-vraisemblance jointe $\ln L\left(\theta; (x_1, z_1), \ldots, (x_n, z_n)\right)$, où $\theta = (\pi, \lambda_1, \lambda_0)$.

5. Calculer les estimateurs du maximum de vraisemblance des paramètres λ_1 et λ_0.

6. On supposera maintenant qu'on ne compte plus le nombre x de spores dans un volume de lait : on se contente seulement d'identifier la présence ($y = 1$) ou l'absence ($y = 0$) de spores dans ce même volume. Comment peut-on adapter la stratégie de décision développée ci-dessus ? (On ne demande de faire aucun calcul.)