Université de Caen	Examen du 20 avril 2018. Durée 2h00	**Analyse de données**
UFR des Sciences	Par F. Chamroukhi	**L3 Maths/Miashs**

Consignes :
- Sont interdits : Documents, calculettes, téléphones, écouteurs, ordinateurs, tablettes.
- Il est interdit de composer avec un crayon.
- Votre feuille double d'examen doit porter, à l'emplacement réservé, vos nom, prénom, et signature.
- Cette zone réservée doit être cachée par collage.
- Vos feuilles intercalaires doivent être toutes numérotées.
- Le barème est donné à titre indicatif.

Exercice 1 *(4 pts)* Soit (X, Y) un coupe de variables aléatoires réelles et soit $((x_1, y_1), \ldots, (x_n, y_n))$ un échantillon de n observations. Chacune des situations présentées dans la Figure 1 représente le nuage de données d'un échantillon de taille $n = 500$. Pour chaque situation, donner une valeur approchée du cœfficient de corrélation linéaire empirique r et justifier votre réponse.

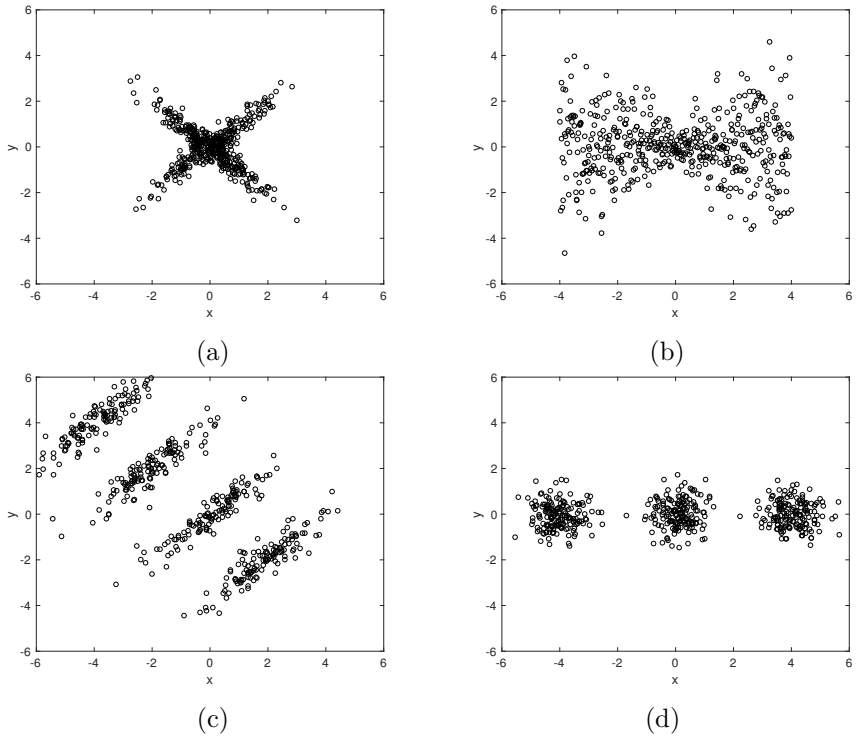

FIGURE 1 – Nuages de données (∘)

Exercice 2 *(6 pts)* On considère un jeu de données multivariées $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ où $\boldsymbol{x}_i \in \mathbb{R}^p$ est un individu décrit par p variables réelles. On suppose que les variables sont potentiellement corrélées et que p est potentiellement grand, et on cherche à réduire la dimension en projetant ce jeu de données dans un espace de dimension M plus réduite et ce au sens de l'ACP. Soient $\boldsymbol{x} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i$ le vecteur de la moyenne empirique des individus et $\mathbf{S} = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top$ la matrice de variances-covariances empirique.

1. On suppose que l'on projette les données dans un espace de dimension 1, i.e., une droite de vecteur directeur $\boldsymbol{u}_1$, que l'on suppose normé (i.e.,$\boldsymbol{u}^\top \boldsymbol{u} = 1$). Donner l'expression de la variance des données projetées dans cet espace et exprimer là en fonction de $\mathbf{S}$ et $\boldsymbol{u}_1$. On la notera v_1.
2. En sachant que l'ACP consiste à maximiser la variance dans l'espace projeté v_1, montrer que le

vecteur $\boldsymbol{u}_1$ définissant cet espace correspond au vecteur propre associé à la plus grande valeur propre de $\mathbf{S}$.

3. On suppose maintenant que l'on projette les données dans un espace de dimension $M \leq p$ déterminé par le sous espace vectoriel de base orthonormée $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M)$, i.e., $\boldsymbol{u}_j^\top \boldsymbol{u}_j = 1 \; \forall j$ et $\boldsymbol{u}_j^\top \boldsymbol{u}_k = 0 \; \forall j \neq k$. Montrer que les vecteurs de cette base sont les vecteurs propres de $\mathbf{S}$ organisés selon l'ordre décroissant des valeurs propres associées.
4. En interprétant le lien entre les valeurs propres et les variances expliquées par chaque axe de l'espace projeté, donner une idée pour choisir la dimension M.

Exercice 3 *(10 pts)* On considère un échantillon aléatoire indépendant $((X_1, Y_1), \ldots, (X_n, Y_n))$ du couple (X, Y) où $X \in \mathbb{R}$ est une variable explicative et $Y \in \mathbb{R}$ une variable expliquée à prédire. On considère le modèle de régression linéaire pondérée suivant

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{1}$$

où les ε_i sont des v.a indépendantes Gaussiennes centrées de variance $\frac{\sigma^2}{w_i}$, $w_i > 0$ est le poids connu de la $i$ème observation, et $(\beta_0, \beta_1) \in \mathbb{R}^2$ sont les deux cœfficients de régression inconnus. On dispose d'un échantillon indépendant d'observations $((x_1, y_1), \ldots, (x_n, y_n))$ et les poids associées $(w_1, \ldots, w_n)$, à partir desquels on cherche à estimer les paramètres (β_0, β_1) par maximum de vraisemblance.
On sait facilement montrer que selon le modèle (1) la loi de Y_i conditionnellement à $X_i = x_i$, est Gaussienne d'espérance $\beta_0 + \beta_1 x_i$ et de variance $\frac{\sigma^2}{w_i}$ de densité notée $f(y_i | X_i = x_i; \beta_0, \beta_1)$.

1. On rappelle que si Z est une v.a Gaussienne d'espérance μ et de variance v^2, sa densité est alors donnée par $f(z; \mu, v^2) = \frac{1}{\sqrt{2\pi v^2}} \exp\left(-\frac{1}{2v^2}(z-\mu)^2\right)$. En déduire la densité $f(y_i | X_i = x_i; \beta_0, \beta_1)$.
2. Donner l'expression de la fonction de log-vraisemblance

$$\ln L(\beta_0, \beta_1) = \ln f(y_1, \ldots, y_n | X_1 = x_1, \ldots, X_n = x_n; \beta_0, \beta_1). \tag{2}$$

3. Montrer en maximisant (2) que l'estimateur du maximum de vraisemblance (EMV) $\widehat{\beta}_0$ de β_0 et l'EMV $\widehat{\beta}_1$ de β_1 sont donnés par :

$$\begin{aligned} \widehat{\beta}_0 &= \overline{Y}_w - \widehat{\beta}_1 \overline{X}_w, \\ \widehat{\beta}_1 &= \sum_{i=1}^n w_i (X_i - \overline{X}_w)(Y_i - \overline{Y}_w) \Big/ \sum_{i=1}^n w_i (X_i - \overline{X}_w)^2, \end{aligned}$$

où $\overline{X}_w = \frac{\sum_{i=1}^n w_i X_i}{\sum_{i=1}^n w_i}$ est la moyenne empirique pondérée des prédicteurs X_i et $\overline{Y}_w = \frac{\sum_{i=1}^n w_i Y_i}{\sum_{i=1}^n w_i}$ celle des réponses Y_i.

4. On considère maintenant la formulation matricielle du problème. Soit $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top \in \mathbb{R}^2$ le vecteur paramètre du modèle. En remarquant que maximiser (2) revient à minimiser la fonction

$$R_w(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top \mathbf{W} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$$

où $\mathbf{W} \in \mathbb{R}^{n \times n}$ est la matrice diagonale de termes les poids $(w_1, \ldots, w_n)$, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$ est le vecteur des réponses, et $\boldsymbol{X} = ((1, X_1)^\top, \ldots, (1, X_n)^\top)^\top \in \mathbb{R}^{n \times 2}$ est la matrice de design, montrer que l'EMV $\widehat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$ est donné par

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \mathbf{W} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \mathbf{W} \boldsymbol{Y}.$$

5. A quoi correspond l'EMV dans ce cas ?