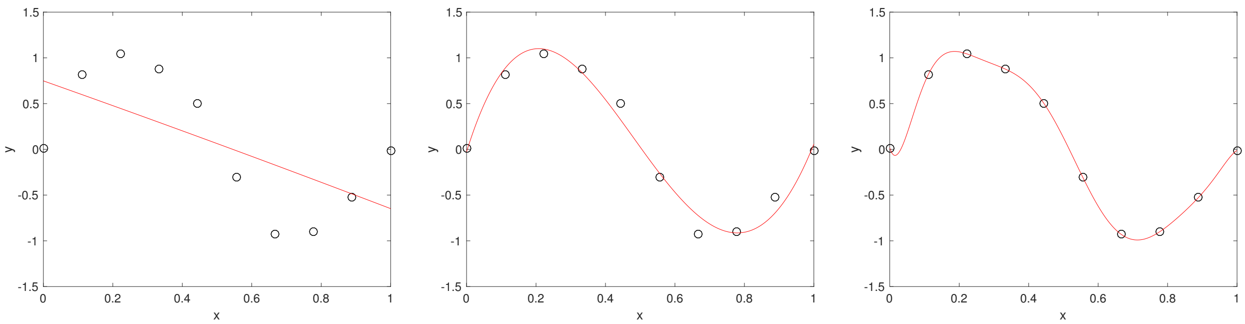
**Université de Caen** | TDs Partie 2 | **Analyse de données**
**UFR des Sciences** | Par Faïcel Chamroukhi | **M1 SAAD/DECIM**

**Exercice 1** On considère un échantillon indépendant $((X_1, Y_1), \ldots, (X_n, Y_n))$ de couples $(X_i, Y_i)$ où $Y_i \in \mathbb{R}$ est la v.a.r expliquée par la v.a.r explicative $X_i \in \mathbb{R}$. On dispose d'un échantillon observé $((x_1, y_1), \ldots, (x_n, y_n))$. On cherche à déterminer un modèle décrivant au mieux la relation entre les observations $y_i$ et les variables $x_i$. Chacune des trois situations présentées dans la Figure 2 montre les données et un modèle ajustée aux données.

1. Donner pour chacune des trois situations une idée sur la valeur du cœfficient de détermination $R^2$. Justifier.
2. Donner une idée sur la valeur du cœfficient de corrélation linéaire empirique $r$. Justifier.
3. Pour la première situation, que peut-on dire sur la valeur du $R^2$ et celle du $r^2$ ? Expliquer pourquoi.
4. Et pour les deux autres situations ?
5. Quel est le meilleur modèle explicatif parmi les trois modèles ? Expliquer pour quelle(s) raison(s).
6. Ce modèle serait-il le meilleur si l'objectif était de prédire $y_i$ pour des nouvelles valeurs $x_i$ ? Expliquer pourquoi



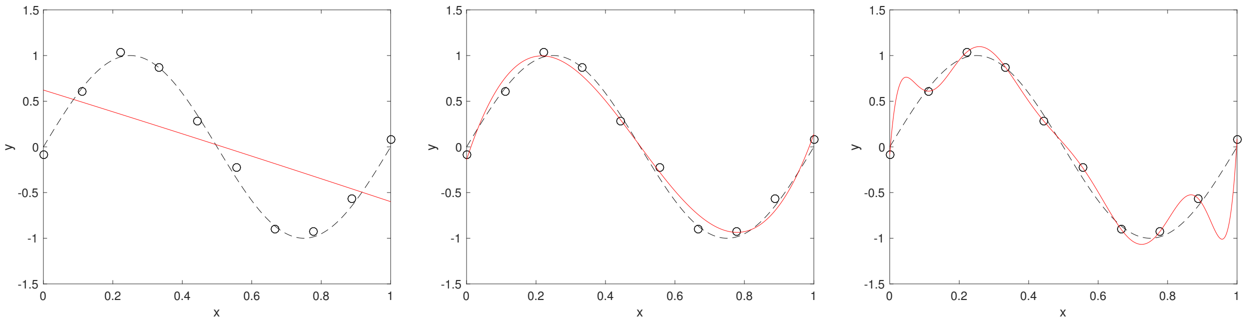

(a) (b) (c)

FIGURE 1 – Nuage de données ($\circ$) et modèle ajusté (—)

**Exercice 2** On considère un échantillon indépendant $((X_1, Y_1), \ldots, (X_n, Y_n))$ de couples $(X_i, Y_i)$ où $Y_i \in \mathbb{R}$ est la v.a.r représentant la prédiction des $X_i \in \mathbb{R}$. On dispose d'un échantillon d'apprentissage $((x_1, y_1), \ldots, (x_n, y_n))$ et d'un échantillon de test $(x_{n+1}, \ldots, x_m)$ issus de cette même population. On cherche à prédire les valeurs $\widehat{y}_i$ étant données des nouvelles valeurs de $x_i$. Chacune des situations présentées dans la Figure 2 montre les données d'apprentissage, le vrai modèle des données, et un modèle de prédiction appris sur ces données en minimisant le risque quadratique.

1. Rappeler la décomposition biais-variance du risque quadratique.
2. Commenter pour chaque situation sur le biais et la variance du modèle.
3. En déduire le meilleur modèle prédictif parmi les trois modèle et expliquer pourquoi
4. Peut-on utiliser le cœfficient $R^2$ pour sélectionner le meilleur modèle parmi les trois ? Expliquer pourquoi.

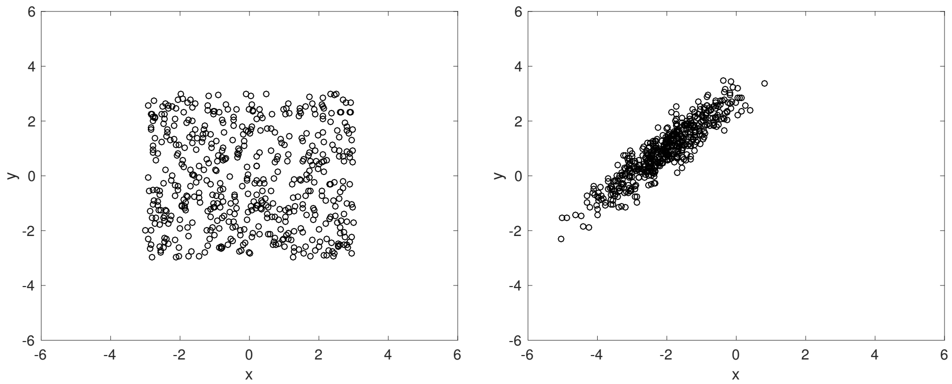
5. Si l'on sait que chacun de ces modèles correspond à un modèle de régression polynômiale de degré $p$. Donner la valeur de $p$ pour la situation (c).
6. Donner l'expression d'un critère pour sélectionner le meilleur modèle, et l'expliquer.



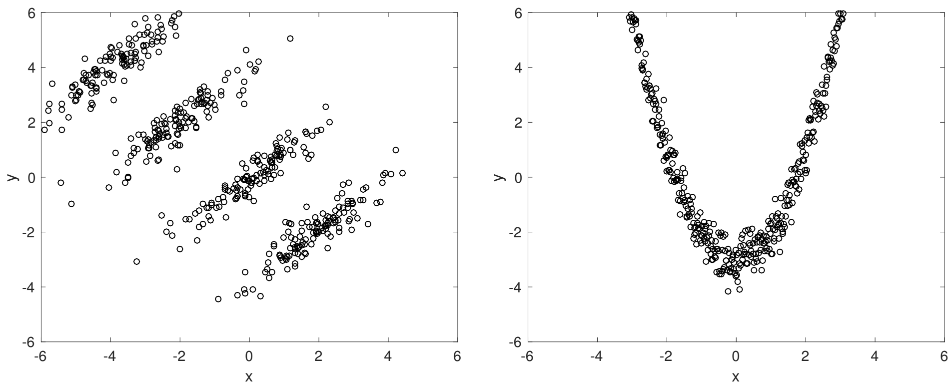

(a) (b) (c)

FIGURE 2 – Nuage de données (○), vrai modèle (--), et modèle de prédiction appris à partir des données (—)

**Exercice 3** On considère un ensemble de données bivariées $((x_1, y_1), \ldots, (x_n, y_n))$ de $n = 500$ observations $(x_i, y_i) \in \mathbb{R}^2$. Analyser les données pour chacune des situations présentées dans la Figure 3.




(a) (b)




(c) (d)

FIGURE 3 – Nuage de données (○)

**Exercice 4** On considère un échantillon indépendant $((\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n))$ d'individus $\boldsymbol{X}_i$ décrits par $d$ variables réelles ($\boldsymbol{X}_i \in \mathbb{R}^d$) d'une population de 2 classes telle que $Y_i \in [\![0, 1]\!]$ est la classe de l'individu $\boldsymbol{X}_i$. On dispose d'un échantillon d'apprentissage $((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n))$ et d'un échantillon de test $((\boldsymbol{x}_{n+1}, y_{n+1}), \ldots, (\boldsymbol{x}_m, y_m))$ issus de cette même population. On s'intéresse à prédire les classes des données des test sur la base d'un modèle probabiliste appris sur les données d'apprentissage. On considère le cas de l'analyse discriminante gaussienne.

La prédiction s'effectue par la règle du maximum a posteriori (MAP) qui consiste à affecter l'individu $\boldsymbol{x}_i$ à la classe $y_i$ maximisant la probabilité a posteriori :

$$y_i = \arg \max_{k \in [\![0,1]\!]} \mathbb{P}(Y_i = k | \boldsymbol{X}_i = \boldsymbol{x}_i; \boldsymbol{\theta}) \tag{1}$$

où $\pi_k = \mathbb{P}(Y_i = k)$ est la probabilité a priori de la classe $k$, $f_{\boldsymbol{X}|Y}(\boldsymbol{x}_i | Y_i = k; \boldsymbol{\theta}_k)$ est sa densité définie par On utilise une modélisation gaussienne pour chacune des classes, i.e.,

$$f_{\boldsymbol{X}|Y}(\boldsymbol{x}_i | Y_i = k; \boldsymbol{\theta}_k) = \frac{1}{(2\pi)^{\frac{d}{2}} |\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^\top {\boldsymbol{\Sigma}_k}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \right\} \tag{2}$$

pour $k \in [\![0, 1]\!]$ avec $\boldsymbol{\theta} = (\pi_1, \boldsymbol{\mu}_0^\top, \boldsymbol{\mu}_0^\top, \mathrm{vech}(\boldsymbol{\Sigma}_0), \mathrm{vech}(\boldsymbol{\Sigma}_1))^T$ étant le vecteur paramètre du modèle.

On suppose que $\pi_0 = \pi_1 = 0.5$ avec $\pi_0 = \mathbb{P}(Y_1 = 0)$ et $\pi_1 = \mathbb{P}(Y_i = 1)$, $\boldsymbol{\mu}_0 = (1, 1)^t$, $\boldsymbol{\mu}_1 = (-1, -1)^t$, et $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \mathbf{I}_2$, $\mathbf{I}_d$ étant la matrice identité de dimension $d$.

La figure 4 montre un jeu de données d'apprentissage ($n = 200$).

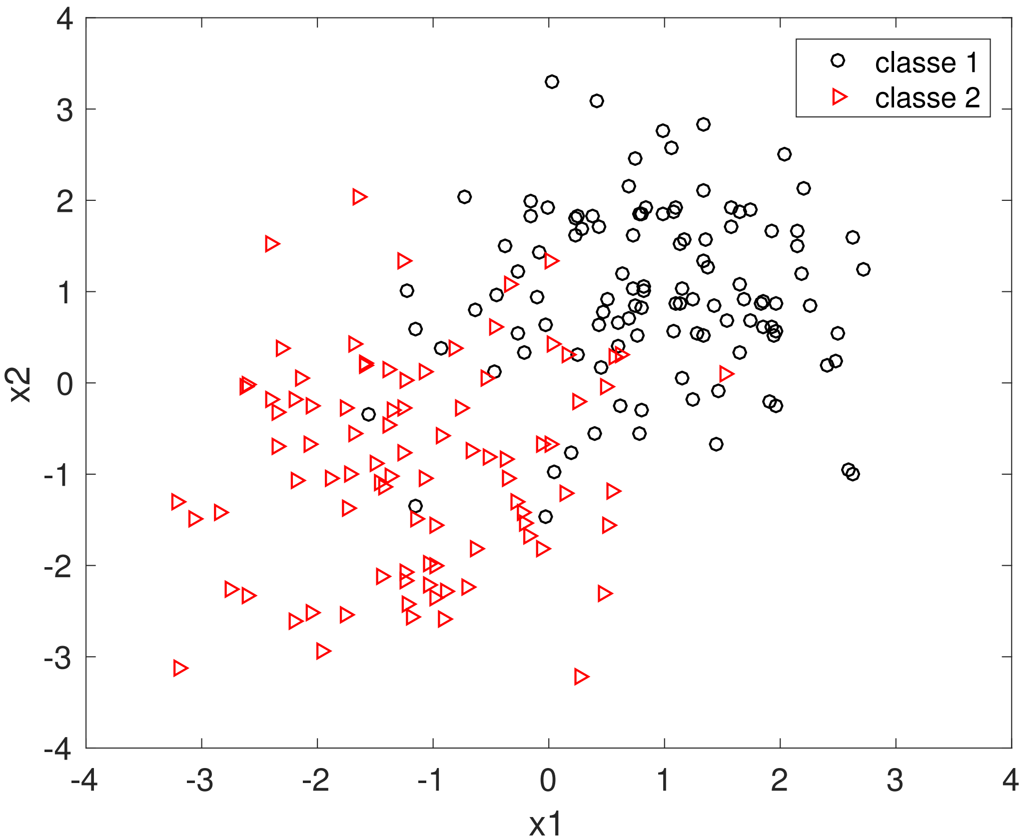



Figure 4 – Nuage des données d'apprentissage ($n = 200$)

1. Donner l'expression de $\mathbb{P}(Y_i = 0|\boldsymbol{X}_i = \boldsymbol{x}_i; \boldsymbol{\theta})$ et $\mathbb{P}(Y_i = 1|\boldsymbol{X}_i = \boldsymbol{x}_i; \boldsymbol{\theta})$
2. Monter que $Y_i = k$ si et seulement si $\boldsymbol{x}_i^\top \boldsymbol{\mu}_k \geq 0$
3. Prédire les classes des observations suivantes :

$$\boldsymbol{x}_1 = (-2, 3)^t, \quad \boldsymbol{x}_2 = (1, -1)^t, \quad \boldsymbol{x}_3 = (-2, 2)^t, \quad \boldsymbol{x}_4 = (1, -2)^t$$

4. Tracer la frontière de décision entre les deux classes
5. Maintenant on suppose que $\boldsymbol{\Sigma}_2 = 2\,\mathbf{I}_2$. Calculer l'équation de la nouvelle frontière de décision et tracer là approximativement
6. Prédire les classes des observations suivantes :

$$\boldsymbol{x}_1 = (-2, 3)^t, \quad \boldsymbol{x}_2 = (1, -1)^t, \quad \boldsymbol{x}_3 = (-2, 2)^t, \quad \boldsymbol{x}_4 = (1, -2)^t$$

**Exercice 5** On considère le cadre de l'exercice précédent. On suppose maintenant que les paramètres sont inconnus et on cherche à les estimer à partir de l'échantillon d'apprentissage par maximum de vraisemblance.

1. Définir les fonction de vraisemblance et de log-vraisemblance

$$\ln L(\boldsymbol{\theta}) = \ln f((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n); \boldsymbol{\theta})$$

2. Calculer l'estimateur du maximum de vraiseblance $\widehat{\boldsymbol{\theta}}$. On traitera le cas où les matrices de covariances sont identiques/différentes
3. Rappeler ses propriétés statistiques

**Exercice 6** On considère le cadre de l'exercice précédent et on suppose que le nombre de classes $K \geq 2$. On suppose que les paramètres du modèle sont inconnus et on cherche à les estimer à partir de l'échantillon d'apprentissage par maximum de vraisemblance.

1. Définir les fonction de vraisemblance et de log-vraisemblance

$$\ln L(\boldsymbol{\theta}) = \ln f((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n); \boldsymbol{\theta})$$

2. Calculer l'estimateur du maximum de vraiseblance $\widehat{\boldsymbol{\theta}}$. On traitera le cas où les matrices de covariances sont identiques/différentes
3. Rappeler ses propriétés statistiques

**Exercice 7** On rappelle que si $(X, Y)$ est un couple de v.a de fonction de densité de probabilité jointe définie pour $|\rho| < 1$ ($\rho$ étant le coefficient de corrélation linéaire) par

$$f_{(X,Y)}(x, y) = \frac{1}{2\,\pi \sigma_X \sigma_Y \sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho^2)} \left( \left( \frac{x-\mu_X}{\sigma_X} \right)^2 - 2\rho \frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X \sigma_Y} + \left( \frac{y-\mu_Y}{\sigma_Y} \right)^2 \right) \right\} \tag{3}$$

alors $X$ est de loi $\mathcal{N}(\mu_X, \sigma_X^2)$ et $Y$ de loi $\mathcal{N}(\mu_Y, \sigma_Y^2)$ avec $(\mu_X, \mu_Y) \in \mathbb{R}^2$ et $(\sigma_X, \sigma_Y) \in {\mathbb{R}_+^\star}^2$.

On considère un échantillon $((X_1, Y_1), \ldots, (X_n, Y_n))$ d'un couple $(X, Y)$ issue d'une population hétérogène à deux classes tel que au sein de la première classe la distribution du couple est définie par la densité

$$f_{(X,Y)}(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right) \tag{4}$$

et pour les données de la deuxième classe la densité est

$$f_{(X,Y)}(x, y) = \frac{1}{2\pi} \exp\left(-\frac{(x-2)^2 + y^2}{2}\right). \tag{5}$$

On note par $(Z_1, \ldots, Z_n)$ les classes des données de l'échantillon observées avec $Z_i \in \{0, 1\}$ la classe du couple $(X_i, Y_i)$ $(i = 1, \ldots, n)$ et on cherche à prédire la classe $Z_i$ pour un nouveau couple $(X_i, Y_i)$ $(i = n+1, \ldots, m)$. On suppose que les classes sont à proportions identiques.

1. Donner la règle de décision de Bayes entre les deux classes
2. Calculer la forntière de décision
3. Discuter la nature de la frontière de décision en fonction de la valeur de $\rho$
4. Tracer la frontière de décision entre les deux classes pour une valeur de $\rho$ de votre choix

**Exercice 8** On considère le modèlle de régression linéaire simple défini par

$$Y_i = \beta_0 + \beta_1 x_i + E_i, \tag{6}$$

où $Y$ est la variable à prédire, $x$ est un prédicteur (déterministe), les $E_i$ sont des v.a centrées décorrélées représentant les résidus (bruit additif). Les deux paramètres $(\beta_0, \beta_1)$ sont inconnus et donc à estimer à partir d'un jeu de données observées $\mathcal{D} = ((x_1, y_1), \ldots, (x_n, y_n))$. L'estimation s'effectue par moindres-carrées.

1. Montrer que les estimations par moindres carrées de $\beta_0$ et $\beta_1$ sont données par :

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \tag{7}$$

$$\hat{\beta}_1 = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) / \sum_{i=1}^{n} (x_i - \bar{x})^2, \tag{8}$$

   où $\bar{x}$ représente la moyenne empirique des $x_i$ et $\bar{y}$ la moyenne moyenne empirique es $y_i$ et sont donnés par $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ et $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$.

2. Soit le jeu de données $\mathcal{D} = \{(2, 6), (6, 5), (8, 4.5)\}$. Prédire la valeur de $y$ pour $x = 10$ sur la base du modèle estimé par moindres carrées.
3. Claculer la valeur du coefficient d'ajustement $R^2$ et celle du coefficient de corrélation linéaire empirique $r$.