Université de Caen | Correction du partiel du 15 mars 2018. Durée 1h30 | **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)* 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 dispose d'un ensemble d'observations $\mathcal{D} = ((x_1, y_1), \ldots, (x_n, y_n))$ et on cherche à prédire les valeurs y pour des nouvelles observations $(x_{n+1}, \ldots, x_m)$. La prédiction s'effectue sur la base d'un modèle de paramètre θ du couple (X, Y) appris sur les données d'apprentissage $\mathcal{D}$ en minimisant le risque quadratique. On note par $\widehat{\Theta}$ un estimateur de θ. Les figures 1-(a, b, c) montrent, respectivement, un échantillon d'apprentissage, le vrai modèle, et un modèle de prédiction. Chacune des figures 1-(d, e, f) montrent le vrai-modèle et 20 réalisations du même modèle de prédiction estimé.

1. Donner la décomposition biais-variance du risque quadratique.
2. Discuter la qualité prédictive de chacun des trois modèles et en déduire le meilleur modèle prédictif.

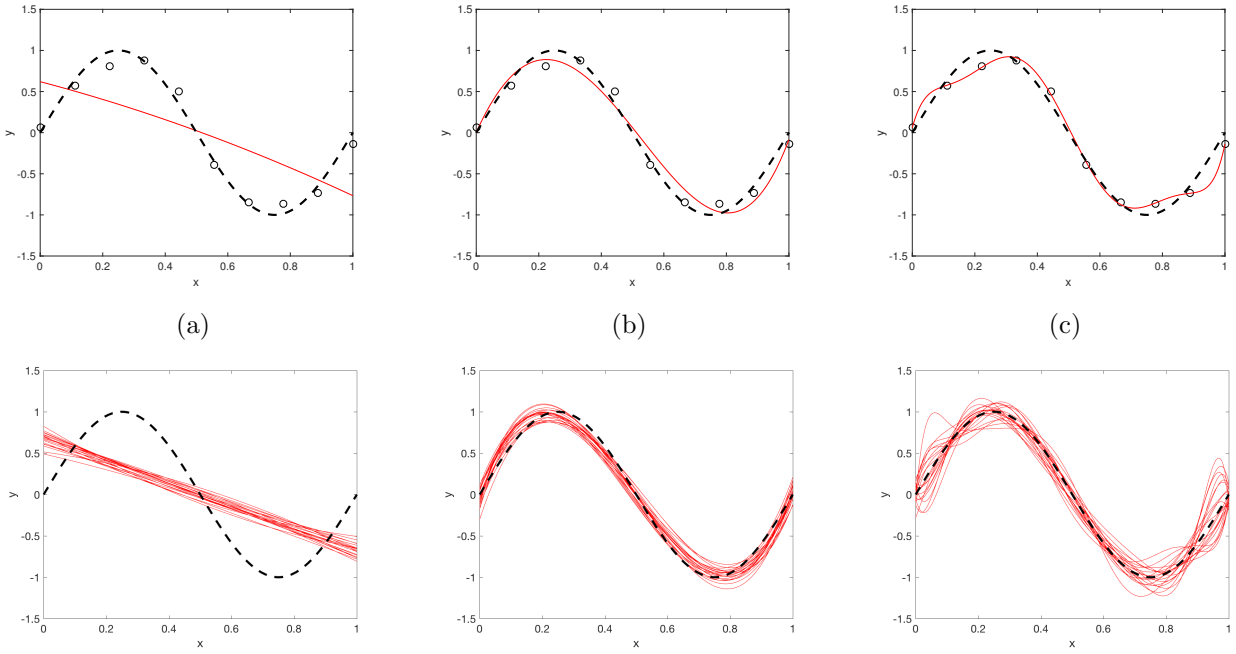

FIGURE 1 – Nuage de données (∘), vrai modèle (--), et modèle de prédiction (—)

Solution 1 Voir TD

Exercice 2 *(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 2 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.

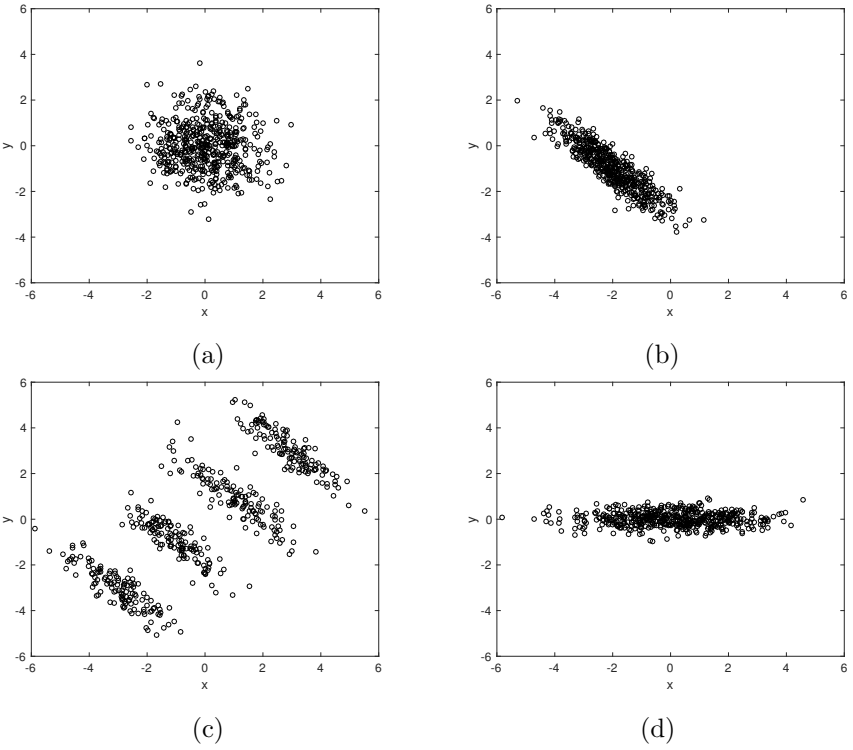

FIGURE 2 – Nuages de données ($\circ$)

Solution 2 Voir TD ((a) proche de 0 ; (b) proche de -0.8 (anti-corrélation) ; (c) proche de 0.7/8 (paradoxe de Simpson) ; (d) proche de 0

Exercice 3 *(6 pts)* On considère un échantillon aléatoire indépendant $((\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n))$ du couple $(\boldsymbol{X}, Y)$ où $\boldsymbol{X}$ est un vecteur de p prédicteurs réels ($\boldsymbol{X} \in \mathbb{R}^p$) et $Y \in [\![0, 1]\!]$ une variable à prédire. On dispose d'un échantillon observé d'apprentissage $((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n))$ et on cherche à prédire les classes de nouvelles observation $(\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_m)$ sur la base d'un modèle probabiliste appris sur les données d'apprentissage. On considère l'analyse discriminante où la densité de la classe $k \in [\![0, 1]\!]$ est définie par :

$$f(\boldsymbol{x}_i | Y_i = k; \boldsymbol{\theta}) = \frac{1}{(2\pi)^{\frac{p}{2}} |\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^\top {\boldsymbol{\Sigma}_k}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)\right). \quad (1)$$

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 :

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

$\boldsymbol{\theta}$ étant le vecteur paramètre du modèle. On note par $\pi_k = \mathbb{P}(Y_i = k)$, la probabilité a priori de la classe k. On suppose que $\pi_0 = 0.5$, $\boldsymbol{\mu}_0 = \boldsymbol{\mu}_1 = \mathbf{g}$, $\boldsymbol{\Sigma}_0 = \mathbf{I}$ et $\boldsymbol{\Sigma}_1 = \lambda \boldsymbol{\Sigma}_0$, où $\lambda > 1$ et $\mathbf{I}$ est la matrice identité.

1. Montrer que $Y = 1$ si et seulement si $||\boldsymbol{x} - \mathbf{g}||_2^2 \geq r$, en déterminant r en fonction de p et λ.
2. On considère le jeu de données bi-variées centrées de la Figure 3 où $\lambda = 3$. En déduire la règle de décision, et représenter approximativement la frontière de décision. On pourra approcher $\sqrt{3 \ln 3}$ par 1.8.

3. On suppose maintenant que les paramètres sont inconnus. Donner sans ou avec calcul les expressions des estimateurs du maximum de vraisemblance des paramètres du modèle.

Solution 3

1. On a $Y = 1$ si et seulement si :

$$\begin{aligned}
\log \frac{\mathbb{P}(y=1|\mathbf{x})}{\mathbb{P}(y=0|\mathbf{x})} &= \log \frac{\pi_1}{\pi_0} - \frac{1}{2}\log \frac{|\mathbf{\Sigma}_1|}{|\mathbf{\Sigma}_0|} - \frac{1}{2}\left\{(\mathbf{x}-\boldsymbol{\mu}_1)^T\mathbf{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) - (\mathbf{x}-\boldsymbol{\mu}_0)^T\mathbf{\Sigma}_0^{-1}(\mathbf{x}-\boldsymbol{\mu}_0)\right\} \geq 0 \\
&= -\frac{1}{2}\log(\lambda^p) - \frac{1}{2}\left\{(\mathbf{x}-\boldsymbol{g})^T\frac{1}{\lambda}(\mathbf{x}-\boldsymbol{g}) - (\mathbf{x}-\boldsymbol{g})^T(\mathbf{x}-\boldsymbol{g})\right\} \geq 0 \\
&= -\frac{p}{2}\log(\lambda) - \frac{1}{2}\left\{||\boldsymbol{x}-\mathbf{g}||_2^2\left(\frac{1}{\lambda}-1\right)\right\} \geq 0 \\
&= \frac{1}{2}||\boldsymbol{x}-\mathbf{g}||_2^2\left(\frac{\lambda-1}{\lambda}\right) \geq \frac{p}{2}\log(\lambda) \\
&= ||\boldsymbol{x}-\mathbf{g}||_2^2 \geq p\log(\lambda)\left(\frac{\lambda}{\lambda-1}\right)
\end{aligned} \tag{2}$$

   Au finale on a donc : $Y = 1$ si et seulement si $||\boldsymbol{x}-\mathbf{g}||_2^2 \geq r$ avec $r = p\log(\lambda)\left(\frac{\lambda}{\lambda-1}\right)$.

2. Pour $p = 2$ et $\lambda = 3$, on a $Y = 1$ si et seulement si $||\boldsymbol{x}-\mathbf{g}||_2^2 \geq 3\log(3)$

   La frontière de décision est donc définie par l'équation : $||\boldsymbol{x}-\mathbf{g}||_2^2 = r = 3\log(3)$ qui est celle d'un cercle de centre $\boldsymbol{g}$ et de rayon $\sqrt{r} = \sqrt{3\log(3)} \approx 1.8$. Donc sur la figure donnée il suffit de tracer approximativement un cercle de centre $(0,0)$ et de rayon 1.8.

3. Voir cours

Exercice 4 *(6 pts)* On suppose que le nombre Y de points marqués par une équipe de basketball peut être modélisé en fonction de si le match se joue à domicile ou non, par le modèle log-linéaire suivant

$$\ln(\lambda(X)) = \beta_0 + \beta_1 X \tag{3}$$

où X est la variable explicative qui vaut 1 si le match se joue à domicile, et 0 sinon, $\lambda(X) = \mathbb{E}_{\beta_0,\beta_1}[Y|X]$ et $Y|X \sim \mathcal{P}\big(\lambda(X)\big)$ de loi $\mathbb{P}(Y = y|X = x;\beta_0,\beta_1) = \frac{e^{-\lambda(x)}\lambda(x)^y}{y!}\ \forall y \in \mathbb{N}$.
On cherche à prédire le nombre de points $\widehat{y}$ marqué lors d'un match futur, avec l'espérance du modèle, i.e. la prédiction est donnée par $\widehat{y} = \mathbb{E}_{\widehat{\beta}_0,\widehat{\beta}_1}[Y|X]$ sur la base d'un échantillon indépendant $((x_1,y_1),\ldots,(x_n,y_n))$.

1. Montrer que la log-vraisemblance $\ln L(\beta_0,\beta_1) = \ln \mathbb{P}(Y_1 = y_1,\ldots,Y_n = y_n|X_1 = x_1,\ldots,X_n = x_n;\beta_0,\beta_1)$ est donnée par

$$\ln L(\beta_0,\beta_1) = -\sum_{i=1}^n e^{\beta_0+\beta_1 x_i} + \sum_{i=1}^n y_i\left(\beta_0+\beta_1 x_i\right) - \sum_{i=1}^n \ln(y_i!)$$

2. Montrer que les estimateurs du maximum de vraisemblance de β_1 et β_0 sont donnés par : $\widehat{\beta}_1 = \ln\left(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i}\right) - \widehat{\beta}_0$ et $\widehat{\beta}_0 = \ln\left(\frac{\sum_{i=1}^n (1-x_i) y_i}{\sum_{i=1}^n (1-x_i)}\right)$. On pourra utiliser notamment le fait que $e^{\beta_0+\beta_1 x_i} = (1-x_i)e^{\beta_0}$ si $x_i = 0$ et $e^{\beta_0+\beta_1 x_i} = x_i e^{\beta_0+\beta_1}$ si $x_i = 1$.

3. On considère le jeu de données de la Table 1. L'équipe joue son prochain match à l'extérieur. Prédire le nombre de points marqués lors de ce match.

x	y
1	64
0	60
1	59
0	62
1	72

TABLE 1 – Nombre de points marqués par l'équipe lors des cinq derniers matchs

4. Même question si le match se joue à domicile.

Solution 4

1. La log-vraisemblance $\ln L(\beta_0, \beta_1)$ est définie par :

$$\begin{aligned} L(\beta_0,\beta_1) &= \ln \mathbb{P}(Y_1 = y_1, \ldots, Y_n = y_n | X_1 = x_1, \ldots, X_n = x_n; \beta_0, \beta_1) \\ &= \ln \prod_{i=1}^{n} \mathbb{P}(Y_i = y_i | X_i = x_i; \beta_0, \beta_1) = \sum_{i=1}^{n} \ln \mathbb{P}(Y_i = y_i | X_i = x_i; \beta_0, \beta_1) \end{aligned}$$

D'après la loi de $Y|X$ (Poisson), on a donc :

$$\begin{aligned} L(\beta_0,\beta_1) &= \sum_{i=1}^{n} \ln \left[\frac{e^{-\lambda(x_i)} \lambda(x_i)^{y_i}}{y_i!} \right] = -\sum_{i=1}^{n} \lambda(x_i) + \sum_{i=1}^{n} y_i \ln \lambda(x_i) - \sum_{i=1}^{n} \ln(y_i!) \\ &= -\sum_{i=1}^{n} e^{\beta_0 + \beta_1 x_i} + \sum_{i=1}^{n} y_i (\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} \ln(y_i!) \end{aligned}$$

car on a $\lambda(x_i) = e^{\beta_0 + \beta_1 x_i}$.

2. Comme on a $e^{\beta_0+\beta_1 x_i} = (1 - x_i)e^{\beta_0}$ si $x_i = 0$ et $e^{\beta_0+\beta_1 x_i} = x_i e^{\beta_0+\beta_1}$ si $x_i = 1$, on peut alors écrire :

$$\begin{aligned} L(\beta_0,\beta_1) &= -\sum_{i=1}^{n} e^{\beta_0+\beta_1 x_i} + \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} \ln(y_i!) \\ &= -\sum_{i=1}^{n} (1 - x_i)e^{\beta_0} - \sum_{i=1}^{n} x_i e^{\beta_0+\beta_1} + \sum_{i=1}^{n} (1 - x_i) y_i \beta_0 + \sum_{i=1}^{n} x_i y_i (\beta_0 + \beta_1) - \sum_{i=1}^{n} \ln(y_i!) \\ &= -e^{\beta_0} \sum_{i=1}^{n} (1 - x_i) - e^{\beta_0+\beta_1} \sum_{i=1}^{n} x_i + \beta_0 \sum_{i=1}^{n} (1 - x_i) y_i + (\beta_0 + \beta_1) \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \ln(y_i!). \end{aligned}$$

Maximiser cette fonction par rapport à β_1 revient à trouver les zéros de l'équation

$$\frac{\partial L(\beta_0,\beta_1)}{\partial \beta_1} = -e^{\beta_0+\beta_1} \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} x_i y_i = 0.$$

On a donc

$$\begin{aligned} e^{\beta_0+\beta_1} \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} x_i y_i \\ e^{\beta_1} &= \frac{\sum_{i=1}^{n} x_i y_i}{e^{\beta_0} \sum_{i=1}^{n} x_i} \\ \beta_1 &= \ln \left(\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i} \right) - \beta_0 \end{aligned}$$

On a alors $\beta_0 + \beta_1 = \ln\left(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i}\right)$ et en intégrant cette expression et celle de β_1 dans celle de la log-vraisemblance $L(\beta_0, \beta_1)$, et en dérivant par rapport à β_0 on obtient :

$$\frac{\partial L(\beta_0, \beta_1)}{\partial \beta_0} = -e^{\beta_0} \sum_{i=1}^n (1 - x_i) + \sum_{i=1}^n (1 - x_i) y_i$$

qui s'annule quand

$$\begin{aligned} e^{\beta_0} \sum_{i=1}^n (1 - x_i) &= \sum_{i=1}^n (1 - x_i) y_i \\ \beta_0 &= \ln\left(\frac{\sum_{i=1}^n (1 - x_i) y_i}{\sum_{i=1}^n (1 - x_i)}\right) \end{aligned}$$

c'est CQFD.

3. D'après le tableau donné, on a :

$$\widehat{\beta}_0 = \ln\left(\frac{\sum_{i=1}^n (1 - x_i) y_i}{\sum_{i=1}^n (1 - x_i)}\right) = \ln((60 + 62)/2) = \ln(61)$$

et

$$\widehat{\beta}_1 = \ln\left(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i}\right) - \widehat{\beta}_0 = \ln((64 + 59 + 72)/3) - \ln(61) = \ln(65) - \ln(61).$$

Le match se joue à l'extérieur, ce qui correspond à $x = 0$. Le nombre de points marqués prédit est alors donné par $\widehat{y} = \mathbb{E}_{\widehat{\beta}_0, \widehat{\beta}_1}[Y|X] = \widehat{\lambda}(x) = e^{\beta_0 + \beta_1 x} = e^{\beta_0} = e^{\ln(61)} = 61$ points.

4. Le match se joue à domicile, ce qui correspond à $x = 1$. Le nombre de points marqués prédit est alors donné par $\widehat{y} = \mathbb{E}_{\widehat{\beta}_0, \widehat{\beta}_1}[Y|X] = \widehat{\lambda}(x) = e^{\beta_0 + \beta_1 x} = e^{\beta_0 + \beta_1} = e^{\ln(65)} = 65$ points.