

Consignes :

- Sont interdits : Documents, calequettes, 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 $((X_1, Y_1), \dots, (X_n, Y_n))$ 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 $\mathcal{D} = ((x_1, y_1), \dots, (x_n, y_n))$ et on cherche à prédire les valeurs y pour des nouvelles observations (x_{n+1}, \dots, 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 $\hat{\theta}$ 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.

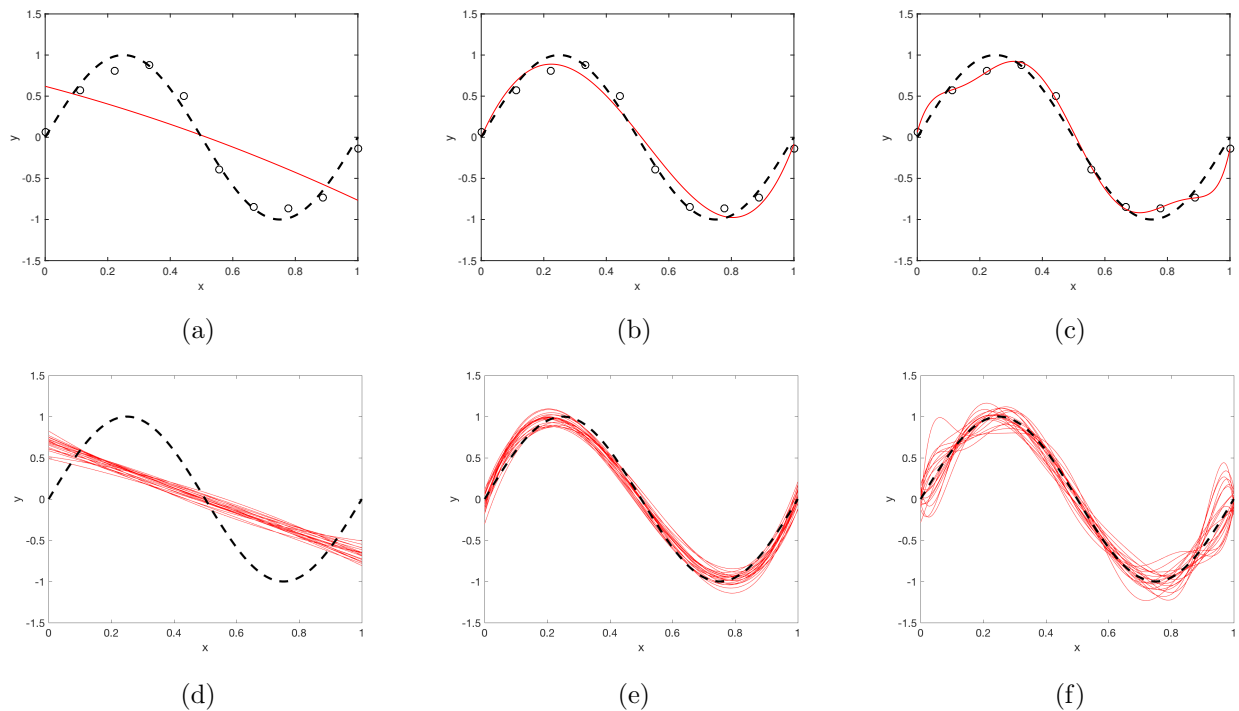


FIGURE 1 – Nuage de données (o), vrai modèle (---), et modèle de prédiction (—)

Exercice 2 (4 pts) Soit (X, Y) un couple de variables aléatoires réelles et soit $((x_1, y_1), \dots, (x_n, y_n))$ 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 coefficient de corrélation linéaire empirique r et justifier votre réponse.

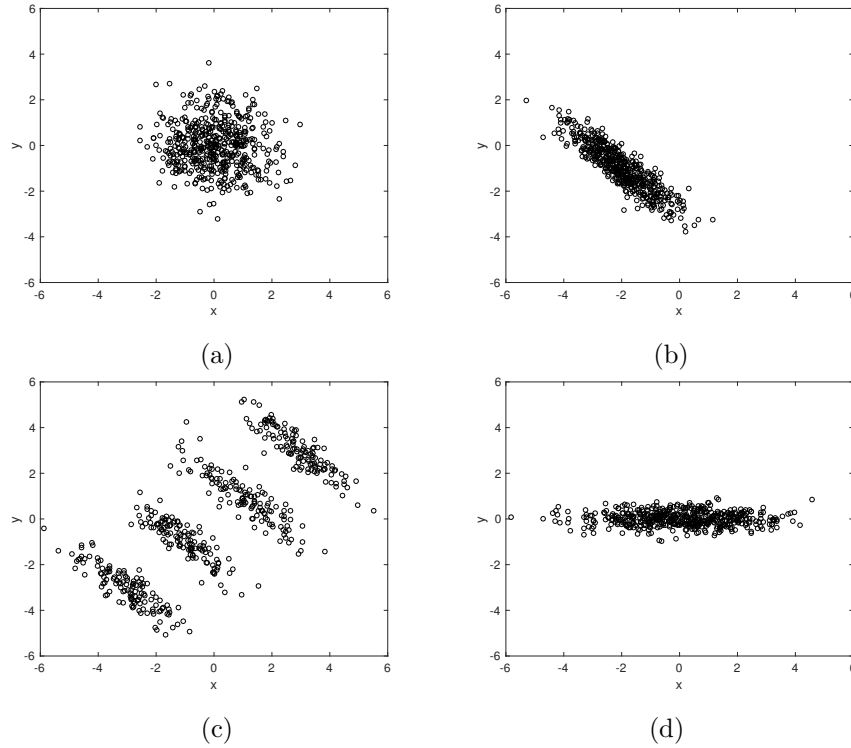


FIGURE 2 – Nuages de données (o)

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

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

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

$$\hat{y}_i = \arg \max_{k \in \llbracket 0, 1 \rrbracket} \mathbb{P}(Y_i = k | \mathbf{X}_i = \mathbf{x}_i; \boldsymbol{\theta}),$$

$\boldsymbol{\theta}$ 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 $\|\mathbf{x} - \mathbf{g}\|_2^2 \geq r$, 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.

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

$$\ln(\lambda(X)) = \beta_0 + \beta_1 X \quad (2)$$

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 $Y|X \sim \mathcal{P}(\lambda(X))$ 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 \hat{y} marqué lors d'un match futur, avec l'espérance du modèle, i.e. la prédiction est donnée par $\hat{y} = \mathbb{E}_{\hat{\beta}_0, \hat{\beta}_1}[Y|X]$ sur la base d'un échantillon indépendant $((x_1, y_1), \dots, (x_n, y_n))$.

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

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

2. Montrer que les estimateurs du maximum de vraisemblance de β_1 et β_0 sont donnés par :
 $\hat{\beta}_1 = \ln \left(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i} \right) - \hat{\beta}_0$ et $\hat{\beta}_0 = \ln \left(\frac{\sum_{i=1}^n (1-x_i) y_i}{\sum_{i=1}^n (1-x_i)} \right)$. 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.
4. Même question si le match se joue à domicile.

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

Feuille à remettre avec la copie d'examen (elle ne devra contenir aucun élément d'identité)

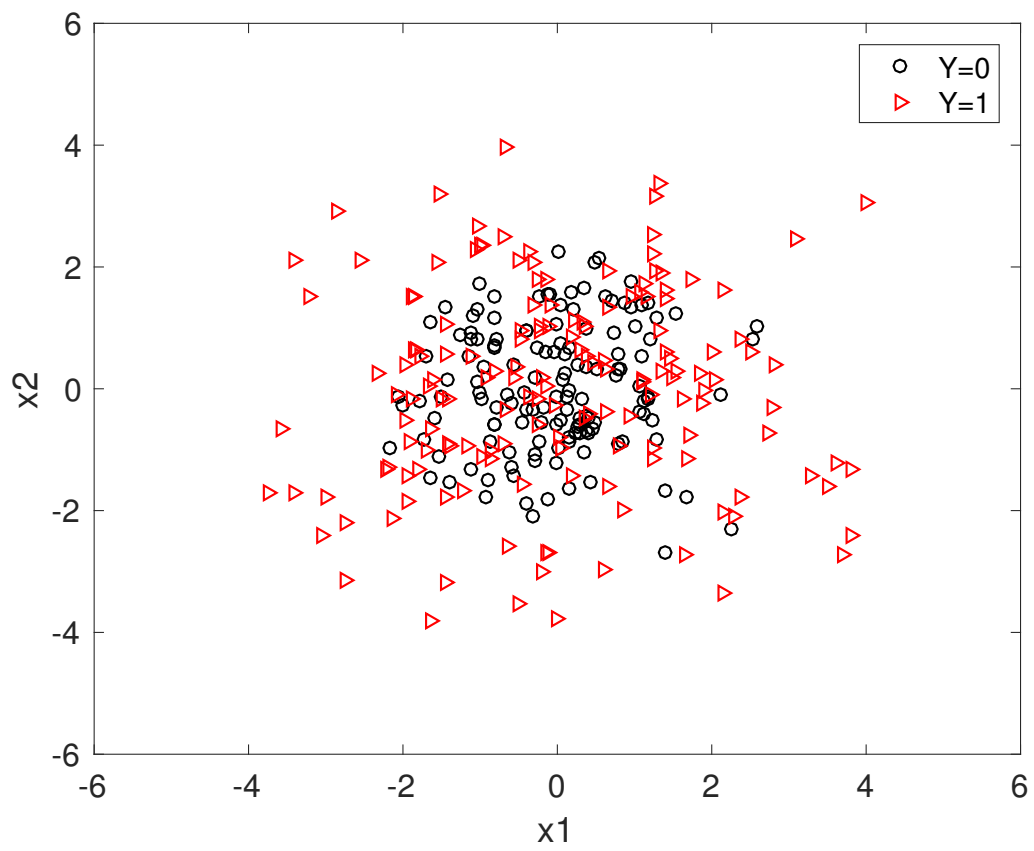


FIGURE 3 – Données