1 Exercices

Leb *et* $\mathrm{Leb}^{\otimes n}$ *désignent respectivement la mesure de Lebesgue sur* $\mathbb{R}$ *et celle sur* $\mathbb{R}^n$*. Par convention, les vecteurs sont des vecteurs colonne ; et* a^T *désigne la transposée de* a

Exercice 1. Soient $x_1, \dots, x_n$ des réels. On considère le modèle statistique

$$\left(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \left\{ p_\theta \cdot \mathrm{dLeb}^{\otimes n} := \bigotimes_{i=1}^n \mathrm{N}(\beta_1 + \beta_2 x_i, \sigma^2) \ : \ \theta := (\beta_1, \beta_2, \sigma^2) \in \mathbb{R}^2 \times \mathbb{R}_+^* \right\}\right) .$$

On note $\mathbf{1}$ et $\underline{\mathbf{x}}$ les vecteurs de $\mathbb{R}^n$ définis par $\mathbf{1} := (1, \dots, 1)^T$ et $\underline{\mathbf{x}} := (x_1, \dots, x_n)^T$. Dans la suite, on suppose qu'il existe au moins deux indices $i \neq j$ tels que $x_i \neq x_j$. Nous utiliserons les notations matricielles suivantes

$$\mathbf{Y} := \begin{bmatrix} Y_1 \\ \dots \\ Y_n \end{bmatrix} \in \mathbb{R}^n , \qquad \mathbf{X} := \begin{bmatrix} \mathbf{1} & \underline{\mathbf{x}} \end{bmatrix} \in \mathbb{R}^{n \times 2} , \qquad \beta := \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \in \mathbb{R}^2 .$$

1. Montrer que la matrice $\mathbf{X^T X}$ est inversible. *Puisque* $u^T \mathbf{X^T X} u \geq 0$ *pour tout* $u \in \mathbb{R}^2$*, on en déduit que la matrice est définie positive.*
2. Déterminer les estimateurs du maximum de vraisemblance $(\hat\beta_1, \hat\beta_2, \hat\sigma^2)$ de $(\beta_1, \beta_2, \sigma^2)$.
3. Déterminer la loi de $(\hat\beta_1, \hat\beta_2)$ sous $p_\theta \cdot \mathrm{dLeb}^{\otimes n}$.
4. Montrer que, sous $p_\theta \cdot \mathrm{dLeb}^{\otimes n}$, $\hat\beta_1$ et $\hat\beta_2$ sont indépendants si et seulement si $n^{-1} \sum_{i=1}^n x_i = 0$.
   Remarque : le modèle dépend de $x_1, \cdots, x_n$ *via les quantités* $\beta_1 + \beta_2 x_i$ *pour tout* $1 \leq i \leq n$ *soit encore, les lignes de* $\mathbf{X}\beta$*. Sans "pré-traitement", il n'y a pas de raisons que* $n^{-1} \sum_{i=1}^n x_i = 0$*. Néanmoins, en écrivant*

$$\begin{aligned} \beta_1 \mathbf{1} + \beta_2 \underline{\mathbf{x}} &= \left(\beta_1 + \beta_2 n^{-1} \sum_{j=1}^n x_j \right) \mathbf{1} + \beta_2 (\underline{\mathbf{x}} - n^{-1} \sum_{j=1}^n x_j \mathbf{1}) \\ &= \tilde\beta_1 \mathbf{1} + \beta_2 \underline{\tilde{\mathbf{x}}} = \begin{bmatrix} \mathbf{1} & \underline{\tilde{\mathbf{x}}} \end{bmatrix} \begin{bmatrix} \tilde\beta_1 \\ \beta_2 \end{bmatrix} \qquad \underline{\tilde{\mathbf{x}}} = \underline{\mathbf{x}} - n^{-1} \sum_{j=1}^n x_j \mathbf{1} = \underline{\mathbf{x}} - \frac{\mathbf{1}^T \underline{\mathbf{x}}}{\mathbf{1}^T \mathbf{1}} \mathbf{1} \end{aligned}$$

   on peut paramétrer le modèle de façon équivalente par $(\tilde\beta_1, \beta_2, \sigma^2)$*, tout en introduisant une matrice d'expérience* $\begin{bmatrix} \mathbf{1} & \underline{\tilde{\mathbf{x}}} \end{bmatrix}$ *dont les colonnes sont* **orthogonales** *(c'est le sens de* $n^{-1} \sum_{i=1}^n \tilde x_i = 0$*).*
5. Déterminer la loi de $\hat\sigma^2$ sous $p_\theta \cdot \mathrm{dLeb}^{\otimes n}$.
6. Montrer que sous $p_\theta \, \mathrm{dLeb}^{\otimes n}$, $\hat\sigma^2$ et $(\hat\beta_1, \hat\beta_2)$ sont indépendants.
7. Soit $\alpha \in]0, 1[$. Proposer un intervalle de confiance de niveau $(1 - \alpha)$ pour σ^2.

Exercice 2 (Estimation d'un coefficient de mélange). Soit un n-échantillon $(X_1, \dots, X_n)$ du modèle $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{ f_\theta \cdot \mathrm{Leb}, \theta \in]0, 1[\})$ où

$$f_\theta(x) = \frac{\theta}{a} \ \mathbb{1}_{\{[0,a]\}}(x) + \frac{(1-\theta)}{b} \ \mathbb{1}_{\{[0,b]\}}(x).$$

On suppose que a et b sont connus et que $0 < a < b$. La fonction de répartition associée à la densité f_θ vaut

$$x \mapsto \begin{cases} 0 & x \le 0 \\ \left(\frac{\theta}{a} + \frac{(1-\theta)}{b}\right) x & 0 \le x \le a \\ \theta + \frac{(1-\theta)}{b} x & a \le x \le b \\ 1 & x \ge 1 \end{cases}$$

1. Exprimer la fonction de vraisemblance à l'aide de $N_a := \sum_{i=1}^n \mathbb{1}_{X_i \in [0,a]}$ le nombre d'observations à valeur dans $[0, a]$.
2. En déduire l'estimateur du maximum de vraisemblance de θ, noté $\hat{\theta}_n^{(1)}$.
3. En utilisant la méthode des moments, proposer un estimateur $\hat{\theta}_n^{(2)}$ de θ.

► **Analyses numériques.** Pour les analyses numériques ci-dessous, on se donne des mesures $x_1, \cdots, x_n$ obtenues comme la réalisation de tirages indépendants sous la loi $f_{0.2}$ dans le cas $a = 1$ et $b = 3$.

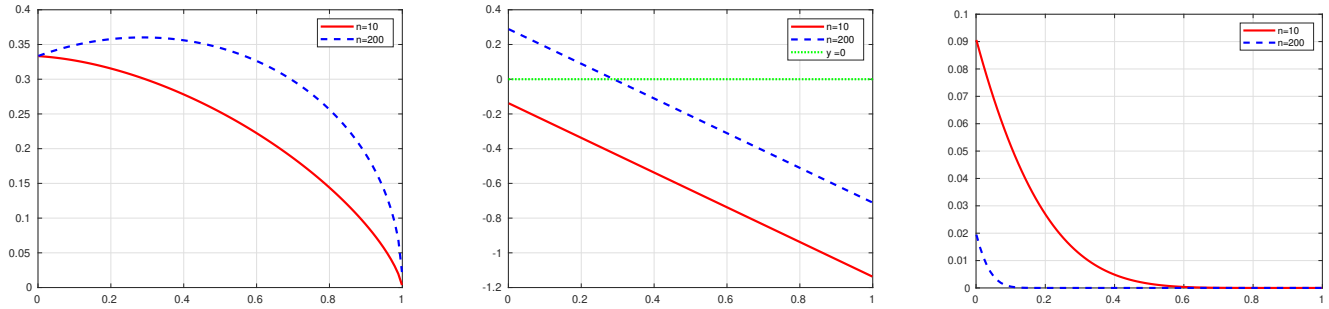

FIGURE 1 – [gauche] Tracé de deux réalisations de la fonction de vraisemblance $\theta \mapsto L(\theta, X_1, \cdots, X_n)$: une pour $n = 10$ puis une pour $n = 200$. [centre] Tracé de deux réalisations de la fonction (aléatoire) dont l'estimateur des moments est un zero : une dans le cas $n = 10$ et l'autre dans le cas $n = 200$. [droite] Evolution du biais de l'estimateur $\hat{\theta}_n^{(1)}$ en fonction de θ. Dans le cas $n = 10$ et $n = 200$.

Sur la figure 7, on visualise la loi de différents estimateurs : $\hat{\theta}_n^{(1)}$, $\hat{\theta}_n^{(2)}$ et

$$\hat{\theta}_n := \frac{b(N_a/n) - a}{b - a}, \qquad \tilde{\theta}_n := \frac{b - 2n^{-1}\sum_{i=1}^n X_i}{b - a} .$$

Exercice 3 (Famille exponentielle de densités). Beaucoup de lois (dont des usuelles) ont une structure particulière pour lesquelles elles sont dites dans *la famille exponentielle*. Soit $(\mathsf{X}, \mathcal{X})$ un espace mesurable (nous prendrons $\mathsf{X} = \mathbb{R}^k$ ou $\mathsf{X} = \mathbb{N}^k$) et μ une mesure σ-finie sur $(\mathsf{X}, \mathcal{X})$. Soit $T : \mathsf{X} \to \mathbb{R}$ et $h : \mathsf{X} \to \mathbb{R}^+$ deux fonctions mesurables.

On appelle `modèle exponentiel canonique associé au couple` (T, h), une famille de lois ayant une densité par rapport à μ de la forme

$$x \mapsto q_\eta(x) = h(x) \exp\left(\eta T(x) - A(\eta)\right) , \quad x \in \mathsf{X} , \tag{1}$$

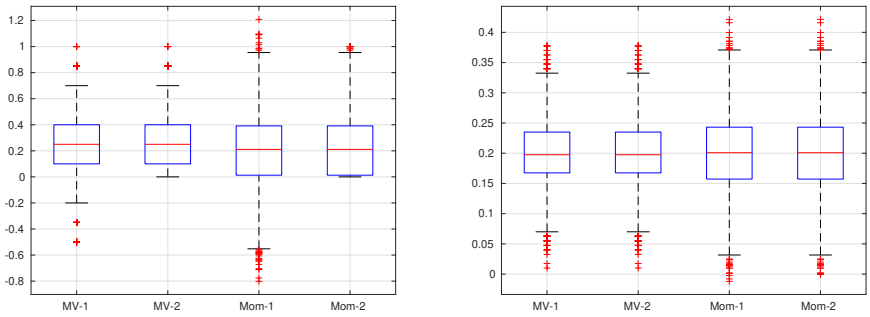

FIGURE 2 – Boxplot de la réalisation de $N = 5000$ estimateurs dans le cas $n = 10$ (gauche) et $n = 200$ (droite). Les boxplots MV-1 et MV-2 correspondent resp. à $\hat{\theta}_n$ et $\hat{\theta}_n^{(1)}$; ceux de Mom-1 et Mom-2 correspondent à $\tilde{\theta}_n$ et $\hat{\theta}_n^{(2)}$. On fera attenton aux échelles en ordonnées pour comparer les schémas de gauche et de droite.

où $A(\eta)$ est défini par :

$$A(\eta) := \log \int h(x) \exp\left(\eta T(x)\right) \mu(\mathrm{d}x) \; . \tag{2}$$

L'*espace des paramètres naturels* de la famille canonique associée à (T, h) est l'ensemble

$$\Xi := \{\eta \in \mathbb{R} \; : \; |A(\eta)| < \infty\} \; .$$

Plus généralement on appelle `famille exponentielle` toute famille de loi de densité

$$x \mapsto p_\theta(x) = h(x) \exp(\varphi(\theta) T(x) - B(\theta))$$

par rapport à une mesure μ σ-finie sur $\mathbb{R}$ ou $\mathbb{N}$. Les valeurs admissibles de θ sont les éléments de

$$\Theta := \{\theta \in \mathbb{R} : \int h(x) \;\; \exp(\varphi(\theta) T(x)) \, \mu(\mathrm{d}x)\}.$$

1. Les lois suivantes sont elles dans la famille exponentielle, et si oui, sont-elles canoniques ?
   (a) La loi exponentielle de densité

   $$x \mapsto p_\eta(x) = \eta \exp(-\eta x) \mathbb{1}_{\mathbb{R}^+}(x)$$

   sur $\mathsf{X} = \mathbb{R}$? Si oui, précisez l'espace des paramètres naturels.
   (b) La loi gaussienne $\mathrm{N}(\eta, 1)$ sur $\mathbb{R}$. Si oui, précisez l'espace des paramètres naturels.
   (c) Soit $\alpha > 0$ fixé. La loi de Weibull sur $]0, \infty[$ de paramètre d'échelle $\lambda > 0$ et de densité

   $$x \mapsto q_\lambda(x) = \frac{\alpha}{\lambda} \left(\frac{x}{\lambda}\right)^{\alpha - 1} \;\; \exp(-(x/\lambda)^\alpha) \qquad x > 0.$$

   Si oui, précisez l'espace des paramètres naturels.
   (d) Montrer que la loi de Poisson définit un modèle exponentiel avec

   $$\varphi(\theta) = \log(\theta), \quad B(\theta) = \theta, \quad T(x) = x, \quad h(x) = 1/x!.$$

   Préciser le modèle exponentiel canonique associé.

(e) Montrer que la loi binomiale définit un modèle exponentiel avec

$$\varphi(\theta) = \log\left(\frac{\theta}{1-\theta}\right), \quad B(\theta) = -n\log(1-\theta), \quad T(x) = x, \quad h(x) = \binom{n}{x}.$$

Préciser le modèle exponentiel canonique associé.

2. Nous allons tout d'abord établir certaines propriétés des modèles exponentiels canoniques.
   (a) Supposons que l'espace des paramètres est ouvert, que A est régulière et que l'on peut permuter dérivée et intégrale. Quelle relation a-t-on entre $A'(\eta)$, $A''(\eta)$ et l'espérance et la variance de $T(X)$ sous q_η ?
   (b) Montrer que la fonction $\eta \mapsto A(\eta)$ est convexe et que l'espace des paramètres naturels est un sous-ensemble convexe de $\mathbb{R}$.
   (c) On suppose dans la suite que Ξ est un intervalle ouvert. Pour $\eta \in \Xi$, on pose

$$G(\eta) := \int h(x) \exp\left(\eta T(x)\right) \mu(\mathrm{d}x) \, .$$

Montrer que G est infiniment différentiable sur Ξ et que, pour tout $k \in \mathbb{N}^*$ et $\eta \in \Xi$,

$$G^{(k)}(\eta) = \int h(x)\, T^k(x)\, \exp\left(\eta T(x)\right) \mu(\mathrm{d}x) \, .$$

   (d) Soit $(X_1, \ldots, X_n)$ un n-échantillon du modèle exponentiel canonique associé à (T, h).
      i. Déterminer l'estimateur des moments associé à la fonction T.
      ii. Déterminer l'estimateur du maximum de vraisemblance. Que remarque-t-on ?
3. Considérons dorénavant le modèle exponentiel général. On supposera que Θ est un intervalle ouvert de $\mathbb{R}$. On suppose que la fonction φ définit un difféomorphisme de Θ sur Ξ l'espace des paramètres naturels associé. On dispose d'un n-échantillon $(X_1, \ldots, X_n)$ du modèle

$$(\mathsf{X}, \mathcal{X}, \{p_\theta \cdot \mu \, : \, \theta \in \Theta\}) \, .$$

Déterminer l'estimateur du paramètre θ, solution des équations de vraisemblance [on pensera à utiliser la paramétrisation canonique].

En conclusion de l' exercice précédent : (i) les modèles exponentiels s'écrivent sous forme de facteurs séparables en la variable x et le paramètre de la famille $\phi(\theta)$ (ou η dans le modèle canonique). (ii) Il n'y a pas unicité de la représentation, y compris du paramètre canonique : on a en effet $\eta T(x) = (\lambda\eta)\,(T(x)/\lambda)$ pour tout $\lambda \in \mathbb{R}$ par exemple. (iii) On a ici présenté une version simplifiée de la définition de ces modèles, mais en toute généralité le paramètre du modèle peut être vectoriel, et le produit $\phi(\theta)T(x)$ peut être un produit scalaire entre des fonctions ϕ, T à valeur dans $\mathbb{R}^q$ i.e. $\sum_{i=1}^q \phi_i(\theta)T_i(x)$. Par exemple, la loi gaussienne $\mathcal{N}_d(\mu, \Sigma)$ pour une matrice de covariance Σ inversible, a une densité proportionnelle à

$$\begin{aligned}
\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) &\propto \exp\left(-\frac{1}{2}(x^T\Sigma^{-1}x - 2\mu^T\Sigma^{-1}x)\right) \\
&= \exp\left(-\frac{1}{2}\mathrm{Tr}\left(\Sigma^{-1}xx^T\right) + \mu^T\Sigma^{-1}x\right) \\
&= \exp\left(-\frac{1}{2}\left\langle\Sigma^{-1}, xx^T\right\rangle + \left\langle\Sigma^{-1}\mu, x\right\rangle\right)
\end{aligned}$$

et est donc un exemple de `famille exponentielle canonique de dimension` $q > 1$. *(iv) Une grande partie des modèles utilisés en pratique sont des modèles exponentiels (modèle gaussien, log-normal, exponentiel, gamma, Bernoulli, Poisson, etc). Ils sont traités dans le Chapitre IV-6 du polycopié.*

Exercice 4 (Durée de vie). On considère un système ne fonctionnant que si deux machines sont toutes les deux en état de marche. On observe les durées de vie des deux machines, que l'on modélise comme des lois exponentielles de paramètres λ_0 et λ_1, $\lambda_i > 0$; et des lois indépendantes.

1. Montrer qu'une variable aléatoire X suit la loi exponentielle $\mathcal{E}(\lambda)$ si et seulement si
$$\forall x > 0 : \ \mathbb{P}(X > x) = \exp(-\lambda x).$$

Le modèle statistique associé à cette expérience est
$$\left(\mathbb{R}_+^2, \mathcal{B}(\mathbb{R}_+^2), \{\mathbb{P}_\theta = \mathcal{E}(\lambda_0) \otimes \mathcal{E}(\lambda_1), \theta = (\lambda_0, \lambda_1) \in (\mathbb{R}_+^\star)^2\}.\right).$$

2. Calculer la probabilité pour que le système ne tombe pas en panne avant la date t. En déduire la loi de la durée de vie Z du système. Le système tombe en panne ; calculer la probabilité pour que la panne soit due à une défaillance de la machine 1.
3. Soit $I = 1$ si la panne du système est due à une défaillance de la machine 1, $I = 0$ sinon. Calculer $\mathbb{P}_\theta(Z > t; I = \delta)$ pour tout $t \geq 0$ et $\delta \in \{0, 1\}$. En déduire que Z et I sont indépendantes.
4. On dispose de n systèmes identiques et fonctionnant indépendamment les uns des autres dont on observe les durées de vie $Z_1, \ldots, Z_n$.
   (a) Écrire le modèle statistique correspondant. Les paramètres λ_0 et λ_1 sont-ils identifiables ?
   (b) Supposons maintenant que l'on observe à la fois les durées de vie des systèmes $Z_1, \ldots, Z_n$ et les causes de la défaillance correspondantes $I_1, \ldots, I_n$, $I_i \in \{0, 1\}$. Écrire le modèle statistique dans ce cas. Les paramètres λ_0 et λ_1 sont-ils identifiables ?

Exercice 5 (Modèle auto-régressif). Soit $\theta := (\phi, \sigma^2) \in \Theta := \mathbb{R} \times \mathbb{R}_+^\star$. On considère l'observation $Z = (X_1, \ldots, X_n)$, qui sous le modèle statistique $p_\theta \cdot \mathrm{dLeb}^{\otimes n}$, a la loi suivante : pour tout $1 \leq k \leq n$, la loi conditionnelle de X_k sachant $(X_1, \cdots, X_{k-1})$ est une loi $\mathrm{N}(\phi\, X_{k-1}, \sigma^2)$. Par convention, $X_0 = 0$.

1. Écrire le modèle statistique engendré par l'observation Z.
2. Déterminer l'estimateur du maximum de vraisemblance de $\theta = (\phi, \sigma^2)$.