TD 1 - Modèle statistique, construction d'estimateurs, propriétés des estimateurs ponctuels.

Exercice 1.1. Construction d'un estimateur par la méthode des moments. Soient $\{X_i\}_{1\le i\le N}$ des variables aléatoires indépendantes de loi uniforme $\mathcal{U}([0,\theta]), \theta > 0$, et soit $(x_1, \cdots, x_N)$ des observations.

1. Proposer une estimation de θ construit par la méthode de substitution en utilisant le moment d'ordre 1.

2. Proposer une estimation de θ construit par la méthode de substitution en utilisant le moment d'ordre 2.

Solution:

1. Soit la fonctionnelle

$$G(P_\theta) \ := \int x dP_\theta = \mathrm{E}_\theta(X_1) = \theta/2,$$

on utilise la mesure empirique $\hat{\mathbb{P}}(x_1, \cdots, x_N) = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$

$$G(\hat{\mathbb{P}}(x_1, \cdots, x_N)) = \int x\hat{\mathbb{P}}(x_1, \cdots, x_N)(dx) = \frac{1}{N}\sum_{i=1}^N x_i = \hat{\theta}^{(1)}/2$$

soit l'estimateur $\hat{\theta}^{(1)} = 2\overline{X}_N$

2. De la fonctionnelle

$$H(P_\theta) \ := \int x^2 dP_\theta = \mathrm{E}_\theta(X_1^2) = \theta^2/3,$$

on déduit

$$H(\hat{\mathbb{P}}(x_1, \cdots, x_N) = \frac{1}{n}\sum_{i=1}^N X_i^2 = (\hat{\theta}^{(2)})^2/3\ .$$

soit $\hat{\theta}^{(2)} = \left(\dfrac{3}{N}\displaystyle\sum_{i=1}^N x_i^2\right)^{1/2}$ (on sait que $\theta > 0$.)

Exercice 1.2. Construction de l'EMV. Supposons que l'on observe n variables aléatoires $X_1, \cdots, X_n$ indépendantes et de même loi. Calculer l'estimateur du maximum de vraisemblance du paramètre θ lorsque la loi des variables X_i est :

1. Une loi de Poisson $\mathcal{P}(\theta)$ de paramètre $\theta \geq 0$.
2. Une loi exponentielle $\mathcal{E}(\theta)$ de paramètre $\theta > 0$.
3. Une loi admettant la densité $\exp\{-(x-\theta)\}\mathbb{I}(x \geq \theta)$, $\theta \in \mathbb{R}$.

On vérifiera dans chaque cas que l'on obtient bien le maximum global de la fonction de vraisemblance.

Solution:

1. On a

$$P_\theta(X_i = x_i) = \frac{\theta^{x_i}}{x_i!}e^{-\theta},$$

d'où

$$L((x_1,\ \ldots,\ x_n),\ \theta) = \prod_{i=1}^{n} \frac{\theta^{x_i}}{x_i!}e^{-\theta}$$

et

$$l_n(\theta) = \ln(L((x_1,\ \ldots,\ x_n),\ \theta)) = -n\theta + \sum_{i=1}^{n}[x_i \ln(\theta) - \ln(x_i!)].$$

En dérivant par rapport à θ, on obtient $l'_n(\theta) = -n + \frac{1}{\theta}\sum_{i=1}^{n} x_i$. Par conséquent, si $\sum_{i=1}^{n} x_i = 0$, on obtient un maximum en $\theta = 0$. Sinon, la solution de l'équation $l'_n(\theta) = 0$ est $\hat{\theta} = \overline{x}$. Cette valeur est bien le maximum global de l_n, car l_n croît sur $]0, \overline{x}[$ ($l'_n(x) > 0$) et décroît sur $]\overline{x}, \infty[$. Dans tous les cas, $\hat{\theta} = \overline{x}$. Soit l'estimateur:

$$\hat{\theta}^{MV}(X_1, \cdots, X_n) = \overline{X}.$$

2. La famille des lois exponentielles est donnée par les densités:

$$p_\theta(x) = \theta e^{-\theta x}\mathbb{I}_{\{x \geq 0\}},\ \text{pour } \theta > 0\ .$$

Pour l'échantillon d'observations $(x_1, \cdots, x_N)$, nous en déduisons donc la vraisemblance:

$$\begin{aligned}
\mathcal{L}((x_1, \cdots, x_n), \theta) &= \prod_{i=1}^{n} p_\theta(x_i) \\
&= \prod_{i=1}^{n} \theta e^{-\theta x_i}\mathbb{I}_{\{x_i > 0\}} \\
&= \theta^n e^{-\theta(x_1 + \cdots + x_n)}\mathbb{I}(\min_{i=1,\cdots,n} x_i \geq 0)
\end{aligned}$$

Soit la log-vraisemblance:

$$l(\theta) = \begin{cases} n\ln\theta - n\theta\overline{x}, & \text{si } \min_{i=1,\ldots,n} x_i \geq 0, \\ -\infty, & \text{si } \min_{i=1,\ldots,n} x_i < 0. \end{cases}$$

Dans le second cas, toute valeur de $\mathbb{R}_+$ peut être considérée comme estimation du max de vraisemblance (mais si une valeur x_i est négative, l'hypothèse que la loi est exponentielle n'est pas valide...).

Si $\min_{i=1,\ldots,n} x_i \geq 0$, on a $l'(\theta) = n\theta^{-1} - n\overline{x}$. Si $\overline{x} = 0$, $l(\theta)$ est strictement croissante et n'admet pas de maximum. Si $\overline{x} > 0$, $l'(\hat{\theta}) = 0$ équivaut à $\hat{\theta} = 1/\overline{x}$. Comme la fonction $l'(\theta)$ est > 0 sur $]0, 1/\overline{x}[$ et < 0 sur $]1/\overline{x}, \infty[$, l'estimateur du maximum de vraisemblance du paramètre θ est bien $\hat{\theta}^{MV}(X_1, \cdots, X_n) = 1/\overline{X}$.

3. On a

$$L((x_1, \ \ldots, \ x_n), \ \theta) = \prod_{i=1}^{n} e^{-(x_i-\theta)} \mathbb{I}_{\{x_i \geq \theta\}} = e^{n\theta-(x_1+\cdots+x_n)} \mathbb{I}(\min_{i=1,\cdots,n} x_i \geq \theta) \ .$$

On note $x_{(1)} = \min_{i=1,\ldots,n} x_i$. On a alors

$$l_n(\theta) = \begin{cases} n(\theta - \overline{x}), & \text{si } x_{(1)} \geq \theta, \\ -\infty, & \text{si } x_{(1)} < \theta. \end{cases}$$

Cette fonction atteint son maximum au point $\hat{\theta} = x_{(1)}$, soit l'estimateur du maximum de vraisemblance: $\hat{\theta}^{MV}(X_1, \cdots, X_n) = X_{(1)}$.

Exercice 1.3. On considère le modèle uniforme

$$\mathcal{M}_\theta = \{\mathcal{U}([0, \theta]), \theta > 0\},$$

Soit $X_1, \cdots, X_n$ un échantillon i.i.d. de loi inconnue dans $\mathcal{M}_\theta$.

1. Montrer que l'estimateur du maximum de vraisemblance est $\hat{\theta}_n = \mathrm{X}_{(n)}$ où

$$X_{(n)} = \max_{1 \leq i \leq n} X_i.$$

2. Calculer $\mathbb{P}\{X_{(n)} \leq x\}$ pour tout x réel. En déduire que l'estimateur $X_{(n)}$ de θ est biaisé, asymptotiquement non-biaisé.

3. Donner $\lim_{n\to+\infty} \mathbb{P}\left(n(\theta - X_{(n)}) \leq x\right)$ et en déduire un résultat de convergence en loi de l'estimateur. Quelle est la vitesse de convergence de l'estimateur ?

Solution:

1. Soit un échantillon d'observations $x = (x_1, \cdots, x_N)$. Nous avons la vraisemblance du paramètre θ en x est donnée par:

$$L(\theta; x) = \prod_{i=1}^{n} \frac{1}{\theta} \mathbb{I}_{[0,\theta]}(x) = \frac{1}{\theta^n} \prod_{i=1}^{n} \mathbb{I}_{[0,\theta]}\left(x_{(n)}\right) \ .$$

avec $x_{(n)} = \max_{1 \leq i \leq n} x_i$.

Si $\theta < x_{(n)}$, $L(\theta;x) = 0$. Si $\theta \geq x_{(n)}$, $L(\theta;x) > 0$, et est décroissante. Le maximum est donc obtenu pour $\hat{\theta} = x_{(n)}$.
(Pb si $x_{(n)} = 0$ mais c'est un événément de probabilité nulle si la loi sous-jacente appartient à $\mathcal{M}_\theta$).
Nous avons donc l'estimateur: $\hat{\theta}_n^{MV} = X_{(n)}$
2. Grâce à l'indépendance des variables aléatoires $X_1, X_2, \ldots, X_n$, on a

$$\mathbb{P}(X_{(n)} \leq x) = \mathbb{P}(\max_{i=1,\ldots,n} X_i \leq x) = \prod_{i=1}^{n} \mathbb{P}(X_i \leq x) = (F(x))^n,$$

où $F(x)$ est la fonction de répartition de la loi $U[0,\ \theta]$, c'est-à-dire

$$F(x) = \frac{x}{\theta}\mathrm{I}_{[0,\theta]}(x) + \mathrm{I}_{]\theta,\infty[}(x)\ .$$

On en déduit que

$$\mathbb{P}(X_{(n)} \leq x) = \left(\frac{x}{\theta}\right)^n \mathrm{I}_{[0,\theta]}(x) + \mathrm{I}_{]\theta,\infty[}(x)\ .$$

On peut reécrire la fonction de répartition sous la forme:

$$\mathbb{P}(X_{(n)} \leq x) = \int_{-\infty}^{x} n\frac{t^{n-1}}{\theta^n}\mathrm{I}_{[0,\theta]}(t)\ dt$$

avec $t \mapsto n\frac{t^{n-1}}{\theta^n}\mathrm{I}_{[0,\theta]}(t)$ la densité (positive et d'intégrale 1 sur $\mathbb{R}$). On peut donc calculer $\mathbb{E}_\theta\left(X_{(n)}\right)$:

$$\mathbb{E}_\theta\left(X_{(n)}\right) = \int_{\mathbb{R}} t\frac{nt^{n-1}}{\theta^n}\mathrm{I}_{[0,\theta]}(t)\ dt = \frac{n}{n+1}\theta\ .$$

L'estimateur est donc biaisé, $\mathbb{E}_\theta\left(X_{(n)}\right) - \theta = \frac{-\theta}{n+1}$, mais asymptotiquement non biaisé.
3. Par ailleurs, on a

$$\mathbb{P}\left(n\left(\theta - X_{(n)}\right) \leq x\right) = \mathbb{P}\left(\theta - X_{(n)} \leq \frac{x}{n}\right) = \mathbb{P}\left(X_{(n)} \geq \theta - \frac{x}{n}\right) = 1 - \mathbb{P}\left(X_{(n)} \leq \theta - \frac{x}{n}\right)$$

où on a utilisé $\mathbb{P}\left(X_{(n)} = \theta - \frac{x}{n}\right) = 0$. Donc

$$\begin{aligned}
\mathbb{P}\left(n\left(\theta - X_{(n)}\right) \leq x\right) &= 1 - [\left(1 - \frac{x}{n\theta}\right)^n \mathrm{I}_{[0,\theta]}\left(\theta - \frac{x}{n}\right) + \mathrm{I}_{]\theta,+\infty[}\left(\theta - \frac{x}{n}\right)] \\
&= 1 - \left[\left(1 - \frac{x}{n\theta}\right)^n \mathrm{I}_{[0,n\theta]}(x) + \mathrm{I}_{]-\infty,0[}(x)\right] \\
&= \left[1 - \left(1 - \frac{x}{n\theta}\right)^n\right] \mathrm{I}_{[0,n\theta]}(x) + \mathrm{I}_{]n\theta,\infty[}(x)\ .
\end{aligned}$$

Comme

$$\left(1 - \frac{x}{n\theta}\right)^n = \exp\left(n\log\left(1 - \frac{x}{n\theta}\right)\right) \underset{n\to+\infty}{\longrightarrow} e^{-x/\theta},$$

on obtient

$$\mathbb{P}\left(n\left(\theta - X_{(n)}\right) \leq x\right) \underset{n\to+\infty}{\longrightarrow} \left(1 - e^{-x/\theta}\right)\mathrm{I}_{[0,\infty[}(x)\ .$$

Comme cette convergence a lieu pour tout $x \in \mathbb{R}$ (la fonction limite est continue sur tout $\mathbb{R}$), nous avons démontré que la suite des variables aléatoires $n(\theta - X_{(n)})$ converge en loi vers une loi exponentielle de paramètre θ.

Exercice 1.4. Soient $X_1, \cdots, X_n$ des variables aléatoires i.i.d. de densité

$$p_\theta(x) = (1+\theta)\mathrm{I}_{\{0\le x\le 1/2\}} + (1-\theta)\mathrm{I}_{\{1/2<x\le 1\}},$$

où $\theta \in [-1,\ 1]$ est un paramètre inconnu que l'on souhaite estimer.

1. Calculer l'estimateur du maximum de vraisemblance $\hat{\theta}_n^{MV}$ de θ.

2. Est-il consistant ? Sans biais ? Déterminer la loi limite de $\sqrt{n}(\hat{\theta}_n^{MV} - \theta)$ quand $n \to \infty$ et en déduire la vitesse de convergence de l'estimateur.

Solution: 1. Soit un échantillon d'observations $(x_1, \cdots, x_n)$, on a:

$$L(x_1,\ \ldots,\ x_n,\ \theta) = \prod_{i=1}^{n}\big((1+\theta)\mathbb{I}_{[0;1/2]}(x_i) + (1-\theta)\mathbb{I}_{]1/2;1]}(x_i)]\big).$$

Soit n_1 le nombre de x_i appartenant à l'intervalle $[0, 1/2]$ et n_2 le nombre de x_i appartenant à $]1/2, 1]$ (évidemment $n_1 + n_2 = n$). Chaque terme du produit de la vraisemblance vaut soit $(1+\theta)$, soit $(1-\theta)$, le nombre de termes qui valent $1+\theta$ est égal à n_1 et le nombre de termes qui valent $1-\theta$ est n_2. Donc

$$L(\theta; (x_1, \cdots, x_n)) = (1+\theta)^{n_1}(1-\theta)^{n_2}.$$

On en déduit la log-vraisemblance:

$$l_n(\theta) = n_1 \ln(1+\theta) + n_2 \ln(1-\theta)\ .$$

Si $n_1 = 0, n_2 = n$, l_n est maximal en $\hat{\theta} = -1$. Si $n_2 = 0, n_1 = n$, l_n est maximal en $\hat{\theta} = 1$. Dans les autres cas, on a

$$l_n'(\hat{\theta}) = \frac{n_1}{(1+\hat{\theta})} - \frac{n_2}{(1-\hat{\theta})} = 0$$

si et seulement si

$$\frac{n_1}{1+\hat{\theta}} = \frac{n_2}{1-\hat{\theta}}.$$

On vérifie que l_n' est positive à gauche et négative à droite de la solution $\theta = \dfrac{n_1 - n_2}{n_1 + n_2}$. Dans tous les cas, on a

$$\hat{\theta} = \frac{n_1 - n_2}{n_1 + n_2} = \frac{n - 2n_2}{n} = 1 - \frac{2n_2}{n}$$

En notant N_2 la statistique associée à n_2:

$$N_2 = \sum_{i=1}^{n} \mathbb{I}_{]1/2;1]}(X_i).$$

nous avons donc l'estimateur:

$$\hat{\theta}_n^{MV} = 1 - \frac{2N_2}{n}$$

2. Nous avons donc:

$$\hat{\theta}_n^{MV} = 1 - \frac{2}{n}\sum_{i=1}^{n} \mathrm{I}_{\{1/2<X_i\leq 1\}}.$$

D'après la loi forte des grands nombres,

$$\hat{\theta}_n^{MV} = 1 - \frac{2}{n}\sum_{i=1}^{n} \mathbb{I}_{]1/2;1]}(X_i) \xrightarrow{p.s.} 1 - 2\mathbb{E}_\theta[\mathrm{I}_{\{1/2<X_1\leq 1\}}]$$

avec $\mathbb{E}_\theta[|\mathrm{I}_{\{1/2<X_1\leq 1\}}|] = \mathbb{E}_\theta[\mathrm{I}_{\{1/2<X_1\leq 1\}}]$ finie:

$$\mathbb{E}_\theta[\mathrm{I}_{\{1/2<X_1\leq 1\}}] = \int_{\mathbb{R}} \mathrm{I}_{\{1/2<x\leq 1\}} p_\theta(x)dx = \int_{1/2}^{1} (1-\theta)dx = \frac{1-\theta}{2}.$$

Donc $\hat{\theta}_n^{MV} \xrightarrow{p.s.} \theta$ presque sûrement lorsque $n \to \infty$, ce qui signifie que $\hat{\theta}_n^{MV}$ est un estimateur fortement consistant. De plus, il est sans biais car

$$\mathbb{E}_\theta[\hat{\theta}_n^{MV}] = 1 - \frac{2}{n}\sum_{i=1}^{n} \mathbb{E}_\theta[\mathbb{I}_{]1/2;1]}(X_i)] = 1 - 2\mathbb{E}_\theta[\mathbb{I}_{]1/2;1]}(X_1)] = \theta.$$

Pour déterminer la loi limite de $\sqrt{n}(\hat{\theta}_n^{MV} - \theta)$, on remplace

$$\theta = 1 - 2\mathbb{E}_\theta[\mathbb{I}_{]1/2;1]}(X_1)]$$

dans l'expression pour obtenir:

$$\sqrt{n}(\hat{\theta}_n^{MV} - \theta) = -\frac{2}{\sqrt{n}}\sum_{i=1}^{n} (\mathbb{I}_{]1/2;1]}(X_1 i) - \mathrm{E}_\theta[\mathbb{I}_{]1/2;1]}(X_i)]) \ .$$

Par conséquent, par application du théorème central limite à la variable aléatoire: $2\mathbb{I}_{]1/2;1]}(X_1)$, on obtient:

$$\sqrt{n}(\hat{\theta}_n^{MV} - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 4\mathrm{Var}(\mathbb{I}_{]1/2;1]}(X_1))) \ .$$

On calcule facilement

$$\begin{aligned} \mathrm{Var}(\mathbb{I}_{]1/2;1]}(X_1)) &= \mathrm{E}(\mathbb{I}_{]1/2;1]}(X_1)) - [\mathrm{E}(\mathbb{I}_{]1/2;1]}(X_1))]^2 \\ &= \mathrm{E}(\mathbb{I}_{]1/2;1]}(X_1)) - [(1-\theta)/2]^2 \\ &= (1-\theta)/2 - [(1-\theta)/2]^2 = \frac{1-\theta^2}{4}. \end{aligned}$$

Donc $\sqrt{n}(\hat{\theta}_n^{MV} - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1-\theta^2)$. La vitesse de convergence est $\sqrt{n}$.

Exercice 1.5. Efficacité des estimateurs

On considère une suite $\{X_i\}_{i\in\mathbb{N}}$ de variables aléatoires de Pareto de paramètres $c > 0$ et $\alpha > 0$, dont la densité est donnée par

$$p_{c,\alpha}(x) = \alpha c^{\alpha} x^{-(\alpha+1)} \mathbb{I}_{[c;+\infty[}(x)$$

On suppose dans un premier temps $c = 1$.

1. Trouver $\hat{\alpha}_n^{MV}$ l'estimateur du maximum de vraisemblance de α.

2. Calculer sa variance.

3. Calculer l'information de Fisher et conclure que l'estimateur $\hat{\alpha}_n^{MV}$ est asymptotiquement efficace.

4. Calculer l'estimateur du maximum de vraisemblance de c lorsque α est connu. Le modèle statistique est-il régulier ?

Solution:

1. Soit $x = (x_1, \cdots, x_n)$, un échantillon d'observations i.i.d. La vraisemblance s'écrit

$$\mathcal{L}(\alpha; x) = \alpha^n \prod_{i=1}^{n} x_i^{-(\alpha+1)} \mathbb{I}_{[1,+\infty)}(\min(x_i)) \ .$$

Donc, en supposant $\min(x_i) \geq 1$,

$$l(\alpha) = \ln \mathcal{L}(\alpha; x) = n \ln \alpha - (\alpha + 1) \sum_{i=1}^{n} \log x_i$$

$$l'(\alpha) = \frac{n}{\alpha} - \sum_{i=1}^{n} \ln x_i \ , \ l''(\alpha) = -\frac{n}{\alpha^2} < 0$$

Donc l est strictement concave et maximale pour

$$\hat{\alpha} = \frac{n}{\sum_{i=1}^{n} \ln(x_i)}.$$

2. Pour ce calcul on remarque que $\mathbb{P}(\ln X_i \leq u) = \mathbb{P}(X_i \leq e^u) = 1 - e^{-\alpha u}$ (en utilisant la fonction de répartition de la loi de Pareto) donc que les $\ln(X)$ sont des variables $\gamma(1, \alpha)$. Comme les variables sont indépendantes, on peut montrer que $\sum_{1\leq i\leq n} \ln(X_i) \sim \gamma(n, \alpha)$ et $Z{=}n\frac{1}{n}\sum_{1\leq i\leq n} \ln(X_i) \sim \gamma(n, n\alpha)$ (propriétés des lois gammas).

On peut alors calculer $\mathbb{E}(\frac{1}{Z_n})$ et $\mathbb{E}(\frac{1}{Z_n^2})$ en utilisant la densité de $\gamma(n, n\alpha)$,

$$p(x) = \frac{(n\alpha)^n}{\Gamma(n)} e^{-n\alpha x} x^{n-1} \mathbb{I}_{[0;+\infty[}(x)$$

et on trouve:

$$\mathbb{E}(\hat{\alpha}_n^{MV}) = \frac{n}{n-1}\alpha \ , \ \mathbb{V}(\hat{\alpha}_n^{MV}) = \frac{n^2\alpha^2}{(n-1)^2(n-2)}$$

sous les conditions $n > 1$ et $n > 2$, respectivement.

3. On est dans le cas d'un modèle régulier donc l'information de Fisher vaut $I(\alpha) = -\mathbb{E}(\partial^2 \ln L/\partial\alpha^2) = 1/\alpha^2$ (pour $n = 1$). Par le TLC, on a $\sqrt{n}(n^{-1}\sum \log X_i - \alpha^{-1}) \to \mathcal{N}(0,\ \alpha^{-2})$, les variables $\ln X_i$ étant i.i.d. de moyenne α^{-1} et de variance α^{-2}. Par la méthode δ (avec la fonction $\phi\ :\ u \to 1/u$, dérivable en $1/\alpha$), on obtient que $\sqrt{n}(n/\sum \ln X_i - \alpha) \to \mathcal{N}(0,\ \alpha^2)$. On conclut que l'EMV est asymptotiquement efficace.

4. Lorsque c est inconnu, le support de la loi des X_i varie avec le paramètre donc le modèle n'est pas régulier (et la question de l'efficacité ne se pose pas). La vraisemblance dans le cas α connu s'écrit

$$L(c;x) = \alpha^n c^{n\alpha} \prod_{i=1}^{n} x_i^{-(\alpha+1)} \mathbb{I}_{[c,+\infty)}(\min x_i)\ .$$

Comme $c \mapsto c^{n\alpha}$ est croissante, L est maximale en $\hat{c} = \min x_i$.

Pour aller plus loin...

Exercice 1.6. EMV pour la régression Linéaire. Soient $\xi_1, \cdots, \xi_n$ des variables aléatoires i.i.d de densité p_ξ par rapport à la mesure de Lebesgue sur $\mathbb{R}$, et soit $x_i \in \mathbb{R}, i = 1, \cdots, n$. On observe les couples $(x_i,\ y_i)$, $1 \le i \le n$, issus du modèle de régression linéaire

$$Y_i = \theta X_i + \xi_i,$$

où $\theta \in \mathbb{R}$ est un paramètre inconnu. On suppose ici que les X_i sont déterministes et non tous nuls.

1. Expliciter la densité jointe de $(Y_1, \cdots, Y_n)$.

2. Montrer que si la loi de ξ_i est $\mathcal{N}(0,1)$, la densité des $(Y_1, \cdots, Y_n)$ est donnée par:

$$p(y_1, \cdots, y_n) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(y_i - \theta x_i)^2\right)\ .$$

En déduire l'estimateur du maximum de vraisemblance $\hat{\theta}^{MV}$ de θ.

3. Quelle est la loi de l'estimateur du maximum de vraisemblance $\hat{\theta}^{MV}$? En déduire son biais, sa variance, son risque quadratique ?

4. On étudie le cas particulier de régression sur le temps: $X_i = i, \forall i \in \mathbb{N}^*$. Quelle est la vitesse de convergence de l'estimateur ?

5. Proposer la prévision linéaire de Y_{n+1} basée sur $(Y_1,\ \ldots,\ Y_n)$.

Solution: 1. Soit: $\varphi : \mathbb{R}^n \to \mathbb{R}^n$, avec $\varphi(s_1, \cdots, s_n) = (s_1 + \theta x_1, \cdots, s_n + \theta x_n)$. Nous avons que $(Y_1, \cdots, Y_n) = \phi(\xi_1, \cdots, \xi_n)$ et que φ est un $\mathcal{C}^1$-difféomorphisme de $\mathbb{R}^n$ dans $\mathbb{R}^n$, $\varphi^{-1}(y_1, \cdots, y_n) = (y_1 - \theta x_1, \cdots, y_n - \theta x_n)$.

D'après le théorème de la loi image, nous avons alors:

$$p_{(Y_1,\cdots,Y_n)}\left(y_1,\cdots,y_n\right)=p_{(\xi_1,\cdots,\xi_n)}\left(\varphi^{-1}\left(y_1,\cdots,y_n\right)\right)|J_{\varphi^{-1}}(y_1,\cdots,y_n)|$$

Or:

$$p_{(\xi_1,\cdots,\xi_n)}\left(s_1,\cdots,s_n\right)=p_\xi\left(s_1\right)\cdots p_\xi\left(s_n\right)$$

comme les ξ_i sont indépendantes de densité p_ξ. Par ailleurs,

$$J_{\phi^{-1}}=Id$$

d'où finalement:

$$p_{(Y_1,\ldots,Y_n)}(y_1,\ \ldots,\ y_n)=\prod_{i=1}^n p_\xi(y_i-\theta x_i)$$

2. Dans la question précédente, en remplacant p_ξ par la densité de la loi normale, on obtient:

$$\begin{aligned}\mathcal{L}_n\left(\theta;\left(x_1,y_1\right),\cdots,\left(x_n,y_n\right)\right)&=p_{(Y_1,\ldots,Y_n)}(y_1,\cdots,y_n)\\&=\frac{1}{(2\pi)^{n/2}}\exp(-\frac{1}{2}\sum_{i=1}^n(y_i-\theta x_i)^2)\end{aligned}$$

On en déduit que

$$l_n(\theta)=-\frac{n}{2}\ln(2\pi)-\frac{1}{2}\sum_{i=1}^n(y_i-\theta x_i)^2.$$

D'où

$$l_n'(\theta)=\sum_{i=1}^n x_i(y_i-\theta x_i)=-\theta\sum_{i=1}^n x_i^2+\sum_{i=1}^n x_iy_i.$$

On a $l_n''=-\sum x_i^2<0$, donc $\hat\theta$ est obtenu en annulant la dérivée, ce qui nous donne l'estimateur:

$$\hat\theta^{MV}=\frac{\sum_{i=1}^n X_iY_i}{\sum_{i=1}^n X_i^2}\ .$$

3. Pour déterminer la loi de l'estimateur, on remarque que

$$\hat\theta_n^{MV}=\frac{\sum_{i=1}^n X_i(\theta X_i+\xi_i)}{\sum_{i=1}^n X_i^2}=\theta+\frac{\sum_{i=1}^n X_i\xi_i}{\sum_{i=1}^n X_i^2}\ .$$

Comme les ξ_i sont i.i.d. de loi normale $\mathcal{N}(0,1)$ et les X_i sont déterministes, on a

$$\frac{\sum_{i=1}^n X_i\xi_i}{\sum_{i=1}^n X_i^2}\sim\mathcal{N}(0,\ \frac{1}{\sum_{i=1}^n X_i^2})\ .$$

D'où

$$\hat\theta_n^{MV}\sim\mathcal{N}(\theta,\ \frac{1}{\sum_{i=1}^n X_i^2})\ .$$

On en déduit que l'estimateur est non biaisé et que le risque quadratique de $\hat{\theta}_n^{MV}$ coincide avec sa variance. Donc

$$\mathbb{E}_\theta[(\hat{\theta}_n^{MV} - \theta)^2] = \mathbb{V}(\hat{\theta}_n) = \frac{1}{\sum_{i=1}^n X_i^2}.$$

4. En remplacant X_i par i et en utilisant la formule $\displaystyle\sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{6}$, on obtient

$$\sqrt{n(n+1)(n+1/2)}\left(\hat{\theta}_n^{MV} - \theta\right) \sim \mathcal{N}(0,3) \ .$$

Par conséquent, l'estimateur a une vitesse de convergence en $n^{3/2}$.

5. L'estimation de θ est basée sur l'échantillon $Y_1, \ . \ . \ . \ , Y_n$, et on définira la prédiction de Y_{n+1} par la formule:

$$\hat{Y}_{n+1} = \mathbb{E}\left(\hat{\theta}_n^{MV} X_{n+1} + \xi_{n+1}\right) \ .$$

Notons que d'autres choix que l'espérance seraient possibles ... Le mode, la médiane... Dans ce cas précis, les différents choix sont équivalents.

$$\hat{Y}_{n+1} = \hat{\theta}_n^{MV} X_{n+1} = (n+1)\hat{\theta}_n^{MV} = \frac{6(n+1)\sum_{i=1}^n iY_i}{n(n+1)(2n+1)} = \frac{6}{n(2n+1)}\sum_{i=1}^n iY_i.$$

Exercice 1.7. Modèle statistique pour une série financière. On mesure le cours d'une action Y_t au cours du temps (toutes les minutes par exemple) et on s'intéresse à la modélisation des $\ln$-retours, c'est-à-dire des quantités $X_t = \ln(Y_{t+1}/Y_t)$. Sur la Figure 1, on a représenté la simulation d'une série financière $x_1, \cdots, x_n$, ainsi que l'histogramme des valeurs observées et le profil de la queue de distribution $G : t \mapsto Card\{i \,|\, x_i > t\}/n$, en coordonnées logarithmiques.

On rappelle qu'une variable de Cauchy de paramètre m et c admet une densité

$$f_{m,c}(x) = \frac{1}{\pi c}\,\frac{1}{1 + (x-m)^2/c^2}$$

où m est le paramètre de position de la loi et c le paramètre d'échelle.

1. Justifier l'utilisation d'une loi de Cauchy plutôt que d'une loi normale pour modéliser les valeurs x_t observées.

2. Pour X une loi de Cauchy de paramètres m et c, que vaut $\mathbb{E}(|X|)$?

3. Que vaut la médiane de X ?

Solution: 1. Si T est une v.a. de loi de Cauchy de paramètre de position $m = 0$ et d'échelle $c = 1$, on a $\mathbb{P}(T > t) = \dfrac{1}{2} - \arctan(t)/\pi \sim 1/(\pi t)$ quand $t \to \infty$.

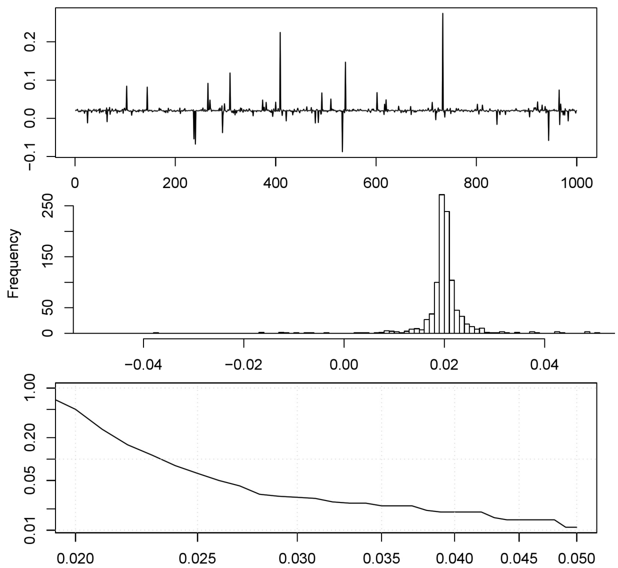

Figure 1: Série financière: (a) valeurs de x_t au cours du temps; (b) histogramme des $(x_t)_t$; (c) $G(t)$ en fonction de t, en échelle logarithmique, avec $G : t \mapsto Card\{i \,|\, x_i > t\}/n$

Pour une v.a. T' de loi de Cauchy de paramètres m, c, $(T' - m)/c$ est une v.a. de Cauchy standard et donc on a également $\mathbb{P}(T' > t) = \mathbb{P}(T > (t-m)/c) \sim c/(\pi t)$. Donc $\ln \mathbb{P}(T' > t) \sim -\ln t$. Or $G : t \mapsto Card\{i|X_i > t\}/n$ correspond à la queue de distribution empirique. L'asymptote linéaire dans le tracé de $t \to \mathbb{P}(T' > t)$ en coordonnées logarithmiques est donc bien compatible avec $\ln \mathbb{P}(T' > t) \sim -\ln t$.

À l'inverse, dans le cas d'une variable gaussienne centrée réduite, on a

$$\mathbb{P}(T > t) = (2\pi)^{-1/2} \int_{x>t} e^{-x^2/2} dx \le (2\pi)^{-1/2} \int_{x>t} \frac{x}{t} e^{-x^2/2} dx = (2\pi)^{-1/2} \frac{1}{t} e^{-t^2/2}.$$

Le changement de moyenne et de variance modifie t en $(t-m)/\sigma$ dans cette borne mais dans tous les cas, $\ln \mathbb{P}(X > t)$ décroît de façon au moins quadratique en t donc exponentielle en $\ln t$ ce qui est loin de l'allure de la fonction de répartition empirique.

2. Une variable de Cauchy n'a pas d'espérance, et donc a fortiori aucun moment fini. En effet, $\forall p \ge 1, \int |x|^p f_{m,c}(x) dx = +\infty$. On ne peut donc pas espérer estimer m en utilisant une moyenne empirique.

3. Une variable à densité admet une unique médiane donnée par $F^{-1}(1/2) = G^{-1}(1/2)$ (avec F la distribution et $G = 1 - F$ la queue de distribution). On vérifie facilement (par le calcul ou un argument de symétrie) que m est cette médiane. Sur le graphe 3, on cherche $\hat{m}$ tel que $G(\hat{m}) = 1/2$, on a que $\hat{m} \simeq 0.02.$, ce qui est cohérent avec l'histogramme.