Estimation ponctuelle

Modèle paramétrique

On observe un échantillon $(X_1, ..., X_n)$ où les X_j sont des v.a. i.i.d.. On parle d'un modèle paramétrique si la loi commune des X_j appartient à une famille paramètre $\mathcal{P} = \{\mathbb{P}_\theta, \theta \in \Theta\}$.

Exemple 1. Modèle de Bernoulli: $\mathcal{P} = \{\mathcal{B}(p), p \in [0,1]\}$, $\theta = p$, $\Theta = [0,1]$.

Modèle Uniforme: $\mathcal{P} = \{\mathcal{U}([0,\theta]), \theta \in \Theta = \mathbb{R}_+ \backslash \{0\}\}$.

Modèle Gaussien: $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+ \backslash \{0\}\}$, $\theta = (\mu, \sigma^2)$, $\Theta = \mathbb{R} \times (\mathbb{R}_+ \backslash \{0\})$

Notations.

- $\mathbb{P}_\theta(X \in A)$: probabilité que $X \in A$ lorsque X suit $\mathbb{P}_\theta$.
- $\mathbb{E}_\theta[h(X)]$: Espérance de $h(X)$ lorsque X suit $\mathbb{P}_\theta$.
- Si X est une v.a. discrète, i.e. X est à valeurs dans un ensemble au plus dénombrable $X \in \{1, 2, ...\}$ on note $\mathbb{P}_\theta(X = x) = p(x, \theta)$.
- Si X est une v.a. continue alors on notera $f(x, \theta)$ la densité de X selon $\mathbb{P}_\theta$.
- Soit $X = (X_1, ..., X_n)$ un échantillon de taille n où les X_j sont i.i.d.

  1. *Dans le cas discret:* $\mathbb{P}_\theta(X = x) = p(x, \theta) = \prod_{j=1}^n p(x_j, \theta)$
  2. *Dans le cas continu:* $f(x, \theta) = \prod_{j=1}^n f(x_j, \theta)$

  ou $x = (x_1, ..., x_n)$ est la réalisation de l'échantillon $X = (X_1, ..., X_n)$.

Exemple 2. Dans le modèle de Bernoulli $(X_1, ..., X_n)$ un échantillon de n v.a. i.i.d. $\sim \mathcal{B}(p)$, où $p \in [0,1]$,

$$p(x, p) = \prod_{j=1}^n p(x_j, p) = \prod_{j=1}^n [p^{x_j}(1-p)^{1-x_j}] = p^{\sum_{j=1}^n x_j}(1-p)^{n - \sum_{j=1}^n x_j}.$$

Dans le modèle Gaussien

$$f(x, \mu, \sigma^2) = \prod_{j=1}^n f(x, \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{j=1}^n (x_j - \mu)^2}$$

Définition 3. *On appelle statistique toute v.a. S qui dépend de l'échantillon $(X_1, ..., X_n)$ mais qui ne fait pas intervenir le paramètre θ.*

Exemple 4. Quelques statistiques:

- $\sum_{j=1}^n X_j$
- $\overline{X}_n$ (la moyenne empirique)
- $\max_{1 \leqslant j \leqslant n}(X_j)$

– $\frac{1}{n}\sum_{j=1}^{n}(X_j-\overline{X}_n)^2$ (variance empirique)

Mais θX_1 et $(\overline{X}_n+\theta^2\max_{1\leqslant j\leqslant n}(X_j))$ ne sont pas des statistique.

Estimation ponctuelle

Définition 5. *Soit g une application sur Θ. On appelle estimateur (ponctuel) de $g(\theta)$ toute statistique T prenant ses valeurs dans $g(\Theta)$.*

Exemple 6. Dans le modèle de Bernoulli $\mathcal{P}=\{\mathcal{B}(p), p\in[0,1]\}$ la moyenne empirique $\overline{X}_n$ est un estimateur de p ($g(p)=p$ c-à-d g est l'identité sur $\Theta=[0,1]$).

Dans le modèle Gaussien $\mathcal{P}=\{\mathcal{N}(\mu,\sigma^2),(\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}_+^*\}$ la variance empirique

$$S_n^2=\frac{1}{n}\sum_{j=1}^{n}(X_j-\overline{X}_n)^2$$

est un estimateur de σ^2 ($g(\mu,\sigma^2)=\sigma^2$) et $S_n=\sqrt{S_n^2}$ (l'écart-type empirique) est un estimateur de σ (l'écart-type théorique).

Quelques propriétés des estimateurs

Définition 7. *Un estimateur T de $g(\theta)$ est dit sans biais (ou non biaisé) si $\mathbb{E}_\theta[T]=g(\theta)$. Autrement, le biais $b(\theta)=\mathbb{E}_\theta[T]-g(\theta)=0$.*

Définition 8. *On appelle risque quadratique d'un estimateur T, et on note $R(T,\theta)$ la quantité $R(T,\theta)=\mathbb{E}_\theta[(T-g(\theta))^2]$. En particulier, si T est sans biais alors $R(T,\theta)=\mathbb{E}_\theta[(T-\mathbb{E}_\theta[T])^2]=\mathrm{Var}_\theta(T)$.*

Remarque 9. On peut toujours écrire

$$R(T,\theta)=\mathrm{Var}_\theta(T)+(b(\theta))^2$$

donc dans le risque il y a une partie due à la variance de la statistique et un autre du à son biais. En effet

$$\begin{aligned}R(T,\theta)&=\mathbb{E}_\theta[(T-g(\theta))^2]=\mathbb{E}_\theta[((T-\mathbb{E}_\theta[T])+(\mathbb{E}_\theta[T]-g(\theta)))^2]\\&=\mathbb{E}_\theta[(T-\mathbb{E}_\theta[T])^2+2(\mathbb{E}_\theta[T]-g(\theta))(T-\mathbb{E}_\theta[T])+(\mathbb{E}_\theta[T]-g(\theta))^2]\\&=\mathbb{E}_\theta[(T-\mathbb{E}_\theta[T])^2]+2b(\theta)\mathbb{E}_\theta[T-\mathbb{E}_\theta[T]]+(b(\theta))^2\\&=\mathrm{Var}_\theta(T)+(b(\theta))^2\end{aligned}$$

Définition 10. *Soient T_1 et T_2 deux estimateurs non biaises de $g(\theta)$. On dira que T_2 est plus efficace de T_1 si $R(T_2,\theta)\leqslant R(T_1,\theta)$ pour tout $\theta\in\Theta$, c-à-d $\mathrm{Var}_\theta(T_2)\leqslant\mathrm{Var}_\theta(T_1)$ pour tout $\theta\in\Theta$.*

Exemple 11. Reprenons l'exemple du modèle de Bernoulli $\mathcal{P}=\{\mathcal{B}(p), p\in[0,1]\}$. Comparons les estimateurs X_1 et $\overline{X}_n$ (Il s'agit d'estimer le paramètre p).

$\mathbb{E}[X_1]=p$ donc X_1 est un estimateur non biaisé de p.

$\mathbb{E}[\overline{X}_n]=p\ \Rightarrow\overline{X}_n$ est aussi un estimateur non biaisé de p.

$R(X_1,p)=\mathrm{Var}_p(X_1)=p(1-p)$

$R(\overline{X}_n, p) = \mathrm{Var}_p(\overline{X}_n) = (\mathrm{Var}_p(X_1) + \cdots + \mathrm{Var}_p(X_n))/n^2 = p(1-p)/n$

On en déduit que $\mathrm{Var}_p(\overline{X}_n) \leqslant \mathrm{Var}_p(X_1)$ pour tout $p \in [0,1]$ donc $\overline{X}_n$ est un estimateur plus efficace que X_1.

Exemple 12. Modèle Uniforme. $\mathcal{P} = \{\mathcal{U}([0,\theta]), \theta \in \mathbb{R}_+^*\}$, $f(x,\theta) = \theta^{-1}\mathbb{I}_{x\in[0,\theta]}$. On considère les estimateurs suivants: $T_1 = 2\,\overline{X}_n$ et $T_2 = [(n+1)/n]\max_{1\leqslant j\leqslant n} X_j$.

On observe un échantillon de taille n. Montrons que ces estimateurs sont non biaises:

$\mathbb{E}_\theta[2\overline{X}_n] = 2\,\mathbb{E}_\theta[X_1] = 2\frac{\theta}{2} = \theta$ (non biaisé)

On pose $Y = \max_{1\leqslant j\leqslant n} X_j$. Soit $y \in \mathbb{R}$,

$$\mathbb{P}_\theta(Y \leqslant y) = \mathbb{P}_\theta(X_1 \leqslant y, ..., X_n \leqslant y) = \mathbb{P}_\theta(X_1 \leqslant y)^n = \begin{cases} 0 & \text{si } y \leqslant 0 \\ (y/\theta)^n & \text{si } y \in [0,\theta] \\ 1 & \text{si } y > \theta \end{cases}$$

Donc Y admet pour densité la fonction $g(y,\theta)$ donnée par

$$g(y,\theta) = \frac{n}{\theta^n} y^{n-1}\mathbb{I}_{0\leqslant y\leqslant\theta}$$

et

$$\mathbb{E}_\theta[Y] = \int_0^\theta \frac{n}{\theta^n} y^n \mathrm{d}y = \frac{n}{n+1}\theta$$

ce qui donne

$$\mathbb{E}_\theta[T_2] = \frac{n+1}{n}\mathbb{E}_\theta[Y] = \theta$$

et par conséquence T_2 est un estimateur non biaisé de θ. Calculons les variances respectives de T_1 et T_2:

$$\mathrm{Var}_\theta(T_1) = 4\mathrm{Var}_\theta(\overline{X}_n) = \frac{4}{n}\mathrm{Var}_\theta(X_1) = \frac{\theta^2}{3n}$$

et

$$\mathrm{Var}_\theta(T_2) = \left(\frac{n+1}{n}\right)^2 \mathrm{Var}_\theta(Y) = \left(\frac{n+1}{n}\right)^2 \left[\mathbb{E}_\theta[Y^2] - \left(\frac{n}{n+1}\right)^2 \theta^2\right]$$

Or

$$\mathbb{E}_\theta[Y^2] = \int_0^\theta \frac{n}{\theta^n} y^{n+1} \mathrm{d}y = \frac{n}{n+2}\theta^2$$

donc

$$\mathrm{Var}_\theta(T_2) = \left(\frac{n+1}{n}\right)^2 \left[\frac{n}{n+2} - \left(\frac{n}{n+1}\right)^2\right]\theta^2 = \frac{\theta^2}{n(n+2)}$$

et

$$\frac{\mathrm{Var}_\theta(T_2)}{\mathrm{Var}_\theta(T_1)} = \frac{3}{n+2} \leqslant 1$$

qui montre que T_2 est plus efficace que T_1.

Définition 13. *Soit l'application $g\colon \Theta \to \mathbb{R}^d$. On dit que la suite $(T_n)_{n\geqslant 1}$ d'estimateurs de $g(\theta)$ est*

1. *Convergente: si $(T_n)_{n\geqslant 1}$ converge en probabilité vers $g(\theta)$ pour tout $\theta \in \Theta$.*

2. *Fortement convergente: si $(T_n)_{n\geqslant 1}$ converge presque sûrement vers $g(\theta)$ pour tout $\theta\in\Theta$.*

3. *Asymptotiquement normale: si pour tout $\theta\in\Theta$ existe une matrice de covariance $\Sigma(\theta)$ telle que $\sqrt{n}(T_n-g(\theta))\xrightarrow{\mathcal{L}}\mathcal{N}_d(0,\Sigma(\theta))$.*

Exemple 14. Soient $X_1,\ldots,X_n$ un échantillon de loi $\mathcal{B}(p)$. D'après la loi forte des grandes nombres

$$\overline{X}_n\xrightarrow{p.s.}p \qquad \text{pour tout } p\in[0,1]$$

la suite $(X_n)_{n\geqslant 1}$ est fortement convergente. De plus $\mathrm{Var}_p(X_1)=p(1-p)<+\infty$ et d'après le TCL

$$\sqrt{n}(\overline{X}_n-p)\xrightarrow{\mathcal{L}}\mathcal{N}(0,p(1-p))$$

la suite $(X_n)_{n\geqslant 1}$ est asymptotiquement normale avec $\Sigma(p)=p(1-p)$.

Exhaustivité

Définition 15. *Une statistique S est dite exhaustive si la loi conditionnelle de $X=(X_1,\ldots,X_n)$ sachant $S=s$ ne dépends pas du paramètre θ pour tout s.*

Théorème 16. *(De factorisation) S est une statistique exhaustive ssi il existe des applications g et h telles que*

$$p(\boldsymbol{x},\theta)=g(\boldsymbol{x})\,h(S(\boldsymbol{x}),\theta) \text{ dans le cas discret}$$

$$f(\boldsymbol{x},\theta)=g(\boldsymbol{x})h(S(\boldsymbol{x}),\theta) \text{ dans le cas continu}$$

Rappel. $p(\boldsymbol{x},\theta)=\prod_{j=1}^n p(x_j,\theta)$ et $f(\boldsymbol{x},\theta)=\prod_{j=1}^n f(x_j,\theta)$

Démonstration. (Uniquement dans le cas discret). ($\Rightarrow$) Supposons que S est exhaustive. $p(\boldsymbol{x},\theta)=\mathbb{P}_\theta(\boldsymbol{X}=\boldsymbol{x})=\mathbb{P}_\theta(\boldsymbol{X}=\boldsymbol{x} \text{ et } S(\boldsymbol{X})=S(\boldsymbol{x}))$ car $\{S(\boldsymbol{X})=S(\boldsymbol{x})\}\supseteq\{\boldsymbol{X}=\boldsymbol{x}\}$. Donc $\mathbb{P}_\theta(\boldsymbol{X}=\boldsymbol{x})=\mathbb{P}_\theta(\boldsymbol{X}=\boldsymbol{x}\,|\,S(\boldsymbol{X})=S(\boldsymbol{x}))\mathbb{P}_\theta(S(\boldsymbol{X})=S(\boldsymbol{x}))$. Or S est exhaustive $\Rightarrow\mathbb{P}_\theta(\boldsymbol{X}=\boldsymbol{x}\,|\,S(\boldsymbol{X})=S(\boldsymbol{x}))$ ne dépends pas de θ: $\mathbb{P}_\theta(\boldsymbol{X}=\boldsymbol{x}\,|\,S(\boldsymbol{X})=S(\boldsymbol{x}))=\mathbb{P}(\boldsymbol{X}=\boldsymbol{x}\,|\,S(\boldsymbol{X})=S(\boldsymbol{x}))$ et on pose $g(\boldsymbol{x})=\mathbb{P}(\boldsymbol{X}=\boldsymbol{x}\,|\,S(\boldsymbol{X})=S(\boldsymbol{x}))$ et $h(s,\theta)=\mathbb{P}_\theta(S(\boldsymbol{X})=s)$. Il vient que

$$p(\boldsymbol{x},\theta)=\mathbb{P}(\boldsymbol{X}=\boldsymbol{x}\,|\,S(\boldsymbol{X})=S(\boldsymbol{x}))\mathbb{P}_\theta(S(\boldsymbol{X})=S(\boldsymbol{x}))=g(\boldsymbol{x})\,h(S(\boldsymbol{x}),\theta).$$

($\Leftarrow$) Réciproquement, supposons qu'il existe g et h telles que

$$p(\boldsymbol{x},\theta)=g(\boldsymbol{x})\,h(S(\boldsymbol{x}),\theta)$$

et montrons que S est exhaustive. Fixons s. On pose $A_s=\{\boldsymbol{y}:S(\boldsymbol{y})=s\}$.

$$\mathbb{P}_\theta(S(\boldsymbol{X})=s)=\sum_{\boldsymbol{x}\in A_s}\mathbb{P}_\theta(\boldsymbol{X}=\boldsymbol{x})=\sum_{\boldsymbol{x}\in A_s}g(\boldsymbol{x})\,h(S(\boldsymbol{x}),\theta)=h(s,\theta)\sum_{\boldsymbol{x}\in A_s}g(\boldsymbol{x})$$

$$\mathbb{P}_\theta(\boldsymbol{X}=\boldsymbol{x}|S(\boldsymbol{X})=s)=\frac{\mathbb{P}_\theta(\boldsymbol{X}=\boldsymbol{x},S(\boldsymbol{X})=s)}{\mathbb{P}_\theta(S(\boldsymbol{X})=s)}=\begin{cases}0 & \text{si } s\neq S(\boldsymbol{x})\\ \frac{\mathbb{P}_\theta(\boldsymbol{X}=\boldsymbol{x})}{\mathbb{P}_\theta(S(\boldsymbol{X})=s)} & \text{si } s=S(\boldsymbol{x})\end{cases}$$

Si $s = S(\boldsymbol{x})$

$$\frac{\mathbb{P}_\theta(\boldsymbol{X} = \boldsymbol{x})}{\mathbb{P}_\theta(S(\boldsymbol{X}) = s)} = \frac{g(\boldsymbol{x})\, h(s,\theta)}{h(s,\theta)\sum_{\boldsymbol{x}\in A_s} g(\boldsymbol{x})} = \frac{g(\boldsymbol{x})}{\sum_{\boldsymbol{x}\in A_s} g(\boldsymbol{x})}$$

donc

$$\mathbb{P}_\theta(\boldsymbol{X} = \boldsymbol{x} | S(\boldsymbol{X}) = s) = \begin{cases} 0 & \text{si } s \neq S(\boldsymbol{x}) \\ \frac{g(\boldsymbol{x})}{\sum_{\boldsymbol{x}\in A_s} g(\boldsymbol{x})} & \text{si } s = S(\boldsymbol{x}) \end{cases}$$

qui ne dépends pas de θ et qui donne l'exhaustivité de S. □

Exemple 17. Soit $\boldsymbol{X} = (X_1, ..., X_n)$ un échantillon de Bernoulli de paramètre p et $S(\boldsymbol{X}) = \sum_{j=1}^n X_j$. Montrons que S est exhaustive pour p.

$$p(\boldsymbol{x},\theta) = \prod_{j=1}^n p^{x_j}(1-p)^{1-x_j} = p^{\sum_{j=1}^n x_j}(1-p)^{n-\sum_{j=1}^n x_j} = g(\boldsymbol{x})\, h(S(\boldsymbol{x}), p)$$

avec $g(\boldsymbol{x}) = 1$ et $h(s, p) = p^s(1-p)^{n-s}$. Par le théorème de factorisation on en déduit que S est exhaustive pour p.

Exemple 18. Soit $\boldsymbol{X} = (X_1, ..., X_n)$ un échantillon Gaussien de moyenne μ et variance σ^2. On pose

$$S(\boldsymbol{X}) = \left(\sum_{j=1}^n X_j, \sum_{j=1}^n X_j^2\right).$$

Montrons que S est exhaustive pour (μ, σ^2):

$$f(\boldsymbol{x},\theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{j=1}^n (x_j-\mu)^2} = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{j=1}^n (x_j^2 - 2\mu x_j + \mu^2)}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}(\sum_{j=1}^n x_j^2 - 2\mu\sum_{j=1}^n x_j + n\mu^2)} = g(\boldsymbol{x})\, h(S(\boldsymbol{X}), (\mu, \sigma^2))$$

où $g(\boldsymbol{x}) = 1$ et

$$h((s_1, s_2), (\mu, \sigma^2)) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}(s_2 - 2\mu s_1 + n\mu^2)}$$

Par le théorème de factorisation on en déduit que S est exhaustive pour (μ, σ^2).

Méthodes d'estimation

Méthode des moments

Définition 19. *Soit $r \in \mathbb{N}^*$. Si X est une v.a. réelle t.a. $\mathbb{E}[|X|^r] < +\infty$ alors on appelle $\mathbb{E}[X^r]$ le moment d'ordre r de X et on le note m_r.*

Définition 20. *On appelle moment empirique d'ordre r la statistique*

$$\frac{1}{n}\sum_{j=1}^n (X_j)^r$$

que on notera $M_{r,n}$.

Exemple 21. $M_{1,n} = \overline{X_n}$ et $M_{2,n} = n^{-1} \sum_{j=1}^{n} X_j^2$.

Définition 22. *Soit $\theta = (\theta_1, \ldots, \theta_d) \in \Theta \subseteq \mathbb{R}^d$. On suppose que X est une v.a. réelle telle que $\mathbb{E}[|X|^d] < +\infty$. S'il existe des applications $\varphi_1, \ldots, \varphi_d$ telles que $\theta_j = \varphi_j(m_1, m_2, \ldots, m_d)$, alors l'estimateur obtenu par la méthode des moments est donnée par*

$$\hat{\theta}_n = (\varphi_1(M_{1,n}, \ldots, M_{d,n}), \ldots, \varphi_d(M_{1,n}, \ldots, M_{d,n})).$$

Exemple 23. $X \sim \mathcal{P}(\lambda)$, $\lambda > 0$. $\lambda = \mathbb{E}[X] = m_1$. L'estimateur de λ obtenu par la méthode des moments est

$$\hat{\lambda_n} = M_{1,n} = \overline{X_n}.$$

Exemple 24. $X \sim \mathcal{N}(\mu, \sigma^2)$. $\mu = \mathbb{E}[X] = m_1$ et $\sigma^2 = \mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = m_2 - m_1^2$. L'estimateur de $\theta = (\mu, \sigma^2)$ obtenu par la méthode des moments est $\hat{\theta_n} = (\hat{\mu}_n, \hat{\sigma}_n^2)$ où $\hat{\mu}_n = M_{1,n} = \overline{X_n}$ et $\hat{\sigma}_n^2 = M_{2,n} - M_{1,n}^2$:

$$\hat{\sigma}_n^2 = M_{2,n} - M_{1,n}^2 = \frac{1}{n} \sum_{j=1}^{n} X_j^2 - \overline{X}_n^2 = \frac{1}{n} \sum_{j=1}^{n} (X_j - \overline{X}_n)^2 = S_n^2$$

la variance empirique.

Méthode de maximum de vraisemblance

Définition 25. *Soit $\boldsymbol{x} = (x_1, \ldots, x_n)$ une réalisation d'un échantillon $\boldsymbol{X} = (X_1, \ldots, X_n)$ de n v.a. iid. La fonction $L_n(\boldsymbol{x}, \theta)$ (ou $L_n(\theta)$) donnée par*

- $L_n(\theta) = L_n(\boldsymbol{x}, \theta) = p(\boldsymbol{x}, \theta) = \prod_{j=1}^{n} p(x_j, \theta)$ *dans le cas discret*
- $L_n(\theta) = L_n(\boldsymbol{x}, \theta) = f(\boldsymbol{x}, \theta) = \prod_{j=1}^{n} f(x_j, \theta)$ *dans le cas continu*

vue comme fonction de θ pour $\boldsymbol{x}$ fixé, s'appelle la vraisemblance de la réalisation $\boldsymbol{x} = (x_1, \ldots, x_n)$.

Définition 26. *On suppose que pour toute réalisation $\boldsymbol{x} = (x_1, \ldots, x_n)$ il existe une unique valeur $\hat{\theta}_n(\boldsymbol{x})$ qui maximise la vraisemblance $L_n(\boldsymbol{x}, \theta)$. Alors la statistique $\hat{\theta}_n(\boldsymbol{X})$ est appelée l'estimateur de maximum de vraisemblance (EMV) du paramètre θ. Il revient au même de maximiser la vraisemblance ou son logarithme, c-à-d la log-vraisemblance, souvent notée $\ell_n(\theta)$ ou $\ell_n(\boldsymbol{x}, \theta)$, i.e.*

$$\ell_n(\boldsymbol{x}, \theta) = \ell_n(\theta) = \log L_n(\theta) = \begin{cases} \sum_{j=1}^{n} \log(p(x_j, \theta)) & \text{dans le cas discret ;} \\ \sum_{j=1}^{n} \log(f(x_j, \theta)) & \text{dans le cas continu .} \end{cases}$$

Remarque 27. Pour chercher l'EMV, on cherchera les solutions de l'équation

$$\left(\frac{\partial \ell_n}{\partial \theta_1}, \ldots, \frac{\partial \ell_n}{\partial \theta_d} \right) = (0, \ldots, 0)$$

et vérifier que la matrice Hessienne

$$\left(\frac{\partial^2 \ell_n}{\partial \theta_i \partial \theta_j}\right)_{i,j=1,\ldots,d}$$

est définie négative (on a supposé que $\theta \mapsto \ell_n(\theta)$ est $C^2(\Theta)$).

Exemple 28. Soit $\boldsymbol{X} = (X_1, \ldots, X_n)$ un échantillon Bernoulli de paramètre $p \in [0, 1]$.

$$L_n(\boldsymbol{x}, p) = p^{\sum_{j=1}^n x_j}(1-p)^{n-\sum_{j=1}^n x_j}$$

$$\ell_n(\boldsymbol{x}, p) = \sum_{j=1}^n x_j \log(p) + \left(n - \sum_{j=1}^n x_j\right)\log(1-p)$$

$$\frac{\partial \ell_n(\boldsymbol{x}, p)}{\partial p} = \sum_{j=1}^n x_j \frac{1}{p} - \left(n - \sum_{j=1}^n x_j\right)\frac{1}{1-p}$$

$$\frac{\partial \ell_n(\boldsymbol{x}, p)}{\partial p} = 0 \Longleftrightarrow \sum_{j=1}^n x_j \frac{1}{p} - \left(n - \sum_{j=1}^n x_j\right)\frac{1}{1-p} = 0$$

qui donne:

$$p = \frac{1}{n}\sum_{j=1}^n x_j$$

Donc $\frac{\partial \ell_n(\boldsymbol{x}, p)}{\partial p} = 0$ admet unique solution $\hat{p}_n = \frac{1}{n}\sum_{j=1}^n x_j = \overline{x}_n$ et

$$\frac{\partial^2 \ell_n(\boldsymbol{x}, p)}{\partial p^2} = -\sum_{j=1}^n x_j \frac{1}{p^2} - \left(n - \sum_{j=1}^n x_j\right)\frac{1}{(1-p)^2} < 0$$

ce qu'implique que $\hat{p}_n$ est un maximum globale de $p \mapsto \ell_n(\boldsymbol{x}, p)$ et que l'EMV de p est $\overline{X}_n$.

Eléments de théorie de l'information

Dans cette section on supposera toujours que $\Theta \subseteq \mathbb{R}^d$, Θ ouvert et $d \geqslant 1$.

Définition 29. *On appelle score la fonction $S(x, \theta) = (S_1(x, \theta), \ldots, S_d(x, \theta))$ à valeurs dans $\mathbb{R}^d$ où*

$$S_i(x, \theta) = \frac{\partial}{\partial \theta_i}\ell_n(x, \theta), \qquad i = 1, \ldots, d.$$

On appelle information de Fisher (apporté par X sur θ) la matrice symétrique $I(\theta)$ de dimension $d \times d$ donnée par

$$I(\theta)_{ij} = \mathbb{E}_\theta[S_i(X, \theta)S_j(X, \theta)] = \mathbb{E}_\theta\left[\frac{\partial}{\partial \theta_i}\ell_n(X, \theta)\frac{\partial}{\partial \theta_i}\ell_n(X, \theta)\right], \qquad i, j = 1, \ldots, d.$$

Remarque 30. On a que $I(\theta) = \mathbb{E}_\theta[S(X, \theta)\, S(X, \theta)^T]$.

Exemple 31. (Exponentielle) $X \sim \mathcal{E}(\lambda)$, $\lambda > 0$. $\lambda = \theta \in \Theta =]0, +\infty[$, $d = 1$.

$$f(x, \lambda) = \lambda e^{-\lambda x} \mathbb{I}_{x>0}$$

Soit $x > 0$:

$$S(x, \lambda) = \frac{d}{d\lambda} \log f(x, \lambda) = \frac{1}{\lambda} - x$$

$$I(\lambda) = \mathbb{E}_\lambda[S(X, \lambda)^2] = \mathbb{E}_\lambda\left[\left(X - \frac{1}{\lambda}\right)^2\right] = \mathrm{Var}_\lambda(X) = \frac{1}{\lambda^2}$$

Exemple 32. (Gaussienne) $X \sim \mathcal{N}(\mu, v)$, $\theta = (\mu, v) \in \Theta = \mathbb{R} \times \mathbb{R}_+^*$, $d = 2$.

$$\ell_n(\mu, v) = \log f(x, \mu, v) = -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log v - \frac{1}{2} \frac{(x-\mu)^2}{v}.$$

$$S_1(\mu, v) = \frac{\partial}{\partial \mu} \ell_n(\mu, v) = \frac{x - \mu}{v}$$

$$S_2(\mu, v) = \frac{\partial}{\partial v} \ell_n(\mu, v) = -\frac{1}{2v} + \frac{1}{2} \frac{(x-\mu)^2}{v^2}$$

$$I(\mu, v)_{11} = \mathbb{E}_{\mu,v}\left[\frac{(x-\mu)^2}{v^2}\right] = \frac{1}{v}$$

$$I(\mu, v)_{11} = \mathbb{E}_{\mu,v}\left[\frac{(x-\mu)}{v}\left(\frac{1}{2v} - \frac{1}{2}\frac{(x-\mu)^2}{v^2}\right)\right] == \mathbb{E}_{\mu,v}\left[\frac{(x-\mu)}{2v^2} - \frac{1}{2}\frac{(x-\mu)^3}{v^3}\right] = 0$$

par symétrie, et

$$I(\mu, v)_{22} = \mathbb{E}_{\mu,v}\left[\left(-\frac{1}{2v} + \frac{1}{2}\frac{(x-\mu)^2}{v^2}\right)^2\right] = \frac{1}{4v^2} - \frac{1}{2v}\mathbb{E}_{\mu,v}\left[\frac{(x-\mu)^2}{v^2}\right] + \frac{1}{4}\mathbb{E}_{\mu,v}\left[\frac{(x-\mu)^4}{v^4}\right]$$

$$= \frac{1}{4v^2} - \frac{1}{2v^2} + \frac{3}{4v^2} = \frac{1}{2v^2}.$$

Donc

$$I(\mu, v) = \begin{pmatrix} \frac{1}{v} & 0 \\ 0 & \frac{1}{2v^2} \end{pmatrix}$$

est l'information de Fisher apportée par X sur (μ, v).

Soit X une variable aléatoire à valeurs dans $\mathcal{X}$. Par exemple $\mathcal{X} = \mathbb{R}$ si X est une v.a. réelle, $\mathcal{X} = \mathbb{R}^n$ si X est une v.a. vectorielle de dimension n.

Théorème 33. *(cas discret) On suppose*

H1. $\{x \in \mathcal{X} : p(x, \theta) > 0\}$ *ne dépends pas de* $\theta \in \Theta$ *;*

H2. *La fonction* $\theta \mapsto p(x, \theta)$ *est* $C^2(\Theta)$ *;*

H3. $\forall A \subseteq \mathcal{X}$

$$\frac{\partial}{\partial\theta_i}\sum_{x\in A} p(x,\theta) = \sum_{x\in A} \frac{\partial}{\partial\theta_i} p(x,\theta), \qquad i=1,\ldots,d;$$

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j}\sum_{x\in A} p(x,\theta) = \sum_{x\in A} \frac{\partial^2}{\partial\theta_i\partial\theta_j} p(x,\theta), \qquad i,j=1,\ldots,d\,.$$

Alors

$$\mathbb{E}[S(X,\theta)]=0 \qquad \text{et} \qquad I(\theta)_{ij} = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p(X,\theta)\right] = -\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta_i} S_j(X,\theta)\right].$$

Remarque 34. Dans le cas continu on remplace partout $p(x,\theta)$ par $f(x,\theta)$ et $\sum_{x\in A} g(x)$ par $\int_A g(x)\mathrm{d}x$.

Cas des familles exponentielles

Définition 35. *Un modèle $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ est dit appartenir à une famille exponentielle si existent un entier $r \geqslant 1$ et des fonctions*

$$c(\theta)>0, \quad h(x)>0, \quad \alpha_j(\theta), \Gamma_j(\theta) \text{ pour } j=1,\ldots,d,$$

telles que

$$p(x,\theta) = c(\theta)\, h(x) \exp\left(\sum_{i=1}^{r} \alpha_i(\theta)\Gamma_i(x)\right).$$

Exemple 36. (Bernoulli) $X \sim \mathrm{Ber}(p)$. $p \in]0,1[$.

$$p(x,\theta) = \theta^x(1-\theta)^{1-x} = e^{x\log\theta + (1-x)\log(1-\theta)} = e^{x\log(\theta/(1-\theta)) + \log(1-\theta)} \qquad \text{pour } x=0,1$$

et donc $r=1$, $c(\theta)=(1-\theta)$, $h(x)=1$, $\alpha_1(\theta)=\log(\theta/(1-\theta))$, $\Gamma_1(x)=x$.

Exemple 37. (Gaussienne) $X \sim \mathcal{N}(\mu,\sigma^2)$.

$$f(x,\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$c(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\mu^2}{2\sigma^2}}, \quad h(x)=1$$

$$\alpha_1(\theta) = \frac{\mu}{\sigma^2}, \qquad \Gamma_1(x)=x$$

$$\alpha_2(\theta) = -\frac{1}{2\sigma^2}, \qquad \Gamma_1(x)=x^2$$

Donc soit $\mathrm{Ber}(p)$ que $\mathcal{N}(\mu,\sigma^2)$ appartiennent à des familles exponentielles.

Exemple 38. $X \sim \mathcal{G}(a,b)$. $(a,b)=\theta \in \Theta =]1,+\infty[\times\mathbb{R}_+^*$

$$f(x,a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} \mathbb{I}_{x>0}$$

$h(x) = \mathbb{I}_{x>0}$, $c(a,b) = b^a/\Gamma(a)$, $\Gamma_1(x) = \log x$, $\Gamma_2(x) = -x$, $\alpha_1(a,b) = a-1$, $\alpha_2(a,b) = b$.

Théorème 39. *Soit $\mathcal{P}$ un modèle paramétrique appartenant à une famille exponentielle. Si les fonctions $\alpha_i(\theta)$, $i = 1, \ldots, d$ sont $C^2(\Theta)$ alors les hypothèses **H1**, **H2**, **H3** sont vérifiées.*

Exemple 40. (Bernoulli, suite de l'exemple 36) $r = 1$, $\alpha_1(p) = \log(p/(1-p))$ est $C^2(]0,1[)$ donc **H1**, **H2**, **H3** sont vérifiées.

$$S(x,p) = \frac{\mathrm{d}}{\mathrm{d}p}\log p(x,p) = \frac{x}{p} - \frac{1-x}{1-p}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}p^2}\log p(x,p) = -\frac{x}{p^2} - \frac{1-x}{(1-p)^2}$$

Par le théorème 33 on a que

$$0 = \mathbb{E}_p[S(X,p)]$$

et que

$$I(p) = \mathbb{E}_p\left[\frac{x}{p^2} + \frac{1-x}{(1-p)^2}\right] = \frac{p}{p^2} + \frac{1-p}{(1-p)^2} = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}$$

pour tout $p \in]0,1[$.

Théorème 41. *(Borne de Cramer-Rao) Ici $d = 1$. Soit et $g\colon \Theta \to \mathbb{R}$ une fonction $C^1(\Theta)$ et $T_n(\boldsymbol{X})$ un estimateur non biaisé de $g(\theta)$. Si $\mathcal{P}$ vérifie les hypothèses **H1**, **H2**, **H3** et en plus on a que*

i. $0 < I(\theta) < +\infty$.

ii.

$$g'(\theta) = \begin{cases} \sum_{\boldsymbol{x}\in\mathcal{X}^n} T(\boldsymbol{x})\,\frac{\partial}{\partial\theta}p(\boldsymbol{x},\theta) & \text{dans le cas discret ;} \\ \int T(\boldsymbol{x})\,\frac{\partial}{\partial\theta}p(\boldsymbol{x},\theta)\,d\boldsymbol{x} & \text{dans le cas continu .} \end{cases}$$

Alors

$$\mathrm{Var}_\theta(T_n(\boldsymbol{X})) \geqslant \frac{(g'(\theta))^2}{n\,I(\theta)}\,.$$

Remarque 42. Si $\mathrm{Var}_\theta(T_n(\boldsymbol{X})) = \frac{(g'(\theta))^2}{n\,I(\theta)}$ alors on dit que l'estimateur T_n est (uniformément) efficace.

Exemple 43. Soit $X \sim \mathcal{N}(\mu, 1)$, $\theta = \mu \in \Theta = \mathbb{R}$. Le modèle appartient à une famille exponentielle avec $r = 1$ et $\alpha_1(x) = x$. Donc α_1 est $C^2(\Theta)$ et les hypothèses **H1**, **H2**, **H3** sont vérifiées et donc

$$I(\mu) = -\mathbb{E}\frac{d^2}{d\mu^2}\log f(X,\mu) = 1$$

car

$$\frac{d^2}{d\mu^2}\log f(x,\mu) = \frac{d^2}{d\mu^2}\left[-\log(2\pi) - \frac{(x-\mu)^2}{2}\right] = -1\,.$$

On considère l'estimateur $T_n(\boldsymbol{X}) = \overline{X}_n$ de μ. On a que $g(\mu) = \mu$ est $C^1(\mathbb{R})$, $g'(\mu) = 1$ et que

$$\int T(\boldsymbol{x}) \frac{\partial}{\partial \mu} p(\boldsymbol{x}, \mu) d\boldsymbol{x} = \frac{1}{n} \sum_{i=1}^{n} \int x_i \frac{\partial}{\partial \mu} e^{-\frac{1}{2}\sum_{j=1}^{n}(x_j-\mu)^2} \frac{d\boldsymbol{x}}{(2\pi)^{n/2}}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{n} \int x_i (x_k - \mu) e^{-\frac{1}{2}\sum_{j=1}^{n}(x_j-\mu)^2} \frac{d\boldsymbol{x}}{(2\pi)^{n/2}}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \int x_i (x_i - \mu) e^{-\frac{1}{2}\sum_{j=1}^{n}(x_j-\mu)^2} \frac{d\boldsymbol{x}}{(2\pi)^{n/2}}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \int (x_i - \mu)^2 e^{-\frac{1}{2}\sum_{j=1}^{n}(x_j-\mu)^2} \frac{d\boldsymbol{x}}{(2\pi)^{n/2}} = 1.$$

Toutes les hypothèses du Théorème 41 sont satisfaites donc la borne de Cramer-Rao est valable

$$\mathrm{Var}(T_n) \geqslant \frac{g'(\mu)^2}{n I(\mu)} = \frac{1}{n}.$$

D'autre part on a aussi par un calcul direct que $\mathrm{Var}(T_n) = 1$ donc l'estimateur $T_n(\boldsymbol{X})$ est uniformément efficace.

Théorème 44. *Dans les mêmes conditions du Théorème 41 on a que l'estimateur $T_n(\boldsymbol{X})$ est uniformément efficace ssi existe une fonction déterministe $a_n(\theta)$ telle que*

$$\sum_{i=1}^{n} S(X_i, \theta) = a_n(\theta)(T_n(\boldsymbol{X}) - g(\theta)).$$

Exemple 45. $X \sim \mathcal{N}(\mu, 1)$, $T_n(\boldsymbol{X}) = \overline{X}_n$, $g(\mu) = \mu$.

$$S(x, \mu) = \frac{d}{d\mu} \log f(x, \mu) = x - \mu.$$

$$\sum_{i=1}^{n} S(X_i, \mu) = \sum_{i=1}^{n} (X_i - \mu) = n(\overline{X}_n - \mu) = n(T_n(\boldsymbol{X}) - g(\mu))$$

et en prenant $a_n(\mu) = n$ le théorème précédent est valable. Il n'est donc pas surprenante que l'estimateur T_n soit efficace.

Lien entre l'information de Fisher et la loi asymptotique de l'EMV

On suppose ici $d = 1$.

Théorème 46. *(cas discret) Soit $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ un modèle paramétrique vérifiant les hypothèses **H1**, **H2**, **H3** tel que $0 < I(\theta) < +\infty$. On suppose que l'EMV $\hat{\theta}_n$ existe et est unique. Soit θ_0 la vraie valeur du paramètre θ inconnu. Si existe $\delta > 0$ et $k(x) \geqslant 0$ tels que*

i. $\left|\frac{d^2}{d\theta^2} \log p(x, \theta)\right| \leqslant k(x) \quad \text{pour tout } \theta \in [\theta_0 - \delta, \theta_0 + \delta]$

ii. $\mathbb{E}_{\theta_0}[k(X)] < +\infty$

alors

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1/I(\theta_0)).$$

Remarque 47. Dans le cas continu remplacer $p(x, \theta)$ par $f(x, \theta)$ dans l'énoncé précédente.

Exemple 48. Soit $X \sim \mathcal{E}(\lambda)$, $\lambda > 0$. $\Theta =]0, +\infty[$.

$$f(x, \lambda) = \lambda e^{-\lambda x} \mathbb{I}_{x>0}$$

$$r = 1, h(x) = \mathbb{I}_{x>0}, c(\lambda) = \lambda, \alpha_1(\lambda) = -\lambda \in C^2(\Theta), \Gamma_1(x) = x$$

$\Rightarrow$ le modèle satisfait **H1**, **H2**, **H3**.

$$S(x, \lambda) = \frac{1}{\lambda} - x$$

$$I(\lambda) = \mathbb{E}_\lambda[(\lambda^{-1} - X)^2] = \lambda^{-2} \in]0, +\infty[.$$

L'EMV de λ est $\hat{\lambda}_n = 1/\overline{X}_n$ (existe et est unique).

Soit $\delta > 0$ assez petit tel que $]\lambda_0 - \delta, \lambda_0 + \delta[\subset]0, +\infty[$. ($\lambda_0$ est le vrai valeur inconnu de λ).

$$\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2} \log f(x, \lambda) = -\frac{1}{\lambda^2}$$

$$\left| \frac{\mathrm{d}^2}{\mathrm{d}\lambda^2} \log f(x, \lambda) \right| = \frac{1}{\lambda^2} \leqslant \frac{1}{(\lambda_0 - \delta)^2} = k(x)$$

On a que $k(x) \geqslant 0$ et que $k(X)$ est intégrable. D'après le théorème précédente on obtient que

$$\sqrt{n}\big((\overline{X}_n)^{-1} - \lambda_0\big) = \sqrt{n}(\hat{\lambda}_n - \lambda_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \lambda_0^2).$$

Directement on aurait pu raisonner comme suive: par le TCL on a

$$\sqrt{n}(\overline{X}_n - \lambda_0^{-1}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \lambda_0^{-2})$$

On pose $g(x) = 1/x$ définie sur $]0, +\infty[$. La fonction g est C^1 dans un voisinage de $\lambda_0 > 0$. D'après la δ-méthode

$$\sqrt{n}(g(\overline{X}_n) - g(\lambda_0^{-1})) \xrightarrow{\mathcal{L}} \mathcal{N}(0, (g'(1/\lambda_0))^2 \lambda_0^{-2})$$

$g'(x) = -1/x^2$

$$(g'(1/\lambda_0))^2 \lambda_0^{-2} = \lambda_0^2$$

et on retrouve que

$$\sqrt{n}\big((\overline{X}_n)^{-1} - \lambda_0\big) = \sqrt{n}(g(\overline{X}_n) - g(\lambda_0^{-1})) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \lambda_0^2).$$