Apprentissage statistique

Chapitre 2 : Estimation ponctuelle

Lucie Le Briquer

15 janvier 2019

Table des matières

1 Définitions

Problème de statistique : on a $y_1, \dots, y_n$ des données. On suppose que ce sont des réalisations d'un modèle statistique $(\mathbb{Z}, \mathcal{Z}, \mathcal{P})$ associées à $Z = (Y_1, \dots, Y_n)$. $Y_1, \dots Y_n$ des variables aléatoires sous $\mathbb{P}^*$. On voudrait $\bar{\mathbb{P}} \in \mathcal{P}$ tel que $\bar{\mathbb{P}}$ est proche de $\mathbb{P}^*$ "en un certain sens".

On va supposer dans tout ce chapitre que l'on est dans le cadre paramétrique :

$$\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\} \quad \text{avec } \Theta \subset \mathbb{R}^d$$

Définition 1 (estimateur)

Soit $(\mathbb{Z}, \mathcal{Z}, \mathcal{P})$. Soit $\theta \mapsto g(\theta)$ tel que g est mesurable de Θ dans $\mathbb{R}^d$. On appelle estimateur de $\theta \mapsto g(\theta)$ toute statistique T de $\mathbb{Z}$ dans $\mathbb{R}^q$.

Définition 2 (erreur en moyenne quadratique)

Soit $(\mathbb{Z}, \mathcal{Z}, \mathcal{P})$ un modèle statistique et T un estimateur de g. On définit l'erreur en moyenne quadratique de T comme :

$$\mathrm{MSE}_\theta(T) = \mathbb{E}_\theta \left[\|T(Z) - g(\theta)\|^2\right]$$

MSE signifie *mean square error*.

Définition 3 (non biaisé)

On dit que T est non biaisé si $\forall \theta \in \Theta$, $\mathbb{E}_\theta[T(Z)] = g(\theta)$.

2 Méthode des moments

$(X_1, \dots, X_n)$ un n-échantillon associé à $(\mathbb{X}, \mathcal{X}, \{\mathbb{Q}_\theta : \theta \in \Theta\})$. Il est associé au modèle $(\mathbb{Z}, \mathcal{Z}, \{\mathbb{P}_\theta : \theta \in \Theta\})$ où $\mathbb{Z} = \mathbb{X}^b$, $\mathcal{Z} = \mathcal{X}^{\otimes n}$ et $\mathbb{P}_\theta = \mathbb{Q}_\theta^{\otimes n}$.

On a supposé qe l'o est dans un modèle paramétrique $\Theta \subset \mathbb{R}^d$. Supposons que l'on veuille estimer $\theta \mapsto \theta$. "Moralement", à partir de $X_1, \dots, X_n$ i.i.d. suivant $\mathbb{P}_{\theta^*}$ comment estimer θ^* ?

Supposons que l'on ait une fonction bijective $F \colon \Theta \longrightarrow U \subset \mathbb{R}^d$ et $F(\theta^*)$. Pour estimer θ^* on aurait juste à résoudre $F(\theta) = F(\theta^*)$ (1).

Dans la méthode des moments, l'idée est de prendre une famille de statistiques :

$$\{T_i \colon \mathbb{X} \longrightarrow \mathbb{R}\}_{i=1,\dots,d}$$

et de poser $F_i(\theta) = \mathbb{E}_\theta[T_i(X_1)]$.

Remarque. En pratique, on obtient donc F mais il faut vérifier que F est bijective.

Le problème est que l'on a pas accès à $F(\theta^*)$ mais seulement à des échantillons $X_1, \dots, X_n$ i.i.d. suivant $\mathbb{Q}_{\theta^*}$. L'idée est donc d'approcher $F_i(\theta^*)$ par la moyenne empirique :

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

L'équation (1) avec l'approximation suggérée donne les équations suivantes à résoudre en θ :

$$\frac{1}{n}\sum_{j=1}^{n} T_i(X_j) = F_i(\theta) \quad \text{pour } i = 1, \dots, d \quad (M)$$

On aboutit à un système à d équations et d inconnues. S'il admet une solution $\hat{\theta}_n$, on appelle $\hat{\theta}_n$ l'estimateur des moments.

Exemple. (loi exponentielle)
On considère le modèle $(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+), \{\mathcal{E}(\theta) : \theta \in \mathbb{R}_+^*\})$ et les deux statistiques suivantes :

$$T^{(1)}(x) = x \qquad T^{(2)}(x) = x^2$$

On calcule $F^{(1)}$ et $F^{(2)}$ associées à $T^{(1)}$ et $T^{(2)}$. $F^{(1)}(\theta) = \mathbb{E}_\theta[X_1] = \frac{1}{\theta}$ et $F^{(2)}(\theta) = \mathbb{E}_\theta[X_1^2] = \frac{2}{\theta^2}$.
Pour $F^{(1)}$ et $T^{(1)}$, (M) devient :

$$\frac{1}{n}\sum_{j=1}^{n} X_j = \frac{1}{\theta}$$

donc $\hat{\theta}_n^{(1)}$ l'estimateur des moments associé à $F^{(1)}$ et $T^{(1)}$ est :

$$\hat{\theta}_n^{(1)} = \frac{n}{\sum_{j=1}^{n} X_j}$$

Pour $F^{(2)}$ et $T^{(2)}$, (M) devient :

$$\frac{1}{n}\sum_{j=1}^{n} X_j^2 = \frac{2}{\theta^2}$$

donc $\hat{\theta}_n^{(2)}$ l'estimateur des moments associé à $F^{(2)}$ et $T^{(2)}$ est :

$$\hat{\theta}_n^{(2)} = \left(\frac{2n}{\sum_{j=1}^{n} X_j^2}\right)^{1/2}$$

Exemple. (modèle de Cauchy)
On considère le modèle $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{\text{Cauchy}(\theta) : \theta \in \mathbb{R}\})$. Où $\text{Cauchy}(\theta) \ll \text{Leb}$ et :

$$\frac{d\text{Cauchy}(\theta)}{d\text{Leb}} = \frac{1}{\pi(1 + (x-\theta)^2)}$$

On ne peut pas utiliser des statistiques aussi simples que précédemment puisque les moyennes ne sont pas définies. On propose d'utiliser $T(x) = \text{sgn}(n)$ (avec la convention $\text{sgn}(0) = 1$).
On peut montrer que :

$$\mathbb{E}_\theta[T(X_1)] = \frac{2}{\pi}\arctan(\theta)$$

L'équation (M) devient :

$$\frac{1}{n}\sum_{j=1}^{n} \text{sgn}(X_j) = \frac{2}{\pi}\arctan(\theta)$$

Donc,

$$\hat{\theta}_n = \tan\left(\frac{\pi}{2n}\sum_{j=1}^{n} \text{sgn}(X_j)\right)$$

3 $Z-$estimateurs

On rappelle l'équation des moments :

$$\frac{1}{n}\sum_{j=1}^{n} T_i(X_j) = F_i(\theta) \quad \text{pour } i = 1, \ldots, d \quad (M)$$

Si on introduit les fonctions $\psi_i^{(m)}(x, \theta) = T_i(x) - F_i(\theta)$, (M) devient :

$$\frac{1}{n}\sum_{j=1}^{n} \psi_i^{(m)}(X_j, \theta) = 0 \quad \text{pour } i = 1, \ldots, d$$

Définition 4 ($Z-$estimateur)

Soit $(X_1, \ldots, X_n)$ un $n-$échantillon associé à $\mathcal{P}$. Soit :

$$\{\psi_i \colon \mathbb{X} \times \Theta \longrightarrow \mathbb{R}\}_{i=1,\ldots,d}$$

que l'on suppose mesurables. Soit alors $\psi \colon \mathbb{X} \times \Theta \longrightarrow \mathbb{R}^d$ définie par $\psi = (\psi_1, \ldots, \psi_d)$.
On appelle $Z-$estimateur $\hat{\theta}_n$ toute solution de l'équation :

$$\frac{1}{n}\sum_{j=1}^{n} \psi(X_j, \theta) = 0$$

Exemples.

1. $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{\mathbb{P}_\theta : \theta \in \mathbb{R}\})$, on suppose que $\forall \theta \in \mathbb{R}$, $\mathbb{P}_\theta$ a pour fonction de répartition :

   $$\mathbb{P}_\theta(]-\infty, x]) = F_\theta(x) = F(x - \theta)$$

   où F est une fonction de répartition fixée supposée symétrique ($F(x) = 1 - F(-x)$) et que :

   $$\int_{\mathbb{R}} x dF(x) = 0 \qquad \int_{\mathbb{R}} x^2 dF(x) < +\infty$$

   On considère $\psi(x, \theta) = x - \theta$. On doit donc résoudre :

   $$\frac{1}{n}\sum_{j=1}^{n}(X_j - \theta) = 0 \qquad \hat{\theta}_n = \frac{1}{n}\sum_{j=1}^{n} X_j$$

2. On considère $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{\text{Cauchy}(\theta), \theta \in \mathbb{R}\})$ et la fonction :

   $$\psi(x, \theta) = \text{sgn}(x - \theta)$$

   Trouver le $Z-$estimateur associé à ψ.

   À faire en exercice.

4 Maximum de vraisemblance (*Maximum Likelihood*)

Exemple. (sondage)
$(X_1, \dots, X_n)$ $n-$échantillon du modèle $\{\text{Ber}(\theta) : \theta \in]0,1[\}$. La loi de $(X_1, \dots, X_n)$ admet une densité p_θ par rapport à la mesure de comptage sur $\{0,1\}^n$

$$p_\theta(x_1, \dots, x_n) = \prod_{i=1}^{n} q_\theta(x_i) = \prod_{i=1}^{n} \{\theta^{x_i}(1-\theta)^{1-x_i}\}$$

La vraisemblance et la log-vraisemblance sont :

$$L_n(\theta) = p_\theta(X_1, \dots, X_n) \qquad l_n(\theta) = \log L_n(\theta)$$

On justifie dans la suite pourquoi L_n et l_n sont intéressantes.
Soit $X_1, \dots, X_n$ distribuées suivant $\mathbb{P}_{\theta^*}$.

$$\begin{aligned} \varphi_{\theta^*}(\theta) &= \mathbb{E}_{\theta^*}\left[\log q_\theta(X_1)\right] = (1-\theta^*)\log q_\theta(0) + \theta^* \log q_\theta(1) \\ &= (1-\theta^*)\log(1-\theta) + \theta^* \log(\theta) \end{aligned}$$

On montre que $\theta \mapsto \varphi_{\theta^*}(\theta)$ est concave et admet un unique maximum : θ^*. En effet :

$$\varphi'_{\theta^*}(\theta) = \frac{\theta^*}{\theta} - \frac{1-\theta^*}{1-\theta} \qquad \varphi''_{\theta^*}(\theta) = -\frac{\theta^*}{\theta^2} - \frac{1-\theta^*}{(1-\theta)^2} < 0$$

Si on avait accès à φ_{θ^*} on aurait juste à la maximiser. Mais φ_{θ^*} est inconnue car la loi de $X_1, \dots, X_n$ ($\mathbb{P}_{\theta^*}$) l'est. On la remplace cependant par la quantité empirique associée :

$$\frac{1}{n}\sum_{j=1}^{n} \log q_\theta(X_j) = \frac{l_n(\theta)}{n}$$

Maintenant au lieu de maximiser φ_{θ^*}, on maximise $\theta \mapsto l_n(\theta)$. Par définition :

$$l_n(\theta) = \sum_{j=1}^{n} \log q_\theta(X_j) = \sum_{j=1}^{n} X_j \log(\theta) + (1-X_j)\log(1-\theta)$$

On trouve $\hat{\theta}_n = \frac{1}{n}\sum_{j=1}^{n} X_j$.

Définition 5 (vraisemblance, log-vraisemblance)

Soit $(X_1, \dots, X_n)$ un $n-$échantillon de $\{\mathbb{Q}_\theta : \theta \in \Theta\}$. On suppose que $\{\mathbb{P}_\theta : \theta \in \Theta\}$ est dominé par μ. On note $\forall \theta \in \Theta$, $f_\theta = d\mathbb{P}_\theta / d\mu$. On définit la vraisemblance $L_n \colon \Theta \longrightarrow \mathbb{R}_+$ pour tout $\theta \in \Theta$ par :

$$L_n(\theta) = f_\theta(X_1, \dots, X_n)$$

On définit la log-vraisemblance $l_n \colon \Theta \longrightarrow [-\infty, +\infty[$ par $l_n(\theta) = \log L_n$.

Définition 6 (estimateur du maximum de vraissemblance, EMV)

Soit $(X_1, \dots, X_n)$ un $n-$échantillon de vraissemblance L_n. On appelle un estimateur du maximum de vraissemblance $\hat{\theta}_n$ toute estimateur qui satisfait :

$$\hat{\theta}_n \in \text{argmax}_\Theta L_n(\theta) = \text{argmax}_\Theta l_n(\theta)$$

Définition 7 (racine de l'équation de vraissemblance)

Soit $(X_1, \ldots, X_n)$ un $n-$échantillon. On suppose que, pour tout θ^*, $\theta \mapsto l_n(\theta)$ est $\mathcal{C}^1$ $\mathbb{P}_{\theta^*}-$presque sûrement. On appelle racine de l'équation de vraissemblance $\hat{\theta}_n$ tout estimateur de θ qui vérifie :

$$\nabla_\theta l_n(\hat{\theta}_n) = 0$$

Exemples.

1. (modèle gaussien)

   $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{N(m, \sigma^2), m \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+^*\})$ modèle dominé par Leb.

   $$q_\theta(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - m)^2}{2\sigma^2}\right)$$

   $$\frac{1}{n} l_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log q_\theta(X_i) = \frac{1}{n} \sum_{i=1}^{n} \frac{(X_i - m)^2}{2\sigma^2} - \frac{\log(2\pi\sigma^2)}{2}$$

   On peut montrer que $\theta \mapsto \frac{1}{n} l_n(\theta))$ admet un unique maximum :

   $$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \qquad \hat{\sigma}_n = \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu}_n)^2$$

2. L'EMV n'existe pas forcément. $(\mathbb{R}, \mathcal{B}(\mathbb{R}), (\mathbb{Q}_\theta)_\Theta)$, $\mathbb{Q}_\theta \ll \text{Leb}\ \forall\theta$

   $$\frac{d\mathbb{Q}_\theta}{d\text{Leb}}(x) = \frac{e^{-|x-\theta|/2}}{2\sqrt{2\pi|x-\theta|}}$$

   On considère la vraissemblance associée à $(X_1, \ldots, X_n)$:

   $$L_n(\theta) = \prod_{i=1}^{n} \frac{e^{-|X_i-\theta|/2}}{2\sqrt{2\pi|X_i-\theta|}}$$

   On a que $\lim_{\theta \to X_i} L_n(\theta) = +\infty$.

5 $M-$estimateurs

On a définit l'EMV comme maximum de $l_n(\theta)$ ou $L_n(\theta)$ qui sont utilisées comme approximation à la place de la fonction $\mathbb{E}_{\theta^*}[\log q_\theta(X_1)]$.

$$\hat{\theta}_n \in \text{argmax} l_n(\theta)$$

L'idée tout d'abord est de considérer d'autres fonctions $m\colon \mathbb{X} \times \Theta \longrightarrow \mathbb{R}$ et on définit alors :

$$M(\theta, \theta^*) = \mathbb{E}_{\theta^*}[m(X_1, \theta)]$$

On suppose (ou on choisit en fait) m, et donc M, tel que $\forall \theta^*$, $\theta \mapsto M(\theta, \theta^*)$ admet un unique maximum atteint en θ^*. On a pas accès à θ^*, donc pas non plus à M mais on remplace M par :

$$M_n(\theta) = \frac{1}{n} \sum_{j=1}^{n} m(X_j, \theta)$$

Définition 8

Soit $(X_1, \ldots, X_n)$ un $n-$échantillon du modèle $\{\mathbb{Q}_\theta : \theta \in \Theta\}$. Soit $m\colon \mathbb{X} \times \Theta \longrightarrow \mathbb{R} \cup \{-\infty\}$ vérifiant :

$$\mathbb{E}_{\theta_1}[|m|(X_1, \theta_2)] < +\infty \quad \forall \theta_1, \theta_2 \in \Theta$$

Un $M-$estimateur associé à m est tout estimeur $\hat{\theta}_n$ vérifiant :

$$\hat{\theta}_n \in \text{argmax}_{\theta \in \Theta} M_n(\theta)$$

où $M_n(\theta) = \frac{1}{n} \sum_{j=1} m(X_j, \theta)$.

Lien avec les $Z-$estimateurs. Si $\forall x \in \mathbb{X}$, $\theta \mapsto m(x, \theta)$ est $\mathcal{C}^1$ sur Θ, alors $\hat{\theta}_n$ vérifie :

$$\nabla_\theta M_n(\hat{\theta}_n) = 0 \quad \Leftrightarrow \quad \frac{1}{n} \sum_{j=1}^{n} \nabla_\theta m(X_j, \hat{\theta}_n) = 0$$

et donc $\hat{\theta}_n$ est un $Z-$estimateur pour $\psi = \nabla_\theta m(.,.)$.

Remarque. Tout $M-$estimateur n'est pas réciproquement un $Z-$estimateur et réciproquement.