Estimation paramétrique

Retour au plan du cours

Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace probabilisé et X une v.a. de $(\Omega, \mathcal{A})$ dans $(E, \mathcal{E})$. La donnée d'un modèle statistique c'est la donnée d'une famille de probabilités sur $(E, \mathcal{E})$, $\{P_\theta,\ \theta \in \Theta\}$. Le modèle étant donné, on suppose alors que la loi de X appartient au modèle $\{P_\theta,\ \theta \in \Theta\}$. Par exemple dans le modèle de Bernoulli, $X = (X_1, \ldots, X_n)$ où les X_i sont i.i.d. de loi de Bernoulli de paramètre $\theta \in]0,1[$. $E = \{0,1\}^n$, $\mathcal{E} = \mathcal{P}(E)$, $\Theta =]0,1[$ et $P_\theta = ((1-\theta)\delta_0 + \theta\delta_1)^{\otimes n}$.

1 Premières définitions

DÉFINITION 1. — *On dit que le modèle $\{P_\theta,\ \theta \in \Theta\}$ est identifiable si l'application*

$$\begin{array}{ccc} \Theta & \to & \{P_\theta,\ \theta \in \Theta\} \\ \theta & \mapsto & P_\theta \end{array}$$

est injective.

DÉFINITION 2. — *Soit $g : \Theta \mapsto \mathbb{R}^k$. On appelle estimateur de $g(\theta)$ au vu de l'observation X, toute application $T : \Omega \mapsto \mathbb{R}^k$ de la forme $T = h(X)$ où $h : E \mapsto \mathbb{R}^k$ mesurable.*

Un estimateur ne doit pas dépendre de la quantité $g(\theta)$ que l'on cherche à estimer. On introduit les proprières suivantes d'un estimateur :

DÉFINITION 3. — *T est un estimateur sans biais de $g(\theta)$ si pour tout $\theta \in \Theta$, $\mathbb{E}_\theta[T] = g(\theta)$. Dans le cas contraire, on dit que l'estimateur T est biaisé et on appelle biais la quantité $\mathbb{E}_\theta(T) - g(\theta)$.*

Généralement X est un vecteur $(X_1, \ldots, X_n)$ d'observations (n étant le nombre d'entre elles). Un exemple important est le cas où $X_1, \ldots, X_n$ forme un *n-échantillon* c'est à dire lorsque que $X_1, \ldots, X_n$ sont i.i.d. On peut alors regarder des propriétés asymptotiques de l'estimateur, c'est à dire en faisant tendre le nombre d'observations n vers $+\infty$. Dans ce cas, il est naturel de noter $T = T_n$ comme dépendant de n. On a alors la définition suivante :

DÉFINITION 4. — *T_n est un estimateur consistant de $g(\theta)$ si pour tout $\theta \in \Theta$, T_n converge en probabilité vers $g(\theta)$ sous P_θ lorsque $n \to \infty$.*

On définit le risque quadratique de l'estimateur dans le cas où $g(\theta) \in \mathbb{R}$.

DÉFINITION 5. — *Soit T_n est un estimateur de $g(\theta)$. Le risque quadratique de T_n est défini par*

$$R(T_n, g(\theta)) = \mathbb{E}_\theta[(T_n - g(\theta))^2].$$

Remarque. — Le risque quadratique est la somme de la variance et du carré du biais de l'estimateur.

L'inégalité de Cramer-Rao et la définition de l'information de Fisher ont été vues en année 3 et ne sont pas rappelées ici.

2 Estimation par la méthode des moments

Dans cette section, X est le vecteur formé par un n-échantillon $X_1, \ldots, X_n$. Les X_i sont à valeurs dans un ensemble $\mathcal{X}$.

Soit $f = (f_1, \ldots, f_k)$ une application de $\mathcal{X}$ dans $\mathbb{R}^k$ telle que l'application

$$\Phi : \begin{array}{ccc} \Theta & \to & \mathbb{R}^k \\ \theta & \mapsto & \mathbb{E}_\theta[f(X_1)] \end{array}$$

soit **injective**. On définit l'estimateur $\hat{\theta}_n$ comme la solution dans Θ (quand elle existe) de l'équation

$$\Phi(\theta) = \frac{1}{n}\sum_{i=1}^n f(X_i). \tag{1}$$

Souvent, lorsque $\mathcal{X} \subset \mathbb{R}$, la fonction on prend $f_i(x) = x^i$ et Φ correspond donc au $i$ème moment de la variables X_1 sous $\mathbb{P}_\theta$. Ce choix justifie le nom donné à la méthode. Voici quelques exemples d'estimateurs bâtis sur par cette méthode.

2.1 Loi exponentielle

Ici $k = 1$, $Q_\theta = \mathcal{E}(\theta)$ pour $\theta \in \mathbb{R}_+^*$. Comme pour tout θ, $\mathbb{E}_\theta[X_1] = 1/\theta$ on prend $\Phi(\theta) = 1/\theta$ et $f = Id : \mathbb{R}_+ \to \mathbb{R}_+$. L'estimateur obtenu par la méthode des moments est

$$\hat{\theta}_n = \frac{1}{\overline{X}_n} \quad \text{où} \quad \overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i.$$

Par continuité de l'application $x \to 1/x$, $\hat{\theta}_n$ est un estimateur consistant de θ. Remarquons que $\overline{X}_n > 0$ p.s. ce qui justifie l'égalité ci-dessus.

2.2 Loi uniforme

Ici $k = 1$, Q_θ est la loi uniforme sur $[0, \theta]$ avec $\theta > 0$. On a que pour tout θ, $\mathbb{E}_\theta[X_1] = \theta/2$, on peut donc prendre par exemple $\Phi(\theta) = \theta/2$ et $f = Id : \mathbb{R} \to \mathbb{R}$. L'estimateur obtenu par la méthode des moments est alors $\hat{\theta}_n = 2\overline{X}_n$. Cet estimateur est sans bias et consistant.

2.3 Loi gaussienne

Ici $k = 2$, on prend $\theta = (m, \tau) \in \mathbb{R} \times \mathbb{R}_+^*$, $Q_\theta = \mathcal{N}(m, \tau)$. Pour tout $\theta = (m, \tau)$, $\mathbb{E}_\theta[X_1] = m$ et $\mathbb{E}_\theta[X_1^2] = m^2 + \tau$. On peut donc prendre, par exemple, $f_1(x) = x$ et $f_2(x) = x^2$ ce qui donne $\Phi(m, \tau) = (m, m^2 + \tau)$. L'estimateur obtenus par la méthode des moments vérifie

$$\hat{m}_n = \overline{X}_n \quad \text{et} \quad \hat{m}_n^2 + \hat{\tau}_n = \frac{1}{n}\sum_{i=1}^n X_i^2.$$

c'est a dire

$$\hat{\theta}_n = \left(\overline{X}_n, \frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2 \right).$$

L'estimateur est consistant mais l'estimateur de la variance est biaisé.

2.4 Propriétés asymptotiques

Notons $\Phi(\theta) = \mathbb{E}_\theta(\frac{1}{n}\sum_{i=1}^n f(X_i))$. Supposons que $X_1, \ldots, X_n$ sont i.i.d. de loi $\mathbb{P}_{\theta_0}$. Les résultats de consistance précédents étaient obtenus grâce au fait que d'une part,

$$\frac{1}{n}\sum_{i=1}^n f(X_i) \xrightarrow{\text{p.s.}} \Phi(\theta_0),$$

et donc, comme Φ^{-1} existe et est continue au voisinage de $\Phi(\theta_0)$, on en déduit que $\hat{\theta}_n$ existe et vérifie

$$\hat{\theta}_n \xrightarrow{\text{p.s.}} \Phi^{-1} \circ \Phi(\theta_0) = \theta_0.$$

Mais que peut-on dire de la distance de $\hat{\theta}_n$ à θ_0 ? Sous l'hypothèse que $\mathbb{E}_{\theta_0}[\|f\|^2] < +\infty$ on a grâce au théorème central limite que

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n f(X_i) - \Phi(\theta_0) \right) \xrightarrow{\text{Loi}} \mathcal{N}_k\left(0, \Gamma(\theta_0)\right),$$

où $\Gamma(\theta_0)$ la matrice covariance de $f(X_1)$ sous $\mathbb{P}_{\theta_0}$. Elle est définie pour $i, j \in \{1, \ldots, k\}$

$$\Gamma_{i,j}(\theta_0) = Cov_{\theta_0}[f_i(X_1)f_j(X_1)].$$

La Delta méthode (cf Proposition 16) va nous permettre de déduire un résultat similaire pour $\hat{\theta}_n$:

THÉORÈME 6. — *Supposons que Φ soit de classe $\mathcal{C}^1$ de $\mathring{\Theta}$ dans $\mathbb{R}^k$ et que $\theta_0 \in \mathring{\Theta}$, et que $D_{\theta_0}\Phi : \mathbb{R}^k \to \mathbb{R}^k$ soit inversible. Supposons de plus que $\mathbb{E}_{\theta_0}[\|f(X_1)\|^2] < +\infty$ et notons $\Gamma(\theta_0)$ la matrice covariance de $f(X_1)$ sous $\mathbb{P}_{\theta_0}$. Alors sous $\mathbb{P}_{\theta_0}$:*

- *$\hat{\theta}_n$ existe avec une probabilité tendant vers 1*
- *on a la convergence en loi*

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{\text{Loi}} \mathcal{N}\left(0, (D_{\theta_0}\Phi)^{-1}\Gamma(\theta_0)\left((D_{\theta_0}\Phi)^{-1}\right)'\right).$$

Démonstration. — Comme $D_{\theta_0}\Phi$ est inversible, $D_\theta\Phi$ reste inversible dans un voisinage de θ_0 et donc, d'après le théorème de l'inversion locale, Φ réalise un difféomorphisme d'un voisinage U de θ_0 dans V un voisinage de $\Phi(\theta_0)$. Par la loi des grands nombres, $\hat{Y}_n = n^{-1}\sum_{i=1}^n f(X_i)$ converge en probabilité (car p.s.) vers $\Phi(\theta_0)$ et donc $\hat{Y}_n$ appartient à V avec une probabilité tendant vers 1 quand n tend vers $+\infty$. Sur cet événement, l'équation (1) admet une unique solution $\hat{\theta}_n$ dans Θ (par injectivité de Φ) et cette solution vérifie $\hat{\theta}_n \in U$ et $\hat{\theta}_n = \Phi^{-1}(\hat{Y}_n)$ où Φ^{-1} est définie de V dans U. On a par ailleurs,

$$\begin{aligned}\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) &= \sqrt{n}\hat{\theta}_n \mathbb{1}_{\hat{Y}_n \notin V} + \sqrt{n}\left(\hat{\theta}_n \mathbb{1}_{\hat{Y}_n \in V} - \theta_0\right) \\ &= \sqrt{n}\hat{\theta}_n 1_{\hat{Y}_n \notin V} + \sqrt{n}[\tilde{\Phi}^{-1}\left(\hat{Y}_n\right) - \tilde{\Phi}^{-1}\left(\Phi(\theta_0)\right)] \quad (2)\end{aligned}$$

où $\tilde{\Phi}^{-1}(y) = \Phi^{-1}(y)\mathbb{1}_{y\in V}$. Or $\sqrt{n}\hat{\theta}_n \mathbb{1}_{\hat{Y}_n \notin V}$ converge vers 0 en probabilité car pour tout $\varepsilon > 0$,

$$\mathbb{P}_{\theta_0}[\mathbf{a}\sqrt{n}\hat{\theta}_n \mathbb{1}_{\hat{Y}_n \notin V} > \varepsilon] \leq \mathbb{P}_{\theta_0}[\hat{Y}_n \notin V] \xrightarrow[n\to+\infty]{} 0.$$

D'après le lemme de Slutsky, il suffit donc de montrer que le second terme du membre de droite de (2) converge en loi vers la limite annoncée. Or par théoreme centrale limite vectoriel

$$\sqrt{n}\left(\hat{Y}_n - \Phi(\theta_0)\right) \xrightarrow{\text{Loi}} \mathcal{N}(0, \Gamma(\theta_0)),$$

et on conclut en utilisant la Proposition 16.

∎

3 Estimation par maximum de vraisemblance

Soit $\{E, \mathcal{E}, \{P_\theta,\ \theta \in \Theta\}\}$ un modèle statistique, où $\Theta \subset \mathbb{R}^k$ (nous sommes dans un cadre paramétrique). On suppose qu'il existe une mesure σ-finie μ qui domine le modèle, c'est à dire que $\forall\theta \in \Theta$, P_θ admet une densité $p(\theta, .)$ par rapport à μ.

DÉFINITION 7. — *Soit X une observation. On appelle vraisemblance de X l'application*

$$\begin{array}{ccc}\Theta & \longrightarrow & \mathbb{R}_+ \\ \theta & \mapsto & p(\theta, X).\end{array}$$

On appelle estimateur du maximum de vraisemblance de θ, tout élément $\hat{\theta}$ de Θ maximisant la vraisemblance , c'est à dire vérifiant

$$\hat{\theta} = \arg\max_{\theta\in\Theta} p(\theta, X).$$

Remarque. — L'estimateur de maximum de vraisemblance n'existe pas toujours et n'est pas toujours unique.

Considérons le cas typique où $X = (X_1, \ldots, X_n)'$, les X_i formant un n-échantillon de loi Q_{θ_0} où Q_{θ_0} est une loi sur $\mathcal{X}$ de paramètre inconnu $\theta_0 \in \Theta \subset \mathbb{R}^k$. On suppose en outre que pour tout $\theta \in \Theta$, Q_θ est absolument continue par rapport à une mesure ν sur $\mathcal{X}$. Dans ce cas, en notant

$$q(\theta, x) = \frac{dQ_\theta}{d\nu}(x),$$

et en prenant $\mu = \nu^{\otimes n}$ on a que la vraisemblance s'écrit

$$p(\theta, X) = \prod_{i=1}^n q(\theta, X_i)$$

et donc

$$\hat{\theta}_n = \arg\max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n \log[q(\theta, X_i)],$$

avec la convention $\log(0) = -\infty$. Voyons quelques exemples.

3.1 Modèle de Bernoulli

Soit $Q_{\theta_0} = \mathcal{B}(\theta)$ avec $\theta \in [0, 1] = \Theta$, et ν la mesure de comptage sur $\mathbb{N}$. Pour tout $\theta \in]0, 1[$ et $x \in \{0, 1\}$

$$q(\theta, x) = \theta^x(1-\theta)^{1-x} = (1-\theta)\exp[x\log\left(\frac{\theta}{1-\theta}\right)]$$

et donc l'estimateur du maximum de vraisemblance doit maximiser dans $[0, 1]$

$$log\left(\theta^{S_n}(1-\theta)^{n-S_n}\right) = S_n\log\left(\frac{\theta}{1-\theta}\right) + n\log(1-\theta),$$

ce qui conduit à $\hat{\theta}_n = \bar{X}$.

3.2 Lois exponentielles

Soit $Q_\theta = \mathcal{E}(\theta)$ avec $\theta \in \theta =]0, +\infty[$, $\nu = \mathbb{1}_{\mathbb{R}_+} dx$. On a pour tout $x > 0$,

$$q(\theta, x) = \theta \exp -\theta x,$$

et le maximum de vraisemblance $\hat{\theta}_n$ maximisant

$$n \log(\theta) - \theta S_n,$$

est donné par $\hat{\theta} = \left(\bar{X}\right)^{-1}$.

3.3 Lois de Poisson

Soit $Q_\theta = \mathcal{P}(\theta)$ avec $\theta \in \Theta =]0, +\infty[$ et ν la mesure de comptage sur $\mathbb{N}$. On a pour tout $x \in \mathbb{N}$,

$$q(\theta, x) = \frac{\theta^x}{x!} e^{-\theta} = e^{-\theta} \exp[x \log(\theta) - \ln(x!)].$$

L'estimateur du maximum de vraisemblance maximisant

$$S_n \log(\theta) - n\theta,$$

est donné par $\hat{\theta}_n = \bar{X}$.

4 Estimateurs exhaustifs complets

DÉFINITION 8. — *On dit qu'un estimateur T est exhaustif (ou suffisant) si pour tout ensemble $B \in \mathcal{E}$, il existe une version de l'espérance conditionnelle $\mathbb{E}_\theta[\mathbb{1}_{X \in B} \mid T]$ qui ne dépend pas de θ.*

Exemple. — Soit $X_1, \ldots, X_n$ i.i.d. de loi de Bernoulli de paramètre θ. $S = X_1 + \ldots + X_n$ est une statistique exhaustive. La loi de $(X_1, \ldots, X_n)/S = s$ ne dépend pas de θ.

THÉORÈME 9. — **Théorème de factorisation**
Si la vraisemblance de l'observation X s'écrit sous la forme

$$f(X, \theta) = h(X) g_\theta(T(X)),$$

alors $T(X)$ est une statistique exhaustive.

Exemples
$X_1, \ldots, X_n$ i.i.d. de loi de Bernoulli de paramètre θ,
$X_1, \ldots, X_n$ i.i.d. de loi de normale $\mathcal{N}(\theta, 1)$.

PROPOSITION 10. — *Soit T une statistique exhaustif et V un estimateur de $g(\theta)$ tel que $\mathbb{E}_\theta(V^2) < +\infty$ $\forall \theta \in \Theta$. Soit $T^* = \mathbb{E}_\theta(V/T)$. T^* est indépendant de θ et pour tout $\theta \in \Theta$,*

$$\mathbb{E}_\theta((T^* - g(\theta))^2) \leq \mathbb{E}_\theta((V - g(\theta))^2),$$

ce qui signifie que le risque quadratique de T^* pour estimer $g(\theta)$ est inférieur ou égal à celui de V.

DÉFINITION 11. — *Soit T est un estimateur sans biais de $g(\theta)$. On dit que T est uniformément de variance minimale parmi les estimateurs sans biais (UMVB) si, pour tout estimateur sans biais U de $g(\theta)$, on a*

$$\forall \theta \in \Theta, R(T, g(\theta)) \leq R(U, g(\theta)),$$

où R désigne le risque quadratique.

Soit T un estimateur exhaustif de $g(\theta)$. On désire introduire une notion qui garantisse l'unicité d'un estimateur $\psi(T)$ qui a les deux propriétés suivantes :
1) $\psi(T)$ est sans biais.
2) Parmi les les estimateurs sans biais, il est de variance minimale.

S'il existe un unique estimateur sans biais de $g(\theta)$ fonction de T alors

$$\mathbb{E}_\theta(\psi_1(T)) = \mathbb{E}_\theta(\psi_2(T)) = g(\theta) \quad \forall \theta \in \Theta$$
$$\Longrightarrow \psi_1(T) = \psi_2(T) \quad P_\theta \; p.s. \; \forall \theta.$$

DÉFINITION 12. — *Une statistique T est complète si pour toute fonction ψ telle que*

$$\mathbb{E}_\theta(|\psi(T)|) < +\infty \quad \forall \theta \in \Theta,$$

on a

$$\mathbb{E}_\theta(\psi(T)) = 0 \quad \forall \theta \in \Theta \Longrightarrow \psi(T) = 0 \quad P_\theta \; p.s. \; \forall \theta.$$

Si T est une statistique complète et si $\psi_1(T)$ et $\psi_2(T)$ sont deux estimateurs sans biais de $g(\theta)$, alors $\psi_1(T) = \psi_2(T)$ p.s.
Exemple des modèles exponentiels :

DÉFINITION 13. — *S'il existe une mesure de référence μ telle que la loi P_θ admette une densité par rapport à μ de la forme*

$$f(\theta, x) = C(\theta) \exp(\langle T(x), Q(\theta) \rangle),$$

le modèle est dit exponentiel.

PROPOSITION 14. — *Dans un modèle exponentiel, $T(X)$ est une statistique exhaustive. De plus, si $Q(\Theta)$ contient un ouvert non vide de $\mathbb{R}^k$, alors $T(X)$ est complète.*

PROPOSITION 15. — *Supposons qu'il existe au moins un estimateur sans biais de $g(\theta)$ et que T soit exhaustive complète. Alors il existe un unique estimateur sans biais de $g(\theta)$ fonction de T, et cet estimateur est UMVB.*

Remarque. — Si l'objectif est de minimiser le risque quadratique, on a parfois intérêt à considérer des estimateurs qui sont biaisés.

5 Annexe : rappels sur les convergences

PROPOSITION 16. — *(Delta-Méthode) Soit X_n une suite de v.a. à valeurs dans $\mathbb{R}^k$ et $\theta \in \mathbb{R}^k$ tels que*

$$r_n (X_n - \theta) \xrightarrow{\text{Loi}} Y,$$

où r_n est une suite de réels positifs tendant vers $+\infty$. Soit f une application de $\mathbb{R}^k$ dans $\mathbb{R}^m$ differentiable en θ. On a alors que

$$f(X_n) \xrightarrow{\text{Prob}} f(\theta) \text{ et } r_n (f(X_n) - f(\theta)) \xrightarrow{\text{Loi}} D_\theta f[Y].$$

Remarque. — Ce résultat permet d'étendre à d'autres v.a. le théorème central limite. Classiquement, Y suit une loi gaussienne $\mathcal{N}(0, \Gamma)$ et dans ce cas

$$D_\theta f[Y] \sim \mathcal{N}\left(0, (D_\theta f) \Gamma (D_\theta f)'\right).$$

Remarquons que l'on ne suppose rien sur la régularité de f ailleurs qu'en θ. Notons aussi que l'on ne suppose pas $D_\theta f \neq 0$ et donc la limite peut-etre nulle éventuellement.