Rappels de statistiques mathématiques : cours 3

Guillaume Lecué

9 septembre 2015

Rappel des cours précédents

- outils : LFGN, TCL multi-dimensionnel, Lemme de Slutsky, méthode Delta
- estimateurs : fonction empirique, quantile empirique, estimateur plug-in, algorithme de Robbins-Monro
- Résultats asymptotiques de convergence p.s. et vitesses de convergence de ces estimateurs :

$$\widehat{F}_n, \quad \widehat{q}_{n,p}, \quad T(\widehat{F}_n) = h\left(\frac{1}{n}\sum_{i=1}^n g(X_i)\right), \quad x_n$$

⚠Jusqu'à maintenant, on n'a pas utilisé la notion de modèle statistique pour construire et étudier des méthodes d'estimation

Aujourd'hui

1. modèle dominé

2. Méthodes d'estimation dans les modèles
   - Méthode des moments
   - Z-estimation
   - M-estimation
   - Principe de maximum de vraisemblance

Rappels : expériences et modèle statistique (1/2)

Définition

Une expérience statistique $\mathcal{E}$ est un triplet

$$\mathcal{E} = \left(\mathfrak{Z}, \mathcal{Z}, \left\{\mathbb{P}_\theta, \theta \in \Theta\right\}\right),$$

avec

- $(\mathfrak{Z}, \mathcal{Z})$ *espace mesurable (souvent* $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$*),*
- $\{\mathbb{P}_\theta, \ \theta \in \Theta\}$ *famille de mesures de probabilités définies sur* $(\mathfrak{Z}, \mathcal{Z})$ *appelée modèle*

Question : Un modèle est une connaissance/intuition a priori sur les données. Comment tirer profit du modèle pour construire et étudier des estimateurs "plus efficaces" que les estimateurs sans modèle $\widehat{F}_n, \widehat{q}_{n,p}, \bar{x}_n, \ldots$?

Exemple : expérience et modèles statistique (2/2)

Probléme : un physicien observe la durée de vie d'atomes radioactifs qu'il décide de modéliser par des variables aléatoires $X_1, \dots, X_n$ i.i.d.. Il souhaite utiliser ces données pour estimer leur loi sous-jacente. Il peut choisir entre deux approches :

- "sans modèle" et estimer la fonction de répartition des X_i par $\widehat{F}_n$
- "avec modèle" : il sait que les durées de vie suivent une loi exponentielle $\in \{\mathcal{E}xp(\theta) : \theta > 0\}$. Dans ce cas, il suffit d'estimer θ par un estimateur $\widehat{\theta}_n$ et d'approcher la fonction de répartition des X_i par $F_{\widehat{\theta}_n}$ où

$$F_\theta(x) = \mathbb{P}[\mathcal{E}xp(\theta) \leq x] = \begin{cases} 0 & \text{si } x \leq 0 \\ 1 - \exp(-\theta x) & \text{sinon.} \end{cases}$$

Expériences dominées

- On fait une hypothèse minimale de « structure » sur le modèle statistique. But : ramener l'étude de la famille

$$\{\mathbb{P}_\theta,\ \theta \in \Theta\}$$

à l'étude d'une famille de fonctions

$$\{z \in \mathfrak{Z} \mapsto f(\theta, z) \in \mathbb{R}_+,\ \theta \in \Theta\}.$$

- Via la notion de domination : si μ, ν sont deux mesures (positives) σ-finies sur $\mathfrak{Z}$, alors μ domine ν (notée $\nu \ll \mu$) quand

$$\forall A \in \mathcal{Z}, \quad \mu\left[A\right] = 0 \Rightarrow \nu\left[A\right] = 0$$

Théorème de Radon-Nikodym

Théorème

Si $\nu \ll \mu$, il existe une fonction positive (μ-p.p.), appelée *densité de ν par rapport à μ*, *notée*

$$z \mapsto \frac{d\nu}{d\mu}(z),$$

définie $\mu-$p.p., $\mu-$ intégrable, telle que, pour tout $A \in \mathcal{Z}$,

$$\nu[A] = \int_A \frac{d\nu}{d\mu}(z)\mu(dz)$$

Expérience dominée

Définition

Une expérience statistique $\mathcal{E} = \left(\mathfrak{Z}, \mathcal{Z}, \left\{ \mathbb{P}_\theta, \theta \in \Theta \right\}\right)$ *est* *dominée* *par la mesure* σ*-finie* μ *définie sur* $(\mathfrak{Z}, \mathcal{Z})$ *si*

$$\forall \theta \in \Theta : \mathbb{P}_\theta \ll \mu$$

On appelle densités de la famille $\{\mathbb{P}_\theta, \theta \in \Theta\}$ par rapport à la mesure dominante μ, la famille de fonctions (définies $\mu-$ p.p.)

$$z \mapsto \frac{d\,\mathbb{P}_\theta}{d\mu}(z), \; z \in \mathfrak{Z}, \; \theta \in \Theta.$$

Dans un modèle dominé, on est ramené à estimer une densité plutôt qu'une mesure de probabilité. De plus l'estimation de la densité peut se réduire à l'estimation du paramétre θ.

modèle d'échantillonnage dominé (sur $\mathbb{R}$)

- On observe un *n-échantillon* de v.a.r. $X_1, \ldots, X_n$.
- La loi des X_i appartient au modèle $\{\mathbb{P}_\theta,\ \theta \in \Theta\}$ (famille de probabilités sur $\mathbb{R}$), dominé par une mesure (σ-finie) μ sur $\mathbb{R}$. On note les densités : $\forall \theta \in \Theta, x \in \mathbb{R}$,

$$f(\theta, x) = \frac{d\,\mathbb{P}_\theta}{d\mu}(x)$$

- La loi du n-uplet $(X_1, \ldots, X_n)$ s'écrit

$$\mathbb{P}^{(X_1,\ldots,X_n)} = \mathbb{P}_\theta^n = \mathbb{P}_\theta^{\otimes n} \ll \mu^{\otimes n}$$

elle admet alors une densité : $\forall x_1, \ldots, x_n \in \mathbb{R}$,

$$\boxed{\frac{d\,\mathbb{P}_\theta^{\otimes n}}{d\mu^{\otimes n}}(x_1, \ldots, x_n) = \prod_{i=1}^n f(\theta, x_i)}$$

Exemple 1 : modèle de densité gaussienne univariée

$X_i \sim \mathcal{N}(m, \sigma^2)$, avec $\theta = (m, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+ \setminus \{0\}$.

- la mesure dominante est λ : $\mathbb{P}_\theta = f.\lambda$ où

$$f(\theta, x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(x-m)^2}{2\sigma^2}\Big)$$

- la densité d'un n-uplet est : pour tout $x_1, \ldots, x_n \in \mathbb{R}$,

$$\frac{d\,\mathbb{P}_\theta^n}{d\mu^{\otimes n}}(x_1, \ldots, x_n) = \prod_{i=1}^n f(\theta, x_i)$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big(- \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - m)^2 \Big)$$

Exemple 2 : modèle de Bernoulli

$X_i \sim \text{Bernoulli}(\theta)$, avec $\theta \in \Theta = [0, 1]$

- la mesure dominante est ici $\mu = \delta_0 + \delta_1$, la mesure de comptage sur $\{0, 1\}$:

$$\mathbb{P}_\theta = (1 - \theta)\delta_0 + \theta\delta_1 \ll \mu$$

et pour tout $x \in \{0, 1\}$

$$\boxed{\frac{d\,\mathbb{P}_\theta}{d\mu}(x) = f(\theta, x) = (1 - \theta)I(x = 0) + \theta I(x = 1) = \theta^x(1 - \theta)^{1-x}}$$

- la loi des observations a pour densité par rapport à $\mu^{\otimes n}$,

$$\frac{d\,\mathbb{P}_\theta^n}{d\mu^{\otimes n}}(x_1 \cdots x_n) = \prod_{i=1}^n \theta^{x_i}(1 - \theta)^{1-x_i},$$

pour $x_1, \ldots, x_n \in \{0, 1\}$

Exemple 3 : temps de panne « arrêtés » (1/3)

- On observe $X_1, \ldots, X_n$, où $X_i = Y_i \wedge T$, avec Y_i lois exponentielles de paramètre θ et T temps fixe (censure).
- Cas 1 : $T = \infty$ (pas de censure). Alors $\theta \in \Theta = \mathbb{R}_+ \setminus \{0\}$ et
$$\mathbb{P}_\theta = f.\lambda \text{ où } f(\theta, x) = \theta \exp(-\theta x) I(x \geq 0)$$
et
$$\frac{d\,\mathbb{P}_\theta^n}{d\mu^{\otimes n}}(x_1, \ldots, x_n) = \theta^n \exp\Big(-\theta \sum_{i=1}^n x_i\Big),$$
pour tout $x_i \in \mathbb{R}_+$ et 0 sinon.
- Cas 2 : Comment s'écrit le modèle dans la cas où $T < \infty$ (présence de censure) ? Comment choisir μ ?

Exemple 3 : temps de panne « arrêtés » (2/3)

- Loi $\mathbb{P}_\theta$ de $X = Y \wedge T$: $Y \sim \mathcal{E}xp(\theta)$:

$$\boxed{X = Y1_{\{Y<T\}} + T1_{\{Y\geq T\}}}$$

d'où, pour $g(\theta, x) = \theta e^{-\theta x} I(0 \leq x < T)$,

$$\begin{aligned}\mathbb{P}_\theta &= g.\lambda + \mathbb{P}[Y \geq T]\delta_T \\ &= g.\lambda + e^{-\theta T}\delta_T \\ &\ll \mu = \lambda + \delta_T \quad \text{(par exemple)}.\end{aligned}$$

Exemple 3 : temps de panne « arrêtés » (3/3)

- Alors, pour ce choix de mesure dominante

$$\boxed{\frac{d\,\mathbb{P}_\theta}{d\mu}(x) = \theta e^{-\theta x} I(0 \leq x < T) + e^{-\theta T} I(x = T)}$$

- Finalement,

$$\mathbb{P}_\theta^n = \mathbb{P}_\theta^{\otimes n} \ll \mu^{\otimes n} = \bigotimes_{i=1}^n \left[\lambda + \delta_T\right]$$

et, pour $N_n(T) = \sum_{i=1}^n I(x_i < T)$,

$$\frac{d\,\mathbb{P}_\theta^n}{d\mu^{\otimes n}}(x_1, \ldots, x_n) = \prod_{i=1}^n \left(\theta e^{-\theta x_i} I(0 \leq x_i < T) + e^{-\theta T} I(x_i = T)\right)$$

$$= \theta^{N_n(T)} e^{-\theta \sum_{i=1}^n x_i I(x_i < T)} e^{-\theta T \left(n - N_n(T)\right)},$$

quand $0 \leq x_i \leq T$ et 0 sinon.

- Méthode de substitution (ou des moments)
- Z-estimation
- M-estimation
- Le principe du maximum de vraisemblance

La notation $\mathbb{E}_\theta$

Soit un modèle statistique $\{\mathbb{P}_\theta : \theta \in \Theta\}$ pour une observation Z. Soit $\theta \in \Theta$, on note $\mathbb{E}_\theta$ l'espérance sous $\mathbb{P}_\theta$: càd pour toute fonction mesurable f,

$$\boxed{\mathbb{E}_\theta f(Z) = \int_{\mathfrak{Z}} f(z)\,\mathbb{P}_\theta(dz)}$$

C'est l'espérance de $f(Z)$ quand Z est supposée être de loi $\mathbb{P}_\theta$.

Remarque : étant donné $\theta \in \Theta$, on ne sait pas si la loi de l'observation Z est bien $\mathbb{P}_\theta$ (on sait seulement qu'elle appartient à $\{\mathbb{P}_\theta : \theta \in \Theta\}$), quand on écrit $\mathbb{E}_\theta$, on fait donc l'hypothèse que Z a pour loi $\mathbb{P}_\theta$ et on en déduit des conséquences (par exemple des constructions d'estimateurs ou des résultats statistiques). Si ce résultat est vrai pour tout les $\theta \in \Theta$ alors il est en particulier vrai pour le "vrai θ" : celui pour lequel Z est vraiment distribuée selon $\mathbb{P}_\theta$.

Méthode des moments en dimension 1

- $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathbb{P}_\theta$, avec $\theta \in \Theta \subset \mathbb{R}$
- pour tout $\theta \in \Theta$, on calcul le moment d'ordre 1 de X (sous $\mathbb{P}_\theta$) :

$$m_1(\theta) = \mathbb{E}_\theta X$$

- la méthode des moments en dimension 1 consiste à "estimer" la quantité inconnue $\mathbb{E}_\theta X$ par la moyenne empirique $\bar{X}_n = \frac{1}{n}\sum X_i$ et à :

$$\text{trouver } \hat{\theta}_n \in \Theta \text{ tel que } m_1(\hat{\theta}_n) = \bar{X}_n$$

- (quand il y a une solution) c'est un estimateur plug-in pour $g(x) = x$ et $h(x) = m_1^{-1}$:

$$\boxed{\theta = m_1^{-1}\left(\mathbb{E}_\theta X\right)} \text{ et } \boxed{\widehat{\theta}_n = m_1^{-1}(\bar{X}_n)}$$

Méthode des moments en dimension 1

- Qualité d'estimation via la méthode Delta : pour $h(x) = m_1^{-1}(x)$,

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}\left(0, h'(\mathbb{E}_\theta X)^2 \mathrm{Var}_\theta(X)\right)$$

en loi sous $\mathbb{P}_\theta$. La variance asymptotique dépend en général de θ $\rightarrow$ remplacer θ par $\widehat{\theta}_n$ via le lemme de Slutsky.

- Exemple : $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{E}xp(\theta)$ pour $\theta > 0$. On a pour tout $\theta > 0$,

$$m_1(\theta) = \mathbb{E}_\theta\left[X\right] = \frac{1}{\theta},$$

l'estimateur par moment associé est solution de $m_1(\widehat{\theta}_n) = \bar{X}_n$, càd

$$\widehat{\theta}_n = \frac{1}{\overline{X}_n}$$

Méthode des moments en dimension d

- $X_1, \dots, X_n \overset{i.i.d.}{\sim} \mathbb{P}_\theta$, avec $\theta \in \Theta \subset \mathbb{R}^d$
- pour tout $\theta \in \Theta$, on calcul les d premiers moments de X (sous $\mathbb{P}_\theta$) :

$$m_1(\theta) = \mathbb{E}_\theta X, m_2(\theta) = \mathbb{E}_\theta X^2, \dots, m_d(\theta) = \mathbb{E}_\theta X^d$$

- la méthode des moments consiste à "estimer" les quantités inconnues $\mathbb{E}_\theta X^k$ par leurs moyennes empiriques $\overline{X^k}_n = \frac{1}{n} \sum X_i^k$ et à :

$$\text{trouver } \hat{\theta}_n \in \Theta \text{ tel que } m_k(\hat{\theta}_n) = \overline{X^k}_n$$
$$\underline{\text{pour tout } k = 1, \dots, d}$$

- il n'y a pas forcément de solution !

Exemple en dimension $d > 1$

- $X_1, \ldots, X_n \overset{i.i.d.}{\sim}$ Béta(α, β), de densité

$$x \mapsto \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} 1_{\{0<x<1\}},$$

- Le paramètre est $\theta = (\alpha, \beta) \in \Theta = \mathbb{R}_+ \setminus \{0\} \times \mathbb{R}_+ \setminus \{0\}$.
- On a

$$\boxed{\mathbb{E}_\theta\left[X\right] = \frac{\alpha}{\alpha+\beta}, \quad \mathbb{E}_\theta\left[X^2\right] = \frac{\alpha(\alpha+1)}{(\alpha+\beta+1)(\alpha+\beta)}}$$

Exemple en dimension $d > 1$

- L'estimateur par moment $\widehat{\theta}_n = (\widehat{\theta}_n^{(1)}, \widehat{\theta}_n^{(2)})$ associé est défini par

$$\left\{\begin{array}{ccc} \overline{X}_n & = & \dfrac{\widehat{\theta}_n^{(1)}}{\widehat{\theta}_n^{(1)} + \widehat{\theta}_n^{(2)}} \\ \frac{1}{n}\sum_{i=1}^n X_i^2 & = & \dfrac{\widehat{\theta}_n^{(1)}(\widehat{\theta}_n^{(1)}+1)}{(\widehat{\theta}_n^{(1)} + \widehat{\theta}_n^{(2)}+1)(\widehat{\theta}_n^{(1)} + \widehat{\theta}_n^{(2)})} \end{array}\right.$$

- Etude asymptotique via le TCL multidimensionnel et la méthode Delta multidimensionnelle.

Limites de la méthode des moments

- Méthode non systématique (pb d'existence)
- Représentation pas toujours explicite
- Choix "optimal" des moments ? (notion d'optimalité parmi une classe d'estimateurs)
- Généralisation : Z-estimation (ou estimation par méthode des moments généralisés, GMM= *generalized method of moments*).

Z-estimation

- La méthode des moments (en dimension 1) est basée sur "l'inversibilité" des fonctions

$$m_k(\theta) = \mathbb{E}_\theta X^k$$

i.e. pour tout $\theta \in \Theta$, on voit θ comme solution de l'équation

$$\mathbb{E}_\theta\left[m_k(\theta) - X^k\right] = 0$$

- Principe de construction d'un Z-estimateur : remplacer les $m_k(\theta) - x^k$ par une fonction $\phi(\theta, x) : \Theta \times \mathbb{R} \to \mathbb{R}$ arbitraire telle que

$$\boxed{\forall \theta \in \Theta, \mathbb{E}_\theta\left[\phi(\theta, X)\right] = 0}$$

Z-estimation

- Résoudre l'équation empirique associée :

$$\boxed{\text{Trouver } a \in \Theta \text{ tel que } \frac{1}{n}\sum_{i=1}^{n} \phi(a, X_i) = 0}$$

Définition

On appelle Z-estimateur (Z : "zéro") associé à ϕ tout estimateur $\widehat{\theta}_n$ satisfaisant

$$\sum_{i=1}^{n} \phi(\widehat{\theta}_n, X_i) = 0$$

quand ϕ est telle que

$$\forall \theta \in \Theta, \mathbb{E}_\theta \left[\phi(\theta, X)\right] = 0$$

Z-estimation : programme

Etablir des conditions sur ϕ et sur le modèle $\{\mathbb{P}_\theta, \theta \in \Theta\}$ pour :

- obtenir l'existence et l'unicité de $\widehat{\theta}_n$
- obtenir la consistance de $\widehat{\theta}_n$: pour tout $\theta \in \Theta$,

$$\widehat{\theta}_n \xrightarrow{\mathbb{P}_\theta} \theta$$

- obtenir la normalité asymptotique de $\widehat{\theta}_n$: pour tout $\theta \in \Theta$,

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \xrightarrow{d} \mathcal{N}(0, v(\theta))$$

sous $\mathbb{P}_\theta$.

Z-estimation : exemple du modèle de localisation "shift model"

$\Theta = \mathbb{R}$, $(d\,\mathbb{P}_{\theta}\,/d\lambda)(x) = f(x-\theta)$ où f est symétrique : $f(-x) = f(x)$, $\forall x \in \mathbb{R}$.

- Il n'y a pas d'hypothèse d'existence de moments !
- On pose
$$\phi(a,x) = \mathrm{Arctg}(x-a)$$
- La fonction
$$a \mapsto \mathbb{E}_{\theta}\left[\phi(a,X)\right] = \int_{\mathbb{R}} \mathrm{Arctg}(x-a)f(x-\theta)dx$$
est strictement décroissante et s'annule seulement en $a=\theta$.
- Z-estimateur associé : unique solution $\widehat{\theta}_n$ de
$$\sum_{i=1}^{n} \mathrm{Arctg}(X_i - \widehat{\theta}_n) = 0$$

Le cas multidimensionnel

Si $\Theta \subset \mathbb{R}^d$ avec $d > 1$, la fonction ϕ est remplacée par

$$\Phi = (\phi_1, \dots, \phi_d) : \Theta \times \mathbb{R} \to \mathbb{R}^d .$$

Definition

On appelle Z-estimateur associé à Φ tout estimateur $\widehat{\theta}_n$ satisfaisant

$$\sum_{i=1}^{n} \nabla\Phi(\widehat{\theta}_n, X_i) = 0$$

c'est-à-dire

$$\sum_{i=1}^{n} \phi_\ell(\widehat{\theta}_n, X_i) = 0, \quad \ell = 1, \dots, d$$

Z-estimation $\to$ M-estimation

- En dimension 1 : si

$$\boxed{\phi(\theta, x) = \partial_\theta \psi(\theta, x)}$$

pour une certaine fonction ψ, résoudre $\sum_{i=1}^n \phi(\theta, X_i) = 0$ revient à chercher un point critique (max ou min local) de

$$\theta \mapsto \sum_{i=1}^n \psi(\theta, X_i)$$

- En dimension $d \geq 1$, il faut $\phi(\theta, x) = \nabla_\theta \psi(\theta, x)$ (moins facile à obtenir).
- Invite à généraliser la recherche d'estimateurs via la maximisation d'un critère $\to$ M-estimation (M : "maximum").

M-estimation

- Principe : Se donner une application $\psi : \Theta \times \mathbb{R} \to \mathbb{R}_+$ telle que, pour tout $\theta \in \Theta \subset \mathbb{R}^d$,
$$a \mapsto \mathbb{E}_\theta\left[\psi(a, X)\right] = \int \psi(a, x)\, \mathbb{P}_\theta(dx)$$
admet un maximum en $a = \theta$.

Définition

On appelle M-estimateur associé à ψ tout estimateur $\widehat{\theta}_n$ satisfaisant
$$\sum_{i=1}^n \psi(\widehat{\theta}_n, X_i) = \max_{a \in \Theta} \sum_{i=1}^n \psi(a, X_i)$$

- Il n'y a pas unicité de $\widehat{\theta}_n$ (à ce niveau).

M-estimation : exemple du modèle de localisation "shift model"

- $\Theta = \mathbb{R}$, $d\,\mathbb{P}_\theta / d\lambda(x) = f(x - \theta)$, et $\int_{\mathbb{R}} xf(x)dx = 0$, $\int_{\mathbb{R}} x^2\,\mathbb{P}_\theta(dx) < +\infty$ pour tout $\theta \in \mathbb{R}$. On pose

$$\boxed{\psi(a, x) = -(a - x)^2}$$

- La fonction

$$a \mapsto \mathbb{E}_\theta\left[\psi(a, X)\right] = -\int_{\mathbb{R}} (a - x)^2 f(x - \theta)dx$$

admet un maximum en $a = \mathbb{E}_\theta\left[X\right] = \int_{\mathbb{R}} xf(x - \theta)dx = \theta$.

- *M*-estimateur associé : $\widehat{\theta}_n$ tel que

$$\sum_{i=1}^n (X_i - \widehat{\theta}_n)^2 = \min_{a \in \mathbb{R}} \sum_{i=1}^n (X_i - a)^2.$$

Paramètre de localisation

- C'est aussi un Z-estimateur associé à $\phi(a,x) = 2(x-a)$: on résout

$$\sum_{i=1}^{n}(a - X_i) = 0 \text{ d'où } \widehat{\theta}_n = \overline{X}_n.$$

- Dans cet exemple très simple, tous les points de vue coïncident.
- Si, dans le même contexte, $\int_{\mathbb{R}} x^2 \, \mathbb{P}_\theta(dx) = +\infty$ et $f(x) = f(-x)$, on peut utiliser Z-estimateur avec $\phi(a,x) = \mathrm{Arctg}(x-a)$.

Lien entre Z- et M- estimateurs

- Pas d'inclusion entre ces deux classes d'estimateurs en général :
  - Si ψ non-régulière, M-estimateur $\not\Rightarrow$ Z-estimateur
  - Si une équation d'estimation admet plusieurs solutions distinctes, Z-estimateur $\not\Rightarrow$ M-estimateur (cas d'un extremum local).
- Toutefois, si ψ est régulière, les M-estimateurs sont des Z-estimateurs : si $\Theta \subset \mathbb{R}$ ($d = 1$), en posant

$$\phi(a, x) = \partial_a \psi(a, x),$$

on a

$$\boxed{\sum_{i=1}^{n} \partial_a \psi(\theta, X_i)\big|_{a=\widehat{\theta}_n} = \sum_{i=1}^{n} \phi(\widehat{\theta}_n, X_i) = 0}.$$

Maximum de vraisemblance

- Principe fondamental et incontournable en statistique. Cas particuliers connus depuis le XVIIIème siècle. Définition générale : Fisher (1922).
- Fournit une première méthode systématique de construction d'un *M*-estimateur (souvent un *Z*-estimateur, souvent aussi *a posteriori* un estimateur par substitution simple).
- Procédure optimale (dans quel sens ?) sous des hypothèses de régularité de la famille $\{\mathbb{P}_\theta, \theta \in \Theta\}$ (Cours 6).
- Parfois difficile à mettre en oeuvre en pratique $\rightarrow$ problème d'optimisation.

Fonction de vraisemblance

Définition

Dans le modèle d'échantillonnage (sur $\mathbb{R}$) dominé de densités

$$f(\theta, x) = \frac{d\,\mathbb{P}_\theta}{d\mu}(x),\ x \in \mathbb{R}$$

la fonction de vraisemblance *du n-échantillon* $(X_1, \ldots, X_n)$ *associée à la famille* $\{f(\theta, \cdot), \theta \in \Theta\}$ *est :*

$$\boxed{\theta \in \Theta \mapsto \mathcal{L}_n(\theta, X_1, \ldots, X_n) = \prod_{i=1}^{n} f(\theta, X_i)}$$

- C'est une fonction aléatoire (définie μ-presque partout)
- c'est la densité des observations prise en les données

Exemples

- Exemple 1 : modèle de Poisson. On observe

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} \text{Poisson}(\theta),$$

$\theta \in \Theta = \mathbb{R}_+ \setminus \{0\}$ et prenons $\mu = \sum_{k \in \mathbb{N}} \delta_k$.

- La densité de $\mathbb{P}_\theta$ par rapport à μ est

$$f(\theta, x) = \frac{\theta^x}{x!} e^{-\theta}, \quad x = 0, 1, 2, \ldots.$$

- La fonction de vraisemblance associée s'écrit

$$\theta \mapsto \mathcal{L}_n(\theta, X_1, \ldots, X_n) = \prod_{i=1}^n e^{-\theta} \frac{\theta^{X_i}}{X_i!}$$
$$= \frac{1}{\prod_{i=1}^n X_i!} e^{-n\theta} \theta^{\sum_{i=1}^n X_i}$$

Exemples

- Exemple 2 Modèle de Cauchy. On observe

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} \text{Cauchy centrée en } \theta,$$

$\theta \in \Theta = \mathbb{R}$ et la mesure dominante est λ.

- On a alors

$$\frac{d\,\mathbb{P}_\theta}{d\lambda}(x) = f(\theta, x) = \frac{1}{\pi\big(1 + (x - \theta)^2\big)}$$

- La fonction de vraisemblance associée s'écrit

$$\theta \mapsto \mathcal{L}_n(\theta, X_1, \ldots, X_n) = \frac{1}{\pi^n} \prod_{i=1}^{n} \frac{1}{\big(1 + (X_i - \theta)^2\big)}$$

Principe de maximum de vraisemblance (1/3)

- Cas d'un modèle à deux lois : $\{\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_1}\}$ restreinte à deux points

  $$\Theta = \{\theta_1, \theta_2\} \subset \mathbb{R},$$

  avec $\mathbb{P}_{\theta_i}$ discrète sur $\mathbb{N}$ et μ la mesure de comptage sur $\mathbb{N}$.
- Pour tout $(x_1, \ldots, x_n) \in \mathbb{N}^n$, et pour $\theta \in \{\theta_1, \theta_2\}$,

  $$\mathbb{P}_\theta\left[X_1 = x_1, \ldots, X_n = x_n\right] = \prod_{i=1}^n \mathbb{P}_\theta\left[X_i = x_i\right] = \prod_{i=1}^n f(\theta, x_i).$$

  C'est la probabilité sous $\mathbb{P}_\theta$ d'observer $(x_1, \ldots, x_n)$.

Principe de maximum de vraisemblance (2/3)

Pour les observations $X_1, \ldots, X_n$, la vraisemblance

$$\theta \in \Theta \mapsto \mathcal{L}_n(\theta, X_1, \ldots, X_n) = \prod_{i=1}^{n} f(\theta, X_i)$$

est donc la probabilité sous $\mathbb{P}_\theta$ d'avoir observé $X_1, \ldots, X_n$.

L'EMV choisit donc le θ le plus vraisemblable : càd le paramétre $\theta \in \Theta$ qui maximise la probabilité d'avoir observé $X_1, \ldots, X_n$

Principe de maximum de vraisemblance (3/3)

1. Cas 1 : "θ_1 est plus vraisemblable que θ_2" quand

$$\prod_{i=1}^n f(\theta_1, X_i) \geq \prod_{i=1}^n f(\theta_2, X_i)$$

2. Cas 2 : "θ_2 est plus vraisemblable que θ_1" quand

$$\prod_{i=1}^n f(\theta_2, X_i) > \prod_{i=1}^n f(\theta_1, X_i)$$

Principe de maximum de vraisemblance :

$$\widehat{\theta}_n^{\,\mathtt{mv}} = \begin{cases} \theta_1 & \text{quand } \theta_1 \text{ est le plus vraisemblable} \\ \theta_2 & \text{quand } \theta_2 \text{ est le plus vraisemblable} \end{cases}$$

Estimateur du maximum de vraisemblance

- On généralise le principe précédent pour une famille de lois et un ensemble de paramètres quelconque.
- Situation : $X_1, \dots, X_n \overset{i.i.d.}{\sim} \mathbb{P}_\theta$, $\{\mathbb{P}_\theta, \theta \in \Theta\}$ dominé, $\Theta \subset \mathbb{R}^d$, $\theta \mapsto \mathcal{L}_n(\theta, X_1, \dots, X_n)$ vraisemblance associée.

Définition

On appelle estimateur du maximum de vraisemblance tout estimateur $\widehat{\theta}_{\mathrm{n}}^{\mathtt{mv}}$ satisfaisant

$$\mathcal{L}_n(\widehat{\theta}_{\mathrm{n}}^{\mathtt{mv}}, X_1, \dots, X_n) = \max_{\theta \in \Theta} \mathcal{L}_n(\theta, X_1, \dots, X_n).$$

- Programme : Existence, unicité, propriétés statistiques

Remarques

- Log-vraisemblance :

$$\theta \mapsto \ell_n(\theta, X_1, \ldots, X_n) = \log \mathcal{L}_n(\theta, X_1, \ldots, X_n) = \sum_{i=1}^{n} \log f(\theta, X_i).$$

Bien défini si $f(\theta, \cdot) > 0$ μ-pp.

Max. vraisemblance = max. log-vraisemblance.

(log-vraisemblance est parfois plus facile à maximiser)

- L'estimateur du maximum de vraisemblance ne dépend pas du choix de la mesure dominante μ.
- Equation de vraisemblance :

$$\nabla_\theta \ell_n(\theta, X_1, \ldots, X_n) = 0$$

Exemple : modèle normal

L'expérience statistique est engendrée par un n-échantillon de loi $\mathcal{N}(\mu, \sigma^2)$, le paramètre est $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+ \setminus \{0\}$.

- Vraisemblance

$$\mathcal{L}_n((\mu, \sigma^2), X_1, \dots, X_n) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2\right).$$

- Log-vraisemblance

$$\ell_n((\mu, \sigma^2), X_1, \dots, X_n) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2.$$

Exemple : modèle normal

Equation(s) de vraisemblance : $\nabla_\theta \ell_n(\theta, X_1, \ldots, X_n) = 0$,

$$\begin{cases} \partial_\mu \ell_n\big((\mu,\sigma^2), X_1, \ldots, X_n\big) & = & \dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^n (X_i - \mu) \\ \partial_{\sigma^2} \ell_n\big((\mu,\sigma^2), X_1, \ldots, X_n\big) & = & -\dfrac{n}{2\sigma^2} + \dfrac{1}{2\sigma^4}\displaystyle\sum_{i=1}^n (X_i - \mu)^2 \end{cases}$$

Solution de ces équations (pour $n \geq 2$) :

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

et on vérifie que c'est bien un maximum global alors $\widehat{\theta}_{\mathrm{n}}^{\,\mathtt{mv}} = (\bar{X}_n, \hat{\sigma}_n)$.

Exemple : modèle de Poisson

- Vraisemblance

$$\mathcal{L}_n(\theta, X_1, \ldots, X_n) = \frac{1}{\prod_{i=1}^n X_i!} e^{-n\theta} \theta^{\sum_{i=1}^n X_i}$$

- Log-vraisemblance

$$\ell_n(\theta, X_1, \ldots, X_n) = c(X_1, \ldots, X_n) - n\theta + \sum_{i=1}^n X_i \log \theta$$

- Equation de vraisemblance

$$-n + \sum_{i=1}^n X_i \frac{1}{\theta} = 0, \quad \text{soit} \quad \boxed{\widehat{\theta}_n^{\,\mathtt{mv}} = \frac{1}{n} \sum_{i=1}^n X_i = \overline{X}_n}$$

Exemple : modèle de Laplace

$X_1, \ldots, X_n \overset{i.i.d.}{\sim}$ Laplace de paramètre $\theta \in \Theta = \mathbb{R}$: densité par rapport à la mesure de Lebesgue :

$$f(\theta, x) = \frac{1}{2\sigma} \exp\Big(- \frac{|x-\theta|}{\sigma}\Big),$$

où $\sigma > 0$ est connu.

- Vraisemblance

$$\mathcal{L}_n(\theta, X_1, \ldots, X_n) = (2\sigma)^{-n} \exp\Big(- \frac{1}{\sigma} \sum_{i=1}^n |X_i - \theta|\Big)$$

- Log-vraisemblance

$$\ell_n(\theta, X_1, \ldots, X_n) = -n \log(2\sigma) - \frac{1}{\sigma} \sum_{i=1}^n |X_i - \theta|$$

Exemple : modèle de Laplace

Maximiser $\mathcal{L}_n(\theta, X_1, \ldots, X_n)$ revient à minimiser la fonction $\theta \mapsto \sum_{i=1}^n \left|X_i - \theta\right|$, dérivable presque partout de dérivée constante par morceaux. Equation de vraisemblance :

$$\sum_{i=1}^n \mathrm{sign}(X_i - \theta) = 0.$$

Soit $X_{(1)} \leq \ldots \leq X_{(n)}$ les statistiques d'ordre.

- n pair : $\widehat{\theta}_{\mathrm{n}}^{\mathtt{mv}}$ n'est pas unique ; tout point de l'intervalle $\left[X_{\left(\frac{n}{2}\right)}, X_{\left(\frac{n}{2}+1\right)}\right]$ est un EMV.
- n impair : $\widehat{\theta}_{\mathrm{n}}^{\mathtt{mv}} = X_{\left(\frac{n+1}{2}\right)}$, l'EMV est unique. Mais $\widehat{\theta}_n^{\mathtt{rv}}$ n'existe pas.
- pour tout n, la médiane empirique est un EMV.

Exemple : modèle de Cauchy

- Vraisemblance

$$\mathcal{L}_n(\theta, X_1, \ldots, X_n) = \pi^{-n} \prod_{i=1}^{n} \frac{1}{1 + (X_i - \theta)^2}$$

- Log-vraisemblance

$$\ell_n(\theta, X_1, \ldots, X_n) = -n \log \pi - \sum_{i=1}^{n} \log\left(1 + (X_i - \theta)^2\right)$$

- Equation de vraisemblance

$$\boxed{\sum_{i=1}^{n} \frac{X_i - \theta}{1 + (X_i - \theta)^2} = 0}$$

pas de solution explicite et admet en général plusieurs solutions.

Choix de modèle statistique

- Le statisticien a le choix de la famille $\{\mathbb{P}_\theta, \theta \in \Theta\}$. L'EMV dépend de ce choix.
- Exemple : on a l'échantillon $(n = 10)$:

$$0.92, -0.20, -1.80, 0.02, 0.49, 1.41, -1.59, -1.29, 0.34, {\color{red}100}$$

On choisit un modèle de localisation $\mathbb{P}_\theta(dx) = f(x-\theta)dx$ pour deux f différents :

1. f densité de la loi normale $\Rightarrow \widehat{\theta}_{\rm n}^{\tt mv} = \overline{X}_n = {\color{red}9.83}$.
2. f densité de loi de Laplace $\Rightarrow$ tout point de l'intervalle $[0.02, 0.34]$ est un $\widehat{\theta}_{\rm n}^{\tt mv}$, en particulier, la médiane :

$$\widehat{\theta}_{\rm n}^{\tt mv} = Med(\hat{F}_n) = \widehat{q}_{n,1/2} = {\color{red}0.02}$$

- Autre choix de modèle...

Maximum de vraisemblance $=$ M-estimateur

- Une inégalité de convexité : μ mesure σ-finie sur $\mathbb{R}$; f, g deux densités de probabilités par rapport à μ. Alors

$$\boxed{\int_{\mathbb{R}} f(x)\log f(x)\mu(dx) \geq \int_{\mathbb{R}} f(x)\log g(x)\mu(dx)}$$

(si les intégrales sont finies) avec égalité ssi $f = g$ μ-pp.

- Preuve : à montrer

$$\int_{\mathbb{R}} f(x)\log\frac{g(x)}{f(x)}\mu(dx) \leq 0.$$

(avec une convention de notation appropriée)

Une inégalité de convexité

- On a $\log(1+x) \leq x$ pour $x \geq -1$ avec égalité ssi $x = 0$.
- Donc

$$\log \frac{g(x)}{f(x)} = \log\Big(1 + (\frac{g(x)}{f(x)} - 1)\Big) \leq \frac{g(x)}{f(x)} - 1$$

  (avec égalité ssi $f(x) = g(x)$).
- Finalement

$$\begin{aligned}\int_{\mathbb{R}} f(x) \log \frac{g(x)}{f(x)} \mu(dx) &\leq \int_{\mathbb{R}} f(x) \Big(\frac{g(x)}{f(x)} - 1\Big) \mu(dx) \\ &= \int_{\mathbb{R}} g(x)\mu(dx) - \int_{\mathbb{R}} f(x)\mu(dx) \\ &= 0.\end{aligned}$$

Conséquence pour l'EMV

- On pose

$$\boxed{\psi(a,x) := \log f(a,x), \quad a \in \Theta, \ x \in \mathbb{R}}$$

(avec une convention pour le cas où on n'a pas $f(a,\cdot) > 0$.)

- La fonction

$$a \mapsto \mathbb{E}_\theta\left[\psi(a,X)\right] = \int_{\mathbb{R}} \log f(a,x) f(\theta,x)\mu(dx)$$

a un maximum en $a = \theta$ d'après l'inégalité de convexité.

- Le M-estimateur associé à ψ maximise la fonction

$$a \mapsto \sum_{i=1}^{n} \log f(a, X_i) = \ell_n(a, X_1, \ldots, X_n)$$

  c'est-à-dire la log-vraisemblance.

  l'estimateur du maximum de vraisemblance est un M-estimateur

- C'est aussi un Z-estimateur si la fonction $\theta \mapsto \log f(\theta, \cdot)$ est régulière, associé à la fonction

$$\phi(\theta, x) = \partial_\theta \log f(\theta, x) = \frac{\partial_\theta f(\theta, x)}{f(\theta, x)}, \ \theta \in \Theta, x \in \mathbb{R}$$

  lorsque $\Theta \subset \mathbb{R}$, à condition que le maximum de log-vraisemblance n'est pas atteint sur la frontière de Θ. (Se généralise en dimension d.)

Un M-estimateur qui n'est pas un Z-estimateur

- On observe $X_1, \ldots, X_n \sim_{\text{i.i.d.}}$ uniformes sur $[0, \theta]$, $\theta \in \Theta = \mathbb{R}_+ \setminus \{0\}$.
- On a
$$\mathbb{P}_\theta(dx) = \theta^{-1} 1_{[0,\theta]}(x) dx$$
et
$$\mathcal{L}_n(\theta, X_1, \ldots, X_n) = \theta^{-n} \prod_{i=1}^{n} 1_{[0,\theta]}(X_i)$$
$$= \theta^{-n} 1_{\{\max_{1 \leq i \leq n} X_i \leq \theta\}}$$
- La fonction de vraisemblance n'est pas régulière.
- L'estimateur du maximum de vraisemblance est $\widehat{\theta}_n^{\mathtt{mv}} = \max_{1 \leq i \leq n} X_i$.