

Statistiques mathématiques : cours 3

Guillaume Lécué

11 septembre 2017

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 :

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

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

Aujourd'hui

modèle dominé

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

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

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 $\hat{F}_n, \hat{q}_{n,p}, x_n, \dots$?

Exemple : expérience et modèles statistique (2/2)

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

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

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

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

Théorème de Radon-Nikodym

Théorème

Soient ν et μ deux mesures σ -finies sur $(\mathfrak{Z}, \mathcal{Z})$.

Si $\nu \ll \mu$ alors 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 μ -p.p., μ -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 $\mathcal{E} = (\mathfrak{Z}, \mathcal{Z}, \{\mathbb{P}_\theta, \theta \in \Theta\})$ 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 μ -p.p.)

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

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 paramètre θ** .

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

- ▶ On observe un n -échantillon de v.a.r. X_1, \dots, 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, \dots, X_n) s'écrit

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

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

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

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 $\lambda : \mathbb{P}_\theta = f \cdot \lambda$ où

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

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

$$\begin{aligned} \frac{d\mathbb{P}_\theta^n}{d\mu^{\otimes n}}(x_1, \dots, x_n) &= \prod_{i=1}^n f(\theta, x_i) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - m)^2\right) \end{aligned}$$

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 $x \in \{0, 1\}$

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

- ▶ 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 $x_1, \dots, x_n \in \{0, 1\}$

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

- ▶ On observe X_1, \dots, 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

$$\frac{d\mathbb{P}_\theta^n}{d\mu^{\otimes n}}(x_1, \dots, x_n) = \theta^n \exp\left(-\theta \sum_{i=1}^n x_i\right),$$

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}_θ de $X = Y \wedge T$: $Y \sim \text{Exp}(\theta)$:

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

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

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

- ▶ Finalement,

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

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

$$\begin{aligned} \frac{d\mathbb{P}_\theta^n}{d\mu^{\otimes n}}(x_1, \dots, x_n) &= \prod_{i=1}^n (\theta e^{-\theta x_i} I(0 \leq x_i < T) + e^{-\theta T} I(x_i = T)) \\ &= \theta^{N_n(T)} e^{-\theta \sum_{i=1}^n x_i I(x_i < T)} e^{-\theta T(n - N_n(T))}, \end{aligned}$$

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}_θ

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

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

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

Remarque : étant donné $\theta \in \Theta$, on ne sait pas si la loi de l'observation Z est bien \mathbb{P}_θ (on sait seulement qu'elle appartient à $\{\mathbb{P}_\theta : \theta \in \Theta\}$), quand on écrit \mathbb{E}_θ , on fait donc **l'hypothèse que Z a pour loi \mathbb{P}_θ** 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}_θ .

Méthode des moments en dimension 1

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

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

- ▶ 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 $h(x) = m_1^{-1}$:

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

Méthode des moments en dimension 1

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

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

en loi sous \mathbb{P}_θ . (La variance asymptotique dépend en général de $\theta \rightsquigarrow$
idée : remplacer θ par $\hat{\theta}_n$ via le lemme de Slutsky)

- ▶ Exemple : $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} \text{Exp}(\theta)$ pour $\theta > 0$. On a pour tout $\theta > 0$,

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

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

$$\hat{\theta}_n = \frac{1}{\bar{X}_n}$$

Méthode des moments en dimension d

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

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

- ▶ la méthode des moments consiste à "estimer" les quantités inconnues $\mathbb{E}_\theta X^k$ par leurs moyennes empiriques $\overline{X}_n^k = \frac{1}{n} \sum X_i^k$ et à :
trouver $\hat{\theta}_n \in \Theta$ solution de $m_k(\hat{\theta}_n) = \overline{X}_n^k$ pour tout $k = 1, \dots, d$
- ▶ il n'y a pas forcément de solution !

Exemple en dimension $d > 1$

- ▶ $X_1, \dots, X_n \stackrel{i.i.d.}{\sim}$ Béta(α, β), de densité

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

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

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

Exemple en dimension $d > 1$

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

$$\begin{cases} \bar{X}_n &= \frac{\hat{\theta}_n^{(1)}}{\hat{\theta}_n^{(1)} + \hat{\theta}_n^{(2)}} \\ \bar{X}_n^2 &= \frac{\hat{\theta}_n^{(1)}(\hat{\theta}_n^{(1)} + 1)}{(\hat{\theta}_n^{(1)} + \hat{\theta}_n^{(2)} + 1)(\hat{\theta}_n^{(1)} + \hat{\theta}_n^{(2)})} \end{cases}$$

- ▶ 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

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

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

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

Z-estimation

- ▶ Résoudre l'équation empirique associée :

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

Définition

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

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

quand ϕ est telle que

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

Z-estimation : programme

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

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

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

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

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

sous \mathbb{P}_θ .

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

$$\phi(a, x) = \text{Arctg}(x - a)$$

- ▶ La fonction

$$a \mapsto \mathbb{E}_\theta [\phi(a, X)] = \int_{\mathbb{R}} \text{Arctg}(x - a) f(x - \theta) dx$$

est strictement décroissante et s'annule seulement en $a = \theta$.

- ▶ **Z-estimateur associé** : unique solution $\hat{\theta}_n$ de

$$\sum_{i=1}^n \text{Arctg}(X_i - \hat{\theta}_n) = 0$$

Le cas multidimensionnel

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

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

Definition

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

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

c'est-à-dire $\sum_{i=1}^n \phi_\ell(\hat{\theta}_n, X_i) = 0$, $\ell = 1, \dots, d$ quand Φ est telle que

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

Z-estimation \rightarrow M-estimation

- ▶ En dimension 1 : si

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

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 \rightarrow M-estimation (M : "maximum").

M-estimation

- Principe : Trouver $\psi : \Theta \times \mathbb{R} \rightarrow \mathbb{R}_+$ telle que, pour tout $\theta \in \Theta \subset \mathbb{R}^d$,

$$a \mapsto \mathbb{E}_\theta [\psi(a, X)] = \int \psi(a, x) \mathbb{P}_\theta(dx)$$

admet un maximum en $a = \theta$.

Définition

On appelle M-estimateur (M = maximum) associé à ψ , tout estimateur $\hat{\theta}_n$ satisfaisant

$$\sum_{i=1}^n \psi(\hat{\theta}_n, X_i) = \max_{a \in \Theta} \sum_{i=1}^n \psi(a, X_i)$$

quand ψ est telle que pour tout $\theta \in \Theta$, $a \mapsto \mathbb{E}_\theta [\psi(a, X)]$ est maximum en θ .

- Il n'y a pas unicité de $\hat{\theta}_n$ (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

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

- ▶ La fonction

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

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

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

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

Paramètre de localisation

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

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

- ▶ 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 $\phi(a, x) = \text{Arctg}(x - a)$.

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 de Z -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

$$\sum_{i=1}^n \partial_a \psi(\theta, X_i) \Big|_{a=\hat{\theta}_n} = \sum_{i=1}^n \phi(\hat{\theta}_n, X_i) = 0$$

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 → **problème d'optimisation**.

Fonction de vraisemblance

Définition

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

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

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

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

- ▶ 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

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

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

- ▶ La densité de \mathbb{P}_θ par rapport à μ est

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

- ▶ La **fonction de vraisemblance** associée s'écrit

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

Exemples

- ▶ Exemple 2 **Modèle de Cauchy**. On observe

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

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

- ▶ On a alors

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

- ▶ La **fonction de vraisemblance** associée s'écrit

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

Principe de maximum de vraisemblance (1/3)

- ▶ Cas d'un modèle à deux lois : $\{\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_2}\}$ **restreinte à deux points**

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

avec \mathbb{P}_{θ_i} discrète sur \mathbb{N} et μ la mesure de comptage sur \mathbb{N} .

- ▶ Pour tout $(x_1, \dots, x_n) \in \mathbb{N}^n$, et pour $\theta \in \{\theta_1, \theta_2\}$,

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

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

Principe de maximum de vraisemblance (2/3)

Pour les observations X_1, \dots, X_n , la vraisemblance

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

est donc la probabilité sous \mathbb{P}_θ d'avoir observé X_1, \dots, X_n .

L'EMV choisit donc le θ le plus vraisemblable : c'ad le paramètre $\theta \in \Theta$ qui maximise la probabilité (sous \mathbb{P}_θ) d'avoir observé X_1, \dots, 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 :

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

Estimateur du maximum de vraisemblance

- ▶ On généralise le principe précédent pour une famille de lois et un ensemble de paramètres **quelconque**.
- ▶ Situation : $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} \mathbb{P}_\theta$, $\{\mathbb{P}_\theta, \theta \in \Theta\}$ 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 $\hat{\theta}_n^{\text{mv}}$ satisfaisant

$$\mathcal{L}_n(\hat{\theta}_n^{\text{mv}}, X_1, \dots, X_n) = \max_{\theta \in \Theta} \mathcal{L}_n(\theta, X_1, \dots, X_n).$$

- ▶ Programme : **Existence, unicité, propriétés statistiques**

Remarques

- ▶ Log-vraisemblance :

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

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** :

$$\nabla_{\theta} \ell_n(\theta, X_1, \dots, X_n) = 0$$

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 : $\nabla_{\theta} \ell_n(\theta, X_1, \dots, X_n) = 0$,

$$\left\{ \begin{array}{l} \partial_{\mu} \ell_n((\mu, \sigma^2), X_1, \dots, X_n) = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu) \\ \partial_{\sigma^2} \ell_n((\mu, \sigma^2), X_1, \dots, X_n) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (X_i - \mu)^2 \end{array} \right.$$

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

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

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

Exemple : modèle de Poisson

► **Vraisemblance**

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

► **Log-vraisemblance**

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

► **Equation de vraisemblance**

$$-n + \sum_{i=1}^n X_i \frac{1}{\theta} = 0, \text{ soit } \boxed{\hat{\theta}_n^{\text{mv}} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}_n}$$

Exemple : modèle de Laplace

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

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

où $\sigma > 0$ est **connu**.

► **Vraisemblance**

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

► **Log-vraisemblance**

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

Exemple : modèle de Laplace

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

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

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

- ▶ n pair : $\hat{\theta}_n^{\text{mv}}$ **n'est pas unique** ; tout point de l'intervalle $[X_{(\frac{n}{2})}, X_{(\frac{n}{2}+1)}]$ est un EMV.
- ▶ n impair : $\hat{\theta}_n^{\text{mv}} = X_{(\frac{n+1}{2})}$, l'EMV est unique.
- ▶ **pour tout** n , la médiane empirique est un EMV.

Exemple : modèle de Cauchy

► **Vraisemblance**

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

► **Log-vraisemblance**

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

► **Equation de vraisemblance**

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

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, 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 $\Rightarrow \hat{\theta}_n^{\text{mv}} = \bar{X}_n = 9.83$.
2. f densité de loi de Laplace \Rightarrow tout point de l'intervalle $[0.02, 0.34]$ est un $\hat{\theta}_n^{\text{mv}}$, en particulier, la médiane :

$$\hat{\theta}_n^{\text{mv}} = \text{Med}(\hat{F}_n) = \hat{q}_{n,1/2} = 0.02$$

- ▶ 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

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

(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

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

(avec égalité ssi $f(x) = g(x)$).

► Finalement

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

Conséquence pour l'EMV

- ▶ On pose

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

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

- ▶ La fonction

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

est maximale en $a = \theta$ d'après **l'inégalité de convexité**.

- ▶ Le M -estimateur associé à ψ maximise la fonction

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

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 $X_1, \dots, X_n \sim_{\text{i.i.d.}}$ uniformes sur $[0, \theta]$, $\theta \in \Theta = \mathbb{R}_+ \setminus \{0\}$.
- ▶ On a

$$\mathbb{P}_\theta(dx) = \theta^{-1} \mathbf{1}_{[0, \theta]}(x) dx$$

et

$$\begin{aligned} \mathcal{L}_n(\theta, X_1, \dots, X_n) &= \theta^{-n} \prod_{i=1}^n \mathbf{1}_{[0, \theta]}(X_i) \\ &= \theta^{-n} \mathbf{1}_{\{\max_{1 \leq i \leq n} X_i \leq \theta\}} \end{aligned}$$

- ▶ La fonction de vraisemblance **n'est pas régulière**.
- ▶ **L'estimateur du maximum de vraisemblance est $\hat{\theta}_n^{\text{mv}} = \max_{1 \leq i \leq n} X_i$.**