MAP 433 : Introduction aux méthodes statistiques. Cours 3

14 février 2014

Aujourd'hui

1. Rappel succint du Cours 2
2. Modélisation statistique
   - Expérience statistique
   - Expériences dominées
   - Modèle de densité
3. Méthodes d'estimation pour le modèle de densité
   - Méthode des moments
   - Z-estimation
   - M-estimation
   - Principe de maximum de vraisemblance

Estimation de fonctionnelles régulières

- $X_1, \ldots, X_n \sim_{\text{i.i.d.}} F$.
- Estimation de $T(F) = h\big(\int_{\mathbb{R}} g(x)dF(x)\big)$.
- Estimateur par substitution ou *plug-in* de $T(F)$:

$$\boxed{T(\widehat{F}_n) = h\Big(\frac{1}{n}\sum_{i=1}^{n} g(X_i)\Big)}$$

Limites de l'approche empirique

L'estimation de $T(F)$ par $T(\widehat{F}_n)$ n'est pas toujours possible :

- La fonctionnelle $F \rightsquigarrow T(F)$ n'est pas « régulière »,
- La paramétrisation $F \rightsquigarrow T(F)$ ne donne pas lieu à une forme analytique simple. $\rightarrow$ autres approches.

Exemple. Hypothèse : F admet une densité f par rapport à le mesure de Lebesgue, continue ($=$ pp à une fonction continue f).

$$T(F) = f(x_0), \;\; x_0 \in \mathbb{R} \;\; \text{(donné)}.$$

On ne peut pas prendre comme estimateur $\widehat{F}_n'(x_0)$ car $\widehat{F}_n$ n'est pas différentiable (constante par morceaux...)

Limites de l'approche empirique

L'estimation de $T(F)$ par $T(\widehat{F}_n)$ n'est pas toujours souhaitable :

- Souvent on dispose d'information a priori supplémentaire : F appartient à une sous-classe particulière de distributions, et il y a des choix plus judicieux que l'estimateur par plug-in.

Conclusion

- L'approche empirique, basée sur $\widehat{F}_n$ permet d'estimer une distribution inconnue F ou une fonctionnelle $T(F) \in \mathbb{R}$ à partir d'un n-échantillon, mais
  - reste très générale, pas toujours adaptée.
  - restreinte à la situation d'un n-échantillon.
- Formalisation de la notion d'expérience statistique
  - incorporation d'information de modélisation supplémentaire.
  - construction de méthodes d'estimation – de décision – systématiques.
  - comparaison et optimalité des méthodes.

Expérience statistique

Consiste à identifier :

- Des observations

$$\mathtt{x}_1, \mathtt{x}_2, \ldots, \mathtt{x}_n$$

  considérées comme des réalisations de variables aléatoires $Z = (X_1, \ldots, X_n)$ de loi $\mathbb{P}^Z$.
- Une famille de lois

$$\{\mathbb{P}_\vartheta, \; \vartheta \in \Theta\}.$$

- Une problématique : retrouver le paramètre ϑ tel que $\mathbb{P}^Z = \mathbb{P}_\vartheta$ (estimation) ou bien prendre une décision sur une propriété relative à ϑ (test).

Expérience statistique

- Approche générale empirique :
  - $\vartheta = F$, Θ est l'ensemble de toutes les lois (s'il s'agit de l'estimation de F) ;
  - $\vartheta = F$, Θ est l'ensemble de toutes les lois vérifiant une hypothèse très générale, par exemple, la bornitude d'un moment (s'il s'agit de l'estimation de $T(F)$).
- Approche paramétrique : on suppose que F appartient à une famille de lois connue indexée par un paramètre ϑ de dimension finie : $\vartheta \in \Theta \subset \mathbb{R}^d$.
  - Exemple : $\Theta = \mathbb{R}$,

$$X_i = \vartheta + \xi_i, \quad i = 1, \ldots, n,$$

ξ_i v.a. i.i.d. de densité connue f sur $\mathbb{R}$ et $\mathbb{E}(X_i) = \vartheta$.
Question : en utilisant cette information supplémentaire, peut-on construire un estimateur plus performant que l'estimateur $\bar{X}_n$ basé sur l'approche empirique ?

Expérience statistique

- En écrivant

$$X_i = \vartheta + \xi_i, \quad i = 1, \ldots, n,$$

ξ_i v.a. i.i.d. de densité connue f, nous précisons la forme de la loi $\mathbb{P}_\vartheta$ de $(X_1, \ldots, X_n)$:

$$\mathbb{P}_\vartheta\left[A\right] = \int_A \left(\prod_{i=1}^n f(x_i - \vartheta)\right) dx_1 \ldots dx_n,$$

pour tout $A \in \mathcal{B}(\mathbb{R}^n)$.

Expérience statistique

Définition

Une expérience (un modèle) statistique $\mathcal{E}$ est le triplet

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

avec

- $(\mathfrak{Z}, \mathcal{Z})$ *espace mesurable (souvent $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$),*
- $\{\mathbb{P}_\vartheta,\ \vartheta \in \Theta\}$ *famille de probabilités définies simultanément sur le même espace $(\mathfrak{Z}, \mathcal{Z})$,*
- ϑ *est le paramètre inconnu, et Θ est l'ensemble des paramètres connu.*

Experience engendrée par $(X_1, \ldots, X_n)$

- Traitement sur un exemple : on observe

$$Z = (X_1, \ldots, X_n), \qquad X_i = \vartheta + \xi_i,$$

ξ_i v.a. i.i.d. de densité connue f.

- La famille de lois $\{\mathbb{P}^n_\vartheta, \vartheta \in \Theta = \mathbb{R}\}$ est définie sur $\mathfrak{Z} = \mathbb{R}^n$ par

$$\mathbb{P}^n_\vartheta\left[A\right] = \int_A \left(\prod_{i=1}^n f(x_i - \vartheta)\right) dx_1 \ldots dx_n,$$

pour $A \in \mathcal{Z} = \mathcal{B}(\mathbb{R}^n)$ (et $\mathbb{P}^Z$ est l'une des $\mathbb{P}^n_\vartheta$).

- Expérience engendrée par l'observation Z :

$$\mathcal{E}^n = \left(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \{\mathbb{P}^n_\vartheta,\ \vartheta \in \Theta\}\right).$$

Expérience (modèle) paramétrique, non-paramétrique

- Si Θ peut être « pris » comme un sous-ensemble de $\mathbb{R}^d$: expérience (=modèle) paramétrique.
- Sinon (par exemple si le paramètre ϑ est un élément d'un espace fonctionnel) : expérience (=modèle) non-paramétrique.

Expériences dominées

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

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

  à l'étude d'une famille de fonctions

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

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

$$\mu\big[A\big] = 0 \Rightarrow \nu\big[A\big] = 0.$$

Théorème de Radon-Nikodym

Théorème

Si $\nu \ll \mu$, il existe une fonction positive

$$z \rightsquigarrow p(z) \stackrel{\text{notation}}{=} \frac{d\nu}{d\mu}(z),$$

définie $\mu-$p.p., $\mu-$ intégrable, telle que

$$\nu[A] = \int_A p(z)\mu(dz) = \int_A \frac{d\nu}{d\mu}(z)\mu(dz), \;\; A \in \mathcal{Z}.$$

Expérience dominée

Définition

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

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

On appelle densités de la famille $\{\mathbb{P}_\vartheta, \vartheta \in \Theta\}$ la famille de fonctions (définies $\mu-$ p.p.)

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

Densité, régression

Deux classes d'expériences statistiques dominées fondamentales :

- Le modèle de densité
- Le modèle de régression

Modèle de densité (paramétrique)

- On observe un n-échantillon de v.a.r. $X_1, \ldots, X_n$.
- La loi des X_i appartient à $\{\mathbb{P}_\vartheta, \vartheta \in \Theta\}$, famille de probabilités sur $\mathbb{R}$, dominée par une mesure (σ-finie) $\mu(dx)$ sur $\mathbb{R}$.
- La loi de $(X_1, \ldots, X_n)$ s'écrit

$$\begin{aligned}\mathbb{P}_\vartheta^n(dx_1 \cdots dx_n) &= \mathbb{P}_\vartheta(dx_1) \otimes \cdots \otimes \mathbb{P}_\vartheta(dx_n) \\ &\ll \mu(dx_1) \otimes \cdots \otimes \mu(dx_n) \\ &\overset{\text{notation}}{=} \mu^n(dx_1 \cdots dx_n)\end{aligned}$$

Modèle de densité (paramétrique)

- Densité du modèle : on part de

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

  et

$$\frac{d\,\mathbb{P}^n_\vartheta}{d\mu^n}(x_1, \ldots, x_n) = \prod_{i=1}^n f(\vartheta, x_i), \quad x_1, \ldots, X_n \in \mathbb{R}\,.$$

- L'expérience statistique engendrée par $(X_1, \ldots, X_n)$ s'écrit :

$$\mathcal{E}^n = \Big(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \big\{\mathbb{P}^n_\vartheta, \vartheta \in \Theta\big\}\Big), \quad \Theta \subset \mathbb{R}^d.$$

Exemple 1 : modèle de densité gaussienne univariée

- $X_i \sim \mathcal{N}(m, \sigma^2)$, avec

$$\vartheta = (m, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+ \setminus \{0\}.$$

$$\mathbb{P}_\vartheta(dx) = f(\vartheta, x)dx = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(x-m)^2}{2\sigma^2}\Big) dx$$
$$\ll \mu(dx) = dx.$$

- Puis

$$\frac{d\,\mathbb{P}_\vartheta^n}{d\mu^n}(x_1, \ldots, x_n) = \prod_{i=1}^n f(\vartheta, x_i)$$
$$= (2\pi\sigma^2)^{-n/2} \exp\Big(- \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - m)^2\Big),$$

avec $x_1, \ldots, x_n \in \mathbb{R}$.

Exemple 2 : modèle de Bernoulli

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

$$\begin{aligned}\mathbb{P}_\vartheta(dx) &= (1-\vartheta)\,\delta_0(dx) + \vartheta\,\delta_1(dx) \\ &\ll \mu(dx) = \delta_0(dx) + \delta_1(dx) \ \ \text{(mesure de comptage)}.\end{aligned}$$

- Puis

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

avec $x \in \{0, 1\}$ (et 0 sinon), et

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

avec $x_i \in \{0, 1\}$ (et 0 sinon).

Exemple 3 : temps de panne « arrêtés »

- 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 $\vartheta \in \Theta = \mathbb{R}_+ \setminus \{0\}$ et

$$\mathbb{P}_\vartheta(dx) = \vartheta \exp(-\vartheta x) 1_{\{x \geq 0\}} dx \ll \mu(dx) = dx$$

et

$$\frac{d\,{\mathbb{P}_\vartheta}^n}{d\mu^n}(x_1, \ldots, x_n) = \vartheta^n \exp\Big(-\vartheta \sum_{i=1}^n x_i\Big),$$

avec $x_i \in \mathbb{R}_+$ (et 0 sinon).

- Cas 2 : Comment s'écrit le modèle dans la cas où $T < \infty$ (présence de censure) ? Comment choisir μ ?

Exemple : temps de panne « arrêtés »

- Loi $\mathbb{P}_\vartheta(dx)$ de $X = Y \wedge T$: $Y \sim$ exponentielle de paramètre ϑ :

$$\boxed{X = Y 1_{\{Y<T\}} + T 1_{\{Y \geq T\}}}$$

d'où

$$\begin{aligned}\mathbb{P}_\vartheta(dx) &= \vartheta e^{-\vartheta x} 1_{\{0 \leq x < T\}} dx + \Big(\int_T^{+\infty} \vartheta e^{-\vartheta y} dy \Big) \delta_T(dx) \\ &= \vartheta e^{-\vartheta x} 1_{\{0 \leq x < T\}} dx + e^{-\vartheta T} \delta_T(dx) \\ &\ll \mu(dx) = dx + \delta_T(dx) \quad \text{(par exemple).}\end{aligned}$$

Exemple : temps de panne « arrêtés » (fin)

- Alors, pour ce choix de mesure dominante

$$\boxed{\frac{d\,\mathbb{P}_\vartheta}{d\mu}(x) = \vartheta e^{-\vartheta x} 1_{\{0 \leq x < T\}} + e^{-\vartheta T} 1_{\{x = T\}}}$$

- Finalement,

$$\mathbb{P}_\vartheta^n(dx_1, \ldots dx_n) \ll \mu^n(dx_1 \ldots dx_n) = \bigotimes_{i=1}^n \big[dx_i + \delta_T(dx_i)\big]$$

et

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

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

avec $0 \leq x_i \leq T$ et 0 sinon, et $N_n(T) = \sum_{i=1}^n 1_{\{x_i < T\}}$.

Méthodes d'estimation

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

Méthode des moments : dimension 1

- $X_1, \ldots, X_n \sim_{\text{i.i.d.}} \mathbb{P}_\vartheta$, avec $\vartheta \in \Theta \subset \mathbb{R}$.
- Principe : trouver $g : \mathbb{R} \to \mathbb{R}$ (en général $g(x) = x^k$) et $h : \mathbb{R} \to \mathbb{R}$ régulières de sorte que

$$\vartheta = h\big(\mathbb{E}_\vartheta\left[g(X)\right]\big) = h\Big(\int_{\mathbb{R}} g(x)dF_\vartheta(x)\Big) = T(F_\vartheta)$$

  et T fonctionnelle régulière de la distribution inconnnue F_ϑ.
- Estimateur : « plug-in »

$$\widehat{\vartheta}_n = h\big(\tfrac{1}{n}\sum_{i=1}^n g(X_i)\big).$$

Méthode des moments

- <u>Précision d'estimation</u> via les techniques empiriques :

$$\sqrt{n}\big(\widehat{\vartheta}_n - \vartheta\big) \xrightarrow{d} \mathcal{N}\big(0, h'(\mathbb{E}_\vartheta[g(X)])^2 \mathrm{Var}_\vartheta[g(X)]\big)$$

  en loi sous $\mathbb{P}_\vartheta$ et la variance asymptotique dépend en général de ϑ $\rightarrow$ élimination par estimation préliminaire licite via le lemme de Slutsky.

- <u>Exemple</u> : $X_1, \ldots, X_n \sim_{\text{i.i.d.}}$ exponentielle de paramètre ϑ. On a

$$\mathbb{E}_\vartheta\left[X\right] = \frac{1}{\vartheta},$$

  l'estimateur par moment associé s'écrit

$$\widehat{\vartheta}_n = \frac{1}{\overline{X}_n}.$$

Exemple en dimension $d > 1$

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

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

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

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

Exemple en dimension $d > 1$

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

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

- Etude asymptotique via le TCL multidimensionnel et la méthode « delta » multidimensionnelle.

Limites de la méthode des moments

- Méthode non systématique
- Représentation pas toujours explicite
- Choix de la fonction g, 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 de la fonction

$$m_g(\vartheta) = \int_{\mathbb{R}} g(x)\,\mathbb{P}_\vartheta(dx)$$

i.e. pour tout $\vartheta \in \Theta$

$$\int_{\mathbb{R}} \big(m_g(\vartheta) - g(x)\big)\,\mathbb{P}_\vartheta(dx) = 0.$$

- Principe de construction d'un Z-estimateur : remplacer $m_g(\vartheta) - g(x)$ par une fonction $\phi(\vartheta, x) : \Theta \times \mathbb{R} \to \mathbb{R}$ arbitraire telle que

$$\boxed{\forall \vartheta \in \Theta, \int_{\mathbb{R}} \phi(\vartheta, x)\,\mathbb{P}_\vartheta(dx) = 0.}$$

Z-estimation

- Résoudre l'équation empirique associée :

$$\boxed{\frac{1}{n}\sum_{i=1}^{n}\phi(a, X_i) = 0 \;\; \text{pour} \;\; a \in \Theta.}$$

Définition

On appelle Z-estimateur associé à ϕ tout estimateur $\widehat{\vartheta}_n$ satisfaisant

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

- Il n'y a pas unicité de $\widehat{\vartheta}_n$ (à ce niveau).
- Programme Etablir des conditions sur ϕ et sur la famille $\{\mathbb{P}_\vartheta, \vartheta \in \Theta\}$ pour obtenir la convergence et les performances asymptotiques de $\widehat{\vartheta}_n$.

Z-estimation : à quoi ça sert ?

- Exemple. $\Theta = \mathbb{R}$, $\mathbb{P}_\vartheta(dx) = f(x-\vartheta)dx$, et f symétrique : $f(-x) = f(x)$, $\forall x \in \mathbb{R}$.
- Il n'y a pas de bornitude des moments !
- On pose
$$\phi(a,x) = \mathrm{Arctg}(x-a).$$
- La fonction
$$a \rightsquigarrow \mathbb{E}_\vartheta\left[\phi(a,X)\right] = \int_{\mathbb{R}} \mathrm{Arctg}(x-a)f(x-\vartheta)dx$$
est strictement décroissante et s'annule seulement en $a=\vartheta$.
- Z-estimateur associé : solution $\widehat{\vartheta}_n$ de
$$\sum_{i=1}^n \mathrm{Arctg}(X_i - \widehat{\vartheta}_n) = 0$$
(unicité).

Le cas multidimensionnel

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

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

Definition

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

$$\sum_{i=1}^{n} \phi_\ell(\widehat{\vartheta}_n, X_i) = 0, \;\; \ell = 1, \ldots, d.$$

Z-estimation $\to$ M-estimation

- <u>En dimension 1</u> : si

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

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

$$\vartheta \rightsquigarrow \sum_{i=1}^n \psi(\vartheta, X_i).$$

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

M-estimation

- Principe : Se donner une application $\psi : \Theta \times \mathbb{R} \to \mathbb{R}_+$ telle que, pour tout $\vartheta \in \Theta \subset \mathbb{R}^d$,

$$a \rightsquigarrow \mathbb{E}_\vartheta\left[\psi(a, X)\right] = \int \psi(a, x)\, \mathbb{P}_\vartheta(dx)$$

admet un maximum en $a = \vartheta$.

Définition

On appelle M-estimateur associé à ψ tout estimateur $\widehat{\vartheta}_n$ satisfaisant

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

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

Un exemple classique : paramètre de localisation

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

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

- La fonction

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

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

- *M*-estimateur associé :

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

Paramètre de localisation

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

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

- Dans cet exemple très simple, tous les points de vue coïncident.
- Si, dans le même contexte, $\int_{\mathbb{R}} x^2 \,\mathbb{P}_\vartheta(dx) = +\infty$ et $f(x) = f(-x)$, on peut utiliser Z-estimateur avec $\phi(a,x) = \mathrm{Arctg}(x-a)$. Méthode robuste, mais est-elle optimale ? Peut-on faire mieux si f est connue ? A suivre...

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

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

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 $\{\mathbb{P}_\vartheta, \vartheta \in \Theta\}$ (Cours 6).
- Parfois difficile à mettre en oeuvre en pratique $\rightarrow$ méthodes numériques, statistique computationnelle.

Fonction de vraisemblance

- La famille $\{\mathbb{P}_\vartheta, \vartheta \in \Theta\}$ est dominée par une mesure σ-finie μ. On se donne, pour $\vartheta \in \Theta$

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

Définition

Fonction de vraisemblance du n-échantillon associée à la famille $\{f(\vartheta, \cdot), \vartheta \in \Theta\}$ *:*

$$\boxed{\vartheta \rightsquigarrow \mathcal{L}_n(\vartheta, X_1, \ldots, X_n) = \prod_{i=1}^{n} f(\vartheta, X_i)}$$

- C'est une fonction aléatoire (définie μ-presque partout).

Exemples

- Exemple 1 : Modèle de Poisson. On observe

$$X_1, \ldots, X_n \sim_{\text{i.i.d.}} \text{Poisson}(\vartheta),$$

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

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

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

- La fonction de vraisemblance associée s'écrit

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

Exemples

- Exemple 2 Modèle de Cauchy. On observe

$$X_1, \ldots, X_n \sim_{\text{i.i.d.}} \text{Cauchy},$$

$\vartheta \in \Theta = \mathbb{R}$ et $\mu(dx) = dx$ (par exemple).

- On a alors

$$\mathbb{P}_\vartheta(dx) = f(\vartheta, x)dx = \frac{1}{\pi\big(1 + (x - \vartheta)^2\big)}dx.$$

- La fonction de vraisemblance associée s'écrit

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

Principe de maximum de vraisemblance

- Cas d'une famille de lois restreinte à deux points

$$\Theta = \{\vartheta_1, \vartheta_2\} \subset \mathbb{R},$$

avec $\mathbb{P}_{\vartheta_i}$ discrète et $\mu(dx)$ la mesure de comptage.

- A priori, pour tout $(x_1, \ldots, x_n)$, et pour $\vartheta \in \{\vartheta_1, \vartheta_2\}$,

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

La probabilité d'avoir la réalisation fixée $(x_1, \ldots, x_n)$.

Principe de maximum de vraisemblance

- A posteriori, on observe $(X_1, \ldots, X_n)$. L'événement

$$\Big\{ \prod_{i=1}^{n} f(\vartheta_1, X_i) > \prod_{i=1}^{n} f(\vartheta_2, X_i) \Big\} \quad \text{(Cas 1)}$$

ou bien l'événement

$$\Big\{ \prod_{i=1}^{n} f(\vartheta_2, X_i) > \prod_{i=1}^{n} f(\vartheta_1, X_i) \Big\} \quad \text{(Cas 2)}$$

est réalisé. (On ignore le cas d'égalité.)

- Principe de maximum de vraisemblance :

$$\widehat{\vartheta}_n^{\,\mathtt{mv}} = \vartheta_1 1_{\{\text{Cas 1}\}} + \vartheta_2 1_{\{\text{Cas 2}\}}.$$

Estimateur du maximum de vraisemblance

- On généralise le principe précédent pour une famille de lois et un ensemble de paramètres quelconques.
- Situation : $X_1, \ldots, X_n \sim_{\text{i.i.d.}} \mathbb{P}_\vartheta$, $\{\mathbb{P}_\vartheta, \vartheta \in \Theta\}$ dominée, $\Theta \subset \mathbb{R}^d$, $\vartheta \rightsquigarrow \mathcal{L}_n(\vartheta, X_1, \ldots, X_n)$ vraisemblance associée.

Définition

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

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

- Existence, unicité...

Remarques

- <u>Log-vraisemblance</u> :

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

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

Max. vraisemblance = max. log-vraisemblance.

- L'estimateur du maximum de vraisemblance ne dépend pas du choix de la mesure dominante μ.
- Notion de racine de l'équation de vraisemblance : tout estimateur $\widehat{\vartheta}_n^{\,\mathtt{rv}}$ vérifiant

$$\nabla_\vartheta \ell_n(\widehat{\vartheta}_n^{\,\mathtt{rv}}, X_1, \ldots, X_n) = 0.$$

Exemple : modèle normal

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

- Vraisemblance

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

- Log-vraisemblance

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

Exemple : modèle normal

Equation(s) de vraisemblance

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

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

$$\boxed{\widehat{\vartheta}_n^{\,\mathtt{rv}} = \Big(\overline{X}_n, \frac{1}{n}\sum_{i=1}^n (X_i-\overline{X}_n)^2\Big)}$$

et on vérifie que $\widehat{\vartheta}_n^{\,\mathtt{rv}} = \widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mv}}$.

Exemple : modèle de Poisson

- Vraisemblance

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

- Log-vraisemblance

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

- Equation de vraisemblance

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

et on vérifie que $\widehat{\vartheta}_n^{\,\mathtt{rv}} = \widehat{\vartheta}_n^{\,\mathtt{mv}}$.

Exemple : modèle de Laplace

L'expérience statistique est engendrée par un n-échantillon de loi de Laplace de paramètre $\vartheta \in \Theta = \mathbb{R}$. La densité par rapport à la mesure de Lebesgue :

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

où $\sigma > 0$ est connu.

- Vraisemblance

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

- Log-vraisemblance

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

Exemple : modèle de Laplace

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

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

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

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

Exemple : modèle de Cauchy

- Vraisemblance

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

- Log-vraisemblance

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

- Equation de vraisemblance

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

pas de solution explicite et admet en général plusieurs solutions.

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

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

(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

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

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

- Le M-estimateur associé à ψ maximise la fonction

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

c'est-à-dire la log-vraisemblance. C'est l'estimateur du maximum de vraisemblance.

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

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

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.)

Choix de modèle statistique

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

$$\underbrace{0.92, -0.20, -1.80, 0.02, 0.49, 1.41, -1.59, -1.29, 0.34}_{\textit{tirage de } \mathcal{N}(0,1)}, {\color{red}100}.$$

- On prend $\mathbb{P}_\vartheta(dx) = f(x - \vartheta)dx$ pour deux f différents :
- f densité de la loi normale $\Rightarrow \widehat{\vartheta}_{\mathrm{n}}^{\mathtt{mv}} = \overline{X}_n = {\color{red}9.83}$.
- f densité de loi de Laplace $\Rightarrow$ tout point de l'intervalle $[0.02, 0.34]$ est un $\widehat{\vartheta}_{\mathrm{n}}^{\mathtt{mv}}$, en particulier, la médiane :

$$\widehat{\vartheta}_{\mathrm{n}}^{\mathtt{mv}} = M_n = (0.02 + 0.34)/2 = {\color{red}0.18}.$$

- Autre choix de modèle...

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

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