MAP 433 : Introduction aux méthodes statistiques. Cours 4

18 Septembre 2015

Aujourd'hui

1. *M*-estimation, rappel du Cours 3
   - Principe de maximum de vraisemblance
2. EMV, asymptotique des *Z*- et *M*- estimateurs
   - Approche asymptotique
3. Modèles réguliers et information de Fisher
   - Construction de l'information de Fisher
   - Modèle régulier
   - Cadre général et interprétation géométrique

M-estimation

- Situation : on observe $X_1, \ldots, X_n$ de loi $\mathbb{P}_\theta$ sur $\mathbb{R}$ et $\theta \in \Theta$.
- Principe : Se donner une application $\psi : \Theta \times \mathbb{R} \to \mathbb{R}_+$ telle que, pour tout $\theta \in \Theta \subset \mathbb{R}^d$,

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

admet un extremum (maximum ou minimum) en $\alpha = \theta$.

Definition

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

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

Au lieu de maximiser, on peut aussi minimiser

Un exemple classique : paramètre de localisation

- $\Theta = \mathbb{R}$, $\mathbb{P}_\theta(dx) = f(x-\theta)dx$, 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

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

- La fonction

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

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

- *M*-estimateur associé :

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

Paramètre de localisation

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

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

- 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 $\phi(\alpha, x) = \mathrm{Arctg}(x-\alpha)$. 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

$$\phi(\alpha, x) = \partial_\theta \psi(\alpha, x),$$

on a

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

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
- Procédure optimale (dans quel sens ?) sous des hypothèses de régularité de la famille $\{\mathbb{P}_\theta, \theta \in \Theta\}$.
- Parfois difficile à mettre en oeuvre en pratique $\rightarrow$ méthodes numériques, statistique computationnelle.

Fonction de vraisemblance

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

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

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

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

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

Exemples

- Exemple 1: Modèle de Poisson. On observe

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

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

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

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

- La fonction de vraisemblance associée s'écrit

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

$$= \frac{1}{\prod_{i=1}^n X_i!} e^{-n\theta} \theta^{\sum_{i=1}^n X_i}.$$

Exemples

- Exemple 2 Modèle de Cauchy. On observe

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

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

- On a alors

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

- La fonction de vraisemblance associée s'écrit

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

Estimateur du maximum de vraisemblance

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

Definition

On appelle estimateur du maximum de vraisemblance tout estimateur $\widehat{\theta}_n^{\,\mathtt{mv}}$ satisfaisant

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

- Existence, unicité...

Remarques

- Log-vraisemblance:

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

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

Max. vraisemblance = max. log-vraisemblance.

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

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

Exemple : modèle normal

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

- 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 $n \geq 2$):

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

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

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

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

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

Exemple : modèle de Laplace

L'expérience statistique est engendrée par un n-échantillon de loi de Laplace de paramètre $\theta \in \Theta = \mathbb{R}$. La 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 $\theta \rightsquigarrow \sum_{i=1}^n \left|X_i - \theta\right|$, dérivable presque partout de dérivée constante par morceaux. Equation de vraisemblance:

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

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

- n pair: $\widehat{\theta}_n^{\,\mathtt{mv}}$ 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: $\widehat{\theta}_n^{\,\mathtt{mv}} = X_{\left(\frac{n+1}{2}\right)}$, 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.

l'EMV est un M-estimateur

- On pose

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

(on suppose que $f(\alpha, \cdot) > 0$.)

- La fonction

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

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

- Le M-estimateur associé à ψ maximise la fonction

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

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

- C'est aussi un Z-estimateur si la fonction $\theta \rightsquigarrow \log f(\theta, \cdot)$ 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.)

Asymptotique des Z- et M-estimateurs

- Problème général délicat. Dans ce cours : conditions suffisantes.
- Convergence : critère simple pour les M-estimateurs.
- Vitesse de convergence : technique simple pour les Z-estimateurs, à condition de savoir que l'estimateur est convergent.
- Sous des hypothèses de régularité, un M-estimateur est un Z-estimateur.

Convergence des M-estimateurs

- Situation: on observe $X_1, \ldots, X_n$ i.i.d. de loi dans la famille $\{\mathbb{P}_\theta, \theta \in \Theta\}$.
- $\psi : \Theta \times \mathbb{R} \to \mathbb{R}$ fonction de contraste.
- Loi des grands nombres :

$$M_n(\alpha) = \frac{1}{n} \sum_{i=1}^{n} \psi(\alpha, X_i)$$

converge en $\mathbb{P}_\theta$-probabilité vers

$$M(\alpha, \theta) = \mathbb{E}_\theta \left[\psi(\alpha, X) \right]$$

qui atteint son maximum en $\alpha = \theta$

- à montrer :

$$\widehat{\theta}_n = \arg\max_{\alpha \in \Theta} M_n(\alpha) \xrightarrow{\mathbb{P}_\theta} \arg\max_{\alpha \in \Theta} \mathbb{E}_\theta \left[\psi(\alpha, X) \right] = \theta.$$

Exemple estimateur de translation

Convergence des M-estimateurs

Proposition

Si le M-estimateur $\widehat{\theta}_n$ associé à la fonction de contraste est bien défini et si

- $\sup_{\alpha \in \Theta} |M_n(\alpha) - M(\alpha, \theta)| \xrightarrow{\mathbb{P}_\theta} 0$,
- $\forall \varepsilon > 0, \ \sup_{|\alpha - \theta| \geq \varepsilon} M(\alpha, \theta) < M(\theta, \theta)$ *(condition de maximum)*

alors

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

- La condition 1 (convergence uniforme) peut être délicate à montrer...

Loi limite des Z-estimateurs

- Situation: on observe $X_1, \ldots, X_n$ i.i.d. de loi dans la famille $\{\mathbb{P}_\theta, \theta \in \Theta\}$, $\Theta \subset \mathbb{R}$.
- $\widehat{\theta}_n$: Z-estimateur associé à $\phi : \Theta \times \mathbb{R} \to \mathbb{R}$ vérifie

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

- Si $\widehat{\theta}_n$ est un M-estimateur associé à la fonction de contraste ψ régulière, alors c'est un Z-estimateur associé à la fonction $\phi(\alpha, x) = \partial_\theta \psi(\vartheta, x)$.
- On suppose $\widehat{\theta}_n$ convergent. Que dire de sa loi limite ?

Loi limite des Z-estimateurs : principe

- Loi des grands nombres

$$Z_n(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\phi(\alpha, X_i) \xrightarrow{\mathbb{P}_\theta} Z(\alpha,\theta) = \mathbb{E}_\theta\left[\phi(\alpha, X)\right]$$

- Principe. Développement de Taylor autour de θ :

$$0 = Z_n(\widehat{\theta}_n) = Z_n(\theta) + (\widehat{\theta}_n - \theta)Z_n'(\theta) + \tfrac{1}{2}(\widehat{\theta}_n - \theta)^2 Z''(\widetilde{\theta}_n).$$

- On néglige le reste :

$$\sqrt{n}(\widehat{\theta}_n - \theta) \approx \frac{-\sqrt{n}Z_n(\theta)}{Z_n'(\theta)}$$

Loi limite des Z-estimateurs : principe

- Convergence du numérateur

$$\sqrt{n}Z_n(\theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \phi(\theta, X_i) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}_\theta\left[\phi(\theta, X)^2\right]\right)$$

si $\mathbb{E}_\theta\left[\phi(\theta, X)\right] = 0$ et $\mathbb{E}_\theta\left[\phi(\theta, X)^2\right] < +\infty$.

- Convergence du dénominateur

$$Z_n'(\theta) = \frac{1}{n}\sum_{i=1}^n \partial_\theta\phi(\theta, X_i) \xrightarrow{\mathbb{P}_\theta} \mathbb{E}_\theta\left[\partial_\theta\phi(\theta, X)\right]$$

$\neq 0$ (à supposer).

- $+$ hypothèses techniques pour contrôler le reste (besoin de la convergence de $\widehat{\theta}_n$).

Loi limite des Z-estimateurs

Proposition (Convergence des Z-estimateurs)

- *Soit* Θ *un ouvert de* $\mathbb{R}$. *Pour tout* $\theta \in \Theta$, $\widehat{\theta}_n \xrightarrow{\mathbb{P}_\theta} \theta$, $\mathbb{E}_\theta\left[\phi(\theta, X)^2\right] < +\infty$ *et*

$$\mathbb{E}_\theta\left[\phi(\theta, X)\right] = 0,\ \mathbb{E}_\theta\left[\partial_\theta \phi(\theta, X)\right] \neq 0.$$

- *(Contrôle reste) pour tout* $\theta \in \Theta$, *pour tout* α *dans un voisinage de* θ,

$$|\partial_\theta^2 \phi(\alpha, x)| \leq g(x),\ \ \mathbb{E}_\theta\left[g(X)\right] < +\infty.$$

Alors

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}\Big(0, \frac{\mathbb{E}_\theta[\phi(\theta, X)^2]}{\left(\mathbb{E}_\theta[\partial_\theta \phi(\theta, X)]\right)^2}\Big).$$

Approche asymptotique

- Hypothèse simplificatrice : $\theta \in \Theta \subset \mathbb{R}$. On se restreint aux estimateurs asymptotiquement normaux c'est-à-dire vérifiant

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

  cf. théorèmes limites obtenus pour les Z-,M-estimateurs.

- Si $\widehat{\theta}_{n,1}$ et $\widehat{\theta}_{n,2}$ as. normaux de variance asymptotique $v_1(\theta) \leq v_2(\theta)$, alors la précision de $\widehat{\theta}_{n,1}$ est asymptotiquement meilleure que celle de $\widehat{\theta}_{n,2}$ au point θ :

$$\widehat{\theta}_{n,1} = \theta + \sqrt{\frac{v_1(\theta)}{n}}\xi^{(n)}$$
$$\widehat{\theta}_{n,2} = \theta + \sqrt{\frac{v_2(\theta)}{n}}\zeta^{(n)}$$

  où $\xi^{(n)}$ et $\zeta^{(n)} \xrightarrow{d} \mathcal{N}(0,1)$.

Comparaison d'estimateurs : cas asymptotique

- Si $v_1(\theta) < v_2(\theta)$, et si $\theta \rightsquigarrow v_i(\theta)$ est continue, on pose

$$\mathcal{C}_{n,\alpha}(\widehat{\theta}_{n,i}) = \left[\widehat{\theta}_{n,i} \pm \sqrt{\frac{v_i(\widehat{\theta}_{n,i})}{n}}\Phi^{-1}(1-\alpha/2)\right], \quad i = 1,2$$

  où $\alpha \in (0,1)$ et $\Phi(\cdot)$ est la fonction de répartition de la loi normale standard.

- $\mathcal{C}_{n,\alpha}(\widehat{\theta}_{n,i})$, $i = 1,2$ sont deux intervalles de confiance asymptotiquement de niveau $1-\alpha$ et on a

$$\frac{|\mathcal{C}_{n,\alpha}(\widehat{\theta}_{n,1})|}{|\mathcal{C}_{n,\alpha}(\widehat{\theta}_{n,2})|} \xrightarrow{\mathbb{P}_\theta^n} \sqrt{\frac{v_1(\theta)}{v_2(\theta)}} < 1.$$

- La notion de longueur minimale possible d'un intervalle de confiance est en général difficile à manipuler.

Conclusion provisoire

- Il est difficile en général de comparer des estimateurs.
- Cadre asymptotique + normalité asymptotique $\rightarrow$ comparaison de la variance asymptotique $\theta \rightsquigarrow v(\theta)$.
- Sous des hypothèses de régularité du modèle $\{\mathbb{P}_\theta^n, \theta \in \Theta\}$

alors

- Il existe une variance asymptotique $v^\star(\theta)$ minimale parmi les variances de la classe des M-estimateurs as. normaux.
- Cette fonction est associée à une quantité d'information intrinsèque au modèle.
- La variance asymptotique de l'EMV est $v^\star(\theta)$.
- Ceci règle partiellement le problème de l'optimalité.

Régularité d'un modèle statistique et information

- Cadre simplificateur : modèle de densité

$$X_1, \ldots, X_n \text{ i.i.d. de loi } \mathbb{P}_\theta$$

dans la famille $\{\mathbb{P}_\theta, \theta \in \Theta\}$ avec $\Theta \subset \mathbb{R}$ pour simplifier.

- Notation :

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

- Hypothèse : la quantité

$$\boxed{\mathbb{I}(\theta) = \mathbb{E}_\theta\left[\left(\partial_\theta \log f(\theta, X)\right)^2\right]}$$

est bien définie.

Information de Fisher

Definition

- $\mathbb{I}(\theta) = \mathbb{E}_\theta\left[\left(\partial_\theta \log f(\theta, X)\right)^2\right]$ *s'appelle l'information de Fisher de la famille* $\{\mathbb{P}_\theta, \theta \in \Theta\}$ *au point* θ. *Elle ne dépend pas de la mesure dominante* μ.
- *Le cadre d'intérêt est celui où*

$$0 < \mathbb{I}(\theta) < +\infty.$$

- $\mathbb{I}(\theta)$ *quantifie l'information qu'apporte chaque observation* X_i *sur le paramètre* θ.

Remarque : on a $\mathbb{P}_\theta\left[f(\theta, X) > 0\right] = 1$, donc la quantité $\log f(\theta, X)$ est bien définie.

Information dans quel sens ? Origine de la notion

- Supposons l'EMV $\widehat{\theta}_n^{\mathtt{mv}}$ bien défini et convergent.
- Supposons l'application $(\theta, x) \rightsquigarrow f(\theta, x)$ possédant toutes les propriétés de régularité et d'intégrabilité voulues.
- Alors

$$\boxed{\sqrt{n}\big(\widehat{\theta}_n^{\mathtt{mv}} - \theta\big) \xrightarrow{d} \mathcal{N}\Big(0, \frac{1}{\mathbb{I}(\theta)}\Big)}$$

en loi sous $\mathbb{P}_\theta$, où encore

$$\widehat{\theta}_n^{\mathtt{mv}} \overset{d}{\approx} \theta + \frac{1}{\sqrt{n\mathbb{I}(\theta)}}\mathcal{N}(0,1)$$

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

Construction de l'information + jeu d'hypothèses attenant

- Heuristique : on établira un jeu d'hypothèses justifiant a posteriori le raisonnement.
- Etape 1 : l'EMV $\widehat{\theta}_n^{\,\mathtt{mv}}$ converge :

$$\widehat{\theta}_n^{\,\mathtt{mv}} \xrightarrow{\mathbb{P}_\theta} \theta$$

  via le théorème de convergence des M-estimateurs.
- Etape 2 : l'EMV $\widehat{\theta}_n^{\,\mathtt{mv}}$ est un Z-estimateur :

$$0 = \partial_\alpha \Big(\sum_{i=1}^{n} \log f(\alpha, X_i) \Big)_{\alpha = \widehat{\theta}_n^{\,\mathtt{mv}}}.$$

Construction de $\mathbb{I}(\theta)$ cont.

- Etape 3 : développement asymptotique autour de θ :

$$0 \approx \sum_{i=1}^{n} \partial_\theta \log f(\theta, X_i) + (\widehat{\theta}_n^{\,\mathtt{mv}} - \theta) \sum_{i=1}^{n} \partial_\theta^2 \log f(\theta, X_i),$$

  soit

$$\widehat{\theta}_n^{\,\mathtt{mv}} - \theta \approx - \frac{\sum_{i=1}^n \partial_\theta \log f(\theta, X_i)}{\sum_{i=1}^n \partial_\theta^2 \log f(\theta, X_i)}$$

- Etape 4 : le numérateur. Normalisation et convergence de $\sum_{i=1}^n \partial_\theta \log f(\theta, X_i)$?

Numérateur

Lemme

On a

$$\mathbb{E}_\theta\left[\partial_\theta \log f(\theta, X)\right] = 0.$$

Proof.

$$\begin{aligned}
\mathbb{E}_\theta\left[\partial_\theta \log f(\theta, X)\right] &= \int_{\mathbb{R}} \partial_\theta \log f(\theta, x) f(\theta, x) \mu(dx) \\
&= \int_{\mathbb{R}} \frac{\partial_\theta f(\theta, x)}{f(\theta, x)} f(\theta, x) \mu(dx) \\
&= \int_{\mathbb{R}} \partial_\theta f(\theta, x) \mu(dx) \\
&= \partial_\theta \int_{\mathbb{R}} f(\theta, x) \mu(dx) = \partial_\theta 1 = 0.
\end{aligned}$$

Dénominateur

De même $\int_{\mathbb{R}} \partial_\theta^2 f(\theta, x)\mu(dx) = 0$. Conséquence :

$$\boxed{\mathbb{I}(\theta) = \mathbb{E}_\theta\left[\big(\partial_\theta \log f(\theta, X)\big)^2\right] = -\,\mathbb{E}_\theta\left[\partial_\theta^2 \log f(\theta, X)\right]}$$

En effet

$$\begin{aligned}
&\mathbb{E}_\theta\left[\partial_\theta^2 \log f(\theta, X)\right] = \\
&= \int_{\mathbb{R}} \frac{\partial_\theta^2 f(\theta,x) f(\theta,x) - \big(\partial_\theta f(\theta,x)\big)^2}{f(\theta,x)^2} f(\theta,x)\mu(dx) \\
&= \int_{\mathbb{R}} \partial_\theta^2 f(\theta,x)\mu(dx) - \int_{\mathbb{R}} \frac{\big(\partial_\theta f(\theta,x)\big)^2}{f(\theta,x)}\mu(dx) \\
&= 0 - \int_{\mathbb{R}} \left(\frac{\partial_\theta f(\theta,x)}{f(\theta,x)}\right)^2 f(\theta,x)\mu(dx) = -\,\mathbb{E}\left[\big(\partial_\theta \log f(\theta, X)\big)^2\right].
\end{aligned}$$

Conséquences

- Les $\partial_\theta \log f(\theta, X_i)$ sont i.i.d. et $\mathbb{E}_\theta\left[\partial_\theta \log f(\theta, X)\right] = 0$. TCL :

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \partial_\theta \log f(\theta, X_i) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}_\theta\left[\left(\partial_\theta \log f(\theta, X)\right)^2\right]\right)$$

$$= \mathcal{N}\left(0, \mathbb{I}(\theta)\right).$$

- Les $\partial_\theta^2 \log f(\theta, X_i)$ sont i.i.d. LGN :

$$\frac{1}{n}\sum_{i=1}^{n} \partial_\theta^2 \log f(\theta, X_i) \xrightarrow{\mathbb{P}_\theta} \mathbb{E}_\theta\left[\partial_\theta^2 \log f(\theta, X)\right]$$

$$\overset{\text{conséquence}}{=} -\mathbb{I}(\theta).$$

Conclusion

- En combinant les deux estimations + lemme de Slutsky :

$$\sqrt{n}(\widehat{\theta}_{\mathrm{n}}^{\,\mathtt{mv}} - \theta) \approx - \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \partial_\theta \log f(\theta, X_i)}{\frac{1}{n} \sum_{i=1}^{n} \partial_\theta^2 \log f(\theta, X_i)}$$
$$\xrightarrow{d} \frac{\mathcal{N}\big(0, \mathbb{I}(\theta)\big)}{\mathbb{I}(\theta)}$$
$$\overset{\text{loi}}{=} \mathcal{N}\Big(0, \frac{1}{\mathbb{I}(\theta)}\Big).$$

- Le raisonnement est rigoureux dès lors que : (i) on a la convergence de $\widehat{\theta}_{\mathrm{n}}^{\,\mathtt{mv}}$, (ii) on peut justifier le lemme et sa conséquence, (iii) $\mathbb{I}(\theta)$ est bien définie et non dégénérée et (iv) on sait contrôler le terme de reste dans le développement asymptotique, partie la plus difficile.

Modèle régulier

Definition

La famille de densités $\{f(\theta,\cdot),\theta\in\Theta\}$*, par rapport à la mesure dominante* μ*,* $\Theta\subset\mathbb{R}$*, est* ***régulière*** *si*

- Θ *ouvert et* $\{f(\theta,\cdot)>0\}=\{f(\theta',\cdot)>0\}$*,* $\forall\theta,\theta'\in\Theta$*.*
- μ*-p.p.* $\theta\rightsquigarrow f(\theta,\cdot)$*,* $\theta\rightsquigarrow\log f(\theta,\cdot)$ *sont* $\mathcal{C}^2$*.*
- $\forall\theta\in\Theta,\exists\mathcal{V}_\theta\subset\Theta$ *t.q. pour* $a\in\mathcal{V}_\theta$

$$|\partial_a^2\log f(a,x)|+|\partial_a\log f(a,x)|+\big(\partial_a\log f(a,x)\big)^2\leq g(x)$$

  où

$$\int_{\mathbb{R}}g(x)\sup_{a\in\mathcal{V}(\theta)}f(a,x)\mu(dx)<+\infty.$$

- *L'information de Fisher est non-dégénérée :*

$$\forall\theta\in\Theta,\ \ \mathbb{I}(\theta)>0.$$

Résultat principal

Proposition

- *Si l'expérience engendrée par l'observation* $X_1, \ldots, X_n \sim_{i.i.d.} \mathbb{P}_\theta$ *est associée à une famille de probabilités* $\{\mathbb{P}_\theta, \theta \in \Theta\}$ *sur* $\mathbb{R}$ *régulière au sens de la définition précédente, alors*

$$\sqrt{n}\big(\widehat{\theta}_{\mathsf{n}}^{\,\mathtt{mv}} - \theta\big) \xrightarrow{d} \mathcal{N}\Big(0, \frac{1}{\mathbb{I}(\theta)}\Big).$$

- *Si* $\widehat{\theta}_n$ *est un Z-estimateur régulier asymptotiquement normal de variance* $v(\theta)$*, alors*

$$\forall \theta \in \Theta, \quad v(\theta) \geq \frac{1}{\mathbb{I}(\theta)}.$$

Preuve de la proposition

- Le premier point consiste à rendre rigoureux le raisonnement précédent. Point délicat : le contrôle du terme de reste.
- Optimalité de la variance de l'EMV parmi celle des Z-estimateurs : on a vu que si $\widehat{\theta}_n$ est un Z-estimateur régulier associé à la fonction ϕ, alors, sa variance asymptotique $v(\theta) = v_\phi(\theta)$ vaut

$$v_\phi(\theta) = \frac{\mathbb{E}_\theta\left[\phi(\theta, X)^2\right]}{\left(\mathbb{E}_\theta\left[\partial_\theta \phi(\theta, X)\right]\right)^2}.$$

- A montrer : pour toute fonction ϕ :

$$\boxed{\frac{\mathbb{E}_\theta\left[\phi(\theta, X)^2\right]}{\left(\mathbb{E}_\theta\left[\partial_\theta \phi(\theta, X)\right]\right)^2} \geq \frac{1}{\mathbb{I}(\theta)}}.$$

Preuve de l'inégalité

- Par construction

$$\partial_a \mathbb{E}_\theta \left[\phi(a, X)\right]\Big|_{a=\theta} = 0.$$

- (avec $\dot\phi(\theta, x) = \partial_\theta \phi(\theta, x)$)

$$\begin{aligned} 0 &= \int_{\mathbb{R}} \left[\dot\phi(\theta, x) f(\theta, x) + \phi(\theta, x) \partial_\theta f(\theta, x)\right] \mu(dx) \\ &= \int_{\mathbb{R}} \left[\dot\phi(\theta, x) f(\theta, x) + \phi(\theta, x) \partial_\theta \log f(\theta, x) f(\theta, x)\right] \mu(dx). \end{aligned}$$

- Conclusion

$$\boxed{\mathbb{E}_\theta \left[\dot\phi(\theta, X)\right] = -\,\mathbb{E}_\theta \left[\phi(\theta, X) \partial_\theta \log f(\theta, X)\right]}$$

Preuve de l'inégalité (fin)

- On a
$$\mathbb{E}_\theta\left[\dot{\phi}(\theta,X)\right] = -\,\mathbb{E}_\theta\left[\phi(\theta,X)\partial_\theta \log f(\theta,X)\right]$$
- Cauchy-Schwarz :
$$\left(\mathbb{E}_\theta\left[\dot{\phi}(\theta,X)\right]\right)^2 \leq \mathbb{E}_\theta\left[\phi(\theta,X)^2\right]\mathbb{E}_\theta\left[\left(\partial_\theta \log f(\theta,X)\right)^2\right],$$
c'est-à-dire
$$\boxed{v_\phi(\theta)^{-1} = \frac{\left(\mathbb{E}_\theta\left[\dot{\phi}(\theta,X)\right]\right)^2}{\mathbb{E}_\theta\left[\phi(\theta,X)^2\right]} \leq \mathbb{I}(\theta).}$$

Information de Fisher dans un modèle général

Definition

- *Situation : suite d'expériences statistiques*

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

dominées par μ_n, associées à l'observation $Z^{(n)}$,

$$f_n(\theta, z) = \frac{d\,\mathbb{P}_\theta^n}{d\mu^n}(z), \quad z \in \mathfrak{Z}^n, \theta \in \Theta \subset \mathbb{R}.$$

- *Information de Fisher (si elle existe) de l'expérience au point θ :*

$$\mathbb{I}(\theta \,|\, \mathcal{E}_n) = \mathbb{E}_\theta^n\left[\left(\partial_\theta \log f_n(\theta, Z^{(n)})\right)^2\right]$$

Le cas multidimensionnel

- Même contexte que précédemment, avec $\Theta \subset \mathbb{R}^d$, et $d \geq 1$.
- Matrice d'information de Fisher

$$\mathbb{I}(\theta) = \mathbb{E}_\theta\left[\nabla_\theta \log f(\theta, Z^n) \nabla_\theta \log f(\theta, Z^n)^T\right]$$

  matrice symétrique positive.
- Si $\mathbb{I}(\theta)$ définie et si $\mathcal{E}^n$ modèle de densité, en généralisant à la dimension d les conditions de régularité, on a

$$\sqrt{n}\left(\widehat{\theta}_n^{\,\mathtt{mv}} - \theta\right) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{I}(\theta)^{-1}\right).$$

Interprétation géométrique

- On pose $\mathbb{D}(\alpha,\theta) = \mathbb{E}_\theta\big[\log f(\alpha, X)\big]$. On a vu (inégalité d'entropie) que

$$\begin{aligned}\mathbb{D}(\alpha,\theta) &= \int_{\mathbb{R}} \log f(\alpha,x) f(\theta,x)\mu(dx)\\ &\leq \int_{\mathbb{R}} \log f(\theta,x) f(\theta,x)\mu(dx) = \mathbb{D}(\theta,\theta).\end{aligned}$$

- On a

$$\boxed{\mathbb{I}(\theta) = \partial^2_\alpha \mathbb{D}(\alpha,\theta)\big|_{\alpha=\theta}.}$$

  - Si $\mathbb{I}(\theta)$ est petite , le rayon de courbure de $\alpha \rightsquigarrow \mathbb{D}(\alpha,\theta)$ est grand dans un voisinage de θ : la stabilisation d'un maximum empirique (l'EMV) est plus difficile, rendant moins précis l'estimation.
  - Si $\mathbb{I}(\theta)$ est grande , le rayon de courbure est petit et le maximum de l'EMV est mieux localisé.

Efficacité à un pas

- Dans un modèle régulier, le calcul numérique de l'EMV peut être difficile à réaliser.
- Si l'on dispose d'un estimateur $\widehat{\theta}_n$ asymptotiquement normal et si les évaluations

$$\ell'_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \partial_\theta \log f(\theta, X_i), \quad \ell''_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \partial_\theta^2 \log f(\theta, X_i)$$

sont faciles, alors on peut corriger $\widehat{\theta}_n$ de sorte d'avoir le même comportement asymptotique que l'EMV :

$$\widetilde{\theta}_n = \widehat{\theta}_n - \frac{\ell'_n(\widehat{\theta}_n)}{\ell''_n(\widehat{\theta}_n)} \quad \text{(algorithme de Newton)}$$

satisfait

$$\boxed{\sqrt{n}\big(\widetilde{\theta}_n - \theta\big) \xrightarrow{d} \mathcal{N}\Big(0, \frac{1}{\mathbb{I}(\theta)}\Big)}$$