Estimateurs obtenus par maximisation ou annulation d'un critère empirique

Lionel Truquet,
lionel.truquet@ensai.fr

1. Maximisation ou annulation d'un critère
   - M-estimateurs
   - Z-estimateurs

2. Consistance
   - Consistance des $M-$estimateurs
   - Consistance des $Z-$estimateurs

3. Normalité asymptotique

Outline

1. Maximisation ou annulation d'un critère
   - M-estimateurs
   - Z-estimateurs
2. Consistance
   - Consistance des $M-$estimateurs
   - Consistance des $Z-$estimateurs
3. Normalité asymptotique

Notations

On dispose de n observations $x_1, \ldots, x_n$ d'un même phénomène. On suppose que ces valeurs sont des éléments d'un espace mesurable $(G, \mathcal{G})$, le cas $G = \mathbb{R}^d$ permet de fixer les idées.

- On considère que ces valeurs sont des réalisations de n variables aléatoires i.i.d. dont la loi dépend d'un paramètre $\theta \in \Theta$ que l'on cherche à estimer. $\Theta \subset \mathbb{R}^k$ le plus souvent dans ce qui suit.

- Pour cela, on se donne une famille de mesures de probabilité $\mathcal{P}$ sur $(G, \mathcal{G})$. Le triplet $(G^n, \mathcal{G}^{\otimes}, \{P^{\otimes n} : P \in \mathcal{P}\})$ est appelé modèle statistique. On posera $X_i(\omega) = \omega_i$ pour $1 \leq i \leq n$.

- On suppose que $\theta = \theta(P)$ est le paramètre à estimer. On supposera que $(x_1, \ldots, x_n)$ est une réalisation de $P_0 \in \mathcal{P}$ et notera $\theta_0 = \theta(P_0)$.

- On admettra que pour tout $P \in \mathcal{P}$, il existe une unique mesure de probabilité $\mathbb{P}^\infty$ sur G^∞ telle que pour tout $n \geq 1$ et $A_1, \ldots, A_n \in \mathcal{G}$,

$$\mathbb{P}^\infty \left(X_1 \in A_1, \ldots, X_n \in A_n\right) = \prod_{i=1}^n P(A_i).$$

On notera $\mathbb{P} = \mathbb{P}_0^\infty$.

Outline

1. Maximisation ou annulation d'un critère
   - M-estimateurs
   - Z-estimateurs
2. Consistance
   - Consistance des $M-$estimateurs
   - Consistance des $Z-$estimateurs
3. Normalité asymptotique

Contexte général

- On se donne une famille de fonctions mesurables $\{m_\theta : G \to \mathbb{R}, \theta \in \Theta\}$. Un estimateur $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ est obtenu en maximisant

$$\theta \mapsto M_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} m_\theta(X_i).$$

- Pour que cette approche soit pertinente, il faut que la fonction $\theta \mapsto \mathbb{E}[m_\theta(X_1)]$ soit maximale en $\theta = \theta_0$. On espère ainsi que $\hat{\theta}_n$ ne soit pas trop éloigné de θ_0 (loi des grands nombres).

- L'exemple le plus simple est celui de la moyenne empirique $\overline{X}_n$ qui est un M-estimateur pour $m_\theta(x) = -(x-\theta)^2$. C'est plus simplement un estimateur obtenu par la méthode des moments.

- L'estimateur du maximum de vraisemblance où $\mathcal{P} = \{p_\theta(x)\mu(dx) : \theta \in \Theta\}$ et $m_\theta(x) = \log p_\theta(x)$.

Outline

Définition

- Il s'agit de définir des estimateurs qui annule $Z_n : \theta \mapsto \frac{1}{n}\sum_{i=1}^n z_\theta(X_i)$. Si l'estimateur θ a k coordonnées, z_θ a souvent k fonctions coordonnées.
- Etudier un $M-$estimateur peut se ramener à étudier un $Z-$estimateur lorsque $\theta \mapsto m_\theta(x)$ est differentiable pour tout x. On a alors $z_\theta(x) = \dot{m}_\theta(x)$ (gradient de $\theta \mapsto m_\theta(x)$ au point θ).
- On peut avoir des M-estimateurs qui ne s'associe pas directement à un $Z-$estimateur. C'est le cas, par exemple lorsque $m_\theta(x) = \log \mathbb{1}_{[0,\theta]}(x) - \log(\theta)$ (max. de vraisemblance lorsque la loi est la loi uniforme sur $[0,\theta]$).
- Estimateurs de localisation. Correspond au cas $z_\theta(x) = z(x-\theta)$. Par example,
    - $z(y) = y$ (estimation de la moyenne),
    - $z(y) = sign(y)$ (estimation de la médiane),
    - $z(y) = -k\mathbb{1}_{y\leq -k} + y\mathbb{1}_{|y|<k} + k\mathbb{1}_{y\geq k}$ (estimateurs de Huber).

  On définit $sign(y) = -\mathbb{1}_{y<0} + \mathbb{1}_{y>0}$.

Outline

Outline

Consistance faible

Definition 1

On dira que $\hat{\theta}_n$ est faiblement consistant si $\hat{\theta}_n \xrightarrow{P} \theta_0$.

On pose $M(\theta) = \mathbb{E}m_\theta(X_1)$.

Theorem 1

On suppose que

1. $\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{P} O$,
2. $\forall \epsilon > 0$, $\sup_{\theta:d(\theta,\theta_0)\geq\epsilon} M(\theta) < M(\theta_0)$.

Alors toute suite d'estimateurs telle que $M_n\left(\hat{\theta}_n\right) \geq \sup_{\theta\in\Theta} M_n(\theta) - o_{\mathbb{P}}(1)$ est faiblement consistante.

- La condition 1 est une loi (faible) des grands nombres uniforme.
- Si θ est compact M est continue, alors la condition 2 est vérifiée. Cette condition demande une "bonne séparation" du point maximum.

Consistance forte

Definition 2

On dira que $\hat{\theta}_n$ *est fortement consistant si* $\lim_{n\to\infty} \hat{\theta}_n = \theta_0$ *p.s.*

Theorem 2

On suppose

1. Θ *compact,*
2. $\theta \mapsto m_\theta(x)$ *continue pour presque tout* x *et* $\mathbb{E} \sup_{\theta\in\Theta} m_\theta(X_1) < \infty$*,*
3. $\theta \mapsto M(\theta)$ *est maximisée en* θ_0 *et* $M(\theta) = M(\theta_0) \Rightarrow \theta = \theta_0$*.*

Alors si $\hat{\theta}_n = \arg\max_{\theta\in\Theta} M_n(\theta)$, $\hat{\theta}_n$ *est fortement consistant.*

La preuve est basé sur le lemme suivant.

Lemma 1

Si les conditions 1 *et* 2 *du théorème précédent sont vérifiées, alors* $\lim_{n\to\infty} \sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| = 0$ *p.s.*

Cas du maximum de vraisemblance, condition d'identifiabilté

- On suppose ici que $m_\theta(x) = \log \frac{p_\theta}{p_{\theta_0}}(x)$ et p_{θ_0} désigne la densité des observations par rapport à la mesure de référence μ.

- Maximiser la vraisemblance est équivalent à la maximisation de $M_n(\theta) = \frac{1}{n}\sum_{i=1}^n m_\theta(X_i)$.

- Si $dP = pd\mu$ et $dQ = qd\mu$, la quantité $K(P,Q) = \int_{p>0} \log\left(\frac{p}{q}\right) pd\mu$ est appelée divergence de Kullback entre P et Q. Il s'agit d'une quantité positive, éventuellement égale à $+\infty$. Ce n'est pas une distance, mais elle caractérise la déviation de Q par rapport à P.

  Ici $M(\theta)$ est l'opposé de la divergence de Kullback entre $p_{\theta_0}d\mu$ et $p_\theta d\mu$.

Lemma 2

Supposons que si $\theta \neq \theta_0$, alors $\mu\left(\{p_{\theta_0} \neq p_\theta\}\right) > 0$. Alors M atteint un unique maximum en θ_0.

Outline

Consistance faible

Un zéro de $\theta \mapsto Z_n(\theta)$ est un maximiseur de $\theta \mapsto -||Z_n(\theta)||$. Le résultat suivant est alors immédiat.

Theorem 3

On suppose les deux conditions suivantes satisfaites.

1. $\sup_{\theta \in \Theta} ||Z_n(\theta) - Z(\theta)|| = o_P(1)$.
2. *Pour tout* $\epsilon > 0$, $\inf_{\theta : d(\theta,\theta_0) > \epsilon} ||Z(\theta)|| > 0 = ||Z(\theta_0)||$.

Alors tout estimateur $\hat{\theta}_n$ *tel que* $Z_n\left(\hat{\theta}_n\right) = o_P(1)$ *est faiblement consitant.*

On peut aussi énoncer des résultats de consistance forte en rajoutant de la compacité comme pour les M-estimateurs.

Example de la médiane

Si $z(x) = sign(x)$, on suppose que $\theta \mapsto Z(\theta)$ admet un zéro seulement en θ_0. Montrer la consistance du $Z-$estimateur. On pourra utiliser le lemme suivant.

Lemma 3

Si $X_1, X_2, \ldots$ sont i.i.d. de fonction de répartition F et $\mathbb{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \leq x}$, alors $\lim_{n\to\infty} \sup_{x\in\mathbb{R}} |\mathbb{F}_n(x) - F(x)| = 0$ p.s.

Il est aussi possible de voir la médiane empirique comme un $M-$estimateur, avec $m_\theta(x) = |x - \theta|$.

Exercices

1. Lorsque $p_\theta(x) = \left(1 + (x-\theta)^2\right)^{-1}$ pour $x \in \mathbb{R}$ (loi de Cauchy), prouver la consistance du maximum de vraisemblance lorsque Θ est un compact de $\mathbb{R}$.

2. Considérons le modèle de régression non linéaire $Y_i = g_\theta(W_i) + \varepsilon_i$ avec $((W_i, \varepsilon_i))_{1 \leq i \leq n}$ i.i.d. telle que $\mathbb{E}\varepsilon_1^2 < \infty$ et $\mathbb{E}\left(\varepsilon_i | W_i\right) = 0$. Lorsque les observations sont $X_i = (Y_i, W_i)$ pour $1 \leq i \leq n$, discuter la consistance d'un $M-$estimateur bien choisi.

Outline

Résultat pour les Z–estimateurs

Pour une fonction $f : \mathbb{R}^k \to \mathbb{R}^k$, on note par $\dot{f}(x)$ la matrice différentielle de f au point x.

Theorem 4

1. *Pour tout x, $\theta \mapsto z_\theta(x)$ est continûment différentiable.*
2. $\mathbb{E}z_{\theta_0}(X_1) = 0$, $\mathbb{E}\|z_{\theta_0}(X_1)\|^2 < \infty$. *De plus,* $\mathbb{E}\dot{z}_{\theta_0}(X_1)$ *existe (intégrabilité) et est inversible.*
3. *Il existe une fonction positive α tel que $\mathbb{E}\alpha(X_1) < \infty$ ainsi qu'un voisinage $\mathcal{V}_{\theta_0}$ de θ_0 tel que pour tout $\theta \in \mathcal{V}_{\theta_0}$ et tout x,*

$$\left|\frac{\partial z_{\theta,h}(x)}{\partial \theta_i}\right| \leq \alpha(x), \quad 1 \leq i, h \leq k.$$

Alors toute suite d'estimateurs $\hat{\theta}_n$ consistante et telle que $Z_n(\hat{\theta}_n) = o_{\mathbb{P}}(1/\sqrt{n})$ vérifie

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) = -\left(\mathbb{E}\dot{z}_{\theta_0}(X_1)\right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} z_{\theta_0}(X_i) + o_P(1) \Rightarrow \mathcal{N}_k\left(0, W_{\theta_0}^{-1} V_{\theta_0} W_{\theta_0}^{-1}\right),$$

avec $W_{\theta_0} = \mathbb{E}\dot{z}_{\theta_0}(X_1)$, $V_{\theta_0} = \mathbb{E}z_{\theta_0}(X_1) z_{\theta_0}(X_1)'$.

Application au M-estimateurs

On donne maintenant une conséquence pour les M-estimateurs dans le cadre compact.

Theorem 5

1. Θ *est un compact de* $\mathbb{R}^k$ *qui contient* θ_0 *dans son intérieur.*

2. $\theta \mapsto \mathbb{E} m_\theta(X_1)$ *est uniquement maximisée en* $\theta = \theta_0$ *et* $\sup_{\theta\in\Theta} |m_\theta(X_1)|$ *est intégrable.*

3. $\theta \mapsto m_\theta(x)$ *est deux fois continûment différentiable. De plus, il existe un voisinnage* $\mathcal{V}_{\theta_0}$ *de* θ_0 *tel que* $\sup_{\theta\in\mathcal{V}_{\theta_0}} \|\ddot{m}_\theta(X_1)\|$ *est integrable.*

4. $\dot{m}_{\theta_0}(X_1)$ *est de carré intégrable et* $\mathbb{E}\left[\ddot{m}_{\theta_0}(X_1)\right]$ *est inversible.*

Alors, $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \Rightarrow \mathcal{N}_k\left(0, W_{\theta_0}^{-1} V_{\theta_0} W_{\theta_0}^{-1}\right)$ *avec*

$$W_{\theta_0} = \mathbb{E}\ddot{m}_{\theta_0}(X_1), \quad V_{\theta_0} = \mathbb{E}\dot{m}_{\theta_0}(X_1)\dot{m}_{\theta_0}(X_1)'.$$

Le cas de l'estimateur du maximum de vraisemblance (EMV)

- Correspond au cas $m_\theta(x) = \log p_\theta(x)$.

- Sous certaines conditions de régularité et d'intégrabilité pour $\theta \mapsto m_\theta(x)$ et ses dérivées, on a normalité asymptotique d'une suite consistante de $M-$estimateurs.

- On a aussi l'égalité $W_{\theta_0} = -V_{\theta_0}$ et la valeur commune, notée I_{θ_0} est appelée information de Fisher au point θ_0. On a alors

$$I_{\theta_0} = \mathbb{E}\left(\frac{\dot{p}_{\theta_0}(X_1)\dot{p}_{\theta_0}(X_1)'}{p_{\theta_0}(X_1)^2}\right) = \mathbb{E}\left(\frac{\ddot{p}_{\theta_0}(X_1)}{p_{\theta_0}(X_1)}\right).$$

- I_{θ_0} s'interprète comme la courbure moyenne de l'opposé de la log-vraisemblance. Plus cette quantité est grande, plus la variance asymptotique de l'EMV est faible. Intuitivement, θ_0 peut être estimé plus précisément.

Deux remarques sur l'EMV

- Dans le cas où les données suivent une loi uniforme sur l'intervalle $[0, \theta_0]$ avec θ_0 inconnu, les conditions de régularité précédentes ne sont pas satisfaites. L'EMV est $\hat{\theta}_n = \max_{1 \le i \le n} X_i$ et $n\left(\theta_0 - \hat{\theta}_n\right) \Rightarrow \mathcal{E}(1/\theta_0)$. Le comportement de l'EMV n'est donc pas standard.

- Si la densité p des données n'est pas dans $\{p_\theta : \theta \in \Theta\}$, et si les conditions de régularité sont satisfaites, il est toujours possible de montrer que $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right)$ est asymptotiquement Gaussien, où θ_0 est l'unique minimiseur de $\theta \mapsto -\mathbb{E} \log \frac{p_\theta(X_1)}{p(X_1)}$. On peut interpréter p_{θ_0} comme une projection de la densité sur le modèle choisi et $p_{\hat{\theta}_n}$ est alors une raisonnable approximation de cette projection.

Example 1 : loi de Weibull

$$p_{\lambda,\sigma}(x) = \frac{\lambda}{\sigma} x^{\lambda-1} \exp\left(-\frac{x^{\lambda}}{\sigma}\right), \quad x > 0, \lambda > 0, \sigma > 0.$$

Vérifier la consistance et la normalité asymptotique de l'EMV lorsque θ_0 est dans l'intérieur d'un compact Θ de $]0, \infty[^2$.

Example 2 : la régression logistique

On suppose observé un échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ de variables aléatoires à valeurs dans $\{0, 1\} \times \mathbb{R}^k$ et tel que

$$\mathbb{P}\left(Y_i = 1 | X_i = x\right) = F\left(\theta_0' x\right)$$

avec $F(x) = \frac{\exp(x)}{1+\exp(x)}$.

Vérifier la consistance et la normalité asymptotique de l'EMV lorsque θ_0 est dans l'intérieur d'un compact Θ de $\mathbb{R}^k$.