Problèmes inverses
Partie 2 — Chapitre 3
Estimation par maximum de vraisemblance

Nicolas Champagnat

Ecole des Mines de Nancy, 05/05/2020

Modèle discret, modèle continu

- On considère le modèle statistique paramétrique

$$\left(\mathcal{H}^n, \{\mathbb{P}_\theta\}_{\theta\in\Theta}\right).$$

- On suppose que le paramètre d'intérêt est θ (cf. δ-méthode du chapitre précédent pour en déduire le cas général)

On se restreindra à deux cas :

- Le cas discret, où $\mathcal{H}$ est un ensemble fini ou dénombrable : $\mathbb{P}_\theta$ est caractérisée par les probabilités des événements élémentaires

$$\mathbb{P}_\theta\{(X_1,\ldots,X_n)=(x_1,\ldots,x_n)\},\quad \forall (x_1,\ldots,x_n)\in\mathcal{H}^n.$$

- Le cas continu, où $\mathcal{H}\subset\mathbb{R}^k$ et on suppose que $\mathbb{P}_\theta$ est absolument continue par rapport à la mesure de Lebesgue : $\mathbb{P}_\theta$ est caractérisée par sa densité

$$f_\theta(x_1,\ldots,x_n),\quad \forall (x_1,\ldots,x_n)\in\mathcal{H}^n.$$

Un exemple discret

- Au jeu de pile-ou-face, supposons qu'on observe les trois lancers PPF, c-à-d $(x_1, x_2, x_3) = (0, 0, 1)$
- Pour la valeur $\theta_1 = 1/2$ du paramètre,

$$\mathbb{P}_{\theta_1}((X_1, X_2, X_3) = (0, 0, 1)) = \left(\frac{1}{2}\right)^3 = \frac{1}{8}.$$

- Pour la valeur $\theta_2 = 1/4$, cette probabilité est

$$\mathbb{P}_{\theta_2}((X_1, X_2, X_3) = (0, 0, 1)) = \frac{1}{4} \times \frac{1}{4} \times \frac{3}{4} = \frac{3}{64}.$$

- Puisque $3/64 < 1/8$, la valeur θ_1 du paramètre est plus vraisemblable que la valeur θ_2.
- Ceci revient à étudier la fonction $\theta \mapsto \mathbb{P}_\theta((X_1, X_2, X_3) = (0, 0, 1))$, et comparer ses valeurs en θ_1 et θ_2 .

Un exemple continu

- Considérons le modèle statistique gaussien d'échantillon de taille 1

$$(\mathbb{R}, \{\mathcal{N}(\theta, 1)\}_{\theta \in \mathbb{R}}).$$

- Supposons que l'observation x est 0.
- On a

$$\mathbb{P}_\theta(X \in [0, dx]) = f_\theta(0)\, dx = \frac{1}{\sqrt{2\pi}} e^{-\frac{\theta^2}{2}}\, dx,$$

où dx peut-être interprété comme une incertitude sur l'observation.

- La valeur $\theta_1 = 0$ du paramètre donne une probabilité $1/\sqrt{2\pi}\, dx$ aux observations.
- La valeur $\theta_2 = 1$ donne une probabilité $e^{-1/2}/\sqrt{2\pi}\, dx$.
- Ainsi, la valeur θ_1 du paramètre est plus vraisemblable que la valeur θ_2.
- Ceci revient à comparer les densités des lois du modèle statistique, évaluées sur les observations (ici $x = 0$), c-à-d étudier la fonction $\theta \mapsto f_\theta(0)$.

Vraisemblance

Définition

On considère le modèle statistique $(\mathcal{H}^n, \{\mathbb{P}_\theta\}_{\theta\in\Theta})$ *et une observation* $(x_1, \ldots, x_n) \in \mathcal{H}^n$.

- *Dans le cas discret, la vraisemblance de l'observation* $(x_1, \ldots, x_n)$ *est l'application de* Θ *dans* $[0, 1]$ *définie par*

$$\theta \mapsto \mathbb{P}_\theta((X_1, \ldots, X_n) = (x_1, \ldots, x_n)).$$

- *Dans le cas continu, la vraisemblance de l'observation* $(x_1, \ldots, x_n)$ *est l'application de* Θ *dans* $[0, 1]$ *définie par*

$$\theta \mapsto f_\theta(x_1, \ldots, x_n),$$

où f_θ *est la densité de la loi* $\mathbb{P}_\theta$.

Dans tous les cas, on note $\theta \mapsto L_n(x_1, \ldots, x_n; \theta)$ *la fonction de vraisemblance définie ci-dessus.*

Notations

On définit également

$$\theta \mapsto \ell_n(x_1, \ldots, x_n; \theta) = \log L_n(x_1, \ldots, x_n; \theta)$$

la fonction de log-vraisemblance.

Pour alléger les écritures, on notera $\mathbf{X} = (X_1, \ldots, X_n)$ et $\mathbf{x} = (x_1, \ldots, x_n)$.
Ainsi, la fonction de vraisemblance s'écrira $L_n(\mathbf{x}; \cdot)$.

Exemples

- Dans le jeu de pile-ou-face

$$L_n(\mathbf{x};\theta) = \theta^{\text{nombre de pile}}(1-\theta)^{\text{nombre de face}} = \theta^{n\bar{x}_n}(1-\theta)^{n(1-\bar{x}_n)}, \quad (1)$$

où $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ est la moyenne empirique des observations.

- Dans le modèle statistique

$$\left(\mathbb{R}^n, \left\{\mathcal{N}(m,\sigma^2)^{\otimes n}\right\}_{m\in\mathbb{R},\ \sigma>0}\right),$$

la vraisemblance de $x_1,\ldots,x_n$ s'écrit

$$L_n(\mathbf{x};\theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x_i-m)^2}{2\sigma^2}} = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\sum_{i=1}^n (x_i-m)^2}{2\sigma^2}\right).$$

Positivité de la vraisemblance

Proposition

Pour tout $\theta \in \Theta$,

$$L_n(\mathbf{X};\theta) = L_n(X_1,\ldots,X_n;\theta) > 0, \quad \mathbb{P}_\theta\text{-}p.s.$$

Démonstration :

- Dans le cas discret, le résultat est évident.
  En effet, pour tout $(x_1,\ldots,x_n) \in \mathcal{H}^n$, $L_n(\mathbf{x};\theta) > 0$ ssi $\mathbb{P}_\theta(\mathbf{X}=\mathbf{x}) > 0$, puisque les deux quantités sont égales.
- Dans le cas continu,

$$\mathbb{P}_\theta(L_n(\mathbf{X};\theta)=0) = \int_{\{L_n(\cdot;\theta)=0\}} f_\theta(\mathbf{x})\,d\mathbf{x} = \int_{\{L_n(\cdot;\theta)=0\}} L_n(\mathbf{x};\theta)\,d\mathbf{x} = 0.$$

Cas i.i.d. : une propriété évidente mais utile

Proposition

Si pour tout $\theta \in \Theta$, $\mathbb{P}_\theta = \mathbb{Q}_\theta^{\otimes n}$, *alors*

$$L_n(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} L(x_i; \theta) \quad et \quad \ell_n(x_1, \ldots, x_n; \theta) = \sum_{i=1}^{n} \ell(x_i; \theta),$$

où $L(x; \theta) = L_1(x; \theta)$ *et* $\ell(x; \theta) = \ell_1(x; \theta)$.

Estimateur du maximum de vraisemblance (EMV)

Abréviation EMV : pour « estimation par maximum de vraisemblance » ou « estimateur du maximum de vraisemblance » (en anglais, MLE=maximum likelihood estimator).

Intuition : la probabilité sous $\mathbb{P}_\theta$ des observations $(x_1, \ldots, x_n)$ doit être élevée pour θ proche de θ_0, la valeur réelle du paramètre.

Définition

Un estimateur du maximum de vraisemblance (EMV) de θ est un estimateur $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ tel que

$$L_n(X_1, \ldots, X_n; \hat{\theta}) = \sup_{\theta \in \Theta} L_n(X_1, \ldots, X_n; \theta).$$

Attention ! pas nécessairement unicité d'un EMV.

Considérations pratiques

- La construction d'un EMV nécessite de trouver un argmax de la vraisemblance, c'est-à-dire la solution d'un problème d'optimisation. Cette solution n'est généralement pas explicite, et nécessite dans les cas pratiques de réaliser une optimisation numérique, par ex. à l'aide des algorithmes de Newton-Raphson ou de Gauss-Newton.
- La vraisemblance n'est pas toujours explicite dans les modèles compliqués, où la denstié des observations est difficile à évaluer $\rightsquigarrow$ utiliser des méthodes d'approximation de densité, comme par exemple l'algorithme de Metropolis-Hastings ou les méthodes MCMC (Markov Chain Monte Carlo).

Cas i.i.d.

Dans le cas d'échantillons i.i.d. ($\mathbb{P}_\theta = \mathbb{Q}_\theta^{\otimes n}$ pour tout $\theta \in \Theta$), on préfère généralement chercher un EMV $\hat{\theta}$ en maximisant la log-vraisemblance :

$$\begin{aligned} \ell_n(X_1, \ldots, X_n; \hat{\theta}) &= \sup_{\theta \in \Theta} \ell_n(X_1, \ldots, X_n; \theta) \\ &= \sup_{\theta \in \Theta} \sum_{i=1}^{n} \ell(X_i; \theta). \end{aligned}$$

En effet, les techniques de calcul différentiel sont souvent plus simples à mettre en œuvre pour des sommes de fonctions que pour des produits.

Exemple du jeu de pile-ou-face

- D'après (1), la log-vraisemblance est donnée par

$$\ell_n(x_1, \ldots, x_n; \theta) = n\bar{x}_n \log\theta + n(1 - \bar{x}_n)\log(1 - \theta).$$

- Ainsi,

$$\frac{\partial}{\partial\theta}\ell_n(x_1, \ldots, x_n; \theta) = \frac{n\bar{x}_n}{\theta} - \frac{n(1 - \bar{x}_n)}{1 - \theta}$$

  s'annule ssi $\theta = \bar{x}_n$, est positive avant et négative après.
- La log-vraisemblance est donc maximale en $\bar{x}_n$, et il existe un unique EMV

$$\hat{\theta}_n = \bar{X}_n.$$

- Remarque : c'est le même estimateur que par la méthode des moments (cf. chapitre précédent).

Information de Kullback-Leibler

Objectif : étudier la consistance de l'EMV.

Définition

Pour tout $\alpha, \theta \in \Theta$, l'information de Kullback-Leibler (ou divergence de Kullback-Leibler, ou entropie relative) entre $\mathbb{P}_\alpha$ et $\mathbb{P}_\theta$ est

$$K_n(\alpha, \theta) = -\mathbb{E}_\theta \log \frac{L_n(\mathbf{X}; \alpha)}{L_n(\mathbf{X}; \theta)} = \mathbb{E}_\theta \left[\ell_n(\mathbf{X}; \theta) - \ell_n(\mathbf{X}; \alpha)\right]$$

si $\ell_n(\mathbf{X}; \alpha)$ et $\ell_n(\mathbf{X}; \alpha)$ appartiennent à $\mathbb{L}^1(\mathbb{P}_\theta)$ et $\mathbb{P}_\alpha$ est absolument continue par rapport à $\mathbb{P}_\theta$
et $K_n(\alpha, \theta) = +\infty$ sinon.

Remarque : lorsque $\mathbb{P}_\alpha$ est absolument continue par rapport à $\mathbb{P}_\theta$, la densité de $\mathbb{P}_\alpha$ par rapport à $\mathbb{P}_\theta$ est donnée par

$$\frac{L_n(\mathbf{x}; \alpha)}{L_n(\mathbf{x}; \theta)}, \quad \forall \mathbf{x} \in \mathcal{H}^n.$$

Exemple

Dans le jeu de pile-ou-face, d'après la formule (1),

$$K_n(\alpha,\theta) = \mathbb{E}_\theta\left[n\bar{X}_n \log\frac{\theta}{\alpha} + n(1-\bar{X}_n)\log\frac{1-\theta}{1-\alpha}\right]$$
$$= n\theta\log\frac{\theta}{\alpha} + n(1-\theta)\log\frac{1-\theta}{1-\alpha}.$$

Propriété fondamentale

L'information de Kullback-Leibler est une mesure de dissimilarité entre les lois $\mathbb{P}_\alpha$ et $\mathbb{P}_\theta$:

Proposition

Pour tout $\alpha, \theta \in \Theta$, $K_n(\alpha, \theta) \geq 0$.
Si de plus le modèle statistique est identifiable, alors $K_n(\alpha, \theta) = 0$ *si et seulement si* $\alpha = \theta$.

Démonstration (1)

- Supposons que $\mathbb{P}_\alpha$ est absolument continue par rapport à $\mathbb{P}_\theta$ (sinon $K_n(\alpha,\theta) = +\infty$ et il n'y a rien à démontrer).
- La fonction $-\log$ est convexe, donc d'après l'inégalité de Jensen

$$K_n(\alpha,\theta) \geq -\log \mathbb{E}_\theta \frac{L_n(\mathbf{X};\alpha)}{L_n(\mathbf{X};\theta)}. \tag{2}$$

- Dans le cas discret,

$$\mathbb{E}_\theta \frac{L_n(\mathbf{X};\alpha)}{L_n(\mathbf{X};\theta)} = \sum_{\mathbf{x}\in\mathcal{H}^n} \frac{L_n(\mathbf{x};\alpha)}{L_n(\mathbf{x};\theta)} \mathbb{P}_\theta(\mathbf{X}=\mathbf{x}) = \sum_{\mathbf{x}\in\mathcal{H}^n} L_n(\mathbf{x};\alpha) = 1.$$

- De même, dans le cas continu,

$$\mathbb{E}_\theta \frac{L_n(\mathbf{X};\alpha)}{L_n(\mathbf{X};\theta)} = \int_{\mathcal{H}^n} \frac{L_n(\mathbf{x};\alpha)}{L_n(\mathbf{x};\theta)} f_\theta(d\mathbf{x}) = \int_{\mathcal{H}^n} L_n(\mathbf{x};\alpha)\, d\mathbf{x} = \int_{\mathcal{H}^n} f_\alpha(\mathbf{x})\, d\mathbf{x} = 1,$$

où f_θ est la densité de $\mathbb{P}_\theta$.

Démonstration (2)

- Dans les deux cas, on obtient que $K_n(\alpha, \theta) \geq 0$.
- Le cas d'égalité de l'inégalité de Jensen implique que $K_n(\alpha, \theta) = 0$ ssi $\frac{L_n(\mathbf{X};\alpha)}{L_n(\mathbf{X};\theta)}$ est constant $\mathbb{P}_\theta$-presque sûrement, c-à-d si

$$L_n(\mathbf{x}; \alpha) = C L_n(\mathbf{x}; \theta) \quad \mathbb{P}_\theta\text{-presque partout.}$$

- Puisque $\mathbb{P}_\alpha$ est absolument continue par rapport à $\mathbb{P}_\theta$, la densité de $\mathbb{P}_\alpha$ par rapport à $\mathbb{P}_\theta$ est la constante C.
- Or $\mathbb{P}_\alpha$ et $\mathbb{P}_\theta$ sont tous deux de masse 1, donc nécessairement $C = 1$ et donc $\mathbb{P}_\alpha = \mathbb{P}_\theta$.
- Par identifiabilité, on en déduit que $\alpha = \theta$.

Résultat principal

Théorème (consistance de l'EMV)

Supposons que le modèle est identifiable, que $\mathbb{P}_\theta = \mathbb{Q}_\theta^{\otimes n}$ *pour tout* $\theta \in \Theta$ *et que* $\Theta \subset \mathbb{R}^d$ *est compact.*

- *Sous l'hypothèse que* $\log L(x;\cdot)$ *est continue sur* Θ *pour tout* $x \in \mathcal{H}$*, il existe un EMV.*
- *Sous l'hypothèse supplémentaire que, pour tout* $\theta \in \Theta$*, il existe un voisinage* V *de* θ *dans* Θ *et il existe une v.a.* $H \in \mathbb{L}^1(\mathbb{Q}_\theta)$ *tel que* $\sup_{\alpha \in V} |\log L(X_1;\alpha)| \leq H$*, alors l'EMV est consistant.*

Remarque : Hypothèses très restrictives : $\log L(x;\cdot)$ continue impose que $L_n(x;\theta) > 0$ pour tout $x \in \mathcal{H}$ et $\theta \in \Theta$. Les conclusions de ce théorème sont valides dans des cas beaucoup plus généraux (cf. exercices).

Exemple 1

- Modèle statistique exponentiel

$$\left(\mathbb{R}_+^n, \left\{\mathrm{Exp}(\theta)^{\otimes n}\right\}_{\theta>0}\right).$$

- La vraisemblance est donnée par

$$L_n(x_1,\ldots,x_n;\theta) = \prod_{i=1}^n L(x_i;\theta) = \prod_{i=1}^n \theta e^{-\theta x_i} = \theta^n e^{-n\theta\,\bar{x}_n}.$$

- La log-vraisemblance est donc donnée par

$$\ell_n(x_1,\ldots,x_n;\theta) = n\log\theta - n\theta\bar{x}_n.$$

- Dérivée $n/\theta - n\bar{x}_n \rightsquigarrow$ la (log-)vraisemblance est maximale en $1/\bar{x}_n$, et l'unique EMV est $\hat{\theta} = 1/\bar{X}_n$.
- Sa consistance découle de la loi des grands nombres.
- On ne pourrait pas obtenir directement ce résultat à l'aide du théorème 4, puisque $\Theta = \mathbb{R}_+^*$ n'est pas compact.

Exemple 2

- Dans le jeu de pile-ou-face, on a

$$\log L(X_1;\theta) = \ell(X_1;\theta) = X_1 \log\theta + (1 - X_1)\log(1-\theta),$$

où $X_1 \in \{0,1\}$.

- Donc $\theta \mapsto \ell(0;\theta)$ et $\theta \mapsto \ell(1;\theta)$ sont continus sur $]0,1[$, et

$$|\ell(X_1;\theta)| \leq \log\theta + \log(1-\theta).$$

- On peut donc appliquer le théorème 4, mais seulement si on réduit l'ensemble des paramètres à un intervalle du type $[\varepsilon, 1-\varepsilon]$ pour $\varepsilon > 0$.
- Pourtant, nous avons vu que l'EMV $\bar{X}_n$ est consistant sur l'ensemble des paramètres complet $\Theta = [0,1]$.

Démonstration

- L'existence d'un EMV $\hat{\theta}_n$ est une conséquence de la continuité de $L_n(\mathbf{x}, \cdot)$ et de la compacité de Θ.
- Soit $\theta \in \Theta$ fixé. Pour tout $\alpha \in \Theta$, on pose

$$U_n(\alpha) = \frac{1}{n} \log L_n(X_1, \ldots, X_n; \alpha) = \frac{1}{n} \sum_{i=1}^{n} \ell(X_i; \alpha)$$

et

$$U(\alpha) = \mathbb{E}_\theta \log L_n(X_1, \ldots, X_n; \alpha).$$

- D'après la loi des grands nombres, $U_n \to U$ en $\mathbb{P}_\theta$-probabilité. De plus, par définition d'un EMV, $U_n(\hat{\theta}_n) = \sup_{\alpha \in \Theta} U_n(\alpha)$.

Lemme

La suite U_n converge uniformément en $\mathbb{P}_\theta$-probabilité vers U : $\forall \varepsilon > 0$,

$$\lim_{n \to +\infty} \mathbb{P}_\theta \left(\sup_{\alpha \in \Theta} |U_n(\alpha) - U(\alpha)| > \varepsilon \right) = 0.$$

Démonstration (suite)

- Puisque

$$\left| \sup_{\alpha\in\Theta} U_n(\alpha) - \sup_{\alpha\in\Theta} U(\alpha) \right| \leq \sup_{\alpha\in\Theta} |U_n(\alpha) - U(\alpha)|,$$

  le lemme implique que $U_n(\hat{\theta}_n)$ cv. en proba. vers $\sup_{\alpha\in\Theta} U(\alpha)$.
- De plus,

$$\sup_{\alpha\in V} |\log L_n(X_1, \ldots, X_n; \alpha)| = \sup_{\alpha\in V} \left| \sum_{i=1}^{n} \log L(X_i; \alpha) \right| \leq n\, H,$$

  et on peut donc appliquer le théorème de continuité sous le signe somme pour en déduire que $\alpha \mapsto U(\alpha)$ est continue.
- Par compacité de Θ, il existe $\tau \in \Theta$ tel que $U(\tau) = \sup_{\alpha\in\Theta} U(\alpha)$.
- Or l'information de Kullback-Leibler vérifie

$$K_n(\tau, \theta) = U(\theta) - U(\tau) = U(\theta) - \sup_{\alpha\in\Theta} U(\alpha) \leq 0.$$

  D'après la proposition 3 et l'identifiabilité du modèle, $K_n(\tau, \theta) = 0$ et donc $\tau = \theta$.

Démonstration (fin)

- Ainsi, $U_n(\hat{\theta}_n) \to U(\theta)$ et $K_n(\hat{\theta}_n, \theta) \to 0$ en $\mathbb{P}_\theta$-probabilité, puisque

$$0 \leq K_n(\hat{\theta}_n, \theta) = U(\theta) - U(\hat{\theta}_n) \leq |U(\theta) - U_n(\hat{\theta}_n)| + \sup_{\alpha \in \Theta} |U_n(\alpha) - U(\alpha)|$$

$$\xrightarrow[n \to +\infty]{\mathbb{P}_\theta} 0$$

  d'après le lemme.
- Or $K_n(\alpha, \theta) = 0$ ssi $\alpha = \theta$ et $\alpha \mapsto K_n(\alpha, \theta)$ continue, donc pour tout $\varepsilon > 0$, il existe $\eta > 0$ tel que, pour tout $\alpha \in \Theta$ tel que $|\alpha - \theta| > \eta$, on a $K_n(\alpha, \theta) > \varepsilon$.
- Ceci implique que

$$\limsup_{n \to +\infty} \mathbb{P}_\theta \left(|\hat{\theta}_n - \theta| > \eta \right) \leq \limsup_{n \to +\infty} \mathbb{P}_\theta \left(K_n(\hat{\theta}_n, \theta) > \varepsilon \right) = 0. \quad \square$$

Démonstration du lemme

- Pour tout $x \in \mathcal{H}$ et $\eta > 0$, posons

$$h(x,\eta) = \sup_{\alpha,\beta\in\Theta,\ \text{t.q. } |\alpha-\beta|\leq\eta} \left|\log L(x;\alpha) - \log L(x;\beta)\right|.$$

- La compactité de Θ et notre hypothèse de domination permettent d'appliquer le théorème de convergence dominée pour en déduire que, pour tout $\varepsilon > 0$, il existe $\eta > 0$ tel que $\mathbb{E}_\theta h(X_1;\eta) < \varepsilon/3$.
- Par compacité de Θ, il existe $N \in \mathbb{N}$ et $\theta_1, \ldots, \theta_N > 0$ tels que

$$\Theta = \bigcup_{j=1}^{N} B(\theta_j, \eta).$$

Démonstration du lemme (suite)

- On a alors

$$\begin{aligned}\sup_{\alpha\in\Theta}|U_n(\alpha)-U(\alpha)| &= \max_{1\le j\le N}\sup_{\alpha\in B(\theta_j,\eta)}|U_n(\alpha)-U(\alpha)|\\ &\le \max_{1\le j\le N}\sup_{\alpha\in B(\theta_j,\eta)}|U_n(\alpha)-U_n(\theta_j)|+\max_{1\le j\le N}|U_n(\theta_j)-U(\theta_j)|\\ &+\max_{1\le j\le N}\sup_{\alpha\in B(\theta_j,\eta)}|U(\theta_j)-U(\alpha)|\\ &\le \frac{1}{n}\sum_{i=1}^{n}h(X_i;\eta)+\max_{1\le j\le N}|U_n(\theta_j)-U(\theta_j)|+\mathbb{E}_\theta h(X_1,\eta).\end{aligned}$$

- D'après la loi des grands nombres, le premier terme converge en $\mathbb{P}_\theta$-probabilité vers $\mathbb{E}_\theta h(X_1,\eta)$, qui est inférieur à $\varepsilon/3$.
- De même, le second terme converge en $\mathbb{P}_\theta$-probabilité vers 0.

Démonstration du lemme (fin)

Finalement

$$\mathbb{P}_\theta\left(\sup_{\alpha\in\Theta}|U_n(\alpha)-U(\alpha)|>\varepsilon\right)\leq\mathbb{P}_\theta\left(\left|\frac{1}{n}\sum_{i=1}^{n}h(X_i;\bar{\eta})-\mathbb{E}_\theta h(X_1,\bar{\eta})\right|>\frac{\varepsilon}{6}\right)$$
$$+\mathbb{P}_\theta\left(\max_{1\leq j\leq N}|U_n(\theta_j)-U(\theta_j)|>\frac{\varepsilon}{6}\right),$$

qui converge vers 0 quand $n\to+\infty$.

Quelques calculs formels

Objectif : normalité asymptotique de l'EMV.

Notations : on note ∇ le gradient par rapport à la variable $\theta \in \Theta$ et ∇^2 la matrice hessienne.

- Soit $\theta \in \Theta$ fixé et K l'application de Θ dans $\mathbb{R}_+$ définie par

$$K : \alpha \mapsto K_n(\alpha, \theta).$$

- On a

$$\mathbb{V}_\theta(\nabla \log L_n(\mathbf{X}; \theta)) = \mathbb{E}_\theta \left(\frac{\nabla L_n(\mathbf{X}; \theta)}{L_n(\mathbf{X}; \theta)} \right)^2 - \left(\mathbb{E}_\theta \frac{\nabla L_n(\mathbf{X}; \theta)}{L_n(\mathbf{X}; \theta)} \right)^2 .$$

Quelques calculs formels (suite)

- Dans le cas discret,

$$\mathbb{E}_\theta \frac{\nabla L_n(\mathbf{X};\theta)}{L_n(\mathbf{X};\theta)} = \sum_{\mathbf{x}\in\mathcal{H}^n} \nabla L_n(\mathbf{x};\theta) = \nabla \sum_{\mathbf{x}\in\mathcal{H}^n} L_n(\mathbf{x};\theta) = \nabla 1 = 0. \quad (3)$$

- De même, dans le cas continu,

$$\mathbb{E}_\theta \frac{\nabla L_n(\mathbf{X};\theta)}{L_n(\mathbf{X};\theta)} = \mathbb{E}_\theta \frac{\nabla L_n(\mathbf{X};\theta)}{f_\theta(\mathbf{X})} = \int_{\mathcal{H}^n} \nabla L_n(\mathbf{x};\theta)\, d\mathbf{x} = \nabla \int_{\mathcal{H}^n} f_\theta(\mathbf{x})\, d\mathbf{x} = 0. \quad (4)$$

- Ainsi,

$$\mathbb{V}_\theta(\nabla \log L_n(\mathbf{X};\theta)) = \mathbb{E}_\theta \left(\frac{\nabla L_n(\mathbf{X};\theta)}{L_n(\mathbf{X};\theta)} \right)^2.$$

Quelques calculs formels (fin)

- Par ailleurs,

$$\nabla K(\alpha) = -\mathbb{E}_\theta \frac{\nabla L_n(\mathbf{X};\alpha)}{L_n(\mathbf{X};\alpha)}$$

- et donc

$$\nabla^2 K(\theta) = \mathbb{E}_\theta \left(\frac{\nabla L_n(\mathbf{X};\theta)}{L_n(\mathbf{X};\theta)} \right)^2 - \mathbb{E}_\theta \frac{\nabla^2 L_n(\mathbf{X};\theta)}{L_n(\mathbf{X};\theta)}.$$

- On démontre de la même façon que dans (3) et (4) que le dernier terme du membre de droite est nul, de sorte que

$$\nabla^2 K(\theta) = \mathbb{V}_\theta(\nabla \log L_n(\mathbf{X};\theta)).$$

Information de Fisher

Ceci nous conduit à la définition suivante.

Définition

On suppose que Θ *est un ouvert* $\mathbb{R}^d$ *et que* $\nabla \log L_n(\mathbf{X};\theta) \in \mathbb{L}^2(\mathbb{P}_\theta)$ *pour tout* $\theta \in \Theta$. *L'information de Fisher est définie pour tout* $\theta \in \Theta$ *par* $I_n(\theta) = \mathbb{V}_\theta(\nabla \log L_n(\mathbf{X};\theta))$, *c'est-à-dire*

$$I_n(\theta) = \left(Cov_\theta \left(\frac{\partial}{\partial \theta_i} \log L_n(\mathbf{X};\theta) \, ; \, \frac{\partial}{\partial \theta_j} \log L_n(\mathbf{X};\theta) \right) \right)_{1 \leq i,j \leq d}.$$

Le calcul précédent montre que $I_n(\theta)$ décrit la courbure de $\alpha \mapsto K_n(\alpha,\theta)$ en son minimum $\alpha = \theta$: au voisinage de θ,

$$K_n(\alpha,\theta) = \frac{1}{2}(\alpha-\theta)^T I_n(\theta)(\alpha-\theta) + o(|\alpha-\theta|^2).$$

L'info. de Fisher permet donc de quantifier le pouvoir de discrimination de l'info. de Kullback-Leibler entre deux valeurs proches du paramètre.

Exemple

Dans le jeu de pile-ou-face, d'après (1),

$$\begin{aligned} I_n(\theta) &= \mathbb{V}_\theta \left(\nabla \left(n\bar{X}_n \log\theta + n(1-\bar{X}_n)\log(1-\theta) \right) \right) \\ &= \left(\frac{1}{\theta} + \frac{1}{1-\theta} \right)^2 \mathbb{V}_\theta(n\bar{X}_n) \\ &= \frac{n}{\theta(1-\theta)}. \end{aligned}$$

Ainsi, l'EMV a une faible incertitude pour θ proche de 0 ou 1 (voir le théorème 8 ci-dessous).

Cas i.i.d.

Proposition

Dans le cas d'échantillons i.i.d., c'est-à-dire si $\mathbb{P}_\theta = \mathbb{Q}_\theta^{\otimes n}$ *pour tout* $\theta \in \Theta$, *on a* $I_n(\theta) = nI(\theta)$, *où*

$$I(\theta) = \mathbb{V}_\theta(\nabla \log L(X_1;\theta)) = \mathbb{E}_{\mathbb{Q}_\theta}\left[\nabla \log L(X_1;\theta)\,(\nabla \log L(X_1;\theta))^T\right]$$

est l'information de Fisher du modèle $(\mathcal{H}, \{\mathbb{Q}_\theta\}_{\theta\in\Theta})$.

Démonstration : Puisque

$$\nabla \log L_n(X_1,\ldots,X_n;\theta) = \sum_{i=1}^n \nabla \log L(X_i;\theta)$$

et que les X_i sont i.i.d., on a

$$I_n(\theta) = \sum_{i=1}^n \mathbb{V}_\theta(\nabla \log L(X_i;\theta)) = n\mathbb{V}_\theta(\nabla \log L(X_1;\theta)).$$

Un résultat technique

On peut justifier les calculs formels précédents.

Proposition

Supposons que, pour tout $\theta \in \Theta$, *il existe un voisinage* V *de* θ *dans* Θ *tel que* $\sup_{\alpha \in V} \left| \frac{\nabla L_n(\mathbf{X};\alpha)}{L_n(\mathbf{X};\theta)} \right| \in \mathbb{L}^1(\mathbb{P}_\theta)$. *Alors*

$$\mathbb{E}_\theta \nabla \log L_n(\mathbf{X};\theta) = 0. \tag{5}$$

Si de plus $\sup_{\alpha \in V} \left| \frac{\nabla^2 L_n(\mathbf{X};\alpha)}{L_n(\mathbf{X};\theta)} \right| \in \mathbb{L}^1(\mathbb{P}_\theta)$, *alors* $I_n(\theta)$ *existe et*

$$I_n(\theta) = -\mathbb{E}_\theta \nabla^2 \log L_n(\mathbf{X};\theta).$$

Hypothèses

Définition

Le modèle statistique $(\mathcal{H}^n, \{\mathbb{P}_\theta\}_{\theta\in\Theta}$ *est dit régulier si*

- *pour tout* $\theta \in \Theta$, $\mathbb{P}_\theta = \mathbb{Q}_\theta^{\otimes n}$;
- *pour tout* $x \in \mathcal{H}$, *l'application* $\theta \mapsto \log L(x;\theta)$ *est continue sur* Θ ;
- *pout tout* $\theta \in \theta$, *il existe un voisinage* V *de* θ *dans* Θ *tel que*

$$\sup_{\alpha\in V}\left|\frac{\nabla L_n(\mathbf{X};\alpha)}{L_n(\mathbf{X};\theta)}\right| + \left|\frac{\nabla^2 L_n(\mathbf{X};\alpha)}{L_n(\mathbf{X};\theta)}\right| \in \mathbb{L}^1(\mathbb{P}_\theta),$$

de sorte que $\mathbb{E}_\theta \nabla \log L_n(\mathbf{X};\theta) = 0$ *et* $I_n(\theta) = -\mathbb{E}_\theta \nabla^2 \log L_n(\mathbf{X};\theta)$, *d'après la proposition précédente.*

- *pour tout* $\theta \in \Theta$, $I(\theta)$ *est une matrice inversible ;*

Exemple : Il est facile de vérifier que le modèle statistique du jeu de pile-ou-face avec $\Theta =]0,1[$ est régulier.

Normalité asymptotique de l'EMV

Théorème

On considère $(\mathcal{H}^n, \{\mathbb{Q}_\theta^{\otimes n}\})_{\theta\in\Theta})$ un modèle statistique régulier.
Si un EMV $\hat{\theta}_n$ existe pour tout n assez grand et est consistant, alors

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n\to+\infty]{\text{loi sous } \mathbb{P}_\theta} \mathcal{N}_d(0, I(\theta)^{-1}), \quad \forall\theta \in \Theta.$$

- Autrement dit, l'EMV a pour vitesse $\sqrt{n}$ et pour loi limite $\mathcal{N}_d(0, I(\theta)^{-1})$.
- Les hypothèses sont nécessaires, car il existe des exemples de modèles statistiques i.i.d. pour lesquels l'EMV est asymptotiquement non-normal (cf. exercices)

Exemple

- Soit le modèle statistique

$$\left(\mathbb{R}_n^n, \left\{\mathcal{U}([0,\theta])^{\otimes n}\right\}_{\theta>0}\right).$$

- Sa vraisemblance est

$$L_n(x_1,\ldots,x_n;\theta) = \theta^{-n}\mathbb{1}_{0\leq x_1,\ldots,x_n\leq\theta}.$$

- L'EMV est donc la plus petite valeur de θ telle que l'indicatrice est non nulle, c-à-d

$$\hat{\theta}_n = \max_{1\leq i\leq n} X_i.$$

- Remarquons que $0 \leq \hat{\theta}_n \leq \theta$ $\mathbb{P}_\theta$-p.s. et, pour tout $t \in [-n\theta, 0]$,

$$\mathbb{P}_\theta\left(n(\hat{\theta}_n-\theta)\leq t\right) = \mathbb{P}_\theta\left(\max_{1\leq i\leq n} X_i \leq \theta + \frac{t}{n}\right) = \left(1+\frac{t}{n\theta}\right)^n.$$

Exemple (suite)

- Donc
$$\mathbb{P}_\theta(n(\theta - \hat{\theta}_n) \leq t) \xrightarrow[n\to+\infty]{} 1 - e^{-t/\theta}, \quad \forall t \geq 0.$$
- On reconnait la fonction de répartition de la loi exponentielle de paramètre $1/\theta$.
- Autrement dit,
$$n(\hat{\theta} - \theta) \xrightarrow[n\to+\infty]{\text{loi sous } \mathbb{P}_\theta} -\text{Exp}(1/\theta).$$
- Ainsi, l'EMV est ici de vitesse n et asymptotiquement exponentiel.

Démonstration

- Fixons $\theta \in \Theta$ dans toute la preuve. On définit

$$U_n(\alpha) = \log L_n(\mathbf{X}; \alpha) = \sum_{i=1}^{n} \log L(X_i; \alpha), \quad \forall \alpha \in \Theta.$$

- Rappelons que $\nabla U_n(\hat{\theta}_n) = 0$ car $\hat{\theta}_n$ est un EMV.
- La formule de Taylor avec reste intégral assure que

$$0 = \nabla U_n(\hat{\theta}_n) = \nabla U(\theta) + (\hat{\theta}_n - \theta) \int_0^1 \nabla^2 U_n(\theta + t(\hat{\theta}_n - \theta))\, dt.$$

- Donc

$$-\frac{1}{\sqrt{n}} \nabla U_n(\theta) = \sqrt{n}(\hat{\theta}_n - \theta) \bar{U}_n,$$

où

$$\bar{U}_n = \frac{1}{n} \int_0^1 \nabla^2 U_n(\theta + t(\hat{\theta}_n - \theta))\, dt.$$

Démonstration (suite)

- Or
$$\mathbb{V}_\theta(\nabla \log L(X_1;\theta)) = I(\theta) \quad \text{et} \quad \mathbb{E}_\theta \nabla \log L(X_1;\theta) = 0.$$
- Donc, d'après le théorème central limite,
$$\frac{1}{\sqrt{n}} \nabla U_n(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \nabla \log L(X_i;\theta) \xrightarrow[n\to+\infty]{\text{loi sous } \mathbb{P}_\theta} \mathcal{N}_d(0, I(\theta)).$$
- Pour terminer la preuve, il nous suffit donc de démontrer que
$$\bar{U}_n \xrightarrow[n\to+\infty]{\mathbb{P}_\theta} -I(\theta),$$
puisqu'on en déduirait par le lemme de Slutsky que
$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n\to+\infty]{\text{loi sous } \mathbb{P}_\theta} I(\theta)^{-1} \mathcal{N}_d(0, I(\theta)) = \mathcal{N}_d(0, I(\theta)^{-1}).$$

Démonstration (suite)

Montrons donc

$$\bar{U}_n \xrightarrow[n\to+\infty]{\mathbb{P}_\theta} -I(\theta),$$

- Pour tout $x \in \mathcal{H}$ et tout $r > 0$, on définit

$$\sigma(x,r) = \sup_{\alpha,\theta\in\Theta \text{ t.q. } |\alpha-\theta|\leq r} \left|\nabla^2 \log L(x;\alpha) - \nabla^2 \log L(x;\theta)\right|.$$

- Le fait que le modèle est régulier assure que $\sigma(X_1, r) \in \mathbb{L}^1(\mathbb{P}_\theta)$ pour $r > 0$ suffisamment petit et, d'après le théorème de convergence dominée, pour $\varepsilon > 0$ fixé, il existe $r > 0$ suffisamment petit tel que $\mathbb{E}_\theta \sigma(X_1, r) < \varepsilon/2$.
- Or

$$\bar{U}_n = \frac{1}{n}\sum_{i=1}^{n} \int_0^1 \nabla^2 \log L(X_i; \theta + t(\hat{\theta}_n - \theta))\, dt.$$

Démonstration (fin)

$$\mathbb{P}_\theta\left(|I(\theta)+\bar{U}_n|\geq\varepsilon\right)\leq\mathbb{P}_\theta\left(\left|I(\theta)+\frac{1}{n}\sum_{i=1}^{n}\int_0^1\nabla^2\log L(X_i;\theta)\,dt\right|\geq\frac{\varepsilon}{2}\right)$$
$$+\mathbb{P}_\theta\left(\frac{1}{n}\sum_{i=1}^{n}\int_0^1\left|\nabla^2\log L(X_i;\theta)-\nabla^2\log L(X_i;\theta+t(\hat{\theta}_n-\theta))\right|\,dt\geq\frac{\varepsilon}{2}\right)$$
$$\leq\mathbb{P}_\theta\left(\left|I(\theta)+\frac{1}{n}\sum_{i=1}^{n}\nabla^2\log L(X_i;\theta)\right|\geq\frac{\varepsilon}{2}\right)$$
$$+\mathbb{P}_\theta\left(|\hat{\theta}_n-\theta|\geq r\right)+\mathbb{P}_\theta\left(\frac{1}{n}\sum_{i=1}^{n}\sigma(X_i,r)\geq\frac{\varepsilon}{2}\right).$$

- Le premier terme cv. vers 0 en proba. par la LGN.
- Le second terme tend vers 0 car l'EMV $\hat{\theta}_n$ est consistant.
- Le troisième terme converge vers 0 puisque, par la LGN,

$$\frac{1}{n}\sum_{i=1}^{n}\sigma(X_i,r)\xrightarrow[n\to+\infty]{\mathbb{P}_\theta}\mathbb{E}_\theta\sigma(X_1,r)<\frac{\varepsilon}{2}.$$

Propriétés théoriques de l'EMV

Dans le cas des échantillons i.i.d., l'EMV satisfait plusieurs « bonnes » propriétés qui montrent qu'il est optimal (en un certain sens) :

- il est consistant ;
- il est asymptotiquement normal, de vitesse $\sqrt{n}$;
- il est asymptotiquement efficace.

Nous avons déjà prouvé les deux premiers points, il nous reste à justifier le dernier.

Théorème de Cramer-Rao

Théorème

Soit un modèle statistique $(\mathcal{H}^n, \{\mathbb{P}_\theta\}_{\theta\in\Theta})$ avec $\mathcal{H} \subset \mathbb{R}^k$ et $\Theta \subset \mathbb{R}^d$, et paramètre d'intérêt $g(\theta)$ pour une fonction $g : \Theta \to \mathbb{R}$ $\mathcal{C}^1$. On suppose que, pour tout $\theta \in \Theta$, il existe un voisinage V de θ dans Θ tel que la v.a.

$$H = \sup_{\alpha\in V} \left| \frac{\nabla L_n(\mathbf{X};\alpha)}{L_n(\mathbf{X};\theta)} \right| \in \mathbb{L}^2(\mathbb{P}_\theta).$$

Sous ces hyp., si $\hat{g}$ est un estimateur d'ordre 2 de $g(\theta)$ sans biais, alors

$$\mathcal{R}(\theta,\hat{g}) \geq \nabla g(\theta)^T I_n(\theta)^{-1} \nabla g(\theta), \tag{6}$$

où

$$\mathcal{R}(\theta,\hat{g}) = \mathbb{E}_\theta|\hat{g} - g(\theta)|^2 = \mathbb{V}_\theta(\hat{g})$$

est le risque quadratique de l'estimateur $\hat{g}$ de $g(\theta)$.

L'équation (6) s'appelle borne de Cramer-Rao.

Exemple

- Dans le jeu de pile-ou-face, puisque l'EMV $\bar{X}_n$ est sans biais,

$$\mathcal{R}(\theta, \bar{X}_n) = \mathbb{V}_\theta(\bar{X}_n) = \frac{1}{n}\mathbb{V}_\theta(X_1) = \frac{\theta(1-\theta)}{n}.$$

- On a vu que $I_n(\theta) = \frac{n}{\theta(1-\theta)}$ dans ce modèle.
- L'EMV du jeu de pile-ou-face réalise donc l'égalité dans la borne de Cramer-Rao.
- Donc il n'existe pas dans le jeu de pile-ou-face de meilleur estimateur d'ordre 2 et sans biais que l'EMV, en terme de risque quadratique.

Démonstration

Admettons pour le moment le

Lemme

Sous les hypothèses du théorème de Cramer-Rao, pour tout $\theta \in \Theta$,

$$\mathbb{E}_\theta \nabla \log L_n(\mathbf{X}; \theta) = 0$$

et

$$\mathbb{E}_\theta \left[\hat{g} \nabla \log L_n(\mathbf{X}; \theta)\right] = \nabla \mathbb{E}_\theta(\hat{g}).$$

Ce résultat implique que

$$\nabla g(\theta) = \mathbb{E}_\theta \left[\hat{g} \nabla \log L_n(\mathbf{X}; \theta)\right] = \mathbb{E}_\theta \left[(\hat{g} - g(\theta)) \nabla \log L_n(\mathbf{X}; \theta)\right].$$

Démonstration (fin)

- Donc, pour tout $u \in \mathbb{R}^d$, en notant $\langle \cdot, \cdot \rangle$ le produit scalaire usuel,

$$\begin{aligned}\langle u, \nabla g(\theta)\rangle^2 &= \left(\mathbb{E}_\theta\left[(\hat{g} - g(\theta))\langle u, \nabla \log L_n(\mathbf{X};\theta)\rangle\right]\right)^2 \\ &\leq \mathcal{R}(\theta, \hat{g})\mathbb{E}_\theta\langle u, \nabla \log L_n(\mathbf{X};\theta)\rangle^2,\end{aligned}$$

  par Cauchy-Schwarz.
- Or, par définition de l'information de Fisher,

$$\begin{aligned}\mathbb{E}_\theta\langle u, \nabla \log L_n(\mathbf{X};\theta)\rangle^2 &= u^T\mathbb{E}_\theta\left[\nabla \log L_n(\mathbf{X};\theta)\nabla \log L_n(\mathbf{X};\theta)^T\right]u \\ &= u^T\mathbb{V}_\theta(\nabla \log L_n(\mathbf{X};\theta))u = u^T I_n(\theta)u.\end{aligned}$$

- Donc, en choisissant $u = I_n(\theta)^{-1}\nabla g(\theta)$,

$$\mathbb{E}_\theta\langle u, \nabla \log L_n(\mathbf{X};\theta)\rangle^2 = \nabla g(\theta)^T I_n(\theta)^{-1}\nabla g(\theta)$$

  et

$$\langle u, \nabla g(\theta)\rangle = \nabla g(\theta)^T I_n(\theta)^{-1}\nabla g(\theta).$$

- D'où

$$\mathcal{R}(\theta, \hat{g}) \geq \nabla g(\theta)^T I_n(\theta)^{-1}\nabla g(\theta).$$

Démonstration du lemme

- La première équation a déjà été démontrée.
- Pour la seconde équation, dans le cas discret,

$$\mathbb{E}_\theta \left[\hat{g}\nabla \log L_n(\mathbf{X};\theta)\right] = \sum_{\mathbf{x}\in\mathcal{H}^n} \hat{g}(\mathbf{x}) \frac{\nabla L_n(\mathbf{x};\theta)}{L_n(\mathbf{x};\theta)} \mathbb{P}_\theta(\mathbf{X}=\mathbf{x}) = \sum_{\mathbf{x}\in\mathcal{H}^n} \hat{g}(\mathbf{x}) \nabla L_n(\mathbf{x};\theta).$$

- Pour tout $\alpha \in V$,

$$|\hat{g}(\mathbf{x})\nabla L_n(\mathbf{x};\alpha)| \leq \frac{1}{2}|\hat{g}(\mathbf{x})|^2 L_n(\mathbf{x};\theta) + \frac{1}{2}H^2 L_n(\mathbf{x};\theta) \qquad (7)$$

  est dans $\mathbb{L}^1(\mu)$, où $\mu = \sum_{\mathbf{x}\in\mathcal{H}^n} \delta_{\mathbf{x}}$.
- D'après le théorème de dérivation sous le signe somme, on a

$$\mathbb{E}_\theta \left[\hat{g}\nabla \log L_n(\mathbf{X};\theta)\right] = \nabla \sum_{\mathbf{x}\in\mathcal{H}^n} \hat{g}(\mathbf{x}) \mathbb{P}_\theta(\mathbf{X}=\mathbf{x}) = \nabla \mathbb{E}_\theta \hat{g}.$$

- La méthode est similaire dans le cas continu.

Estimateur efficace

Définition

- *Un estimateur $\hat{g}$ sans biais et d'ordre 2 est efficace si son risque quadratique atteint (c'est-à-dire réalise l'égalité dans) la borne de Cramer-Rao* (6).
- *Dans la suite de modèle statistiques $(\mathcal{H}^n, \{\mathbb{Q}_\theta^{\otimes n}\}_{\theta\in\Theta})$, une suite $\hat{g}_n$ d'estimateurs sans biais et d'ordre 2 est asymptotiquement efficace si*

$$\lim_{n\to+\infty} n\mathcal{R}(\theta, \hat{g}_n) = \nabla g(\theta)^T I(\theta)^{-1} \nabla g(\theta).$$

Efficacité asymptotique de l'EMV

On déduit des résultats précédents et de la normalité asymptotique de l'EMV

Théorème

Si $\Theta \subset \mathbb{R}$ et si les hypothèses du théorème de normalité asymptotique de l'EMV sont vérifiées, l'EMV $\hat{\theta}_n$ est asymptotiquement efficace, c'est-à-dire

$$\lim_{n \to +\infty} n\mathcal{R}(\theta; \hat{\theta}_n) = \frac{1}{I(\theta)}.$$

Intervalle de confiance asymptotique : cas de la dimension 1

Rappel : pour tout $\alpha \in]0,1[$, q_α est le quantile d'ordre $1-\alpha/2$ de $\mathcal{N}(0,1)$.

Théorème (intervalle de confiance asymptotique)

Si $\Theta \subset \mathbb{R}$, et les hypothèses précédentes sont satisfaites, et l'application qui à $x \in \mathcal{H}$ associe

$$\sup_{\alpha \in \Theta} \frac{(\nabla L(x;\alpha))^2}{L(x;\alpha)}$$

est $\mathbb{L}^1$ par rapport à la mesure de Lebesgue sur $\mathbb{R}^k$ dans le cas continu (resp. par rapport à $\mu = \sum_{x \in \mathcal{H}^n} \delta_x$ dans le cas discret), alors $\forall \alpha \in]0,1[$,

$$\left[\hat{\theta}_n - \frac{q_\alpha}{\sqrt{nI(\hat{\theta}_n)}} , \hat{\theta}_n + \frac{q_\alpha}{\sqrt{nI(\hat{\theta}_n)}} \right]$$

est un intervalle de confiance de niveau asympt. $1-\alpha$ pour le paramètre θ.

Détermination du nombre de données n à collecter

- On se donne une précision d'estimation $\varepsilon > 0$ du paramètre θ et un niveau de confiance $1 - \alpha > 0$.
- On réalise une première estimation grossière de θ avec un nombre limité n_0 de données $\rightsquigarrow$ premier intervalle de confiance I_0 de niveau α.
- on calcule la taille maximale de l'intervalle de confiance sur I_0 :

$$\sup_{\theta \in I_0} \frac{q_\alpha}{\sqrt{nI(\theta)}} = \frac{1}{\sqrt{n}} \sup_{\theta \in I_0} \frac{q_\alpha}{\sqrt{I(\theta)}}$$

- On choisit pour n le premier entier tel que cette largeur soit inférieure au seuil de précision ε. C'est le nombre de mesures à réaliser afin d'obtenir une estimation avec précision inférieure à ε et niveau de confiance $1 - \alpha$.

Démonstration

- D'après le théorème de normalité asymptotique de l'EMV, on a

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n\to+\infty]{\text{loi sous } \mathbb{P}_\theta} \mathcal{N}(0, I(\theta)^{-1}). \tag{8}$$

- Difficulté : $I(\theta)$ est inconnu.
- Mais on peut utiliser la consistance de l'estimateur $\hat{\theta}_n$ pour démontrer que $I(\hat{\theta}_n)$ est un estimateur consistant de $I(\theta)$.
- Pour cela, nous avons besoin de montrer que la fonction I est continue. Or, dans le cas continu (le cas discret se démontre de la même façon),

$$I(\theta) = -\mathbb{E}_\theta(\nabla \log L(X_1;\theta))^2 = -\int_{\mathcal{H}} \frac{(\nabla L(x;\theta))^2}{L(x;\theta)}\, dx.$$

- Nos hypothèses permettent d'appliquer le théorème de continuité sous le signe somme afin de déduire la continuité de I.
- Ainsi, $I(\hat{\theta}_n)$ converge en probabilité vers $I(\theta)$.

Démonstration (fin)

- Le lemme de Slutsky permet d'en déduire que

$$\sqrt{nI(\hat{\theta}_n)}(\hat{\theta}_n - \theta) \xrightarrow[n\to+\infty]{\text{loi sous } \mathbb{P}_\theta} \mathcal{N}(0,1).$$

- La construction d'un intervalle de confiance est ensuite classique : si $G \sim \mathcal{N}(0,1)$, $\mathbb{P}(|G| \leq q_\alpha) = 1 - \alpha$, et donc, d'après le théorème de Portmanteau,

$$\lim_{n\to+\infty} \mathbb{P}\left(\theta \in \left[\hat{\theta}_n - \frac{q_\alpha}{\sqrt{nI(\hat{\theta}_n)}}, \hat{\theta}_n + \frac{q_\alpha}{\sqrt{nI(\hat{\theta}_n)}}\right]\right)$$

$$\lim_{n\to+\infty} \mathbb{P}\left(\sqrt{nI(\hat{\theta}_n)}\,|\hat{\theta}_n - \theta| \leq q_\alpha\right) = \mathbb{P}(|G| \geq q_\alpha) = 1 - \alpha.$$

Région de confiance en dimension $d \geq 2$

- Supposons $d \geq 2$ et $I(\theta)$ inversible.
- Il existe une matrice $A(\theta)$ symétrique définie positive telle que $A^2(\theta) = I(\theta)$, appelée *racine carrée matricielle*.
- On a alors
$$\sqrt{n}A(\theta)(\hat{\theta}_n - \theta) \xrightarrow[n\to+\infty]{\text{loi sous } \mathbb{P}_\theta} \mathcal{N}_d(0, \mathrm{Id}).$$
- Si $G \sim \mathcal{N}_d(0, \mathrm{Id})$, alors $|G|^2$ suit la loi du $\chi^2(d)$ à d degrés de liberté.
- Soit $q_{d,\alpha}$ le quantile de niveau $1-\alpha$ pour la loi $\chi^2(d)$, c'est-à-dire l'unique solution q de
$$\int_0^q \frac{(1/2)^{k/2}}{\Gamma(k/2)} r^{k/2-1} e^{-r/2}\, dr = 1 - \alpha.$$
- La consistance de l'estimateur $A(\hat{\theta}_n)$ de $A(\theta)$ découle de la continuité de la racine carrée matricielle et de la continuité de $\theta \mapsto I(\theta)$, qui se démontre comme dans le dernier théorème.

Région de confiance en dimension $d \geq 2$ (suite)

On obtient donc le résultat

Théorème (ellipsoïde de confiance asymptotique)

On suppose que $\Theta \subset \mathbb{R}^d$ avec $d \geq 2$, que les hypothèses précédentes sont satisfaites, et que l'application qui à $x \in \mathcal{H}$ associe la matrice

$$\sup_{\alpha \in \Theta} \left| \frac{\nabla L(x;\alpha)(\nabla L(x;\alpha))^T}{L(x;\alpha)} \right|$$

est $\mathbb{L}^1$ par rapport à la mesure de Lebesgue sur $\mathbb{R}^k$ dans le cas continu (resp. par rapport à $\mu = \sum_{x \in \mathcal{H}^n} \delta_x$ dans le cas discret), alors pour tout $\alpha \in]0,1[$,

$$\hat{\theta}_n + \sqrt{\frac{q_{d,\alpha}}{n}} A(\hat{\theta}^n)^{-1} B(0,1) = \left\{ \hat{\theta}_n + \sqrt{\frac{q_{d,\alpha}}{n}} A(\hat{\theta}^n)^{-1} u, \ t.q. \ |u| \leq 1 \right\}$$

est une région de confiance de niveau asympt. $1-\alpha$ pour le paramètre θ.

Rappels sur les tests statistiques

- Un test d'hypothèse statistique vise à répondre par oui ou non à une question formulée en terme d'une hypothèse.
- Les tests non-paramétriques sont souvent non-asymptotiques et ne font pas intervenir de distributions connues, voire pas d'hypothèse de modèle paramétrique du tout.
- Les tests paramétriques supposent que les observations suivent un modèle paramétrique, et les hypothèses sont formulées en terme des paramètres du modèle. On s'intéresser le plus souvent à des propriétés asymptotiques en la taille n de l'échantillon.

Un test statistique consiste à définir deux hypothèses :

- l'hypothèse nulle H_0, qui est l'hypothèse communément admise pour laquelle on souhaite savoir si les observations permettent de la réfuter ;
- l'hypothèse alternative H_1, qui peut être par exemple la négation de l'hypothèse nulle.

Rappels sur les tests statistiques (suite)

- On se fixe un seuil de confiance $\alpha \in]0, 1[$ (valeur typique $\alpha = 5\%$).
- Un test repose sur une statistique T, appelée statistique de test, à partir de laquelle un critère de rejet est défini.
- Ce critère prend la forme d'une région de rejet telle que, si la valeur observée de T est dans cette région, l'hypothèse H_0 est rejetée ; sinon, elle est dite acceptée.
- Le plus souvent, $T \in \mathbb{R}$ et la région de rejet a pour forme $\{T \geq t\}$ pour un seuil de rejet $t \in \mathbb{R}$ à déterminer.
- On appelle erreur de première espèce la probabilité de rejeter H_0 lorsque H_0 est vraie.
- Le test est de niveau α si l'erreur de première espèce est α, et de niveau asymptotique α si la limite de l'erreur de première espèce quand la taille de l'échantillon tend vers l'infini est α

Rappels sur les tests statistiques (fin)

- Si $T \in \mathbb{R}$ et la région de rejet est de la forme $\{T \geq t\}$, on appelle p-valeur la probabilité sous l'hypothèse H_0 que T soit supérieure ou égale à t_0, la valeur observée de la statistique T.
- C'est une façon de mesurer la certitude avec laquelle on rejette l'hypothèse H_0 : plus la p-valeur est petite, plus la valeur observée t_0 est irréaliste sous l'hypothèse H_0.
- Le risque de première espèce et la p-valeur ne dépendent que de H_0.
- Le risque de seconde espèce β est la probabilité d'accepter H_0 alors que H_1 est vraie, et la puissance du test est la probabilité de rejeter H_0 lorsque H_1 est vraie.
- Une fois le niveau du test fixé (en choisissant une région de rejet qui ne dépend que de H_0), la qualité d'un test se mesure à la petitesse de son risque de seconde espèce (qui dépend à la fois de H_1 et H_0).

Test de Wald

- Soit un modèle statistique paramétrique i.i.d. de la forme $(\mathcal{H}^n, \{\mathbb{Q}_\theta^{\otimes n}\}_{\theta\in\Theta})$.
- Le test de Wald est un test paramétrique visant à tester une valeur précise du paramètre θ, exploitant la normalité asymptotique de l'estimateur du maximum de vraisemblance.
- Dans ce cas, l'hypothèse nulle est

$$(H_0) \quad \theta = \theta_0$$

pour un θ_0 fixé.

- L'hypothèse alternative peut être

$$(H_1) \quad \theta \neq \theta_0$$

ou

$$(H_1') \quad \theta = \theta_1$$

pour un certain $\theta_1 \neq \theta_0$.

Test de Wald : statistique de test et niveau asymptotique

- Soit $\alpha \in]0,1[$. Nous avons vu que

$$\sqrt{n}A(\hat{\theta}_n)(\hat{\theta}_n - \theta) \xrightarrow[n\to+\infty]{\text{loi sous } \mathbb{P}_\theta} G,$$

où $G \sim \mathcal{N}_d(0, \mathrm{Id})$.

- Donc

$$\left|\sqrt{n}A(\hat{\theta}_n)(\hat{\theta}_n - \theta)\right|^2 \xrightarrow[n\to+\infty]{\text{loi sous } \mathbb{P}_\theta} \chi^2(d).$$

- On définit donc la statistique de test

$$T_n = \left|\sqrt{n}A(\hat{\theta}_n)(\hat{\theta}_n - \theta_0)\right|^2,$$

le seuil de rejet $t = q_{d,\alpha}$ et la région de rejet $\{T_n \geq q_{d,\alpha}\}$.

- Le test de Wald ainsi construit est asymptotiquement de niveau α, puisque l'erreur de première espèce est $\mathbb{P}_{\theta_0}(T_n \geq q_{d,\alpha})$ et

$$\lim_{n\to+\infty} \mathbb{P}_{\theta_0}(T_n \geq q_{d,\alpha}) = \mathbb{P}(|G|^2 \geq q_{d,\alpha}) = \alpha.$$

Test de Wald : erreur de seconde espèce

- Si l'hypothèse alternative est $(H_1')\ \theta = \theta_1$, l'erreur de seconde espèce est

$$\beta_n = \mathbb{P}_{\theta_1}(T_n \leq q_{d,\alpha}).$$

- Or, sous $\mathbb{P}_{\theta_1}$,

$$\begin{aligned}\sqrt{T_n} &= \left|\sqrt{n}A(\hat{\theta}_n)(\hat{\theta}_n - \theta_1) + \sqrt{n}A(\hat{\theta}_n)(\theta_1 - \theta_0)\right| \\ &\geq \left|\sqrt{n}A(\hat{\theta}_n)(\theta_1 - \theta_0)\right| - \left|\sqrt{n}A(\hat{\theta}_n)(\hat{\theta}_n - \theta_1)\right|,\end{aligned}$$

où le second terme du membre de droite converge en loi sous $\mathbb{P}_{\theta_1}$ vers $|G|$, et le premier terme du membre de droite est une constante de la forme $a\sqrt{n}$ pour un certain $a > 0$.

Test de Wald (fin)

- Donc

$$\beta_n = \mathbb{P}_{\theta_1}(\sqrt{T_n} \leq \sqrt{q_{d,\alpha}}) \leq \mathbb{P}_{\theta_1}\left(\left|\sqrt{n}A(\hat{\theta}_n)(\hat{\theta}_n - \theta_1)\right| \geq a\sqrt{n} - \sqrt{q_{d,\alpha}}\right).$$

- En particulier, pour toute constant $C > 0$, à partir d'un certain rang n,

$$\beta_n \leq \mathbb{P}_{\theta_1}\left(\left|\sqrt{n}A(\hat{\theta}_n)(\hat{\theta}_n - \theta_1)\right| \geq C\right) \xrightarrow[n\to+\infty]{} \mathbb{P}(|G| \geq C).$$

- Ceci est vrai pour tout $C > 0$ et $\mathbb{P}(|G| \geq C) \to 0$ quand $C \to +\infty$, donc

$$\lim_{n\to+\infty} \beta_n = 0.$$

- Ainsi, la puissance asymptotique du test $1 - \beta_n$ est 1.

- Estimations plus fines : on peut en fait montrer que $\beta_n \leq Ce^{-bn}$, de sorte qu'on peut déterminer n tel que la puissance du test soit supérieure à tout seuil fixé dans $]0, 1[$.