Estimation ponctuelle: Estimation par maximum de vraisemblance

Statistique mathématique
M2 santé publique, université Paris-Sud

16 octobre 2018

La fonction de vraisemblance

Définition

Soit $\mathbf{X} = (X_1, \ldots, X_n)$ des variables aléatoires de densité jointe $f(\mathbf{x}; \theta)$ (ou fonction de masse) où $\theta \in \Theta \subseteq \mathbb{R}^p$. La fonction de vraisemblance $L(\theta)$ est la fonction aléatoire

$$L(\theta) = f(\mathbf{X}; \theta).$$

Notons qu'on considère L comme une fonction de θ et pas de $\mathbf{X}$.
Interprétation dans le cas discret ?

La fonction de vraisemblance

Définition

Soit $\mathbf{X} = (X_1, \ldots, X_n)$ des variables aléatoires de densité jointe $f(\mathbf{x}; \theta)$ (ou fonction de masse) où $\theta \in \Theta \subseteq \mathbb{R}^p$. La fonction de vraisemblance $L(\theta)$ est la fonction aléatoire

$$L(\theta) = f(\mathbf{X}; \theta).$$

Notons qu'on considère L comme une fonction de θ et pas de $\mathbf{X}$.
Interprétation dans le cas discret ?
Lorsque $\mathbf{X}$ est un échantillon (c'est à dire des v.a iid) de densité $f(\cdot; \theta)$, alors la vraisemblance est

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta).$$

Estimateur du maximum de vraisemblance (MLE)

Définition (estimateur du maximum de vraisemblance)

Soit $\mathbf{X} = (X_1, \ldots, X_n)$ un échantillon aléatoire issu de F_θ, et soit $\widehat{\theta}$ tel que

$$L(\widehat{\theta}) \geq L(\theta), \quad \forall \theta \in \Theta.$$

Alors $\widehat{\theta}$ est appelé *un estimateur du maximum de vraisemblance* de θ.

Estimateur du maximum de vraisemblance (MLE)

Définition (estimateur du maximum de vraisemblance)

Soit $\mathbf{X} = (X_1, \ldots, X_n)$ un échantillon aléatoire issu de F_θ, et soit $\widehat{\theta}$ tel que

$$L(\widehat{\theta}) \geq L(\theta), \quad \forall \theta \in \Theta.$$

Alors $\widehat{\theta}$ est appelé *un estimateur du maximum de vraisemblance* de θ.

On appelle $\widehat{\theta}$ l' estimateur du maximum de vraisemblance, s'il est l'unique maximum de $L(\theta)$,

$$\widehat{\theta} = \underset{\theta \in \Theta}{\text{argmax}} L(\theta).$$

Commentaires sur le MLE

Les estimateurs par la méthode des moments et plug-in ne dépendent pas que de statistiques exhaustives
$\hookrightarrow$ mais utilisent aussi de l'information non pertinente (libre).

- Si T est une statistique exhaustive pour θ, alors le théorème de factorisation nous dit que

$$L(\theta) = g\big(T(\mathbf{X}); \theta\big) h\big(\mathbf{X}\big) \propto g\big(T(\mathbf{X}); \theta\big)$$

  i.e. **tout** MLE ne dépend des données qu'à travers la statistique exhaustive

- Le MLE est **invariant**. Si $g : \Theta \to \Theta'$ une bijection, et $\widehat{\theta}$ le MLE de θ, alors $g\big(\widehat{\theta}\big)$ est le MLE de $g(\theta)$.

Commentaires sur le MLE

- Lorsque le support de la densité dépend du paramètre, la maximisation se fait en général directement (voir l'exemple de la loi uniforme).
- Pour une large classe de modèles statistique, la vraisemblance peut être maximisée via un calcul différentiel. Si Θ est ouvert, le support de la vraisemblance ne dépend pas de θ et la vraisemblance est différentiable, alors le MLE vérifie l'équation de la vraisemblance

$$\nabla_\theta \log L(\theta) = 0.$$

- Notons que maximiser la vraisemblance $\log L(\theta)$ revient (équivalent) à maximiser $L(\theta)$
- Lorsque Θ n'est pas un ouvert, l'équation de la vraisemblance peut être utilisée à condition de vérifier que la maximum n'est pas sur le bord de Θ ! ! !.

Exemples

Exemple : loi uniforme

Soit $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{U}[0, \theta]$. La vraisemblance est

$$L(\theta) = \theta^{-n} \prod_{i=1}^{n} \mathbf{1}\{0 \leq X_i \leq \theta\} = \theta^{-n} \mathbf{1}\{\theta \geq X_{(n)}\}.$$

Si $\theta \leq X_{(n)}$ la vraisemblance est nulle. Sur le domaine $[X_{(n)}, \infty[$, la vraisemblance est une fonction décroissante de θ. Donc $\widehat{\theta} = X_{(n)}$.

Exemples

Exemple : loi uniforme

Soit $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{U}[0, \theta]$. La vraisemblance est

$$L(\theta) = \theta^{-n} \prod_{i=1}^{n} \mathbf{1}\{0 \leq X_i \leq \theta\} = \theta^{-n} \mathbf{1}\{\theta \geq X_{(n)}\}.$$

Si $\theta \leq X_{(n)}$ la vraisemblance est nulle. Sur le domaine $[X_{(n)}, \infty[$, la vraisemblance est une fonction décroissante de θ. Donc $\widehat{\theta} = X_{(n)}$.

Exemple : loi de Poisson

Soit $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$. Alors

$$L(\lambda) = \prod_{i=1}^{n} \left\{ \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \right\} \Rightarrow \log L(\lambda) = -n\lambda + \log \lambda \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \log(x_i!).$$

On a $\nabla_\lambda \log L(\lambda) = -n + \lambda^{-1} \sum_{i=1}^{n} x_i = 0$, on obtient $\widehat{\lambda} = \bar{x}$ car $\nabla^2_\lambda \log L(\lambda) = -\lambda^{-2} \sum x_i < 0$.

Plan

- Le problème d'estimation ponctuelle
- Estimateur du maximum de vraisemblance
- Relation avec la divergence de Kullback-Leibler
- Propriétés asymptotiques du MLE

Le folklore

Rappelons le point de départ :

- Collection de v.a (un vecteur aléatoire) $X = (X_1, \ldots, X_n)$
- $X \sim F_\theta \in \mathcal{F}$
- $\mathcal{F}$ une famille paramétrique de paramètre $\theta \in \Theta \subseteq \mathbb{R}^d$

Le problème de l'estimation ponctuelle

- Supposons que F est complètement définie par son paramètre θ qui inconnu
- Soit $(x_1, \ldots, x_n)$ des réalisations de $X \sim F_\theta$
- Estimer la valeur de θ qui a *généré* les réalisations $(x_1, \ldots, x_n)$

Estimaton ponctuelle dans les familles paramétriques

Nous avons vu

- Méthode d'estimation plug-in
- Méthode des moments
- Méthode d'estimation par maximum de vraisemblance MLE

Maintenant : zoom sur le maximum de vraisemblance. Pourquoi fait-il sens ? Quelle sont ses propriétés ?

MLE

Rappelons notre définition de l'estimateur par maximum de vraisemblance

Définition (estimateur du maximum de vraisemblance)

Soit $\mathbf{X} = (X_1, \ldots, X_n)$ un échantillon aléatoire issu de F_θ, et soit $\widehat{\theta}$ tel que

$$L(\widehat{\theta}) \geq L(\theta), \quad \forall \theta \in \Theta.$$

Alors $\widehat{\theta}$ est appelé *un estimateur du maximum de vraisemblance* de θ.

On appelle $\widehat{\theta}$ l' estimateur du maximum de vraisemblance, s'il est l'unique maximum de $L(\theta)$,

$$\widehat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}} L(\theta).$$

MLE

Rappelons notre définition de l'estimateur par maximum de vraisemblance

Définition (estimateur du maximum de vraisemblance)

Soit $\mathbf{X} = (X_1, \ldots, X_n)$ un échantillon aléatoire issu de F_θ, et soit $\widehat{\theta}$ tel que

$$L(\widehat{\theta}) \geq L(\theta), \quad \forall \theta \in \Theta.$$

Alors $\widehat{\theta}$ est appelé *un estimateur du maximum de vraisemblance* de θ.

On appelle $\widehat{\theta}$ l' estimateur du maximum de vraisemblance, s'il est l'unique maximum de $L(\theta)$,

$$\widehat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}} L(\theta).$$

- $\rightarrow \widehat{\theta}$ rend ce qu'on observe le plus probable, le plus vraisemblable.
- Il fait sens intuitivement. Mais pourquoi fait-il sens mathématiquement ?

La divergence de Kullback-Leibler

Définition : divergence de Kullback-Leibler

Soit $p(x)$ et $q(x)$ deux fonctions de densité (ou fonctions de masse) sur $\mathbb{R}$. La divergence de Kullback-Leibler de q par rapport à p est donnée par

$$KL[q||p] = \int_{-\infty}^{+\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx.$$

- On a $KL[p||p] = \int_{-\infty}^{+\infty} p(x) \log(1) dx = 0$.
- Par l'inégalité de Jensen, pour $X \sim p(\cdot)$, nous avons

$$KL[q||p] = \mathbb{E}\left\{-\log\left[q(X)/p(X)\right]\right\} \geq -\log\left\{\mathbb{E}\left[\frac{q(X)}{p(X)}\right]\right\} = 0.$$

- $p \neq q \Longrightarrow KL[q||p] > 0$.
- KL mesure *la distance* entre les lois de probabilités.
- KL n'est pas une métrique : non symétrique et pas d'inégalité triangulaire ! ! !

Vraisemblance et divergence KL

Lemme

Un estimateur $\widehat{\theta}$ basé sur un échantillon iid $X_1, \ldots, X_n$ est un MLE si et seulement si $KL\left[F\left(x;\widehat{\theta}\right)||\widehat{F}_n(x)\right] \leq KL\left[F\left(x;\theta\right)||\widehat{F}_n(x)\right] \quad \forall \theta \in \Theta$.

Vraisemblance et divergence KL

Lemme

Un estimateur $\widehat{\theta}$ basé sur un échantillon iid $X_1, \ldots, X_n$ est un MLE si et seulement si $KL\left[F\left(x;\widehat{\theta}\right)||\widehat{F}_n(x)\right] \leq KL\left[F\left(x;\theta\right)||\widehat{F}_n(x)\right] \quad \forall \theta \in \Theta$.

Preuve : Rappelons que $\int h(X) d\widehat{F}_n(x) = n^{-1} \sum_{i=1}^{n} h\left(X_i\right)$ donc,

$$KL\left[F_\theta||\widehat{F}_n\right] = \int_{-\infty}^{+\infty} \log\left(\frac{n^{-1}\sum_{i=1}^{n} \mathbf{1}_{X_i}(x)}{f(x;\theta)}\right) d\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{n^{-1}}{f(X_i;\theta)}\right)$$

$$= -\frac{1}{n}\sum_{i=1}^{n} \log n - \frac{1}{n}\sum_{i=1}^{n} \log f(X_i;\theta)$$

$$= -\log n - \frac{1}{n}\log\left[\prod_{i=1}^{n} f(X_i;\theta)\right] = -\log n - \frac{1}{n}\log L(\theta)$$

MLE est divergence KL

Intuitions :

- $\widehat{F}_n$ est (avec probabilité 1) la meilleure approximation uniforme quand $n \uparrow \infty$ de F_{θ_0} où θ_0 est le vrai paramètre.
- F_{θ_0} est *très proche* de $\widehat{F}_n$ quand $n \uparrow \infty$.
- Regarder la *projection* de $\widehat{F}_n$ sur la famille $\left\{F_\theta\right\}_{\theta \in \Theta}$ comme un estimateur de F_{θ_0}.
- *Projection* au sens de Kullback-Leibler ! ! !

MLE est divergence KL

Intuitions :

- $\widehat{F}_n$ est (avec probabilité 1) la meilleure approximation uniforme quand $n \uparrow \infty$ de F_{θ_0} où θ_0 est le vrai paramètre.
- F_{θ_0} est *très proche* de $\widehat{F}_n$ quand $n \uparrow \infty$.
- Regarder la *projection* de $\widehat{F}_n$ sur la famille $\left\{F_\theta\right\}_{\theta \in \Theta}$ comme un estimateur de F_{θ_0}.
- *Projection* au sens de Kullback-Leibler ! ! !
- $KL(q||p)$ permet de distinguer si une observation X provient de q ou p sachant qu'elle provient de p ! !

Comportement asymptotique du MLE

- Les conditions de convergence du MLE ?
- Comment la loi de $\widehat{\theta}_{\text{MLE}}$ se concentre autour de θ quand $n \uparrow \infty$?

Il arrive souvent que le MLE soit le même que l'estimateur par la méthode des moments. On peut montrer la consistence directement dans ces cas.

Comportement asymptotique du MLE

- Les conditions de convergence du MLE ?
- Comment la loi de $\widehat{\theta}_{\text{MLE}}$ se concentre autour de θ quand $n \uparrow \infty$?

Il arrive souvent que le MLE soit le même que l'estimateur par la méthode des moments. On peut montrer la consistence directement dans ces cas.

Exemple : loi géométrique

Soit $X_1, \ldots, X_n$ iid de loi géométrique de paramètre θ. La fonction de masse est donnée par

$$f(x;\theta) = \theta(1-\theta)^{x-1} \text{ où } x = 1, 2, 3 \ldots$$

Notons que $\mathbb{E}(X) = \dfrac{1}{\theta}$ et $\text{Var}(X) = \dfrac{1-\theta}{\theta^2}$.

- Calculer le MLE de θ.
- Est-il consistent ?
- Quelle est sa loi limite ?

Exemple : loi uniforme

Exemple : loi uniforme

Soit $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{U}[0, \theta]$. Le MLE de θ est $\widehat{\theta} = X_{(n)} = \max\{X_1, \ldots, X_n\}$ de fonction de répartition $\mathbb{P}\left[\widehat{\theta} \leq x\right] = (x/\theta)^n \mathbf{1}\left\{x \in [0, \theta]\right\}$. Pour $\varepsilon > 0$,

$$\mathbb{P}\left[|\widehat{\theta} - \theta| > \varepsilon\right] = \mathbb{P}\left[\widehat{\theta} < \theta - \varepsilon\right] = \left(\frac{\theta - \varepsilon}{\theta}\right)^n \overset{n \to \infty}{\longrightarrow} 0,$$

donc, le MLE est convergent.

Exemple : loi uniforme

Loi uniforme

Montrons la concentration asymptotique de dist$\left(\widehat{\theta}\right)$ autour θ,

$$\begin{aligned}\mathbb{P}\left[n(\theta-\widehat{\theta}_n)\leq x\right] &= \mathbb{P}\left[\widehat{\theta}\geq \theta-\frac{x}{n}\right]\\ &= 1-\left(1-\frac{x}{n\theta}\right)^n\\ &\overset{n\to\infty}{\longrightarrow} 1-\exp\left[-\frac{x}{\theta}\right]\end{aligned}$$

donc $n\left(\theta-\widehat{\theta}_n\right)$ converge en loi vers v.a exponentielle. On déduit la concentration de dist$\left(\theta-\widehat{\theta}_n\right)$ autour de 0 quand $n\uparrow\infty$. à la vitesse n.

Maintenant, on suppose que $X_1,\ldots,X_n$ sont iid de densité (ou fonction de masse) $f(x;\theta),\theta\in\mathbb{R}$. On note

- $\ell(x;\theta)=\log f(x;\theta)$
- $\ell'(x;\theta),\ell''(x;\theta)$ et $\ell'''(x;\theta)$ ses dérivées successives par rapport à θ.

Comportement asymptotique du MLE

Conditions de régularité

(A1) Θ est ouvert de $\mathbb{R}$.

(A2) Le support de f est indépendant de θ.

(A3) f est trois fois dérivables par rapport à θ pour tout $x \in \text{Supp} f$.

(A4) $\mathbb{E}\left[\ell'(X_i;\theta)\right] = 0 \forall\theta$ et $\text{Var}_\theta\left[\ell'(X_i;\theta)\right] = I(\theta) \in (0,\infty), \forall\theta$.

(A5) $-\mathbb{E}\left[\ell''(X_i;\theta)\right] = J(\theta) \in (0,\infty) \forall\theta$.

(A6) Pour tout θ et $\delta > 0$, $\quad \exists M(x) > 0$ tels que $\mathbb{E}\left[M(X_i)\right] < \infty$ et

$$\left|\theta - \theta_0\right| < \delta \Longrightarrow \left|\ell'''(x;\theta_0)\right| \leq M(x).$$

Regardons de près ces différentes conditions ...
Si Θ est un ouvert, et si θ_0 est la vraie valeur du paramètre, la loi de $\widehat{\theta}$ peut être symétrique autour de θ_0 (une gaussienne).

Comportement asymptotique du MLE

Sous la condition (A2) nous avons $\frac{\partial}{\partial\theta}\int_{\text{Supp}f} f(x;\theta)dx = 0$ pour tout $\theta \in \Theta$ donc, on peut échanger les signes $\int$ et $\partial/\partial\theta$,

$$0 = \int \frac{\partial}{\partial\theta} f(x;\theta)dx = \int \ell'(x;\theta)f(x;\theta)dx = \mathbb{E}_\theta\left[\ell'(X_i;\theta)\right]$$

donc en présence de (A2), (A4) est essentiellement une condition qui permet la dérivation sous le signe $\int$ et assure que la v.a ℓ' possède un moment d'ordre deux pour tout θ. De manière similaire (A5) assure que ℓ'' possède un moment d'ordre un pour tout θ.
On peut aussi dériver une seconde fois sous le signe $\int$

$$0 = \int \frac{\partial}{\partial\theta}\left[\ell'(x;\theta)f(x;\theta)\right]dx = \int \ell''(x;\theta)f(x;\theta)dx + \int \left[\ell'(x;\theta)\right]^2 f(x;\theta)dx$$

donc $I(\theta) = J(\theta)$.

Famille exponentielle

Soit $X_1, \ldots, X_n$ des v.a iid de loi de la famille exponentielle à un paramètre $f(x;\theta) = \exp\left[c(\theta)T(x) - d(\theta) + S(x)\right], \quad x \in \text{Supp} f$.
Il suit que

$$\ell'(x;\theta) = c'(\theta)T(x) - d'(\theta)$$

et

$$\ell''(x;\theta) = c''(\theta)T(x) - d''(\theta).$$

Famille exponentielle

Soit $X_1, \ldots, X_n$ des v.a iid de loi de la famille exponentielle à un paramètre $f(x;\theta) = \exp\big[c(\theta)T(x) - d(\theta) + S(x)\big], \quad x \in \text{Supp} f$.
Il suit que

$$\ell'(x;\theta) = c'(\theta)T(x) - d'(\theta)$$

et

$$\ell''(x;\theta) = c''(\theta)T(x) - d''(\theta).$$

On sait que

$$\mathbb{E}\big[T(X_i)\big] = \frac{d'(\theta)}{c'(\theta)}$$

et

$$\text{Var}\big[T(X_i)\big] = \frac{1}{\big[c'(\theta)\big]^2}\left(d''(\theta) - c''(\theta)\frac{d'(\theta)}{c'(\theta)}\right).$$

Donc $\mathbb{E}\big[\ell'(X_i;\theta)\big] = c'(\theta)\mathbb{E}\big[T(X_i)\big] - d'(\theta) = 0.$

Famille exponentielle

De plus,

$$I(\theta) = \left[c'(\theta)\right]^2 \mathrm{Var}\Big[T(X_i)\Big]$$
$$= d''(\theta) - c''(\theta)\frac{d'(\theta)}{c'(\theta)}$$

et

$$J(\theta) = d''(\theta) - c''(\theta)\mathbb{E}\Big[T(X_i)\Big]$$
$$= d''(\theta) - c''(\theta)\frac{d'(\theta)}{c'(\theta)}$$

donc $I(\theta) = J(\theta)$.

Normalité asymptotique du MLE

Théorème : loi asymptotique du MLE

Soit $X_1, \ldots, X_n$ des v.a iid de densité (fonction de masse) $f(x;\theta)$ vérifiant les hypothèses (A1)-(A6). Supposons que la suite des MLE $\widehat{\theta}_n$ vérifie $\widehat{\theta}_n \xrightarrow{p} \theta$ où

$$\sum_{i=1}^{n} \ell'\left(X_i;\widehat{\theta}_n\right) = 0, n = 1, 2, ...$$

Alors,

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{I(\theta)}{J^2(\theta)}\right).$$

Lorsque $I(\theta) = J(\theta)$, nous avons

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{I(\theta)}\right)$$

Preuve

Sous les conditions (A1)-(A6), si $\widehat{\theta}_n$ maximise la vraisemblance, nous avons

$$\sum_{i=1}^{n} \ell'(X_i; \widehat{\theta}_n) = 0.$$

Développement de Taylor autour de $\widehat{\theta}_n$, nous avons

$$0 = \sum_{i=1}^{n} \ell'(X_i; \widehat{\theta}_n) = \sum_{i=1}^{n} \ell'(X_i; \theta) + (\widehat{\theta}_n - \theta) \sum_{i=1}^{n} \ell''(X_i; \theta) + \frac{1}{2}(\widehat{\theta}_n - \theta)^2 \sum_{i=1}^{n} \ell'''(X_i; \theta_n^*)$$

où θ_n^* se situe entre θ et $\widehat{\theta}_n$.

Preuve : suite

On divise par $\sqrt{n}$

$$0 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \ell'(X_i;\theta) + \sqrt{n}(\widehat{\theta}_n - \theta) \frac{1}{n} \sum_{i=1}^{n} \ell''(X_i;\theta) + \frac{1}{2}\sqrt{n}(\widehat{\theta}_n - \theta)^2 \frac{1}{n} \sum_{i=1}^{n} \ell'''(X_i;\theta_n^*)$$

on déduit que $\sqrt{n}(\widehat{\theta}_n - \theta)$ est égal à

$$\frac{-n^{-1/2} \sum_{i=1}^{n} \ell'(X_i;\theta)}{n^{-1} \sum_{i=1}^{n} \ell''(X_i;\theta) + (\widehat{\theta}_n - \theta)(2n)^{-1} \sum_{i=1}^{n} \ell'''(X_i;\theta_n^*)}.$$

TCL + (A4) nous donne

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \ell'(X_i;\theta) \xrightarrow{d} \mathcal{N}(0, I(\theta)).$$

Preuve : suite

Loi des grands nombres + (A5)

$$\frac{1}{n}\sum_{i=1}^{n}\ell''(X_i;\theta) \xrightarrow{p} -J(\theta).$$

Maintenant, il nous reste à montrer que le reste tend vers 0 en probabilité

$$R_n = \frac{(\widehat{\theta}_n - \theta)}{(2n)}\sum_{i=1}^{n}\ell'''(X_i;\theta_n^*).$$

Pour tout $\varepsilon > 0$, nous avons

$$\mathbb{P}\big[|R_n| > \varepsilon\big] = \mathbb{P}\big[|R_n| > \varepsilon, |\widehat{\theta}_n - \theta| > \delta\big] + \mathbb{P}\big[|R_n| > \varepsilon, |\widehat{\theta}_n - \theta| \leq \delta\big]$$

et $\mathbb{P}\big[|R_n| > \varepsilon, |\widehat{\theta}_n - \theta| > \delta\big] \leq \mathbb{P}\big[|\widehat{\theta}_n - \theta| > \delta\big] \xrightarrow{p} 0.$

Preuve : fin

Sous l'événement $\left|\widehat{\theta}_n - \theta\right| < \delta$ (A6) nous assure

$$\left|R_n\right| \leq \frac{\delta}{2n} \sum_{i=1}^{n} M(X_i).$$

Donc

$$\begin{aligned}
\mathbb{P}\left[|R_n| > \varepsilon, |\widehat{\theta}_n - \theta| \leq \delta\right] &\leq \mathbb{P}\left[|R_n| > \varepsilon, |R_n| \leq \frac{\delta}{2n} \sum_{i=1}^{n} M(X_i)\right] \\
&\leq \mathbb{P}\left[\frac{1}{n} \sum_{i=1}^{n} M(X_i) > \frac{2\varepsilon}{\delta}\right] \\
&\leq \frac{\delta}{2\varepsilon} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} M(X_i)\right] \\
&\leq \frac{\delta}{2\varepsilon} \mathbb{E}\left[\sum_{i=1}^{n} M(X_1)\right]
\end{aligned}$$

donc, en prenant δ suffisamment petit, on conclut que $R_n \xrightarrow{p} 0$.

Bilan de la preuve

Rappelons que grâce à un développement de Taylor et des hypothèses (A1)-(A6), nous avons pu écrire

$$\sqrt{n}\big(\widehat{\theta}_n - \theta\big) = \frac{\underbrace{-n^{-1/2}\sum_{i=1}^{n} \ell'\big(X_i;\theta\big)}_{(1)}}{\underbrace{n^{-1}\sum_{i=1}^{n} \ell''\big(X_i;\theta\big)}_{(2)} + \underbrace{\big(\widehat{\theta}_n - \theta\big)(2n)^{-1}\sum_{i=1}^{n} \ell'''\big(X_i;\theta_n^*\big)}_{(3)}}.$$

Bilan de la preuve

Rappelons que grâce à un développement de Taylor et des hypothèses (A1)-(A6), nous avons pu écrire

$$\sqrt{n}\big(\widehat{\theta}_n - \theta\big) = \frac{\underbrace{-n^{-1/2}\sum_{i=1}^{n} \ell'\big(X_i;\theta\big)}_{(1)}}{\underbrace{n^{-1}\sum_{i=1}^{n} \ell''\big(X_i;\theta\big)}_{(2)} + \underbrace{\big(\widehat{\theta}_n - \theta\big)(2n)^{-1}\sum_{i=1}^{n} \ell'''\big(X_i;\theta_n^*\big)}_{(3)}}.$$

Or, on vient de montrer que

- $(1) \xrightarrow{d} \mathcal{N}\big(0, I(\theta)\big)$
- $(2) \xrightarrow{p} -J(\theta)$.
- $(3) \xrightarrow{p} 0$.

Bilan de la preuve

Rappelons que grâce à un développement de Taylor et des hypothèses (A1)-(A6), nous avons pu écrire

$$\sqrt{n}\big(\widehat{\theta}_n - \theta\big) = \frac{\underbrace{-n^{-1/2}\sum_{i=1}^{n} \ell'\big(X_i;\theta\big)}_{(1)}}{\underbrace{n^{-1}\sum_{i=1}^{n} \ell''\big(X_i;\theta\big)}_{(2)} + \underbrace{\big(\widehat{\theta}_n - \theta\big)(2n)^{-1}\sum_{i=1}^{n} \ell'''\big(X_i;\theta_n^*\big)}_{(3)}}.$$

Or, on vient de montrer que

- $(1) \xrightarrow{d} \mathcal{N}\big(0, I(\theta)\big)$
- $(2) \xrightarrow{p} -J(\theta)$.
- $(3) \xrightarrow{p} 0$.

En appliquant le théorème de Slutsky, on conclut

$$\sqrt{n}\big(\widehat{\theta}_n - \theta\big) \xrightarrow{d} \mathcal{N}\left(0, \frac{I(\theta)}{J^2(\theta)}\right).$$

Convergence du MLE

La démonstration précédente suppose la convergence en probabilité du MLE vers la vraie valeur du paramètre ! ! !

Convergence du MLE

La démonstration précédente suppose la convergence en probabilité du MLE vers la vraie valeur du paramètre ! ! !
On considère la fonction aléatoire

$$\phi_n(t) = \frac{1}{n}\sum_{i=1}^{n}\Big[\log f(X_i;t) - \log f(X_i;\theta)\Big]$$

Cette fonction est maximale en $t = \widehat{\theta}_n$.
Par la loi des grands nombres, nous avons, pour tout $t \in \Theta$

$$\begin{aligned}\phi_n(t) \xrightarrow{p} \phi(t) &= \mathbb{E}\left[\log\left(\frac{f(X_i;t)}{f(X_i;\theta)}\right)\right] \\ &= -KL\big[f(\cdot;t)||f(\cdot;\theta)\big]\end{aligned}$$

- On sait que le maximum de $-KL\big[f(\cdot;t)||f(\cdot;\theta)\big]$ est 0 en $t = \theta$.

Convergence du MLE

- Est ce que le fait que $\phi_n(t) \xrightarrow{p} \phi(t)$ pour tout t où ϕ admet un unique maximum $\theta \Longrightarrow \quad \widehat{\theta}_n \xrightarrow{p} \theta$?

Hélas, la réponse est non ! ! !

Plus d'hypothèses sur ϕ_n

Nous avons besoin de plus d'hypothèses sur ϕ_n

Théorème

Supposons que $\{\phi_n(t)\}$ et $\phi(t)$ des fonctions aléatoires définies sur $\mathbb{R}$. Supposons que

- Pour tout $M > 0$, $\sup_{|t| \leq M} \left|\phi_n(t) - \phi(t)\right| \xrightarrow{p} 0$
- T_n maximise $\phi_n(t)$ et T_0 est l'unique maximum de $\phi(t)$
- $\forall \varepsilon > 0$, il existe M_ε tel que $\mathbb{P}\big[|T_n| > M_\varepsilon\big] < \varepsilon \; \forall n$

alors, $T_n \xrightarrow{p} T_0$.

Plus d'hypothèses sur ϕ_n

Nous avons besoin de plus d'hypothèses sur ϕ_n

Théorème

Supposons que $\{\phi_n(t)\}$ et $\phi(t)$ des fonctions aléatoires définies sur $\mathbb{R}$. Supposons que

- Pour tout $M > 0$, $\sup_{|t| \leq M} \left|\phi_n(t) - \phi(t)\right| \xrightarrow{p} 0$
- T_n maximise $\phi_n(t)$ et T_0 est l'unique maximum de $\phi(t)$
- $\forall \varepsilon > 0$, il existe M_ε tel que $\mathbb{P}\big[|T_n| > M_\varepsilon\big] < \varepsilon \ \forall n$

alors, $T_n \xrightarrow{p} T_0$.

Si ϕ_n est concave, alors ces hypothèses peuvent être allégées

Théorème

Supposons que $\{\phi_n(t)\}$ et $\phi(t)$ des fonctions aléatoires concaves définies sur $\mathbb{R}$. Supposons que

- $\phi_n(t) \xrightarrow{p} \phi(t)$ pour tout t.
- T_n maximise ϕ_n et T_0 est l'unique maximum de ϕ.

Alors, $T_n \xrightarrow{p} T_0$.

Cas de la famille exponentielle

Soit $X_1, \ldots, X_n$ des v.a iid d'une famille paramétrique à un seul paramètre

$$f(x;\theta) = \exp\big[c(\theta)T(x) - d(\theta) + S(x)\big], \quad x \in \text{Supp} f.$$

Le MLE de θ maximise en t la fonction

$$\phi_n(t) = \frac{1}{n}\sum_{i=1}^{n}\big[c(t)T(X_i) - d(t)\big] + \text{ un terme qui ne dépend pas de } t.$$

Si $c(\cdot)$ est continue et bijective de fonction inverse $c^{-1}(\cdot)$, on peut définir $u = c(t)$ et considérer

$$\phi_n^*(u) = \frac{1}{n}\sum_{i=1}^{n}\big[uT(X_i) - d_0(u)\big],$$

avec $d_0(u) = d\big(c^{-1}(u)\big)$. On obtient ϕ_n^* une fonction concave car sa dérivée seconde est $\big(\phi_n^*\big)''(u) = -d_0''(u)$ est négative.
petit rappel : $d_0''(u) = \text{Var}\, T(X_i)$! !

Cas de la famille exponentielle

La loi des grands nombres implique pour tout u, nous avons

$$\phi_n^*(u) \xrightarrow{p} u\mathbb{E}\big[T(X_1)\big] - d_0(u) = \phi^*(u).$$

On dérive $\phi^*(u)$ par rapport à u, on obtient un maximum solution de l'équation

$$d_0'(u) = \mathbb{E}\big[T(X_1)\big].$$

Or, on sait que

$$\mathbb{E}\big[T(X_1)\big] = d_0'\big(c(\theta)\big),$$

on a ϕ^* maximale quand $d_0'(u) = d_0'(c(\theta))$. Cette équation est vérifiée si on pose $u = c(\theta)$, donc $c(\theta)$ est un maximum de ϕ^*. La concavité de ϕ^* nous garantit son unicité. On obtient en appliquant notre théorème que si $\hat{u}_n = c\big(\widehat{\theta}_n\big)$ alors $\widehat{u}_n = c\big(\widehat{\theta}_n\big) \xrightarrow{p} c(\theta)$. Il suffit d'appliquer c^{-1} CQFD.

Cas de la famille exponentielle

La loi des grands nombres implique pour tout u, nous avons

$$\phi_n^*(u) \xrightarrow{p} u\mathbb{E}\big[T(X_1)\big] - d_0(u) = \phi^*(u).$$

On dérive $\phi^*(u)$ par rapport à u, on obtient un maximum solution de l'équation

$$d_0'(u) = \mathbb{E}\big[T(X_1)\big].$$

Or, on sait que

$$\mathbb{E}\big[T(X_1)\big] = d_0'\big(c(\theta)\big),$$

on a ϕ^* maximale quand $d_0'(u) = d_0'(c(\theta))$. Cette équation est vérifiée si on pose $u = c(\theta)$, donc $c(\theta)$ est un maximum de ϕ^*. La concavité de ϕ^* nous garantit son unicité. On obtient en appliquant notre théorème que si $\hat{u}_n = c\big(\widehat{\theta}_n\big)$ alors $\widehat{u}_n = c\big(\widehat{\theta}_n\big) \xrightarrow{p} c(\theta)$. Il suffit d'appliquer c^{-1} CQFD.

Le MLE est convergent dans le cas de la famille exponentille

Optimalité

- On s'intéresse à des estimateurs qu'on va noter $\widehat{g}$ de $g(\theta)$ il suffit de prendre $g(\theta) = \theta$ (on se place dans le cas général ! ! !)

Optimalité

- On s'intéresse à des estimateurs qu'on va noter $\widehat{g}$ de $g(\theta)$ il suffit de prendre $g(\theta) = \theta$ (on se place dans le cas général ! ! !)
- Supposons qu'on possède une collection d'estimateurs sans biais (ou asymptotiquement sans biais).

Optimalité

- On s'intéresse à des estimateurs qu'on va noter $\widehat{g}$ de $g(\theta)$ il suffit de prendre $g(\theta) = \theta$ (on se place dans le cas général ! ! !)
- Supposons qu'on possède une collection d'estimateurs sans biais (ou asymptotiquement sans biais).
- On cherche dans cette collection un estimateur qui possède une variance minimale $\Rightarrow$ minimise le MSE dans la collection.

Optimalité

- On s'intéresse à des estimateurs qu'on va noter $\widehat{g}$ de $g(\theta)$ il suffit de prendre $g(\theta) = \theta$ (on se place dans le cas général ! ! !)
- Supposons qu'on possède une collection d'estimateurs sans biais (ou asymptotiquement sans biais).
- On cherche dans cette collection un estimateur qui possède une variance minimale $\Rightarrow$ minimise le MSE dans la collection.
- On cherche une borne inférieure qui va minorer la variance de tout estimateur sans biais de $g(\theta)$. Cette borne sera une fonction de θ et un estimateur dont la variance atteint cette borne sera appelé *Uniformly Mnimum Variance Unbiased Estimator*.

Optimalité

Objectif formel

But

Pour $X_1, \ldots, X_n$ iid de densité $f(x;\theta)$ (ou fonction de masse). θ est le paramètre inconnu. On veut établir des conditions sous lesquelles nous avons

$$\text{Var}\left[\widehat{g}\right] \geq \phi(\theta), \quad \forall\theta$$

pour tout estimateur sans biais (de $g(\theta)$) $\widehat{g}$. On veut aussi déterminer la borne $\phi(\theta)$.

Inégalité de Cauchy-Schwarz

Théorème

Soit U et V deux variables aléatoires. Alors,

$$\mathrm{Cov}(U,V) \leq \sqrt{\mathrm{Var}(U)\mathrm{Var}(V)}.$$

Inégalité de Cauchy-Schwarz

Théorème

Soit U et V deux variables aléatoires. Alors,

$$\mathrm{Cov}(U, V) \leq \sqrt{\mathrm{Var}(U)\mathrm{Var}(V)}.$$

Une première utilisation directe de ce théorème nous donne la borne inférieure

$$\mathrm{Var}(\widehat{g}) \geq \frac{\mathrm{Cov}^2(\widehat{g}, U)}{\mathrm{Var}(U)}$$

valable pour toute variable aléatoire U qui possède une variance finie pour tout θ.

Inégalité de Cauchy-Schwarz

Théorème

Soit U et V deux variables aléatoires. Alors,

$$\mathrm{Cov}(U,V) \leq \sqrt{\mathrm{Var}(U)\mathrm{Var}(V)}.$$

Une première utilisation directe de ce théorème nous donne la borne inférieure

$$\mathrm{Var}(\widehat{g}) \geq \frac{\mathrm{Cov}^2(\widehat{g},U)}{\mathrm{Var}(U)}$$

valable pour toute variable aléatoire U qui possède une variance finie pour tout θ.

Question ?

Existe-t-il une v.a U où $\mathrm{Cov}^2(\widehat{g},U)$ ne dépend que de $g(\theta)$ (donc ne dépend pas spécialement de $\widehat{g}$) ?

Conditions de régularité

Supposons que θ est réel et que les conditions suivantes sont vérifiées.

C1 Le support $\left\{x : f(x;\theta) > 0\right\}$ est indépendant de θ

C2 $f(x;\theta)$ est dérivable par rapport à θ, $\forall\theta \in \Theta$.

C3
$$\mathbb{E}\left[\frac{\partial}{\partial\theta} f(X;\theta)\right] = 0.$$

C4 Pour une statistique $T = T(\mathbf{X})$ où $\mathbb{E}|T| < \infty$ et $g(\theta) = \mathbb{E}T$ dérivable, où
$$g'(\theta) = \mathbb{E}\left[T\frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)\right], \quad \forall\theta$$

Conditions de régularité

Supposons que θ est réel et que les conditions suivantes sont vérifiées.

C1 Le support $\left\{x : f(x;\theta) > 0\right\}$ est indépendant de θ

C2 $f(x;\theta)$ est dérivable par rapport à θ, $\forall\theta \in \Theta$.

C3
$$\mathbb{E}\left[\frac{\partial}{\partial\theta} f(X;\theta)\right] = 0.$$

C4 Pour une statistique $T = T(\mathbf{X})$ où $\mathbb{E}|T| < \infty$ et $g(\theta) = \mathbb{E}T$ dérivable, où
$$g'(\theta) = \mathbb{E}\left[T\frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)\right], \quad \forall\theta$$

Pour donner un sens à C3 et C4, supposons qu'on a des v.a continues

$$\frac{\partial}{\partial\theta}\int T(\mathbf{x})f(\mathbf{x};\theta)d\mathbf{x} \overset{!}{=} \int T(\mathbf{x})\frac{f(\mathbf{x};\theta)}{f(\mathbf{x};\theta)}\frac{\partial}{\partial\theta}f(\mathbf{x};\theta)d\mathbf{x}$$
$$= \int T(\mathbf{x})f(\mathbf{x};\theta)\frac{\partial}{\partial\theta}\log f(\mathbf{x};\theta)d\mathbf{x}$$

$\longrightarrow$ on peut interchanger les signes $\int$ et ∂

Borne de Cramer-Rao

Théorème

Soit $\mathbf{X} = (X_1, \ldots, X_n)$ de densité jointe $f(\mathbf{x}, \theta)$ satisfaisant les conditions C1, C2 et C3. Si une statistique T vérifie la condition C4, alors

$$\mathrm{Var}(T) \geq \frac{\left[g'(\theta)\right]^2}{I(\theta)},$$

où $I(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta} \log f(\mathbf{X}; \theta)\right)^2\right]$.

Borne de Cramer-Rao : preuve

Par l'inégalité de Cauchy-Schwarz où $U = \frac{\partial}{\partial \theta} \log f\left(\mathbf{X}; \theta\right)$,

$$\mathrm{Var}\left(T\right) \geq \frac{\mathrm{Cov}^2\left(T, \frac{\partial}{\partial \theta} \log f\left(\mathbf{X}; \theta\right)\right)}{\mathrm{Var}\left(\frac{\partial}{\partial \theta} \log f\left(\mathbf{X}; \theta\right)\right)}$$

Puisque $\mathbb{E}\left[\frac{\partial}{\partial \theta} \log f\left(\mathbf{X}; \theta\right)\right] = 0$, nous avons $\mathrm{Var}\left(\frac{\partial}{\partial \theta} \log f\left(\mathbf{X}; \theta\right)\right) = I(\theta)$.

Borne de Cramer-Rao : preuve

Par l'inégalité de Cauchy-Schwarz où $U = \frac{\partial}{\partial\theta} \log f\left(\mathbf{X};\theta\right)$,

$$\mathrm{Var}\left(T\right) \geq \frac{\mathrm{Cov}^2\left(T, \frac{\partial}{\partial\theta} \log f\left(\mathbf{X};\theta\right)\right)}{\mathrm{Var}\left(\frac{\partial}{\partial\theta} \log f\left(\mathbf{X};\theta\right)\right)}$$

Puisque $\mathbb{E}\left[\frac{\partial}{\partial\theta} \log f\left(\mathbf{X};\theta\right)\right] = 0$, nous avons $\mathrm{Var}\left(\frac{\partial}{\partial\theta} \log f\left(\mathbf{X};\theta\right)\right) = I(\theta)$.
Remarquons que

$$\begin{aligned}
\mathrm{Cov}\left(T, \frac{\partial}{\partial\theta} \log f\left(\mathbf{X};\theta\right)\right) &= \mathbb{E}\left[T \frac{\partial}{\partial\theta} \log f\left(\mathbf{X};\theta\right)\right] - \mathbb{E}\left[T\right]\mathbb{E}\left[\frac{\partial}{\partial\theta} \log f\left(\mathbf{X};\theta\right)\right] \\
&= \mathbb{E}\left[T \frac{\partial}{\partial\theta} \log f\left(\mathbf{X};\theta\right)\right] \\
&= g'(\theta)
\end{aligned}$$

Borne de Cramer-Rao

La borne de Cramer-Rao est-t-elle atteinte ? quand ?

si

$$\mathrm{Var}\big[T\big] = \frac{\big[g'(\theta)\big]^2}{I(\theta)},$$

Borne de Cramer-Rao

La borne de Cramer-Rao est-t-elle atteinte ? quand ?
si

$$\mathrm{Var}\big[T\big] = \frac{\big[g'(\theta)\big]^2}{I(\theta)},$$

alors

$$\mathrm{Var}\big[T\big] = \frac{\mathrm{Cov}^2\left(T, \frac{\partial}{\partial\theta}\log f\left(\mathbf{X};\theta\right)\right)}{\mathrm{Var}\left(\frac{\partial}{\partial\theta}\log f\left(\mathbf{X};\theta\right)\right)}.$$

Ce n'est vrai que si et seulement si $\frac{\partial}{\partial\theta}\log f\left(\mathbf{X};\theta\right)$ est une fonction linéaire (corrélation 1) de T. C'est à dire

$$\frac{\partial}{\partial\theta}\log f\left(\mathbf{x};\theta\right) = A(\theta)\,T\left(\mathbf{x}\right) + B(\theta).$$

On résout l'équation différentielle, on obtient

$$\log f\left(\mathbf{x};\theta\right) = A^*(\theta)\,T\left(\mathbf{x}\right) + B^*(\theta).$$

Borne de Cramer-Rao

La borne de Cramer-Rao est-t-elle atteinte ? quand ?
si

$$\mathrm{Var}\big[T\big] = \frac{\big[g'(\theta)\big]^2}{I(\theta)},$$

alors

$$\mathrm{Var}\big[T\big] = \frac{\mathrm{Cov}^2\left(T, \frac{\partial}{\partial\theta}\log f\left(\mathbf{X};\theta\right)\right)}{\mathrm{Var}\left(\frac{\partial}{\partial\theta}\log f\left(\mathbf{X};\theta\right)\right)}.$$

Ce n'est vrai que si et seulement si $\frac{\partial}{\partial\theta}\log f\left(\mathbf{X};\theta\right)$ est une fonction linéaire (corrélation 1) de T. C'est à dire

$$\frac{\partial}{\partial\theta}\log f\left(\mathbf{x};\theta\right) = A(\theta)\,T\left(\mathbf{x}\right) + B(\theta).$$

On résout l'équation différentielle, on obtient

$$\log f\left(\mathbf{x};\theta\right) = A^*(\theta)\,T\left(\mathbf{x}\right) + B^*(\theta).$$

Conclusion ?