Statistique Asymptotique

Agathe Guilloux et Marie-Luce Taupin
agathe.guilloux@univ-evry.fr et marieluce.taupin@univ-evry.fr

Laboratoire LaMME, Université d'Evry val d'Essonne
http ://www.math-evry.cnrs.fr/members/aguilloux/welcome et
http ://www.math-evry.cnrs.fr/members/mtaupin/welcome

2021-2022

Agresti, A. (2015).
Foundations of linear and generalized linear models.
John Wiley & Sons.

Bickel, P.-J. and K.-A. Doksum (2016).
Mathematical statistics—basic ideas and selected topics.
CRC Press, Boca Raton, FL.

Cadre, B. and C. Vial (2012, June).
Statistique mathématique, cours et exercices corrigés.
Références sciences. Ellipse.

Fahrmeir, L. and H. Kaufmann (1985).
Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models.
The Annals of Statistics, 342–368.

Fourdrinier, D. (2002).
Statistique inférentielle.
Dunod.

Gaudouin, O.
Statistique inférentielle avancée.
https ://www-ljk.imag.fr/membres/Olivier.Gaudoin/SIA.pdf.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013).
An introduction to statistical learning, Volume 6.
Springer.

Lindsey, J.-K. (2000).
Applying generalized linear models.
Springer Science & Business Media.

McCullagh, P. and J.-A. Nelder (1989).
Generalized linear models 2nd edition chapman and hall.
London, UK.

Rivoirard, V. and G. Stoltz (2012).
Statistique mathématique en action.
Vuibert.

Références bibliographiques I

Quelques références sur les deux parties

- (Rivoirard and Stoltz 2012)
- (Cadre and Vial 2012)
- (Fourdrinier 2002)
- (Gaudouin)
- (Bickel and Doksum 2016)
- (Agresti 2015) : GLM
- (Fahrmeir and Kaufmann 1985) : GLM
- (James, Witten, Hastie, and Tibshirani 2013) : GLM
- (Lindsey 2000) : GLM
- (McCullagh and Nelder 1989) : GLM

Organisation des séances

- FA+FI : inférence statistique, EMV (ML Taupin)
- FI : preuve EMV+ exhaustivité, admissibilité (ML Taupin)
- FA+FI : exercices, cas de la famille exponentielle (preuve dans ce cas là) (ML Taupin)
- FI : information de Fisher, estimation paramétrique optimale (ML Taupin)
- FA+FI : tests basés sur la vraisemblance (ML Taupin)
- FI : exercices et compléments (ML Taupin)
- FA+FI GLM (A. Guilloux)
- FI : GLM (A. Guilloux)
- FA+FI GLM (A. Guilloux)
- GLM (A. Guilloux)

Inférence statistique

Modèle statistique

Soit X une variable aléatoire de loi P_{θ^*} appartenant à une famille de loi $\{P_\theta, \theta \in \Theta\}$. On note $\mathcal{X}$ l'ensemble des valeurs possibles pour X. La loi P_{θ^*} dépendant d'un paramètre θ^* inconnu à valeurs dans un ensemble $\Theta \subset \mathbb{R}^d$. On cherche à estimer θ^* à partir de n variables aléatoires $X_1, \cdots, X_n$ indépendantes et identiquement distribuées (i.i.d.) de loi P_{θ^*} où

$$P_{\theta^*} \in \{P_\theta, \theta \in \Theta\}.$$

On parle aussi du n-échantillon $(X_1, \cdots, X_n)$ de X de loi P_{θ^*} où

$$P_{\theta^*} \in \{P_\theta, \theta \in \Theta\}.$$

Modèle statistique (2)

Definition

Le modèle statistique (ou la structure statistique) associé à cette expérience est le triplet $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ où

- $\mathcal{X}$ est l'espace des observations, ensemble de toutes les observations possibles.
- $\mathcal{A}$ est la tribu des évènements observables associée.
- $\mathcal{P}$ est une famille de lois de probabilités possibles définie sur $\mathcal{A}$.

L'intérêt de ce formalisme est de permettre de traiter la même façon tous les types d'observations possibles.
On dit que le modèle est discret quand $\mathcal{X}$ est fini ou dénombrable. Dans ce cas, la tribu $\mathcal{A}$ est l'ensemble des parties de $\mathcal{X}$: $\mathcal{A} = \mathcal{P}(\mathcal{X})$. Cette situation correspond au cas où la variable aléatoire est discrète.

Modèle statistique (3)

On dit que le modèle est continu quand $\mathcal{X}$ est inclus dans $\mathbb{R}^d$. Dans ce cas, $\mathcal{P}$ est un ensemble de lois probabilité admettant une densité par rapport à mesure de Lebesgue dans $\mathbb{R}^d$ et $\mathcal{A}$ est la tribu des boréliens de $\mathcal{X}$ (tribu engendrée par les ouverts de $\mathcal{X}$) : $\mathcal{A} = \mathcal{B}(\mathcal{X})$.
On peut aussi envisager des modèles ni continus ni discrets, par exemple si l'observation a certains éléments continus et d'autres discrets et $\mathcal{X}$ et $\mathcal{A}$ sont alors plus complexes.

Exemples (1)

• Exemple : Loi exponentielle En exercice.

• Exemple : Loi uniforme Considérons $(X_1, \cdots, X_n)$, un n-échantillon de loi uniforme continue $\mathcal{U}_{[0,\theta^*]}$. On souhaite estimer le paramètre θ^*, inconnu. Au moins deux estimateurs sont possibles :

- l'estimateur intuitif, $T_n^{(1)} = \max_{1 \leq i \leq n} X_i$
- l'estimateur des moments, $T_n^{(2)} = 2\bar{X}_n = 2n^{-1} \sum_{i=1}^n X_i$.

Exemples (2)

• Etude de $T_n^{(1)}$ Le premier estimateur $T_n^{(1)}$ est un estimateur biaisé, car $\mathbb{E}(T_n^{(1)}) = \frac{\theta^* n}{n+1} \neq \theta^*$, mais asymptotiquement sans biais puisque

$$|\mathbb{E}(T_n^{(1)}) - \theta^*| = \frac{\theta^*}{n+1} \underset{n\to\infty}{\longrightarrow} 0,$$

$$\text{EQM}(T_n^{(1)}) = \mathbb{E}[(T_n^{(1)} - \theta^*)^2] = \frac{(\theta^*)^2 n}{(n+1)^2(n+2)} \underset{n\to\infty}{\longrightarrow} 0.$$

L'estimateur $T_n^{(1)}$ est donc un estimateur consistant de θ^*, c'est-à-dire $T_n^{(1)} \underset{n\to\infty}{\overset{\mathbb{L}^2}{\longrightarrow}} \theta^*$ et $T_n^{(1)} \underset{n\to\infty}{\overset{\mathbb{P}}{\longrightarrow}} \theta^*$. Nous avons de plus que

$$n[\theta^* - T_n^{(1)}]/\theta^* \underset{n\to\infty}{\overset{\mathcal{L}}{\longrightarrow}} \mathcal{E}(1), \text{ (variable aléatoire de loi exponentielle).}$$

Nous dirons que $T_n^{(1)}$ converge vers θ^* à la vitesse $1/n$.

Exemples (3)

• Etude de $T_n^{(2)}$ Le deuxième estimateur $T_n^{(2)}$ est un estimateur sans biais de θ^*, car

$$E(T_n^{(2)}) = \theta^*.$$

De plus

$$\mathbb{E}[(T_n^{(2)} - \theta^*)^2] = \text{Var}(T_n^{(2)}) = \frac{(\theta^*)^2}{3n} \xrightarrow[n\to\infty]{} 0.$$

L'estimateur $T_n^{(2)}$ est donc un estimateur <u>consistant</u> de θ^*, $T_n^{(2)} \xrightarrow[n\to\infty]{\mathbb{L}^2} \theta^*$ et $T_n^{(2)} \xrightarrow[n\to\infty]{\mathbb{P}} \theta^*$. Par ailleurs, d'après le théorème central limite,

$$\frac{\sqrt{n}(T_n^{(2)} - \theta^*)}{\sqrt{\theta^{*2}/3}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1) \ .$$

Nous dirons que $T_n^{(2)}$ converge vers θ^* <u>à la vitesse</u> $1/\sqrt{n}$.
La vitesse de convergence d'un estimateur est donc une mesure de la qualité d'approximation de cet estimateur.

Exemples (4)

L'estimateur $T_n^{(1)}$, bien que biaisé, converge plus rapidement vers θ^* que l'estimateur $T_n^{(2)}$, qui lui est sans biais.

Outils (1)

Definition

(Convergence en probabilité)
Soient $Y_1, Y_2, \cdots, Y_n, \cdots$ une suite de v.a. et Y une v.a. On dira que Y_n converge en probabilité vers la v.a. Y si

$$\forall \epsilon > 0 \ , \lim_{n \to \infty} P\left(|Y_n - Y| > \epsilon\right) = 0$$

On note alors $Y_n \xrightarrow[n \to \infty]{\mathbb{P}} Y$.

Outils (2)

Propriété

(Loi (faible) des grands nombres)
Soient n variables aléatoires i.i.d. $X_1, X_2, \cdots, X_n$ *de même loi qu'une variable aléatoire* X *telle que* $E(X) = \mu$ *et* $Var(X) = \sigma^2$ *sont finis. Alors,*

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{\mathbb{P}} \mu$$

La loi faible des grands nombres se démontre en utilisant l'inégalité de Tchebyschev, et en remarquant que $\mathbb{E}(\bar{X}_n) = \mu$ et $\text{Var}(\bar{X}_n) = \sigma^2/n$. Rq : La loi faible des grands nombre reste vraie si $\mathbb{E}|X| < \infty$.

Outils (3)

Propriété

(Inégalité de Tchebyschev)

Soit Y une v.a d'espérance $E(Y)$ et de variance $Var(Y)$. Alors

$$\forall \epsilon > 0 \ , P\left(|Y - E(Y)| \geq \epsilon\right) \leq \frac{Var(Y)}{\epsilon^2}$$

Outils (4) : TLC

Propriété

(Théorème central limite)

Soient $X_1, X_2, \cdots, X_n, \cdots$, une suite de v.a. i.i.d. d'espérance μ et de variance $\sigma^2 < \infty$. Notons $\bar{X}_n$ la moyenne empirique (aléatoire) des n premières v.a. X_i de la suite, i.e.

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n},$$

et la variable Z_n centrée réduite,

$$Z_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}.$$

Alors pour Z de loi $\mathcal{N}(0,1)$, on a $Z_n \xrightarrow[n\to\infty]{\mathcal{L}} Z$.

Outils (5) : TLC

Plus précisément, si l'on note $F_n(x)$ la fonction de répartition de Z_n, alors

$$\lim_{n\to\infty} F_n(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{z^2}{2}} dz = F(x).$$

soit encore

$$\lim_{n\to\infty} P\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} < x \right) = P(Z < x) \ , \text{ où } Z \sim \mathcal{N}(0,1)$$

Outils (6)

Definition

Si $\mathbb{E}(|X|^p) < \infty$ pour $p \geq 1$, alors pour tout $q \leq p$, le moment d'ordre q de X est par définition

$$\mathbb{E}(X^q) < \infty.$$

Propriété

(Moment) Soit p un entier non nul, tel que $\mathbb{E}(|X|^p) < \infty$. Alors pour tout $1 \leq q \leq p$

$$\mathbb{E}(|X|^q) < \infty.$$

Outils (7)

Propriété

(delta-method)

Soit T_n un estimateur consistant d'un paramètre θ^*, et soit g une fonction de $\mathbb{R}$ dans $I \subset \mathbb{R}$, continue et dérivable en θ^*. Alors $g(T_n)$ est un estimateur consistant de $g(\theta^*)$. De plus si $v_n(T_n - \theta^*) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1)$, alors

$$v_n(g(T_n) - g(\theta^*)) \xrightarrow[n\to\infty]{\mathcal{L}} g'(\theta^*)\mathcal{N}(0,1).$$

Remarque : Si T_n est sans biais, alors $g(T_n)$ n'est pas nécessairement sans biais ; $E(T_n) = \theta^*$ n'implique pas nécessairement que $E(g(T_n)) = g(\theta^*)$.

Outils (8)

Propriété

Si X_n converge en loi vers X, et si Y_n converge en probabilité vers une constante c, alors le couple (X_n, Y_n) converge en loi vers le couple (X, c).

La conséquence immédiate est que si l'on dispose d'un estimateur consistant de la variance, $\widehat{\sigma}^2 \xrightarrow[n\to\infty]{\mathbb{P}} \sigma^2$, le TLC devient

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\widehat{\sigma}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \times \frac{\sigma}{\widehat{\sigma}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Modèle dominé

Soit X une variable aléatoire de loi $P_{\theta^*} \in \{P_\theta, \theta \in \Theta\}$. On note $\mathcal{X}$ l'ensemble des valeurs possibles pour X. La loi P_{θ^*} dépendant d'un paramètre θ^* inconnu à valeurs dans un ensemble $\Theta \subset \mathbb{R}^d$.
On suppose que $\forall \theta \in \Theta$ il existe une mesure dominante ν σ-finie, et une fonction $L_\theta(x)$ telle que pour tous $\theta \in \Theta$

$$\frac{dP_\theta}{d\nu} = L_\theta, \text{ avec } \mathbb{P}(X \in A) = \int_A L_{\theta^*}(x)\nu(dx).$$

Ceci est supposé vrai pour toutes les lois P_θ, $\forall \theta \in \Theta$. On dit alors que P_θ admet une densité $L_\theta(x)$ par rapport à la mesure dominante ν.

Modèle dominé : exemples

• Si la variable X est une variable aléatoire discrète, alors la mesure ν est la mesure de comptage, $L_{\theta^*}(x) = P(X = x)$ et on écrit

$$\mathbb{P}(X \in A) = \sum_{x \in A} P(X = x) = \int_A L_{\theta^*}(x)\nu(dx).$$

• Si la variable X est une variable aléatoire continue de densité f_X, alors la mesure ν est la mesure de Lebesgue, $L_{\theta^*}(x) = f_X(x)$ et

$$\mathbb{P}(X \in A) = \int_A L_{\theta^*}(x)\nu(dx) = \int_A f_X(x)dx.$$

Modèle paramétrique

Dans ce qui suit, le paramètre θ^* inconnu (espérance,variance,...), appartient à $\Theta \subset \mathbb{R}^d$, $d \geq 1$. Le paramètre θ^* est donc éventuellement multi-dimensionnel. Par exemple, si θ^* est l'espérance de la loi de X, θ^* peut représenter la moyenne (ν) ou la probabilité d'un évènement (p) du caractère (si X est qualitatif),... Le paramètre θ^* peut être la variance (σ^2) de la loi de X, et correspond à alors la dispersion du caractère X,$\cdots$).

Estimateur-estimation

On suppose que l'on dispose de n observations $x_1, x_2, \cdots, x_n$, mesures du caractère faites sur un échantillon de taille n : on considère que $x_1, x_2, \cdots, x_n$ est la réalisation d'un n-uplet de variables aléatoires $X_1, X_2, \cdots, X_n$ indépendantes et identiquement distribuées, de même loi que X (i.i.d.). On dit que $(X_1, X_2, \cdots, X_n)$ est un échantillon aléatoire simple, ou encore que $(X_1, X_2, \cdots, X_n)$ est un n-échantillon de la loi de X. Le problème que l'on se pose est de savoir comment estimer θ^* à partir des n-observations $(x_1, x_2, \cdots, x_n)$.

Definition

Soit $(X_1, X_2, \cdots, X_n)$ un n-échantillon d'une loi P_θ^* dépendant d'un paramètre réel inconnu θ^*. On appelle estimateur de θ^* une variable aléatoire T_n obtenue comme fonction du n-échantillon aléatoire $(X_1, X_2, \cdots, X_n)$; autrement dit $T_n = f(X_1, X_2, \cdots, X_n)$.

Remarque : Un estimateur est en fait une suite de v.a. $(T_n)_n$, $n \geq 1$; on assimile souvent l'estimateur avec le terme général de la suite T_n.

Qualité d'un estimateur

La qualité d'un estimation t_n (fonction des observations $x_1, \cdots, x_n$) est établie à partir des propriétés (probabilistes) de la variable aléatoire associée qu'est l'estimateur T_n. Les propriétés recherchées pour un estimateur sont de deux types. Nous allons d'une part nous intéresser aux propriétés dites "asymptotiques", autrement dit quand la taille n de l'échantillon tend vers l'infini, et d'autre part aux propriétés non asymptotiques.

Consistance en probabilité

La propriété principale requise sera la convergence de l'estimateur vers la valeur du paramètre θ^* à estimer, quand la taille n de l'échantillon tend vers l'infini. Cette propriété s'appelle la consistance de l'estimateur. L'estimateur T_n étant une variable aléatoire, il existe plusieurs notions de convergence, ou de consistance d'un estimateur.

Definition

L'estimateur T_n de θ^* sera consistant si il converge en probabilité vers θ^* quand n tend vers l'infini c'est-à-dire pour tout $\varepsilon > 0$ on a

$$\mathbb{P}(|T_n - \theta^*| \geq \varepsilon) \underset{n\to\infty}{\longrightarrow} 0.$$

On note alors

$$T_n \underset{n\to\infty}{\xrightarrow{\mathbb{P}}} \theta^*.$$

Consistance en moyenne quadratique

Definition

L'estimateur T_n de θ^* converge en moyenne qudratique vers θ^* si

$$EQM(T_n) = \mathbb{E}\left[(T_n - \theta^*)^2\right] \underset{n\to\infty}{\longrightarrow} 0.$$

On note alors

$$T_n \underset{n\to\infty}{\overset{\mathbb{L}^2}{\longrightarrow}} \theta^*.$$

Rq : la convergence en moyenne quadratique implique la convergence en probabilité ;

Décomposition biais-variance

La convergence en moyenne quadratique, plus facile en général à établir que la convergence en probabilité, se montre en utilisant la décomposition biais-variance.

Théorème

Soit T_n une variable aléatoire. Alors on a l'égalité

$$\mathbb{E}\left[(T_n - \theta^*)^2\right] = Var(T_n) + [\mathbb{E}(T_n) - \theta^*]^2.$$

Definition

On appelle biais de l'estimateur T_n, la quantité $B(T_n) = \mathbb{E}(T_n) - \theta^*$.

Convergence en probabilité - convergence en moyenne quadratique

D'après la décomposition biais-variance, une condition nécessaire et suffisante pour que l'erreur quadratique moyenne converge vers 0 est que

- $\text{Var}(T_n)$ tend vers 0 quand n tend vers l'infini,

et

- $[\mathbb{E}(T_n) - \theta]^2$ tend vers 0 quand n tend vers l'infini.

Autrement dit pour qu'un estimateur converge en moyenne quadratique, il faut et il suffit que sa variance tende vers 0 quand n tend vers l'infini, **et** qu'il soit asymptotiquement sans biais.

Convergence en probabilité - convergence en moyenne quadratique

Bien souvent , la convergence en probabilité se montre à l'aide de l'inégalité de Markov et se déduit de la convergence en moyenne quadratique. En effet, l'inégalité de Markov implique que

$$\mathbb{P}(|T_n - \theta| \geq \varepsilon) \quad = \quad \mathbb{P}(|T_n - \theta|^2 \geq \varepsilon^2) \leq \frac{\mathbb{E}[(T_n - \theta)^2]}{\varepsilon^2}.$$

Si $\mathbb{E}[(T_n - \theta)^2]$ converge vers 0 quand n tend vers l'infini, alors pour tout $\varepsilon > 0$, $\mathbb{P}(|T_n - \theta| \geq \varepsilon)$ tend également vers 0 quand n tend vers l'infini. Par conséquent, la convergence en moyenne quadratique implique la convergence en probabilité.

Biais d'un estimateur

Soit T_n est un estimateur de θ^*. Une propriété non asymptotique d'un estimateur est l'absence ou non de biais. Autrement dit T_n est-il un estimateur sans biais ? On calcule $E(T_n)$. On dit que T_n est un estimateur sans biais si $\mathbb{E}(T_n) = \theta^*$. Sinon T_n est dit biaisé et le biais de T_n est donné par $B(T_n) = E(T_n) - \theta^*$.
Si $B(T_n)$ tend vers 0 quand n tend vers l'infini, alors T_n est dit asymptotiquement sans biais.

Comparaison d'estimateurs

Si T_n^1 et T_n^2 sont deux estimateurs de θ^*, on choisira celui d'erreur quadratique la plus petite. Si $EQM(T_n^1) \leq EQM(T_n^2)$, on dit que T_n^1 est un meilleur estimateur de θ^* que T_n^2.

Notion de vitesse de convergence sur des exemples (1)

• Exemple : Estimation d'une probabilité Pour estimer une probabilité p^* à partir d'un n-échantillon $(X_1, \cdots, X_n)$ de la loi $\mathcal{B}(p^*)$, l'estimateur intuitif, $T_n = n^{-1} \sum_{i=1}^n X_i$, qui représente la proportion aléatoire de l'échantillon, est un estimateur sans biais de p car $E(T_n) = p^*$, et de variance $\text{Var}(T_n) = \dfrac{p^*(1-p^*)}{n} \underset{n\to\infty}{\longrightarrow} 0$. Par conséquent il converge en moyenne quadratique vers p^*, et est donc un estimateur consistant de p^*, $T_n \underset{n\to\infty}{\overset{\mathbb{L}^2}{\longrightarrow}} p^*$. D'après le théorème central limite,

$$\frac{\sqrt{n}(T_n - p^*)}{\sqrt{p^*(1-p^*)}} \underset{n\to\infty}{\overset{\mathcal{L}}{\longrightarrow}} \mathcal{N}(0,1) \ .$$

Nous dirons que T_n converge vers p^* à la vitesse $1/\sqrt{n}$.

Notion de vitesse de convergence sur des exemples (2)

• Exemple : loi uniforme Considérons $(X_1, \cdots, X_n)$, un n-échantillon de loi uniforme continue $\mathcal{U}_{[0,\theta^*]}$. On souhaite estimer le paramètre θ^*, inconnu. Nous avons vu qu'au moins deux estimateurs sont possibles :

- l'estimateur intuitif, $T_n^{(1)} = \max_{1 \leq i \leq n} X_i$
- l'estimateur des moments, $T_n^{(2)} = 2\bar{X}_n = 2n^{-1} \sum_{i=1}^n X_i$.

Notion de vitesse de convergence sur des exemples (2)

Ces estimateurs satisfont

- $$\mathbb{E}(T_n^{(1)}) = \frac{\theta^* n}{n+1} \neq \theta^*,$$

  et

  $$|\mathbb{E}(T_n^{(1)}) - \theta^*| = \frac{\theta^*}{n+1} \xrightarrow[n\to\infty]{} 0.$$

  De plus

  $$\mathbb{E}[(T_n^{(1)} - \theta^*)^2] = \frac{\theta^2 n}{(n+1)^2(n+2)} \xrightarrow[n\to\infty]{} 0,$$

  et

  $$n[\theta^* - T_n^{(1)}]/\theta \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{E}(1),$$

  où $\mathcal{E}(1)$ est une variable aléatoire de loi exponentielle de paramètre $\lambda = 1$.

Notion de vitesse de convergence sur des exemples (3)

- $$E(T_n^{(2)}) = \theta^*,$$

  et

  $$\mathbb{E}[(T_n^{(2)} - \theta^*)^2] = \mathrm{Var}(T_n^{(2)}) = \frac{\theta^2}{3n} \xrightarrow[n\to\infty]{} 0,$$

  avec

  $$\frac{\sqrt{n}(T_n - \theta^*)}{\sqrt{\theta^{*2}/3}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1) \ .$$

Nous dirons que :

la vitesse de convergence de $T_n^{(1)}$ vers θ^* à la vitesse $1/n$.

la vitesse de convergence de $T_n^{(2)}$ vers θ^* à la vitesse $1/\sqrt{n}$.

Méthode des moments

Definition

Soit $p \neq 0$ un entier. On appelle moment d'ordre p d'une variable aléatoire X, la quantité $\mathbb{E}(X^p)$.

Soient $(X_1, \cdots, X_n)$ un n-échantillon de la loi d'une v.a. X telle que $\mathbb{E}(|X|^p) < \infty$. D'après la loi des grands nombres, si $\mathbb{E}(|X|^p) < \infty$, il est naturel d'estimer $\mathbb{E}(X^p)$ par $n^{-1} \sum_{i=1}^n X_i^p$. L'estimateur ainsi construit est sans biais et consistant. D'après les propriétes 1, 3, 4, et 5, on a

Propriété

Alors $\overline{X_n^p} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} X_i^p$ *est un estimateur sans biais et consistant de* $E(X_1^p)$.

De plus

$$\sqrt{n} \left(\frac{\overline{X_n^p} - \mathbb{E}(X^p)}{\sqrt{Var(X^p)}} \right) \xrightarrow[n \to \infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Méthode des moments : contexte général (1)

Soit $(X_1, \cdots, X_n)$ un n-échantillon de la loi d'une v.a. X telle que $\mathbb{E}(|X|^p) < \infty$. On suppose que la loi de X dépend d'un paramètre θ^* qui s'écrit comme une fonction continue des p premiers moments. Autrement dit, il existe une fonction continue de $\mathbb{R}^p$ dans $\mathbb{R}$, telle que

$$\theta^* = F(\mathbb{E}(X), \mathbb{E}(X^2), \cdots, \mathbb{E}(X^p)).$$

Un estimateur de θ^* obtenu par la méthode des moments est

$$\widehat{\theta} = F\left(\overline{X}, \frac{1}{n}\sum_{i=1}^n X_i^2, \cdots, \frac{1}{n}\sum_{i=1}^n X_i^p\right).$$

Les 2 propriétés qui précédent peuvent s'étendre au cas de $p \geq 1$. On en déduit alors que $\widehat{\theta}$ est un estimateur consistant de θ^*, $\sqrt{n}$-consistant et asymptotiquement gaussien de θ^*.

Méthode des moments : contexte général (2)

Remarque : la méthode des moments permet d'estimer n'importe quel moment non centré $E(X_1^k)$, ou toute fonction continue g de $E(X_1^k)$, par les équivalents empiriques dans l'échantillon. On peut également estimer les moments centrés de la loi de X, de la forme $E\left[(X_1 - E(X_1))^k\right]$, ou toute fonction continue des moments centrés du type $g\left(E\left[(X_1 - E(X_1))^k\right]\right)$, par le moment empirique correspondant $g\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^k\right)$; si $T_n^k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^k$ est un estimateur consistant de $E\left[(X_1 - E(X_1))^k\right]$, alors $g(T_n^k)$ sera un estimateur consistant de $g\left(E\left[(X_1 - E(X_1))^k\right]\right)$.

Méthode des moments : Exemple de la loi de géométrique de paramètre p^*

Soit $(X_1, \cdots, X_n)$ un n-échantillon de loi géométrique de paramètre p^*, avec $0 < p^* < 1$.On cherche un estimateur consistant de p^*. On sait que $\overline{X_n}$ est un estimateur consistant de $E(X_1)$ avec

$$\mathbb{E}(X_1) = \frac{1}{p^*},$$

donc $T_n = \dfrac{1}{\overline{X_n}} = g\left(\overline{X_n}\right)$ est un estimateur consistant de p^*. La quantité $T_n = \dfrac{1}{\overline{X_n}}$ est un estimateur biaisé de p^*, mais asymptotiquement sans biais.

Estimation par maximum de vraisemblance

Densité par rapport à la mesure dominante

Considérons $X_1, \cdots, X_n$ n variables aléatoires i.i.d. de même loi $P_{\theta^*} \in \{P_\theta, \theta \in \Theta\}$. On suppose pour tout θ, P_θ admet une densité par rapport à la mesure dominante $d\nu(.)$. On a donc

$$\mathbb{P}(X \in A) = \int_A L_{\theta^*}(x)\nu(dx).$$

De la même façon la densité du n-uplet $(X_1, \cdots, X_n)$ notée $L_{\theta^*}(x_1, \cdots, x_n)$ satisfait

$$\mathbb{P}((X_1, \cdots, X_n) \in \mathbb{A}) = \int_{\mathbb{A}} L_{\theta^*}(x_1, \cdots, x_n)\nu(dx_1)\cdots\nu(dx_n).$$

Vraisemblance (1)

Si les variables X_i, $i = 1, \cdots, n$ sont indépendantes et identiquement distribuées (i.i.d.), la loi du n-uplet $(X_1, \cdots, X_n)$ est $L_{\theta^*}(x_1, \cdots, x_n)$ satisfaisant la relation

$$L_{\theta^*}(x_1, \cdots, x_n) = \prod_{i=1}^{n} L_{\theta^*}(x_i). \quad (*)$$

Définition La vraisemblance de X_1, la variable aléatoire (fonction de θ et de la variable aléatoire X_1,

$$\theta \mapsto L_\theta(X_1). \quad (1)$$

La vraisemblance de $(X_1, \cdots, X_n)$, la variable aléatoire (fonction de θ et des variables aléatoires X_i, $i = 1, \cdots, n$)

$$\theta \mapsto L_n(\theta) = L_\theta(X_1, \cdots, X_n) = \prod_{i=1}^{n} L_\theta(X_i). \quad (2)$$

Vraisemblance pour un échantillon de loi discrète (1)

Soient $(X_1, \cdots, X_n)$ un n-échantillon de loi P_{θ^*} discrète. La mesure dominante est alors la mesure de comptage avec

$$\mathbb{P}(X \in A) = \sum_{x \in A} \mathbb{P}(X = x).$$

La densité par rapport à cette mesure est donc est caractérisée par $L_{\theta^*}(x) = \mathbb{P}(X = x)$. La vraisemblance de X_1 est

$$\theta \mapsto L_\theta(X_1) \text{ où } L_{\theta^*}(x_1) = P(X_1 = x_1).$$

Vraisemblance pour un échantillon de loi discrète (2)

La loi de $(X_1, \cdots, X_n)$ est alors donnée par

$$P(X_1 = x_1 \cap \cdots \cap X_n = x_n).$$

Puisque les variables aléatoires $X_1, \cdots, X_n$ sont indépendantes, nous avons

$$P(X_1 = x_1 \cap X_2 = x_2 \cap \cdots \cap X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i) = \prod_{i=1}^{n} L_{\theta^*}(x_i). \quad (3)$$

La vraisemblance de $(X_1, \cdots, X_n)$ est la v.a.

$$\theta \mapsto L_n(\theta) = \prod_{i=1}^{n} L_\theta(X_i) \quad \text{où } L_{\theta^*}(x) = P(X = x).$$

Vraisemblance pour un échantillon de loi continue (1)

Soient $X_1, \cdots, X_n$ un n-échantillon de loi absolument continue (par rapport à la mesure de Lesbesgue). Les variables admettent donc une densité $L_{\theta^*}(x_i)$ avec

$$\mathbb{P}(X \in A) = \int_A L_{\theta^*}(x)dx.$$

La variable aléatoire X a pour densité $L_{\theta^*}(x)$ pour $x \in \mathcal{X}$. La vraisemblance de X_1 est la variable aléatoire

$$\theta \mapsto L_\theta(X_1).$$

Vraisemblance pour un échantillon de loi continue (2)

La loi de $(X_1, \cdots, X_n)$ est déterminée par sa densité, qui d'après l'indépendance des v.a. $X_1, \cdots, X_n$, est donnée par

$$f_{\theta^*}(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} L_{\theta^*}(x_i).$$

La vraisemblance de $(X_1, \cdots, X_n)$ est la variable aléatoire

$$\theta \mapsto L_n(\theta) = \prod_{i=1}^{n} L_\theta(X_i) \quad \text{où } L_{\theta^*} \text{ est la densité de } X.$$

Exemple : loi de Bernoulli

Soit $(X_1, \cdots, X_n)$ un n-échantillon de la loi $\mathcal{B}(\theta^*)$, avec $0 < \theta^* < 1$. $\Theta = [0, 1]$. L'ensemble des valeurs possibles x_i de X_i est $\mathcal{X} = \{0, 1\}$ (indépendant de θ). On a $P(X_i = 1) = \theta^*$ et $P(X_i = 0) = 1 - \theta^*$. Et donc

$$P(X_i = x_i) = (\theta^*)^{x_i}(1 - \theta^*)^{1-x_i} = L_{\theta^*}(x_i) \text{ avec } x_i = 0 \text{ ou } 1$$

En utilisant l'indépendance des variables X_i, $i = 1 \cdots, n$, on a

$$\begin{aligned} P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) &= P(X_1 = x_1)P(X_2 = x_2) \dots P(X_n = x_n) \\ &= (\theta^*)^{\sum_{i=1}^n x_i}(1 - \theta^*)^{n - \sum_{i=1}^n x_i} \end{aligned}$$

Ainsi, la vraisemblance d'un n-échantillon de loi $\mathcal{B}(\theta^*)$ est

$$\theta \mapsto L_n(\theta) = \theta^{\sum_{i=1}^n x_i}(1 - \theta)^{n - \sum_{i=1}^n x_i}.$$

Exemple : n-échantillon de loi $\mathcal{N}(m^*, \sigma^{*2})$.

Soit $\mathbf{X} = (X_1, \cdots, X_n)$ un n-échantillon de la loi $\mathcal{N}(m^*, \sigma^{*2})$ alors la densité de $(X_1, \cdots, X_n)$ est

$$x \in \mathbb{R}^n \mapsto \frac{1}{(\sqrt{2\pi\sigma^{*2}})^n} \exp\left[\frac{-1}{2\sigma^{*2}}(x - m^*)^T(x - m^*)\right].$$

La vraisemblance est alors la fonction

$$(m, \sigma^2) \mapsto \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\left[\frac{-1}{2\sigma^2}(\mathbf{X} - m)^T(\mathbf{X} - m)\right].$$

Exemple : vecteur gaussien

Si $\mathbf{X}$ est un vecteur gaussien de $\mathbb{R}^n$, $\mathbf{X} \sim \mathcal{N}_n(m^*, \Gamma^*)$, où Γ^* est une matrice $n \times n$ de plein rang, alors $\mathbf{X}$ admet pour densité

$$x \in \mathbb{R}^n \mapsto \frac{1}{(\sqrt{2\pi})^n (\det(\Gamma^*))^{1/2}} \exp\left[\frac{-1}{2}(x - m^*)^T (\Gamma^*)^{-1}(x - m^*)\right].$$

La vraisemblance de $\mathbf{X}$ est alors

$$(m, \Gamma) \mapsto \frac{1}{(\sqrt{2\pi})^n \det(\Gamma)^{1/2}} \exp\left[\frac{-1}{2}(\mathbf{X} - m)^T (\Gamma)^{-1}(\mathbf{X} - m)\right].$$

Log-vraisemblance et notations (1)

Dans ce qui suit on appelle $log - vraisemblance$ de X_1, quand elle existe, la fonction

$$\theta \mapsto \ell_\theta(X) = \log(L_\theta(X_1)). \tag{4}$$

De la même façon, on appelle $log - vraisemblance$ de $(X_1, \cdots, X_n)$, quand elle existe, la fonction

$$\theta \mapsto \ell_\theta(X_1, \cdots, X_n) = \log(L_\theta(X_1, \cdots, X_n)). \tag{5}$$

Si les variables X_i sont indépendantes et identiquement distribuées, on a

$$\ell_\theta(X_1, \cdots, X_n) = \log(L_\theta(X_1, \cdots, X_n)) = \sum_{i=1}^{n} \ell_\theta(X_i) = \sum_{i=1}^{n} \log(L_\theta(X_i)).$$

Log-vraisemblance et notations (2)

Pour tout $\theta \in \Theta$, on note $\dot{\ell}_\theta \in \mathbb{R}^d$ le vecteur des dérivées premières (quand elles existent), défini pour $1 \leq i \leq d$ par

$$(\dot{\ell}_\theta)_i = \frac{\partial}{\partial \theta_i} \ell_\theta = \frac{\partial}{\partial \theta_i} \log(L_\theta).$$

De même on note $\ddot{\ell}_\theta$ la matrice des dérivées secondes définie pour $1 \leq i, j \leq d$,

$$(\ddot{\ell}_\theta)_{i,j} = \frac{\partial^2}{\partial \theta_i \partial \theta_j} \ell_\theta = \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log(L_\theta).$$

Log-vraisemblance et propriétés (1)

Sous de bonnes hypothèses (régularité+justification interversion $\int$ et dérivation), et en utilisant que

$$\int L_\theta(x)\nu(dx) = 1 \ \forall \theta \in \Theta,$$

on a

$$0 = \int \dot{L}_{\theta^*}(x)\nu(dx) = \int \frac{\dot{L}_{\theta^*}(x)}{L_{\theta^*}(x)} L_{\theta^*}(x)\nu(dx) = \mathbb{E}_{\theta^*}(\dot{\ell}_{\theta^*}(X)),$$

où $\dot{L}_\theta(x)$ désigne la dérivée première par rapport à θ et où $\mathbb{E}_{\theta^*}(\Psi(X)) = \int \Psi(x) L_{\theta^*}(x)\nu(dx)$.

Log-vraisemblance et propriétés (2)

Propriété

Pour tout θ,

$$\int L_\theta(x)\nu(dx) = 1.$$

Propriété

Sous de bonnes conditions de régularité,

$$\mathbb{E}_{\theta^*}(\dot{\ell}_{\theta^*}(X)) = 0.$$

De façon analogue à la propriété (9), on a

$$0 = \int \ddot{L}_{\theta^*}(x)\nu(dx) = \int \frac{\ddot{L}_{\theta^*}(x)}{L_{\theta^*}(x)} L_{\theta^*}(x)\nu(dx).$$

Log-vraisemblance et propriétés (3)

Et l'on en déduit la propriété suivante :

Propriété

$$\mathbb{E}_{\theta^*}\left(\ddot{\ell}_{\theta^*}(X)\right) = -\mathbb{E}_{\theta^*}\left[\left(\dot{\ell}_{\theta^*}(X)\right)^T\left(\dot{\ell}_{\theta^*}(X)\right)\right].$$

Log-vraisemblance et propriétés (4)

Enfin on peut montrer que

Propriété

Supposons que si $\theta \neq \theta^*$ alors $P_\theta \neq P_{\theta^*}$. On dit dans ce cas que le modèle est identifiable. Sous cette hypothèse d'identifiabilité, la fonction

$$\theta \mapsto \mathbb{E}_{\theta^*}\left(\log \frac{L_\theta}{L_{\theta*}}(X)\right) = \mathbb{E}_{\theta^*}\left(\ell_\theta(X)\right) - \mathbb{E}_{\theta^*}\left(\ell_{\theta^*}(X)\right)$$

est maximum en $\theta = \theta^*$.

Log-vraisemblance et propriétés (5)

Propriété

Notons $\mathbb{K}$ la distance (divergence) de Kulback entre les deux lois de densité L_θ et L_{θ^*} par rapport à la mesure dominante ν définie par

$$\mathbb{K}(L_\theta^*, L_\theta) = \mathbb{E}_{\theta^*}\left(\log \frac{L_\theta^*(X)}{L_\theta(X)}\right) = \int \log \frac{L_\theta^*(x)}{L_\theta(x)} L_{\theta^*}(x) d\nu(x).$$

Alors $\mathbb{K}(L_\theta^*, L_\theta) \geq 0$ et sous l'hypothèse d'identifiabilité, la fonction

$$\theta \mapsto \mathbb{K}(L_\theta^*, L_\theta)$$

atteint son minimum en un unique point θ^*.

Propriété

$$\mathbb{E}_{\theta^*}(\dot{\ell}_\theta(X)) > 0 \text{ si } \theta < \theta^* \text{ et } \mathbb{E}_{\theta^*}(\dot{\ell}_\theta(X)) < 0 \text{ si } \theta > \theta^*.$$

Hypothèses de modèle régulier

$\forall \theta \in \Theta, P_\theta$ est absolument continue par rapport à la mesure ν. $(\mathbf{H_1})$

Le paramètre θ^* appartient à l'intérieur de Θ, compact de $\mathbb{R}^d$. $(\mathbf{H_2})$

La fonction $\theta \mapsto \log(L_\theta(X))$ est 2 fois continuement différentiable. $(\mathbf{H_3})$

$\forall 1 \leq i, j \leq d$ et $\forall \theta, \exists$ un voisinage $\nu_{\theta,i,j}$ et $h_{\theta,i,j} \in \mathbb{L}_1(P_\theta)$

$$\text{tq} \sup_{\theta \in \nu_{\theta,i,j}} |(\ddot{\ell}_\theta(X))_{i,j}| \leq h_{\theta,i,j}. \quad (\mathbf{H_4})$$

$$\forall \theta, \exists \delta(\theta^*) \text{ tel que } \sup_{\theta' \, tq \|\theta' - \theta^*\| \leq \delta(\theta^*)} L_{\theta'}(x) \in \mathbb{L}_1(\nu). \quad (\mathbf{H_5})$$

Si $\theta \neq \theta'$ alors $L_\theta \neq L_{\theta'} - P_{\theta^*}$ presque surement . $(\mathbf{H_6})$

$\exists! \theta^*$ telle que $\mathbb{E}_{\theta^*}(\ell_\theta(X))$ est maximum en θ^*. $(\mathbf{H_7})$

Les hypothèses $(\mathbf{H_1})$-$(\mathbf{H_5})$ assurent que le modèle est régulier et qu'on intervertir dérivation et intégrale. Les hypothèses $(\mathbf{H_6})$-$(\mathbf{H_7})$ assurent que le modèle est identifiable.

Justification théorique

Si on note

$$\Psi(\theta) = \mathbb{E}_{\theta^*}(\ell_\theta(X)) \text{ et } \widehat{\Psi}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell_\theta(X_i),$$

d'après la loi des Grands Nombres, pour tout $\theta \in \Theta$ on a

$$\widehat{\Psi}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell_\theta(X_i) \xrightarrow[n\to\infty]{\mathbb{P}} \Psi(\theta) = \mathbb{E}_{\theta^*}(\ell_\theta(X)). \tag{6}$$

Sous de bonnes hypothèses

$$\widehat{\theta} = \underset{\theta\in\Theta}{\text{argmax}}\, \widehat{\Psi}(\theta) \xrightarrow[n\to\infty]{\mathbb{P}} \underset{\theta\in\Theta}{\text{argmax}}\, \Psi(\theta) = \theta^*.$$

On propose donc d'estimer

$$\theta^* = \underset{\theta\in\Theta}{\text{argmax}}\, \Psi(\theta) \text{ par } \widehat{\theta} = \underset{\theta\in\Theta}{\text{argmax}}\, \widehat{\Psi}(\theta).$$

Estimateur du maximum de vraisemblance (1)

Ces différentes propriétés justifient le choix de la vraisemblance (ou de son log quand il existe) comme critère d'estimation. En effet $\mathbb{E}_{\theta^*}(\log(L_\theta))$ est naturellement estimée de façon empirique par

$$\frac{1}{n}\sum_{i=1}^{n}\ell_\theta(X_i) = \frac{1}{n}\log[L_\theta(X_1,\cdots,X_n)] \xrightarrow[n\to\infty]{\mathbb{P}} \mathbb{E}_{\theta^*}(\log(L_\theta))$$

En remarquant que

$$\theta^* = \arg\max \mathbb{E}_{\theta^*}(\log(L_\theta)),$$

On propose donc d'estimer θ^* par par l'estimateur du maximum de vraisemblance de θ^*, noté $\widehat{\theta}$:

$$\begin{aligned}\widehat{\theta} &= \arg\max_\theta \frac{1}{n}\sum_{i=1}^{n}\log[L_\theta(X_i)] = \arg\max_\theta \frac{1}{n}\ell_\theta(X_1,\cdots,X_n)\\ &= \arg\max_\theta L_\theta(X_1,\cdots,X_n).\end{aligned}$$

Estimateur du maximum de vraisemblance (4)

Definition

On dit que $\widehat{\theta}_n$ est un estimateur du maximum de vraisemblance de θ^* (EMV) si $\widehat{\theta}_n$ maximise (en θ) la vraisemblance $L_n(\theta)$ (définie par (2)), c'est-à-dire tel que

$$L_{\widehat{\theta}}(X_1, \cdots, X_n) \geq L_\theta(X_1, \cdots, X_n) \quad \text{pour tout } \theta \in \Theta. \tag{7}$$

Rq : la justification théorique a été faite à partir de la log-vraisemblance mais l'EMV existe même si la log-vraisemblance n'existe pas !

Estimateur du maximum de vraisemblance (5)

Heuristique : La méthode du maximum de vraisemblance consiste à à estimer θ^* par la valeur de θ qui rend l'échantillon observé le plus vraisemblable. On va donc chercher la valeur de θ qui donne à l'échantillon qui a été observé la plus grande probabilité possible si X est une variable aléatoire discrète, ou la plus grande densité si X est une variable aléatoire continue, c'est-à dire la plus grande vraisemblance, et à choisir cette valeur pour estimer θ^*, d'où son nom.

Remarque : L'EMV n'existe pas toujours, il n'est pas toujours unique et n'a pas toujours une expression analytique. On est alors amené à utiliser des méthodes de maximisation numériques.

Propriétés de l'EMV

Théorème

Sous les hypothèses $(\mathbf{H_1})$-$(\mathbf{H_7})$*, l'estimateur du maximum de vraisemblance* $\widehat{\theta}$ *défini par (7) est consistant, asymptotiquement sans biais c'est-à-dire*

$$\widehat{\theta} \xrightarrow[n\to\infty]{\mathbb{P}} \theta^*, \quad \sqrt{n}(\widehat{\theta} - \theta^*) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, I^{-1}(\theta^*)),$$

où $I(\theta^*)$ *est une matrice* $d \times d$*, appelée matrice d'information de Fisher pour* X_1*, et définie par*

$$I(\theta) = -\mathbb{E}_{\theta^*}(\ddot{\ell}_\theta(X)) = \mathbb{E}_{\theta^*}\left[(\dot{\ell}_\theta(X))^T(\dot{\ell}_\theta(X))\right]. \tag{8}$$

On dit alors que l'estimateur du maximum de vraisemblance est consistant, $\sqrt{n}$-consistant, asymptotiquement gaussien et asymptotiquement efficace.

Remarques sur les hypothèses $(\mathbf{H_1})$-$(\mathbf{H_7})$

Remarque

Quand $X_1, \cdots, X_n$ sont i.i.d. de loi $\mathcal{U}_{[0,\theta^*]}$ les hypothèses ne sont pas vérifiées. Ceci vient pour une grande partie du fait que l'application $\theta \mapsto L_\theta(x) = (1/\theta)\mathbb{1}_{x\in[0,\theta]}$ n'est pas continue. Le modèle n'est donc pas régulier.

Remarque sur l'information de Fisher

Remarque

Quand les variables aléatoires $X_1, \cdots, X_n$ sont i.i.d. alors l'information de Fisher pour $(X_1, \cdots, X_n)$ qui est définie par

$$\begin{aligned} I_n(\theta) &= -\mathbb{E}_{\theta^*}(\ddot{\ell}_\theta(X_1, \cdots, X_n)) = \mathbb{E}_{\theta^*}\left[\dot{\ell}_\theta(X, \cdots, X_n))^T \dot{\ell}_\theta(X, \cdots, X_n))\right] \\ &= = nI(\theta^*), \end{aligned} \tag{9}$$

avec

$$I(\theta) = -\mathbb{E}_{\theta^*}(\ddot{\ell}_\theta(X_1)) = \mathbb{E}_{\theta^*}\left[(\dot{\ell}_\theta(X))^T(\dot{\ell}_\theta(X))\right].$$

Calcul de l'EMV (1)

1. Déterminer la loi de $(X_1, \cdots X_n)$ (en utilisant éventuellement l'indépendance des X_i).
2. Calculer l'expression de la vraisemblance $L_\theta(X_1, \cdots, X_n)$.
3. On cherche ensuite la valeur de θ, notée $\widehat{\theta}$ telle que

$$L_{\widehat{\theta}}(X_1, \cdots, X_n) \geq L_\theta(X_1, \cdots, X_n), \text{ pour tout } \theta \in \Theta.$$

4. Remplacer $(X_1, \cdots, X_n)$ par sa réalisation $(x_1, \cdots, x_n)$ dans $\widehat{\theta}_n$ pour obtenir l'estimation par maximum de vraisemblance de θ.

Calcul de l'EMV (2)

Deux cas peuvent se produire.

a) La fonction $\theta \mapsto L_\theta$ ne s'annule pas sur un domaine dépendant de θ. On calcule alors la log-vraisemblance, sur le domaine où la vraisemblance ne s'annule pas, $\ell_n(\theta) = \log\left(L_\theta(X_1, \cdots, X_n)\right)$. En effet, les 2 fonctions ont les mêmes maxima car la fonction $x \to \log(x) = \ln(x)$ est strictement croissante. Si $\theta \mapsto \ell_n(\theta)$ est de classe $\mathcal{C}^2$, alors $\widehat{\theta}$ qui maximise la log-vraisemblance satisfait

$$\frac{\partial}{\partial\theta}\ell_n(\widehat{\theta}) = 0, \text{ et la matrice } \frac{\partial^2}{\partial\theta^2}\ell_n(\widehat{\theta}) \text{ est définie négative.}$$

L'EMV est alors $\widehat{\theta}$ solution des équations précédentes.

b) La fonction $\theta \mapsto L_\theta(X_1, \cdots, X_n)$ s'annule sur un domaine dépendant de θ et/ ou n'est pas de classe $\mathcal{C}^2$. On trace alors le graphe de la fonction $\theta \mapsto L_\theta(X_1, \cdots, X_n)$ et déterminer graphiquement $\widehat{\theta}$.

Optimisation numérique (1)

Sous les hypothèses ($\mathbf{H_1}$)-($\mathbf{H_7}$) l'estimateur du maximum de vraisemblance (7) est unique et il satisfait par définition les équations de vraisemblance

$$\nabla \ell_n(\theta)|_{\theta=\widehat{\theta}} = \left.\frac{\partial \ell_n(\theta)}{\partial \theta}\right|_{\theta=\widehat{\theta}} = \vec{0}. \tag{10}$$

Il faut ensuite vérifier que la matrice des dérivées secondes est définie négative c'est-à-dire $H(\theta)$ la matrice des dérivées secondes (Hessienne de dimension $p \times p$) est définie négative avec

$$H(\theta) = \frac{\partial^2 \ell_n(\theta)}{\partial \theta^2}, \text{ et } (H(\theta))_{i,j} = \frac{\partial^2 \ell_n(\theta)}{\partial \theta_i \partial \theta_j}.$$

Optimisation numérique (2)

Definition

On appelle fonction de score le vecteur gradient des dérivées premières de la log-vraisemblance définie par

$$\nabla \ell_n(\theta) = \frac{\partial \ell_n(\theta)}{\partial \theta} \tag{11}$$

Avec cette notation les equations de vraisemblance (10) s'écrivent

$$\nabla \ell_n(\widehat{\theta}) = \vec{0}.$$

Optimisation numérique (3)

Dans les modèles simples, la résolution de (10) fournit des estimateurs explicites. Mais dans la plupart des modèles, la résolution de (10) ne fournit pas d'expressions analytiques de $\widehat{\theta}$. On a alors recours à des méthodes numériques itératives de maximisation de la log-vraisemblance ou résoudre les équations de vraisemblance. Une méthode classique est basée sur l'algorithme de Newton-Raphson, fondé sur l'approximation du vecteur de score par une fonction linéaire de θ dans un voisinage de θ, jusqu'à la réalisation d'un critère d'arrêt défini par

$$|\theta^{(m+1)} - \theta^{(m)}| \leq \text{ seuil } .$$

Optimisation numérique (4)

L'idée de départ repose sur les équations de vraisemblance

$$\nabla \ell_n(\widehat{\theta}) = \vec{0}.$$

Pour tout m, à l'étape $m+1$, on va choisir $\theta^{(m+1)}$ tel que

$$\nabla \ell_n(\theta^{(m+1)}) = 0.$$

En faisant un développement limité de $\nabla \ell_n(\theta^{(m+1)})$, on va approcher $\nabla \ell_n(\theta^{(m+1)})$ par

$$\nabla \ell_n(\theta^{(m)}) + (\theta^{(m+1)} - \theta^{(m)}) H(\theta^{(m)}).$$

Optimisation numérique (5)

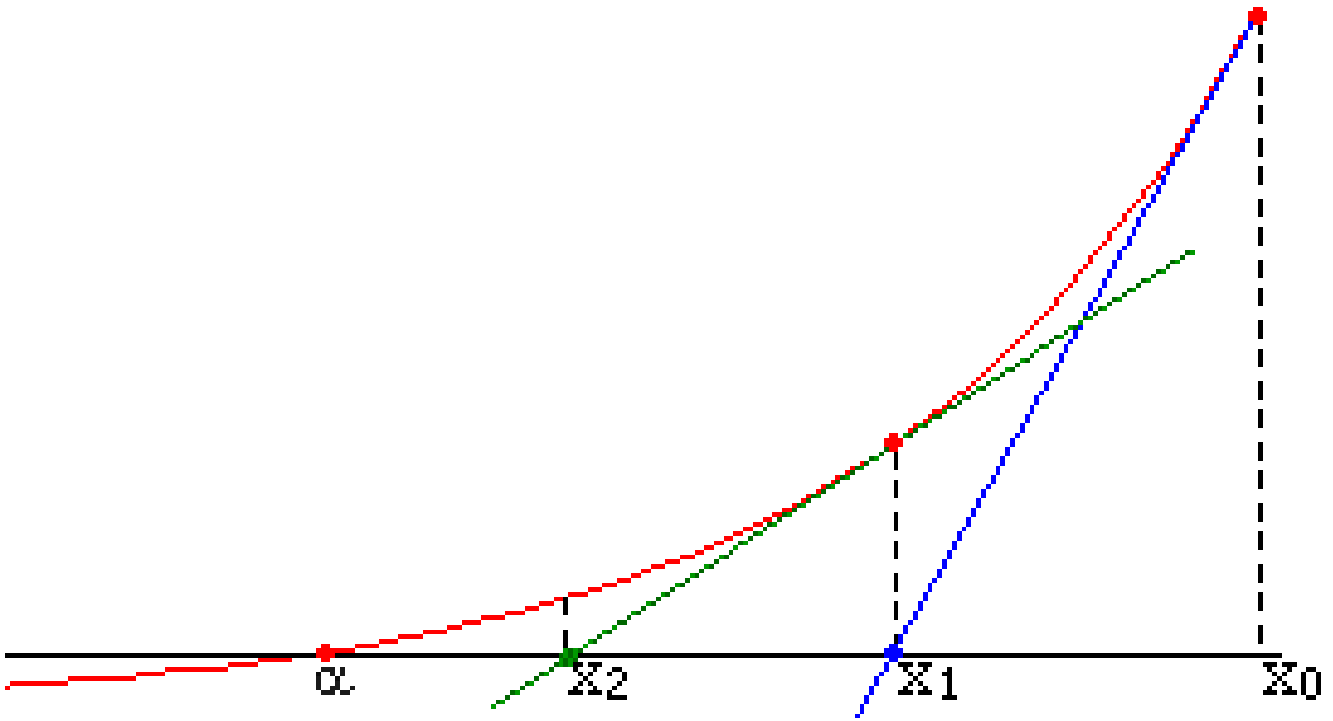
On va donc choisir $\theta^{(m+1)}$ telle que cette approximation soit égale à 0. On choisit alors $\theta^{(m+1)}$ tel que

$$\nabla \ell_n(\theta^{(m)}) + H(\theta^{(m)})(\theta^{(m+1)} - \theta^{(m)}) = 0.$$

Dans ces conditions $\theta^{(m+1)}$ est l'intersection de la tangeante à la vraisemblance en $\theta^{(m)}$ avec l'axe des abscisses. On a donc

$$\theta^{(m+1)} = \theta^{(m)} - (H(\theta^{(m)}))^{-1} \nabla \ell_n(\theta^{(m)}).$$

Optimisation numérique (6)



Optimisation numérique (7)

Plusieurs points sur lesquels il faut être vigilant :

- Le choix de la valeur initiale :cet aspect est crucial pour assurer la convergence de l'algorithme itératif vers un maximum global et non un maximum local.
- La justification de la convergence de l'algorithme
- Il faut vérifier la convergence de l'algorithme
- Il faut que $\theta \mapsto \ell_n(\theta)$ soit concave, strictement concave, ce qui est assuré si le modèle est identifiable (voir hypothèses $\mathbf{H_6}$ et $\mathbf{H_7}$).
- L'unicité de l'EMV $\widehat{\theta}$ comme maximum global.

Exhausitivité

Exhausitivité (1)

Soit un modèle statistique paramétrique $(\mathcal{X}; \mathcal{A}; P_{\theta^*})$, où

$$P_{\theta^*} \in \left\{ P_\theta, \theta \in \Theta \subset \mathbb{R}^d \right\}.$$

Soit n observations de variables aléatoires $X_1, \cdots, X_n$ de loi P_{θ^*} avec θ^* inconnu. On cherche à obtenir le maximum d'information sur θ^* à partir de $(X_1, \cdots, X_n)$. Cela passe par la construction d'un estimateur de θ^*, (une statistique), autrement dit une fonction mesurable des $X_1, \cdots, X_n$. Cette statistique est en fait un résumé des observations qui si elle est bien choisie, contient toute l'information possible sur θ^* que l'on peut obtenir à partir des observations. A priori $T(X_1, \cdots, X_n)$ des observations contient moins d'information sur θ^* que $(X_1, \cdots, X_n)$. On cherche des statistiques qui résument les observations tout en conservant l'intégralité de l'information sur θ^*, ce sont les statistiques exhaustives.

Exhausitivité (2)

Definition

Une statistique T est exhaustive pour si et seulement si la loi de probabilité conditionnelle de X sachant $[T = t]$ ne dépend pas de θ^*.

Cette définition exprime que si cette loi conditionne de X sachant $T = t$ ne dépend pas de θ^*, ce la signifie que la connaissance de l'intégralité de l'observation x n'apporte pas d'information suplémentaire sur θ^*. Par conséquent il faut s'attendre à ne se servir que de $t(x)$(au lieu de x tout entier) pour estimer θ^*.

Exhausitivité (3) : exemple

Soit un échantillon $X_1, \cdots, X_n$ i.i.d. de loi de $\mathcal{B}(\theta^*)$ la varisemblance de X_1 est

$$\theta \mapsto L_\theta(X_1) = \theta^{X_1}(1-\theta)^{(1-X_1)},$$

la vraisemblance de $(X_1, \cdots, X_n)$ est

$$\theta \mapsto L_\theta(X_1, \cdots, X_n) = \theta^{\sum_{i=1}^n X_i}(1-\theta)^{(n-\sum_{i=1}^n X_i)}.$$

Cette vraisemblance ne dépend que que $T = \sum_{i=1}^n X_i \sim \mathcal{B}(n, \theta^*)$. De plus

$$P(X_1 = x_1, \cdots, X_n = x_n | T = t) = \begin{cases} 0 \text{ si } \sum_{i=1}^n X_i \neq t \\ \frac{1}{C_n^T} \text{ si } \sum_{i=1}^n X_i = t. \end{cases} \quad (12)$$

Ainsi $P(X_1 = x_1, \cdots, X_n = x_n | T = t)$ ne dépend pas de $\theta^* t$ donc $T = \sum_{i=1}^n X_i$ est une statistique exhaustive pour θ^*.

Théorème de factorisation de Fisher-Neyman

La vérification de cette propriété n'est pas toujours facile. Il est parfois plus pratique d'utiliser une autre caractérisation de l'exhaustivité décrite dans le théoreme de factorisation de Fisher-Neyman.

Théorème

Théorème de factorisation de Fisher-Neyman *Pour qu'une statistique T soit exhaustive pour θ, il faut et il suffit qu'il existe deux fonctions mesurables g et h telles que*

$$\forall x \in \mathcal{X}, \forall \theta \in \Theta, \quad L_\theta(x) = g(t(x); \theta) h(x).$$

Théorème

Si T est exhaustive et si $T = \varphi(S)$ alors S est exhaustive.

Exhaustivité et EMV

Théorème

Si T est exhaustive et $\widehat{\theta}$ est l'estimateur du maximum de vraisemblance de θ^*, alors il existe une fonction φ telle que $\widehat{\theta} = \varphi(T)$.

Complétude

Definition

complétude Une statistique T est complète ou totale si et seulement si pour toute fonction mesurable φ on a

$\mathbb{E}_\theta \varphi(T)) = 0, \forall \theta \in \Theta \Longrightarrow \varphi = 0$ presque partout sur le support de la loi de T

Le fait de dire "$\varphi = 0$ presque partout sur le support de la loi de T" veut dire partout sauf sur un ensemble de mesure nulle.

Attention : T exhaustive n'implique pas que $\varphi(T)$ est exhaustive.

Admissibilité

Definition

• Soient T et T', deux estimateurs de $g(\theta)$.
L'estimateur T est dit meilleur que T' si

$$\mathbb{E}_\theta(T - g(\theta)) \leq \mathbb{E}_\theta(T' - g(\theta)) \; \forall \theta \in \Theta.$$

• L'estimateur T est dit strictement meilleur que T' si de plus il existe $\bar{\theta}$ tel que

$$\mathbb{E}_{\bar{\theta}}[T - g(\bar{\theta})] < \mathbb{E}_{\bar{\theta}}[T' - g(\bar{\theta})].$$

Definition

Un estimateur T de $g(\theta^*)$ est dit **admissible** s'il n'existe pas d'estimateur de $g(\theta^*)$ qui soit strictement meilleur que T.

Famille exponentielle

Famille exponentielle : exemples de lois discrètes

- Loi de Bernoulli (à densité par rapport à la mesure de comptage sur $\{0,1\}$), $\mathcal{B}(p)$ avec $p \in]0,1[$ de densité

$$p^y(1-p)^{1-y} = \exp\left[y \log\left(\frac{p}{1-p}\right) + \log(1-p)\right].$$

- Loi de Poisson (à densité par rapport à la mesure de comptage sur $\mathbb{N}$) $\mathcal{P}(\lambda)$ avec $\lambda > 0$ de densité

$$\exp(-\lambda)\frac{\lambda^y}{(y)!} = \exp[y \log(\lambda) - \lambda - \log((y)!)].$$

Famille exponentielle : exemples de lois continues

- Loi normale (à densité par rapport la mesure de Lebesgue sur $\mathbb{R}$), $\mathcal{N}(\mu, 1)$ avec $\mu \in \mathbb{R}$ de densité

$$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right)=\exp\left[y\mu-\frac{1}{2}\mu^2-\frac{1}{2}y^2-\log(\sqrt{2\pi})\right].$$

- Loi exponentielle (à densité par rapport la mesure de Lebesgue sur $\mathbb{R}_+$) $\mathcal{E}(\lambda)$ avec $\lambda > 0$ de densité

$$\lambda \exp(-\lambda y) = \exp(-\lambda y + \log(\lambda)).$$

- Loi log-normale (à densité par rapport à la mesure de Lebesgue sur $\mathbb{R}_+$), $Y \sim \log -\mathcal{N}(\mu, 1) \Longleftrightarrow \log Y \sim \mathcal{N}(\mu, 1)$ avec $\mu \in \mathbb{R}$ a pour densité

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(\log(y)-\mu)^2} = e^{\log(y)\mu-\frac{1}{2}\mu^2-\frac{1}{2}(\log(y))^2-\log(y)-\log(\sqrt{2\pi})}.$$

Famille exponentielle : exemples

On peut écrire toutes ses densités sous

$$\exp(\eta T(y) - A(\eta) - c(y)),$$

ou

$$\exp(Q(\theta) T(y) - A(\theta) - c(y)),$$

ou presque...

Préliminaire d'analyse

Soit $\mathcal{Y}$ un espace mesurable, ν une mesure positive sur $\mathcal{Y}$, T une fonction mesurable sur $\mathcal{Y}$ (suposée non ν-p.p. constante) et c une fonction mesurable. On définit

$$\mathcal{H} = \left\{ \eta \in \mathbb{R} \text{ tq } \int_{\mathcal{Y}} \exp(\eta T(y) - c(y)) d\nu(y) < \infty \right\},$$

et l'on suppose que $\overset{\circ}{\mathcal{H}} \neq \emptyset$ alors la fonction

$$\eta \mapsto \int_{\mathcal{Y}} \exp(\eta T(y) - c(y)) d\nu(y)$$

est infiniment différentiable sur $\overset{\circ}{\mathcal{H}}$ et de dérivée k-ième

$$\int_{\mathcal{Y}} T^k(y) \exp(\eta T(y) - c(y)) d\nu(y).$$

Famille exponentielle

Soit un modèle statistique paramétrique $(\mathcal{X}; \mathcal{A}; P_{\theta^*})$ où $P_{\theta^*} \in \left\{P_\theta, \theta \in \Theta \subset \mathbb{R}^d\right\}$, et on suppose que pour tout θ, P_θ admet une densité L_θ par rapport à une mesure dominante ν. On suppose que l'on dispose de n observations de variables aléatoires $X_1, \cdots, X_n$ i.i.d. de loi P_{θ^*}.

On considère ici des familles de densités ayant une structure particulière : les familles associées au modèle exponentiel (famille plus large que la loi exponentielle).

Famille exponentielle naturelle dans $\mathbb{R}$ (1)

Definition

On dit que la famille de lois P_η est une famille exponentielle si sa densité par rapport à ν s'écrit sous la forme

$$L_\eta(y) = \exp(\eta T(y) - A(\eta))h(y) = \exp(\eta T(y) - A(\eta) - c(y))$$

où $A(\eta)$ est une constante de normalisation définie par

$$A(\eta) = \log \int \exp(\eta(T(y))h(y)d\nu(y) = \log \int \exp(\eta(T(y) - c(y))d\nu(y).$$

La famille de densités g_η est définie pour η dans $\mathcal{H}$ où

$$\mathcal{H} = \left\{ \eta \in \mathbb{R} \text{ tel que } \int \exp(\eta T(y))h(y)d\nu(y) < \infty \right\}.$$

Famille exponentielle naturelle dans $\mathbb{R}$ (2)

Dans ce qui suit $\mathcal{H}$ est supposé d'intérieur non vide.

Proposition

- $\eta \mapsto L_\eta$ *est* > 0.
- *la fonction* $\eta \mapsto J(\eta) = \exp(A(\eta)) = \int \exp(\eta(T(y))h(y)d\nu(y)$ *est* $\mathcal{C}^\infty$ *sur* $\mathcal{H}$.
- *Pour tout k entier* $J^{(k)} = \int (T(y))^k \exp(\eta(T(y))h(y)d\nu(y)$.
- *la fonction* $\eta \mapsto A(\eta) = \log \int \exp(\eta(T(y))h(y)d\nu(y)$ *est* $\mathcal{C}^\infty$ *sur* $\mathcal{H}$.
- $A^{'}(\eta) = \mathbb{E}_\eta(T(X))$ *et* $A^{''}(\eta) = Var_\eta(T(X))$.
- T *non presque surement constante implique que le modèle est identifiable c'est-à-dire que* $\eta \neq \eta' \Longrightarrow L_\eta \neq L_{\eta'}$.
- *si* $(X_1, \cdots, X_n)$ *sont i.i.d. de densité* $L_\eta(y_1, \cdots, y_n)$ *alors la densité de* $(X_1, \cdots, X_n)$ *appartient aussi au modèle exponentiel avec*

$$L_\eta(y_1, \cdots, y_n) = \exp(\eta \sum_{i=1}^{n} T(y_i) - A(\eta)) \prod_{i=1}^{n} h(y_i).$$

Famille exponentielle naturelle dans $\mathbb{R}$ (3)

Proposition

Dans une famille exponentielle naturelle, pour tout $\eta \in \overset{\circ}{\mathcal{H}}$ *on a*

$$\mathbb{E}_\eta(T(Y)) = A'(\eta) \text{ et } Var_\eta(T(Y)) = A''(\eta).$$

Exemples : loi normale $\mathcal{N}(\mu, 1)$, loi log-normale $\log - \mathcal{N}(\mu, 1)$

Estimation par EMV et modèle exponentiel naturel

Considérons un n-échantillon $X_1, \cdots, X_n$ de densité L_{η^*} (avec T non ν presque surement constante) dans une famille exponentielle naturelle avec $\eta^* \in \overset{\circ}{\mathcal{H}}$ inconnu.
Alors la log-vraisemblance de $X_1, \cdots, X_n$ divisée par n vaut

$$\ell_n(\eta)/n = \eta \frac{1}{n} \sum_{i=1}^{n} T(Y_i) - A(\eta) - \frac{1}{n} \sum_{i=1}^{n} c(Y_i),$$

La fonction qui à $\eta \mapsto \ell_n(\eta)/n$ est strictement concave sur $\overset{\circ}{\mathcal{H}}$.
Donc s'il existe $\widehat{\eta} \in \overset{\circ}{\mathcal{H}}$ tel que

$$\ell_n'(\widehat{\eta})/n = 0 \Longleftrightarrow \frac{1}{n} \sum_{i=1}^{n} T(Y_i) = A'(\widehat{\eta}),$$

c'est l'unique estimateur du maximum de vraisemblance de η^*.

Propriétés de l'EMV dans modèle exponentiel naturel

Pour estimer $A'(\eta^*)$ on utilise $\widehat{A'(\eta^*)} = A'(\widehat{\eta}) = \frac{1}{n}\sum_{i=1}^n T(X_i)$.

- La loi des grands nombres assure que

$$\widehat{A'(\eta^*)} = A'(\widehat{\eta}) = \frac{1}{n}\sum_{i=1}^n T(X_i) \xrightarrow[n\to\infty]{\mathbb{P}} \mathbb{E}_{\eta^*}(T(X_1)) = A'(\eta^*).$$

- Le TLC implique

$$\frac{\sqrt{n}(\overline{T} - A'(\eta^*))}{\sqrt{A''(\eta^*)}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

- avec le lemme de Slutsky, on obtient

$$\frac{\sqrt{n}(\overline{T} - A'(\eta^*))}{\sqrt{A''(\widehat{\eta})}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Si on veut remonter à des résultats sur η^* il faut utiliser la continuité de $\eta \mapsto A(\eta)$ ainsi que la *delta − method* (cf outils du début du cours).

Famille exponentielle générale dans $\mathbb{R}$ (1)

Definition

On dit que la famille de lois P_θ est une famille exponentielle si sa densité par rapport à ν s'écrit sous la forme

$$L_\theta(y) = \exp(Q(\theta)T(y) - \alpha(\theta))h(y) = \exp(Q(\theta)T(y) - \alpha(\theta) - c(y)),$$

où $\alpha(\eta)$ est une constante de normalisation définie par

$$\alpha(\theta) = \log\int \exp(Q(\theta)(T(y))h(y)d\nu(x) = \log\int \exp(Q(\theta)(T(y) - c(y))d\nu(y).$$

La famille de densités L_θ est définie pour θ dans Θ où

$$\Theta = \left\{\theta \in \mathbb{R} \text{ tel que } \int \exp(Q(\theta)T(y))h(y)d\nu(y) < \infty\right\}.$$

Famille exponentielle générale dans $\mathbb{R}$ (2)

Dans ce qui suit on suppose que

- Θ est un intervalle non vide,
- $\theta \mapsto Q(\theta)$ est $\mathcal{C}^\infty$
- $Q(\theta) \neq 0$ presque partout,
- $\theta \mapsto Q(\theta)$ est bijective de Θ dans $Q(\Theta)$,
- Q^{-1} dérivable.
- T presque surement non constante.

Famille exponentielle générale dans $\mathbb{R}$ (3)

On a de même

Proposition

- *$\theta \mapsto L_\theta$ est > 0. On peut donc bien définir la log-vraisemblance ;*
- *la fonction $\theta \mapsto J(\theta) = \exp(\alpha(\theta)) = \int \exp(Q(\theta)(T(y))h(y)d\nu(y)$ est $\mathcal{C}^\infty$ sur Θ.*
- *la fonction $\theta \mapsto \alpha(\theta) = \log \int \exp(Q(\theta)(T(y))h(y)d\nu(y)$ est $\mathcal{C}^\infty$ sur Θ.*
- *T non presque surement constante implique que le modèle est identifiable c'est-à-dire que $\theta \neq \theta' \Longrightarrow L_\theta \neq l_{\theta'}$.*
- *si $(X_1, \cdots, X_n)$ sont i.i.d. de densité L_θ alors la densité de $(X_1, \cdots, X_n)$ appartient aussi au modèle exponentiel avec*

$$L_\theta(y_1, \cdots, y_n) = \exp(Q(\theta) \sum_{i=1}^{n} T(y_i) - \alpha(\theta)) \prod_{i=1}^{n} h(y_i).$$

- *On retrouve la famille exponentielle naturelle en posant $\eta = Q(\theta)$ et $A(\eta) = \alpha(Q^{-1}(\eta))$.*

Famille exponentielle générale dans $\mathbb{R}$ (4)

Exemples : loi exponentielle, Benoulli, Poisson...

Famille exponentielle naturelle dans $\mathbb{R}^d$ (1)

Definition

Soit $\eta \in \mathbb{R}^d$. On dit que la famille de lois P_η est une famille exponentielle si sa densité par rapport à ν s'écrit sous la forme

$$L_\eta(x) = \exp(\sum_{j=1}^{d} \eta_j T_j(x) - A(\eta))h(x) = \exp(\eta^T T(x) - A(\eta))h(x)$$

où $T(x) \in \mathbb{R}^d$, et $A(\eta)$ est une constante de normalisation définie par

$$A(\eta) = \log \int \exp(\eta^T T(x)h(x)d\nu(x).$$

La famille de densités L_η est définie pour η dans $\mathcal{H}$ où

$$\mathcal{H} = \left\{ \eta \in \mathbb{R} \text{ tel que } \int \exp(\eta^T T(x))h(x)d\nu(x) < \infty \right\}.$$

Famille exponentielle naturelle dans $\mathbb{R}^d$ (2)

Dans ce qui suit $\mathcal{H}$ est supposé d'intérieur non vide.

Proposition

- $\eta \mapsto L_\eta$ *est* > 0. *La log-vraisemblance est donc bien définie.*
- *la fonction* $\eta \mapsto J(\eta) = \exp(A(\eta)) = \int \exp(\eta^T(T(x))h(x)d\nu(x)$ *est* $\mathcal{C}^\infty$ *sur* $\mathcal{H}$.
- *Pour tout* $j_1, \cdots, j_k$,

$$\frac{\partial^k J(\eta)}{\partial \eta_{j_1} \cdots \partial \eta_{j_k}} = \int \prod_{j=j_1}^{j_k} T_j(x) \exp(\eta^T T(x))h(x)d\nu(x).$$

- *la fonction* $\eta \mapsto A(\eta) = \log \int \exp(\eta^T(T(x))h(x)d\nu(x)$ *est* $\mathcal{C}^\infty$ *sur* $\mathcal{H}$.
- $A'(\eta) = \mathbb{E}_\eta(T(x))$ *et* $A'(\eta) = Var_\eta(T(x))$.

Famille exponentielle naturelle dans $\mathbb{R}^d$ (3)

Proposition

- *si* $(X_1, \cdots, X_n)$ *sont i.i.d. de densité* L_η *alors la densité de* $(X_1, \cdots, X_n)$ *appartient aussi au modèle exponentiel avec*

$$L_\eta(x_1, \cdots, x_n) = \exp(\eta^T \sum_{i=1}^{n} T(x_i) - A(\eta)) \prod_{i=1}^{n} h(x_i).$$

- T *non presque surement constante implique que le modèle est identifiable c'est-à-dire que* $\eta \neq \eta' \Longrightarrow L_\eta \neq L_{\eta'}$.

Famille exponentielle générale dans $\mathbb{R}^d$ (1)

Definition

On dit que la famille de lois P_θ est une famille exponentielle si sa densité par rapport à ν s'écrit sous la forme

$$L_\theta(x) = \exp(Q(\theta)^T T(x) - \alpha(\theta))h(x),$$

où $\alpha(\theta)$ est une constante de normalisation définie par

$$\alpha(\theta) = \log \int \exp(Q(\theta)^T (T(x))h(x)d\nu(x).$$

La famille de densités L_θ est définie pour θ dans Θ où

$$\Theta = \left\{\theta \in \mathbb{R} \text{ tel que } \int \exp(Q(\theta)^T T(x))h(x)d\nu(x) < \infty\right\}.$$

Famille exponentielle générale dans $\mathbb{R}^d$ (2)

Dans ce qui suit on suppose que

- Θ est un intervalle non vide,
- $\theta \mapsto Q(\theta)$ est $\mathcal{C}^\infty$
- $Q(\theta) \neq 0$ presque partout,
- $\theta \mapsto Q(\theta)$ est bijective de Θ dans $Q(\Theta)$,
- Q^{-1} dérivable.
- $T(x)$ presque surement non constante.

On a de même

Proposition

- *$\theta \mapsto L_\theta$ est > 0.*
- *la fonction $\theta \mapsto J(\theta) = \exp(\alpha(\theta)) = \int \exp(\theta(T(x))h(x)d\nu(x)$ est $\mathcal{C}^\infty$ sur Θ.*
- *la fonction $\eta \mapsto \alpha(\theta) = \log \int \exp(\theta(T(x))h(x)d\nu(x)$ est $\mathcal{C}^\infty$ sur Θ.*

Famille exponentielle générale dans $\mathbb{R}^d$ (3)

Proposition

- *si* $(X_1, \cdots, X_n)$ *sont i.i.d. de densité* L_θ *alors la densité de* $(X_1, \cdots, X_n)$ *appartient aussi au modèle exponentiel avec*

$$L_\theta(x_1, \cdots, x_n) = \exp(Q(\theta) \sum_{i=1}^{n} T(x_i) - \alpha(\theta)) \prod_{i=1}^{n} h(x_i).$$

- T *non presque surement constante implique que le modèle est identifiable c'est-à-dire que* $\theta \neq \theta' \Longrightarrow L_\theta \neq L_{\theta'}$.

Lien entre famille exponentielle et exhaustivité

On suppose qu'on a un modèle statistique $(\mathcal{X}; \mathcal{A}; P_{\theta^*})$ où $P_{\theta^*} \in \left\{P_\theta, \theta \in \Theta \subset \mathbb{R}^d\right\}$. On dispose de n observations de variables aléatoires $X_1, \cdots, X_n$ de loi P_{θ^*}. On suppose que pour tout θ, P_θ admet une densité par rapport à une mesure dominante ν.

Théorème

Théorème de Darmois *Dans un modèle* $(\mathcal{X}; \mathcal{A}; P_{\theta^*})$ *où* $P_{\theta^*} \in \left\{P_\theta, \theta \in \Theta \subset \mathbb{R}^d\right\}$. *On suppose que pour tous* $\theta \in \Theta$ *le support de la loi* P_θ *ne dépend pas de* θ. *Alors, il existe une statistique exhaustive si et seulement si cette loi appartient à la famille exponentielle avec* $L_n(\theta) = \Pi_{i=1}^n L_\theta(X_i) = g(T(X_1, \cdots, X_n); \theta) h(X_1, \cdots, X_n)$ *où*

$$g_\theta(T(x); \theta) = \exp(Q(\theta)^T T(x) - A(\theta)),$$

et $T(X) = (\sum_{i=1}^n T_1(X_i), \cdots, \sum_{i=1}^n T_d(X_i))$ *est une statistique exhaustive.*

Estimation paramétrique optimale

Estimation paramétrique optimale : introduction (1)

Soit X une variable aléatoire de loi P_{θ^*} appartenant à une famille de loi $\{P_\theta, \theta \in \Theta\}$. On note $\mathcal{X}$ l'ensemble des valeurs possibles pour X. La loi P_{θ^*} dépendant d'un paramètre θ^* inconnu à valeurs dans un ensemble Θ. On cherche à estimer $g(\theta^*)$ à partir d'un n-échantillon $(X_1, \cdots, X_n)$ de X de loi P_{θ^*} où

$$P_{\theta^*} \in \{P_\theta, \theta \in \Theta\}.$$

On suppose que pour tout $\theta \in \Theta \in \mathbb{R}^d$ il existe une mesure dominante ν σ-finie, et une fonction $L_\theta(x)$ telle que pour tous $\theta \in \Theta$,

$$\mathbb{P}(X \in A) = \int_A L_{\theta^*}(x)\nu(dx).$$

Estimation paramétrique optimale : introduction (2)

Dans $\mathbb{R}$, la qualité d'un estimateur T_n de $g(\theta^*)$ est mesurée par l'erreur quadratique moyenne définie par

$$EQM(T_n) = \mathbb{E}_{\theta^*}[T_n - g(\theta^*)]^2.$$

On a la décomposition *biais* − *variance* suivante

$$EQM(T_n) = \mathrm{Var}_{\theta^*}(T_n) + (\mathbb{E}_{\theta^*}(T_n) - g(\theta^*))^2 . \quad (13)$$

Estimation paramétrique optimale : introduction (3)

Dans $\mathbb{R}^d$, c'est plus complexe. Si $g(\theta) \in \mathbb{R}^d$,

$$g(\theta) = \begin{pmatrix} g_1 \\ . \\ . \\ . \\ g_d \end{pmatrix}, \text{ et } T_n = \begin{pmatrix} T_{n,1} \\ . \\ . \\ . \\ T_{n,d} \end{pmatrix}.$$

On peut regarder

$$\mathbb{E}_{\theta^*} \parallel T_n - g(\theta^*) \parallel_{\ell_2}^2 = \sum_{i=1}^{d} (T_{n,i} - g_i)^2,$$

et on a la décomposition biais-variance

$$\sum_{i=1}^{d} \text{Var}(T_{n,i}) + (\mathbb{E}_{\theta^*}(T_{n,i}) - g_i)^2.$$

Estimation paramétrique optimale : introduction (4)

On va donc plutot regarder la matrice suivante

$$\mathbb{E}_{\theta^*}[(T_n - g(\theta^*))(T_n - g(\theta^*))^T] + (\mathbb{E}_{\theta^*}(T_n) - g(\theta^*))(\mathbb{E}_{\theta^*}(T_n) - g(\theta^*))^T .$$

On a la décomposition *biais* − *variance* suivante

$$\begin{aligned} EQM(T_n) &= \mathbb{E}_{\theta^*}[(T_n - g(\theta^*))^T (T_n - g(\theta^*))] + \\ &\quad (\mathbb{E}_{\theta^*}(T_n) - g(\theta^*))^T (\mathbb{E}_{\theta^*}(T_n) - g(\theta^*)) . \end{aligned} \tag{14}$$

Estimation paramétrique optimale : introduction (5)

On va chercher l'estimateur de $g(\theta^*)$ qui a l'erreur quadratique moyenne la plus petite possible. Si l'estimateur T_n est sans biais, $\mathbb{E}_{\theta^*}(T_n) = g(\theta^*)$, alors le meilleur estimateur est celui qui "minimise" la covariance

$$\mathbb{E}_{\theta^*}[(T_n - g(\theta^*))(T_n - g(\theta^*))^T] = \Lambda_{T_n}.$$

Pour $\theta \in \mathbb{R}$, un estimateur optimal sera un estimateur sans biais et de variance minimale **ESBVM**. Dans ce chapitre nous allons voir quelle procédure utiliser pour construire un estimateur ESBVM. Cette procédure est liée à la notion d'exhaustivité et de complétude.
Par ailleurs cette notion de qualité d'un estimateur peut-être évaluée par des propriétés non asymptotiques (pour tout n) ou par des propriétés asymptotiques (quand $n \to \infty$).

Réduction de la variance

Soit X de loi P_{θ^*} appartenant à une famille de loi $\{P_\theta, \theta \in \Theta\}$. La loi P_{θ^*} dépend d'un paramètre θ^* inconnu à valeurs dans un ensemble $\Theta \in \mathbb{R}^d$. On cherche à estimer θ^* à partir d'un n-échantillon $(X_1, \cdots, X_n)$ de X de loi P_{θ^*} où $P_{\theta^*} \in \{P_\theta, \theta \in \Theta\}$. On suppose que pour tout $\theta \in \Theta \in \mathbb{R}^d$ il existe une mesure dominante ν σ-finie, et une fonction $L_\theta(x)$ telle que $dP_\theta/d\nu = L_\theta$ et $\mathbb{P}(X \in A) = \int_A L_{\theta^*}(x)\nu(dx)$.
La théorème suivant permet, à partir d'un estimateur sans biais de θ^*, de construire un autre estimateur sans biais de θ^* de variance inférieure, dans le contexte où il existe une statistique exhaustive.

Théorème

Théorème de Rao-blackwell

Supposons que la densité L_θ est telle qu'il existe une statistique exhaustive pour θ. Supposons qu'il existe un estimateur sans biais de θ^*, noté $\widehat{\theta}$. Alors $\widetilde{\theta} = \mathbb{E}_{\theta^*}(\widehat{\theta}|T)$ est un estimateur sans biais de θ^*, de variance inférieure à celle de $\widehat{\theta}$ i.e.

$$Var(\widetilde{\theta}) \leq Var(\widehat{\theta}).$$

Estimation sans biais de variance minimale

Théorème

Théorème de Lehmann-Scheffé

Si $\widehat{\theta}$ est un estimateur sans biais de θ^* et T est une statistique exhaustive complète, alors $\widetilde{\theta} = \mathbb{E}_{\theta^*}(\widehat{\theta}|\, T)$ est l'unique estimateur sans biais de θ^*, de variance minimale parmi tous les estimateurs sans biais de θ^*.

Corollaire

Pour trouver un estimateur optimal, il suffit de trouver un estimateur sans biais, fonction d'une statistique exhaustive et complète.

Théorème

Le théorème de Lehmann-Scheffé reste valable si on remplace θ par $\varphi(\theta)$ où φ est une fonction mesurable quelconque. Autrement dit, l'ESBVM de $\varphi(\theta)$ est un estimateur sans biais de $\varphi(\theta)$, fonction d'une statistique exhaustive et complète.

Estimation paramétrique optimale

Information de Fisher : cas $d = 1$

Nous avons vu dans les propriétés du l'estimateur du maximum de vraisemblance que la variance limite est l'inverse de l'information de Fisher

$$I_n(\theta) = -\mathbb{E}_{\theta^*}(\ddot{\ell}_n(\theta^*)) = \mathbb{E}_{\theta^*}\left[(\dot{\ell}_n(\theta^*))^T(\dot{\ell}_n(\theta^*))\right].$$

Revenons ici sur cette quantité quand $d = 1$

$$I_n(\theta) = -\mathbb{E}_{\theta^*}(\ddot{\ell}_n(\theta^*)) = \mathbb{E}_{\theta^*}\left[(\dot{\ell}_n(\theta^*))^2\right]. \tag{15}$$

Information de Fisher : cas $d = 1$ (1)

Considérons dans un premier temps que $d = 1$ cad $\theta^* \in \mathbb{R}$. Partons de la première propriété de la log-vraisemblance à savoir

$$\mathbb{E}_{\theta^*}(\ell'_{\theta^*}(X)) = 0.$$

Soit $T_n = \Psi(X_1, \cdots, X_n)$ un estimateur de $g(\theta^*)$ et notons $\varphi(\theta^*) = \mathbb{E}_{\theta^*}(T_n)$. En utilisant que

$$\mathbb{E}_{\theta^*}(\ell'_{\theta^*}(X)) = 0,$$

on obtient que

$$\mathbb{E}_{\theta^*}(\ell'_n(\theta^*)) = 0 = \varphi(\theta)\mathbb{E}_{\theta^*}(\ell'_n(\theta^*)).$$

Information de Fisher : cas $d = 1$ (2)

Par conséquent

$$\begin{aligned}
\varphi'(\theta^*) = (\mathbb{E}_{\theta^*}(T_n))' &= \frac{\partial}{\partial\theta^*}\int \Psi(x_1,\cdots,x_n)L_n(\theta^*)\nu(dx_1)\cdots\nu(dx_n) \\
&= \int \Psi(x_1,\cdots,x_n)L_n'(\theta^*)\nu(dx_1)\cdots\nu(dx_n) \\
&= \int \Psi(x_1,\cdots,x_n)\frac{L_n'(\theta^*)}{L_n(\theta^*)}L_n(\theta^*)\nu(dx_1)\cdots\nu(dx_n) \\
&= \mathbb{E}_{\theta^*}(T_n\ell_n'(\theta^*)) = \mathbb{E}_{\theta^*}\left[(T_n-\varphi(\theta^*))\ell_n'(\theta^*)\right].
\end{aligned}$$

Information de Fisher : cas $d = 1$ (3)

On applique l'inégalité de Cauchy-Schwarz ($|\mathbb{E}(XY)|^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2)$) et on obtient

$$(\varphi'(\theta^*))^2 \leq \mathbb{E}_{\theta^*}\left[(T_n - \varphi(\theta^*))^2\right] \mathbb{E}_{\theta^*}\left[(\ell'_n(\theta^*))^2\right],$$

soit aussi

$$(\varphi'(\theta^*))^2 \leq \mathbb{E}_{\theta^*}\left[(T_n - \mathbb{E}_{\theta^*}(T_n))^2\right] \mathbb{E}_{\theta^*}\left[(\ell'_n(\theta^*))^2\right],$$

Information de Fisher : cas $d = 1$ (4)

On obtient donc

$$\mathrm{Var}_{\theta^*}(T_n) \geq \frac{(\varphi'(\theta^*))^2}{I_n(\theta^*)}$$

soit aussi

$$\mathrm{Var}_{\theta^*}(T_n) \geq \frac{\left(\frac{\partial}{\partial\theta^*}\mathbb{E}_{\theta^*}(T_n)\right)^2}{nI(\theta^*)}.$$

Cette inégalité s'écrit aussi

$$b^2(T_n) + \mathrm{Var}_{\theta^*}(T_n) \geq b^2(T_n) + \frac{(b'(T_n) + g'(\theta^*))^2}{nI(\theta^*)},$$

où $b(T_n) = \mathbb{E}_{\theta^*}(T_n) - g(\theta^*)$ est le biais de l'estimateur T_n.

Efficacité, efficacité asymptotique : cas $d = 1$

On dira que T_n est un estimateur efficace de $g(\theta^*)$ si

$$\mathrm{Var}_{\theta^*}(T_n) = \frac{\left(\frac{\partial}{\partial\theta^*}\mathbb{E}_{\theta^*}(T_n)\right)^2}{nI(\theta^*)}.$$

On dira que T_n est un estimateur asymptotiquement efficace de $g(\theta^*)$ si

$$\sqrt{n}(T_n - g(\theta^*)) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, I^{-1}(\theta^*)).$$

rq : Si T_n est un estimateur asymptotiquement efficace de T_n et si $\psi \in \mathcal{C}^1$ alors $\psi(T_n)$ est un estimateur asymptotiquement efficace de $\psi(\theta^*)$ avec

$$\sqrt{n}(\psi(T_n) - \psi(\theta^*)) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, (\psi'(\theta^*))^2 I^{-1}(\theta^*)).$$

Information de Fisher : cas général (1)

Notons $U(\theta, X)$ la fonction de score pour X_1 définie par

$$U(\theta, X) = \begin{pmatrix} U_1(\theta, X) \\ . \\ . \\ . \\ U_d(\theta, X) \end{pmatrix} = \nabla \ell_\theta(X), \text{ avec } U_j(\theta, X) = \frac{\partial \ell_\theta}{\partial \theta_j}(X).$$

On note $(D_\theta \ell_\theta)(X)$, alors $U(\theta, X) = (D_\theta \ell_\theta)(X)^T$, et

$$(D^2_\theta \ell_\theta)(x) = \begin{pmatrix} \frac{\partial^2}{\partial \theta_1^2} \ell_\theta(x) & \cdots & \frac{\partial^2}{\partial \theta_1 \partial \theta_d} \ell_\theta(x) \\ . & . & . \\ \frac{\partial^2}{\partial \theta_1 \partial \theta_d} \ell_\theta(x) & . & \frac{\partial^2}{\partial \theta_d^2} \ell_\theta(x) \end{pmatrix}$$

Information de Fisher : cas général (2)

On sait que :

- $\mathbb{E}_{\theta^*}(U(\theta^*, X_1)) = 0$,
- La matrice d'information de Fisher pour X_1 notée $I(\theta^*)$ vérifie

$$I_{j,k}(\theta^*) = \text{Cov}(U_j(\theta^*, X_1), U_k(\theta^*, X_1)) = -\mathbb{E}_{\theta^*}(\frac{\partial^2}{\partial\theta_J \partial\theta_k}\ell_{\theta^*}(X_1)),$$

- $I(\theta^*) = -\mathbb{E}_{\theta^*}[(D^2_\theta \ell_{\theta^*})(X)] = \mathbb{E}_{\theta^*}[(D_{\theta^*}\ell_{\theta^*})(D_{\theta^*}\ell_{\theta^*})^T]$
- La matrice d'information de Fisher pour $(X_1, \cdots, X_n)$ notée $I_n(\theta^*)$ vérifie $I_n(\theta^*) = nI(\theta^*)$.

Information de Fisher : cas général (3)

Soit T_n une application mesurable $\mathbb{R}^d \mapsto \mathbb{R}^q$ et soit Δ telle que

$$\Delta_{i,j} = \frac{\partial}{\partial \theta_j^*} \mathbb{E}_{\theta^*}(T_{n,i}), \text{ pour } 1 \leq i \leq q, \ 1 \leq j \leq d.$$

Soit Λ_{T_n} la matrice de variance-covariance de T_n. Si le modèle est régulier, alors, pour tous $\theta^* \in \mathbb{R}^d$, on a

$$\Lambda_{T_n} - \Delta I_n^{-1}(\theta^*) \Delta^T \text{ est une matrice semi-definie positive,}$$

soit aussi si $\text{Cov}(T_n) - [D_{\theta^*} \mathbb{E}_{\theta^*}(T_n)] I_n^{-1}(\theta^*) [D_{\theta^*} \mathbb{E}_{\theta^*}(T_n)]^T$ est une matrice semi-définie positive.
Rappel : on dit que M est semi-définie positive si $\forall x \neq 0, x^T M x \geq 0$.
Quand $d = q = 1$ on retrouve

$$\text{Var}_{\theta^*}(T_n) \geq \frac{(\frac{\partial}{\partial \theta^*} \mathbb{E}_{\theta^*}(T_n))^2}{I_n(\theta^*)}.$$

Information de Fisher : cas général (4)

Pour tous $i = 1, \cdot, q$ on a

$$\mathrm{Var}(T_{n,i}) \geq \sum_{j=1}^{d} \sum_{k=1}^{d} [(I_n(\theta^*))^{-1}]_{j,k} \frac{\partial \mathbb{E}_{\theta^*}(T_{n,i})}{\partial \theta_j^*} \frac{\partial \mathbb{E}_{\theta^*}(T_{n,i})}{\partial \theta_k^*}.$$

En particulier, si T_n est un estimateur sans biais de θ^* ; $\mathbb{E}_{\theta^*}(T_{n,i}) = \theta_i^*$, $\frac{\partial \mathbb{E}_{\theta^*}(T_{n,i})}{\partial \theta_j^*} = 1$ si $i = j$ et 0 sinon . Ceci implique que dans ce cas là,

$$\mathrm{Var}_{\theta^*}(T_{n,i}) \geq (I_n(\theta^*))_{i,i}^{-1}.$$

L'estimateur $T_{n,i}$ est un estimateur efficace de θ_i^* si

$$\mathrm{Var}_{\theta^*}(T_{n,i}) = (I_n(\theta^*))_{i,i}^{-1}.$$

Information de Fisher : cas général (5)

De façon plus générale, l'estimateur T_n est un estimateur efficace de $g(\theta^*)$ si

$$\mathrm{Cov}(T_n) = [D_{\theta^*}\mathbb{E}_{\theta^*}(T_n)]I_n(\theta^*)[D_{\theta^*}\mathbb{E}_{\theta^*}(T_n)]^T.$$

Un estimateur T_n sans biais est efficace si et seulement si $\Lambda_{T_n} = (I_n(\theta^*))^{-1}$. Dans ce cas on

$$\mathrm{Var}_{\theta^*}(T_{n,i}) = [(I_n(\theta^*)^{-1}]_{i,i}.$$

Information de Fisher : cas général (6)

Considérons deux cas :

• Dans le vecteur $\theta^* \in \mathbb{R}^d$, seule la coordonnée i est inconnue. Dans ce cas là, l'information de Fisher pour estimer θ_i^* est $1/I_{i,i}(\theta^*)$ avec

$$I_{i,i}(\theta^*) = \mathbb{E}_{\theta^*}\left[\left(\frac{\partial \ell_{\theta^*}(X_1)}{\partial \theta_i}\right)^2\right].$$

• Dans le cas où le vecteur θ^* complet est inconnu, l'information de Fisher pour estimer θ_i^* est $([I(\theta^*)^{-1}]_{i,i}$.
Or

$$[I(\theta^*)^{-1}]_{i,i} \geq \frac{1}{I_{i,i}(\theta^*)}.$$

Information de Fisher : cas général (7)

La variance asymptotique d'une estimateur asymptotiquement efficace de θ_i^* quand tout le vecteur θ^* est inconnu est toujours plus grande que la variance asymptotique d'un estimateur asymptotiquement efficace de θ_i^* quand les autres coordonnées de θ^* sont connues.
Pour faire aussi bien dans les deux cas, il faut que la matrice $I(\theta^*)$ soit diagonale.

Information de Fisher : cas général (8)

La borne de Cramer-Rao ne peut être atteinte que si P_{θ^*} appartient à une famille exponentielle avec

$$L_n(\theta) = \exp\left(Q(\theta)^T \sum_{i=1}^{n} T(X_i) - n\alpha(\theta) \right) \Pi_{i=1}^{n} h(X_i).$$

Alors, à une transformation linéaire près, la seule fonction de θ^* qui peut être estimée efficacement est

$$H(\theta^*) = -\left(\frac{\partial Q_i(\theta^*)}{\partial \theta_j^*} \right)^{-1} \nabla\alpha(\theta).$$

Information de Fisher : cas général (9)

Dans le cas $q = d = 1$,

$$H(\theta) = -\frac{\alpha'(\theta)}{Q'(\theta)}.$$

Un estimateur efficace de $H(\theta^*)$ est alors

$$\widehat{H(\theta)^*} = \frac{1}{n}\sum_{i=1}^{n} T(X_i).$$

Et la variance minimale est

$$\mathrm{Var}_{\theta^*}(\widehat{H(\theta^*)}) = \frac{H'(\theta^*)}{nQ'(\theta^*)}.$$

Tests basés sur la vraisemblance

Exemple (1)

Considérons le modèle de régression linéaire gaussien

$$\mathbf{Y} = m + \varepsilon = \mathbf{X}\theta + \varepsilon,$$

où $\mathbf{X}$ est une matrice $n \times q$ déterministe de plein rang $q = p + 1$, avec p le nombre de variables explicatives et l'on peut écrire

$$\mathbf{X}\theta = \beta_0 V_1 + \cdots + \beta_p V_{p+1} = \beta_0 V_1 + \cdots + \beta_p V_q,$$

$$\text{avec } V_1 = \begin{pmatrix} 1 \\ . \\ 1 \end{pmatrix}, \; V_2 = \begin{pmatrix} X_1^{(1)} \\ . \\ X_n^{(1)} \end{pmatrix}, \; V_3 = \begin{pmatrix} X_1^{(2)} \\ . \\ X_n^{(2)} \end{pmatrix}, \cdots, V_q = \begin{pmatrix} X_1^{(p)} \\ . \\ X_n^{(p)} \end{pmatrix}.$$

On suppose :

1. (A_1) Les ε_i sont indépendants ;
2. (A_2)Les ε_i sont de même loi
3. (A_3)Les ε_i sont i.i.d. $\sim \mathcal{N}(0, \sigma^2)$, même variance

Exemple (2)

1. On peut tester la nullité d'une coordonnée : $(H_0) : \beta_j = 0$ contre $(H_1) : \beta_j \neq 0$. Cette hypothèse s'écrit aussi $m \in W$ où W est engendré par les $q-1$ colonnes de $\mathbf{X}$ (on a enlevé celle correspondant à l'indice j.
2. On peut tester la nullité de plusieurs coordonnées : (H_0) : $\beta_1 = \beta_2 = 0$ contre $(H_1) : \theta \in \mathbb{R}^q$. Cette hypothèse s'écrit aussi $m \in W$ où W est engendré par les $q-2$ colonnes de $\mathbf{X}$ (on a enlevé celle correspondant aux indices 1 et 2).

Pour le test de nullité d'une coordonnée de θ on peut faire un test de Student ou un test de comparaison de modèle emboités (test de Fisher). Pour le test de nullité de plusieurs coordonnées de θ on fera un test de comparaison de modèles (test de Fisher)

Tests de sous-hypothèses (2)

Tester la nullité d'une coordonnée de θ

On souhaite tester $(H_0) : \theta_j = 0$ contre $(H_1) : \theta_j \neq 0$. Soit W le sev engendré par toutes les colonnes de $\mathbf{X}$ sauf la colonne correspondant à θ_j et V est engendré par toutes les colonnes de $\mathbf{X}$. On a $W \subset V$ avec $dim(W) = q - 1$ et $dim(V) = q$.
Ce test revient à tester $(H_0) : m \in W$ contre $(H_1) : m \in V \setminus W$.
On sait que

$$\hat{\theta} \sim \mathcal{N}_q(\theta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}),$$

et aussi que

$$\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}_{j,j}),$$

soit aussi

$$\frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{j,j}}} \sim \mathcal{N}(0,1).$$

Tests de sous-hypothèses (3)

Tester la nullité d'une coordonnée de θ

D'après le théorème de Cochran, $\hat{\theta}$ et $S^2 = (n-q)^{-1} \parallel \mathbf{Y} - \Pi_V(\mathbf{Y}) \parallel^2$ sont indépendantes et

$$\frac{\hat{\beta}_j - \beta_j}{S\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{j,j}}} \sim \mathcal{S}t(n-q).$$

La statistique de test est donc

$$W_n = \frac{\hat{\beta}_j}{S\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{j,j}}} \sim_{H_0} \mathcal{S}t(n-q).$$

On rejette alors (H_0) au niveau α si $|W_n| > t_\alpha$ où t_α est tel que

$$\mathbb{P}(|\mathcal{S}t(n-1)| > t_\alpha) = \alpha.$$

Tests de sous-hypothèses (4)

Notons V est un sous-espace vectoriel de $\mathbb{R}^n$ de dimension q et W un sous-espace vectoriel de V de dimension q' avec $q' < q$. On se propose de faire le test suivant (H_0) : $q - q'$ coordonnées de θ sont nulles contre (H_1) les q sont à priori non nulles.
Sous (H_0) W est engendré par q' colonnes de $\mathbf{X}$ correspondant aux q' coordonnées de θ qui sont non nulles. On note $\mathbf{X}_{H_0}$ la matrice contenant ces q' colonnes
Sous (H_1), V est engendré par les q colonnes de la matrices $\mathbf{X}$. On note $\mathbf{X}_{H_1}$ la matrice contenant ces q colonnes
Ce test équivaut à écrire sous (H_0), $\mathbf{X}_{H_0}\theta \in W \subset V$ avec $\dim(W) = q' < q$ et sous (H_1), $\mathbf{X}_{H_1}\theta \in V \setminus W$ avec $\dim(V) = q$.
Sous (H_0), $\mathbf{X}_{H_0}\hat{\theta}_{H_0} = \Pi_W(\mathbf{Y})$ et sous (H_1), $\mathbf{X}_{H_1}\hat{\theta}_{H_1} = \Pi_V(\mathbf{Y})$.
Le test du rapport de vraisemblance a pour statistique de test

$$T_n = \frac{\| \Pi_V(\mathbf{Y}) - \Pi_W(\mathbf{Y}) \|^2 /(q - q')}{\| \mathbf{Y} - \Pi_V(\mathbf{Y}) \|^2 /(n - p)} \sim^{H_0} \mathcal{F}(q - q', n - q).$$

Contexte

Considérons $X_1, \cdots, X_n$ n variables aléatoires i.i.d. de même loi $P_{\theta^*} \in \{P_\theta, \theta \in \Theta\}$. On suppose pour tout θ, P_θ admet une densité par rapport à la mesure dominante $d\nu(.)$. On a donc

$$\mathbb{P}(X \in A) = \int_A L_{\theta^*}(x)\nu(dx).$$

De la même façon la densité du n-uplet $(X_1, \cdots, X_n)$ notée $L_{\theta^*}(x_1, \cdots, x_n)$ satisfait

$$\mathbb{P}((X_1, \cdots, X_n) \in \mathbb{A}) = \int_{\mathbb{A}} L_{\theta^*}(x_1, \cdots, x_n)\nu(dx_1)\cdots\nu(dx_n).$$

On souhaite faire des tests sur θ^*. On note $\widehat{\theta}$ l'EMV de θ^* ;
Le fait de ne pas avoir un contexte gaussien va impliquer que les résultats vont être asymptotiques.

Outils

1. $\mathbb{E}_{\theta^*}(\dot{\ell}_{\theta^*}(X_1)) = 0$ et $\ell_n(\widehat{\theta}) = 0 \longrightarrow$ Test du Score
2. $\sqrt{n}(\widehat{\theta} - \theta^*) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, I^{-1}(\theta^*)) \longrightarrow$ Test de Wald
3. $2(\ell_n(\widehat{\theta}_{H_1}) - \ell_n(\widehat{\theta}_{H_0}) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{X}(dim(H_1) - dim(H_0)) \longrightarrow$ Test du rapport de vraisemblance

Test du score : $d = 1$

Dans le cas $d = 1$, $(H_0) : \theta^* = \theta_0$ contre $(H_1) : \theta^* \neq \theta_0$. L'outil est $\ell_n(\widehat{\theta}) = 0$ et $\mathbb{E}_{\theta^*}(\dot{\ell}_n(\theta^*)) = 0$.
De plus d'après le TLC

$$\frac{\sqrt{n}\left(\frac{\dot{\ell}_n(\theta^*)}{n} - \mathbb{E}_{\theta^*}\left(\frac{\dot{\ell}_n(\theta^*)}{n}\right)\right)}{\sqrt{\mathrm{Var}_{\theta^*}(\dot{\ell}_{\theta^*})}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

soit aussi

$$\frac{\sqrt{n}\left(\frac{\dot{\ell}_n(\theta^*)}{n} - \mathbb{E}_{\theta^*}\left(\frac{\dot{\ell}_n(\theta^*)}{n}\right)\right)}{\sqrt{I(\theta^*)}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

La statistique de test est donc

$$S_n^2 = \left[\frac{\left(\dot{\ell}_n(\theta_0)\right)}{\sqrt{I_n(\theta_0)}}\right]^2 \xrightarrow[n\to\infty]{H_0,\ \mathcal{L}} (\mathcal{N}(0,1))^2 = \mathcal{X}^2(1)$$

Test du score : cas général

Soit $\theta^* \in \mathbb{R}^d$ et $\varphi : \mathbb{R}^d \mapsto \mathbb{R}^\ell$, une application différentiable dans $\mathbb{R}d$ dont l'application tangente est de plein rang. On souhaite tester

$$(H_0) : \varphi(\theta^*) = 0 \text{ contre } (H_1) : \varphi(\theta^*) \neq 0.$$

La statistique de test est

$$S_n = \left(\ell_n(\widehat{\theta}^{(0)})\right)^T [I_n(\widehat{\theta}^{(0)})]^{-1} \left(\ell_n(\widehat{\theta}^{(0)})\right) \overset{H_0}{\underset{n\to\infty}{\overset{\mathcal{L}}{\longrightarrow}}} \mathcal{X}^2(\ell),$$

où $\widehat{\theta}^{(0)}$ est l'estimateur du maximum de vraisemblance de θ^* sous (H_0). On rejette (H_0) si $S_n > x_\alpha$ où

$$\mathbb{P}(\mathcal{X}(\ell) > x_\alpha) = \alpha.$$

Test de Wald : $d = 1$

Dans le cas $d = 1$, $(H_0) : \theta^* = \theta_0$ contre $(H_1) : \theta^* \neq \theta_0$. L'outil est

$$\sqrt{n}(\widehat{\theta} - \theta^*) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, I^{-1}(\theta^*)).$$

La statistique de test est

$$W_n = \left[\frac{\sqrt{n}(\widehat{\theta} - \theta_0)}{\sqrt{I^{-1}(\theta_0)}}\right]^2 \xrightarrow[n\to\infty]{H_0, \mathcal{L}} \mathcal{X}^2(1).$$

Test de Wald : cas général

Soit $\theta^* \in \mathbb{R}^d$ et $\varphi : \mathbb{R}^d \mapsto \mathbb{R}^\ell$, une application différentiable dans $\mathbb{R}d$ dont l'application tangente est de plein rang. On souhaite tester

$$(H_0) : \varphi(\theta^*) = 0 \text{ contre } (H_1) : \varphi(\theta^*) \neq 0.$$

La statistique de test est

$$W_n = \left(\varphi(\widehat{\theta}^{(0)})\right)^T \left[\varphi'(\widehat{\theta}^{(0)}) I_n^{-1}(\widehat{\theta}^{(0)}) \left(\varphi'(\widehat{\theta}^{(0)})\right)^T\right]^{-1} \left(\varphi(\widehat{\theta}^{(0)})\right) \overset{H_0}{\underset{n\to\infty}{\xrightarrow{\mathcal{L}}}} \mathcal{X}^2(\ell).$$

On rejette (H_0) si $W_n > x_\alpha$ où

$$\mathbb{P}(\mathcal{X}(\ell) > x_\alpha) = \alpha.$$

Si $\varphi(\theta) = Id$ on a $W_n = (\widehat{\theta} - \theta_0)^T I_n^{-1}(\theta_0)(\widehat{\theta} - \theta_0) \overset{H_0}{\underset{n\to\infty}{\xrightarrow{\mathcal{L}}}} \mathcal{X}^2(d)$.

Test du rapport de vraisemblance : $d = 1$

Dans le cas $d = 1$, $(H_0) : \theta^* = \theta_0$ contre $(H_1) : \theta^* \neq \theta_0$. L'outil est

$$2(\ell_n(\widehat{\theta}) - \ell_n(\theta_0) \overset{H_0}{\underset{n\to\infty}{\xrightarrow{\mathcal{L}}}} \mathcal{X}^2(1).$$

Test du rapport de vraisemblance : cas général

Soit $\theta^* \in \mathbb{R}^d$ et $\varphi : \mathbb{R}^d \mapsto \mathbb{R}^\ell$, une application différentiable dans $\mathbb{R}d$ dont l'application tangente est de plein rang. On souhaite tester

$$(H_0) : \varphi(\theta^*) = 0 \text{ contre } (H_1) : \varphi(\theta^*) \neq 0.$$

La statistique de test est

$$\Lambda_n = 2[\ell_n(\widehat{\theta}^{(1)}) - \ell_n(\widehat{\theta}^{(0)})] \xrightarrow[n\to\infty]{\overset{H_0}{\mathcal{L}}} \mathcal{X}^2(\ell),$$

où $\widehat{\theta}^{(1)}$ est l'estimateur de θ^* sous (H_1) et $\widehat{\theta}^{(0)}$ est l'estimateur de θ^* sous (H_0). On rejette (H_0) si $\Lambda_n > x_\alpha$ où

$$\mathbb{P}(\mathcal{X}(\ell) > x_\alpha) = \alpha.$$