Université de Rennes

ENSAI École nationale de la statistique et de l'analyse de l'information

Chapitre 3. Estimation optimale

Claude Petit, université de Rennes - claude.petit@univ-rennes.fr

Sept. 2025

Overview

1. Exhaustivité, minimalité, complétude

2. Information de Fisher

3. Estimateurs optimaux

1. Exhaustivité, minimalité, complétude

Exhaustivité : un premier exemple

$X = (X_1, ..., X_n)$ n-échantillon de bernoulli de paramètre $\theta \in]0, 1[$.
$S = X_1 + X_2 + ... + X_n$. $S \sim \mathcal{B}(n, \theta)$, loi binomiale.
$\mathbb{X} = \{x = (x_1, ..., x_n) : x_i = 0, 1\}$, de cardinal 2^n. Partition de $\mathbb{X}$:

$$\mathbb{X} = \bigcup_{i=0}^{n} [S = i] \tag{1}$$

$$\mathbb{P}[X_i = x_i] = \theta^{x_i}(1-\theta)^{1-x_i}\mathbb{1}_{\{0,1\}}(x_i)$$

$$\mathbb{P}[X = x | S = s] = \frac{\mathbb{P}([X = x] \cap [S = s])}{\mathbb{P}[S = s]} = \frac{\prod_{i=1}^{n}\left(\theta^{x_i}(1-\theta)^{1-x_i}\right)}{\binom{n}{s}\theta^s(1-\theta)^{n-s}}\mathbb{1}_{B_s}(x)$$

$$= \frac{1}{\binom{n}{s}}\mathbb{1}_{B_s}(x)$$

avec $B_s = \{(x_1, ..., x_n) : \sum_{i=1}^{n} x_i = s\}$. Ne dépend pas de θ !

Exhaustivité : illustration

Cette expression ne dépend pas de θ: le fait de savoir la valeur de l'observation x lorsque l'on connaît la somme n'apporte pas plus d'information sur $\theta \Rightarrow$ S condense l'information apportée par le vecteur $(X_1, ..., X_n)$ sans la dégrader. Les données originelles de l'échantillon ne contiennent pas d'information supplémentaire sur la loi de X et peuvent être écartées. S suffit pour apporter toute l'information (en anglais, exhaustif $=$ sufficient).

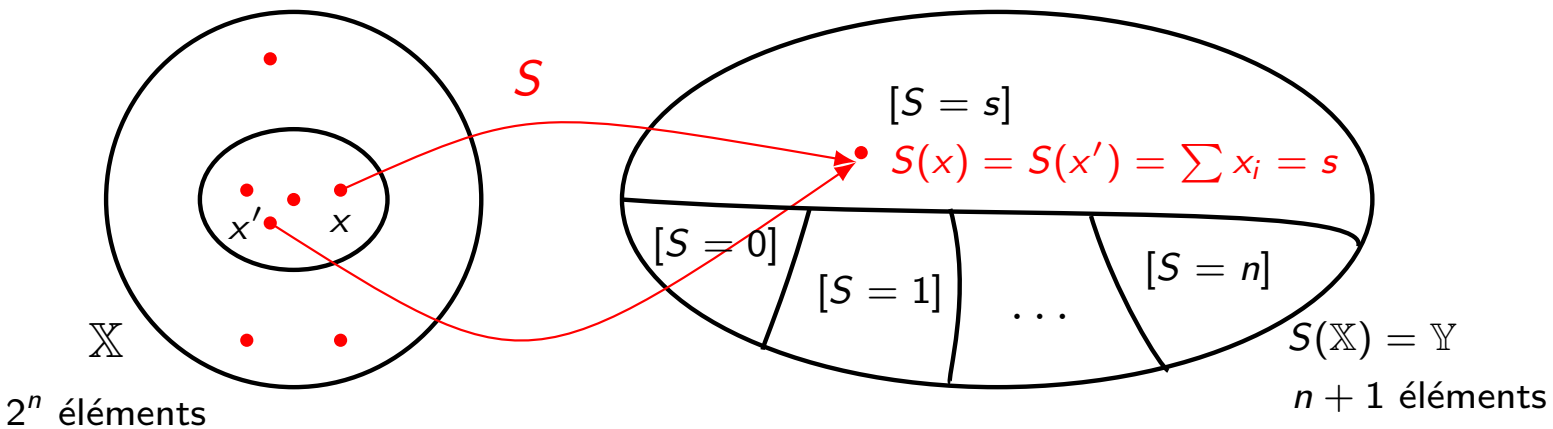

Figure 1: Exhaustivité et tribu engendrée

Exhaustivité : définition

Definition

Une statistique S est exhaustive pour X si la loi conditionnelle de X sachant $[S = s]$ ne dépend pas de θ.

Si la valeur prise par une statistique exhaustive S est connue, égale à s, l'échantillon X ne fournit plus d'information sur θ car sa loi ne dépend plus de θ.
$\Rightarrow$ S condense parfaitement la totalité de l'information relative à θ. Elle contient toute l'information sur le paramètre, sans la dégrader.

Considérons une observation x au travers de la statistique $S(x)$.
L'information sur θ sera la même pour une autre observation y si S est exhaustive et $S(x) = S(y)$.
$\Rightarrow$ La connaissance de S suffit à obtenir toute l'information sur θ.

Critère de factorisation de Neyman

Theorem (Critère de factorisation de Neyman)

T est exhaustive si, et seulement si, la vraisemblance du modèle s'écrit

$$L(x, \theta) = h(x) \times g(S(x), \theta) \tag{2}$$

- Exemple : modèle d'échantillonnage de Bernoulli de vraisemblance

$$L(x, \theta) = \theta^{S(x)}(1-\theta)^{n-S(x)} = h(x) \times g(s, \theta) \tag{3}$$

On a la décomposition avec $h(x) = 1$, $S(x) = \sum_{i=1}^{n} x_i$ et $g(s, \theta) = \theta^s(1-\theta)^{n-s}$. $\Rightarrow$ S est exhaustive.

Exhaustivité : exemples -1-

• $X = (X_1, ..., X_n)$ n-échantillon de loi uniforme sur $[\theta - 1/2, \theta + 1/2]$.
Modèle dominé (mesure de Lebesgue), pas homogène, vraisemblance :

$$L(x, \theta) = \prod_{i=1}^{n} \mathbb{1}_{[\theta-1/2, \theta+1/2]}(x_i) = \mathbb{1}_{[\theta - 1/2 \leq x_{(1)} \leq x_{(n)} \leq \theta + 1/2]}(x)$$
$$= \mathbb{1}_{J(\theta)}(x)$$

de la forme $g((x_{(1)}, x_{(n)}), \theta) \times 1 \Rightarrow (X_{(1)}, X_{(n)})$ exhaustive. L'une de ces deux statistiques seule ne serait pas exhaustive.

• Pour un n-échantillon $X = (X_1, ..., X_n)$ quelconque de vraisemblance

$$L(x, \theta) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} f(x_{(i)}) \tag{4}$$

La statistique d'ordre $X = (X_{(1)}, ..., X_{(n)})$ est exhaustive. Elle ne condense par contre pas du tout l'information...

Exhaustivité : exemples -2-

Dans un modèle exponentiel, la vraisemblance s'écrit

$$L(x,\theta) = h(x) \times g(t,\theta) = h(x) \exp\left(\langle T(x), \lambda(\theta)\rangle - \beta(\theta)\right) \qquad (5)$$

$\Rightarrow$ la statistique naturelle (canonique) est exhaustive.

• Exemple : n-échantillon gaussien $\mathcal{N}(m,\sigma^2)$ avec $\theta = (m,\sigma^2)$.
$T(X) = (T_1(X), T_2(X)) = (\overline{X}, S^2)$ est exhaustive pour θ et

$$L(x,\theta) = (2\pi)^{-n/2} \exp\left(-\frac{n}{2\sigma^2}\left[(t_1 - m)^2 + t_2\right] - \frac{n}{2}\ln\left(\sigma^2\right)\right) \qquad (6)$$

avec $h(x) = (2\pi)^{-n/2}$, $t = (t_1, t_2) = (\overline{x}, s^2)$.

 Dans cet exemple, c'est le théorème de factorisation qui permet de conclure (en l'état, pas de produit scalaire entre les (.)).

Exhaustivité : propriétés -1-

Theorem

Soit T une statistique exhaustive et ϕ une fonction mesurable. Alors la statistique S vérifiant $T = \phi(S)$ est exhaustive.

Si ϕ est bijective, S et T sont alors équivalentes et T exhaustive $\iff$ S exhaustive.

Autrement dit, l'image réciproque d'une statistique exhaustive est exhaustive. L'existence d'une fonction ϕ telle que $T = \phi(S)$ induit une relation d'ordre partiel sur les statistiques. On notera $T < S$. $T < S$ et T exhaustive entraîne S exhaustive, car S est plus fine que T.

Exhaustivité : propriétés -2-

• Si ϕ n'est pas bijective, alors la tribu induite va être plus grossière : $\sigma(T) \subset \sigma(S)$ et l'information va être plus condensée.

• On peut se demander s'il existe une statistique qui condense l'information de façon maximale sans la dégrader.

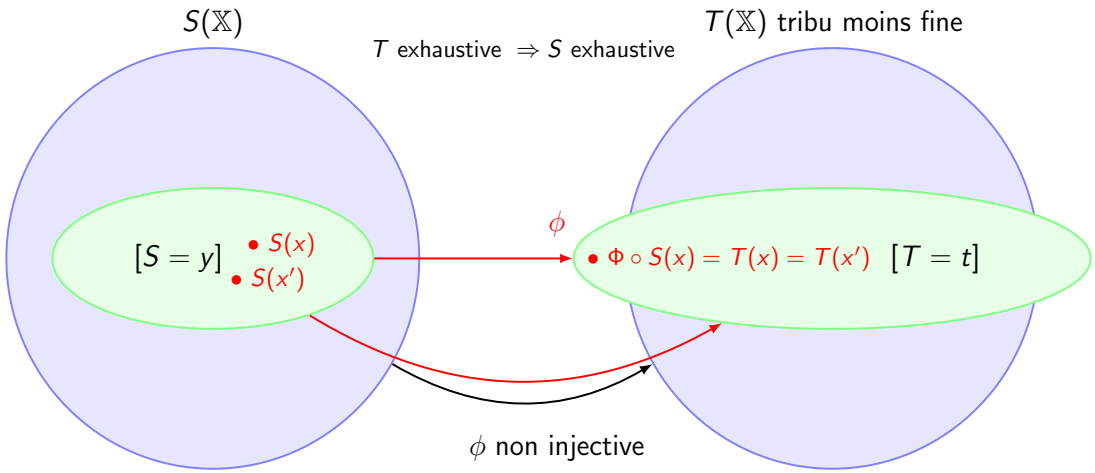

Figure 2: Illustration de l'exhaustivité qui condense l'information.

Minimalité et rapport de vraisemblance : définition

Definition

Une statistique exhaustive S est minimale si elle est fonction de toute autre statistique exhaustive. Autrement dit, $\forall \theta \in \Theta$,

$$T \text{ exhaustive minimale } \iff \forall S \text{ exhaustive }, \ \exists \ \phi : T = \phi(S), \mathbb{P}_\theta - \text{p.s.}$$

- T est fonction d'une statistique S ssi $S(x) = S(x') \Rightarrow T(x) = T(x')$.

Autrement dit, les évènements de la forme $[S(x) = y']$ sont chacun inclus dans un évènement de la forme $[T(x) = y]$.

- La partition associée à une statistique exhaustive minimale est la plus grossière possible et la réduction est maximale.

- Le rapport de vraisemblance (likelihood ratio) est, pour x, y fixés:

$$\theta \longrightarrow LR(x, y, \theta) = \frac{L(x, \theta)}{L(y, \theta)} \tag{7}$$

Critère de minimalité

Theorem (Critère de minimalité)

S statistique d'un modèle dominé, telle que

$$LR(x, y, \theta) \text{ ne dépend pas de } \theta \iff S(x) = S(y) \tag{8}$$

Alors S est exhaustive et minimale.

• Exemple : reprenons l'exemple du début de chapitre:

$$LR(x, y, \theta) = \theta^{S(x)-S(y)}(1-\theta)^{S(y)-S(x)} = \left(\frac{\theta}{1-\theta}\right)^{S(x)-S(y)} \tag{9}$$

qui est constant ssi $S(x) = S(y)$. Donc $S(X) = \sum_{i=1}^{n} X_i$ statistique exhaustive minimale.

• Exemple : échantillon uniforme sur $[\theta - 1/2, \theta + 1/2]$. En calculant LR, on voit (à faire) que $(X_{(1)}, X_{(n)})$ est une statistique exhaustive minimale.

Exemple du modèle exponentiel

Theorem

Si le modèle est exponentiel, alors:

- *La statistique canonique S est exhaustive.*
- *Si l'espace Λ des paramètres canoniques contient un ouvert de $\mathbb{R}^k$ ou un repère affine de $\mathbb{R}^k$, alors T est minimale*

$$LR(x, y, \theta) = g(\theta) = \frac{h(x)}{h(y)} \exp\left(\langle \lambda(\theta), S(x) - S(y) \rangle\right) \tag{10}$$

- Exemple: $(X_1, ..., X_n)$ n-échantillon $\mathcal{N}(\theta, \theta^2)$. Modèle exponentiel avec paramètre $(1/\theta, -1/\theta^2)$ et statistique $(n\overline{x}, -n\overline{x^2}/2)$.

Espace canonique Λ formé par une courbe d'équation $u = -\lambda^2$ fermée dans $\mathbb{R}^2$ (ne peut contenir aucun ouvert de $\mathbb{R}^2$). Mais on peut trouver trois points non alignés sur cette courbe. $\Rightarrow$ statistique exhaustive minimale.

$$\Lambda = \left\{ \left(\frac{1}{\theta}, -\frac{1}{\theta^2} \right) ; \theta > 0 \right\} = \{(\lambda, -\lambda^2); \lambda > 0\} \subset \mathbb{R}^2 \tag{11}$$

Statistique libre et complète

Une statistique exhaustive apporte toute l'information sur θ contenue dans X. Une statistique dont la loi ne dépend pas de θ n'apporte aucune information sur θ.

Definition

- Une statistique S est libre vis à vis de θ si sa loi ne dépend pas de θ.
- Elle est libre du premier ordre si la fonction (de θ) $\mathbb{E}_\theta[S]$ est constante.
- Une statistique U est ignorable s'il existe une statistique exhaustive S indépendante de U. S apporte alors toute l'information sur θ et U apporte une information complémentaire à S.

Si deux statistiques sont dépendantes, alors les informations qu'elles apportent sur un paramètre sont redondantes.

Les différents types d'information

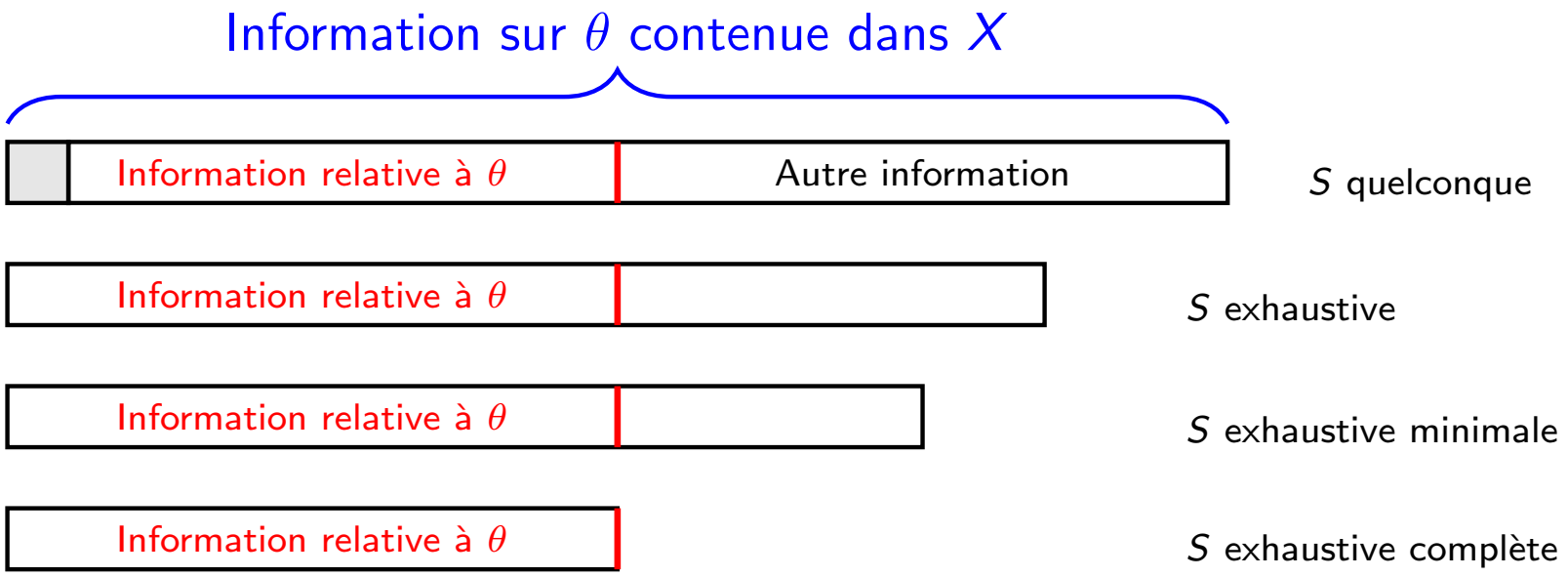

Figure 3: Information, exhaustivité, minimalité, complétude.

• Exemple: modèle d'échantillonnage gaussien $\mathcal{N}(\theta, 1)^{\otimes n}$. La statistique $S^2 = \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 / n$ est libre, car sa loi ne dépend pas de θ (poser $Y_i = X_i - \theta$). $Y \sim \mathcal{N}(0, I_n)$ et

$$S^2 = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 / n$$

Liberté : exemples

• Exemple : X échantillon de loi uniforme sur $[\theta - 1/2, \theta + 1/2]$.
$\Rightarrow S = (S_1, S_2) = (X_{(1)}, X_{(n)})$ exhaustive minimale de densité

$$g(s,t) = n(n-1)(t-s)^{n-2}\mathbb{1}_{[\theta-1/2\leq s\leq t\leq\theta+1/2]} \tag{12}$$

$U = X_{(n)} - X_{(1)}$ statistique libre de θ de densité (exercice 😱)

$$h(u) = n(n-1)u^{n-2}\mathbb{1}_{[0,1]}(u) \tag{13}$$

Pourtant, U non indépendante de S. S apporte donc une information inutile U sur θ.

• Exemple : $X = (X_1, ..., X_n)$ n-échantillon dont la loi a pour densité

$$f(x) = \frac{1}{\ln\alpha}\frac{1}{x}\mathbb{1}_{]\theta,\alpha\theta[}(x) \tag{14}$$

$\alpha > 1$, $\theta > 0$ inconnu. $(X_{(1)}, X_{(n)})$ exhaustive minimale. Idem pour $S(X) = (X_{(n)}/X_{(1)}, X_{(1)}X_{(n)})$. $U(X) = X_{(n)}/X_{(1)}$ est libre et il est clair que S et U ne sont pas indépendantes.

Statistique complète : motivation

• Une statistique exhaustive minimale réduit au maximum les données sans perdre d'information sur θ.

• Mais une telle statistique n'est pas nécessairement indépendante de toute statistique libre, car une statistique libre peut compléter l'information apportée par une statistique minimale.

• La complétude assure que la statistique est bien indépendante de toute statistique libre: toute la liberté résiduelle a été extraite de la statistique et toute transformation la rend non libre.

• On cherche des statistiques exhaustives S ne contenant aucune composante libre, c'est à dire que pour toute fonction g non constante, $g(S)$ ne sera pas libre au premier ordre. Ou encore, si l'espérance est constante, alors g est constante. Ces statistiques ne doivent contenir aucun matériel (c.-à-d. aucune fonction de S) libre, même du premier ordre. 😱

$\Rightarrow$ Cette notion est apportée par la complétude.

Statistique complète : définition

Definition

Une statistique S est complète s'il n'existe pas de fonction g non constante, intégrable, à valeurs dans $\mathbb{R}$ qui soit libre. Autrement dit,

$$\forall \theta,\ \mathbb{E}_\theta[g(S)] = \text{constante} \Rightarrow g = \text{constante}\ \mathbb{P}_\theta - \text{p.s.} \tag{15}$$

Theorem (Lehmann & Scheffé (1950), Bahadur (1957))

S exhaustive complète $\Rightarrow$ S exhaustive minimale.

Statistique complète : exemples -1-

• Exemple : on considère (encore) un n-échantillon de loi de Bernoulli de paramètre θ inconnu et l'on pose $S(X) = \sum_{i=1}^{n} X_i$ dont on a vu qu'elle est une statistique exhaustive minimale. Est-elle complète ? Considérons $g : \{0, ..., n\} \longrightarrow \mathbb{R}$ telle que $\mathbb{E}_\theta[g(S)] = 0$. Comme la loi de T est une loi binomiale, on a

$$h(\theta) = \mathbb{E}_\theta[g(S)] = \sum_{s=0}^{n} g(s) \left(\begin{array}{c} n \\ s \end{array} \right) \theta^s (1-\theta)^{n-s} = 0 \qquad (16)$$

h est donc un polynôme de degré n avec une infinité de zéros: il est nul et par suite S est complète.

• Exemple : on considère un n-échantillon de loi uniforme sur $[\theta - 1/2, \theta + 1/2]$ et $S = (X_{(1)}, X_{(n)})$. S n'est pas complète 😱 (à faire en exercice).

Statistique complète : exemples -2-

$(X_1, ..., X_n)$ n-échantillon de densité $f(x) = e^{x-\theta}\mathbb{1}_{]-\infty,\theta]}(x)$.

$X_{(n)} = \max_{i=1}^{n} X_i$ statistique exhaustive pour θ de densité

$$h(u) = ne^{n(u-\theta)}\mathbb{1}_{]-\infty,\theta]}(u) \tag{17}$$

Soit g fonction d'intégrale nulle par rapport à cette densité, $\forall\theta$.

$$\forall\theta \in \mathbb{R}, \int_{-\infty}^{\theta} g(u)ne^{n(u-\theta)}du = 0 \tag{18}$$

$$\Rightarrow \forall\theta \in \mathbb{R}, \int_{-\infty}^{\theta} g(u)e^{nu}du = 0 \tag{19}$$

Ainsi, sur tout intervalle $[\theta, \theta']$,

$$\int_{\theta}^{\theta'} g(u)e^{nu}du = 0 \tag{20}$$

 La fonction g est donc nulle presque partout (pourquoi ?).

Statistique complète : exemples -3-

On considère le modèle exponentiel de vraisemblance

$$L(x,\theta) = h(x)\exp\left(\langle\lambda(\theta), S(x)\rangle - \beta(\theta)\right) \tag{21}$$

Avec $\Lambda = \{\lambda(\theta); \theta \in \Theta\}$ espace canonique des paramètres. Si Λ contient un ouvert non vide, alors S est exhaustive complète. En effet, $\forall g$,

$$\mathbb{E}_\theta[g(S)] = \int_{\mathbb{R}^k} g(s)e^{<\lambda,s>}d\mu(s) \tag{22}$$

Transformée de Laplace de g. Ses propriétés (dans $\mathbb{C}$) $\Rightarrow g \equiv 0$ (principe des zéros isolés).

• Exemple : échantillon de Bernoulli ; statistique canonique
$S(x) = \sum_i X_i$, paramètre canonique $\lambda(\theta) = \ln(\theta/(1-\theta))$.
$\Lambda = \{\lambda(\theta) = \ln(\theta/(1-\theta)), \theta \in]0,1[\}$.

Étude de fonction $\Rightarrow \Lambda(\Theta) = \mathbb{R}$ ouvert de $\mathbb{R}$. $\Rightarrow S$ complète.

Statistique complète : exemples -4-

• Exemple : $(X_1, ..., X_n)$ n-échantillon de loi $\mathcal{N}(m, \sigma^2)$ avec $\theta = (m, \sigma^2)$.

On pose $S_1 = \sum X_i$, $S_2 = -\sum X_i^2$, $S = (S_1, S_2)$ et

$$\lambda = \left(\frac{m}{\sigma^2}, \frac{1}{2\sigma^2} \right) \tag{23}$$

Le modèle est celui d'une famille exponentielle de rang plein.

$\Rightarrow$ S est complète pour λ et également pour θ par bijectivité de $\lambda(\theta)$.
Ainsi, $(\overline{X}, S^2)$ est une statistique complète pour $\theta = (m, \sigma^2)$.

• Exercice : trouver une statistique exhaustive minimale complète pour modèle le multinomial.

Théorème de Basu

Theorem (de Basu)

Soit S une statistique exhaustive complète et U une statistique libre. Alors $S \perp\!\!\!\perp U$.

Le théorème de Basu donne un moyen rapide de démontrer que la moyenne et la variance empirique d'un échantillon gaussien sont indépendantes:

$\overline{X}$ est exhaustive et complète pour la moyenne θ (c'est un modèle exponentiel). Par ailleurs S^2 est libre. D'après le théorème de Basu, les deux statistiques sont donc indépendantes.

Exhaustivité, minimalité, liberté, complétude : conclusion

- Ce sont des notions difficiles à bien comprendre.
- En anglais, exhaustif = sufficient, libre = ancillary.
- À voir :
  - Minimalité : https://www.youtube.com/watch?v=IsgteDaNTFk
  - Complétude : https://www.youtube.com/watch?v=GF8nFqEbqkI

2. Information de Fisher

Information de Fisher : introduction et notations

Problème : quantitier la quantité d'information sur θ contenu dans un échantillon. Plusieurs définitions différentes de la notion de quantité d'information.

- Information au sens de Shannon (1948).
- Entropie de Boltzmann en thermodynamique.
- Information au sens de Kullback-Leibler.
- Ici : information au sens de Fisher.

Intérêt de l'information de Fisher :

- Notion locale.
- Pouvoir de discrimination du modèle entre deux valeurs proches de θ.

Score d'un modèle statistique

- $\theta = (\theta_1, ..., \theta_p) \in \Theta \subset \mathbb{R}^p$ ouvert.
- $g : \Theta \longrightarrow \mathbb{R}$ fonction $2\times$ différentiable.
- $\nabla g(\theta)$ gradient de g en θ
- $H_g(\theta)$ ou $\nabla^2 g(\theta)$ matrice hessienne.
- $\mathbb{V}(S)$ matrice de covariance d'un vecteur $S = (S_1, ..., S_p)^T \in \mathbb{R}^p$.

$$\mathbb{V}(S) = \mathbb{E}\left[(S - \mathbb{E}[S]) \times (S - \mathbb{E}[S])^T\right] = \begin{pmatrix} \mathrm{cov}(S_1, S_1) & \dots & \mathrm{cov}(S_1, S_p) \\ \vdots & & \vdots \\ \mathrm{cov}(S_p, S_1) & \dots & \mathrm{cov}(S_p, S_p) \end{pmatrix}$$

- $l(X, \theta) = \ln L(X, \theta)$ log-vraisemblance.

Le score du modèle est (sous réserve d'existence),

$$S(X, \theta) = \frac{\partial}{\partial \theta} \ln L(X, \theta) = \nabla l(\theta) \tag{24}$$

Fonction de X donc v.a. mais dépend de θ :  pas une statistique !

Notion de modèle régulier

Definition

Un modèle paramétrique est régulier si, et seulement si,

- Il est dominé, homogène et $\Theta \subset \mathbb{R}^p$ est un ouvert.
- $L(x, \theta) > 0 \ \forall x \in \mathbb{X}, \forall \theta \in \Theta$.
- $\theta \longrightarrow L(x, \theta)$ est de classe C^2 pour presque tout x.
- $\forall B \in \mathbb{X}$, $\theta \longrightarrow \int_B L(x|\theta) d\mu(x)$ est $2\times$ différentiable sous le signe $\int$.
- Le score $S(X, \theta) \in L^2(\mathbb{P}_\theta)$.

Les modèles gaussiens ou de Poisson sont réguliers, mais pas le modèle uniforme sur $[0, \theta]$.

Information de Fisher : définition

Definition

L'information de Fisher $\mathbb{I}(\theta)$ du modèle est la variance du score:

$$\mathbb{I}(\theta) = \mathbb{V}_\theta(S) \tag{25}$$

Theorem (Information d'un modèle régulier)

- $\mathbb{E}_\theta[S] = 0$
- $\mathbb{I}(\theta) = \mathbb{E}_\theta[SS^T] = -\mathbb{E}_\theta\left[\frac{\partial^2 l}{\partial \theta^2}\right] = -\mathbb{E}_\theta\left[H_l\right]$

En dimension 1, $\mathbb{I}(\theta) = \mathbb{E}\left[S^2\right] = -\mathbb{E}\left[l''(\theta)\right]$

Les modèles gaussiens ou de Poisson sont réguliers, mais pas le modèle uniforme sur $[0, \theta]$.

Information de Fisher : illustration -1-

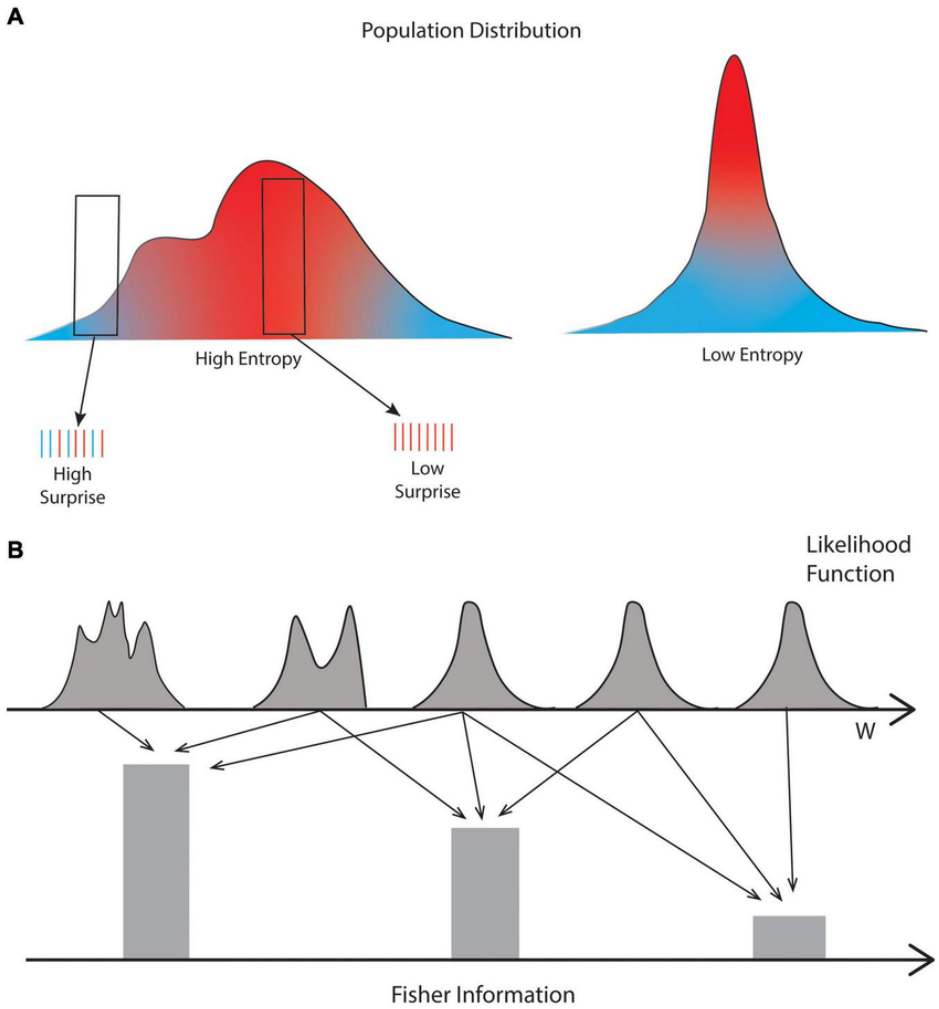

Figure 4: Information de Fisher et de Shannon. A: l'entropie de Shannon capture l'étalement (en rouge et bleu) de la statistique en terme de vraisemblance. B: l'information de Fisher capture la quantité d'information que la vraisemblance apporte sur la vraie valeur du paramètre, grâce à la variation de la fonction de vraisemblance. $\mathbb{I}(\theta) = -\mathbb{E}\left[I''(\theta)\right]$ (R. Grzywacz, H. Aleem).

Information de Fisher : illustration -2-

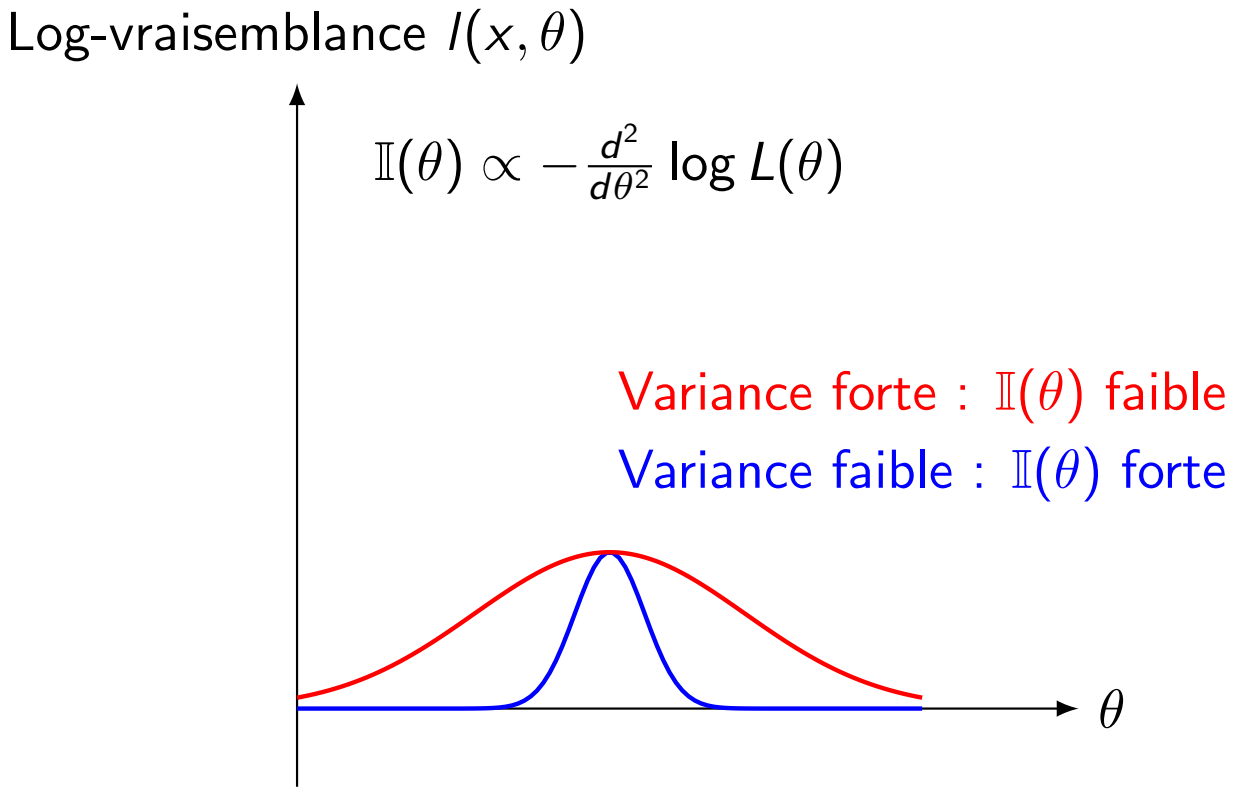

Figure 5: Information de Fisher : $\mathbb{I}(\theta) = -\mathbb{E}\left[I''(\theta)\right]$. En rouge : grande variance donc peu d'information en θ, en bleu : variance faible, donc forte information sur θ au voisinage du maximum de $I(\theta)$.

Information de Fisher : exemple

Considérons un n-échantillon $X = (X_1, ..., X_n)$ de loi de Benoulli de paramètre θ et considérons la statistique $T(X) = \sum_{i=1}^{n} X_i$. La vraisemblance du modèle est

$$L(X, \theta) = \exp\left(T(X) \ln \theta + (n - T(X) \ln(1 - \theta)\right) \tag{26}$$

Considérons une observation $x \in \mathbb{R}^n$. x vecteur de dimension n et $\theta \in]0, 1[$ paramètre scalaire. Le score du modèle est

$$S(x, \theta) = \frac{\partial l(x, \theta)}{\partial \theta} = \frac{T(x)}{\theta} - \frac{n - T(x)}{1 - \theta} = \frac{T(x)}{\theta(1 - \theta)} - \frac{n}{1 - \theta} \tag{27}$$

On en déduit l'information du modèle :

$$\mathbb{I}(\theta) = \mathbb{V}\left(S(X, \theta)\right) = \frac{n}{\theta(1 - \theta)} \tag{28}$$

Propriétés de l'information de Fisher -1-

- Information de Fisher apportée par une statistique.

Soient $(\mathbb{X}, \mathfrak{F}, (\mathbb{P}_\theta, \theta \in \Theta))$ un modèle régulier et T une statistique pour laquelle le modèle image est également régulier.

Soit $g(t, \theta)$ la densité de T.

L'information de Fisher apportée par T est l'information de Fisher du modèle image:

$$I_T(\theta) = \mathbb{V}\left(S(T, \theta)\right) \tag{29}$$

Le score du modèle image par T est :

$$\frac{d}{d\theta} \ln g(T, \theta) \tag{30}$$

Conséquence directe du théorème de transfert.

Propriétés de l'information de Fisher -2-

- Additivité de l'information.

Si T_1 et T_2 sont deux statistiques indépendantes alors

$$I_{(T_1,T_2)}(\theta) = I_{T_1}(\theta) + I_{T_2}(\theta) \tag{31}$$

Dans un modèle d'échantillonnage régulier $X = (X_1, ..., X_n)$,

$$\mathbb{I}_X(\theta) = n.\mathbb{I}_{X_1}(\theta) \tag{32}$$

 Résultat fondamental ! L'information a été construite sur la base de cette propriété : 2 phénomènes aléatoires indépendants doivent apporter une quantité d'information totale égale à la somme de leur quantité d'information respective.

Propriétés de l'information de Fisher -3-

- Information conditionnelle.

Soit T statistique de densité. Alors

$$I(\theta) = I_T(\theta) + I_{X|T}(\theta) \tag{33}$$

où $I_{X|T}$ est l'information de Fisher de X sachant T.

Illustre également la dégradation de la quantité d'information :

$\mathbb{I}(\theta) \geq \mathbb{I}_T(\theta)$. Si $\theta \in \mathbb{R}$, $\mathbb{I}_{X|T}(\theta) > 0$. Si $\theta \in \mathbb{R}^p$, $\mathbb{I}_{X|T}(\theta) \in \mathrm{Sym}^+(\mathbb{R}^p)$.

Pour des matrices, $A \geq B \iff \forall \theta \in \mathbb{R}^p,\ \theta^T A \theta \geq \theta^T B \theta$.

T statistique exhaustive $\Rightarrow I_X(\theta) = I_T(\theta)$.

C'est même une équivalence, sous hypothèse de régularité:
T exhaustive $\iff I_X(\theta) = I_T(\theta)$

Pour une statistique T libre de θ, $I_T(\theta) = 0$ (c'est aussi une $\iff$).

Exemple d'un échantillon gaussien -1-

- $(X_1, ..., X_n)$ n-échantillon gaussien d'une v.a. $X \sim \mathcal{N}(m, \sigma^2)$.
- $\theta = (m, \sigma^2)$ paramètre vectoriel.
- Modèle régulier de log-vraisemblance d'une observation x

$$I(x, \theta) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(x - m)^2 \tag{34}$$

$$\begin{cases} \dfrac{\partial^2 I}{\partial m^2}(x, \theta) = -\dfrac{1}{\sigma^2} \\ \dfrac{\partial^2 I}{\partial (\sigma^2)^2}(x, \theta) = \dfrac{1}{2\sigma^4} - \dfrac{(x - m)^2}{\sigma^6} \\ \dfrac{\partial^2 I}{\partial m \partial (\sigma^2)}(x, \theta) = \dfrac{x - m}{\sigma^4} \end{cases} \tag{35}$$

Exemple d'un échantillon gaussien -2-

Ainsi $\mathbb{I}_X(\theta)$

$$= -\mathbb{E}\left[\begin{pmatrix} \frac{\partial^2 l}{\partial m^2}(X,\theta) & \frac{\partial^2 l}{\partial m \partial(\sigma^2)}(X,\theta) \\ \frac{\partial^2 l}{\partial m \partial(\sigma^2)}(X,\theta) & \frac{\partial^2 l}{\partial(\sigma^2)^2}(X,\theta) \end{pmatrix}\right] = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{pmatrix}$$

et l'information de Fisher associée au n-échantillon sera donc

$$\mathbb{I}(\theta) = \begin{pmatrix} n/\sigma^2 & 0 \\ 0 & n/\sigma^4 \end{pmatrix} \tag{36}$$

Propriétés de l'information de Fisher -4-

- Information de Fisher et reparamétrisation.

Considérons un nouveau paramètre $\lambda = \phi(\theta)$. On peut calculer l'information de Fisher du modèle relativement à ce nouveau paramètre via la formule suivante :

$$\mathbb{I}_X(\lambda) = J^T \mathbb{I}_X(\theta) J = \left(\nabla \phi^{-1}(\lambda)\right)^T \times \mathbb{I}_X(\theta) \times \left(\nabla \phi^{-1}(\lambda)\right) \tag{37}$$

où J matrice jacobienne de l'application inverse de ϕ.

Pour exprimer cette nouvelle information en fonction de λ, il faut pouvoir exprimer θ en fonction de λ en inversant ϕ.

En dimension 1, si $\theta, \lambda \in \mathbb{R}$, la formule devient:

$$\mathbb{I}_X(\lambda) = \frac{\mathbb{I}_X(\theta)}{\phi'(\theta)^2} \tag{38}$$

Information de Kullback

Definition

$$K(\theta_0, \theta_1) = \mathbb{E}_{\theta_0}\left[\ln \frac{L(X, \theta_0)}{L(X, \theta_1)}\right] = \int_{\mathbb{R}^n} \ln \frac{L(x, \theta_0)}{L(x, \theta_1)} L(x, \theta_0) dx \tag{39}$$

$\ln\left(L(X, \theta_0)/L(X, \theta_1)\right)$ est le pouvoir discriminant de θ_0 contre θ_1.

Theorem

$$\left.\frac{\partial^2 K(\theta_0, \theta)}{\partial \theta^2}\right|_{\theta=\theta_0} = \mathbb{I}(\theta_0) \tag{40}$$

- Exemple:

si X suit une loi de Poisson $\mathcal{P}(\theta)$, alors $K(\theta_0, \theta_1) = \theta_1 - \theta_0 + \theta_0 \ln(\theta_0/\theta_1)$.

Information de Kullback : illustration

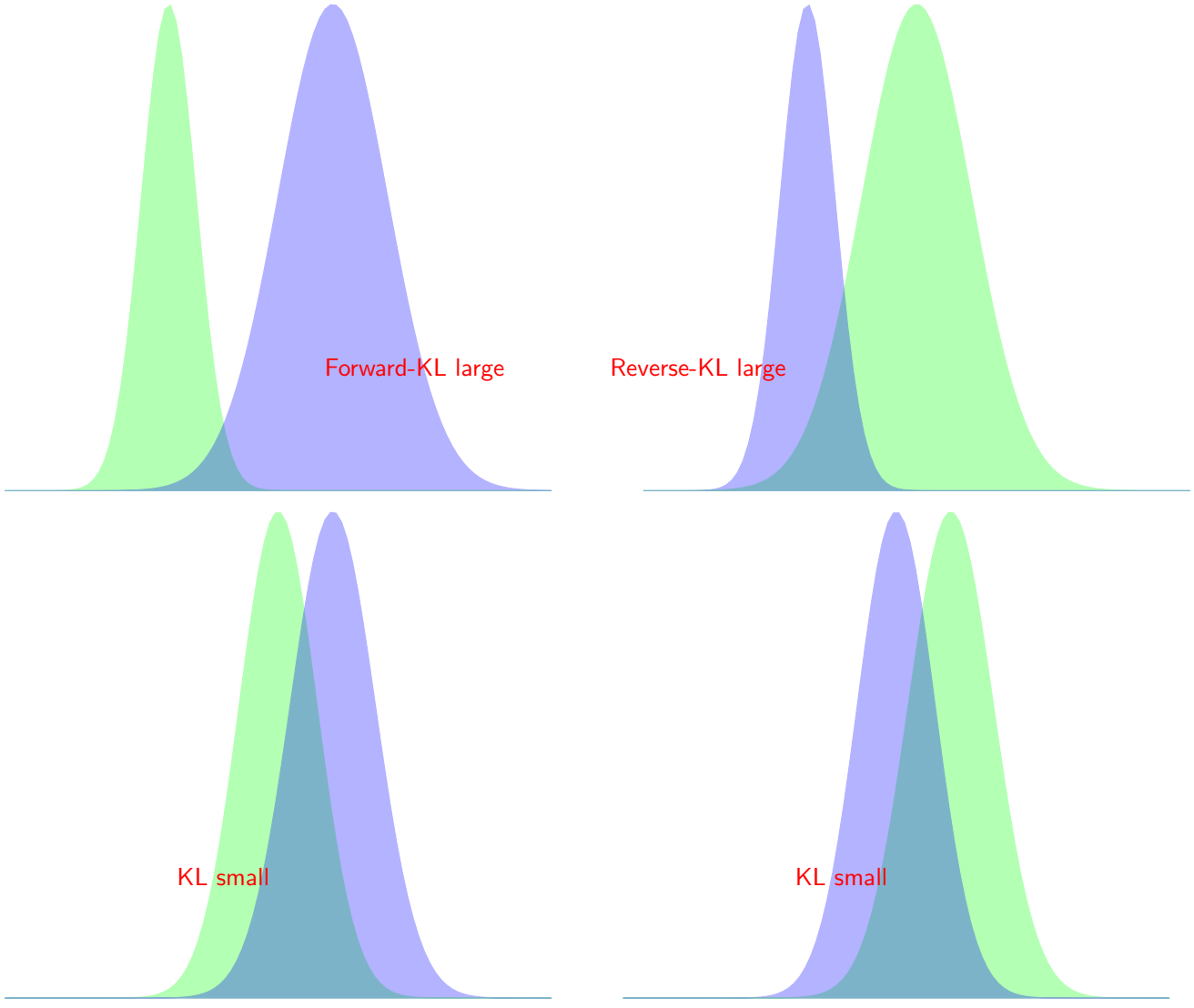

Figure 6: Information de Kullabck-Leibler.

3. Estimateurs optimaux

Fonction de coût et risque moyen

Une fonction de coût mesure l'écart (la perte) entre deux quantités.

Qualité d'un estimateur S appréciée en fonction de l'écart entre S et θ.

Une fonction de coût c doit être positive et vérifier $\theta = \theta' \Rightarrow c(\theta, \theta') = 0$.

Exemples : toute distance sur Θ, coût quadratique c_2, absolu c_1 ou 0-1 noté c_0:

$$c_2(\theta, \theta') = (\theta - \theta')^2 \qquad c_1(\theta, \theta') = |\theta - \theta'| \qquad c_0(\theta, \theta') = \mathbb{1}_{[|\theta - \theta'| > \epsilon|]}$$

Definition

Le risque moyen d'un estimateur T de θ est

$$R(T, \theta) = \mathbb{E}[c(T, \theta)]$$

Estimateur préférable et admissible

Un estimateur est préférable à un autre (à θ fixé) si son risque moyen est inférieur.

Il est uniformément préférable si le risque est inférieur quelque soit $\theta \in \Theta$.

Definition

Étant donné une classe $\mathcal{T}$ d'estimateurs de θ, $T \in \mathcal{T}$ est admissible dans $\mathcal{T}$ pour θ s'il est uniformément préférable à tout autre estimateur de $\mathcal{T}$:

$$\forall T' \in \mathcal{T}, \forall \theta \in \Theta,\ R(T, \theta) \leq R(T', \theta) \tag{41}$$

Estimateur VUMSB

Objectif : déterminer l'estimateur admissible pour la classe des estimateurs sans biais.

La formule reliant le biais et la variance nous donne

$$R(T,\theta) = \mathbb{V}_\theta(T) + (\mathbb{E}_\theta[T] - \theta)^2 \tag{42}$$

Definition

Un estimateur T^* est de variance uniformément minimale parmi les estimateurs sans biais de θ (on dira VUMSB) si T^* est admissible pour le risque quadratique, dans la classe des estimateurs sans biais :

- $\mathbb{E}[T^*] = \theta$ (43)
- $\forall\ T$ sans biais , $\mathbb{V}(T^*) \leq \mathbb{V}(T)$ (44)

Lorsqu'il existe, l'estimateur VUMSB est unique $\mathbb{P}_\theta$-p.s. pour tout θ.

Théorèmes de Rao-Blackwell et Lehmann-Scheffé

Theorem (de Rao-Blackwell)

Soit T un estimateur sans biais et S une statistique exhaustive. Alors

$$T^* = \mathbb{E}_\theta[T|S]$$

estimateur sans biais de θ préférable à T pour le risque quadratique.

Lorsque T VUMSB et S exhaustive, T^* uniformément préférable à T et T^* est aussi VUMSB. $\Rightarrow T = T^*$.

Theorem (de Lehmann-Scheffé)

T estimateur sans biais et S statistique complète.
Alors $T^* = \mathbb{E}_\theta[T|S]$ est VUMSB.

Estimateur VUMSB : exemple gaussien

$(X_1, ..., X_n)$ n-échantillon gaussien $\sim \mathcal{N}(m, \sigma^2)$, $\theta = (m, \sigma^2)$.

$$L(x, \theta) = (2\pi)^{-n/2} \exp\left[-\frac{n}{2\sigma^2}\overline{x^2} + \frac{nm}{\sigma^2}\overline{x} - \frac{nm^2}{2\sigma^2} - \frac{n}{2}\ln\sigma^2\right] \qquad (45)$$

Modèle exponentiel (homogène), statistique naturelle : $\left(\overline{X^2}, \overline{X}\right)$.

Paramètre canonique : $(-n/(2\sigma^2), nm/\sigma^2)$.

Espace des paramètres canoniques : $\Lambda =]-\infty, 0[\times\mathbb{R}$.

Ouvert $\Rightarrow$ statistique canonique exhaustive et complète.

$T = \left(\overline{X}, \overline{X^2}\right)$. Estimateur sans biais : $\left(\overline{X}, \widetilde{S}^2\right)$,
$\widetilde{S}^2 = n/(n-1) \times \left(\overline{X^2} - \overline{X}^2\right)$. Théorème de Lehmann-Scheffé :

$$\Rightarrow \left(\mathbb{E}_\theta\left[\overline{X}|\, T\right], \mathbb{E}_\theta\left[\widetilde{S}^2|\, T\right]\right) \text{ VUMSB}$$

Mais $\overline{X}$ et $\widetilde{S}^2$ fonctions de T, $\Rightarrow \left(\overline{X}, \widetilde{S}^2\right)$ VUMSB pour θ.

Borne de Fréchet, Darmois, Cramer, Rao

Risque quadratique + classe estimateurs sans biais $\Rightarrow$ borne inférieure pour la variance.

Theorem (Borne FDCR)

$X \sim \mathbb{P}_\theta$ modèle régulier avec $\mathbb{I}(\theta)$ inversible $\forall \theta \in \Theta$.
T estimateur sans biais de $\phi(\theta) \in \mathbb{R}^d$ tel que :

- $\phi(\theta) = \mathbb{E}_\theta[T]$ *est $\neq$tiable en θ.*
- $\mathbb{E}_\theta[T]$ *est $\neq$tiable en θ sous le signe $\int$.*

$$\Rightarrow \mathbb{V}_\theta(T) \geq \nabla\phi(\theta) \times \mathbb{I}(\theta)^{-1} \times \nabla\phi(\theta)' \tag{46}$$

Si T sans biais, $\mathbb{E}_\theta[T] = \theta$ ($\phi = Id$) et FDCR : $\mathbb{V}_\theta(T) \geq \mathbb{I}(\theta)^{-1}$

Definition

T estimateur de $\phi(\theta)$ efficace si sans biais et atteint FDCR.

Borne FDCR : exemple gaussien

- Pour $\theta = (m, \sigma^2)$. La borne de Cramer-Rao est donnée par

$$\mathbb{I}(\theta)^{-1} = \begin{pmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/n \end{pmatrix} \tag{47}$$

- Rappel : matrice de covariance de l'estimateur $T = (\overline{X}, \widetilde{S}^2)$:

$$\mathbb{V}_\theta(T) = \begin{pmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/(n-1) \end{pmatrix} \tag{48}$$

 : le calcul de la variance de la variance empirique est lourd....

L'e.m.v. n'est donc pas efficace pour θ.

Par contre $\overline{X}$ efficace pour $m = \phi(\theta)$ avec ϕ projection sur (Ox). Mais S^2 non efficace pour $\sigma^2 = \psi(\theta)$ avec ψ projection sur (Oy).

Efficacité asymptotique de l'e.m.v.

Theorem (Efficacité asymptotique de l'e.m.v.)

Soit $X_1, ..., X_n$ *n-échantillon de* $X \sim \mathbb{P}_\theta$, $\theta \in \Theta$,
$f(x, \theta)$ *densité de* $\mathbb{P}_\theta$ *par rapport à* μ. *On suppose que:*

- *Le modèle est homogène et identifiable.*
- $\ln f(x, \theta)$ *de classe* C^2 *en* θ *et de* $\neq$*tielle* 2^e *localement holdérienne :*
$\forall \theta \in \Theta, \exists\ V$ *voisinage de* θ *et* $M : \mathbb{X} \to \mathbb{R}_+$ *tel que*

$$\forall \theta_1, \theta_2 \in V : \left|\left| \frac{d^2}{d\theta^2} \ln f(x, \theta_1) - \frac{d^2}{d\theta^2} \ln f(x, \theta_2) \right|\right| \leq M(x) || \theta_1 - \theta_2 ||,\ \mathbb{P}_\theta p.s.$$

- Θ *ouvert de* $\mathbb{R}^p$ *et* $\mathbb{I}(\theta)$ *inversible* $\forall \theta \in \Theta$

ALORS l'e.m.v. $\widehat{\theta}_n$ *de* θ *est* $\sqrt{n}$*-consistant et A.N. :*

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \overset{\mathbb{P}_\theta}{\rightsquigarrow} \mathcal{N}(0, \mathbb{I}(\theta)^{-1}) \tag{49}$$

En particulier $\widehat{\theta}_n$ asymptotiquement sans biais et efficace.

Estimateur bayésiens optimaux -1-

Si cadre bayésien, comparaison de 2 estimateurs en intégrant le risque sur l'espace des paramètres. On parle de risque intégré.

Definition

Le risque moyen *a posteriori* de l'estimateur T associé à la loi *a priori* Π et à l'observation x est

$$R(\Pi, T, x) = \mathbb{E}[c(H, T(x))|X = x] = \int_{\Theta} c(\theta, T(x))L(\theta, x)d\theta \qquad (50)$$

Définit un ordre total sur tous les estimateurs ($\neq$ cas fréquentiste).

Definition

Le risque intégré de l'estimateur T associé à la loi *a priori* Π est

$$R(\Pi, T) = \mathbb{E}[c(H, T)] = \int_{\mathbb{X}} \int_{\Theta} c(\theta, T(x))L(\theta, x)d\theta d\mu(x) \qquad (51)$$

Si $\widehat{\theta}(x)$ minimise le risque moyen *a posteriori*, $\widehat{\theta}$ minimise le risque intégré.

Estimateur bayésiens optimaux -2-

Definition

Soit $\mathcal{T}$ une classe d'estimateurs de θ. Le risque de Bayes est:

$$R(\Pi) = \inf_{T \in \mathcal{T}} R(\Pi, T) \tag{52}$$

L'estimateur de Bayes $\widehat{\theta}_\Pi$ associé à la loi *a priori* Π est l'estimateur qui minimise le risque intégré :

- $R(\Pi, \widehat{\theta}) = R(\Pi)$ (53)
- $\widehat{\theta} = \text{argmin}_T R(\Pi, T)$ (54)

Estimateur bayésiens optimaux -3-

- Si coût quadratique, estimateur bayésien = moyenne *a posteriori*:

$$\widehat{\theta}(x) = \mathbb{E}[H|X = x] \tag{55}$$

En effet, si $t = T(x)$ et on veut minimiser en t

$$g(t) = R(\Pi, T|x) = \mathbb{E}[(H - t)^2|X = x] = \mathbb{E}[H^2|X = x] - 2t\mathbb{E}[H|X = x] + t^2$$

$$g'(t) = 0 \Rightarrow t = \widehat{\theta}(x).$$

- Si risque absolu, estimateur bayésien =médiane *a posteriori*:

$$\widehat{\theta}(x) \text{ vérifie } \mathbb{P}[H \leq \widehat{\theta}(x)|X = x] = 1/2 \tag{56}$$

Second principle fondamental de la statistique

 Pour terminer la leçon :

Theorem (Second principe fondamental de la statistique)

Mieux vaut un petit biais qu'une grosse variance.

Fin du chapitre 3

À retenir : ce qu'est....

- un estimateur exhaustif, minimum, libre, complet.
- l'information de Fischer et comment la calculer.
- un modèle régulier.
- une fonction de coût et un risque moyen.
- un estimateur VUMSB, efficace.
- la borne FDCR.
- un estimateur bayésien optimal.

 La notion de risque de Bayes optimal est fondamentale en Machine Learning et dans les techniques d'intelligence artificielle actuelles.