MAP 433 : Introduction aux méthodes statistiques. Cours 6

21 mars 2014

Aujourd'hui

Régression linéaire multiple ($=$ Modèle linéaire)

- La fonction de régression est $r(\vartheta, \mathbf{x}_i) = \vartheta^T \mathbf{x}_i$. On observe

$$(\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_n, Y_n)$$

avec

$$\boxed{Y_i = \vartheta^T \mathbf{x}_i + \xi_i, \quad i = 1, \ldots, n}$$

où $\vartheta \in \Theta = \mathbb{R}^k, \quad \mathbf{x}_i \in \mathbb{R}^k$.

- Matriciellement

$$\boxed{\mathbf{Y} = \mathbb{M}\vartheta + \boldsymbol{\xi}}$$

avec $\mathbf{Y} = (Y_1 \cdots Y_n)^T$, $\boldsymbol{\xi} = (\xi_1 \cdots \xi_n)^T$ et $\mathbb{M}$ la matrice $(n \times k)$ dont les lignes sont les $\mathbf{x}_i$.

EMC en régression linéaire multiple

- Estimateur des moindres carrés en régression linéaire multiple : tout estimateur $\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}$ satisfaisant

$$\sum_{i=1}^{n}\big(Y_i - (\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}})^T\,\mathbf{x}_i\,\big)^2 = \min_{\vartheta\in\mathbb{R}^k}\sum_{i=1}^{n}\big(Y_i - \vartheta^T\,\mathbf{x}_i\,\big)^2.$$

- En notation matricielle :

$$\begin{aligned}\|\mathbf{Y} - \mathbb{M}\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}\,\|^2 &= \min_{\vartheta\in\mathbb{R}^k}\|\mathbf{Y} - \mathbb{M}\,\vartheta\|^2 \\ &= \min_{v\in V}\|\mathbf{Y} - v\|^2\end{aligned}$$

où $V = \mathrm{Im}(\mathbb{M}) = \{v \in \mathbb{R}^n : v = \mathbb{M}\,\vartheta,\ \vartheta \in \mathbb{R}^k\}$.
Projection orthogonale sur V.

Géométrie de l'EMC

- L'EMC vérifie

$$\boxed{\mathbb{M}\,\widehat{\vartheta}_{\mathsf{n}}^{\,\mathtt{mc}} = P_V \mathbf{Y}}$$

  où P_V est le projecteur orthogonal sur V.

- Mais $\mathbb{M}^T P_V = \mathbb{M}^T P_V^T = (P_V \mathbb{M})^T = \mathbb{M}^T$. On en déduit les équations normales des moindres carrés :

$$\boxed{\mathbb{M}^T \mathbb{M}\,\widehat{\vartheta}_{\mathsf{n}}^{\,\mathtt{mc}} = \mathbb{M}^T \mathbf{Y}.}$$

- Remarques.
  - L'EMC est un Z-estimateur.
  - Pas d'unicité de $\widehat{\vartheta}_{\mathsf{n}}^{\,\mathtt{mc}}$ si la matrice $\mathbb{M}^T \mathbb{M}$ n'est pas inversible.

Géométrie de l'EMC

Proposition

Si $\mathbb{M}^T\,\mathbb{M}$ (matrice $k \times k$) inversible, alors $\widehat{\vartheta}_{\mathsf{n}}^{\,\mathtt{mc}}$ est unique et

$$\boxed{\widehat{\vartheta}_{\mathsf{n}}^{\,\mathtt{mc}} = \left(\mathbb{M}^T\,\mathbb{M}\right)^{-1}\mathbb{M}^T\,\mathbf{Y}}$$

- Contient la droite de régression simple.
- Résultat géometrique, non stochastique.

Cadre gaussien : loi des estimateurs

- Hyp. 1 : $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 \mathrm{Id}_n)$.
- Hyp. 2 : $\mathbb{M}^T \mathbb{M} > 0$.

Proposition

(i) $\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}} \sim \mathcal{N}\left(\vartheta, \sigma^2 \left(\mathbb{M}^T \mathbb{M}\right)^{-1}\right)$

(ii) $\|\mathbf{Y} - \mathbb{M}\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}\|^2 \sim \sigma^2 \chi^2(n-k)$ *loi du Chi 2 à $n-k$ degrés de liberté*

(iii) $\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}$ *et* $\mathbf{Y} - \mathbb{M}\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}$ *sont indépendants.*

- Preuve : Thm. de Cochran (Poly, page 18). Si $\boldsymbol{\xi} \sim \mathcal{N}(0, \mathrm{Id}_n)$ et A_j matrices $n \times n$ projecteurs t.q. $A_j A_i = 0$ pour $i \neq j$, alors : $A_j\,\boldsymbol{\xi} \sim \mathcal{N}\left(0, A_j\right)$, indépendants, $\|A_j \boldsymbol{\xi}\|^2 \sim \chi^2(\mathrm{Rang}(A_j))$.

Propriétés de l'EMC : cadre gaussien

Estimateur de la variance σ^2 :

$$\boxed{\widehat{\sigma}_n^2 = \frac{\|\mathbf{Y} - \mathbb{M}\,\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}\|^2}{n-k} = \frac{1}{n-k}\sum_{i=1}^n \left(Y_i - \left(\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}\right)^T \mathbf{x}_i\right)^2}$$

D'après la dernière Proposition :

- $\widehat{\sigma}_n^2/\sigma^2 \sim \chi^2(n-k)$ loi du Chi 2 à $n-k$ degrés de liberté
- C'est un estimateur sans biais :

$$\mathbb{E}_\vartheta\left[\widehat{\sigma}_n^2\right] = \sigma^2.$$

- $\widehat{\sigma}_n^2$ est indépendant de $\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}$.

Propriétés de l'EMC : cadre gaussien

- Lois des coordonnées de $\widehat{\vartheta}_n^{\,\mathtt{mc}}$:

$$(\widehat{\vartheta}_n^{\,\mathtt{mc}})_j - \vartheta_j \sim \mathcal{N}(0, \sigma^2 b_j)$$

où b_j est le $j$ème élément diagonal de $\left(\mathbb{M}^T\mathbb{M}\right)^{-1}$.

$$\frac{(\widehat{\vartheta}_n^{\,\mathtt{mc}})_j - \vartheta_j}{\widehat{\sigma}_n\sqrt{b_j}} \sim t_{n-k}$$

loi de Student à $n-k$ degrés de liberté.

$$t_q = \frac{\xi}{\sqrt{\eta/q}}$$

où $q \geq 1$ un entier, $\xi \sim \mathcal{N}(0,1)$, $\eta \sim \chi^2(q)$ et ξ indépendant de η.

Exemple de données de régression

Données de diabète

Patient	age	sex	bmi	map	tc	ldl	hdl	tch	ltg	glu	Response
1	59	2	32.1	101	157	93.2	38	4	4.9	87	151
2	48	1	21.6	87	183	103.2	70	3	3.9	69	75
3	72	2	30.5	93	156	93.6	41	4	4.7	85	141
4	24	1	25.3	84	198	131.4	40	5	4.9	89	206
5	50	1	23.0	101	192	125.4	52	4	4.3	80	135
6	23	1	22.6	89	139	64.8	61	2	4.2	68	97
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
441	36	1	30.0	95	201	125.2	42	5	5.1	82	220
442	36	1	19.6	71	250	132.2	97	3	4.6	92	57

n=442, k=10

bmi = Body Mass Index

map = Blood Pressure

tc, ldl, tch, ltg, glu = Blood Serum Measurements

Response Y = a quantitative measure of disease progression 1 year after baseline

Résultats de traitement statistique initial

	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
(Intercept)	152.133	2.576	59.061	$< 2e-16 ***$
age	-10.012	59.749	-0.168	0.867000
sex	-239.819	61.222	-3.917	$0.000104 ***$
bmi	519.840	66.534	7.813	$4.30e-14 ***$
map	324.390	65.422	4.958	$1.02e-06 ***$
tc	-792.184	416.684	-1.901	0.057947
ldl	476.746	339.035	1.406	0.160389
hdl	101.045	212.533	0.475	0.634721
tch	177.064	161.476	1.097	0.273456
ltg	751.279	171.902	4.370	$1.56e-05 ***$
glu	67.625	65.984	1.025	0.305998

Questions statistiques

- **Sélection de variables.** Lesquelles parmi les 10 variables :

  `age,sex,bmi,map,tc,ldl,hdl,tch,ltg,glu`

  sont significatives ? Formalisation mathématique : trouver (estimer) l'ensemble $N = \{j : \vartheta_j \neq 0\}$.
- **Prévison.** Un nouveau patient arrive avec son vecteur des 10 variables $\mathbf{x}_0 \in \mathbb{R}^{10}$. Donner la prévison de la réponse Y $=$état du patient dans 1 an.

RSS (Residual Sum of Squares)

Modèle de régression

$$Y_i = r(\vartheta, \mathbf{x}_i) + \xi_i, \quad i = 1, \ldots, n.$$

- **Résidu :** si $\widehat{\vartheta}_n$ est un estimateur de ϑ,

$$\widehat{\xi}_i = Y_i - r(\widehat{\vartheta}_n, \mathbf{x}_i) \text{ résidu au point } i.$$

- **RSS :** Residual Sum of Squares, somme résiduelle des carrés. Caractérise la qualité d'approximation.

$$\mathrm{RSS}(= \mathrm{RSS}_{\widehat{\vartheta}_n}) = \|\widehat{\xi}\|^2 = \sum_{i=1}^{n} \big(Y_i - r(\widehat{\vartheta}_n, \mathbf{x}_i)\big)^2.$$

- En régression linéaire : $\boxed{\mathrm{RSS} = \|\mathbf{Y} - \mathbb{M}\,\widehat{\vartheta}_n\|^2.}$

Sélection de variables : Backward Stepwise Regression

- On se donne un critère d'élimination de variables (plusieurs choix de critère possibles...).
- On élimine une variable, la moins significative du point de vue du critère choisi.
- On calcule l'EMC $\widehat{\vartheta}^{\mathrm{mc}}_{n,k-1}$ dans le nouveau modèle, avec seulement les $k-1$ paramétres restants, ainsi que le RSS :
$$\mathrm{RSS}_{k-1} = \|\mathbf{Y} - \mathbb{M}\,\widehat{\vartheta}^{\mathrm{mc}}_{n,k-1}\|^2.$$
- On continue à éliminer des variables, une par une, jusqu'à la stabilisation de RSS : $\mathrm{RSS}_m \approx \mathrm{RSS}_{m-1}$.

Données de diabète : Backward Regression

- **Sélection "naïve"** : $\{\texttt{sex,bmi,map,ltg}\}$
- **Sélection par Backward Regression** :
  Critère d'élimination : plus grande valeur de $\Pr(>|t|)$.

	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
(Intercept)	152.133	2.576	59.061	$< 2e-16$ ***
age	−10.012	59.749	−0.168	0.867000
sex	−239.819	61.222	−3.917	0.000104 ***
bmi	519.840	66.534	7.813	$4.30e-14$ ***
map	324.390	65.422	4.958	$1.02e-06$ ***
tc	−792.184	416.684	−1.901	0.057947
ldl	476.746	339.035	1.406	0.160389
hdl	101.045	212.533	0.475	0.634721
tch	177.064	161.476	1.097	0.273456
ltg	751.279	171.902	4.370	$1.56e-05$ ***
glu	67.625	65.984	1.025	0.305998

Données de diabète : Backward Regression

Backward Regression : Itération 2.

Critère d'élimination : plus grande valeur de $\Pr(>|t|)$.

	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
(Intercept)	152.133	2.573	59.128	$< 2e-16$
sex	-240.835	60.853	-3.958	0.000104
bmi	519.905	64.156	5.024	$8.85e-05$
map	322.306	65.422	4.958	$7.43e-07$
tc	-790.896	416.144	-1.901	0.058
ldl	474.377	338.358	1.402	0.162
hdl	99.718	212.146	0.470	0.639
tch	177.458	161.277	1.100	0.272
ltg	749.506	171.383	4.373	$1.54e-05$
glu	67.170	65.336	1.013	0.312

Données de diabète : Backward Regression

Backward Regression : Itération 5 (dernière).

Variables sélectionnées :

{sex,bmi,map,tc,ldl,ltg}

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	152.133	2.572	59.159	$< 2e-16$
sex	−226.511	59.857	−3.784	0.000176
bmi	529.873	65.620	8.075	$6.69e-15$
map	327.220	62.693	5.219	$2.79e-07$
tc	−757.938	160.435	−4.724	$3.12e-06$
ldl	538.586	146.738	3.670	0.000272
ltg	804.192	80.173	10.031	$< 2e-16$

Discussion de `Backward Regression` :

- Méthode de sélection purement empirique, pas de justification théorique.
- Application d'autres critères d'élimination en `Backward Regression` peut amener aux résultats différents. Exemple. Critère C_p de Mallows–Akaike : on élimine la variable j qui réalise

$$\min_j \Big(\mathrm{RSS}_{m,(-j)} + 2\widehat{\sigma}_n^2 m \Big).$$

Sélection de variables : LASSO

LASSO = Least Absolute Shrinkage and Selection Operator

- Estimateur LASSO : tout estimateur $\widehat{\vartheta}_n^L$ vérifiant

$$\widehat{\vartheta}_n^L \in \arg\min_{\vartheta \in \mathbb{R}^k} \left(\sum_{i=1}^n \left(Y_i - \vartheta^T \mathbf{x}_i \right)^2 + \lambda \sum_{j=1}^k |\vartheta_j| \right) \quad \text{avec } \lambda > 0.$$

- Si $\mathbb{M}^T \mathbb{M} > 0$, l'estimateur LASSO $\widehat{\vartheta}_n^L$ est unique.
- Estimateur des moindres carrés pénalisé. Pénalisation par $\sum_{j=1}^k |\vartheta_j|$, la norme ℓ_1 de ϑ.

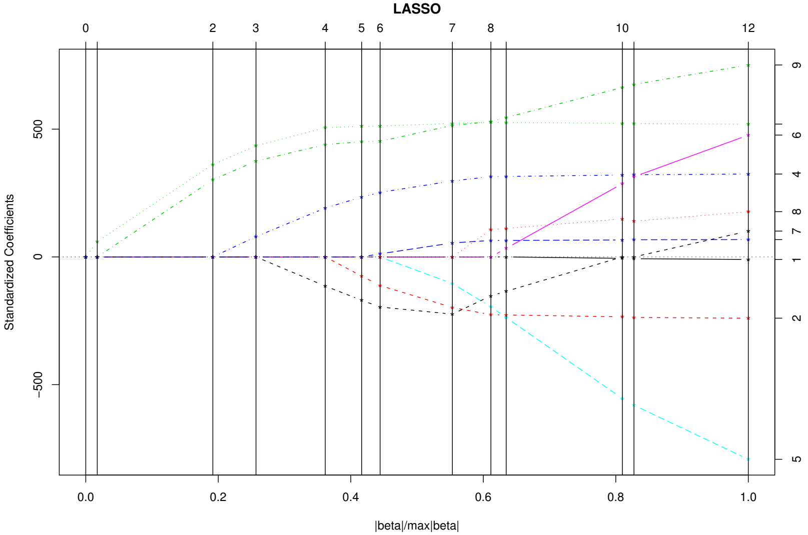
Sélection de variables : LASSO

- Deux utilisations de LASSO :
  - Estimation de ϑ : alternative à $\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}$ si $k > n$.
  - Sélection de variables : on ne retient que les variables qui correspondent aux coordonnées non-nulles du vecteur $\widehat{\vartheta}_n^L$.
- LASSO admet une justification théorique : sous certaines hypothèses sur la matrice $\mathbb{M}$,

$$\lim_{n\to\infty} \mathbb{P}\{\widehat{N}_n = N\} = 1,$$

où $N = \{j : \vartheta_j \neq 0\}$ et $\widehat{N}_n = \{j : \widehat{\vartheta}_{n,j}^L \neq 0\}$.

Application de LASSO : "regularization path"



Données de diabète : LASSO

Application aux données de diabète.

- L'ensemble de variables sélectionné par LASSO :

$$\{\texttt{sex}, \texttt{bmi}, \texttt{map}, \texttt{tc}, \texttt{hdl}, \texttt{ltg}, \texttt{glu}\}$$

- `Backward Regression` :

$$\{\texttt{sex}, \texttt{bmi}, \texttt{map}, \texttt{tc}, \texttt{ldl}, \texttt{ltg}\}$$

- Sélection naïve :

$$\{\texttt{sex}, \texttt{bmi}, \texttt{map}, \texttt{tc}\}$$

Limites des moindres carrés et du cadre gaussien

- Calcul explicite (et efficace) de l'EMC limité à une fonction de régression linéaire.
- Modèle linéaire donne un cadre assez général :
  - Modèle polynomial,
  - Modèles avec interactions...
- Hypothèse de gaussianité = cadre asymptotique implicite.
- Besoin d'outils pour les modèles à réponse Y discrète.

Régression linéaire non-gaussienne

Modèle de régression linéaire

$$Y_i = \vartheta^T \mathbf{x}_i + \xi_i, \quad i = 1, \ldots, n.$$

- Hyp. 1' : ξ_i i.i.d., $\mathbb{E}[\xi_i] = 0$, $\mathbb{E}[\xi_i^2] = \sigma^2 > 0$.
- Hyp. 2' : $\mathbb{M}^T\mathbb{M} > 0$, $\lim_n \max_{1 \leq i \leq n} \mathbf{x}_i^T \left(\mathbb{M}^T\mathbb{M}\right)^{-1} \mathbf{x}_i = 0$.

Proposition (Normalité asymptotique de l'EMC)

$$\sigma^{-1}\left(\mathbb{M}^T\mathbb{M}\right)^{1/2}(\widehat{\vartheta}_n^{\,\mathrm{mc}} - \vartheta) \xrightarrow{d} \mathcal{N}\big(0, \mathrm{Id}_k\big), \quad n \to \infty.$$

- A comparer avec le cadre gaussien :

$$\sigma^{-1}\left(\mathbb{M}^T\mathbb{M}\right)^{1/2}(\widehat{\vartheta}_n^{\,\mathrm{mc}} - \vartheta) \sim \mathcal{N}\big(0, \mathrm{Id}_k\big) \text{ pour tout } n.$$

Régression non-linéaire

- On observe
$$(\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_n, Y_n),$$
où
$$\boxed{Y_i = r(\vartheta, \mathbf{x}_i) + \xi_i, \quad i = 1, \ldots, n}$$
avec
$$\mathbf{x}_i \in \mathbb{R}^k, \quad \text{et} \quad \vartheta \in \Theta \subset \mathbb{R}^d.$$
- Si $\xi_i \sim_{\text{i.i.d.}} \mathcal{N}(0, \sigma^2)$,
$$\mathcal{L}_n(\vartheta, Y_1, \ldots, Y_n) \propto \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^n \big(Y_i - r(\vartheta, \mathbf{x}_i)\big)^2\Big)$$
et l'estimateur du maximum de vraisemblance est obtenu en minimisant la fonction
$$\vartheta \rightsquigarrow \sum_{i=1}^n \big(Y_i - r(\vartheta, \mathbf{x}_i)\big)^2.$$

Moindre carrés non-linéaires

Définition

- *M-estimateur associé à la fonction de contraste* $\psi : \Theta \times \mathbb{R}^k \times \mathbb{R} \to \mathbb{R}$ *: tout estimateur* $\widehat{\vartheta}_n$ *satisfaisant*

$$\sum_{i=1}^n \psi(\widehat{\vartheta}_n, \mathbf{x}_i, Y_i) = \max_{a \in \Theta} \sum_{i=1}^n \psi(a, \mathbf{x}_i, Y_i).$$

- *Estimateur des moindres carrés non-linéaires : associé au contraste* $\psi(a, \mathbf{x}, y) = -\big(y - r(a, \mathbf{x})\big)^2$.

- Extension des résultats en densité $\to$ théorèmes limites pour des sommes de v.a. indépendantes non-équidistribuées.

Modèle à réponse binaire

- On observe

$$(\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_n, Y_n), \quad Y_i \in \{0, 1\}, \; \mathbf{x}_i \in \mathbb{R}^k \, .$$

- Modélisation via la fonction de régression

$$\mathbf{x} \rightsquigarrow p_{\mathbf{x}}(\vartheta) = \mathbb{E}_\vartheta\left[Y | \mathbf{X} = \mathbf{x}\right] = \mathbb{P}_\vartheta\left[Y = 1 | \mathbf{X} = \mathbf{x}\right]$$

- Représentation

$$\begin{aligned} Y_i &= p_{\mathbf{x}_i}(\vartheta) + \left(Y_i - p_{\mathbf{x}_i}(\vartheta)\right) \\ &= r(\vartheta, \mathbf{x}_i) + \xi_i \end{aligned}$$

  avec $r(\vartheta, \mathbf{x}_i) = p_{\mathbf{x}_i}(\vartheta)$ et $\xi_i = Y_i - p_{\mathbf{x}_i}(\vartheta)$.
- $\mathbb{E}_\vartheta\left[\xi_i\right] = 0$ mais structure des ξ_i compliquée (dépendance en ϑ).

Modèle à réponse discrète

- Y_i v.a. de Bernoulli de paramètre $p_{\mathbf{x}_i}(\vartheta)$.
  Vraisemblance

$$\mathcal{L}_n(\vartheta, Y_1, \ldots, Y_n) = \prod_{i=1}^{n} p_{\mathbf{x}_i}(\vartheta)^{Y_i}\big(1 - p_{\mathbf{x}_i}(\vartheta)\big)^{1-Y_i}$$

  $\rightarrow$ méthodes de résolution numérique.

- Régression logistique (très utile dans les applications)

$$p_{\mathbf{x}}(\vartheta) = \psi(\mathbf{x}^T\,\vartheta),$$

$$\psi(t) = \frac{e^t}{1+e^t},\ \ t \in \mathbb{R}\ \ \text{fonction logistique.}$$

Régression logistique et modèles latents

- Représentation équivalente de la régression logistique : on observe

$$\boxed{Y_i = 1_{\{Y_i^\star > 0\}}, \quad i = 1, \ldots, n}$$

(les $\mathbf{x}_i$ sont donnés), et $Y_i^\star$ est une variable latente ou cachée,

$$\boxed{Y_i^\star = \vartheta^T \mathbf{x}_i + U_i, \quad i = 1, \ldots, n}$$

avec $U_i \sim_{\text{i.i.d.}} F$, où

$$F(t) = \frac{1}{1+e^{-t}}, \quad t \in \mathbb{R}.$$

-

$$\begin{aligned} \mathbb{P}_\vartheta\left[Y_i^\star > 0\right] &= \mathbb{P}_\vartheta\left[\mathbf{x}_i^T\,\vartheta + U_i > 0\right] \\ &= 1 - \mathbb{P}_\vartheta\left[U_i \leq -\mathbf{x}_i^T\,\vartheta\right] \\ &= 1 - \left(1 + \exp(-\mathbf{x}_i^T\,\vartheta)\right)^{-1} = \psi(\mathbf{x}_i^T\,\vartheta). \end{aligned}$$

Bilan provisoire : modèles paramétriques dominés

- Modèle de densité : on observe

$$X_1, \ldots, X_n \sim_{\text{i.i.d.}} \mathbb{P}_\vartheta, \quad \vartheta \in \Theta \subset \mathbb{R}^d .$$

Estimateurs : moments, Z- et M-estimateurs, EMV.

- Modèle de régression : on observe

$$Y_i = r(\vartheta, \mathbf{x}_i) + \xi_i, \quad i = 1, \ldots, n, \quad \xi_i \text{ i.i.d.}, \ \vartheta \in \Theta \subset \mathbb{R}^d .$$

Estimateurs :

  - Si $r(\vartheta, \mathbf{x}) = \mathbf{x}\,\vartheta^T$, EMC (coïncide avec l'EMV si les ξ_i gaussiens)
  - Sinon, M-estimateurs, EMV...
  - Autres méthodes selon des hypothèses sur le « design »...

$\widehat{\vartheta}_n$ estimateur de ϑ : précision, qualité de $\widehat{\vartheta}_n$? En pratique, on a « souvent »

- une information non-asymptotique de type

$$\mathbb{E}\left[\| \widehat{\vartheta}_n - \vartheta\|^2\right] \leq c_n(\vartheta)^2,$$

- ou bien asymptotique de type

$$v_n(\widehat{\vartheta}_n - \vartheta) \xrightarrow{d} Z_\vartheta, \quad v_n \to \infty.$$

Permet « souvent » de construire un(e) région-intervalle de confiance...

Region-intervalle de confiance : définition formelle

$\{\mathbb{P}_\vartheta^n, \vartheta \in \Theta\}$, $\Theta \subset \mathbb{R}^d$, engendrée par l'observation $Z^{(n)}$.

- Densité : $Z^{(n)} = (X_1, \ldots, X_n)$, $\mathbb{P}_\vartheta^n = \mathbb{P}_\vartheta \otimes \ldots \otimes \mathbb{P}_\vartheta$
- Régression (à design déterministe) : $Z^{(n)} = (Y_1, \ldots, Y_n)$, $\mathbb{P}_\vartheta^n = \mathbb{P}_{\vartheta,\mathbf{x}_1} \otimes \ldots \otimes \mathbb{P}_{\vartheta,\mathbf{x}_n}$, où $\mathbb{P}_{\vartheta,\mathbf{x}_i}$ loi de $Y_i = r(\vartheta, \mathbf{x}_i) + \xi_i$.

Définition

Région de confiance de niveau $1-\alpha$, $\alpha \in (0,1)$, (resp. asymptotiquement de niveau α) : sous-ensemble observable $\mathcal{C}_{n,\alpha}(Z^{(n)})$ de $\mathbb{R}^d$ t.q.

$$\forall \vartheta \in \Theta : \mathbb{P}_\vartheta^n \left[\vartheta \in \mathcal{C}_{n,\alpha}(Z^{(n)})\right] \geq 1 - \alpha$$

resp.

$$\forall \vartheta \in \Theta : \liminf_{n\to\infty} \mathbb{P}_\vartheta^n \left[\vartheta \in \mathcal{C}_{n,\alpha}(Z^{(n)})\right] \geq 1 - \alpha.$$

Comparaison d'estimateurs

Etant donné $\{\mathbb{P}_\vartheta^n, \vartheta \in \Theta\}$ comment construire le meilleur estimateur ? Dans quel sens ?

- Intuitivement : $\widehat{\vartheta}_n$ fournit une précision optimale si on peut lui associer une région de confiance de longueur (moyenne) minimale.
- Différence entre point de vue asymptotique et non-asymptotique.
- Dans ce cours, nous étudions les deux points de vue sous un angle –un peu réducteur– particulier :
  - Non-asymptotique : contrôle du risque quadratique
  - Asymptotique : comparaison des estimateurs asymptotiquement normaux.

Risque quadratique, admissibilité

Situation : $\widehat{\vartheta}_{n,i} = \widehat{\vartheta}_{n,i}(Z^{(n)})$, $i = 1, 2$ deux estimateurs basés sur l'observation $Z^{(n)}$ qui engendre l'expérience $\{\mathbb{P}^n_\vartheta, \vartheta \in \Theta\}$, $\Theta \subset \mathbb{R}^1$.

Définition

Risque quadratique de l'estimateur $\widehat{\vartheta}_n$ au point $\vartheta \in \Theta$:

$$\mathcal{R}(\widehat{\vartheta}_n, \vartheta) = \mathbb{E}^n_\vartheta \left[\left(\widehat{\vartheta}_n - \vartheta \right)^2 \right].$$

Définition

L'estimateur $\widehat{\vartheta}_{n,1}$ est préférable – au sens du risque quadratique – à l'estimateur $\widehat{\vartheta}_{n,2}$ si

$$\forall \vartheta \in \Theta, \;\; \mathcal{R}\big(\widehat{\vartheta}_{n,1}, \vartheta\big) \leq \mathcal{R}\big(\widehat{\vartheta}_{n,2}, \vartheta\big).$$

Absence d'optimalité

- Existe-t-il un estimateur optimal $\vartheta_n^\star$ au sens où

$$\forall \vartheta \in \Theta, \ \ \mathcal{R}\big(\vartheta_n^\star, \vartheta\big) \leq \inf_{\widehat{\vartheta}_n} \mathcal{R}\big(\widehat{\vartheta}_n, \vartheta\big) \ ?$$

- Si $\Theta = \{\vartheta_1, \vartheta_2\}$ et s'il n'existe pas d'événement observable A tel que, simultanément :

$$\mathbb{P}_{\vartheta_1}^n\big[A\big] = 0 \ \text{ et } \ \mathbb{P}_{\vartheta_2}^n\big[A\big] = 1,$$

(on dit que $\mathbb{P}_{\vartheta_1}^n$ et $\mathbb{P}_{\vartheta_2}^n$ ne sont pas étrangères), alors il n'existe pas d'estimateur optimal.

- Condition suffisante pour que $\mathbb{P}_{\vartheta_1}^n$ et $\mathbb{P}_{\vartheta_2}^n$ ne soient pas étrangères : $\mathbb{P}_{\vartheta_1}^n \ll \mathbb{P}_{\vartheta_2}^n$ et $\mathbb{P}_{\vartheta_2}^n \ll \mathbb{P}_{\vartheta_1}^n$.

Absence d'optimalité (cont.)

- Preuve : Pour tout estimateur $\vartheta_n^\star$, on a

$$\max\left\{\mathcal{R}(\vartheta_n^\star,\vartheta_1),\mathcal{R}(\vartheta_n^\star,\vartheta_2)\right\}>0 \qquad (\star).$$

- Supposons $\vartheta_n^\star$ estimateur optimal et $\mathcal{R}(\vartheta_n^\star,\vartheta_1)>0$. Alors $\widehat{\vartheta}_n^{\,\text{trivial}}:=\vartheta_1$ vérifie

$$0=\mathcal{R}\big(\widehat{\vartheta}_n^{\,\text{trivial}},\vartheta_1\big)<\mathcal{R}\big(\vartheta_n^\star,\vartheta_1\big) \quad \text{contradiction !}$$

et contredit l'optimalité de $\vartheta_n^\star$.

Absence d'optimalité (fin.)

- Preuve de $(\star)$: si $\mathcal{R}\big(\vartheta_n^\star, \vartheta_1\big) = \mathcal{R}\big(\vartheta_n^\star, \vartheta_2\big) = 0$, alors

$$\vartheta_n^\star = \vartheta_1 \;\; \mathbb{P}_{\vartheta_1}^n - \text{p.s.} \quad \text{et} \quad \vartheta_n^\star = \vartheta_2 \;\; \mathbb{P}_{\vartheta_2}^n - \text{p.s.}.$$

  Soient $A = \{\omega, \vartheta_n^\star(\omega) = \vartheta_1\}$ et $B = \{\omega, \vartheta_n^\star(\omega) = \vartheta_2\}$. Alors $\mathbb{P}_{\vartheta_1}^n[A] = 1$ et donc $\mathbb{P}_{\vartheta_2}^n[A] > 0$. Aussi, $\mathbb{P}_{\vartheta_2}^n[B] = 1$. Donc $A \cap B \neq \emptyset$. Il existe ω_0 tel que $\vartheta_1 = \vartheta_n^\star(\omega_0) = \vartheta_2$ contradiction !

- Attention ! La propriété $\mathbb{P}_{\vartheta_1}^n$ et $\mathbb{P}_{\vartheta_2}^n$ non étrangères est minimale. Mais elle disparaît en général lorsque $n \to \infty$.

Notions d'optimalité

- Différentes notions existent. Deux exemples extrêmes :

Définition (Admissibilité et critère minimax)

- *Un estimateur $\vartheta_n^\star$ est admissible s'il n'existe pas d'estimateur $\widehat{\vartheta}_n$ préférable à $\vartheta_n^\star$ tel que, pour un point $\vartheta_0 \in \Theta$*
$$\mathcal{R}(\widehat{\vartheta}_n, \vartheta_0) < \mathcal{R}(\vartheta_n^\star, \vartheta_0).$$
- *Un estimateur $\vartheta_n^\star$ est minimax si*
$$\sup_{\vartheta \in \Theta} \mathcal{R}(\vartheta_n^\star, \vartheta) = \inf_{\widehat{\vartheta}_n} \sup_{\vartheta \in \Theta} \mathcal{R}(\widehat{\vartheta}_n, \vartheta).$$

- Admissibilité : permet d'éliminer des estimateurs absurdes (mais pas tous).
- Minimaxité : notion très robuste mais conservatrice, à suivre...

Approche asymptotique

- Hypothèse simplificatrice : $\vartheta \in \Theta \subset \mathbb{R}$. On se restreint aux estimateurs asymptotiquement normaux c'est-à-dire vérifiant
$$\sqrt{n}\big(\widehat{\vartheta}_n - \vartheta\big) \xrightarrow{d} \mathcal{N}\big(0, v(\vartheta)\big)$$
cf. théorèmes limites obtenus pour les Z-,M-estimateurs.
- Si $\widehat{\vartheta}_{n,1}$ et $\widehat{\vartheta}_{n,2}$ as. normaux de variance asymptotique $v_1(\vartheta) \leq v_2(\vartheta)$, alors la précision de $\widehat{\vartheta}_{n,1}$ est asymptotiquement meilleure que celle de $\widehat{\vartheta}_{n,2}$ au point ϑ :
$$\widehat{\vartheta}_{n,1} = \vartheta + \sqrt{\frac{v_1(\vartheta)}{n}}\xi^{(n)}$$
$$\widehat{\vartheta}_{n,2} = \vartheta + \sqrt{\frac{v_2(\vartheta)}{n}}\zeta^{(n)}$$
où $\xi^{(n)}$ et $\zeta^{(n)} \xrightarrow{d} \mathcal{N}(0,1)$.

Comparaison d'estimateurs : cas asymptotique

- Si $v_1(\vartheta) < v_2(\vartheta)$, et si $\vartheta \rightsquigarrow v_i(\vartheta)$ est continue, on pose

$$\mathcal{C}_{n,\alpha}(\widehat{\vartheta}_{n,i}) = \left[\widehat{\vartheta}_{n,i} \pm \sqrt{\frac{v_i(\widehat{\vartheta}_{n,i})}{n}}\Phi^{-1}(1-\alpha/2)\right], \quad i = 1, 2$$

où $\alpha \in (0,1)$ et $\Phi(\cdot)$ est la fonction de répartition de la loi normale standard.

- $\mathcal{C}_{n,\alpha}(\widehat{\vartheta}_{n,i})$, $i = 1, 2$ sont deux intervalles de confiance asymptotiquement de niveau $1-\alpha$ et on a

$$\frac{|\mathcal{C}_{n,\alpha}(\widehat{\vartheta}_{n,1})|}{|\mathcal{C}_{n,\alpha}(\widehat{\vartheta}_{n,2})|} \xrightarrow{\mathbb{P}_\vartheta^n} \sqrt{\frac{v_1(\vartheta)}{v_2(\vartheta)}} < 1.$$

- La notion de longueur minimale possible d'un intervalle de confiance est en général difficile à manipuler.

Conclusion provisoire

- Il est difficile en général de comparer des estimateurs.
- Cadre asymptotique + normalité asymptotique $\rightarrow$ comparaison de la variance asymptotique $\vartheta \rightsquigarrow v(\vartheta)$.
- Sous des hypothèses de régularité du modèle $\{\mathbb{P}_\vartheta^n, \vartheta \in \Theta\}$ alors
  - Il existe une variance asymptotique $v^\star(\vartheta)$ minimale parmi les variances de la classe des M-estimateurs as. normaux.
  - Cette fonction est associée à une quantité d'information intrinsèque au modèle.
  - La variance asymptotique de l'EMV est $v^\star(\vartheta)$.
- Ceci règle partiellement le problème de l'optimalité.

Régularité d'un modèle statistique et information

- Cadre simplificateur : modèle de densité

$$X_1, \ldots, X_n \ \text{ i.i.d. de loi } \ \mathbb{P}_\vartheta$$

dans la famille $\{\mathbb{P}_\vartheta, \vartheta \in \Theta\}$ avec $\Theta \subset \mathbb{R}$ pour simplifier.

- Notation :

$$f(\vartheta, x) = \frac{d\,\mathbb{P}_\vartheta}{d\mu}(x), \quad x \in \mathbb{R}, \vartheta \in \Theta.$$

- Hypothèse : la quantité

$$\boxed{\mathbb{I}(\vartheta) = \mathbb{E}_\vartheta\left[\left(\partial_\vartheta \log f(\vartheta, X)\right)^2\right]}$$

est bien définie.

Information de Fisher

Définition

- $\mathbb{I}(\vartheta) = \mathbb{E}_\vartheta\left[\left(\partial_\vartheta \log f(\vartheta, X)\right)^2\right]$ *s'appelle l'information de Fisher de la famille* $\{\mathbb{P}_\vartheta, \vartheta \in \Theta\}$ *au point* ϑ*. Elle ne dépend pas de la mesure dominante* μ*.*
- *Le cadre d'intérêt est celui où*

$$0 < \mathbb{I}(\vartheta) < +\infty.$$

- $\mathbb{I}(\vartheta)$ *quantifie « l'information » qu'apporte chaque observation* X_i *sur le paramètre* ϑ*.*

Remarque : on a $\mathbb{P}_\vartheta\left[f(\vartheta, X) > 0\right] = 1$, donc la quantité $\log f(\vartheta, X)$ est bien définie.

Information dans quel sens ? Origine de la notion

- Supposons l'EMV $\widehat{\vartheta}_n^{\,\mathtt{mv}}$ bien défini et convergent.
- Supposons l'application $(\vartheta, x) \rightsquigarrow f(\vartheta, x)$ possédant toutes les propriétés de régularité et d'intégrabilité voulues.
- Alors

$$\boxed{\sqrt{n}\big(\widehat{\vartheta}_n^{\,\mathtt{mv}} - \vartheta\big) \xrightarrow{d} \mathcal{N}\Big(0, \frac{1}{\mathbb{I}(\vartheta)}\Big)}$$

en loi sous $\mathbb{P}_\vartheta$, où encore

$$\widehat{\vartheta}_n^{\,\mathtt{mv}} \overset{d}{\approx} \vartheta + \frac{1}{\sqrt{n\mathbb{I}(\vartheta)}}\mathcal{N}(0,1)$$

en loi sous $\mathbb{P}_\vartheta$.

Construction de l'information + jeu d'hypothèses attenant

- Heuristique : on établira un jeu d'hypothèses justifiant a posteriori le raisonnement.
- Etape 1 : l'EMV $\widehat{\vartheta}_n^{\,\mathtt{mv}}$ converge :

$$\widehat{\vartheta}_n^{\,\mathtt{mv}} \xrightarrow{\mathbb{P}_\vartheta} \vartheta$$

  via le théorème de convergence des M-estimateurs.
- Etape 2 : l'EMV $\widehat{\vartheta}_n^{\,\mathtt{mv}}$ est un Z-estimateur :

$$0 = \partial_\vartheta \Big(\sum_{i=1}^n \log f(\vartheta, X_i) \Big)_{\vartheta = \widehat{\vartheta}_n^{\,\mathtt{mv}}}.$$

Construction de $\mathbb{I}(\vartheta)$ cont.

- Etape 3 : développement asymptotique autour de ϑ :

$$0 \approx \sum_{i=1}^{n} \partial_\vartheta \log f(\vartheta, X_i) + (\widehat{\vartheta}_n^{\,\mathtt{mv}} - \vartheta) \sum_{i=1}^{n} \partial_\vartheta^2 \log f(\vartheta, X_i),$$

soit

$$\widehat{\vartheta}_n^{\,\mathtt{mv}} - \vartheta \approx - \frac{\sum_{i=1}^{n} \partial_\vartheta \log f(\vartheta, X_i)}{\sum_{i=1}^{n} \partial_\vartheta^2 \log f(\vartheta, X_i)}$$

- Etape 4 : le numérateur. Normalisation et convergence de $\sum_{i=1}^{n} \partial_\vartheta \log f(\vartheta, X_i)$?

Numérateur

Lemme

On a

$$\mathbb{E}_\vartheta\left[\partial_\vartheta \log f(\vartheta, X)\right] = 0.$$

Preuve.

$$\begin{aligned}
\mathbb{E}_\vartheta\left[\partial_\vartheta \log f(\vartheta, X)\right] &= \int_{\mathbb{R}} \partial_\vartheta \log f(\vartheta, x) f(\vartheta, x)\mu(dx) \\
&= \int_{\mathbb{R}} \frac{\partial_\vartheta f(\vartheta, x)}{f(\vartheta, x)} f(\vartheta, x)\mu(dx) \\
&= \int_{\mathbb{R}} \partial_\vartheta f(\vartheta, x)\mu(dx) \\
&= \partial_\vartheta \int_{\mathbb{R}} f(\vartheta, x)\mu(dx) = \partial_\vartheta 1 = 0.
\end{aligned}$$

Dénominateur

De même $\int_{\mathbb{R}} \partial_\vartheta^2 f(\vartheta, x)\mu(dx) = 0$. Conséquence :

$$\boxed{\mathbb{I}(\vartheta) = \mathbb{E}_\vartheta\left[\left(\partial_\vartheta \log f(\vartheta, X)\right)^2\right] = -\mathbb{E}_\vartheta\left[\partial_\vartheta^2 \log f(\vartheta, X)\right]}$$

En effet

$$\mathbb{E}_\vartheta\left[\partial_\vartheta^2 \log f(\vartheta, X)\right]$$

$$= \int_{\mathbb{R}} \frac{\partial_\vartheta^2 f(\vartheta, x) f(\vartheta, x) - \left(\partial_\vartheta f(\vartheta, x)\right)^2}{f(\vartheta, x)^2} f(\vartheta, x)\mu(dx)$$

$$= \int_{\mathbb{R}} \partial_\vartheta^2 f(\vartheta, x)\mu(dx) - \int_{\mathbb{R}} \frac{\left(\partial_\vartheta f(\vartheta, x)\right)^2}{f(\vartheta, x)}\mu(dx)$$

$$= 0 - \int_{\mathbb{R}} \left(\frac{\partial_\vartheta f(\vartheta, x)}{f(\vartheta, x)}\right)^2 f(\vartheta, x)\mu(dx) = -\mathbb{E}\left[\left(\partial_\vartheta \log f(\vartheta, X)\right)^2\right].$$

Conséquences

- Les $\partial_\vartheta \log f(\vartheta, X_i)$ sont i.i.d. et $\mathbb{E}_\vartheta\left[\partial_\vartheta \log f(\vartheta, X)\right] = 0$. TCL :

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \partial_\vartheta \log f(\vartheta, X_i) \xrightarrow{d} \mathcal{N}\Big(0, \mathbb{E}_\vartheta\left[\left(\partial_\vartheta \log f(\vartheta, X)\right)^2\right]\Big)$$

$$= \mathcal{N}\big(0, \mathbb{I}(\vartheta)\big).$$

- Les $\partial_\vartheta^2 \log f(\vartheta, X_i)$ sont i.i.d. LGN :

$$\frac{1}{n}\sum_{i=1}^{n} \partial_\vartheta^2 \log f(\vartheta, X_i) \xrightarrow{\mathbb{P}_\vartheta} \mathbb{E}_\vartheta\left[\partial_\vartheta^2 \log f(\vartheta, X)\right]$$

$$\overset{\text{conséquence}}{=} -\mathbb{I}(\vartheta).$$

Conclusion

- En combinant les deux estimations + lemme de Slutsky :

$$\sqrt{n}(\widehat{\vartheta}_n^{\,\mathtt{mv}} - \vartheta) \approx -\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n \partial_\vartheta \log f(\vartheta, X_i)}{\frac{1}{n}\sum_{i=1}^n \partial_\vartheta^2 \log f(\vartheta, X_i)}$$

$$\xrightarrow{d} \frac{\mathcal{N}\big(0, \mathbb{I}(\vartheta)\big)}{\mathbb{I}(\vartheta)}$$

$$\overset{\text{loi}}{=} \mathcal{N}\Big(0, \frac{1}{\mathbb{I}(\vartheta)}\Big).$$

- Le raisonnement est rigoureux dès lors que : i) on a la convergence de $\widehat{\vartheta}_n^{\,\mathtt{mv}}$, ii) on peut justifier le lemme et sa conséquence, iii) $\mathbb{I}(\vartheta)$ est bien définie et non dégénérée et iv) on sait contrôler le terme de reste dans le développement asymptotique, partie la plus difficile.

Modèle régulier

Définition

La famille de densités $\{f(\vartheta,\cdot), \vartheta \in \Theta\}$, *par rapport à la mesure dominante* μ, $\Theta \subset \mathbb{R}$, *est* *régulière* *si*

- Θ *ouvert et* $\{f(\vartheta,\cdot) > 0\} = \{f(\vartheta',\cdot) > 0\}$, $\forall \vartheta, \vartheta' \in \Theta$.
- μ*-p.p.* $\vartheta \rightsquigarrow f(\vartheta,\cdot)$, $\vartheta \rightsquigarrow \log f(\vartheta,\cdot)$ *sont* $\mathcal{C}^2$.
- $\forall \vartheta \in \Theta, \exists \mathcal{V}_\vartheta \subset \Theta$ *t.q. pour* $a \in \mathcal{V}_\vartheta$

$$|\partial_a^2 \log f(a,x)| + |\partial_a \log f(a,x)| + \big(\partial_a \log f(a,x)\big)^2 \leq g(x)$$

où

$$\int_{\mathbb{R}} g(x) \sup_{a \in \mathcal{V}(\vartheta)} f(a,x)\mu(dx) < +\infty.$$

- *L'information de Fisher est non-dégénérée :*

$$\forall \vartheta \in \Theta, \quad \mathbb{I}(\vartheta) > 0.$$

Résultat principal

Proposition

- *Si l'expérience engendrée par l'observation $X_1, \ldots, X_n \sim_{i.i.d.} \mathbb{P}_\vartheta$ est associée à une famille de probabilités $\{\mathbb{P}_\vartheta, \vartheta \in \Theta\}$ sur $\mathbb{R}$ régulière au sens de la définition précédente, alors*

$$\sqrt{n}\big(\widehat{\vartheta}_{\mathtt{n}}^{\mathtt{mv}} - \vartheta\big) \xrightarrow{d} \mathcal{N}\Big(0, \frac{1}{\mathbb{I}(\vartheta)}\Big).$$

- *Si $\widehat{\vartheta}_n$ est un Z-estimateur régulier asymptotiquement normal de variance $v(\vartheta)$, alors*

$$\forall \vartheta \in \Theta, \quad v(\vartheta) \geq \frac{1}{\mathbb{I}(\vartheta)}.$$

Preuve de la proposition

- Le premier point consiste à rendre rigoureux le raisonnement précédent. Point délicat : le contrôle du terme de reste.
- Optimalité de la variance de l'EMV parmi celle des Z-estimateurs : on a vu que si $\widehat{\vartheta}_n$ est un Z-estimateur régulier associé à la fonction ϕ, alors, sa variance asymptotique $v(\vartheta) = v_\phi(\vartheta)$ vaut

$$v_\phi(\vartheta) = \frac{\mathbb{E}_\vartheta\left[\phi(\vartheta, X)^2\right]}{\left(\mathbb{E}_\vartheta\left[\partial_\vartheta \phi(\vartheta, X)\right]\right)^2}.$$

- A montrer : pour toute fonction ϕ :

$$\boxed{\frac{\mathbb{E}_\vartheta\left[\phi(\vartheta, X)^2\right]}{\left(\mathbb{E}_\vartheta\left[\partial_\vartheta \phi(\vartheta, X)\right]\right)^2} \geq \frac{1}{\mathbb{I}(\vartheta)}}.$$

Preuve de l'inégalité

- Par construction

$$\partial_a \mathbb{E}_\vartheta \left[\phi(a, X)\right]\big|_{a=\vartheta} = 0.$$

- (avec $\dot{\phi}(\vartheta, x) = \partial_\vartheta \phi(\vartheta, x)$)

$$\begin{aligned} 0 &= \int_{\mathbb{R}} \left[\dot{\phi}(\vartheta, x) f(\vartheta, x) + \phi(\vartheta, x) \partial_\vartheta f(\vartheta, x)\right] \mu(dx) \\ &= \int_{\mathbb{R}} \left[\dot{\phi}(\vartheta, x) f(\vartheta, x) + \phi(\vartheta, x) \partial_\vartheta \log f(\vartheta, x) f(\vartheta, x)\right] \mu(dx). \end{aligned}$$

- <u>Conclusion</u>

$$\boxed{\mathbb{E}_\vartheta \left[\dot{\phi}(\vartheta, X)\right] = -\mathbb{E}_\vartheta \left[\phi(\vartheta, X) \partial_\vartheta \log f(\vartheta, X)\right]}$$

Preuve de l'inégalité (fin)

- On a

$$\mathbb{E}_\vartheta\left[\dot{\phi}(\vartheta, X)\right] = -\mathbb{E}_\vartheta\left[\phi(\vartheta, X)\partial_\vartheta \log f(\vartheta, X)\right]$$

- Cauchy-Schwarz :

$$\left(\mathbb{E}_\vartheta\left[\dot{\phi}(\vartheta, X)\right]\right)^2 \leq \mathbb{E}_\vartheta\left[\phi(\vartheta, X)^2\right]\mathbb{E}_\vartheta\left[\left(\partial_\vartheta \log f(\vartheta, X)\right)^2\right],$$

c'est-à-dire

$$\boxed{v_\phi(\vartheta)^{-1} = \frac{\left(\mathbb{E}_\vartheta\left[\dot{\phi}(\vartheta, X)\right]\right)^2}{\mathbb{E}_\vartheta\left[\phi(\vartheta, X)^2\right]} \leq \mathbb{I}(\vartheta).}$$

Information de Fisher dans un modèle général

Définition

- *Situation : suite d'expériences statistiques*

$$\mathcal{E}^n = \left(\mathfrak{Z}^n, \mathcal{Z}^n, \{\mathbb{P}_\vartheta^n, \vartheta \in \Theta\}\right)$$

dominées par μ_n, associées à l'observation $Z^{(n)}$,

$$f_n(\vartheta, z) = \frac{d\,\mathbb{P}_\vartheta^n}{d\mu^n}(z), \quad z \in \mathfrak{Z}^n, \vartheta \in \Theta \subset \mathbb{R}.$$

- *Information de Fisher (si elle existe) de l'expérience au point ϑ :*

$$\mathbb{I}(\vartheta \,|\, \mathcal{E}_n) = \mathbb{E}_\vartheta^n\left[\left(\partial_\vartheta \log f_n(\vartheta, Z^{(n)})\right)^2\right]$$

Le cas multidimensionnel

- Même contexte que précédemment, avec $\Theta \subset \mathbb{R}^d$, et $d \geq 1$.
- Matrice d'information de Fisher

$$\mathbb{I}(\vartheta) = \mathbb{E}_\vartheta \left[\nabla_\vartheta \log f(\vartheta, Z^n) \nabla_\vartheta \log f(\vartheta, Z^n)^T\right]$$

  matrice symétrique positive.
- Si $\mathbb{I}(\vartheta)$ définie et si $\mathcal{E}^n$ modèle de densité, en généralisant à la dimension d les conditions de régularité, on a

$$\sqrt{n}\left(\widehat{\vartheta}_n^{\,\mathtt{mv}} - \vartheta\right) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{I}(\vartheta)^{-1}\right).$$

Interprétation géométrique

- On pose $\mathbb{D}(a, \vartheta) = \mathbb{E}_\vartheta \left[\log f(a, X) \right]$. On a vu (inégalité d'entropie) que

$$\mathbb{D}(a, \vartheta) = \int_{\mathbb{R}} \log f(a, x) f(\vartheta, x) \mu(dx)$$
$$\leq \int_{\mathbb{R}} \log f(\vartheta, x) f(\vartheta, x) \mu(dx) = \mathbb{D}(\vartheta, \vartheta).$$

- On a

$$\boxed{\mathbb{I}(\vartheta) = \partial_a^2 \mathbb{D}(a, \vartheta)\big|_{a=\vartheta}.}$$

  - Si $\mathbb{I}(\vartheta)$ est « petite », le rayon de courbure de $a \rightsquigarrow \mathbb{D}(a, \vartheta)$ est grand dans un voisinage de ϑ : la stabilisation d'un maximum empirique (l'EMV) est plus difficile, rendant moins précis l'estimation.
  - Si $\mathbb{I}(\vartheta)$ est « grande », le rayon de courbure est petit et le maximum de l'EMV est mieux localisé.

Information de Fisher et régression

- $\mathcal{E}^n$ expérience engendrée par $(\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_n, Y_n)$ avec

$$Y_i = r(\vartheta, \mathbf{x}_i) + \xi_i,$$

  ξ_i : densité g par rapport à la mesure de Lebesgue + « design » déterministe.

- Observation : $Z^n = (Y_1, \ldots, Y_n)$, $\mu^n = dy_1 \ldots dy_n$, $z = (y_1, \ldots, y_n)$ et

$$f_n(\vartheta, Z^n) = \prod_{i=1}^{n} g\big(Y_i - r(\vartheta, \mathbf{x}_i)\big)$$

- Information de Fisher

$$\mathbb{I}(\vartheta | \mathcal{E}^n) = \mathbb{E}_\vartheta\left[\big(\partial_\vartheta \log f_n(\vartheta, Z^n)\big)^2\right]$$

Information de Fisher et régression

- Formule explicite pour la log-vraisemblance

$$\partial_\vartheta \log f_n(\vartheta, Z^n) = \sum_{i=1}^n \partial_\vartheta \log g\big(Y_i - r(\vartheta, \mathbf{x}_i)\big)$$

- Propriété analogue avec le modèle de densité : $\mathbb{E}_\vartheta \left[\partial_\vartheta \log g\big(Y_i - r(\vartheta, \mathbf{x}_i)\big)\right] = 0.$
- Information de Fisher par indépendance + centrage :

$$\mathbb{I}(\vartheta|\mathcal{E}^n) = \sum_{i=1}^n \mathbb{E}_\vartheta^n \left[\Big(\partial_\vartheta \log g\big(Y_i - r(\vartheta, \mathbf{x}_i)\big)\Big)^2\right]$$

$$= \ldots$$

Exemples et applications

A titre d'exercice, savoir calculer l'information de Fisher pour :

- L'estimation du paramètre d'une loi de Poisson dans le modèle de densité.
- L'estimation de la moyenne-variance pour un échantillon gaussien.
- La régression logistique
- L'estimation du paramètre d'une loi exponentielle avec ou sans censure.

Efficacité à un pas

- Dans un modèle régulier, le calcul numérique de l'EMV peut être difficile à réaliser.
- Si l'on dispose d'un estimateur $\widehat{\vartheta}_n$ asymptotiquement normal et si les évaluations

$$\ell_n'(\vartheta) = \frac{1}{n}\sum_{i=1}^n \partial_\vartheta \log f(\vartheta, X_i), \quad \ell_n''(\vartheta) = \frac{1}{n}\sum_{i=1}^n \partial_\vartheta^2 \log f(\vartheta, X_i)$$

sont faciles, alors on peut corriger $\widehat{\vartheta}_n$ de sorte d'avoir le même comportement asymptotique que l'EMV :

$$\widetilde{\vartheta}_n = \widehat{\vartheta}_n - \frac{\ell_n'(\widehat{\vartheta}_n)}{\ell_n''(\widehat{\vartheta}_n)} \quad \text{(algorithme de Newton)}$$

satisfait

$$\boxed{\sqrt{n}\big(\widetilde{\vartheta}_n - \vartheta\big) \xrightarrow{d} \mathcal{N}\Big(0, \frac{1}{\mathbb{I}(\vartheta)}\Big)}$$