MAP 433 : Introduction aux méthodes statistiques. Cours 4

21 février 2014

Aujourd'hui

M-estimation

- <u>Situation</u> : on observe $X_1, \ldots, X_n$ de loi $\mathbb{P}_\vartheta$ sur $\mathbb{R}$ et $\vartheta \in \Theta$.
- <u>Principe</u> : Se donner une application $\psi : \Theta \times \mathbb{R} \to \mathbb{R}_+$ telle que, pour tout $\vartheta \in \Theta \subset \mathbb{R}^d$,

$$a \rightsquigarrow \mathbb{E}_\vartheta \left[\psi(a, X)\right] = \int \psi(a, x)\, \mathbb{P}_\vartheta(dx)$$

admet un maximum en $a = \vartheta$.

Définition

On appelle M-estimateur associé à ψ tout estimateur $\widehat{\vartheta}_n$ satisfaisant

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

Un exemple classique : paramètre de localisation

- $\Theta = \mathbb{R}$, $\mathbb{P}_\vartheta(dx) = f(x-\vartheta)dx$, et $\int_{\mathbb{R}} xf(x)dx = 0$, $\int_{\mathbb{R}} x^2\,\mathbb{P}_\vartheta(dx) < +\infty$ pour tout $\vartheta \in \mathbb{R}$. On pose

$$\boxed{\psi(a,x) = -(a-x)^2}$$

- La fonction

$$a \rightsquigarrow \mathbb{E}_\vartheta\left[\psi(a,X)\right] = -\int_{\mathbb{R}} (a-x)^2 f(x-\vartheta)dx$$

admet un maximum en $a = \mathbb{E}_\vartheta\left[X\right] = \int_{\mathbb{R}} xf(x-\vartheta)dx = \vartheta$.

- *M*-estimateur associé :

$$\sum_{i=1}^n (X_i - \widehat{\vartheta}_n)^2 = \min_{a\in\mathbb{R}} \sum_{i=1}^n (X_i - a)^2.$$

Paramètre de localisation

- C'est aussi un Z-estimateur associé à $\phi(a,x) = 2(x-a)$: on résout

$$\sum_{i=1}^{n}(a - X_i) = 0 \ \text{ d'où } \ \widehat{\vartheta}_n = \overline{X}_n.$$

- Dans cet exemple très simple, tous les points de vue coïncident.
- Si, dans le même contexte, $\int_{\mathbb{R}} x^2 \, \mathbb{P}_\vartheta(dx) = +\infty$ et $f(x) = f(-x)$, on peut utiliser Z-estimateur avec $\phi(a,x) = \mathrm{Arctg}(x-a)$. Méthode robuste, mais est-elle optimale ? Peut-on faire mieux si f est connue ? A suivre...

Lien entre Z- et M- estimateurs

- Pas d'inclusion entre ces deux classes d'estimateurs en général :
  - Si ψ non-régulière, M-estimateur $\not\Rightarrow$ Z-estimateur
  - Si une équation d'estimation admet plusieurs solutions distinctes, Z-estimateur $\not\Rightarrow$ M-estimateur (cas d'un extremum local).
- Toutefois, si ψ est régulière, les M-estimateurs sont des Z-estimateurs : si $\Theta \subset \mathbb{R}$ ($d = 1$), en posant

$$\phi(a,x) = \partial_a \psi(a,x),$$

on a

$$\boxed{\sum_{i=1}^{n} \partial_a \psi(\vartheta, X_i)\big|_{a=\widehat{\vartheta}_n} = \sum_{i=1}^{n} \phi(\widehat{\vartheta}_n, X_i) = 0}.$$

Maximum de vraisemblance

- Principe fondamental et incontournable en statistique. Cas particuliers connus depuis le XVIIIème siècle. Définition générale : Fisher (1922).
- Fournit une première méthode systématique de construction d'un M-estimateur (souvent un Z-estimateur, souvent aussi *a posteriori* un estimateur par substitution simple).
- Procédure optimale (dans quel sens ?) sous des hypothèses de régularité de la famille $\{\mathbb{P}_\vartheta, \vartheta \in \Theta\}$ (Cours 6).
- Parfois difficile à mettre en oeuvre en pratique $\rightarrow$ méthodes numériques, statistique computationnelle.

Fonction de vraisemblance

- La famille $\{\mathbb{P}_\vartheta, \vartheta \in \Theta\}$ est dominée par une mesure σ-finie μ. On se donne, pour $\vartheta \in \Theta$

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

Définition

Fonction de vraisemblance du n-échantillon associée à la famille $\{f(\vartheta,\cdot), \vartheta \in \Theta\}$:

$$\boxed{\vartheta \rightsquigarrow \mathcal{L}_n(\vartheta, X_1, \ldots, X_n) = \prod_{i=1}^n f(\vartheta, X_i)}$$

- C'est une fonction aléatoire (définie μ-presque partout).

Exemples

- Exemple 1 : Modèle de Poisson. On observe

$$X_1, \ldots, X_n \sim_{\text{i.i.d.}} \text{Poisson}(\vartheta),$$

$\vartheta \in \Theta = \mathbb{R}_+ \setminus \{0\}$ et prenons $\mu(dx) = \sum_{k \in \mathbb{N}} \delta_k(dx)$.

- La densité de $\mathbb{P}_\vartheta$ par rapport à μ est

$$f(\vartheta, x) = e^{-\vartheta} \frac{\vartheta^x}{x!}, \quad x = 0, 1, 2, \ldots.$$

- La fonction de vraisemblance associée s'écrit

$$\vartheta \leadsto \mathcal{L}_n(\vartheta, X_1, \ldots, X_n) = \prod_{i=1}^n e^{-\vartheta} \frac{\vartheta^{X_i}}{X_i!}$$
$$= \frac{1}{\prod_{i=1}^n X_i!} e^{-n\vartheta} \vartheta^{\sum_{i=1}^n X_i}.$$

Exemples

- Exemple 2 Modèle de Cauchy. On observe

$$X_1, \ldots, X_n \sim_{\text{i.i.d.}} \text{Cauchy},$$

$\vartheta \in \Theta = \mathbb{R}$ et $\mu(dx) = dx$ (par exemple).

- On a alors

$$\mathbb{P}_\vartheta(dx) = f(\vartheta, x)dx = \frac{1}{\pi\big(1 + (x - \vartheta)^2\big)}dx.$$

- La fonction de vraisemblance associée s'écrit

$$\vartheta \rightsquigarrow \mathcal{L}_n(\vartheta, X_1, \ldots, X_n) = \frac{1}{\pi^n}\prod_{i=1}^n \big(1 + (X_i - \vartheta)^2\big)^{-1}.$$

Principe de maximum de vraisemblance

- Cas d'une famille de lois restreinte à deux points

$$\Theta = \{\vartheta_1, \vartheta_2\} \subset \mathbb{R},$$

avec $\mathbb{P}_{\vartheta_i}$ discrète et $\mu(dx)$ la mesure de comptage.

- A priori, pour tout $(x_1, \ldots, x_n)$, et pour $\vartheta \in \{\vartheta_1, \vartheta_2\}$,

$$\begin{aligned}\mathbb{P}_\vartheta\left[X_1 = x_1, \ldots, X_n = x_n\right] &= \prod_{i=1}^n \mathbb{P}_\vartheta\left[X_i = x_i\right] \\ &= \prod_{i=1}^n f(\vartheta, x_i).\end{aligned}$$

La probabilité d'avoir la réalisation fixée $(x_1, \ldots, x_n)$.

Principe de maximum de vraisemblance

- A posteriori, on observe $(X_1, \ldots, X_n)$. L'événement

$$\Big\{\prod_{i=1}^n f(\vartheta_1, X_i) > \prod_{i=1}^n f(\vartheta_2, X_i)\Big\} \quad \text{(Cas 1)}$$

ou bien l'événement

$$\Big\{\prod_{i=1}^n f(\vartheta_2, X_i) > \prod_{i=1}^n f(\vartheta_1, X_i)\Big\} \quad \text{(Cas 2)}$$

est réalisé. (On ignore le cas d'égalité.)

- Principe de maximum de vraisemblance :

$$\widehat{\vartheta}_n^{\,\mathtt{mv}} = \vartheta_1 1_{\{\text{Cas } 1\}} + \vartheta_2 1_{\{\text{Cas } 2\}}.$$

Estimateur du maximum de vraisemblance

- On généralise le principe précédent pour une famille de lois et un ensemble de paramètres quelconques.
- <u>Situation</u> : $X_1, \ldots, X_n \sim_{\text{i.i.d.}} \mathbb{P}_\vartheta$, $\{\mathbb{P}_\vartheta, \vartheta \in \Theta\}$ dominée, $\Theta \subset \mathbb{R}^d$, $\vartheta \rightsquigarrow \mathcal{L}_n(\vartheta, X_1, \ldots, X_n)$ vraisemblance associée.

Définition

On appelle estimateur du maximum de vraisemblance tout estimateur $\widehat{\vartheta}_n^{\,\mathtt{mv}}$ *satisfaisant*

$$\mathcal{L}_n(\widehat{\vartheta}_n^{\,\mathtt{mv}}, X_1, \ldots, X_n) = \max_{\vartheta \in \Theta} \mathcal{L}_n(\vartheta, X_1, \ldots, X_n).$$

- Existence, unicité...

Remarques

- Log-vraisemblance :

$$\vartheta \rightsquigarrow \ell_n(\vartheta, X_1, \ldots, X_n) = \log \mathcal{L}_n(\vartheta, X_1, \ldots, X_n) = \sum_{i=1}^{n} \log f(\vartheta, X_i).$$

Bien défini si $f(\vartheta, \cdot) > 0$ μ-pp.

Max. vraisemblance = max. log-vraisemblance.

- L'estimateur du maximum de vraisemblance ne dépend pas du choix de la mesure dominante μ.
- Notion de racine de l'équation de vraisemblance : tout estimateur $\widehat{\vartheta}_n^{\,\mathtt{rv}}$ vérifiant

$$\nabla_\vartheta \ell_n(\widehat{\vartheta}_n^{\,\mathtt{rv}}, X_1, \ldots, X_n) = 0.$$

Exemple : modèle normal

L'expérience statistique est engendrée par un n-échantillon de loi $\mathcal{N}(\mu, \sigma^2)$, le paramètre est $\vartheta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+ \setminus \{0\}$.

- Vraisemblance

$$\mathcal{L}_n((\mu, \sigma^2), X_1, \ldots, X_n) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\tfrac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2\right).$$

- Log-vraisemblance

$$\ell_n((\mu, \sigma^2), X_1, \ldots, X_n) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2.$$

Exemple : modèle normal

Equation(s) de vraisemblance

$$\begin{cases} \partial_\mu \ell_n\big((\mu,\sigma^2),X_1,\ldots,X_n\big) & = & \dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^n (X_i-\mu) \\ \partial_{\sigma^2} \ell_n\big((\mu,\sigma^2),X_1,\ldots,X_n\big) & = & -\dfrac{n}{2\sigma^2}+\dfrac{1}{2\sigma^4}\displaystyle\sum_{i=1}^n (X_i-\mu)^2. \end{cases}$$

Solution de ces équations (pour $n \geq 2$) :

$$\boxed{\widehat{\vartheta}_n^{\,\mathtt{rv}} = \Big(\overline{X}_n, \frac{1}{n}\sum_{i=1}^n (X_i-\overline{X}_n)^2\Big)}$$

et on vérifie que $\widehat{\vartheta}_n^{\,\mathtt{rv}} = \widehat{\vartheta}_n^{\,\mathtt{mv}}$.

Exemple : modèle de Poisson

- Vraisemblance

$$\mathcal{L}_n(\vartheta, X_1, \ldots, X_n) = \frac{1}{\prod_{i=1}^n X_i!} e^{-n\vartheta} \vartheta^{\sum_{i=1}^n X_i}.$$

- Log-vraisemblance

$$\ell_n(\vartheta, X_1, \ldots, X_n) = c(X_1, \ldots, X_n) - n\vartheta + \sum_{i=1}^n X_i \log \vartheta.$$

- Equation de vraisemblance

$$-n + \sum_{i=1}^n X_i \frac{1}{\vartheta} = 0, \quad \text{soit} \quad \boxed{\widehat{\vartheta}_n^{\,\mathtt{rv}} = \frac{1}{n} \sum_{i=1}^n X_i = \overline{X}_n}$$

et on vérifie que $\widehat{\vartheta}_n^{\,\mathtt{rv}} = \widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mv}}$.

Exemple : modèle de Laplace

L'expérience statistique est engendrée par un n-échantillon de loi de Laplace de paramètre $\vartheta \in \Theta = \mathbb{R}$. La densité par rapport à la mesure de Lebesgue :

$$f(\vartheta, x) = \frac{1}{2\sigma} \exp\big(- \frac{|x-\vartheta|}{\sigma} \big),$$

où $\sigma > 0$ est connu.

- Vraisemblance

$$\mathcal{L}_n(\vartheta, X_1, \ldots, X_n) = (2\sigma)^{-n} \exp\big(- \frac{1}{\sigma} \sum_{i=1}^{n} \big|X_i - \vartheta\big| \big)$$

- Log-vraisemblance

$$\ell_n(\vartheta, X_1, \ldots, X_n) = -n \log(2\sigma) - \frac{1}{\sigma} \sum_{i=1}^{n} \big|X_i - \vartheta\big|.$$

Exemple : modèle de Laplace

Maximiser $\mathcal{L}_n(\vartheta, X_1, \ldots, X_n)$ revient à minimiser la fonction $\vartheta \rightsquigarrow \sum_{i=1}^n \left|X_i - \vartheta\right|$, dérivable presque partout de dérivée constante par morceaux. Equation de vraisemblance :

$$\sum_{i=1}^n \mathrm{sign}(X_i - \vartheta) = 0.$$

Soit $X_{(1)} \leq \ldots \leq X_{(n)}$ la statistique d'ordre.

- n pair : $\widehat{\vartheta}_n^{\,\mathtt{mv}}$ n'est pas unique ; tout point de l'intervalle $\left[X_{\left(\frac{n}{2}\right)}, X_{\left(\frac{n}{2}+1\right)}\right]$ est un EMV.
- n impair : $\widehat{\vartheta}_n^{\,\mathtt{mv}} = X_{\left(\frac{n+1}{2}\right)}$, l'EMV est unique. Mais $\widehat{\vartheta}_n^{\,\mathtt{rv}}$ n'existe pas.
- pour tout n, la médiane empirique est un EMV.

Exemple : modèle de Cauchy

- **Vraisemblance**

$$\mathcal{L}_n(\vartheta, X_1, \ldots, X_n) = \pi^{-n} \prod_{i=1}^{n} \frac{1}{1+(X_i-\vartheta)^2}.$$

- **Log-vraisemblance**

$$\ell_n(\vartheta, X_1, \ldots, X_n) = -n \log \pi - \sum_{i=1}^{n} \log\left(1+(X_i-\vartheta)^2\right)$$

- **Equation de vraisemblance**

$$\boxed{\sum_{i=1}^{n} \frac{X_i-\vartheta}{1+(X_i-\vartheta)^2} = 0}$$

pas de solution explicite et admet en général plusieurs solutions.

Maximum de vraisemblance = M-estimateur

- Une inégalité de convexité : μ mesure σ-finie sur $\mathbb{R}$; f, g deux densités de probabilités par rapport à μ. Alors

$$\boxed{\int_{\mathbb{R}} f(x) \log f(x) \mu(dx) \geq \int_{\mathbb{R}} f(x) \log g(x) \mu(dx)}$$

(si les intégrales sont finies) avec égalité ssi $f = g$ μ-pp.

- Preuve : à montrer

$$\int_{\mathbb{R}} f(x) \log \frac{g(x)}{f(x)} \mu(dx) \leq 0.$$

(avec une convention de notation appropriée)

Une inégalité de convexité

- On a $\log(1+x) \leq x$ pour $x \geq -1$ avec égalité ssi $x = 0$.
- Donc
$$\log \frac{g(x)}{f(x)} = \log \Big(1 + \big(\frac{g(x)}{f(x)} - 1\big)\Big) \leq \frac{g(x)}{f(x)} - 1$$
(avec égalité ssi $f(x) = g(x)$).
- Finalement
$$\begin{aligned}\int_{\mathbb{R}} f(x) \log \frac{g(x)}{f(x)} \mu(dx) &\leq \int_{\mathbb{R}} f(x) \Big(\frac{g(x)}{f(x)} - 1\Big) \mu(dx) \\ &= \int_{\mathbb{R}} g(x)\mu(dx) - \int_{\mathbb{R}} f(x)\mu(dx) \\ &= 0.\end{aligned}$$

Conséquence pour l'EMV

- On pose

$$\boxed{\psi(a,x) := \log f(a,x), \quad a \in \Theta, \ x \in \mathbb{R}}$$

(avec une convention pour le cas où on n'a pas $f(a,\cdot) > 0$.)

- La fonction

$$a \rightsquigarrow \mathbb{E}_\vartheta\left[\psi(a,X)\right] = \int_{\mathbb{R}} \log f(a,x) f(\vartheta,x)\mu(dx)$$

a un maximum en $a = \vartheta$ d'après l'inégalité de convexité.

- Le M-estimateur associé à ψ maximise la fonction

$$a \rightsquigarrow \sum_{i=1}^{n} \log f(a, X_i) = \ell_n(a, X_1, \ldots, X_n)$$

c'est-à-dire la log-vraisemblance. C'est l'estimateur du maximum de vraisemblance.

- C'est aussi un Z-estimateur si la fonction $\vartheta \rightsquigarrow \log f(\vartheta, \cdot)$ est régulière, associé à la fonction

$$\phi(\vartheta, x) = \partial_\vartheta \log f(\vartheta, x) = \frac{\partial_\vartheta f(\vartheta, x)}{f(\vartheta, x)}, \ \vartheta \in \Theta, x \in \mathbb{R}$$

lorsque $\Theta \subset \mathbb{R}$, à condition que le maximum de log-vraisemblance n'est pas atteint sur la frontière de Θ. (Se généralise en dimension d.)

Choix de modèle statistique

- Le statisticien a le choix de la famille $\{\mathbb{P}_\vartheta, \vartheta \in \Theta\}$. L'EMV dépend de ce choix.
- Exemple : on a l'échantillon $(n = 10)$:

$$\underbrace{0.92, -0.20, -1.80, 0.02, 0.49, 1.41, -1.59, -1.29, 0.34}_{\textit{tirage de } \mathcal{N}(0,1)}, 100.$$

- On prend $\mathbb{P}_\vartheta(dx) = f(x - \vartheta)dx$ pour deux f différents :
- f densité de la loi normale $\Rightarrow \widehat{\vartheta}_n^{\,\mathtt{mv}} = \overline{X}_n = 9.83$.
- f densité de loi de Laplace $\Rightarrow$ tout point de l'intervalle $[0.02, 0.34]$ est un $\widehat{\vartheta}_n^{\,\mathtt{mv}}$, en particulier, la médiane :

$$\widehat{\vartheta}_n^{\,\mathtt{mv}} = M_n = (0.02 + 0.34)/2 = 0.18.$$

- Autre choix de modèle...

Un M-estimateur qui n'est pas un Z-estimateur

- On observe $X_1, \ldots, X_n \sim_{\text{i.i.d.}}$ uniformes sur $[0, \vartheta]$, $\vartheta \in \Theta = \mathbb{R}_+ \setminus \{0\}$.
- On a
$$\mathbb{P}_{\vartheta}(dx) = \vartheta^{-1} 1_{[0,\vartheta]}(x) dx$$
et
$$\mathcal{L}_n(\vartheta, X_1, \ldots, X_n) = \vartheta^{-n} \prod_{i=1}^{n} 1_{[0,\vartheta]}(X_i) = \vartheta^{-n} 1_{\{\max_{1 \leq i \leq n} X_i \leq \vartheta\}}$$
- La fonction de vraisemblance n'est pas régulière.
- L'estimateur du maximum de vraisemblance est $\widehat{\vartheta}_{\mathtt{n}}^{\mathtt{mv}} = \max_{1 \leq i \leq n} X_i$.

Asymptotique des Z- et M-estimateurs

- Problème général délicat. Dans ce cours : conditions suffisantes.
- Convergence : critère simple pour les M-estimateurs.
- Vitesse de convergence : technique simple pour les Z-estimateurs, à condition de savoir que l'estimateur est convergent.
- Sous des hypothèses de régularité, un M-estimateur est un Z-estimateur.

Convergence des M-estimateurs

- Situation : on observe $X_1, \ldots, X_n$ i.i.d. de loi dans la famille $\{\mathbb{P}_\vartheta, \vartheta \in \Theta\}$.
- $\psi : \Theta \times \mathbb{R} \to \mathbb{R}$ fonction de contraste.
- Loi des grands nombres :

$$M_n(a) = \frac{1}{n} \sum_{i=1}^{n} \psi(a, X_i)$$

converge en $\mathbb{P}_\vartheta$-probabilité vers

$$M(a, \vartheta) \ = \mathbb{E}_\vartheta \left[\psi(a, X)\right]$$

qui atteint son maximum en $a = \vartheta$

- « à montrer » :

$$\widehat{\vartheta}_n = \arg\max_{a \in \Theta} M_n(a) \xrightarrow{\mathbb{P}_\vartheta} \arg\max_{a \in \Theta} \mathbb{E}_\vartheta \left[\psi(a, X)\right] = \vartheta.$$

Convergence des M-estimateurs

Proposition

Si le M-estimateur $\widehat{\vartheta}_n$ associé à la fonction de contraste est bien défini et si

- $\sup_{a \in \Theta} |M_n(a) - M(a, \vartheta)| \xrightarrow{\mathbb{P}_\vartheta} 0,$
- $\forall \varepsilon > 0, \ \sup_{|a-\vartheta| \geq \varepsilon} M(a, \vartheta) < M(\vartheta, \vartheta)$ *(condition de maximum)*

alors $\widehat{\vartheta}_n \xrightarrow{\mathbb{P}_\vartheta} \vartheta$.

- La condition 1 (convergence uniforme) peut être délicate à montrer...

Loi limite des Z-estimateurs

- Situation : on observe $X_1, \ldots, X_n$ i.i.d. de loi dans la famille $\{\mathbb{P}_\vartheta, \vartheta \in \Theta\}$, $\Theta \subset \mathbb{R}$.
- $\widehat{\vartheta}_n$: Z-estimateur associé à $\phi : \Theta \times \mathbb{R} \to \mathbb{R}$ vérifie

$$\boxed{\sum_{i=1}^{n} \phi(\widehat{\vartheta}_n, X_i) = 0}$$

- Si $\widehat{\vartheta}_n$ est un M-estimateur associé à la fonction de contraste ψ régulière, alors c'est un Z-estimateur associé à la fonction $\phi(a, x) = \partial_a \psi(a, x)$.
- On suppose $\widehat{\vartheta}_n$ convergent. Que dire de sa loi limite ?

Loi limite des Z-estimateurs : principe

- Loi des grands nombres

$$Z_n(a) = \frac{1}{n}\sum_{i=1}^{n}\phi(a, X_i) \xrightarrow{\mathbb{P}_\vartheta} Z(a,\vartheta) = \mathbb{E}_\vartheta\left[\phi(a, X)\right]$$

- Principe. Développement de Taylor autour de ϑ :

$$0 = Z_n(\widehat{\vartheta}_n) = Z_n(\vartheta) + (\widehat{\vartheta}_n - \vartheta)Z_n'(\vartheta) + \tfrac{1}{2}(\widehat{\vartheta}_n - \vartheta)^2 Z''(\widetilde{\vartheta}_n).$$

- On néglige le reste :

$$\sqrt{n}(\widehat{\vartheta}_n - \vartheta) \approx \frac{-\sqrt{n}Z_n(\vartheta)}{Z_n'(\vartheta)}$$

Loi limite des Z-estimateurs : principe

- Convergence du numérateur

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

si $\mathbb{E}_\vartheta\left[\phi(\vartheta, X)\right] = 0$ et $\mathbb{E}_\vartheta\left[\phi(\vartheta, X)^2\right] < +\infty$.

- Convergence du dénominateur

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

$\neq 0$ (à supposer).

- $+$ hypothèses techniques pour contrôler le reste (besoin de la convergence de $\widehat{\vartheta}_n$).

Loi limite des Z-estimateurs

Proposition (Convergence des Z-estimateurs)

- *Soit* Θ *un ouvert de* $\mathbb{R}$. *Pour tout* $\vartheta \in \Theta$, $\widehat{\vartheta}_n \xrightarrow{\mathbb{P}_\vartheta} \vartheta$, $\mathbb{E}_\vartheta\left[\phi(\vartheta, X)^2\right] < +\infty$ *et*

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

- *(Contrôle reste) pour tout* $\vartheta \in \Theta$, *pour tout* a *dans un voisinage de* ϑ,

$$|\partial_a^2 \phi(a, x)| \leq g(x),\ \ \mathbb{E}_\vartheta\left[g(X)\right] < +\infty.$$

Alors

$$\sqrt{n}(\widehat{\vartheta}_n - \vartheta) \xrightarrow{d} \mathcal{N}\Big(0, \frac{\mathbb{E}_\vartheta[\phi(\vartheta, X)^2]}{\big(\mathbb{E}_\vartheta[\partial_\vartheta \phi(\vartheta, X)]\big)^2}\Big).$$

Influence d'une variable sur une autre

- Principe : on part de l'observation d'un n-échantillon

$$Y_1, \ldots, Y_n \quad (Y_i \in \mathbb{R})$$

- A chaque observation Y_i est associée une observation auxiliaire $\mathbf{X}_i \in \mathbb{R}^k$.
- On suspecte l'échantillon

$$\mathbf{X}_1, \ldots, \mathbf{X}_n \quad (\mathbf{X}_i \in \mathbb{R}^k)$$

de contenir la « majeure partie de la variabilité des Y_i ».

Modélisation de l'influence

- Si $\mathbf{X}_i$ contient toute la variabilité de Y_i, alors Y_i est mesurable par rapport à $\mathbf{X}_i$: il existe $r : \mathbb{R}^k \to \mathbb{R}$ telle que

$$Y_i = r(\mathbf{X}_i),$$

  mais peu réaliste (ou alors problème d'interpolation numérique).

- Alternative : représentation précédente avec erreur additive : on postule

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

  ξ_i erreur aléatoire centrée (pour des raisons d'identifiabilité).

Motivation : meilleure approximation L^2

- Meilleure approximation L^2. Si $\mathbb{E}\left[Y^2\right] < +\infty$, la meilleure approximation de Y par une variable aléatoire $\mathbf{X}$-mesurable est donnée par l'espérance conditionnelle $\mathbb{E}\left[Y | \mathbf{X}\right]$:

$$\mathbb{E}\left[\left(Y - r(\mathbf{X})\right)^2\right] = \min_h \mathbb{E}\left[\left(Y - h(\mathbf{X})\right)^2\right]$$

- où

$$r(\mathbf{x}) = \mathbb{E}\left[Y | \mathbf{X} = \mathbf{x}\right], \quad \mathbf{x} \in \mathbb{R}^k .$$

- On appelle $r(\cdot)$ **fonction de régression de Y sur $\mathbf{X}$**.

Régression

- On définit :

$$\xi = Y - \mathbb{E}\left[Y \middle| \mathbf{X}\right] \implies \mathbb{E}\left[\xi\right] = 0.$$

- On a alors naturellement la représentation désirée

$$Y = r(\mathbf{X}) + \xi, \quad \mathbb{E}\left[\xi\right] = 0$$

si l'on pose

$$\boxed{r(\mathbf{x}) = \mathbb{E}\left[Y \middle| \mathbf{X} = \mathbf{x}\right], \ \ \mathbf{x} \in \mathbb{R}^k}$$

- On observe alors un n-échantillon

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

où

$$Y_i = r(\mathbf{X}_i) + \xi_i, \ \ \mathbb{E}\left[\xi_i\right] = 0$$

avec comme paramètre la fonction $r(\cdot)$+ un jeu d'hypothèses sur la loi des ξ_i.

Modèle de régression à design aléatoire

Définition

Modèle de régression **à design aléatoire** = *donnée de l'observation*

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

avec $(Y_i, \mathbf{X}_i) \in \mathbb{R} \times \mathbb{R}^k$ *i.i.d.*, *et*

$$Y_i = r(\vartheta, \mathbf{X}_i) + \xi_i, \quad \mathbb{E}\left[\xi_i | \mathbf{X}_i\right] = 0, \quad \vartheta \in \Theta \subset \mathbb{R}^d.$$

- $\mathbf{x} \rightsquigarrow r(\vartheta, \mathbf{x})$ *fonction de régression, connue au paramètre* ϑ *près.*
- $\mathbf{X}_i$ = *variables explicatives, co-variables, prédicteurs ;* $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ = **design**.

Modèle alternatif : signal+bruit

- Principe : sur un exemple. On observe

$$Y_i = r(\vartheta, i/n) + \xi_i, \quad i = 1, \ldots, n$$

  où $r(\vartheta, \cdot) : [0,1] \to \mathbb{R}$ est une fonction connue au paramètre $\vartheta \in \Theta \subset \mathbb{R}^d$ près, et les ξ_i sont i.i.d., $\mathbb{E}\left[\xi_i\right] = 0$.

- But : reconstruire $r(\vartheta, \cdot)$ c'est-à-dire estimer ϑ.
- Plus généralement, on observe

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

  où $\mathbf{x}_1, \ldots, \mathbf{x}_n$ sont des points de $\mathbb{R}^k$ déterministes.

Modèle de régression à design déterministe

Définition

Modèle de régression **à design déterministe** = *donnée de l'observation*

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

avec $Y_i \in \mathbb{R}, \mathbf{x}_i \in \mathbb{R}^k$, *et*

$$Y_i = r(\vartheta, \mathbf{x}_i) + \xi_i, \ \ \mathbb{E}\left[\xi_i\right] = 0, \ \ \vartheta \in \Theta \subset \mathbb{R}^d.$$

- $\mathbf{x}_i$ *déterministes, donnés (ou choisis) : plan d'expérience, points du « design ».*
- *Hypothèses sur les* ξ_i *: à débattre. Pour simplifier, les* ξ_i *sont i.i.d. (hypothèse restrictive).*
- $\Longrightarrow$ *les* Y_i *ne sont pas i.i.d.*

Question : Comment estimer θ dans ce modèle ?

Régression gaussienne

- Modèle de régression à design déterministe :

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

- Supposons : $\xi_i \sim \mathcal{N}(0, \sigma^2)$, i.i.d.
- On a alors le modèle de régression gaussienne. Comment estimer ϑ ? On sait explicier la loi de l'observation $Z = (Y_1, \ldots, Y_n) \Longrightarrow$ appliquer le principe du maximum de vraisemblance.
- La loi de Y_i :

$$\mathbb{P}^{Y_i}(dy) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\big(-\frac{1}{2\sigma^2}(y - r(\vartheta, \mathbf{x}_i))^2\big) dy$$
$$\ll dy.$$

EMV pour régression gaussienne

- Le modèle $\{\mathbb{P}_\vartheta^n = \text{loi de } (Y_1, \ldots, Y_n), \vartheta \in \mathbb{R}^k\}$ est dominé par $\mu^n(dy_1 \ldots dy_n) = dy_1 \ldots dy_n$.
- D'où

$$\frac{d\,\mathbb{P}_\vartheta^n}{d\mu^n}(y_1, \ldots, y_n) = \prod_{i=1}^n \tfrac{1}{\sqrt{2\pi\sigma^2}} \exp\big(-\tfrac{1}{2\sigma^2}(y_i - r(\vartheta, \mathbf{x}_i))^2 \big)$$

$$= \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\big(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - r(\vartheta, \mathbf{x}_i))^2 \big).$$

- La fonction de vraisemblance

$$\boxed{\mathcal{L}_n(\vartheta, Y_1, \ldots, Y_n) \propto \exp\Big(-\frac{1}{2\sigma^2} \sum_{i=1}^n \big(Y_i - r(\vartheta, \mathbf{x}_i)\big)^2 \Big)}$$

Estimateur des moindres carrés

Maximiser la vraisemblance en régression gaussienne = minimiser la somme des carrés :

$$\sum_{i=1}^{n} \big(Y_i - r(\vartheta, \mathbf{x}_i)\big)^2 \to \min_{\vartheta \in \Theta} .$$

Définition

Estimateur des moindres carrés : tout estimateur $\widehat{\vartheta}_n^{\,\mathtt{mc}}$ t.q.

$$\widehat{\vartheta}_n^{\,\mathtt{mc}} \in \arg\min_{\vartheta \in \Theta} \sum_{i=1}^{n} \big(Y_i - r(\vartheta, \mathbf{x}_i)\big)^2.$$

- L'EMC est un M-estimateur. Pour le modèle de régression gaussienne : $\boxed{\mathrm{EMV} = \mathrm{EMC}}$.
- Existence, unicité.
- Propriétés remarquables si la régression est linéaire : $r(\vartheta, \mathbf{x}_i) = \vartheta^T \mathbf{x}_i$.

Droite de régression

- Modèle le plus simple $\boxed{r(\vartheta, x) = a + bx}$

$$\boxed{Y_i = a + b\,x_i + \xi_i, \quad i = 1, \ldots, n}$$

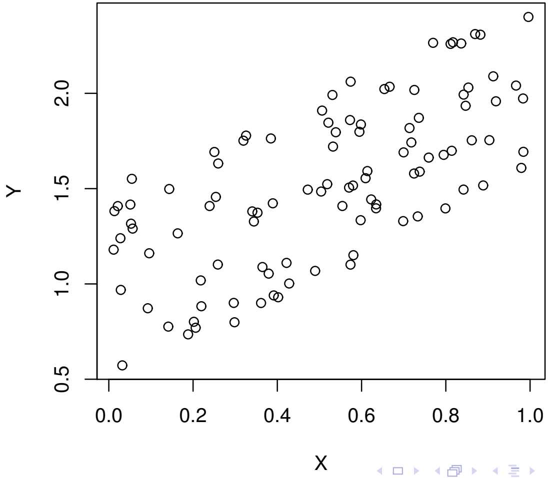
avec $\vartheta = (a, b)^T \in \Theta = \mathbb{R}^2$ et les $(x_1, \ldots, x_n)$ donnés.

- L'estimateur des moindres carrés :

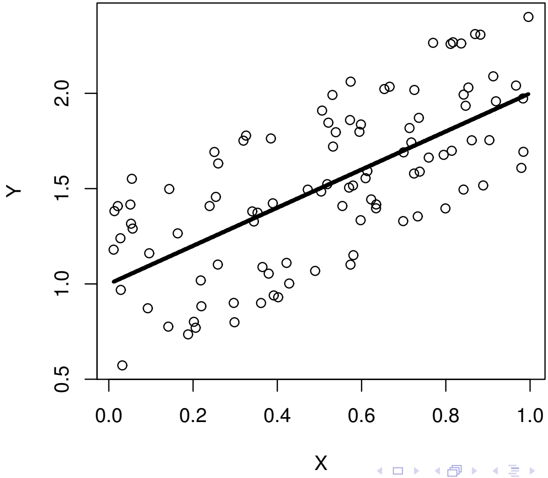
$$\widehat{\vartheta}_n^{\,\mathtt{mc}} = (\hat{a}, \hat{b}) = \arg \min_{(a,b)\in\mathbb{R}^2} \sum_{i=1}^{n} \big(Y_i - a - bx_i\big)^2.$$

- Solution explicite existe toujours, sauf cas pathologique quand tous les x_i sont les mêmes (Poly, page 112).

Régression linéaire simple



Régression linéaire simple



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 $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^T$ 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 $\widehat{\vartheta}_n^{\,\mathtt{mc}}$ satisfaisant

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

- En notations matricielles :

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

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