MDI220, Statistique
Cours 2: Estimation ponctuelle

Anne Sabourin

19 Septembre 2017

Cadre de l'estimation

- $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ un modèle statistique sur l'espace d'obervations $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$.
- $\mathbf{X}$: les données, $\mathbf{X} = (X_1, \ldots, X_n) \overset{\text{i.i.d.}}{\sim} P_\theta$
- **But :** estimer $g(\theta) \in \mathcal{A}$ une quantité d'intérêt.
- **Estimateur** (Rappel) :
  une fonction $\mathcal{X}^n \to \mathcal{A}$, *i.e.* une **statistique**.

Idée directrice

- M-estimateur : construit en **M**inimisant (par rapport à θ) une fonction qui dépend de θ et de $\mathbf{X}$.

  $\rightarrow \widehat{\theta}$ est un $\arg\max$

- Z-estimateur : construit en annulant (*i.e.* en trouvant d'un **Z**éro) une fonction dépendant de θ et de $\mathbf{X}$, en faisant varier θ.

  $\rightarrow \widehat{\theta}$ est une racine.

Exemple type de M-estimateur

- Modèle dominé (par la mesure de Lebesgue) sur $\mathbb{R}$: P_θ a une densité $p_\theta(x)$.
- à $x = (x_1, \ldots, x_n)$ fixé, $t \mapsto p_t(x)$ est la fonction de vraisemblance
- **But** : estimer $g(\theta) = \theta$. (donc $\mathcal{A} = \Theta$).
- Supposons que $\forall x \in \mathcal{X}^n$, $\exists! \widehat{\theta}(x)$ tel que

$$\forall t \in \Theta, p_t(x) \leq p_{\widehat{\theta}}(x).$$

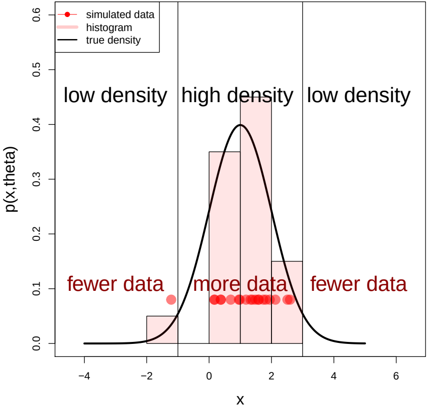
- On pose $M(x, t) = -\log p(x, t)$, $t \in \Theta$

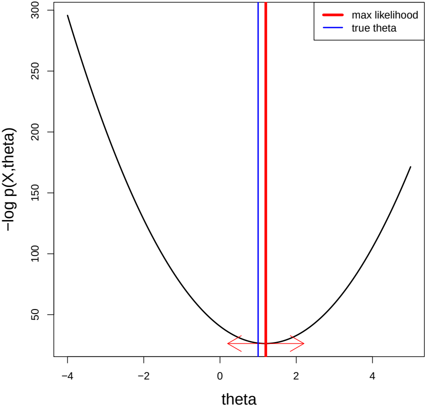
$$\widehat{\theta}(x) = \operatorname*{argmin}_{t \in \Theta} M(x, t),$$

  M est appelée 'fonction de contraste' ou 'contraste'.
- ex : modèle gaussien $\mathcal{N}(\theta, \sigma^2)$, σ^2 connu. $-\log p_t(x) =?$ $\quad \widehat{\theta}(x) =?$

Max de vraisemblance : justification heuristique.

Simulation : $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta = 1, 1)$, $1 \le i \le n = 20$.





X_i plus fréquents où $p_\theta(x)$ grande

$p_\theta(x)\mathrm{d}x \approx \mathbb{P}_\theta(X_i \in \mathrm{d}x)$

θ plus vraisemblable si $p_\theta(X_{1:n})$ grand

Définition : M-estimateur

Soit $M : \mathcal{X}^n \times \mathcal{A} \to \mathbb{R}^+ \cup \{+\infty\}$ un contraste et

$$\operatorname*{argmin}_{t \in \mathcal{A}} M(x,t) = \Big\{ t \in \mathcal{A} : \forall t', M(x,t') \geq M(x,t) \Big\}.$$

Un M-estimateur est une statistique $\widehat{g}(X)$ telle que

$$\widehat{g}(X) = \operatorname*{argmin}_{t \in \mathcal{A}} M(X,t),$$

pour un contraste M admettant un unique minimiseur en t.

- **Notations** : pour $f : \mathcal{A} \mapsto \mathbb{R}$,
  $\operatorname{argmin}_{\mathcal{A}} f = \operatorname{argmin}_{t \in \mathcal{A}} f(t) = \{ t \in \mathcal{A} : \forall t' \in \mathcal{A}, f(t') \geq f(t) \}$.
  Lorsque $\operatorname{argmin} f = \{t_0\}$, on écrit pour simplifier $\operatorname{argmin} f = t_0$.
- La définition suppose l'existence et l'unicité du minimum.
- C'est le cas Si M est strictement convexe en t.

Z-estimateur

- Si $\widehat{g}(X)$ est un M-estimateur et si le contraste M est différentiable p.r.à. t, on a $\nabla_t M(X, \widehat{g}(X)) = 0$.
  $\rightarrow \widehat{g}(X)$ est un **Z**éro de $\nabla_t M(X, \cdot)$.

définition : Z-estimateur

Soit $\Psi : \mathcal{X}^n \times \mathcal{A} \rightarrow \mathbb{R}^d$ telle que

$$\forall x \in \mathcal{X}^n, \exists! \widehat{g}(x) \text{ tel que } \Psi(x, \widehat{g}(x)) = 0.$$

La satistique $\widehat{g}(X)$ est alors appelée Z-estimateur.

Question

Définitions d'un M- et d'un Z- estimateurs très générales, les propriétés de $\widehat{g}$ dépendent du choix de M ou Ψ.

Comment choisir M ou Ψ pour 'bien' estimer $g(\theta)$?

- Dans ce cours : pas de réponse absolue.
- On donne des exemples de construction et on vérifiera qu'elles ont de bonnes propriétés pour le coût quadratique, à taille d'échantillon n fixé.
- Propriétés asymptotiques : cf. le cours de statistiques asymptotiques (MACS 203, P2)

justification II (heuristique) de l'estimateur de max de vraisemblance

- $M(x,t) = -\log p_\theta(x)$. Si $\widehat{\theta}_{MV}(X) = \mathrm{argmin}_{t\in\Theta} M(X,t)$ est unique, $\widehat{\theta}$ est *l'estimateur du maximum de vraisemblance* pour le paramètre θ.
- raison du bon comportement de $\widehat{\theta}_{MV}$: Si $X_{1:n} \overset{\text{i.i.d.}}{\sim} P_{\theta_0}$,

$$\widehat{\theta}(X_{1:n}) = \underset{t}{\mathrm{argmin}} \sum_{i=1}^{n} -\log p_t(X_i) = \underset{t}{\mathrm{argmin}} \frac{1}{n}\sum_{i=1}^{n} \log \frac{p_{\theta_0}(X_i)}{p_t(X_i)}$$
$$\approx_{n\to\infty} \underset{t}{\mathrm{argmin}}\, \mathbb{E}_{\theta_0} \log \frac{p_{\theta_0}(X_1)}{p_t(X_1)} = \underbrace{\int_{\mathcal{X}} p_{\theta_0}(x) \log \frac{p_{\theta_0}(X_1)}{p_t(X_1)} \mathrm{d}x}_{KL(P_{\theta_0}, P_t)}$$

- $KL(P_{\theta_0}, P_t) \geq 0$ mesure la divergence entre P_{θ_0} et P_t.
- cas d'égalité : $P_t = P_{\theta_0}$ *i.e.*(si modèle identifiable) $t = \theta_0$.
- Justification du '$\approx_{n\to\infty}$' : cours de stats asymptotiques.

Max de vraisemblance : Exemple II

- $X_{1:n} \overset{\text{i.i.d.}}{\sim} \mathcal{P}oiss(\theta)$, $\theta \in \Theta = \mathbb{R}_+^*$.
- Pour $t > 0$, $-\log p_t(X_{1:n}) = \dots$

  au tableau

- Résultat : $\widehat{\theta}_{MV}(X_{1:n}) = \frac{1}{n}\sum_{i=1}^n X_i$
- Pourquoi est-ce rassurant ?

Limites de l'estimateur de maximum de vraisemblance

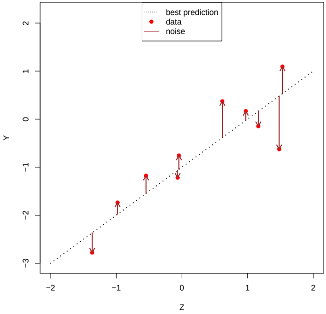
- Souvent pas d'expression explicite pour $\widehat{\theta}_{MV}$
- Alors : recours obligatoire à des méthodes d'optimisation numérique
- $\rightarrow$ Coûteux en temps de calcul et pas exact.

Cadre de la régression

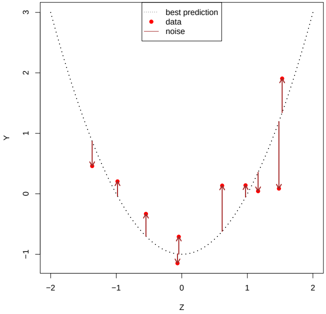
- Observations : $X_i = (Y_i, z_i)$, $Y_i \in \mathbb{R}$ (aléatoire), $z_i \in \mathbb{R}^d$ (donnée du problème, non aléatoire), telles que

$$Y_i = \varphi(\theta, z_i) + \epsilon_i, \qquad \epsilon_{1:n} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \text{ (bruit)} \qquad \theta \text{ à estimer}$$

- Cas fréquent : régression linéaire,
$\varphi(\theta, z_i) = \langle \theta, \Phi(z_i) \rangle = \sum_{j=1}^{d} \theta_j \Phi(z_i)_j$.



$\varphi(\theta, z) = \theta_0 + \theta_1 z$



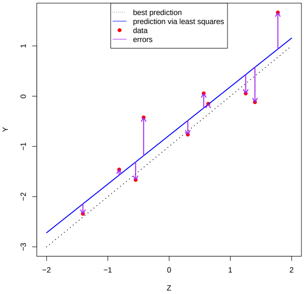
$\varphi(\theta, z) = \theta_0 + \theta_1 z^2$

Estimateur des moindres carrés

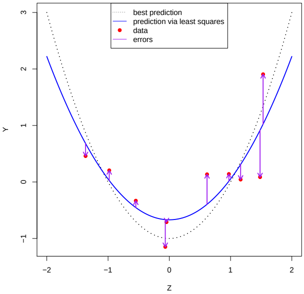
Méthode très ancienne (Gauss).

- Contraste : erreur quadratique entre les Y_i et leur meilleur prédiction $\varphi(z_i, t)$:

$$M(X_{1:n}, t) = \sum_{i=1} \left(\varphi(z_i, t) - Y_i\right)^2 \qquad \widehat{\theta}_{MC}(X) = \operatorname*{argmin}_{t \in \Theta} M(X_{1:n}, t).$$



$$\varphi(\theta, z) = \theta_0 + \theta_1 z$$



$$\varphi(\theta, z) = \theta_0 + \theta_1 z^2$$

Méthode des moments : principe de substitution

But : estimer θ. Supposons que

- on dispose d'une fonction $h = (h_1, \ldots, h_p) : \mathcal{X}^n \to \mathbb{R}^p$ telle que $\Phi(\theta) := \mathbb{E}_\theta\Big[h(X)\Big]$ soit calculable (en fonction de θ)
- on peut retrouver θ à partir de $\Phi(\theta)$, *i.e.* $\theta \mapsto \Phi(\theta)$ est injective. Alors $\exists \Phi^{-1} : \mathrm{Im}(\Phi) \subset \mathbb{R}^p \to \Theta$.

principe de substitution

Remplacer $\Phi(\theta) = \mathbb{E}_\theta(h(X))$ (inconnu car θ inconnnu) par

$$\Phi_n(X_{1:n}) := \frac{1}{n}\sum_{i=1}^{n} h(X_i)$$

- (si $\Phi_n(X) \in \mathrm{Im}(\Phi)$), on pose $\widehat{\theta}(X_{1:n}) = \Phi^{-1} \circ \Phi_n(X_{1:n})$
- cas général : on minimise le contraste $M(X_{1:n}, t) = \|\Phi_n(X_{1:n}) - \Phi(t)\|$

Principe de substitution et minimisation de contraste

- On définit le contraste

$$M(X_{1:n}, t) = \|\Phi_n(X_{1:n}) - \Phi(t)\|$$

- Si $\exists!$ minimiseur, l'estimateur par la méthode des moments est

$$\widehat{\theta}(X_{1:n}) = \operatorname*{argmin}_t M(X_{1:n}, t).$$

Lemme : Condition suffisante pour que $\exists!$ minimiseur

Sous l'hypothèse d'injectivité de $\theta \mapsto \Phi(\theta)$, s'il existe t^* tel que $M(X_{1:n}, t^*) = 0$, alors t^* est l'unique minimiseur de M

- Sous l'hypothèse d'injectivité, si $\Phi_n(X_{1:n}) \in \mathrm{Im}(\Phi)$, le lemme s'applique

Exemple I : paramètre d'une loi Gamma

- $\theta = (\alpha, \lambda) := (\theta_1, \theta_2), \alpha > 0, \lambda > 0.$
- $X_{1:n} \overset{\text{i.i.d.}}{\sim} P_\theta = \mathcal{G}amma(\alpha, \lambda).$
- Modèle dominé par la mesure de Lebesgue, densité

$$p_{(\alpha,\lambda)}(x) = \mathbb{1}_{x>0} \ \ \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$$

- On choisit $h(X) = (X, X^2)$ (méthode des *moments*)
- On montre que $\Phi(\theta) := \mathbb{E}_\theta(h(X)) = (\frac{\theta_1}{\theta_2}, \frac{\theta_1(1+\theta_1)}{\theta_2^2}) := (m_1, m_2)$
- sur $\mathrm{Im}(\Phi) = \{(m_1, m_2) : m_1 > 0, m_2 > m_1^2\}$,

$$\Phi^{-1}(m) = \Big(\frac{m_1^2}{m_2 - m_1^2}, \frac{m_1}{m_2 - m_1^2} \Big).$$

Exemple I : paramètre d'une loi Gamma (suite)

- Contraste : $M(X_{1:n}, \alpha, \lambda) = \|\Phi_n(X_{1:n}) - \Phi(\alpha, \lambda)\|$ avec

$$\Phi_n(X_{1:n}) = \Big(\frac{1}{n}\sum_i X_i, \frac{1}{n}\sum_i X_i^2\Big)$$

- on montre que $\Phi_n(X_{1:n}) \in \mathrm{Im}(\Phi) \to$ le lemme s'applique
- on obtient

$$\widehat{\theta}_M(X_{1:n}) = \Phi^{-1}(\Phi_n(X_{1:n})) = \Big(\frac{\overline{X_n}^2}{\widehat{\sigma}_n^2}, \frac{\overline{X_n}}{\widehat{\sigma}_n^2}\Big)$$

avec $\overline{X_n} = \frac{1}{n}\sum_i X_i$, $\widehat{\sigma}_n^2 = \frac{1}{n}\sum_i X_i - \overline{X_n}^2$.

Exemple II : paramètres d'une loi normale

- $\theta = (\mu, \sigma^2), X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$
- on choisit $h(X) = (X, X^2)$. On a immmédiatement

$$\Phi(\theta) = \mathbb{E}_\theta(h(X)) = (\mu, \mu^2 + \sigma^2) \; ; \quad \mathrm{Im}(\Phi = \{(m_1, m_2) : m_2 > m_1^2\}$$
$$\Phi^{-1}(m) = (m_1, m_2 - m_1^2)$$

- On vérifie que $\Phi_n(X_{1:n}) = \left(\frac{1}{n}\sum_i X_i, \frac{1}{n}\sum_i X_i^2\right) \in \mathrm{Im}(\Phi)$ (comme précédemment)
- On peut poser $\widehat{\theta}_M(X_{1:n}) = \Phi^{-1}(\Phi_n(X))$;

$$\widehat{\theta}_M(X_{1:n}) = (\widehat{\mu}_M, \widehat{\sigma^2}_M) = \left(\frac{1}{n}\sum_i X_i, \frac{1}{n}\sum_i X_i^2 - (\frac{1}{n}\sum_i X_i)^2\right)$$

(moyenne et variance empiriques)