APM-41033-EP : Introduction aux méthodes statistiques

Cours 2 — Estimation ponctuelle

Aymeric Dieuleveut

École polytechnique

Septembre 2026

Rappel et objectifs du cours 2/10

Rappel du cours 1 :

1. Notion de modèle statistique $(\mathsf{Z}, \mathcal{Z}, \{\mathbb{P}_\theta\,,\ \theta \in \Theta\})$
   - Motivation, Définition
   - Identifiabilité, Modèle induit, modèle paramétrique, etc.
   - Notion de $n-$échantillon
2. plusieurs tâches :
   - estimation ponctuelle : donner une valeur plausible du paramètre θ,
   - estimation par région : déterminer un sous-ensemble $\Theta_0(Z) \subset \Theta$ auquel le paramètre θ appartient de façon plausible
   - test : décider si le paramètre θ appartient à un sous-ensemble Θ_0 et évaluer la plausibilité de la décision.
   - prédire : donner une valeur plausible pour une quantité non observée, par exemple prédire la valeur d'une sortie Y à partir d'une entrée X

$\rightarrow$ Objectifs :

1. Décrire les deux grandes techniques d'estimation (Moments et Maximum de Vraisemblance)
2. Comprendre leur intuition et les résultats génériques que nous pouvons obtenir
3. Avancé : Notion de M et Z-estimateurs.

$\rightarrow$ 2 techniques : pour chacune :

Intuition **Définition** **Exemples** **Propriétés** **Généralisation***

Estimateur ponctuel

Définition : Estimateur ponctuel

Soient $(\mathsf{Z}, \mathcal{Z}, \{\mathbb{P}_\theta\,,\ \theta \in \Theta\})$ un modèle statistique paramétrique et $g : \Theta \to \mathbb{R}^q$ une fonction. Un estimateur ponctuel T de $g(\theta)$ est une statistique à valeurs dans $\mathbb{R}^q$, c'est-à-dire une application mesurable $T : (\mathsf{Z}, \mathcal{Z}) \to (\mathbb{R}^q, \mathcal{B}(\mathbb{R}^q))$.

On impose parfois à un estimateur de $g(\theta)$ de prendre ses valeurs dans $g(\Theta)$ lui-même ; cette convention apporte peu et exclut des estimateurs parfaitement admissibles.

Considérons un n-échantillon du modèle

$$(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+), \{\mathcal{E}(\theta)\ :\ \theta \in \mathbb{R}_+^*\}).$$

On note les observation $(X_1, \ldots, X_n)$

Quels sont des estimateurs de θ ?

1. X_1
2. $\bar{X}_n$
3. θ
4. 1
5. -1

Un estimateur :

1. N'est pas nécessairement "bon"
2. Aura des (mesures de) qualité : sans biais, erreur quadratique (minimale), consistant (convergent).

   cf cours 3-4, 6-7

Définition : Méthode des moments, cas simple (informel)

Si $\mathbb{E}_\theta[T(X_1)] = e(\theta)$ [a], l'estimateur des moments basé sur la statistique $T(X_1)$ consiste à identifier le moment empirique $\frac{1}{n}\sum_{i=1}^n T(X_i)$ et le moment exact, pour proposer $\hat{\theta}_n$ solution de $e(\hat{\theta}_n) = \frac{1}{n}\sum_{i=1}^n T(X_i)$.

a. On suppose $\mathbb{E}_\theta[|T(X_1)|] < \infty$ pour tout $\theta \in \Theta$; l'énoncé précis est dans le polycopié.

Exemple : Modèle de Bernoulli

Le modèle est $\big(\{0,1\}^n, \mathcal{P}(\{0,1\}^n), \{\mathcal{B}\mathrm{er}(p)^{\otimes n}, p \in [0,1]\}\big)$. Estimer p ?

$\rightarrow$ Candidat naturel : la moyenne $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, c'est-à-dire la proportion empirique de 1.

Formellement, $\mathbb{E}_p[\ X_1\ ] = p \qquad \rightsquigarrow \qquad \hat{p}_n = \frac{1}{n}\sum_{i=1}^n X_i$

Définition : Méthode des moments, cas simple (informel)

Si $\mathbb{E}_\theta[T(X_1)] = e(\theta)$ [a], l'estimateur des moments basé sur la statistique $T(X_1)$ consiste à identifier le moment empirique $\frac{1}{n}\sum_{i=1}^n T(X_i)$ et le moment exact, pour proposer $\hat{\theta}_n$ solution de $e(\hat{\theta}_n) = \frac{1}{n}\sum_{i=1}^n T(X_i)$.

a. On suppose $\mathbb{E}_\theta[|T(X_1)|] < \infty$ pour tout $\theta \in \Theta$; l'énoncé précis est dans le polycopié.

Exemple : Modèle de Poisson, moment d'ordre 2

Le modèle est $(\mathbb{N}^n, \mathcal{P}(\mathbb{N}^n), \mathcal{C} = \{\mathcal{P}\text{oi}(\lambda)^{\otimes n} \ : \ \lambda \in \mathbb{R}_+^*\})$.

On note que $\mathbb{E}_\lambda[\ X_1{}^2\ ] = \lambda^2 + \lambda = e(\lambda) \quad \rightsquigarrow \quad e(\hat{\lambda}_n) = \frac{1}{n}\sum_{i=1}^n X_i^2 \quad \Rightarrow \hat{\lambda}_n^2 + \hat{\lambda}_n = \frac{1}{n}\sum_{i=1}^n X_i^2$

Méthode : Méthode des moments

- Soit $X_1, \ldots, X_n$ un n-échantillon d'un modèle $(\mathsf{X}, \mathcal{X}, \{\mathbb{Q}_\theta\,,\ \theta \in \Theta\})$ où $\Theta \subset \mathbb{R}^d$.
- Choisissons d fonctions $T_j : \mathsf{X} \to \mathbb{R}, j \in \{1, \ldots, d\}$ mesurables telles que, pour tout $\theta \in \Theta$, $\mathbb{E}_\theta\left[|T_j(X_1)|\right] = \int_{\mathsf{X}} |T_j(x)| \mathbb{Q}_\theta(\mathrm{d}x) < \infty$. On calcule les moments exacts : pour tout $\theta \in \Theta$ et $j = 1, \ldots, d$,

$$\mathbb{E}_\theta\left[T_j(X_1)\right] =: e_j(\theta)\ .$$

- Les moments $e_j(\theta)$ ne sont pas connus, on les estime par les moments empiriques

$$n^{-1} \sum_{i=1}^{n} T_j(X_i).$$

→ Pour estimer le paramètre inconnu $\theta \in \Theta$, on considère alors le système de d équations à d inconnues que l'on résout par rapport à la variable ϑ

$$\forall j \in \{1, \ldots, d\} \qquad n^{-1} \sum_{i=1}^{n} T_j(X_i) = e_j(\vartheta)\ .$$

- En supposant que ce système admette une solution unique $\hat{\theta}_n$, on appelle $\hat{\theta}_n$ l'*estimateur des moments* de θ associé aux fonctions T_j, $j = 1, \ldots, d$.

→L'estimateur des moments est donc égal à la valeur du paramètre ϑ pour laquelle les moments exacts et les moments empiriques sont égaux.

- Modèle de base pour les systèmes à événements discrets (arrivée à des services, occurrence de panne, survie)
- Inter-Arrivées $X_1 \ldots, X_n$ i.i.d. de loi exponentielle de paramètre $\theta \in \Theta = \mathbb{R}_+^*$:

$$p(\theta, x) = \theta \mathrm{e}^{-\theta x} \mathbb{1}_{\{x>0\}} \, .$$

- Propriété : loi sans mémoire - pour tout $a, b > 0$ et $\theta \in \mathbb{R}_+^*$,

$$\mathbb{P}_\theta \left(X_1 \geq a + b \,|\, X_1 \geq a \right) = \frac{\mathrm{e}^{-(a+b)\theta}}{\mathrm{e}^{-a\theta}} = \mathbb{P}_\theta(X_1 \geq b)$$

- Calcul des moments par intégrations par parties successives :

$$\mathbb{E}_\theta[X_1^k] = \frac{k!}{\theta^k} \, , \quad k \in \mathbb{N}^* \, .$$

Considérons les fonctions $T(x) = x$ et $\tilde{T}(x) = x^2$.

$$e(\theta) = \int T(x)\mathbb{Q}_\theta(\mathrm{d}x) = \int_0^{+\infty} x \quad \theta \exp(-\theta x)\mathrm{d}x = \frac{1}{\theta}$$

$$\tilde{e}(\theta) = \int \tilde{T}(x)\mathbb{Q}_\theta(\mathrm{d}x) = \int_0^{+\infty} x^2 \quad \theta \exp(-\theta x)\mathrm{d}x = \frac{2}{\theta^2} \ .$$

- Les estimateurs des moments associés sont les solutions respectives des équations

$$e(\vartheta) = \frac{1}{\vartheta} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{et} \quad \tilde{e}(\vartheta) = \frac{2}{\vartheta^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \ .$$

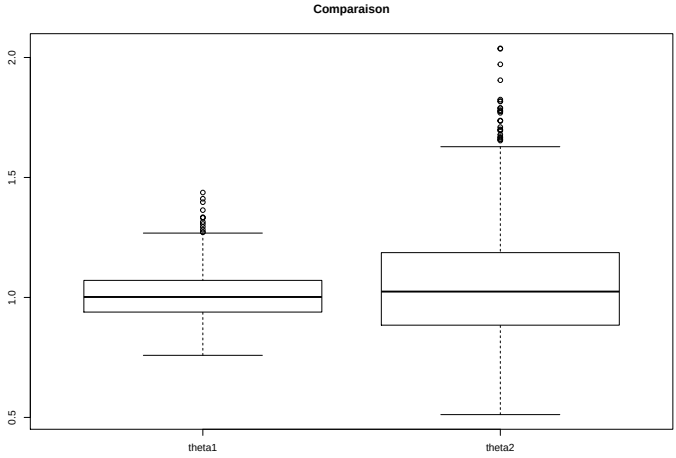


Figure – Comparaison des performances des estimateurs $\hat{\theta}_{n,1} = 1/(n^{-1}\sum_{i=1}^{n} X_i)$ et $\hat{\theta}_{n,2} = \left(2/(n^{-1}\sum_{i=1}^{n} X_i^2)\right)^{1/2}$

c. Méthode des moments - Mise en pratique - loi uniforme $[a, b]$

Modèle simple utilisé pour représenter une variable aléatoire bornée dans un intervalle

- Observations $X_1, \dots, X_n$ i.i.d. de loi uniforme sur $[a, b]$ où le paramètre est $\theta = (a, b) \in \Theta = \{(a, b) \in \mathbb{R}^2 : a < b\}$
- Densité : $p_\theta(x) = \frac{1}{b-a} \mathbb{1}_{\{a \le x \le b\}}$.
- Considérons les fonctions $T(x) = x$ et $\tilde{T}(x) = x^2$. On a les moments exacts :

$$e(\theta) = \mathbb{E}_\theta[X] = \frac{a+b}{2}, \qquad \tilde{e}(\theta) = \mathbb{E}_\theta[X^2] = \frac{b^3 - a^3}{3(b-a)} = \frac{a^2 + ab + b^2}{3} .$$

- Les estimateurs des moments sont définis comme solutions des équations

$$e(\vartheta) = \tfrac{a+b}{2} = \frac{1}{n} \sum_{i=1}^{n} X_i, \qquad \tilde{e}(\vartheta) = \tfrac{a^2+ab+b^2}{3} = \frac{1}{n} \sum_{i=1}^{n} X_i^2 .$$

- En posant $s = a + b$ et $p = ab$, on obtient

$$\hat{s}_n = 2\overline{X}_n,$$
$$\hat{p}_n = \hat{s}_n^2 - 3\,\overline{X^2}_n ,$$

et $\hat{a}_n \le \hat{b}_n$ sont les racines de $t^2 - \hat{s}_n t + \hat{p}_n = 0$ (calcul détaillé dans le polycopié).

⚠ Ici, plusieurs moments sont nécessaires pour identifier les deux paramètres a et b.

- **Résultat 1 :** (cas $d = 1$) Si pour tout θ on a

$$\mathbb{E}_\theta\left[T^2(X)\right] = \int_{\mathsf{X}} T^2(x)\mathbb{Q}_\theta(\mathrm{d}x) < \infty,$$

alors,

$$\mathbb{P}_\theta\left(\left|n^{-1}\sum_{j=1}^{n} T(X_j) - e(\theta)\right| \geq \varepsilon\right) \leq \frac{\sigma^2(\theta)}{n\varepsilon^2} \ .$$

où

$$\sigma^2(\theta) = \mathrm{Var}_\theta(T(X)) = \int \left\{T(x) - e(\theta)\right\}^2 \mathbb{Q}_\theta(\mathrm{d}x)$$

Donc si $\hat{\theta}_n$ EMM associé à T,

$$\mathbb{P}_\theta\left(\left|e(\hat{\theta}_n) - e(\theta)\right| \geq \varepsilon\right) \leq \frac{\sigma^2(\theta)}{n\varepsilon^2}.$$

- **Résultat 2 :** si $\mathbb{E}_\theta[|T(X_1)|] < \infty$, la loi (faible) des grands nombres donne $n^{-1}\sum_{j=1}^{n} T(X_j) \overset{\mathbb{P}_\theta-\mathrm{prob}}{\longrightarrow} e(\theta)$.
Donc si $\hat{\theta}_n$ EMM associé à T,

$$e(\hat{\theta}_n) \overset{\mathbb{P}_\theta-\mathrm{prob}}{\longrightarrow} e(\theta) \ .$$

Si e est injective et si e^{-1} est continue au point $e(\theta)$, alors $\hat{\theta}_n = e^{-1}\left(n^{-1}\sum_{i=1}^{n} T(X_i)\right) \overset{\mathbb{P}_\theta-\mathrm{prob}}{\longrightarrow} \theta$.

e. Méthode des moments - résumé et généralisation

Résumé :

- L'estimateur des moments basé sur les statistiques $\mathbf{T} = (T_1, \ldots, T_d)$ consiste à identifier les moments empiriques et les moments exacts.
- L'estimateur des moments est la solution d'un système de d équations à d inconnues

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\psi}(\theta, X_i) = 0 \,, \qquad \boldsymbol{\psi}(\theta, x) = \mathbf{T}(x) - \mathbb{E}_\theta[\mathbf{T}(X_1)] \,.$$

- Très souvent, pour un paramètre unidimensionnel, une seule fonction de moment suffit.
- Les EMM ont des propriétés de convergence qui découlent de la LGN et du TCL.

Extension

- Il est souvent intéressant de considérer des estimateurs définis comme solution de systèmes d'équations mais pour des fonctions d'estimation $\boldsymbol{\psi}$ plus générales.
- Une telle approche est appelée la méthode des Z-estimateurs.

Une partie de ces propriétés se démontre directement pour les Z-estimateurs.

Définition : Z-estimateur

- Soit $X_1, \ldots, X_n$ un n-échantillon du modèle $(\mathsf{X}, \mathcal{X}, \{\mathbb{Q}_\theta\ ,\ \theta \in \Theta\})$ avec $\Theta \subset \mathbb{R}^d$.
- Soient $\psi_j : \Theta \times \mathsf{X} \to \mathbb{R}$, $j = 1, \ldots, d$ des fonctions mesurables telles que, pour tout $\theta \in \Theta$ et tout j, $\mathbb{E}_\theta[|\psi_j(\theta, X_1)|] < \infty$ et

$$\mathbb{E}_\theta[\boldsymbol{\psi}(\theta, X_1)] = \mathbf{0}_{d\times 1}\ , \quad \boldsymbol{\psi} = (\psi_1, \ldots, \psi_d)^\top\ .$$

- On appelle Z-estimateur associé à $\boldsymbol{\psi}$ tout estimateur $\hat{\theta}_n$ vérifiant

$$\Psi_n(\hat{\theta}_n) = 0\ , \quad \text{où} \quad \Psi_n(\vartheta) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\psi}(\vartheta, X_i)$$

Exemple de Z-estimateur (non EMM) - Modèle de translation

- Soit $(X_1, \dots, X_n)$ un n-échantillon du modèle

$$\{q_\theta \cdot \mathrm{d}\mu, \qquad \theta \in \Theta = \mathbb{R}\}$$

avec $q_\theta(x) = f(x-\theta)$, où f est une densité de probabilité paire : $f(x) = f(-x)$.

- Exemple : loi gaussienne centrée réduite translatée de θ.
- Deux choix de fonctions d'estimation ψ sont possibles :

Moyenne empirique :

$$\psi(\theta, x) = x - \theta$$

Si $\int_{\mathbb{R}} |x| f(x) \mathrm{d}x < \infty$:

$$\begin{aligned}\mathbb{E}_\theta[X_1 - \theta] &= \int (x-\theta) f(x-\theta)\, \mathrm{d}x \\ &= \int z f(z)\, \mathrm{d}z = 0 \quad \text{(parité de } f\text{)}\end{aligned}$$

Z- estim : $\frac{1}{n}\sum_{i=1}^n \psi(\hat{\theta}_n, X_i) = 0$

Solution $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i$.

Médiane empirique :

$$\psi(\theta, x) = \mathrm{signe}(x - \theta)$$

$$\begin{aligned}\mathbb{E}_\theta[\mathrm{signe}(X-\theta)] &= \int \mathrm{signe}(x-\theta) f(x-\theta)\, \mathrm{d}x \\ &= \int \mathrm{signe}(z) f(z)\, \mathrm{d}z = 0\end{aligned}$$

Z- estim : $\frac{1}{n}\sum_{i=1}^n \psi(\hat{\theta}_n, X_i) = 0$

Solution $\hat{\theta}_n = \mathrm{med}(X_1, \dots, X_n)$: toute médiane empirique est un Z-estimateur.

Ainsi, la moyenne empirique est un estimateur des moments, tandis que la médiane empirique est un Z-estimateur mais pas un estimateur des moments.

a. Intuition de l'EMV - Exemple 1

💡 Le terme de vraisemblance est transparent : on se demande quelle valeur du paramètre rendrait nos observations les plus *vraisemblables*.

Considérons un modèle gaussien $\{\mathcal{N}(\mu, 1)^{\otimes n}, \mu \in \mathbb{R}\}$
J'observe les valeurs : $X_1 = 180$, $X_2 = 170$, $X_3 = 155, 172, 183, 194, 164, \ldots$

Est-il vraisemblable que $\mu = 10$? que $\mu = 200$?
Quel paramètre rend l'observation la plus vraisemblable ?

Si j'ai une seule observation ?

Si j'ai plusieurs observations ?

Considérons le modèle uniforme $\{\mathcal{U}([0,\theta])^{\otimes n}, \theta \in \mathbb{R}_+^*\}$
J'observe les valeurs : $X_1 = 8$, $X_2 = 7$, ... 11, 17, 2, 5, 14, ...

Est-il vraisemblable que $\theta = 5$? que $\theta = 17$? que $\theta = 100$?
Quel paramètre rend l'observation la plus vraisemblable ?

Si j'ai une seule observation ?

Si j'ai plusieurs observations ?

Modèle à densité $\rightarrow$

Définition : Vraisemblance

Soit $(X_1, \ldots, X_n)$ un n-échantillon du modèle

$$(\mathsf{X}, \mathcal{X}, \{q_\theta \cdot \mu\,,\ \theta \in \Theta\})$$

On appelle fonction de vraisemblance (ou vraisemblance) associée au n-échantillon $(X_1, \ldots, X_n)$ l'application

$$\vartheta \in \Theta \mapsto \mathrm{L}_n(\vartheta, X_1, \ldots, X_n) := \prod_{i=1}^{n} q_\vartheta(X_i).$$

On appelle estimateur du maximum de vraisemblance tout estimateur $\hat{\theta}_n^{\mathrm{MV}}$ satisfaisant

$$\hat{\theta}_n^{\mathrm{MV}} \in \arg\max_{\vartheta \in \Theta} \mathrm{L}_n(\vartheta, X_1, \ldots, X_n).$$

Définition : Log-Vraisemblance

Soit $(X_1, \ldots, X_n)$ un n-échantillon du modèle

$$(\mathsf{X}, \mathcal{X}, \{q_\theta \cdot \mu\,,\ \theta \in \Theta\})$$

où μ est une mesure σ-finie. La log-vraisemblance normalisée est l'application

$$\vartheta \mapsto \ell_n(\vartheta, X_1, \ldots, X_n) = \frac{1}{n} \log \mathrm{L}_n(\vartheta, X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} \log q_\vartheta(X_i),$$

En posant $\log 0 = -\infty$, on pourra parler de log-vraisemblance (normalisée) en toute généralité. Bien entendu, on a :

$$\hat{\theta}_n^{\mathrm{MV}} \in \arg\max_{\vartheta \in \Theta} \ell_n(\vartheta, X_1, \ldots, X_n).$$

c. Exemples : Loi de Poisson

Pour $\theta \in \Theta = \mathbb{R}_+^*$, nous appelons loi de Poisson de paramètre θ, la loi de densité par rapport à la mesure de comptage

$$q_\theta(x) = \mathrm{e}^{-\theta} \frac{\theta^x}{x!} \ , \quad x \in \mathbb{N}, \theta \in \mathbb{R}_+^* \ .$$

Soit $X_1, \ldots, X_n$ un n-échantillon d'une loi de Poisson de paramètre θ. La vraisemblance s'écrit alors, pour tout $\vartheta > 0$,

$$\mathrm{L}_n(\vartheta, X_1, \ldots, X_n) = \prod_{i=1}^n \mathrm{e}^{-\vartheta} \frac{\vartheta^{X_i}}{X_i!} = \frac{1}{\prod_{i=1}^n X_i!} \exp\left(-n\vartheta + n \log(\vartheta) \bar{X}_n\right) \ .$$

$$\ell_n(\vartheta, X_1, \ldots, X_n) = \frac{1}{n} \log \mathrm{L}_n(\vartheta, X_1, \ldots, X_n) = \frac{-\log(\prod_{i=1}^n X_i!)}{n} - \vartheta + \log(\vartheta) \bar{X}_n \ .$$

En isolant ce qui ne dépend pas de ϑ et en considérant ℓ_n plutôt que L_n, on constate que l'estimateur du maximum de vraisemblance (fonction concave, dérivée nulle) est donné par [1]

$$\bar{X}_n := n^{-1} \sum_{i=1}^n X_i \ .$$

💡 Une fois encore, on remarque que l'EMV correspond à un EMM simple !
C'est un phénomène typique, par exemple pour la famille exponentielle (canonique) - cf exo 3 PC 2 !

1. Si toutes les observations sont nulles, la fonction décroît sur $\mathbb{R}_+^*$ et l'estimateur n'est pas défini.

Équations de vraisemblance

Si le maximum de $\vartheta \mapsto \ell_n(\vartheta)$ est atteint en un point intérieur de Θ où ℓ_n est différentiable, alors

$$\nabla_\vartheta \ \ell_n(\vartheta, X_1, \ldots, X_n)|_{\vartheta=\hat{\theta}_n^{\mathrm{MV}}} = 0 \ .$$

Ceci fournit un système de d équations si $\Theta \subset \mathbb{R}^d$.

Définition : Équations de vraisemblance

Le système d'équations $\nabla_\vartheta \ \ell_n(\vartheta, X_1, \ldots, X_n)|_{\vartheta=\hat{\theta}_n^{\mathrm{MV}}} = 0$ est appelé système des équations de vraisemblance.

L'estimateur du maximum de vraisemblance est donc obtenu soit

- en résolvant un problème d'optimisation,
- en résolvant un système d'équations, le plus souvent non linéaires (en vérifiant que la solution retenue est bien un maximum global !).

Sauf dans les modèles simples (voir le modèle statistique des lois de Poisson par exemple), ces estimateurs ne sont pas explicites.

c. Exemples : moyenne et variance d'une gaussienne

Soit $X_1, \dots, X_n$ un n-échantillon de loi $\mathcal{N}(\mu, \sigma^2)$ de densité

$$f(\theta, x) = (2\pi\sigma^2)^{-1/2} \exp\big(-\tfrac{1}{2\sigma^2}(x-\mu)^2\big)\,, \quad \theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+^*\,.$$

La log-vraisemblance normalisée associée s'écrit

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

Les dérivées partielles s'écrivent

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

et les équations de vraisemblance consistent à les annuler toutes les deux. Pour $n \geq 2$, l'E.M.V. est unique et donné par

$$\hat{\theta}_n^{\mathrm{MV}} := \left(\overline{X}_n, \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2\right).$$

c. Exemples : modèle non-différentiable - Modèle uniforme

- Soit $(X_1, \ldots, X_n)$ un n-échantillon de la loi uniforme sur $[0, \theta]$, où $\theta \in \Theta = \mathbb{R}_+^*$ de densité par rapport à la mesure de Lebesgue :

$$f(\theta, x) = \frac{1}{\theta} \mathbb{1}_{[0,\theta]}(x).$$

- Les observations étant positives $\mathbb{P}_\theta$-presque sûrement, la condition « $X_i \in [0, \vartheta]$ pour tout i » équivaut à $X_{n:n} \leq \vartheta$ et la vraisemblance s'écrit

$$\mathrm{L}_n(\vartheta, X_1, \ldots, X_n) = \frac{1}{\vartheta^n} \prod_{i=1}^{n} \mathbb{1}_{[0,\vartheta]}(X_i) = \vartheta^{-n} \mathbb{1}_{[X_{n:n}, +\infty[}(\vartheta),$$

où

$$X_{n:n} = \max_{i=1,\ldots,n} X_i.$$

- La valeur maximale de $\mathrm{L}_n(\vartheta, X_1, \ldots, X_n)$ est obtenue pour $\vartheta = X_{n:n}$:

$$\hat{\theta}_n^{\mathrm{MV}} := X_{n:n}.$$

- En revanche, la fonction L_n n'est pas dérivable en $X_{n:n}$, et la log-vraisemblance vaut $-\infty$ sur $]0, X_{n:n}[$.

Résumé EMV

Méthode : Explicitation de l'EMV (si possible)

1. Écrire la densité q_ϑ par rapport à la mesure dominante, puis la vraisemblance $\mathrm{L}_n(\vartheta) = \prod_{i=1}^n q_\vartheta(X_i)$
2. **Cas 1 :** modèles « réguliers » (gaussien, de Poisson, exponentiel, Gamma, de Bernoulli) :
   a. Écrire la log-vraisemblance
   b. Trouver les points critiques
   (résoudre le système si besoin)
   c. Vérifier que le point critique est un maximum.
3. **Cas 2 :** Modèles non réguliers (Uniforme par exemple)
   a. Déterminer le maximum directement.

   ⚠ Ne pas essayer de différencier ! ! !

What's next :

1. Quelques warnings
2. Éléments théoriques

Warning 1 : l'E.M.V. n'est pas toujours défini

- Soit $(X_1, \ldots, X_n)$ un n-échantillon de variables aléatoires réelles de loi de densité $\{f_0(x-\theta), \theta \in \Theta = \mathbb{R}\}$ par rapport à la mesure de Lebesgue où

$$f_0(x) = \frac{\mathrm{e}^{-\frac{|x|}{2}}}{2\sqrt{2\pi|x|}}, \quad x \in \mathbb{R},$$

- La fonction de vraisemblance s'écrit

$$\mathrm{L}_n(\vartheta, X_1, \ldots, X_n) = \prod_{i=1}^{n} f_0(X_i - \vartheta).$$

- Pour tout $i = 1, \ldots, n$, on a

$$\lim_{\vartheta \to X_i} \mathrm{L}_n(\vartheta, X_1, \ldots, X_n) = +\infty.$$

- Pour cette expérience statistique, l'estimateur du maximum de vraisemblance n'est pas défini.

Warning 2 : l' E.M.V. n'est pas toujours unique

- Soit $(X_1, \ldots, X_n)$ un n-échantillon de la loi de Laplace de paramètre $\theta \in \Theta = \mathbb{R}$, dont la densité par rapport à la mesure de Lebesgue est donnée par

$$q_\theta(x) = \frac{1}{2} \exp\big(- |x - \theta|\big),$$

- La log-vraisemblance normalisée est donnée par

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

- Maximiser

$$\vartheta \mapsto \ell_n(\vartheta, X_1, \ldots, X_n)$$

revient à minimiser la fonction

$$\vartheta \mapsto \sum_{i=1}^{n} \big|X_i - \vartheta\big| \, .$$

Warning 2 : l'E.M.V. n'est pas toujours unique

- Cette fonction est dérivable presque partout, de dérivée

$$-\sum_{i=1}^{n} \text{signe}(X_i - \vartheta) \ .$$

  La dérivée (définie presque partout) est constante par morceaux.
- Si n est impair, la dérivée s'annule en un point unique,

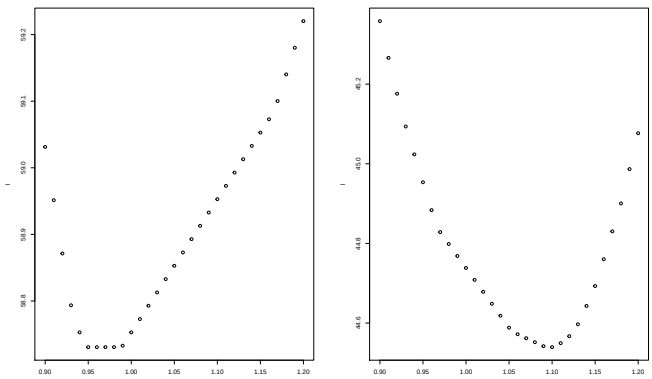
$$X_{(n+1)/2:n}$$

  , qui est la médiane empirique et l'unique maximum de la vraisemblance, où $X_{1:n} \leq \ldots \leq X_{n:n}$ désigne la statistique d'ordre associée à l'échantillon (voir les statistiques d'ordre, chapitre 1 du polycopié)
- Si n est pair, il y a une infinité de solutions : tout point de l'intervalle fermé

$$\left[X_{n/2:n}, X_{n/2+1:n}\right]$$

  est un estimateur du maximum de vraisemblance.

Pour le modèle de Laplace translatée de paramètre $\theta = 1$, fonction $\vartheta \mapsto \sum_{i=1}^{n} |X_i - \vartheta|$, dont les minimiseurs sont les estimateurs du maximum de vraisemblance, pour $n = 50$ (à gauche) et $n = 51$ (à droite). À gauche, le plateau de minimiseurs est visible.



L'E.M.V. est invariant par changement de paramétrisation

- Soit $\{\mathbb{P}_\theta \,:\, \theta \in \Theta\}$ un modèle paramétrique dominé.
- G une transformation bijective de Θ sur $G(\Theta) = \Xi$.
- Pour tout $\xi \in \Xi$, posons $\mathbb{Q}_\xi = \mathbb{P}_{G^{-1}(\xi)}$ et considérons le modèle statistique $\{\mathbb{Q}_\xi \,:\, \xi \in \Xi\}$.

Proposition : Invariance par reparamétrisation

Si $\hat{\theta}_n^{\mathrm{MV}}$ est un E.M.V. pour le modèle $\{\mathbb{P}_\theta \,:\, \theta \in \Theta\}$, alors $\hat{\xi}_n = G(\hat{\theta}_n^{\mathrm{MV}})$ est un E.M.V. pour le modèle $\{\mathbb{Q}_\xi \,:\, \xi \in \Xi\}$.

Exemple : Modèle gaussien : variance ou écart type

- Considérons le modèle gaussien paramétré en moyenne-variance $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+^*$.
- L'E.M.V. pour ce modèle est $\hat{\mu}_n = \bar{X}_n = n^{-1} \sum_{i=1}^n X_i$, $\hat{\sigma}_n^2 = n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$
- Considérons le modèle gaussien paramétré en moyenne-écart type $\xi = (\mu, \sigma) \in \Xi = \mathbb{R} \times \mathbb{R}_+^*$.
- La fonction $G : \Theta \to \Xi$, $(\mu, \sigma^2) \mapsto (\mu, \sqrt{\sigma^2})$ est une bijection de Θ sur Ξ .
- L'E.M.V. pour ξ est $\hat{\mu} = \bar{X} = n^{-1} \sum_{i=1}^n X_i$, $\hat{\sigma} = \left(n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2\right)^{1/2}$.

Comprendre le M.V. - Intuition

- Soit $X_1, \ldots, X_n$ un n-échantillon du modèle stat. $(\mathsf{X}, \mathcal{X}, \{q_\theta \cdot \mu\,,\ \theta \in \Theta\})$.
- Dans la suite, pour ne pas se perdre :
  $\theta^\star \to$ paramètre « hypothèse » : la valeur sous laquelle les observations sont tirées, on travaille sous $\mathbb{P}_{\theta^\star}$
  $\vartheta \to$ paramètre libre : la variable par rapport à laquelle on exprime $\mathrm{L}_n(\vartheta)$ et on maximise
- La loi (faible) des grands nombres (sous hypothèse d'intégrabilité) montre que, pour tous $\vartheta, \theta^\star \in \Theta$, sous $\mathbb{P}_\theta^\star$

$$\ell_n(\vartheta) = n^{-1} \sum_{i=1}^{n} \log q_\vartheta(X_i) \overset{\mathbb{P}_{\theta^\star}-\text{prob}}{\longrightarrow} \mathbb{E}_{\theta^\star}[\log q_\vartheta(X_1)]$$

  L'espérance est sous la loi des observations, *i.e.* la loi de densité $q(\theta^\star, .)$.
- La suite $\{\hat{\theta}_n^{\mathrm{MV}}, n \in \mathbb{N}\}$ maximise $\vartheta \mapsto \ell_n(\vartheta, X_1, \ldots, X_n)$.

- Intuition : la suite $\{\hat{\theta}_n^{\mathrm{MV}}, n \in \mathbb{N}\}$ converge vers

$$\arg\max_{\vartheta \in \Theta} \mathbb{E}_{\theta^\star}[\log q_\vartheta(X_1)]\ .$$

On pose

$$\vartheta \to M_{\theta^\star}(\vartheta) := \mathbb{E}_{\theta^\star}[\log q_\vartheta(X)]$$

La fonction $M_{\theta^\star}$ est la limite de la log-vraisemblance ; nous allons montrer qu'elle admet un maximum unique, atteint en $\theta^\star$.

Vraie valeur $\theta^\star$ — **0,62**

Taille de l'échantillon n — **1**

Valeur candidate ϑ — **0,32**

↻ **Résimuler**

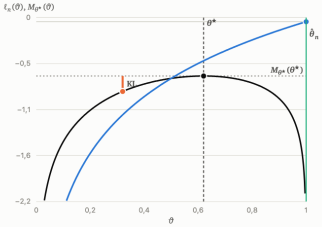
☑ Tracer, à partir des observations, les log-vraisemblances normalisées ℓ_n des tailles successives et marquer leurs maxima $\hat{\theta}_n = \bar{X}_n$

Observations $x_1, x_2, \ldots, x_n$ ($n = 1$; 40 premières valeurs) — nombre de 1 observés : $S_n = 1$

1

La log-vraisemblance normalisée ℓ_n et sa limite $M_{\theta^\star}$

En trait noir, $M_{\theta^\star}$; au point ϑ, l'écart vertical entre le maximum $M_{\theta^\star}(\theta^\star)$ et $M_{\theta^\star}(\vartheta)$ est la divergence de Kullback-Leibler. Les courbes des tailles précédentes restent tracées, de la plus pâle ($n = 1$) à la plus soutenue (taille courante), sur le même échantillon emboîté.

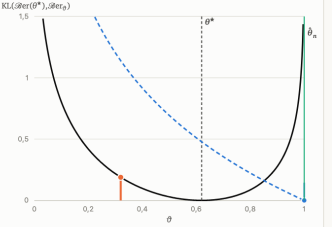


$M_{\theta^\star}(\theta^\star)$, maximum de $M_{\theta^\star}$
−0,664

$M_{\theta^\star}(\vartheta)$
−0,853

La divergence $\vartheta \mapsto \mathrm{KL}(\mathcal{B}\mathrm{er}(\theta^\star), \mathcal{B}\mathrm{er}(\vartheta))$

$\mathrm{KL}(\mathcal{B}\mathrm{er}(\theta^\star), \mathcal{B}\mathrm{er}(\vartheta)) = \theta^\star \log \frac{\theta^\star}{\vartheta} + (1-\theta^\star) \log \frac{1-\theta^\star}{1-\vartheta} = M_{\theta^\star}(\theta^\star) - M_{\theta^\star}(\vartheta)$, positive, nulle en $\vartheta = \theta^\star$ seulement.



$\mathrm{KL}(\mathcal{B}\mathrm{er}(\theta^\star), \mathcal{B}\mathrm{er}(\vartheta))$
0,189

$\bar{X}_n$, maximum de ℓ_n
1

$\ell_n(\bar{X}_n) - \ell_n(\vartheta)$
1,139

Vraie valeur $\theta^\star$ — **0,62**

Taille de l'échantillon n — **4**

Valeur candidate ϑ — **0,32**

↻ **Résimuler**

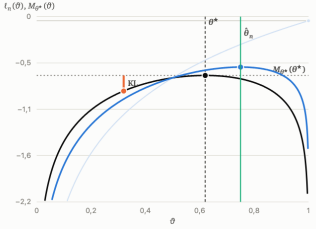
☑ Tracer, à partir des observations, les log-vraisemblances normalisées ℓ_n des tailles successives et marquer leurs maxima $\hat{\theta}_n = \bar{X}_n$

Observations $x_1, x_2, \dots, x_n$ ($n = 4$; 40 premières valeurs) — nombre de 1 observés : $S_n = 3$

`1 1 0 1`

La log-vraisemblance normalisée ℓ_n et sa limite $M_{\theta^\star}$

En trait noir, $M_{\theta^\star}$; au point ϑ, l'écart vertical entre le maximum $M_{\theta^\star}(\theta^\star)$ et $M_{\theta^\star}(\vartheta)$ est la divergence de Kullback-Leibler. Les courbes des tailles précédentes restent tracées, de la plus pâle ($n = 1$) à la plus soutenue (taille courante), sur le même échantillon emboîté.

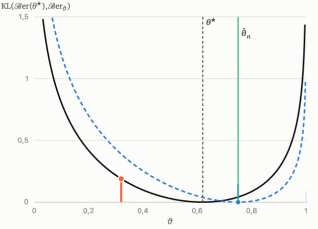


$M_{\theta^\star}(\theta^\star)$, maximum de $M_{\theta^\star}$
−0,664

$M_{\theta^\star}(\vartheta)$
−0,853

La divergence $\vartheta \mapsto \mathrm{KL}(\mathcal{B}\mathrm{er}(\theta^\star), \mathcal{B}\mathrm{er}(\vartheta))$

$\mathrm{KL}(\mathcal{B}\mathrm{er}(\theta^\star), \mathcal{B}\mathrm{er}(\vartheta)) = \theta^\star \log \frac{\theta^\star}{\vartheta} + (1-\theta^\star) \log \frac{1-\theta^\star}{1-\vartheta} = M_{\theta^\star}(\theta^\star) - M$
, positive, nulle en $\vartheta = \theta^\star$ seulement.



$\mathrm{KL}(\mathcal{B}\mathrm{er}(\theta^\star), \mathcal{B}\mathrm{er}(\vartheta))$
0,189

$\bar{X}_n$, maximum de ℓ_n
0,75

$\ell_n(\bar{X}_n) - \ell_n(\vartheta)$
0,389

Vraie valeur $\theta^\star$ — **0,62** | Taille de l'échantillon n — **16** | Valeur candidate ϑ — **0,32**

↻ Résimuler

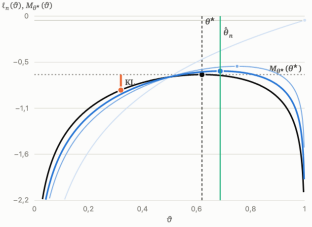
☑ Tracer, à partir des observations, les log-vraisemblances normalisées ℓ_n des tailles successives et marquer leurs maxima $\hat{\theta}_n = \bar{X}_n$

Observations $x_1, x_2, \dots, x_n$ ($n = 16$; 40 premières valeurs) — nombre de 1 observés : $S_n = 11$

`1 1 0 1 1 1 1 0 0 1 0 1 1 1 1 0`

La log-vraisemblance normalisée ℓ_n et sa limite $M_{\theta^\star}$

En trait noir, $M_{\theta^\star}$; au point ϑ, l'écart vertical entre le maximum $M_{\theta^\star}(\theta^\star)$ et $M_{\theta^\star}(\vartheta)$ est la divergence de Kullback-Leibler. Les courbes des tailles précédentes restent tracées, de la plus pâle ($n = 1$) à la plus soutenue (taille courante), sur le même échantillon emboîté.

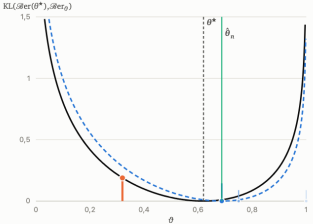


$M_{\theta^\star}(\theta^\star)$, maximum de $M_{\theta^\star}$: **−0,664**

$M_{\theta^\star}(\vartheta)$: **−0,853**

La divergence $\vartheta \mapsto \mathrm{KL}(\mathcal{B}\mathrm{er}(\theta^\star), \mathcal{B}\mathrm{er}(\vartheta))$

$\mathrm{KL}(\mathcal{B}\mathrm{er}(\theta^\star), \mathcal{B}\mathrm{er}(\vartheta)) = \theta^\star \log \frac{\theta^\star}{\vartheta} + (1-\theta^\star) \log \frac{1-\theta^\star}{1-\vartheta} = M_{\theta^\star}(\theta^\star) - M_{\theta^\star}(\vartheta)$, positive, nulle en $\vartheta = \theta^\star$ seulement.



$\mathrm{KL}(\mathcal{B}\mathrm{er}(\theta^\star), \mathcal{B}\mathrm{er}(\vartheta))$: **0,189**

$\bar{X}_n$, maximum de ℓ_n : **0,688**

$\ell_n(\bar{X}_n) - \ell_n(\vartheta)$: **0,283**

Vraie valeur $\theta^\star$ — **0,62**

Taille de l'échantillon n — **1 024**

Valeur candidate ϑ — **0,32**

↻ **Résimuler**

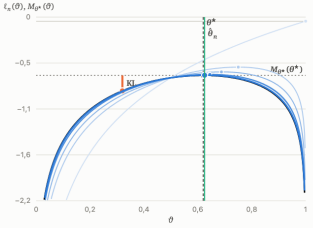
☑ Tracer, à partir des observations, les log-vraisemblances normalisées ℓ_n des tailles successives et marquer leurs maxima $\hat{\theta}_n = \bar{X}_n$

Observations $x_1, x_2, \dots, x_n$ ($n = 1\,024$; 40 premières valeurs) — nombre de 1 observés : $S_n = 640$

`1 1 0 1 1 1 1 0 0 1 0 1 1 1 1 0 0 1 1 1 0 0 1 0 0 0 0 0 0 0 1 1 1 1 1 0 1 1 1 0 …`

La log-vraisemblance normalisée ℓ_n et sa limite $M_{\theta^\star}$

En trait noir, $M_{\theta^\star}$; au point ϑ, l'écart vertical entre le maximum $M_{\theta^\star}(\theta^\star)$ et $M_{\theta^\star}(\vartheta)$ est la divergence de Kullback-Leibler. Les courbes des tailles précédentes restent tracées, de la plus pâle ($n = 1$) à la plus soutenue (taille courante), sur le même échantillon emboîté.

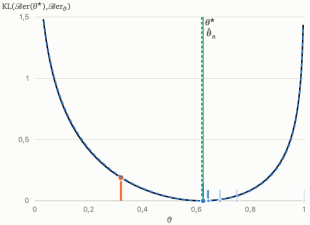


$M_{\theta^\star}(\theta^\star)$, maximum de $M_{\theta^\star}$
−0,664

$M_{\theta^\star}(\vartheta)$
−0,853

La divergence $\vartheta \mapsto \mathrm{KL}(\mathcal{B}\mathrm{er}(\theta^\star), \mathcal{B}\mathrm{er}(\vartheta))$

$\mathrm{KL}(\mathcal{B}\mathrm{er}(\theta^\star), \mathcal{B}\mathrm{er}(\vartheta)) = \theta^\star \log \frac{\theta^\star}{\vartheta} + (1-\theta^\star) \log \frac{1-\theta^\star}{1-\vartheta} = M_{\theta^\star}(\theta^\star) - M_{\theta^\star}(\vartheta)$, positive, nulle en $\vartheta = \theta^\star$ seulement.



$\mathrm{KL}(\mathcal{B}\mathrm{er}(\theta^\star), \mathcal{B}\mathrm{er}(\vartheta))$
0,189

$\bar{X}_n$, maximum de ℓ_n
0,625

$\ell_n(\bar{X}_n) - \ell_n(\vartheta)$
0,195

Divergence de Kullback-Leibler

On décompose :

$$\begin{aligned}\vartheta \to M_\theta^\star(\vartheta) &:= \mathbb{E}_{\theta^\star}[\log q_\vartheta(X_1)] \\ &= \mathbb{E}_{\theta^\star}\left[\log \frac{q_\vartheta(X_1)}{q_{\theta^\star}(X_1)}\right] + \mathbb{E}_{\theta^\star}[\log q_{\theta^\star}(X_1)] \ . \\ &= -\operatorname{KL}(\mathbb{Q}_{\theta^\star}, \mathbb{Q}_\vartheta) + \mathbb{E}_{\theta^\star}[\log q_{\theta^\star}(X_1)] \ , \quad \text{avec}\end{aligned}$$

Définition : Divergence de Kullback-Leibler

Soient $\mathbb{P}_0$ et $\mathbb{P}_1$ deux probabilités définies sur un espace mesurable $(\mathsf{X}, \mathcal{X})$ admettant des densités f_0 et f_1 par rapport à une mesure σ-finie μ

On appelle divergence de Kullback–Leibler entre les distributions $\mathbb{P}_0$ et $\mathbb{P}_1$ la quantité

$$\operatorname{KL}(\mathbb{P}_0, \mathbb{P}_1) := \int_{\{x\,:\, f_0(x)>0\}} f_0(x) \log \frac{f_0(x)}{f_1(x)} \mu(\mathrm{d}x) \ ,$$

avec la convention $\log(1/0) = +\infty$.

La divergence de Kullback-Leibler est toujours définie mais vaut éventuellement $+\infty$ (voir le théorème sur la divergence de Kullback-Leibler, compléments du chapitre 2).

La divergence s'exprime comme une espérance :

$$\operatorname{KL}(\mathbb{P}_0, \mathbb{P}_1) = \mathbb{E}_{\mathbb{P}_0}\left[\log \frac{f_0(X)}{f_1(X)}\right] \quad \text{où } X \sim \mathbb{P}_0$$

Théorème : Propriété KL

Soient $\mathbb{P}_0$ et $\mathbb{P}_1$ deux probabilités de densités f_0 et f_1 par rapport à μ. La divergence $\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_1)$ est positive, et elle est nulle si et seulement si $\mathbb{P}_0 = \mathbb{P}_1$, c'est-à-dire si et seulement si $f_0 = f_1$ μ-presque partout.

Si le modèle est identifiable : pour tous $(\vartheta, \theta^\star) \in \Theta^2$, $q_\vartheta(\cdot) = q_{\theta^\star}(\cdot)$ μ-presque partout si et seulement si $\vartheta = \theta^\star$ alors, sous l'hypothèse d'intégrabilité précédente, la fonction $M_{\theta^\star}$, limite de la log-vraisemblance, admet un maximum unique, atteint en $\theta^\star$!

Preuve

- La fonction $s \mapsto -\log(s)$ est strictement convexe sur $\mathbb{R}_+^*$.

$$\vartheta \to M_{\theta^\star}(\vartheta) = -\mathrm{KL}(\mathbb{P}_{\theta^\star}, \mathbb{P}_\vartheta) + \mathbb{E}_{\theta^\star}\left[\log q_{\theta^\star}(X)\right],$$

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_1) &= \mathbb{E}_{\mathbb{P}_0}\left[-\log \frac{f_1(X)}{f_0(X)}\right] \\
&\geq -\log\left(\mathbb{E}_{\mathbb{P}_0}\left[\frac{f_1(X)}{f_0(X)}\right]\right) = -\log\left(\int f_0(x)\left(\frac{f_1(x)}{f_0(x)}\right)\mu(\mathrm{d}x)\right) = 0 \\
&\geq 0.
\end{aligned}$$

- Égalité si et seulement si $f_1 = c_0 f_0$ $\mathbb{P}_0$-p.s., donc si $f_1 = f_0$ $\mathbb{P}_0$-p.s.

Théorème : Inégalité de Jensen

Soit I un intervalle ouvert de $\mathbb{R}$, f une fonction convexe sur I, Y une variable aléatoire réelle telle que $\mathbb{P}(Y \in \mathrm{I}) = 1$ et $\mathbb{E}[|Y|] < \infty$. Alors

$$f(\mathbb{E}[Y]) \leq \mathbb{E}[f(Y)] \, .$$

Si la fonction f est strictement convexe, alors l'inégalité est stricte dès que la variable aléatoire Y n'est pas presque sûrement constante (i.e. $\mathbb{P}(Y \neq \mathbb{E}[Y]) > 0$).

Résumé et généralisation : l'E.M.V. est un M-estimateur

Résumé :

- L'E.M.V. est défini comme la solution d'un problème de maximisation. C'est un cas particulier d'une très vaste classe de méthodes, où l'estimateur est obtenu comme la solution d'un problème d'optimisation.
- Soit $(X_1, \dots, X_n)$ un n-échantillon du modèle $(\mathsf{X}, \mathcal{X}, \{\mathbb{Q}_\theta\ ,\ \theta \in \Theta\})$ et $m : \Theta \times \mathsf{X} \to \mathbb{R}$, $(\vartheta, x) \mapsto m(\vartheta, x)$ une fonction telle que, pour tous $\vartheta, \theta^\star \in \Theta$,

$$\mathbb{E}_{\theta^\star}[|m(\vartheta, X_1)|] < \infty$$

.

- On a vu comment le calculer, ses propriétés, le lien avec la KL.

Généralisation

- Pour tout $\theta^\star \in \Theta$, posons

$$M_{\theta^\star} : \Theta \to \mathbb{R}\ , \qquad \vartheta \mapsto M_{\theta^\star}(\vartheta) := \mathbb{E}_{\theta^\star}[m(\vartheta, X_1)]\ .$$

- Hypothèse Pour tout $\theta^\star \in \Theta$, la fonction $M_{\theta^\star}$ a un maximum au point $\theta^\star$.
- Pour l'E.M.V., $m(\vartheta, x) = \log q_\vartheta(x)$ et la fonction limite est $M_{\theta^\star}(\vartheta) = M_{\theta^\star}(\theta^\star) - \mathrm{KL}(\mathbb{Q}_{\theta^\star}, \mathbb{Q}_\vartheta)$, dont le maximum est atteint en $\theta^\star$ par positivité de la divergence.
- On peut très bien imaginer utiliser d'autres fonctions que la $\log$-vraisemblance pour définir nos estimateurs...

- Idée La fonction $M_{\theta^\star}$ n'est pas connue mais nous pouvons l'estimer par

$$\vartheta \mapsto M_n(\vartheta) = n^{-1} \sum_{i=1}^{n} m(\vartheta, X_i) \ .$$

- Par la loi des grands nombres, pour tous $\vartheta, \theta^\star \in \Theta$,

$$M_n(\vartheta) \overset{\mathbb{P}_{\theta^\star}-\text{prob}}{\longrightarrow} M_{\theta^\star}(\vartheta)$$

Définition : M-estimateur

Nous appelons M-estimateur toute solution du problème d'optimisation

$$\hat{\theta}_n \in \underset{\vartheta \in \Theta}{\arg\max}\, M_n(\vartheta) \ .$$

Deux exemples dans le modèle de translation : $m(\vartheta, x) = -(x - \vartheta)^2$ donne la moyenne empirique (moindres carrés), $m(\vartheta, x) = -|x - \vartheta|$ donne la médiane empirique.

Conclusion

→ Objectifs :

1. Décrire les deux grandes techniques d'estimation (Moments et Maximum de Vraisemblance)
2. Comprendre leur intuition et les résultats génériques que nous pouvons obtenir
3. Avancé : Notion de M et Z-estimateurs.

→ 2 techniques : pour chacune :
Intuition **Définition** **Exemples** **Propriétés** **Généralisation***
En PC :

1. Exercice 1 : modèle de régression linéaire gaussien !
2. Exercice 2 : estimation d'un coefficient de mélange.
3. Exercice 3 : famille exponentielle, EMV = un EMM.

Théorème : Théorème de Cochran

Soit X un vecteur aléatoire gaussien de $\mathbb{R}^n$ de loi $\mathcal{N}\left(0_{\mathbb{R}^n}, \sigma^2 \mathrm{Id}_n\right)$. Soit F un sev de $\mathbb{R}^n$ de dimension d, $F^\perp$ son orthogonal et $P_F, P_{F^\perp}$ les matrices des projections orthogonales sur F et $F^\perp$. Alors :

- les vecteurs aléatoires $P_F X, P_{F^\perp} X$ sont indépendants et de lois respectives $\mathcal{N}\left(0_{\mathbb{R}^n}, \sigma^2 P_F\right), \mathcal{N}\left(0_{\mathbb{R}^n}, \sigma^2 P_{F^\perp}\right)$;
- les variables aléatoires réelles $\|P_F X\|^2, \|P_{F^\perp} X\|^2$ sont indépendantes et de lois respectives $\sigma^2\chi^2(d), \sigma^2\chi^2(n-d)$.

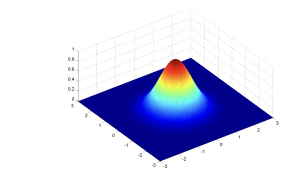


Figure – Densité d'une gaussienne isotrope $\mathcal{N}(0, \sigma^2 \mathrm{I}_2)$

Schéma de preuve :

1. Si $F = \mathrm{Vect}(e_1, \ldots, e_d)$, $P_F X = (X_1, \ldots, X_d, 0, \ldots, 0)$ et $P_{F^\perp} X = (0, \ldots, 0, X_{d+1}, \ldots, X_n)$
   Le théorème est évident.
2. Soit $(f_1, \ldots, f_d, f_{d+1}, \ldots, f_n)$ une base orthonormée adaptée à $F \oplus F^\perp$. Soit U la matrice de changement de base de $(e_1, \ldots, e_n)$ à $(f_1, \ldots, f_n)$:
   1. $UX \sim \mathcal{N}(0, \sigma^2 U \mathrm{I}_n U^T) = \mathcal{N}(0, \sigma^2 \mathrm{I}_n)$
   2. On peut se ramener au cas précédent par changement de base.

Organisation

1. Les groupes de DM sont les groupes d'explorations numériques. Une première composition sert pour le DM 1 et l'EN 1, une seconde pour le DM 2, l'EN 2 et l'EN 3. Les groupes sont visibles sur Moodle.
2. Le DM 1 (régression linéaire multiple) est en ligne depuis le lundi 31 août, sur Moodle et sur la page du cours ; à rendre le mercredi 23 septembre (un seul rendu par groupe).
3. La feuille de consignes des DM et EN (objectifs, travail en groupe, modalités de rendu) est en ligne : merci d'en prendre connaissance, c'est important ! :)
4. L'EN 1 sera mise en ligne le lundi 14 septembre ; à rendre aussi le mercredi 23 septembre.
5. Un sondage pour les modalités de polycopié est en ligne, merci d'y participer.

6. Le tutorat a lieu mercredi de 16h à 17h (détails sur Moodle) – Lucas est génial !

My *2 cents* — comment travailler

1. Construire sur vos forces
2. Rien n'est difficile mais il y a du formalisme et du vocabulaire !
3. Ne prenez pas de retard !
4. Cours connu pour demander du travail - mais qui vous donnera d'excellentes bases !

Annexes

Définition : Lois Gamma

(i) Pour tout réel $p > 0$, on appelle loi Gamma réduite de paramètre de forme p (et l'on note $\mathrm{Gamma}(p)$) la loi sur $\mathbb{R}_+^*$ de densité

$$f_p(x) = \frac{1}{\Gamma(p)} \exp(-x)\ x^{p-1}, \quad x > 0\ .$$

(ii) Pour $\lambda > 0$, on appelle loi Gamma $\mathrm{Gamma}(p, \lambda)$, la loi de la variable aléatoire $X = Z/\lambda$ où Z suit une loi $\mathrm{Gamma}(p)$ et où λ est le paramètre d'intensité. La densité de la loi $\mathrm{Gamma}(p, \lambda)$ est donnée par

$$f_{p,\lambda}(x) = \frac{\lambda^p}{\Gamma(p)} \exp(-\lambda x)\ x^{p-1}\ , \quad x > 0\ .$$

La loi exponentielle d'intensité $\lambda > 0$, de densité $x \mapsto \lambda \mathrm{e}^{-\lambda x}$ sur $\mathbb{R}_+$, coïncide avec la loi $\mathrm{Gamma}(1, \lambda)$.
Si Z suit la loi $\mathrm{Gamma}(p)$, nous avons, pour tout $r > -p$,

$$\mathbb{E}[Z^r] = \frac{\Gamma(p+r)}{\Gamma(p)}\ . \tag{1}$$

Stabilité par addition

Lemme : Somme de variables Gamma indépendantes

Soit $(X_1, \dots, X_n)$ n variables aléatoires indépendantes distribuées suivant des lois $\mathrm{Gamma}(p_i, \theta)$ avec $\theta > 0$ et $p_i > 0$, $i \in \{1, \dots, n\}$. Alors, $\sum_{i=1}^{n} X_i$ est distribuée suivant une loi $\mathrm{Gamma}(\sum_{i=1}^{n} p_i, \theta)$.

La fonction génératrice des moments de X_i est donnée, pour $t < \theta$,

$$\begin{aligned}\mathbb{E}[\mathrm{e}^{tX_i}] &= \int_0^\infty \mathrm{e}^{tx} \frac{\theta^{p_i} x^{p_i-1} \mathrm{e}^{-\theta x}}{\Gamma(p_i)} \mathrm{d}x \\ &= \int_0^\infty \frac{\theta^{p_i} x^{p_i-1} \mathrm{e}^{-(\theta-t)x}}{\Gamma(p_i)} \mathrm{d}x \\ &= \frac{\theta^{p_i}}{(\theta-t)^{p_i}}\end{aligned}$$

Preuve, seconde étape

La fonction génératrice de la somme $S = X_1 + \cdots + X_n$ est donnée par

$$\begin{aligned}\mathbb{E}[\mathrm{e}^{tS}] &= \mathbb{E}[\mathrm{e}^{t(X_1+\cdots+X_n)}] \\ &= \prod_{i=1}^{n} \mathbb{E}[\mathrm{e}^{tX_i}] = \prod_{i=1}^{n} \frac{\theta^{p_i}}{(\theta-t)^{p_i}} \\ &= \left(\frac{\theta}{\theta-t}\right)^{p_1+\cdots+p_n}\end{aligned}$$

Application

1. Supposons que $X_1, \ldots, X_n$ soient i.i.d. de loi exponentielle de paramètre θ – $\mathrm{Gamma}(1, \theta)$
2. Alors $S = X_1 + \cdots + X_n$ suit la loi $\mathrm{Gamma}(n, \theta)$.
3. Si f est une fonction mesurable positive, alors

$$\mathbb{E}[f(S)] = \int_0^\infty f(s) \frac{\theta^n s^{n-1}}{\Gamma(n)} \mathrm{e}^{-\theta s} \mathrm{d}s$$

Propriétés de l'estimateur des moments (loi exponentielle) $n/\sum_{i=1}^n X_i$

- On sait que sous $\mathbb{P}_\theta$, $\sum_{i=1}^n X_i$ suit la loi $\mathrm{Gamma}(n,\theta)$. Donc Pour $n \geq 2$:

$$\mathbb{E}_\theta[\hat{\theta}_{n,1}] = \mathbb{E}_\theta\left[n/\sum_{i=1}^n X_i\right] = \int_0^\infty \frac{\theta^n}{(n-1)!}\mathrm{e}^{-\theta y} y^{n-1}\frac{n}{y}\mathrm{d}y = \frac{n}{n-1}\theta$$

$$\begin{aligned}\mathbb{E}_\theta[(\hat{\theta}_{n,1}-\theta)^2] &= \mathbb{E}_\theta[\hat{\theta}_{n,1}^2] - 2\theta\mathbb{E}_\theta[\hat{\theta}_{n,1}] + \theta^2 \\ &= \theta^2\frac{n^2}{(n-1)(n-2)} - 2\theta^2\frac{n}{n-1} + \theta^2 = \frac{n+2}{n-1}\frac{\theta^2}{n-2}\end{aligned}$$

- On obtient comme erreur quadratique moyenne

$$\mathbb{E}_\theta[(\hat{\theta}_{n,1}-\theta)^2] = \frac{n+2}{(n-1)(n-2)}\theta^2$$

- On peut mener des calculs similaires pour le second estimateur des moments $\hat{\theta}_{n,2}$.

Modèle de translation et d'échelle

Soit $(X_1, \ldots, X_n)$ un n-échantillon d'une loi de densité $q_\theta(x) = \frac{1}{\sigma} p\left(\frac{x-\mu}{\sigma}\right)$, $\quad \theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+^*$, où p est une densité sur $\mathbb{R}$ vérifiant $\int x p(x) \mathrm{d}x = 0, \quad \int x^2 p(x) \mathrm{d}x = m_2 > 0, \quad \int x^4 p(x) \mathrm{d}x = m_4 < \infty.$

- la densité d'une gaussienne centrée réduite, $p(x) = (2\pi)^{-1/2} \exp(-x^2/2)$: dans ce cas, $m_2 = 1$ et $m_4 = 3$.
- la loi de Laplace $p(x) = (1/2) \exp(-|x|)$ auquel cas $m_2 = 2$ et $m_4 = 4!$.

Estimateur des moments

On pose $T_1(x) = x$ et $T_2(x) = x^2$ et on identifie l'estimateur des moments à la solution du système d'équations

$$\begin{cases} e_1(\mu, \sigma^2) & = \mu = \frac{1}{n}\sum_{i=1}^n X_i \ , \\ e_2(\mu, \sigma^2) & = m_2\sigma^2 + \mu^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 \ . \end{cases}$$ Estimateur des moments $$\begin{cases} \hat{\mu}_n & = \frac{1}{n}\sum_{i=1}^n X_i \ , \\ \hat{\sigma}_n^2 & = \frac{1}{nm_2}\sum_{i=1}^n (X_i - \hat{\mu}_n)^2 \ . \end{cases}$$

Estimateur de translation et d'échelle

- Soit $(X_1, \ldots, X_n)$ un n-échantillon de $(\mathsf{X}, \mathcal{X}, \{q_\theta \cdot \mu\,,\ \theta \in \Theta\})$

$$q_\theta(x) = \sigma^{-1} q\left((x-\mu)/\sigma\right)\ , \quad \theta = (\mu, \sigma) \in \Theta = \mathbb{R} \times \mathbb{R}_+^*\ ,$$

où q est une densité de probabilité paire : $q(x) = q(-x)$.

- Fonctions d'estimation

$$\psi_1(\mu, \sigma, x) := \mathrm{signe}(x-\mu) \quad \text{et} \quad \psi_2(\mu, \sigma, x) := \mathrm{signe}(|x-\mu|/\sigma - a)$$

où $a = Q^{-1}(3/4)$ avec
  - Q est la fonction de répartition de q : $Q(x) = \int_{-\infty}^{x} q(z)\mathrm{d}z$.
  - $\mathrm{signe}(z) = -1$ si $z < 0$, $\mathrm{signe}(z) = +1$ si $z > 0$ et $\mathrm{signe}(0) = 0$.
- Par exemple, si q est la densité d'une Gaussienne centrée réduite :

$$\Phi(a) = 3/4 \ \Rightarrow a \approx 0,6745$$

Vérification 1 : $\mathbb{E}_\theta[\psi_1(\theta, X_1)] = 0$

$$\begin{aligned}\mathbb{E}_\theta[\psi_1(\theta, X_1)] &= \int_{-\infty}^{\infty} \text{signe}(x-\mu)\frac{1}{\sigma} q\left(\frac{x-\mu}{\sigma}\right) \mathrm{d}x \\ &= \int_{-\infty}^{\infty} \text{signe}(z) q\left(z\right) \mathrm{d}z \\ &= 0\end{aligned}$$

car $q(z) = q(-z)$

Vérification 2 : $\mathbb{E}_\theta[\psi_2(\theta, X_1)] = 0$

$$
\begin{aligned}
\mathbb{E}_\theta[\psi_2(\theta, X_1)] &= \int_{-\infty}^{\infty} \operatorname{signe}\left(\frac{|x-\mu|}{\sigma} - a\right) \frac{1}{\sigma} q\left(\frac{x-\mu}{\sigma}\right) \mathrm{d}x \\
&= \int_{-\infty}^{\infty} \operatorname{signe}(|z| - a) q(z) \mathrm{d}z \\
&= 2 \int_{0}^{\infty} \operatorname{signe}(z - a) q(z) \mathrm{d}z \\
&= 2 \left(- \int_{0}^{a} q(z) \mathrm{d}z + \int_{a}^{\infty} q(z) \mathrm{d}z \right) \\
&= 2\{-Q(a) + Q(0) + 1 - Q(a)\} = 4\{3/4 - Q(a)\} = 0 \quad \text{car } Q(0) = 1/2 \text{ par parité de } q
\end{aligned}
$$

Par simplicité, on suppose que n est impair et que $X_i \neq X_j$ pour $i \neq j$

$$\frac{1}{n}\sum_{i=1}^{n} \text{signe}(X_i - \mu) = 0 \quad \frac{1}{n}\sum_{i=1}^{n} \text{signe}(|X_i - \mu|/\sigma - a) = 0$$

Solution

$$\hat{\mu}_n = \text{med}(X_1, \ldots, X_n) = X_{(n+1)/2:n}$$
$$\hat{\sigma}_n = \frac{1}{a}\,\text{med}(|X_1 - \hat{\mu}_n|, \ldots, |X_n - \hat{\mu}_n|)$$

L'estimateur $\hat{\sigma}_n$ est appelé estimateur MAD (*median absolute deviation about the median*) ; il est très utilisé en statistique robuste, car il est peu sensible aux valeurs aberrantes.

Preuve de l'inégalité de Jensen

- Si f est une fonction convexe sur un intervalle ouvert I, alors pour tout $t \in$ I, il existe une constante c_t telle que

$$\Phi(t) + c_t(y - t) \leq \Phi(y) \ , \quad \text{pour tout } y \in \text{I}.$$

- En appliquant cette inégalité avec $t = \mathbb{E}[Y]$,

$$\Phi(\mathbb{E}[Y]) + c_{\mathbb{E}[Y]}(y - \mathbb{E}[Y]) \leq \Phi(y) \ , \quad \text{pour tout } y \in \text{I}$$

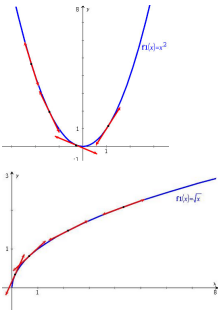
Une fois $t = \mathbb{E}[Y]$ choisi, l'inégalité est vraie pour tout y.

Aussi

$$\Phi(\mathbb{E}[Y]) + c(Y - \mathbb{E}[Y]) \leq \Phi(Y) \ , \mathbb{P} - \text{p.s.} \ .$$

On obtient alors en prenant l'espérance :

$$\Phi(\mathbb{E}[Y]) \leq \mathbb{E}[\Phi(Y)]$$



- L'inégalité de la tangente est stricte dès que Φ est strictement convexe et $y \neq t$. L'inégalité

$$\Phi(\mathbb{E}[Y]) + c_{\mathbb{E}[Y]}(Y - \mathbb{E}[Y]) \leq \Phi(Y) \ , \mathbb{P} - \text{p.s.} \ ,$$

est donc stricte sur $\{Y \neq \mathbb{E}[Y]\}$, événement de probabilité strictement positive dès que Y n'est pas presque sûrement constante.