Chapitre I-2

Estimation ponctuelle

Ce chapitre est consacré à l'estimation ponctuelle : un modèle statistique étant fixé, comment proposer, à partir des observations, une valeur plausible du paramètre inconnu ? Deux méthodes générales de construction sont étudiées, la **méthode des moments** et le **maximum de vraisemblance**. Pour chacune, on suit le même canevas : l'intuition, la définition, des exemples, quelques propriétés, puis une généralisation – les *Z-estimateurs* pour la première, les *M-estimateurs* pour la seconde. L'étude du maximum de vraisemblance conduit à la divergence de Kullback-Leibler, qui décrit la limite de la log-vraisemblance normalisée et le point où cette limite atteint son maximum.
La partie principale du chapitre suit le déroulé du cours. Elle est suivie d'une partie *Outils mathématiques*, qui rassemble les prérequis utilisés (inégalités de Jensen et de Bienaymé-Tchebychev, loi faible des grands nombres, fonctions convexes) et rappelle où sont démontrés les résultats empruntés au chapitre précédent, puis de la feuille d'exercices, et enfin d'une partie *Compléments*, hors programme d'examen, qui traite de modèles moins standards (Cauchy, translation et échelle, censure, régression générale) et donne l'énoncé complet du théorème sur la divergence de Kullback-Leibler.

Comment construire un estimateur ?

Au chapitre précédent, nous avons proposé quelques estimateurs sans méthode générale : la proportion empirique de réponses positives pour le paramètre d'une loi de Bernoulli, une correction affine de cette proportion pour le protocole des réponses randomisées, une droite ajustée par moindres carrés pour une tendance de température (premiers estimateurs du chapitre 1). Dans chaque cas, l'estimateur a été trouvé au cas par cas, en remarquant que le paramètre s'exprimait à l'aide d'une moyenne que les observations permettaient d'approcher.

Ce chapitre présente deux procédés systématiques, applicables dès que le modèle est écrit. Le premier, la méthode des moments, prolonge le raisonnement précédent : on égale des moyennes théoriques, fonctions du paramètre, à leurs contreparties empiriques. Le second, le maximum de vraisemblance, retient la valeur du paramètre qui rend les observations les plus vraisemblables, c'est-à-dire qui maximise la densité des observations vue comme une fonction du paramètre. Les deux méthodes ne coïncident pas en général, et aucune des deux ne garantit à elle seule que l'estimateur obtenu est bon : la mesure de la qualité d'un estimateur fait l'objet des chapitres suivants.

I-2.1 Les fondamentaux

I-2.1.1 Estimation ponctuelle

Le chapitre précédent a distingué quatre tâches de l'inférence statistique (Section I-1.1.3) : l'estimation ponctuelle, l'estimation par région, le test et la prédiction. Ce chapitre est consacré à la première. Estimer ponctuellement, c'est répondre par une seule valeur : on calcule un nombre, ou un vecteur, à partir des observations, et on le propose comme valeur plausible de la quantité inconnue. Cette quantité n'est pas nécessairement le paramètre lui-même ; ce peut être une fonction du paramètre, par exemple la variance $1/\theta^2$ d'une loi exponentielle d'intensité θ.

Définition I-2.1.1: 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 un espace mesurable contenant $g(\Theta)$, typiquement $(\mathbb{R}^q, \mathcal{B}(\mathbb{R}^q))$: une application mesurable $T : (\mathsf{Z}, \mathcal{Z}) \to (\mathbb{R}^q, \mathcal{B}(\mathbb{R}^q))$ [1].

Comme toute statistique (Définition I-1.3), un estimateur ne dépend que des observations : le paramètre inconnu n'intervient pas dans son expression. C'est la seule contrainte imposée par la définition.

Exemple I-2.1.1: Qu'est-ce qui est un estimateur ?

Considérons un n-échantillon $(X_1, \dots, X_n)$ du modèle exponentiel $\big(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+), \{\mathcal{E}(\theta)\,,\ \theta \in \mathbb{R}_+^*\}\big)$, et cherchons à estimer θ. Parmi les cinq quantités

$$X_1\,, \qquad \bar{X}_n\,, \qquad \theta\,, \qquad 1\,, \qquad -1\,,$$

quatre sont des estimateurs de θ : X_1, $\bar{X}_n$, la constante 1 et la constante -1 sont des fonctions mesurables des observations. La cinquième, θ, n'en est pas un : c'est la quantité inconnue que l'on cherche à estimer, et elle n'est pas calculable à partir des données. En revanche, rien dans la définition n'interdit à -1 d'être un estimateur d'un paramètre strictement positif : c'est un très mauvais estimateur, mais c'en est un.

Pour départager des estimateurs, il faut donc mesurer leur qualité. Trois mesures seront introduites dans la suite du cours : le *biais* $\mathbb{E}_\theta[T(Z)] - g(\theta)$, qui compare l'estimateur à sa cible en moyenne ; l'*erreur quadratique moyenne* $\mathbb{E}_\theta\big[\|T(Z) - g(\theta)\|^2\big]$; et, pour une suite d'estimateurs indexée par le nombre d'observations, la *consistance*, c'est-à-dire la convergence de $T_n(Z)$ vers $g(\theta)$ lorsque n tend vers l'infini. Les deux premières sont étudiées avec la théorie de la décision, aux cours 3 et 4, la troisième avec le comportement asymptotique des estimateurs, aux cours 6 et 7. Nous les rencontrerons néanmoins dès ce chapitre, en particulier dans l'Exercice 2 (PC2), où le biais de deux estimateurs concurrents est calculé et comparé.

Intuition. *Un estimateur n'est pas nécessairement « bon ». La définition ne demande qu'une chose, qu'il soit calculable à partir des données, et c'est ce qui la rend utile : elle sépare la question de la* construction*, objet de ce chapitre, de celle de l'*évaluation*, objet des chapitres suivants. Une méthode de construction fournit un candidat ; c'est ensuite l'étude de son biais, de son erreur quadratique et de son comportement quand n grandit qui dit s'il faut le retenir.*

I-2.1.2 Méthode des moments

La méthode des moments part d'une remarque élémentaire : dans beaucoup de modèles, le paramètre s'exprime à l'aide d'une espérance, et une espérance s'approche par une moyenne em-

1. On impose parfois à un estimateur de $g(\theta)$ de prendre ses valeurs dans $g(\Theta)$ lui-même. Cette restriction apporte peu sur le plan conceptuel et exclut des estimateurs pourtant parfaitement admissibles dans la théorie classique ; nous ne la retenons pas.

pirique. L'exemple qui suit, calculé avant toute définition, en montre le principe.

Exemple I-2.1.2: Modèle de Bernoulli : un premier estimateur des moments

Soit $(X_1, \dots, X_n)$ un n-échantillon du modèle $\big(\{0,1\}, \mathcal{P}(\{0,1\}), \{\mathcal{B}\mathrm{er}(p)\,,\ p \in [0,1]\}\big)$, où l'on note p le paramètre. Le candidat naturel pour estimer p est la proportion empirique de 1, c'est-à-dire la moyenne empirique $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$. La raison en est que $\mathbb{E}_p[X_1] = p$: le paramètre est une espérance, et l'on estime une espérance par sa version empirique, $\hat{p}_n := \bar{X}_n$. C'est l'estimateur proposé au chapitre 1 (Section I-1.1.8).

On a choisi une fonction T des observations (ici $T(x) = x$), calculé son espérance $e(\theta) = \mathbb{E}_\theta[T(X_1)]$ en fonction du paramètre, puis résolu l'équation $e(\hat{\theta}_n) = n^{-1}\sum_{i=1}^n T(X_i)$. Le cas d'un paramètre réel et d'une seule fonction T se formule ainsi.

Définition I-2.1.2: Méthode des moments, cas simple

Soit $(X_1, \dots, X_n)$ un n-échantillon d'un modèle paramétrique $(\mathsf{X}, \mathcal{X}, \{\mathbb{Q}_\theta\,,\ \theta \in \Theta\})$ avec $\Theta \subseteq \mathbb{R}$, et soit $T : \mathsf{X} \to \mathbb{R}$ une fonction mesurable telle que $\mathbb{E}_\theta[|T(X_1)|] < \infty$ pour tout $\theta \in \Theta$. On appelle *moment exact* associé à T la fonction $e : \theta \mapsto e(\theta) := \mathbb{E}_\theta[T(X_1)]$, et *moment empirique* associé à T la statistique $n^{-1}\sum_{i=1}^n T(X_i)$. L'*estimateur des moments* associé à T est la valeur $\hat{\theta}_n$ du paramètre qui égale le moment exact et le moment empirique,

$$e(\hat{\theta}_n) = \frac{1}{n}\sum_{i=1}^n T(X_i)\ ,$$

lorsque cette équation, d'inconnue $\vartheta \in \Theta$, admet une solution unique.

Les deux exemples qui suivent appliquent cette définition à des modèles où le paramètre n'est pas lui-même une espérance.

Exemple I-2.1.3: Modèles exponentiel et de Poisson : deux applications

1. *Modèle exponentiel.* Soit $(X_1, \dots, X_n)$ un n-échantillon du modèle $\big(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+), \{\mathcal{E}(\theta)\,,\ \theta \in \mathbb{R}_+^*\}\big)$. Ici $\mathbb{E}_\theta[X_1] = 1/\theta$: le paramètre n'est pas lui-même une espérance, mais s'en déduit par une fonction inversible. On retient la valeur $\hat{\theta}_n$ telle que $1/\hat{\theta}_n = \bar{X}_n$, soit $\hat{\theta}_n = 1/\bar{X}_n$, définie dès que $\bar{X}_n > 0$, ce qui est le cas $\mathbb{P}_\theta$-presque sûrement.
2. *Modèle de Poisson, moment d'ordre* 2. Soit $(X_1, \dots, X_n)$ un n-échantillon du modèle $\big(\mathbb{N}, \mathcal{P}(\mathbb{N}), \{\mathcal{P}\mathrm{oi}(\lambda)\,,\ \lambda \in \mathbb{R}_+^*\}\big)$. Rien n'oblige à utiliser le moment d'ordre 1 : comme $\mathbb{E}_\lambda[X_1^2] = \mathrm{Var}_\lambda(X_1) + \mathbb{E}_\lambda[X_1]^2 = \lambda + \lambda^2$, on peut retenir la valeur $\hat{\lambda}_n$ telle que $\hat{\lambda}_n^2 + \hat{\lambda}_n = n^{-1}\sum_{i=1}^n X_i^2$, c'est-à-dire l'unique racine dans $\mathbb{R}_+$ de cette équation du second degré,

$$\hat{\lambda}_n = \frac{1}{2}\Big(-1 + \sqrt{1 + 4\,n^{-1}\textstyle\sum_{i=1}^n X_i^2}\Big)\ .$$

   Cette racine vaut 0 exactement lorsque toutes les observations sont nulles, événement de probabilité $\mathrm{e}^{-n\lambda} > 0$. L'espace des paramètres ayant été pris égal à $\mathbb{R}_+^*$, l'estimateur n'est alors pas défini ; il le serait si l'on adjoignait à Θ la valeur $\lambda = 0$, à laquelle correspond la masse de Dirac en 0. Le premier estimateur des moments, $\bar{X}_n$, rencontre exactement la même limite.

Lorsque le paramètre est de dimension d, une seule équation ne suffit plus : il en faut autant que d'inconnues, donc autant de fonctions de moment.

Méthode I-2.1.1: Méthode des moments

Soit $(X_1, \dots, X_n)$ un n-échantillon d'un modèle paramétrique $(\mathsf{X}, \mathcal{X}, \{\mathbb{Q}_\theta\,,\ \theta \in \Theta\})$, où $\Theta \subseteq \mathbb{R}^d$. Pour tout $\theta \in \Theta$, on note $\mathbb{P}_\theta := \mathbb{Q}_\theta^{\otimes n}$ la loi des observations.

1. Choisir d fonctions mesurables $T_j : \mathsf{X} \to \mathbb{R}$, $j \in \{1, \dots, d\}$, telles que
$$\mathbb{E}_\theta[|T_j(X_1)|] = \int_{\mathsf{X}} |T_j(x)|\, \mathbb{Q}_\theta(\mathrm{d}x) < \infty\,, \qquad \theta \in \Theta\,.$$
2. Calculer les *moments exacts* : pour tout $\theta \in \Theta$ et tout $j \in \{1, \dots, d\}$,
$$e_j(\theta) := \mathbb{E}_\theta[T_j(X_1)]\,.$$
3. Les moments exacts $e_j(\theta)$ ne sont pas connus, puisque θ ne l'est pas ; on les estime par les *moments empiriques* $n^{-1}\sum_{i=1}^n T_j(X_i)$.
4. Résoudre, par rapport à la variable ϑ, le système de d équations à d inconnues
$$\forall j \in \{1, \dots, d\} \qquad \frac{1}{n}\sum_{i=1}^n T_j(X_i) = e_j(\vartheta)\,.$$

En supposant que ce système admette une solution unique $\hat\theta_n$ dans Θ, on appelle $\hat\theta_n$ l'*estimateur des moments* de θ associé aux fonctions T_j, $j = 1, \dots, d$.

L'estimateur des moments est donc égal à la valeur du paramètre pour laquelle les moments exacts et les moments empiriques sont égaux. Deux points méritent d'être soulignés dès maintenant. D'une part, la méthode ne produit un estimateur que si le système admet une solution unique : ni l'existence ni l'unicité ne sont automatiques, et la solution peut exister sans appartenir à l'espace des paramètres, comme dans le modèle de Poisson lorsque toutes les observations sont nulles. D'autre part, le résultat dépend des fonctions T_j choisies ; un même modèle admet donc en général plusieurs estimateurs des moments, comme le montre l'exemple qui suit.

Exemple I-2.1.4: Modèle exponentiel : deux estimateurs des moments

Le modèle exponentiel est le modèle de base des systèmes à événements discrets : arrivées à un service, occurrences de panne, durées de survie. Soit $(X_1, \dots, X_n)$ un n-échantillon du modèle
$$\big(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+), \{\mathcal{E}(\theta)\,,\ \theta \in \Theta := \mathbb{R}_+^*\}\big)\,,$$
où la loi exponentielle d'intensité $\theta > 0$ admet, par rapport à la mesure de Lebesgue Leb sur $\mathbb{R}$, la densité
$$q_\theta(x) := \theta \mathrm{e}^{-\theta x}\mathbb{1}_{\mathbb{R}_+}(x)\,. \tag{I-2.1}$$
Cette loi est sans mémoire : pour tous $a, b > 0$ et tout $\theta \in \Theta$,
$$\mathbb{P}_\theta\left(X_1 \geq a + b \,|\, X_1 \geq a\right) = \frac{\mathrm{e}^{-(a+b)\theta}}{\mathrm{e}^{-a\theta}} = \mathbb{P}_\theta(X_1 \geq b)\,,$$
et ses moments se calculent par intégrations par parties successives, ou par la fonction génératrice des moments (lemme des moments de la loi exponentielle, chapitre 1) :
$$\mathbb{E}_\theta[X_1^k] = \frac{k!}{\theta^k}\,, \qquad k \in \mathbb{N}^*\,.$$

Considérons les deux fonctions $T(x) = x$ et $\tilde{T}(x) = x^2$. Les moments exacts associés sont

$$e(\theta) = \mathbb{E}_\theta\left[T(X_1)\right] = \int_0^{+\infty} x\,\theta\exp(-\theta x)\,\mathrm{d}x = \frac{1}{\theta}\,,$$
$$\tilde{e}(\theta) = \mathbb{E}_\theta\left[\tilde{T}(X_1)\right] = \int_0^{+\infty} x^2\,\theta\exp(-\theta x)\,\mathrm{d}x = \frac{2}{\theta^2}\,.$$

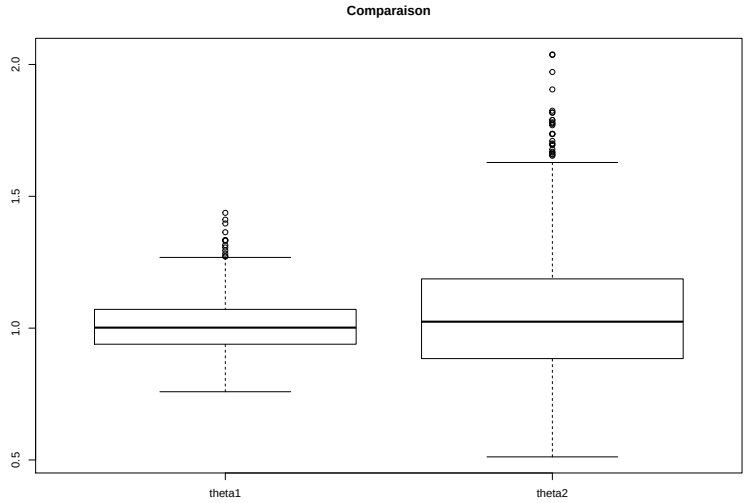
Les estimateurs des moments associés sont les solutions respectives des équations

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

Sous la contrainte $\vartheta > 0$, et dès que les observations ne sont pas toutes nulles – ce qui est vrai $\mathbb{P}_\theta$-presque sûrement –, chacune de ces équations a une solution unique :

$$\hat{\theta}_{n,1} := \frac{1}{n^{-1}\sum_{i=1}^n X_i}\,, \qquad \text{et} \qquad \hat{\theta}_{n,2} := \left(\frac{2}{n^{-1}\sum_{i=1}^n X_i^2}\right)^{1/2}\,. \tag{I-2.2}$$

Remarque. Il y a donc, ici, deux estimateurs des moments, et il y en aurait autant que de fonctions T admissibles. On ne dit pas « l'estimateur des moments » du modèle exponentiel, mais *un* estimateur des moments, ou l'estimateur des moments *associé au moment* T. Les deux estimateurs de (I-2.2) ne se valent pas : la figure ci-dessous compare leurs lois sur des échantillons simulés. Le biais et l'erreur quadratique moyenne de $\hat{\theta}_{n,1}$ sont calculés au chapitre 1 à l'aide des lois Gamma (Section I-1.4) : $\mathbb{E}_\theta[\hat{\theta}_{n,1}] = n\theta/(n-1)$ pour $n \geq 2$ et $\mathbb{E}_\theta[(\hat{\theta}_{n,1} - \theta)^2] = (n+2)\theta^2/((n-1)(n-2))$ pour $n \geq 3$.



Comparaison des estimateurs $\hat{\theta}_{n,1} = 1/(n^{-1}\sum_{i=1}^n X_i)$ *et* $\hat{\theta}_{n,2} = \left(2/(n^{-1}\sum_{i=1}^n X_i^2)\right)^{1/2}$ *: boîtes à moustaches des valeurs prises par chacun des deux estimateurs sur un grand nombre d'échantillons simulés de taille* n *sous* $\mathbb{P}_\theta$*, pour* $\theta = 1$*. Les deux estimateurs se concentrent autour de la vraie valeur du paramètre, mais* $\hat{\theta}_{n,2}$ *est nettement plus dispersé que* $\hat{\theta}_{n,1}$*.*

Remarque. Une boîte à moustaches résume une série de valeurs par cinq nombres calculés sur la série ordonnée. La boîte s'étend du premier quartile q_1 au troisième quartile q_3, et le trait qui la coupe marque la médiane : elle contient donc la moitié centrale des valeurs, et sa

hauteur est l'écart interquartile $\mathrm{IQR} = q_3 - q_1$. Les moustaches ne vont pas jusqu'au minimum et au maximum : suivant la convention de Tukey, la moustache supérieure s'arrête à la plus grande valeur inférieure à $q_3 + 1{,}5\,\mathrm{IQR}$, la moustache inférieure à la plus petite valeur supérieure à $q_1 - 1{,}5\,\mathrm{IQR}$, et les valeurs situées au-delà sont tracées une à une. Pour des observations gaussiennes, la boîte couvre environ 50 % des valeurs et les moustaches environ 99,3 %. Ici, chaque boîte résume les valeurs prises par un estimateur sur un grand nombre d'échantillons simulés sous la même loi : la position de la médiane par rapport à la vraie valeur du paramètre renseigne sur le biais, la hauteur de la boîte et la longueur des moustaches sur la dispersion.

Le modèle uniforme sur un intervalle inconnu demande, lui, deux fonctions de moment, puisque le paramètre est de dimension 2.

Exemple I-2.1.5: Loi uniforme sur $[a,b]$: deux moments nécessaires

La loi uniforme est le modèle simple d'une variable aléatoire bornée dans un intervalle. Soit $(X_1, \dots, X_n)$ un n-échantillon de la loi uniforme sur $[a,b]$, où le paramètre est $\theta = (a,b) \in \Theta := \big\{(a,b) \in \mathbb{R}^2\,,\ a < b\big\}$, de densité par rapport à la mesure de Lebesgue Leb

$$q_\theta(x) = \frac{1}{b-a}\,\mathbb{1}_{[a,b]}(x)\ .$$

Avec $T(x) = x$ et $\tilde{T}(x) = x^2$, les moments exacts valent

$$e(\theta) = \mathbb{E}_\theta[X_1] = \frac{a+b}{2}\ , \qquad \tilde{e}(\theta) = \mathbb{E}_\theta[X_1^2] = \frac{b^3-a^3}{3(b-a)} = \frac{a^2+ab+b^2}{3}\ .$$

L'estimateur des moments est la solution du système

$$\frac{a+b}{2} = \bar{X}_n\ , \qquad \frac{a^2+ab+b^2}{3} = \overline{X^2}_n := \frac{1}{n}\sum_{i=1}^n X_i^2\ .$$

Le système n'est pas linéaire, mais le changement d'inconnues $s = a+b$ et $p = ab$ le résout : comme $a^2 + ab + b^2 = s^2 - p$, il vient

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

et $\hat{a}_n$, $\hat{b}_n$ sont les racines de $t^2 - \hat{s}_n t + \hat{p}_n = 0$. Ces racines sont réelles : en notant $V_n := \overline{X^2}_n - \bar{X}_n^2 = n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2 \geq 0$ la variance empirique non corrigée, le discriminant vaut

$$\hat{s}_n^2 - 4\hat{p}_n = 4\bar{X}_n^2 - 4\big(4\bar{X}_n^2 - 3\overline{X^2}_n\big) = 12\big(\overline{X^2}_n - \bar{X}_n^2\big) = 12\,V_n \geq 0\ ,$$

de sorte que

$$\hat{a}_n = \bar{X}_n - \sqrt{3V_n}\ , \qquad \hat{b}_n = \bar{X}_n + \sqrt{3V_n}\ .$$

Pour $n \geq 2$, les observations ne sont pas toutes égales $\mathbb{P}_\theta$-presque sûrement, donc $V_n > 0$, $\hat{a}_n < \hat{b}_n$ et $(\hat{a}_n, \hat{b}_n) \in \Theta$: le système admet une solution unique dans Θ, la contrainte $a < b$ imposant l'ordre des deux racines. Ici, plusieurs moments sont *nécessaires* pour identifier les deux paramètres a et b : le nombre d'équations doit être au moins égal au nombre d'inconnues.

Remarque. L'intervalle $[\hat{a}_n, \hat{b}_n]$ ne contient pas nécessairement toutes les observations : cela se produit dès que $X_{n:n} - \bar{X}_n > \sqrt{3V_n}$, où $X_{n:n}$ est la plus grande observation. Par exemple, pour un échantillon de taille 10 dont neuf observations sont proches de 0 et la dixième proche de 1, on trouve $\bar{X}_n \simeq 0{,}1$, $V_n \simeq 0{,}09$ et $\hat{b}_n \simeq 0{,}1 + \sqrt{0{,}27} \simeq 0{,}62$, alors que la plus grande observation vaut environ 1. La méthode des moments n'utilise que les deux premiers moments empiriques, et ne tient pas compte de la contrainte $a \leq X_i \leq b$ imposée par chaque observation.

Le maximum de vraisemblance, lui, retiendra $[X_{1:n}, X_{n:n}]$ (Exemple I-2.1.11).

Illustration / animation : L'intervalle des moments ne contient pas toujours toutes les observations

On illustre ici, pour un échantillon uniforme sur $[a, b]$, l'intervalle $[\hat{a}_n, \hat{b}_n]$ obtenu par la méthode des moments, à comparer à l'intervalle $[X_{1:n}, X_{n:n}]$, qui contient nécessairement toutes les observations ; les répétitions permettent d'observer la fréquence à laquelle le premier échoue.

link

Convergence des moments empiriques. Une fois l'estimateur construit, la première question est de savoir s'il se rapproche du paramètre lorsque le nombre d'observations augmente. Pour la méthode des moments, la réponse repose sur un fait de probabilités : une moyenne empirique se concentre autour de l'espérance qu'elle estime.

Proposition I-2.1.1: Convergence des moments empiriques

Soient $(X_1, \dots, X_n)$ un n-échantillon du modèle $(\mathsf{X}, \mathcal{X}, \{\mathbb{Q}_\theta\,,\ \theta \in \Theta\})$ et $T : \mathsf{X} \to \mathbb{R}$ une fonction mesurable telle que $\mathbb{E}_\theta[|T(X_1)|] < \infty$ pour tout $\theta \in \Theta$; on note $e(\theta) = \mathbb{E}_\theta[T(X_1)]$.

(i) Si de plus $\mathbb{E}_\theta[T(X_1)^2] < \infty$, alors, en notant $\sigma^2(\theta) := \mathrm{Var}_\theta(T(X_1)) = \int_{\mathsf{X}} \{T(x) - e(\theta)\}^2\, \mathbb{Q}_\theta(\mathrm{d}x)$, pour tout $\varepsilon > 0$,

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

(ii) Dans tous les cas, $n^{-1}\sum_{i=1}^{n} T(X_i) \overset{\mathbb{P}_\theta-\mathrm{prob}}{\longrightarrow} e(\theta)$ (loi faible des grands nombres).

En particulier, si pour tout n l'estimateur des moments $\hat{\theta}_n$ associé à T est défini $\mathbb{P}_\theta$-presque sûrement, alors $e(\hat{\theta}_n) = n^{-1}\sum_{i=1}^{n} T(X_i)$ presque sûrement, de sorte que $\mathbb{P}_\theta(|e(\hat{\theta}_n) - e(\theta)| \geq \varepsilon) \leq \sigma^2(\theta)/(n\varepsilon^2)$ sous l'hypothèse de (i), et $e(\hat{\theta}_n) \overset{\mathbb{P}_\theta-\mathrm{prob}}{\longrightarrow} e(\theta)$.

Démonstration

Sous $\mathbb{P}_\theta$, les variables $T(X_1), \dots, T(X_n)$ sont indépendantes, de même loi, d'espérance $e(\theta)$. Le point (i) est l'inégalité de Bienaymé-Tchebychev (Proposition I-2.2.1) appliquée à la moyenne empirique, dont la variance vaut $\sigma^2(\theta)/n$ par indépendance. Le point (ii) est la loi faible des grands nombres pour des variables i.i.d. intégrables (Théorème I-2.2.1) ; sous l'hypothèse de (i), il découle aussi directement de la borne de (i). La dernière affirmation résulte de la définition de $\hat{\theta}_n$. □

Remarque. La proposition porte sur $e(\hat{\theta}_n)$ et non sur $\hat{\theta}_n$. Pour conclure sur l'estimateur lui-même, il faut pouvoir inverser e : si e est injective et si son inverse $e^{-1} : e(\Theta) \to \Theta$ est continue au point $e(\theta)$, alors $\hat{\theta}_n = e^{-1}\big(n^{-1}\sum_{i=1}^{n} T(X_i)\big)$ converge en probabilité vers θ (Lemme I-2.2.1). C'est le cas, par exemple, lorsque Θ est un intervalle et que e est continue et strictement monotone : dans le modèle exponentiel, $e(\theta) = 1/\theta$ et $\hat{\theta}_{n,1} = 1/\bar{X}_n \overset{\mathbb{P}_\theta-\mathrm{prob}}{\longrightarrow} \theta$. La consistance et la vitesse de convergence des estimateurs des moments, qui découlent de la loi des grands nombres et du théorème central limite, sont étudiées dans la suite du cours.

Généralisation : les Z-estimateurs. Résumons. L'estimateur des moments associé aux fonctions $\mathbf{T} = (T_1, \dots, T_d)$ est la solution du système de d équations à d inconnues

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\psi}(\vartheta, X_i) = 0 \;, \qquad \text{avec} \qquad \boldsymbol{\psi}(\vartheta, x) := \mathbf{T}(x) - \mathbb{E}_\vartheta[\mathbf{T}(X_1)] \;. \tag{I-2.3}$$

Très souvent, pour un paramètre réel, une seule fonction de moment suffit. Il est cependant intéressant de considérer des estimateurs définis comme solutions de systèmes d'équations de la forme (I-2.3) pour des *fonctions d'estimation* $\boldsymbol{\psi}$ plus générales, qui ne s'écrivent pas comme la différence entre une statistique et son espérance. La seule propriété de $\boldsymbol{\psi}$ utilisée est que son espérance s'annule au vrai paramètre.

Définition I-2.1.3: Z-estimateur

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

$$\mathbb{E}_\theta\big[\boldsymbol{\psi}(\theta, X_1)\big] = \mathbf{0}_{d\times 1} \;, \qquad \boldsymbol{\psi} := (\psi_1, \dots, \psi_d)^\top \;.$$

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

$$\Psi_n(\hat\theta_n) = 0 \;, \qquad \text{où} \qquad \Psi_n(\vartheta) := \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\psi}(\vartheta, X_i) \;.$$

Un estimateur des moments est un Z-estimateur, associé à la fonction d'estimation (I-2.3). La lettre Z rappelle que l'estimateur est un *zéro* de la fonction aléatoire Ψ_n. Une partie des propriétés de convergence des estimateurs des moments se démontre directement pour les Z-estimateurs, avec les mêmes outils (loi des grands nombres et théorème central limite) ; ce sera l'objet de la suite du cours. L'exemple qui suit montre un Z-estimateur qui n'est pas un estimateur des moments.

Exemple I-2.1.6: Modèle de translation : moyenne empirique et médiane empirique

Soit f une densité de probabilité sur $\mathbb{R}$, par rapport à la mesure de Lebesgue Leb, *paire* : $f(x) = f(-x)$ pour tout $x \in \mathbb{R}$. Soit $(X_1, \dots, X_n)$ un n-échantillon du modèle de translation

$$\big(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{q_\theta \cdot \mathrm{Leb} \,,\, \theta \in \Theta = \mathbb{R}\}\big) \;, \qquad q_\theta(x) := f(x - \theta) \;,$$

par exemple la loi gaussienne centrée réduite translatée de θ, pour $f(x) = (2\pi)^{-1/2}\mathrm{e}^{-x^2/2}$. Deux choix de fonction d'estimation sont possibles.

1. *Moyenne empirique.* Supposons $\int_{\mathbb{R}} |x| f(x)\,\mathrm{d}x < \infty$ et posons $\psi(\theta, x) = x - \theta$. Par le changement de variable $z = x - \theta$ et la parité de f,

$$\mathbb{E}_\theta[X_1 - \theta] = \int_{\mathbb{R}} (x - \theta) f(x - \theta)\,\mathrm{d}x = \int_{\mathbb{R}} z f(z)\,\mathrm{d}z = 0 \;.$$

   L'équation $n^{-1}\sum_{i=1}^n (X_i - \vartheta) = 0$ a pour unique solution $\hat\theta_n = \bar X_n$: la moyenne empirique est le Z-estimateur associé à ψ. C'est aussi l'estimateur des moments associé à $T(x) = x$, puisque $\mathbb{E}_\theta[X_1] = \theta$.

2. *Médiane empirique.* Posons $\psi(\theta, x) = \mathrm{signe}(x - \theta)$, où $\mathrm{signe}(z) = -1$ si $z < 0$, $\mathrm{signe}(0) = 0$ et $\mathrm{signe}(z) = 1$ si $z > 0$. Aucune hypothèse de moment n'est nécessaire, la fonction signe étant bornée, et

$$\mathbb{E}_\theta[\mathrm{signe}(X_1 - \theta)] = \int_{\mathbb{R}} \mathrm{signe}(z) f(z)\,\mathrm{d}z = \int_0^{+\infty} f(z)\,\mathrm{d}z - \int_{-\infty}^0 f(z)\,\mathrm{d}z = 0 \;,$$

toujours par parité de f. Notons $X_{1:n} \leq \cdots \leq X_{n:n}$ les statistiques d'ordre de l'échantillon, notées $X_{(1)} \leq \cdots \leq X_{(n)}$ au chapitre 1 (Section I-1.1.5). Pour ϑ distinct de toutes les observations, $\sum_{i=1}^n \text{signe}(X_i - \vartheta)$ est la différence entre le nombre d'observations strictement supérieures à ϑ et le nombre d'observations strictement inférieures à ϑ. Supposons les observations deux à deux distinctes, ce qui est $\mathbb{P}_\theta$-presque sûr puisque leur loi a une densité. Si $n = 2m+1$ est impair, l'équation $\sum_{i=1}^n \text{signe}(X_i - \vartheta) = 0$ a pour unique solution $\vartheta = X_{m+1:n}$, la médiane empirique ; si $n = 2m$ est pair, ses solutions sont tous les points de l'intervalle ouvert $]X_{m:n}, X_{m+1:n}[$, parmi lesquels la médiane empirique $\frac{1}{2}(X_{m:n} + X_{m+1:n})$. Dans les deux cas, la médiane empirique $\text{med}(X_1, \ldots, X_n)$ est un Z-estimateur associé à ψ.

Ainsi, la moyenne empirique est un estimateur des moments, tandis que la médiane empirique est un Z-estimateur qui n'est pas un estimateur des moments : on vérifie que la fonction $(\theta, x) \mapsto \text{signe}(x - \theta)$ ne s'écrit pas sous la forme $T(x) - \mathbb{E}_\theta[T(X_1)]$.

I-2.1.3 Maximum de vraisemblance

La seconde méthode retient la valeur du paramètre pour laquelle les observations obtenues avaient la plus grande probabilité, ou, pour un modèle à densité, la plus grande densité. Le terme de *vraisemblance* est transparent : on se demande quelle valeur du paramètre rendrait nos observations les plus vraisemblables. Avant la définition, deux exemples fixent l'idée.

Intuition. *Considérons le modèle gaussien* $\{\mathcal{N}(\mu, 1)^{\otimes n}\,,\ \mu \in \mathbb{R}\}$ *pour des tailles, en centimètres, et les observations* 180, 170, 155, 172, 183, 194, 164, … *Est-il vraisemblable que* $\mu = 10$ *? que* $\mu = 200$ *? Sous* $\mathcal{N}(10, 1)$*, une taille de* 170 *est à* 160 *écarts-types de la moyenne, et la densité y vaut* $(2\pi)^{-1/2}\mathrm{e}^{-12\,800}$*. Ces valeurs de* μ *sont possibles au sens strict, la densité gaussienne ne s'annulant jamais, mais si peu vraisemblables qu'on ne les retiendrait pas. Avec une seule observation* X_1*, la densité* $\mu \mapsto (2\pi)^{-1/2}\exp(-(X_1 - \mu)^2/2)$ *est maximale en* $\mu = X_1$ *: la valeur la plus vraisemblable est l'observation elle-même. Avec plusieurs observations indépendantes, la densité du vecteur est le produit des densités,* $\prod_{i=1}^n (2\pi)^{-1/2}\exp(-(X_i - \mu)^2/2)$*, maximal lorsque* $\sum_{i=1}^n (X_i - \mu)^2$ *est minimal, c'est-à-dire en* $\mu = \bar{X}_n$ *: c'est le calcul de l'Exemple I-2.1.10.*

Intuition. *Considérons maintenant le modèle uniforme* $\{\mathcal{U}([0, \theta])^{\otimes n}\,,\ \theta \in \mathbb{R}_+^*\}$ *et les observations* 8, 7, 11, 17, 2, 5, 14, … *La valeur* $\theta = 5$ *est impossible : une observation vaut* $17 > 5$*, et la densité* $\theta^{-1}\mathbb{1}_{[0,\theta]}(17)$ *est nulle. La valeur* $\theta = 100$ *est possible, mais chaque observation a alors la densité* $1/100$*, et le vecteur des* n *observations la densité* 100^{-n}*. La valeur* $\theta = 17$*, la plus grande observation, est possible et donne la densité* 17^{-n}*, la plus grande de toutes : toute valeur inférieure à* 17 *annule la densité, toute valeur supérieure la diminue. Ici, une seule observation renseigne déjà, puisque* θ *est au moins égal à cette observation. Dans les deux exemples, le raisonnement ne fait intervenir que la densité des observations, vue comme une fonction du paramètre : c'est cette fonction que l'on appelle vraisemblance.*

Définitions. Le modèle est supposé dominé : chaque loi $\mathbb{Q}_\theta$ admet une densité q_θ par rapport à une même mesure σ-finie μ (Définition I-1.1). La vraisemblance est la densité du n-échantillon, produit des densités individuelles (Section I-1.1.7), évaluée aux observations et considérée comme une fonction du paramètre.

Définition I-2.1.4: Vraisemblance et estimateur du maximum de vraisemblance

Soient $(\mathsf{X}, \mathcal{X})$ un espace mesurable, μ une mesure σ-finie sur $(\mathsf{X}, \mathcal{X})$ et $(X_1, \ldots, X_n)$ un n-échantillon du modèle dominé $(\mathsf{X}, \mathcal{X}, \{q_\theta \cdot \mu\,,\ \theta \in \Theta\})$, où, pour tout $\theta \in \Theta$, q_θ est la densité de $\mathbb{Q}_\theta$ par rapport à μ. On appelle *fonction de vraisemblance* (ou *vraisemblance*) associée au

n-échantillon $(X_1, \dots, X_n)$ l'application

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

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

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

c'est-à-dire

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

Quand il n'y a pas de risque de confusion, on note $\mathrm{L}_n(\vartheta)$ pour $\mathrm{L}_n(\vartheta, X_1, \dots, X_n)$.

La variable ϑ de la vraisemblance parcourt Θ : c'est un paramètre libre, distinct de la valeur inconnue du paramètre sous laquelle les observations ont été tirées. Cette distinction sera rendue explicite par la notation à la Section I-2.1.4. La vraisemblance dépend du choix de la mesure dominante μ, mais deux mesures dominantes conduisent à des densités qui diffèrent d'un facteur multiplicatif indépendant du paramètre (Section I-1.2) : les points de maximum, donc l'estimateur, n'en dépendent pas [2].

Définition I-2.1.5: Log-vraisemblance normalisée

Sous les hypothèses de la Définition I-2.1.4, la *log-vraisemblance normalisée* est l'application

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

avec la convention $\log 0 = -\infty$.

Remarque. Dans tout ce cours, log désigne le logarithme népérien, de base e, et non le logarithme décimal. On écrit log plutôt que ln parce que cette dernière notation se confond à la lecture avec celles de la vraisemblance L_n et de la log-vraisemblance normalisée ℓ_n.

Le logarithme étant strictement croissant sur $\mathbb{R}_+^*$, et la convention $\log 0 = -\infty$ prolongeant cette monotonie à $\mathbb{R}_+$, les fonctions L_n et ℓ_n ont les mêmes points de maximum :

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

Le passage au logarithme transforme le produit en somme, ce qui simplifie les calculs ; la normalisation par n fait de ℓ_n une moyenne empirique, ce qui servira à comprendre le comportement de l'estimateur lorsque n augmente (Section I-2.1.4).

Exemples.

Exemple I-2.1.7: Modèle exponentiel

Soit $(X_1, \dots, X_n)$ un n-échantillon du modèle exponentiel $\big(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+), \{\mathcal{E}(\theta) \, , \ \theta \in \Theta = \mathbb{R}_+^*\}\big)$,

2. Pour des versions fixées des densités. Une densité n'est définie qu'à un ensemble négligeable près, et la vraisemblance s'évalue aux points observés : en modifiant chaque densité sur un ensemble négligeable qui dépend du paramètre, on peut déplacer les points de maximum, au point de faire perdre sa consistance à l'estimateur du maximum de vraisemblance dans un modèle gaussien (Y. Baraud et L. Birgé, *Annals of Statistics*, 2018, proposition 1). Ce point est précisé dans les compléments du chapitre 1 (Section I-1.4) ; les versions continues utilisées dans ce cours écartent la difficulté. Merci à Matthieu Lerasle de l'avoir signalé.

dont la densité par rapport à la mesure de Lebesgue est $q_\theta(x) = \theta \mathrm{e}^{-\theta x} \mathbb{1}_{\mathbb{R}_+}(x)$. Les observations étant positives, la vraisemblance s'écrit, pour tout $\vartheta > 0$,

$$\mathrm{L}_n(\vartheta) = \prod_{i=1}^n \vartheta \, \mathrm{e}^{-\vartheta X_i} = \vartheta^n \exp\Big(-\vartheta \sum_{i=1}^n X_i \Big) > 0 \, , \qquad \ell_n(\vartheta) = \log \vartheta - \vartheta \, \bar{X}_n \, .$$

La fonction ℓ_n est dérivable sur $\mathbb{R}_+^*$, de dérivée $\ell_n'(\vartheta) = 1/\vartheta - \bar{X}_n$ et de dérivée seconde $-1/\vartheta^2 < 0$: elle est strictement concave. Dès que $\bar{X}_n > 0$, ce qui est $\mathbb{P}_\theta$-presque sûr, elle admet un unique point critique, $\vartheta = 1/\bar{X}_n$, qui est son maximum global : la dérivée ℓ_n' étant strictement décroissante et nulle en ce point, elle est strictement positive avant et strictement négative après, de sorte que ℓ_n croît puis décroît. Ainsi

$$\hat{\theta}_n^{\mathrm{MV}} = \frac{1}{\bar{X}_n} = \hat{\theta}_{n,1} \ :$$

l'estimateur du maximum de vraisemblance coïncide avec le premier estimateur des moments de l'Exemple I-2.1.4.

Exemple I-2.1.8: Modèle de Bernoulli

Soit $(X_1, \ldots, X_n)$ un n-échantillon du modèle $\big(\{0,1\}, \mathcal{P}(\{0,1\}), \{\mathcal{B}\mathrm{er}(p) \, , \ p \in \Theta = [0,1]\}\big)$, où l'on note p le paramètre ; ce modèle est dominé par la mesure de comptage sur $\{0,1\}$, avec la densité $q_p(x) = p^x(1-p)^{1-x}$, en convenant que $0^0 = 1$. En notant $S_n := \sum_{i=1}^n X_i = n\bar{X}_n$ le nombre de 1 observés, la vraisemblance s'écrit, pour $\vartheta \in [0,1]$,

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

Trois cas se présentent. Si $S_n = 0$, alors $\mathrm{L}_n(\vartheta) = (1-\vartheta)^n$ est strictement décroissante et atteint son maximum en $\vartheta = 0$ seulement. Si $S_n = n$, $\mathrm{L}_n(\vartheta) = \vartheta^n$ atteint son maximum en $\vartheta = 1$ seulement. Si $0 < S_n < n$, alors $\mathrm{L}_n(0) = \mathrm{L}_n(1) = 0$ et $\mathrm{L}_n > 0$ sur $]0,1[$, où la log-vraisemblance normalisée vaut

$$\ell_n(\vartheta) = \bar{X}_n \log \vartheta + (1 - \bar{X}_n) \log(1-\vartheta) \, ,$$
$$\ell_n'(\vartheta) = \frac{\bar{X}_n}{\vartheta} - \frac{1-\bar{X}_n}{1-\vartheta} \, , \qquad \ell_n''(\vartheta) = -\frac{\bar{X}_n}{\vartheta^2} - \frac{1-\bar{X}_n}{(1-\vartheta)^2} < 0 \, .$$

La fonction ℓ_n est strictement concave sur $]0,1[$, et son unique point critique $\vartheta = \bar{X}_n$ est son maximum global. Dans les trois cas, l'estimateur du maximum de vraisemblance existe, est unique, et vaut

$$\hat{\theta}_n^{\mathrm{MV}} = \bar{X}_n \, ,$$

qui est aussi l'estimateur des moments associé à $T(x) = x$. Le comportement de la log-vraisemblance lorsque n augmente est étudié en détail sur cet exemple dans les compléments (Section I-2.4).

Exemple I-2.1.9: Loi de Poisson : à vous

Avant de lire la suite, reprenez sur ce modèle la démarche des deux exemples précédents : écrire la vraisemblance, passer à la log-vraisemblance normalisée, isoler les termes qui dépendent du paramètre, puis étudier la fonction obtenue. La version en ligne du cours propose le même exercice à trois niveaux de difficulté, sur les modèles de Poisson, uniforme et de Pareto.
Pour $\theta \in \Theta = \mathbb{R}_+^*$, la loi de Poisson de paramètre θ est la loi sur $\mathbb{N}$ de densité, par rapport à

la mesure de comptage μ sur $\mathbb{N}$,

$$q_\theta(x) = \mathrm{e}^{-\theta}\frac{\theta^x}{x!}\,, \qquad x \in \mathbb{N}\,.$$

Soit $(X_1, \dots, X_n)$ un n-échantillon de cette loi. La vraisemblance s'écrit, pour tout $\vartheta > 0$,

$$\mathrm{L}_n(\vartheta) = \prod_{i=1}^n \mathrm{e}^{-\vartheta}\frac{\vartheta^{X_i}}{X_i!} = \frac{1}{\prod_{i=1}^n X_i!}\exp\big(-n\vartheta + n\bar{X}_n \log\vartheta\big)\,,$$
$$\ell_n(\vartheta) = -\frac{1}{n}\log\Big(\prod_{i=1}^n X_i!\Big) - \vartheta + \bar{X}_n\log\vartheta\,.$$

En isolant ce qui ne dépend pas de ϑ, et en considérant ℓ_n plutôt que L_n, on constate que la fonction à maximiser est $\vartheta \mapsto -\vartheta + \bar{X}_n\log\vartheta$. Si $\bar{X}_n > 0$, elle est strictement concave sur $\mathbb{R}_+^*$, de dérivée $-1 + \bar{X}_n/\vartheta$, qui s'annule en $\vartheta = \bar{X}_n$ seulement : l'estimateur du maximum de vraisemblance est

$$\hat{\theta}_n^{\mathrm{MV}} = \bar{X}_n = n^{-1}\sum_{i=1}^n X_i\,.$$

Si $\bar{X}_n = 0$, c'est-à-dire si toutes les observations sont nulles, événement de probabilité $\mathrm{e}^{-n\theta} > 0$, la fonction $\vartheta \mapsto -\vartheta$ n'a pas de maximum sur $\mathbb{R}_+^*$ et l'estimateur du maximum de vraisemblance n'est pas défini.

Remarque. Sur les modèles de Bernoulli et de Poisson, l'estimateur du maximum de vraisemblance coïncide, lorsqu'il est défini, avec l'estimateur des moments associé à $T(x) = x$, puisque le paramètre y est l'espérance d'une observation. C'est un phénomène typique des familles exponentielles, étudié dans l'Exercice 3 (PC2).

Équations de vraisemblance. Dans les trois exemples précédents, l'estimateur a été obtenu en annulant la dérivée de la log-vraisemblance. Cette démarche a une portée générale, sous deux conditions : que le maximum soit atteint en un point intérieur de Θ, et que la log-vraisemblance y soit différentiable.

Définition I-2.1.6: Équations de vraisemblance

Soit $\Theta \subseteq \mathbb{R}^d$. On suppose que la fonction $\vartheta \mapsto \ell_n(\vartheta, X_1, \dots, X_n)$ est différentiable sur l'intérieur de Θ. Le système de d équations

$$\nabla_\vartheta\ \ell_n(\vartheta, X_1, \dots, X_n) = 0\,, \qquad \vartheta \in \mathring{\Theta}\,,$$

est appelé *équation de vraisemblance* si $d = 1$ et *système d'équations de vraisemblance* si $d > 1$. Tout estimateur $\hat{\theta}_n^{\mathrm{rv}}$ solution de ce système est appelé *racine de l'équation de vraisemblance*.

Si le maximum de ℓ_n, ou ce qui revient au même celui de L_n, est atteint en un point intérieur de Θ où ℓ_n est différentiable, alors l'estimateur du maximum de vraisemblance est une racine de l'équation de vraisemblance : en un point de maximum intérieur, le gradient s'annule. Là où $\mathrm{L}_n > 0$, les fonctions L_n et $\ell_n = n^{-1}\log\mathrm{L}_n$ ont d'ailleurs les mêmes points critiques, puisque $\nabla\,\ell_n = \nabla\,\mathrm{L}_n\,/(n\,\mathrm{L}_n)$. La réciproque est fausse : résoudre le système fournit tous les points critiques de ℓ_n, c'est-à-dire ses maxima et minima locaux, et une racine de l'équation de vraisemblance n'est pas nécessairement un maximum global. L'estimateur du maximum de vraisemblance est donc obtenu soit en résolvant un problème d'optimisation, soit en résolvant un système d'équations, le plus souvent non linéaires, puis en vérifiant que la solution retenue est bien un maximum. Sauf dans les modèles simples, comme le modèle de Poisson, ces estimateurs ne sont pas explicites (voir

le modèle de Cauchy, Section I-2.4).

Exemple I-2.1.10: Moyenne et variance d'une gaussienne

Soit $(X_1, \dots, X_n)$, avec $n \geq 2$, un n-échantillon du modèle gaussien $\big(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{\mathcal{N}(\mu, \sigma^2),\ (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+^*\}\big)$, de densité par rapport à la mesure de Lebesgue

$$q_{(\mu,\sigma^2)}(x) = (2\pi\sigma^2)^{-1/2} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big) .$$

La log-vraisemblance normalisée s'écrit, pour $\vartheta = (m, v) \in \mathbb{R} \times \mathbb{R}_+^*$,

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

et les équations de vraisemblance sont

$$\begin{cases} \dfrac{\partial\, \ell_n}{\partial m}\big((m, v), X_1, \dots, X_n\big) &= \dfrac{1}{nv}\displaystyle\sum_{i=1}^n (X_i - m) = 0 , \\ \dfrac{\partial\, \ell_n}{\partial v}\big((m, v), X_1, \dots, X_n\big) &= -\dfrac{1}{2v} + \dfrac{1}{2nv^2}\displaystyle\sum_{i=1}^n (X_i - m)^2 = 0 . \end{cases}$$

Elles admettent une unique solution, $m = \bar{X}_n$ et $v = n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$, dès que les observations ne sont pas toutes égales, ce qui est $\mathbb{P}_\theta$-presque sûr pour $n \geq 2$. Il reste à vérifier que ce point critique est le maximum global. À v fixé, $m \mapsto \ell_n((m, v))$ est une fonction quadratique strictement concave de m, maximale en $m = \bar{X}_n$. En reportant, $v \mapsto \ell_n((\bar{X}_n, v)) = -\frac{1}{2}\log(2\pi v) - \frac{S}{2v}$, avec $S := n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2 > 0$, a pour dérivée $(S - v)/(2v^2)$, strictement positive pour $v < S$ et strictement négative pour $v > S$: elle atteint son maximum en $v = S$ seulement. Comme $\ell_n((m, v)) \leq \ell_n((\bar{X}_n, v)) \leq \ell_n((\bar{X}_n, S))$ pour tout (m, v), avec égalité si et seulement si $(m, v) = (\bar{X}_n, S)$, l'estimateur du maximum de vraisemblance existe, est unique et vaut

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

Remarque. L'estimateur du maximum de vraisemblance de σ^2 est la variance empirique *non corrigée* $n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{n-1}{n} S_n^2$, et non la variance empirique corrigée S_n^2 du théorème de Gosset (Section I-1.1.5) ; d'après ce théorème, $\mathbb{E}_\theta[\hat{\sigma}^2_{\mathrm{MV}}] = \frac{n-1}{n}\sigma^2$, de sorte que l'estimateur du maximum de vraisemblance est biaisé. Le modèle de régression linéaire gaussienne de l'Exercice 1 (PC2) conduit au même calcul, avec $n - 2$ à la place de $n - 1$.

Exemple I-2.1.11: Modèle uniforme : une vraisemblance non dérivable

Soit $(X_1, \dots, X_n)$ un n-échantillon du modèle uniforme $\big(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{\mathcal{U}([0, \theta]),\ \theta \in \Theta = \mathbb{R}_+^*\}\big)$, de densité par rapport à la mesure de Lebesgue

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

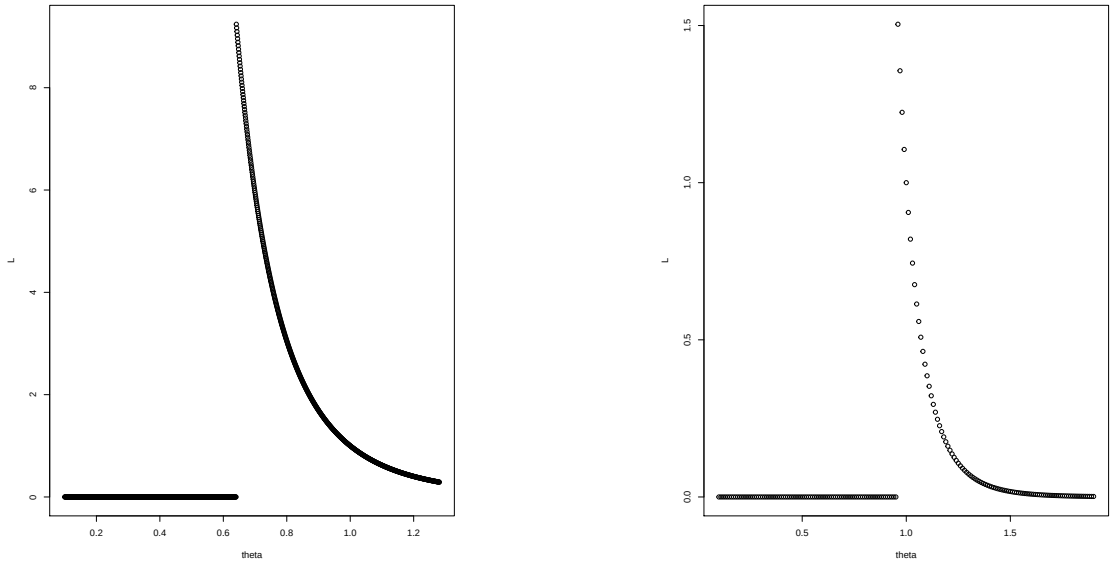
Les observations sont positives $\mathbb{P}_\theta$-presque sûrement, de sorte que la condition « $X_i \in [0, \vartheta]$ pour tout i » équivaut à $X_{n:n} \leq \vartheta$, où $X_{n:n} = \max_{1 \leq i \leq n} X_i$. La vraisemblance s'écrit donc

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

Elle est nulle pour $\vartheta < X_{n:n}$, et strictement décroissante sur $[X_{n:n}, +\infty[$: sa valeur maximale, $X_{n:n}^{-n}$, est atteinte en $\vartheta = X_{n:n}$ seulement, et

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

La fonction L_n n'est pas dérivable en $X_{n:n}$, et la log-vraisemblance vaut $-\infty$ sur $]0, X_{n:n}[$: l'équation de vraisemblance n'a pas de sens ici, et le maximum se détermine directement. Pour $n = 5$ et $n = 10$ observations, la figure ci-dessous montre le saut de la vraisemblance en $X_{n:n}$, puis sa décroissance en ϑ^{-n}, d'autant plus rapide que n est grand.



Vraisemblance $\vartheta \mapsto \mathrm{L}_n(\vartheta)$ du modèle uniforme $\mathcal{U}([0,\theta])$ pour un échantillon de taille $n = 5$ (à gauche) et $n = 10$ (à droite) : la fonction est nulle à gauche de la plus grande observation $X_{n:n}$, puis décroît comme ϑ^{-n}. Le maximum est atteint au point de saut.

Illustration / animation : La vraisemblance uniforme se raidit en $X_{n:n}$ bien plus vite que la vraisemblance exponentielle en $\bar{Y}_n$

L'animation accessible par le QR code ci-contre permet de comparer, sur le même échantillon uniforme sur $[0, \theta^\star]$, le saut de la vraisemblance du modèle uniforme en $X_{n:n}$ et le pic lisse de la vraisemblance du modèle exponentiel en $\bar{Y}_n$, à mesure que n augmente.

link

Méthode I-2.1.2: Explicitation de l'estimateur du maximum de vraisemblance

Lorsque c'est possible, l'estimateur s'explicite en suivant les étapes ci-dessous.

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, Poisson, exponentiel, Gamma, Bernoulli) : la vraisemblance est strictement positive et dérivable sur l'intérieur de Θ.
   a. Écrire la log-vraisemblance $\ell_n(\vartheta) = n^{-1} \log \mathrm{L}_n(\vartheta)$.

   b. Trouver les points critiques, c'est-à-dire résoudre les équations de vraisemblance, un système si $d > 1$.
   c. Vérifier que le point critique retenu est un maximum global (par concavité, ou par une étude directe), et traiter les valeurs des observations pour lesquelles il n'y a pas de point critique.
3. *Cas 2, modèles non réguliers* (uniforme, par exemple) : la vraisemblance n'est pas dérivable, ou s'annule sur une partie de Θ. Déterminer le maximum directement, par une étude de la fonction ; ne pas essayer de différencier.

Les exemples précédents illustrent chacun des deux cas : les modèles exponentiel, de Bernoulli, de Poisson et gaussien (Exemples I-2.1.7 à I-2.1.10) relèvent du cas 1, le modèle uniforme (Exemple I-2.1.11) du cas 2. Les deux exemples qui suivent montrent que l'estimateur du maximum de vraisemblance peut ne pas exister, ou ne pas être unique.

Propriétés : existence, unicité, invariance.

Exemple I-2.1.12: L'estimateur du maximum de vraisemblance n'est pas toujours défini

Soit q_0 la densité, par rapport à la mesure de Lebesgue, définie par

$$q_0(x) = \frac{\mathrm{e}^{-|x|/2}}{2\sqrt{2\pi|x|}} , \qquad x \in \mathbb{R} \setminus \{0\} ,$$

et prolongée par 0 en $x = 0$; c'est une densité de probabilité, puisque $\int_{\mathbb{R}} q_0 \, \mathrm{dLeb} = \int_0^{+\infty} (2\pi x)^{-1/2} \mathrm{e}^{-x/2} \, \mathrm{d}x = 1$ (c'est l'intégrale de la densité de la loi $\chi^2(1)$, Section I-1.2). Considérons un n-échantillon $(X_1, \ldots, X_n)$ du modèle de translation $\big(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{q_0(\cdot - \theta) \cdot \mathrm{Leb} \,, \ \theta \in \Theta = \mathbb{R}\}\big)$. La vraisemblance s'écrit

$$\mathrm{L}_n(\vartheta) = \prod_{i=1}^n q_0(X_i - \vartheta) .$$

Les observations sont $\mathbb{P}_\theta$-presque sûrement deux à deux distinctes. Pour tout i, lorsque ϑ tend vers X_i, le facteur $q_0(X_i - \vartheta)$ tend vers $+\infty$, tandis que les autres facteurs $q_0(X_j - \vartheta)$, $j \neq i$, tendent vers $q_0(X_j - X_i) \in]0, +\infty[$: ainsi $\lim_{\vartheta \to X_i} \mathrm{L}_n(\vartheta) = +\infty$. La vraisemblance n'est pas majorée, l'ensemble $\arg\max_{\vartheta \in \Theta} \mathrm{L}_n(\vartheta)$ est vide, et l'estimateur du maximum de vraisemblance n'est pas défini pour ce modèle.

Exemple I-2.1.13: Modèle de Laplace : non-unicité du maximum de vraisemblance

Soit $(X_1, \ldots, X_n)$ un n-échantillon du modèle de Laplace $\big(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{q_\theta \cdot \mathrm{Leb} \,, \ \theta \in \Theta = \mathbb{R}\}\big)$, de densité par rapport à la mesure de Lebesgue

$$q_\theta(x) = \frac{1}{2} \exp\big(- |x - \theta|\big) , \qquad x \in \mathbb{R} .$$

La log-vraisemblance normalisée s'écrit

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

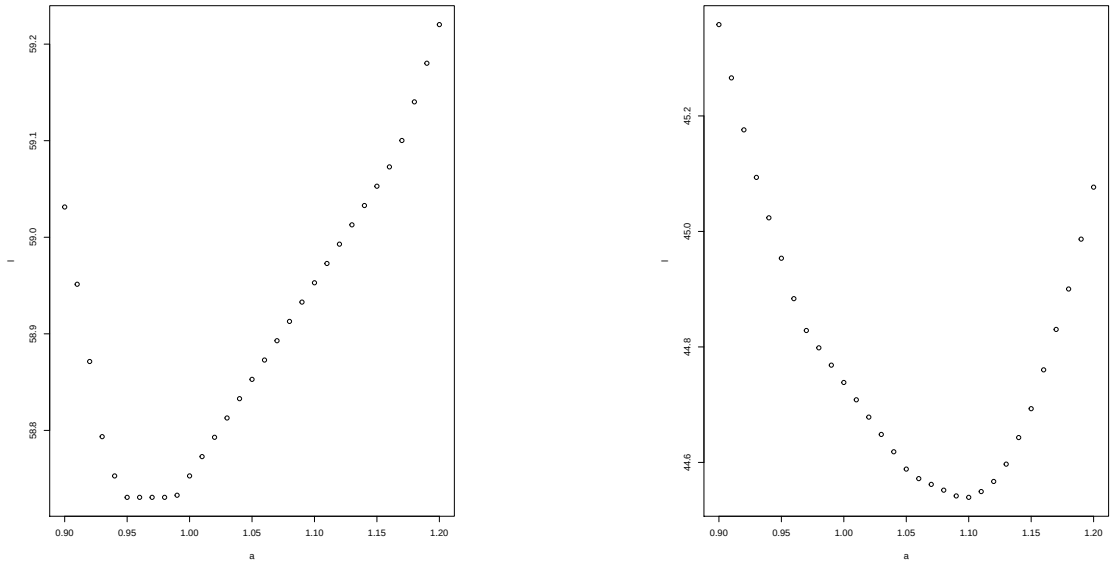
et maximiser ℓ_n revient à minimiser la fonction $\varphi : \vartheta \mapsto \sum_{i=1}^n |X_i - \vartheta|$. Cette fonction est continue, convexe et affine par morceaux ; elle est dérivable en tout point distinct des observations, de dérivée

$$\varphi'(\vartheta) = - \sum_{i=1}^n \mathrm{signe}(X_i - \vartheta) ,$$

constante par morceaux. Avec les notations de l'Exemple I-2.1.6, pour $\vartheta \in]X_{k:n}, X_{k+1:n}[$, la dérivée vaut $k-(n-k) = 2k-n$: elle est strictement négative tant que $k < n/2$ et strictement positive dès que $k > n/2$.

— Si $n = 2m+1$ est impair, φ décroît strictement sur $]-\infty, X_{m+1:n}]$ et croît strictement sur $[X_{m+1:n}, +\infty[$: son minimum est atteint en un point unique, $\hat{\theta}_n^{\mathrm{MV}} = X_{m+1:n}$, la médiane empirique.

— Si $n = 2m$ est pair, φ décroît strictement sur $]-\infty, X_{m:n}]$, est constante sur $[X_{m:n}, X_{m+1:n}]$ (la dérivée y vaut $2m-n = 0$) et croît strictement sur $[X_{m+1:n}, +\infty[$: il y a une infinité de solutions, tout point de l'intervalle fermé $[X_{m:n}, X_{m+1:n}]$ est un estimateur du maximum de vraisemblance, et la médiane empirique $\frac{1}{2}(X_{m:n} + X_{m+1:n})$ en est un parmi d'autres.



Fonction $\vartheta \mapsto \sum_{i=1}^n |X_i - \vartheta|$, dont les minimiseurs sont les estimateurs du maximum de vraisemblance, pour un échantillon du modèle de Laplace de paramètre $\theta = 1$ de taille $n = 50$ (à gauche) et $n = 51$ (à droite). Pour n pair, la fonction présente un plateau de minimiseurs ; pour n impair, le minimiseur est unique.

Illustration / animation : Un nombre pair d'observations laisse un plateau de maximiseurs de vraisemblance, un nombre impair n'en laisse qu'un

Vous pouvez comparer ici, pour deux échantillons de même loi de Laplace, l'un de taille paire et l'autre de taille impaire, la fonction $\vartheta \mapsto \sum_i |X_i - \vartheta|$ et l'ensemble de ses minimiseurs.

link

Reparamétrer un modèle, c'est décrire la même famille de lois avec un autre paramètre, image du premier par une bijection : le modèle gaussien peut être paramétré par la variance ou par l'écart-type. L'estimateur du maximum de vraisemblance suit ce changement de paramètre.

Proposition I-2.1.2: Invariance par reparamétrisation

Soient $\{\mathbb{P}_\theta\,,\ \theta\in\Theta\}$ un modèle paramétrique dominé, $G:\Theta\to\Xi:=G(\Theta)$ une bijection de Θ sur son image, et $\{\mathbb{Q}_\xi\,,\ \xi\in\Xi\}$ le modèle reparamétré, défini par $\mathbb{Q}_\xi:=\mathbb{P}_{G^{-1}(\xi)}$ pour tout $\xi\in\Xi$. Si $\hat\theta_n^{\mathrm{MV}}$ est un estimateur du maximum de vraisemblance pour le modèle $\{\mathbb{P}_\theta\,,\ \theta\in\Theta\}$, alors $\hat\xi_n:=G(\hat\theta_n^{\mathrm{MV}})$ est un estimateur du maximum de vraisemblance pour le modèle $\{\mathbb{Q}_\xi\,,\ \xi\in\Xi\}$. De plus, l'ensemble des estimateurs du maximum de vraisemblance du modèle reparamétré est l'image par G de celui du modèle initial.

Démonstration

Notons $\widetilde{\mathrm{L}_n}$ la vraisemblance du modèle reparamétré. Pour tout $\xi\in\Xi$, la loi $\mathbb{Q}_\xi$ a pour densité celle de $\mathbb{P}_{G^{-1}(\xi)}$, donc $\widetilde{\mathrm{L}_n}(\xi,X_1,\dots,X_n)=\mathrm{L}_n(G^{-1}(\xi),X_1,\dots,X_n)$. Ainsi, pour tout $\xi\in\Xi$, en posant $\vartheta=G^{-1}(\xi)\in\Theta$,

$$\widetilde{\mathrm{L}_n}(\xi,X_1,\dots,X_n)=\mathrm{L}_n(\vartheta,X_1,\dots,X_n)\le\mathrm{L}_n(\hat\theta_n^{\mathrm{MV}},X_1,\dots,X_n)=\widetilde{\mathrm{L}_n}\big(G(\hat\theta_n^{\mathrm{MV}}),X_1,\dots,X_n\big)\,,$$

ce qui montre que $G(\hat\theta_n^{\mathrm{MV}})$ maximise $\widetilde{\mathrm{L}_n}$. La même chaîne d'inégalités, lue dans l'autre sens à l'aide de G^{-1}, montre que tout maximum de $\widetilde{\mathrm{L}_n}$ est l'image par G d'un maximum de L_n. □

Exemple I-2.1.14: Reparamétrisations des modèles gaussien et exponentiel

1. Considérons le modèle gaussien paramétré en *moyenne et variance*, $\theta=(\mu,\sigma^2)\in\Theta=\mathbb{R}\times\mathbb{R}_+^*$, dont l'estimateur du maximum de vraisemblance est, pour $n\ge2$, $(\hat\mu_n,\hat\sigma_n^2)=\big(\bar X_n,n^{-1}\sum_{i=1}^n(X_i-\bar X_n)^2\big)$ (Exemple I-2.1.10). Le même modèle paramétré en *moyenne et écart-type*, $\xi=(\mu,\sigma)\in\Xi=\mathbb{R}\times\mathbb{R}_+^*$, est l'image du premier par la bijection $G:(\mu,\sigma^2)\mapsto(\mu,\sqrt{\sigma^2})$ de Θ sur Ξ. Sans nouveau calcul, son estimateur du maximum de vraisemblance est
$$\hat\mu_n=\bar X_n\,,\qquad\hat\sigma_n=\Big(\frac1n\sum_{i=1}^n(X_i-\bar X_n)^2\Big)^{1/2}\,.$$
2. Dans le modèle exponentiel, l'estimateur du maximum de vraisemblance de l'intensité θ est $1/\bar X_n$ (Exemple I-2.1.7). Le modèle paramétré par l'espérance $\tau=1/\theta\in\mathbb{R}_+^*$ est l'image du premier par la bijection $G:\theta\mapsto1/\theta$; l'estimateur du maximum de vraisemblance de τ est donc $\hat\tau_n=\bar X_n$.

I-2.1.4 Comprendre le maximum de vraisemblance

Pourquoi maximiser la vraisemblance conduit-il vers la vraie valeur du paramètre ? La réponse tient en deux faits : la log-vraisemblance normalisée est une moyenne empirique, donc elle converge vers une fonction limite ; et cette fonction limite atteint son maximum au vrai paramètre, ce qui s'exprime à l'aide de la divergence de Kullback-Leibler.

Deux paramètres à ne pas confondre. Soit $(X_1,\dots,X_n)$ un n-échantillon du modèle dominé $(\mathsf{X},\mathcal{X},\{q_\theta\cdot\mu\,,\ \theta\in\Theta\})$. Dans la suite, pour ne pas se perdre, on distingue deux rôles :

— $\theta^\star\in\Theta$ est le paramètre « hypothèse », la valeur du paramètre sous laquelle les observations sont tirées : on travaille sous $\mathbb{P}_{\theta^\star}$;

— $\vartheta\in\Theta$ est le paramètre libre, la variable par rapport à laquelle on exprime $\mathrm{L}_n(\vartheta)$ et on maximise.

Les espérances sont prises sous la loi des observations, c'est-à-dire sous $\mathbb{P}_{\theta^\star}$, dont la densité est $q_{\theta^\star}$.

La log-vraisemblance normalisée converge. Pour ϑ fixé, $\ell_n(\vartheta) = n^{-1}\sum_{i=1}^n \log q_\vartheta(X_i)$ est la moyenne empirique des variables aléatoires $\log q_\vartheta(X_i)$, qui sont indépendantes et de même loi sous $\mathbb{P}_{\theta^\star}$. La loi faible des grands nombres s'applique dès qu'elles sont intégrables, et donne la limite $\mathbb{E}_{\theta^\star}[\log q_\vartheta(X_1)]$. Il reste à comprendre où cette limite, vue comme fonction de ϑ, atteint son maximum. En écrivant $\log q_\vartheta = \log(q_\vartheta/q_{\theta^\star}) + \log q_{\theta^\star}$, on fait apparaître l'espérance sous $\mathbb{P}_{\theta^\star}$ du logarithme d'un rapport de densités : c'est, au signe près, la divergence de Kullback-Leibler entre $\mathbb{Q}_{\theta^\star}$ et $\mathbb{Q}_\vartheta$.

Définition I-2.1.7: Divergence de Kullback-Leibler

Soient $\mathbb{P}_0$ et $\mathbb{P}_1$ deux probabilités sur un espace mesurable $(\mathsf{X}, \mathcal{X})$, admettant des densités f_0 et f_1 par rapport à une mesure σ-finie μ sur $(\mathsf{X}, \mathcal{X})$. On appelle *divergence de Kullback-Leibler* entre les lois $\mathbb{P}_0$ et $\mathbb{P}_1$ la quantité

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

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

Cette intégrale est toujours bien définie, à valeurs dans $]-\infty, +\infty]$: la partie négative de l'intégrande est μ-intégrable, ce qui est démontré dans les compléments (Section I-2.4), où l'on montre aussi que $\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_1)$ ne dépend pas du choix de la mesure dominante μ. Si X est une variable aléatoire de loi $\mathbb{P}_0$, alors $f_0(X) > 0$ presque sûrement et la divergence s'écrit comme une espérance,

$$\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_1) = \mathbb{E}\Big[\log \frac{f_0(X)}{f_1(X)}\Big]\ .$$

La divergence n'est pas une distance : elle n'est pas symétrique en $(\mathbb{P}_0, \mathbb{P}_1)$ et ne vérifie pas l'inégalité triangulaire. Elle en a cependant la propriété essentielle.

Théorème I-2.1.1: Positivité de la divergence de Kullback-Leibler

Soient $\mathbb{P}_0$ et $\mathbb{P}_1$ deux probabilités sur $(\mathsf{X}, \mathcal{X})$, de densités f_0 et f_1 par rapport à une mesure σ-finie μ. Alors $\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_1) \in [0, +\infty]$, et $\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_1) = 0$ 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.

Démonstration

Notons $A := \{x \in \mathsf{X} :\ f_0(x) > 0\}$; on a $\mathbb{P}_0(A) = \int_A f_0\, \mathrm{d}\mu = 1$. Soit X une variable aléatoire de loi $\mathbb{P}_0$: presque sûrement $X \in A$, et l'on peut poser $Y := f_1(X)/f_0(X) \in [0, +\infty[$. Par le théorème de transfert, $\mathbb{E}[Y] = \int_A f_1\, \mathrm{d}\mu = \mathbb{P}_1(A) \le 1$, de sorte que Y est intégrable, et, par définition, $\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_1) = \int_A f_0 \log(f_0/f_1)\, \mathrm{d}\mu = \mathbb{E}[-\log Y]$, avec $-\log 0 = +\infty$.
Premier cas : $\mathbb{P}(Y = 0) > 0$. L'intégrande de la définition vaut $+\infty$ sur l'ensemble $\{f_0 > 0, f_1 = 0\}$, qui est de $\mathbb{P}_0$-probabilité $\mathbb{P}(Y = 0) > 0$, donc de μ-mesure strictement positive. Sa partie négative étant μ-intégrable, $\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_1) = +\infty$, et l'inégalité est vraie.
Second cas : $\mathbb{P}(Y = 0) = 0$. Alors Y est presque sûrement à valeurs dans l'intervalle ouvert $]0, +\infty[$, sur lequel la fonction $-\log$ est strictement convexe, et Y est intégrable. L'inégalité de Jensen (Théorème I-2.2.2) donne

$$\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_1) = \mathbb{E}[-\log Y] \ge -\log \mathbb{E}[Y] = -\log \mathbb{P}_1(A) \ge 0\ ,$$

la dernière inégalité venant de $\mathbb{P}_1(A) \le 1$.
Cas d'égalité. Si $\mathbb{P}_0 = \mathbb{P}_1$, alors $f_0 = f_1$ μ-presque partout, et $\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_1) = \int_A f_0 \log 1\, \mathrm{d}\mu = 0$. Réciproquement, supposons $\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_1) = 0$. On est nécessairement dans le second cas, et les deux inégalités ci-dessus sont des égalités : d'une part $\mathbb{P}_1(A) = 1$, donc $\mathbb{E}[Y] = 1$; d'autre part l'inégalité de Jensen est une égalité, ce qui, la fonction $-\log$ étant strictement convexe, impose

que Y soit presque sûrement constante, égale à $\mathbb{E}[Y] = 1$. Ainsi $f_1 = f_0$ $\mathbb{P}_0$-presque sûrement, c'est-à-dire μ-presque partout sur A. Il en résulte $\int_{A^c} f_1 \,\mathrm{d}\mu = 1 - \int_A f_1 \,\mathrm{d}\mu = 1 - \int_A f_0 \,\mathrm{d}\mu = 0$, donc $f_1 = 0 = f_0$ μ-presque partout sur A^c. Finalement $f_0 = f_1$ μ-presque partout, et $\mathbb{P}_0 = \mathbb{P}_1$. □

Proposition I-2.1.3: Limite de la log-vraisemblance normalisée

Soient $(X_1, \dots, X_n)$ un n-échantillon du modèle dominé $(\mathsf{X}, \mathcal{X}, \{q_\theta \cdot \mu\,,\ \theta \in \Theta\})$ et $\theta^\star \in \Theta$. On suppose que, pour tout $\vartheta \in \Theta$, $\mathbb{E}_{\theta^\star}\big[|\log q_\vartheta(X_1)|\big] < \infty$, et l'on pose

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

(i) Pour tout $\vartheta \in \Theta$, $\ell_n(\vartheta, X_1, \dots, X_n) \overset{\mathbb{P}_{\theta^\star}-\text{prob}}{\longrightarrow} M_{\theta^\star}(\vartheta)$.

(ii) Pour tout $\vartheta \in \Theta$, la divergence $\mathrm{KL}(\mathbb{Q}_{\theta^\star}, \mathbb{Q}_\vartheta)$ est finie et

$$M_{\theta^\star}(\vartheta) = M_{\theta^\star}(\theta^\star) - \mathrm{KL}(\mathbb{Q}_{\theta^\star}, \mathbb{Q}_\vartheta) \le M_{\theta^\star}(\theta^\star)\,.$$

(iii) Si le modèle $(\mathsf{X}, \mathcal{X}, \{q_\theta \cdot \mu\,,\ \theta \in \Theta\})$ est identifiable (Définition I-1.2), alors $M_{\theta^\star}(\vartheta) < M_{\theta^\star}(\theta^\star)$ pour tout $\vartheta \neq \theta^\star$: la fonction $M_{\theta^\star}$ admet un maximum unique, atteint en $\theta^\star$.

Démonstration

(i) Sous $\mathbb{P}_{\theta^\star}$, les variables aléatoires $\log q_\vartheta(X_i)$, $i = 1, \dots, n$, sont indépendantes, de même loi et intégrables par hypothèse ; la loi faible des grands nombres (Théorème I-2.2.1) donne la convergence de leur moyenne empirique $\ell_n(\vartheta)$ vers leur espérance commune $M_{\theta^\star}(\vartheta)$.
(ii) Sous $\mathbb{P}_{\theta^\star}$, on a $q_{\theta^\star}(X_1) > 0$ presque sûrement, car $\mathbb{P}_{\theta^\star}(q_{\theta^\star}(X_1) = 0) = \int_{\{q_{\theta^\star}=0\}} q_{\theta^\star} \,\mathrm{d}\mu = 0$, et de même $q_\vartheta(X_1) > 0$ presque sûrement, sans quoi $|\log q_\vartheta(X_1)|$ vaudrait $+\infty$ avec probabilité strictement positive, contredisant l'intégrabilité. La variable $\log\big(q_{\theta^\star}(X_1)/q_\vartheta(X_1)\big) = \log q_{\theta^\star}(X_1) - \log q_\vartheta(X_1)$ est donc presque sûrement finie et intégrable, d'espérance $M_{\theta^\star}(\theta^\star) - M_{\theta^\star}(\vartheta)$. Or, par le théorème de transfert, cette espérance est $\int_{\{q_{\theta^\star}>0\}} q_{\theta^\star} \log(q_{\theta^\star}/q_\vartheta) \,\mathrm{d}\mu = \mathrm{KL}(\mathbb{Q}_{\theta^\star}, \mathbb{Q}_\vartheta)$, qui est donc finie, et positive d'après le Théorème I-2.1.1.
(iii) Si le modèle est identifiable et $\vartheta \neq \theta^\star$, alors $\mathbb{Q}_\vartheta \neq \mathbb{Q}_{\theta^\star}$, donc $\mathrm{KL}(\mathbb{Q}_{\theta^\star}, \mathbb{Q}_\vartheta) > 0$ d'après le cas d'égalité du Théorème I-2.1.1, et (ii) donne $M_{\theta^\star}(\vartheta) < M_{\theta^\star}(\theta^\star)$. □

La quantité $M_{\theta^\star}(\theta^\star) = \int_{\mathsf{X}} q_{\theta^\star} \log q_{\theta^\star} \,\mathrm{d}\mu$ est notée $\mathrm{Ent}(\mathbb{Q}_{\theta^\star})$ dans le chapitre d'origine du polycopié, qui l'appelle entropie de la loi $\mathbb{Q}_{\theta^\star}$; c'est l'opposé de l'entropie de Shannon, et, contrairement à la divergence, elle dépend de la mesure dominante. La décomposition du point (ii) s'écrit alors $M_{\theta^\star} = -\mathrm{KL}(\mathbb{Q}_{\theta^\star}, \mathbb{Q}_\cdot) + \mathrm{Ent}(\mathbb{Q}_{\theta^\star})$: la fonction limite de la log-vraisemblance est, à une constante près, l'opposé de la divergence de Kullback-Leibler entre la vraie loi et la loi candidate.

Intuition. *La suite des estimateurs du maximum de vraisemblance $\hat\theta_n^{\mathrm{MV}}$ maximise ℓ_n, qui converge vers $M_{\theta^\star}$, dont le maximum est atteint en $\theta^\star$ et en $\theta^\star$ seulement. On s'attend donc à ce que $\hat\theta_n^{\mathrm{MV}}$ converge vers $\theta^\star$: c'est la consistance de l'estimateur du maximum de vraisemblance. Ce raisonnement est heuristique : la proposition donne la convergence de $\ell_n(\vartheta)$ pour chaque ϑ fixé, alors que la convergence du point de maximum demande que l'approximation soit uniforme en ϑ, au moins au voisinage de $\theta^\star$, ce qui requiert des hypothèses supplémentaires (loi uniforme des grands nombres). Ces hypothèses et la démonstration font l'objet des cours 6 et 7. Le modèle de Bernoulli, pour lequel $M_{\theta^\star}$ et la divergence sont explicites, est traité en détail dans les compléments (Section I-2.4), avec une figure qui montre ℓ_n se rapprocher de $M_{\theta^\star}$ lorsque n augmente.*

Illustration / animation : La log-vraisemblance normalisée converge vers une fonction dont l'écart au maximum est la divergence de Kullback-Leibler

On illustre ici, pour $\theta^\star$ et n choisis dans le modèle de Bernoulli, la log-vraisemblance normalisée et sa limite, dont l'écart vertical en ϑ est la divergence $\mathrm{KL}(\mathcal{B}\mathrm{er}(\theta^\star), \mathcal{B}\mathrm{er}(\vartheta))$.

link

Résumé et généralisation : les M-estimateurs. L'estimateur du maximum de vraisemblance est défini comme la solution d'un problème de *maximisation* : on maximise la moyenne empirique $\ell_n(\vartheta) = n^{-1}\sum_{i=1}^n \log q_\vartheta(X_i)$, qui converge vers une fonction limite $M_{\theta^\star}$ dont le maximum est atteint en $\theta^\star$. Rien, dans ce mécanisme, n'est propre au logarithme de la densité : on peut remplacer $\log q_\vartheta(x)$ par une autre fonction $m(\vartheta, x)$, pourvu que la fonction limite correspondante atteigne son maximum au vrai paramètre. C'est une très vaste classe de méthodes, celle des M-estimateurs.

Définition I-2.1.8: M-estimateur

Soit $(X_1, \dots, X_n)$ un n-échantillon du modèle $(\mathsf{X}, \mathcal{X}, \{\mathbb{Q}_\theta\,,\ \theta \in \Theta\})$. Soit $m : \Theta \times \mathsf{X} \to \mathbb{R}$, $(\vartheta, x) \mapsto m(\vartheta, x)$, une fonction mesurable en x telle que $\mathbb{E}_{\theta^\star}[|m(\vartheta, X_1)|] < \infty$ pour tous $\vartheta, \theta^\star \in \Theta$. Pour tout $\theta^\star \in \Theta$, on pose

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

et l'on suppose que, pour tout $\theta^\star \in \Theta$, la fonction $M_{\theta^\star}$ atteint son maximum au point $\theta^\star$. La fonction $M_{\theta^\star}$ n'est pas connue, mais on peut l'estimer par la moyenne empirique

$$M_n(\vartheta) := \frac{1}{n}\sum_{i=1}^n m(\vartheta, X_i)\,, \qquad \vartheta \in \Theta\,,$$

qui vérifie $M_n(\vartheta) \overset{\mathbb{P}_{\theta^\star}-\mathrm{prob}}{\longrightarrow} M_{\theta^\star}(\vartheta)$ pour tous $\vartheta, \theta^\star \in \Theta$, par la loi faible des grands nombres. On appelle *M-estimateur* associé à m toute solution du problème d'optimisation

$$\hat\theta_n \in \operatorname*{arg\,max}_{\vartheta\in\Theta} M_n(\vartheta)\,.$$

Sous l'hypothèse d'intégrabilité de la Proposition I-2.1.3, l'estimateur du maximum de vraisemblance est le M-estimateur associé à $m(\vartheta, x) = \log q_\vartheta(x)$, la fonction limite étant $M_{\theta^\star} = M_{\theta^\star}(\theta^\star) - \mathrm{KL}(\mathbb{Q}_{\theta^\star}, \mathbb{Q}_\cdot)$, dont le maximum est atteint en $\theta^\star$ par positivité de la divergence. Dans de très nombreux cas, le problème d'optimisation n'admet pas de solution explicite, et un M-estimateur s'obtient en pratique par une procédure numérique. Dans certains cas, il est plus naturel de considérer un problème de *minimisation*, ce qui revient à changer m en $-m$. Enfin, si Θ est ouvert et si $\vartheta \mapsto m(\vartheta, x)$ est différentiable pour tout x, un M-estimateur vérifie $n^{-1}\sum_{i=1}^n \nabla_\vartheta m(\hat\theta_n, X_i) = 0$: c'est un Z-estimateur associé à la fonction d'estimation $\boldsymbol{\psi} = \nabla_\vartheta m$, sous réserve que $\mathbb{E}_{\theta^\star}[\nabla_\vartheta m(\theta^\star, X_1)] = 0$, ce qui est le cas lorsque l'on peut dériver sous l'espérance, puisque $M_{\theta^\star}$ est maximale en $\theta^\star$. Il existe des Z-estimateurs qui ne dérivent d'aucun problème d'optimisation, et des M-estimateurs associés à des fonctions m non différentiables, qui ne sont donc pas des Z-estimateurs au sens ci-dessus.

Exemple I-2.1.15: Moindres carrés et médiane pour le paramètre de translation

Reprenons le modèle de translation de l'Exemple I-2.1.6 : f est une densité paire sur $\mathbb{R}$ et $q_\theta(x) = f(x-\theta)$, $\theta \in \Theta = \mathbb{R}$. Sous $\mathbb{P}_{\theta^\star}$, la variable $X_1 - \theta^\star$ a pour densité f, donc a une loi symétrique.

1. *Moindres carrés.* Supposons $\sigma^2 := \int_{\mathbb{R}} x^2 f(x)\,\mathrm{d}x < \infty$ et posons $m(\vartheta, x) := -(x-\vartheta)^2$.

Pour tous $\vartheta, \theta^\star \in \mathbb{R}$, comme $\mathbb{E}_{\theta^\star}[X_1] = \theta^\star$ et $\mathrm{Var}_{\theta^\star}(X_1) = \sigma^2$,

$$M_{\theta^\star}(\vartheta) = -\mathbb{E}_{\theta^\star}\left[(X_1 - \vartheta)^2\right] = -(\theta^\star - \vartheta)^2 - \sigma^2 \,,$$

fonction qui admet un maximum unique au point $\theta^\star$. Le M-estimateur associé à m maximise $M_n(\vartheta) = -n^{-1}\sum_{i=1}^n (X_i - \vartheta)^2$, fonction quadratique strictement concave, dont l'unique maximum est $\hat\theta_n = \bar X_n$. Il est plus naturel ici de parler de *minimisation* de $n^{-1}\sum_{i=1}^n (X_i - \vartheta)^2$: l'estimateur obtenu est appelé *estimateur des moindres carrés*. La fonction $\vartheta \mapsto m(\vartheta, x)$ est dérivable, de dérivée $2(x - \vartheta)$, et la moyenne empirique est aussi le Z-estimateur associé à $\psi(\vartheta, x) = x - \vartheta$: elle est à la fois un estimateur des moments, un Z-estimateur et un M-estimateur.

2. *Médiane.* Supposons $\int_{\mathbb{R}} |x| f(x)\,\mathrm{d}x < \infty$ et posons $m(\vartheta, x) := -|x - \vartheta|$. La fonction $g : \vartheta \mapsto \mathbb{E}_{\theta^\star}|X_1 - \vartheta|$ vérifie $g(\vartheta) = g(2\theta^\star - \vartheta)$, par symétrie de la loi de $X_1 - \theta^\star$, et, pour $\theta^\star \le \vartheta$,

$$g(\vartheta) - g(\theta^\star) = \mathbb{E}_{\theta^\star}\Big[\int_{\theta^\star}^{\vartheta} \left(\mathbb{1}_{\{X_1 < t\}} - \mathbb{1}_{\{X_1 > t\}}\right) \mathrm{d}t\Big] = \int_{\theta^\star}^{\vartheta} \left(2F_{\theta^\star}(t) - 1\right) \mathrm{d}t \ge 0 \,,$$

où $F_{\theta^\star}(t) = \mathbb{P}_{\theta^\star}(X_1 \le t)$ est la fonction de répartition de X_1, continue puisque X_1 a une densité, et vérifie $F_{\theta^\star}(t) \ge F_{\theta^\star}(\theta^\star) = 1/2$ pour $t \ge \theta^\star$; on a utilisé l'identité $|x - \vartheta| - |x - \theta^\star| = \int_{\theta^\star}^{\vartheta} (\mathbb{1}_{\{x<t\}} - \mathbb{1}_{\{x>t\}})\,\mathrm{d}t$, valable pour tout réel x, puis le théorème de Fubini. Ainsi $M_{\theta^\star} = -g$ atteint son maximum en $\theta^\star$. Le M-estimateur associé à m maximise $M_n(\vartheta) = -n^{-1}\sum_{i=1}^n |X_i - \vartheta|$, dont les points de maximum ont été déterminés à l'Exemple I-2.1.13 : la médiane empirique est un M-estimateur, unique si n est impair. La fonction $\vartheta \mapsto |x - \vartheta|$ n'étant pas dérivable en x, ce M-estimateur ne s'obtient pas en annulant un gradient ; c'est cependant un Z-estimateur, associé à $\psi(\vartheta, x) = \mathrm{signe}(x - \vartheta)$, comme on l'a vu à l'Exemple I-2.1.6.

I-2.1.5 En pratique : programme des exercices

Les notions de ce chapitre sont mises en œuvre dans la feuille d'exercices (PC 2), intégrée au chapitre (Section I-2.3). Exercices traités en petite classe :

Exercice	Notions du cours	Rappels
Exo 1 – *Modèle linéaire gaussien*	EMV, équations de vraisemblance	Vecteurs gaussiens, th. de Cochran, loi du χ^2
Exo 2 – *Coefficient de mélange*	EMV et EMM comparés	Loi binomiale, biais
Exo 3 – *Famille exponentielle*	EMV = EMM, invariance	Dérivation sous l'intégrale, convexité

L'Exercice 1 (PC2) poursuit l'Exercice 4 (PC1) sur le modèle de régression linéaire gaussienne, sujet récurrent du cours : on y calcule l'estimateur du maximum de vraisemblance des coefficients et de la variance, puis la loi de ces estimateurs. Pour les questions 5 et 6, la loi de $\hat\sigma^2$ et son indépendance avec $(\hat\beta_1, \hat\beta_2)$ résultent du théorème de Cochran, énoncé et démontré au chapitre 1 (Théorème I-1.1.1) : le vecteur $\mathbf{Y} - \mathbf{X}\beta$ suit la loi $\mathcal{N}(0, \sigma^2 \mathrm{I}_n)$, l'estimateur $\hat\beta$ est une fonction de la projection orthogonale de $\mathbf{Y}$ sur l'image de $\mathbf{X}$, sous-espace de dimension 2, tandis que $n\hat\sigma^2 = \|\mathbf{Y} - \mathbf{X}\hat\beta\|^2$ est le carré de la norme de la projection sur l'orthogonal de ce sous-espace, de dimension $n - 2$; le théorème donne l'indépendance des deux projections et la loi $\chi^2(n-2)$ de $n\hat\sigma^2/\sigma^2$. La question 7 demande un intervalle de confiance pour σ^2, notion définie a minima dans les outils (Section I-2.2) et développée au chapitre suivant.

L'Exercice 2 (PC2) compare, sur un modèle de mélange de deux lois uniformes, l'estimateur du maximum de vraisemblance et un estimateur des moments : les deux méthodes donnent des estimateurs différents, dont le biais et la dispersion sont comparés sur des simulations. L'Exercice 3

(PC2) étudie les familles exponentielles, pour lesquelles l'estimateur du maximum de vraisemblance est un estimateur des moments, comme on l'a observé pour les lois de Bernoulli et de Poisson.

Illustration / animation : L'estimateur seuillé du mélange prend souvent la valeur 0 quand n est petit

L'animation accessible par le QR code ci-contre permet de comparer, pour les quatre estimateurs du modèle de mélange, les boîtes à moustaches obtenues sur des échantillons répétés à mesure que n augmente, et d'observer la fréquence à laquelle l'estimateur seuillé vaut exactement 0.

link

Les exercices bonus (durée de vie, qui ne mobilise que le chapitre 1, et modèle auto-régressif, première vraisemblance hors du cadre du n-échantillon) sont accompagnés d'un guide de lecture dans le fascicule.

Notions de probabilités clés utilisées cette semaine, à revoir : espérance et variance, indépendance, loi des grands nombres, inégalité de Bienaymé-Tchebychev, fonctions convexes et inégalité de Jensen, vecteurs gaussiens et théorème de Cochran (voir la partie *Outils mathématiques* ci-dessous).