I-2.4 Les compléments

Les contenus de cette partie ne sont pas exigibles ; ils prolongent les fondamentaux sur des modèles moins standards et donnent les énoncés complets de résultats dont seule une partie a été utilisée dans le cours.

Méthode des moments sans moments : le modèle de Cauchy

La méthode des moments n'exige pas que les observations aient des moments : il suffit de trouver une fonction T bornée dont l'espérance dépende du paramètre de façon inversible. Le modèle de Cauchy, qui n'a pas d'espérance, en est l'exemple.

Exemple I-2.4.1: Estimateur des moments du modèle de Cauchy

Soit $(X_1, \dots, X_n)$ un n-échantillon du modèle de Cauchy $\big(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{\text{Cauchy}(\theta)\,,\ \theta \in \Theta = \mathbb{R}\}\big)$, où la loi de Cauchy de paramètre de translation θ et d'échelle 1 a pour densité par rapport à la mesure de Lebesgue

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

Cette loi n'a pas de moment d'ordre $k \geq 1$, puisque $\int_{\mathbb{R}} |x|^k q_\theta(x)\,\mathrm{d}x = +\infty$: le choix $T(x) = x^k$ est exclu. Prenons $T(x) = \text{signe}(x)$, fonction bornée, avec la convention $\text{signe}(0) = 0$, sans incidence puisque la loi a une densité. Notons $F(t) = \frac{1}{\pi}\int_{-\infty}^{t} \frac{\mathrm{d}x}{1+x^2} = \frac{1}{\pi}\arctan(t) + \frac{1}{2}$ la fonction de répartition de la loi de Cauchy de paramètre 0 ; celle de X_1 sous $\mathbb{P}_\theta$ est $t \mapsto F(t-\theta)$, et

$$\begin{aligned} e(\theta) = \mathbb{E}_\theta[\text{signe}(X_1)] &= \mathbb{P}_\theta(X_1 > 0) - \mathbb{P}_\theta(X_1 < 0) \\ &= \big(1 - F(-\theta)\big) - F(-\theta) = 1 - 2F(-\theta) = \frac{2}{\pi}\arctan(\theta)\,, \end{aligned}$$

où l'on a utilisé $\arctan(-\theta) = -\arctan(\theta)$. La fonction e est une bijection continue et strictement croissante de $\mathbb{R}$ sur $]-1,1[$, d'inverse $y \mapsto \tan(\pi y/2)$. En posant $Y_i := \text{signe}(X_i)$, l'équation $e(\vartheta) = \bar{Y}_n$ admet une solution unique dès que $\bar{Y}_n \in]-1,1[$, c'est-à-dire dès que les observations ne sont pas toutes de même signe,

$$\hat{\theta}_n := \tan\Big(\frac{\pi}{2}\,\bar{Y}_n\Big) = \tan\Big(\frac{\pi}{2n}\sum_{i=1}^{n} \text{signe}(X_i)\Big)\,.$$

D'après la Proposition I-2.1.1, $\bar{Y}_n \overset{\mathbb{P}_\theta-\text{prob}}{\longrightarrow} e(\theta)$; l'événement $\{|\bar{Y}_n| = 1\}$, de probabilité $F(-\theta)^n + (1-F(-\theta))^n$, devient négligeable lorsque n croît, et l'argument du Lemme I-2.2.1, la fonction e^{-1} étant continue sur $]-1,1[$, montre que $\hat{\theta}_n \overset{\mathbb{P}_\theta-\text{prob}}{\longrightarrow} \theta$, quelle que soit la valeur, arbitraire, donnée à $\hat{\theta}_n$ sur cet événement.

Modèle de translation et d'échelle : estimateur des moments

Le modèle de translation et d'échelle décrit une observation obtenue à partir d'une loi de référence connue par un décalage et un changement d'unité inconnus ; les modèles gaussien et de Laplace en sont des cas particuliers. Il demande deux fonctions de moment, et le système obtenu se résout explicitement.

Exemple I-2.4.2: Estimateur des moments du modèle de translation et d'échelle

Soit q une densité de probabilité sur $\mathbb{R}$, par rapport à la mesure de Lebesgue, vérifiant

$$\int_{\mathbb{R}} x\,q(x)\,\mathrm{d}x = 0\,, \qquad m_2 := \int_{\mathbb{R}} x^2 q(x)\,\mathrm{d}x \in]0, +\infty[\ .$$

On peut prendre pour q la densité gaussienne centrée réduite, $q(x) = (2\pi)^{-1/2}\exp(-x^2/2)$, auquel cas $m_2 = 1$, ou la densité de Laplace $q(x) = \frac{1}{2}\exp(-|x|)$, auquel cas $m_2 = 2$. Considérons un n-échantillon $(X_1, \dots, X_n)$ du modèle $(\mathsf{X}, \mathcal{X}, \{q_\theta \cdot \mu\,,\ \theta \in \Theta\})$ avec

$$q_\theta(x) := \frac{1}{\sigma}\, q\Big(\frac{x-m}{\sigma}\Big)\,, \qquad \theta = (m, \sigma^2) \in \Theta := \mathbb{R} \times \mathbb{R}_+^*\,,$$

où $\sigma = \sqrt{\sigma^2}$. Par la formule de changement de variable (Section I-1.2), sous $\mathbb{P}_\theta$, X_1 a même loi que $m + \sigma Z$ où Z a pour densité q ; donc

$$\mathbb{E}_\theta[X_1] = m\,, \qquad \mathrm{Var}_\theta(X_1) = \sigma^2 m_2\,, \qquad \mathbb{E}_\theta[X_1^2] = \sigma^2 m_2 + m^2\,.$$

Avec $T_1(x) = x$ et $T_2(x) = x^2$, l'estimateur des moments est solution du système

$$\begin{cases} e_1(m, \sigma^2) = m = \dfrac{1}{n}\displaystyle\sum_{i=1}^n X_i\,, \\ e_2(m, \sigma^2) = \sigma^2 m_2 + m^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^n X_i^2\,. \end{cases}$$

L'application $\mathbf{e} : (m, \sigma^2) \mapsto (m,\ \sigma^2 m_2 + m^2)$ est injective sur $\mathbb{R} \times \mathbb{R}_+^*$, d'inverse explicite $(e_1, e_2) \mapsto \big(e_1, (e_2 - e_1^2)/m_2\big)$; sa matrice jacobienne

$$\begin{bmatrix} 1 & 0 \\ 2m & m_2 \end{bmatrix}$$

est inversible en tout point, de déterminant $m_2 > 0$. Le système admet donc une solution unique, qui appartient à Θ dès que $V_n := n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2 > 0$, ce qui est $\mathbb{P}_\theta$-presque sûr pour $n \geq 2$:

$$\hat{m}_n = \frac{1}{n}\sum_{i=1}^n X_i\,, \qquad \hat{\sigma}_n^2 = \frac{1}{n\, m_2}\sum_{i=1}^n (X_i - \hat{m}_n)^2\,.$$

Dans le cas gaussien ($m_2 = 1$), cet estimateur des moments coïncide avec l'estimateur du maximum de vraisemblance de l'Exemple I-2.1.10. L'hypothèse supplémentaire $\int_{\mathbb{R}} x^4 q(x)\,\mathrm{d}x < \infty$, faite dans le chapitre d'origine et dans les slides, n'intervient pas dans la construction ; elle sert à l'étude de la loi limite de $\hat{\sigma}_n^2$, par le théorème central limite appliqué aux X_i^2.

Estimateurs robustes de translation et d'échelle : médiane et MAD

Les fonctions d'estimation d'un Z-estimateur ne sont pas tenues d'être des différences de moments : on peut les choisir bornées, ce qui rend l'estimateur peu sensible aux observations aberrantes. La médiane et l'écart absolu médian en sont l'exemple de base.

Exemple I-2.4.3: Médiane et écart absolu médian comme Z-estimateurs

Soit $(X_1, \dots, X_n)$ un n-échantillon du modèle $(\mathsf{X}, \mathcal{X}, \{q_\theta \cdot \mu\,,\ \theta \in \Theta\})$ avec

$$q_\theta(x) = \frac{1}{\sigma}\, q\Big(\frac{x-m}{\sigma}\Big)\,, \qquad \theta = (m, \sigma) \in \Theta = \mathbb{R} \times \mathbb{R}_+^*\,,$$

où q est une densité de probabilité paire sur $\mathbb{R}$, de fonction de répartition $Q(x) = \int_{-\infty}^x q(z)\,\mathrm{d}z$, continue, avec $Q(0) = 1/2$ par parité. Soit $a > 0$ un réel tel que $Q(a) = 3/4$; si q est la densité gaussienne centrée réduite, $a \simeq 0{,}6745$. Considérons les fonctions d'estimation

$$\psi_1(\theta, x) := \mathrm{signe}(x - m) \qquad \text{et} \qquad \psi_2(\theta, x) := \mathrm{signe}\,\Big(\frac{|x-m|}{\sigma} - a\Big)\,,$$

avec $\mathrm{signe}(0) = 0$. Sous $\mathbb{P}_\theta$, la variable $Z := (X_1 - m)/\sigma$ a pour densité q, et

$$\mathbb{E}_\theta[\psi_1(\theta, X_1)] = \int_{\mathbb{R}} \mathrm{signe}(z)\, q(z)\, \mathrm{d}z = 0 \,,$$

par parité de q ; de même,

$$\begin{aligned}\mathbb{E}_\theta[\psi_2(\theta, X_1)] &= \int_{\mathbb{R}} \mathrm{signe}(|z| - a)\, q(z)\, \mathrm{d}z = 2\int_0^{+\infty} \mathrm{signe}(z - a)\, q(z)\, \mathrm{d}z \\ &= 2\Big(-\int_0^a q(z)\, \mathrm{d}z + \int_a^{+\infty} q(z)\, \mathrm{d}z\Big) \\ &= 2\big(-(Q(a) - \tfrac{1}{2}) + (1 - Q(a))\big) = 3 - 4Q(a) = 0 \,.\end{aligned}$$

Les Z-estimateurs associés à $\boldsymbol{\psi} = (\psi_1, \psi_2)^\top$ sont les solutions du système

$$\frac{1}{n}\sum_{i=1}^n \mathrm{signe}(X_i - m) = 0 \,, \qquad \frac{1}{n}\sum_{i=1}^n \mathrm{signe}\Big(\frac{|X_i - m|}{\sigma} - a\Big) = 0 \,.$$

Supposons, pour simplifier, n impair et les observations deux à deux distinctes. La première équation ne fait intervenir que m et a pour unique solution la médiane empirique, $\hat{m}_n = \mathrm{med}(X_1, \dots, X_n)$ (Exemple I-2.1.6). La seconde exprime que le nombre d'indices i tels que $|X_i - \hat{m}_n| > a\sigma$ égale le nombre d'indices tels que $|X_i - \hat{m}_n| < a\sigma$; si les écarts $|X_i - \hat{m}_n|$ sont deux à deux distincts, ce qui est presque sûr, elle a pour unique solution

$$\hat{\sigma}_n = \frac{1}{a}\,\mathrm{med}\big(|X_1 - \hat{m}_n|, \dots, |X_n - \hat{m}_n|\big) \,.$$

L'estimateur $\hat{\sigma}_n$ est appelé estimateur *MAD* (*median absolute deviation*, écart absolu médian) ; il est très utilisé en statistique robuste, car remplacer une observation par une valeur arbitrairement grande ne le modifie que de façon bornée, contrairement à l'écart-type empirique.

Le modèle de Bernoulli au microscope

Le modèle de Bernoulli permet de suivre en détail, avec des formules explicites, ce que la Proposition I-2.1.3 affirme en général : la log-vraisemblance normalisée converge vers une fonction limite, strictement concave, dont le maximum est atteint au vrai paramètre, et l'écart entre les deux se contrôle par l'inégalité de Bienaymé-Tchebychev.

Exemple I-2.4.4: Log-vraisemblance et divergence dans le modèle de Bernoulli

Reprenons le modèle de l'Exemple I-2.1.8, et fixons $\theta^\star \in]0,1[$. Pour $\vartheta \in]0,1[$, la log-vraisemblance normalisée s'écrit

$$\ell_n(\vartheta) = \frac{1}{n}\sum_{i=1}^n \big\{X_i \log\vartheta + (1 - X_i)\log(1 - \vartheta)\big\} = \bar{X}_n \log\vartheta + (1 - \bar{X}_n)\log(1 - \vartheta) \,.$$

La variable $\log q_\vartheta(X_1) = X_1 \log\vartheta + (1 - X_1)\log(1 - \vartheta)$ est bornée, donc intégrable, et comme $\mathbb{E}_{\theta^\star}[X_1] = \theta^\star$,

$$M_{\theta^\star}(\vartheta) = \mathbb{E}_{\theta^\star}[\log q_\vartheta(X_1)] = \theta^\star \log\vartheta + (1 - \theta^\star)\log(1 - \vartheta) \,.$$

Cette fonction est deux fois dérivable sur $]0,1[$, de dérivée $\theta^\star/\vartheta - (1 - \theta^\star)/(1 - \vartheta)$, nulle en $\vartheta = \theta^\star$, et de dérivée seconde $-\theta^\star/\vartheta^2 - (1 - \theta^\star)/(1 - \vartheta)^2 < 0$: elle est strictement concave

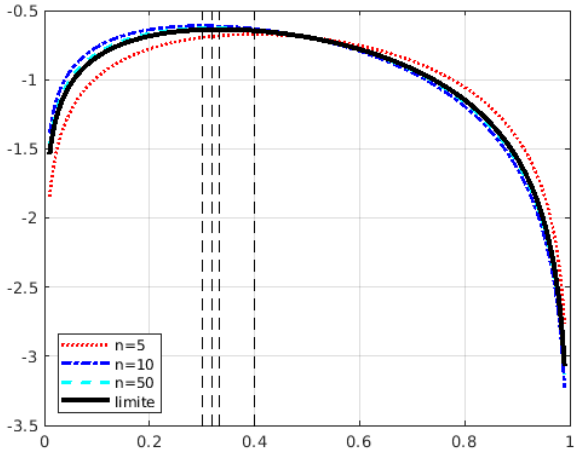
et atteint son maximum unique en $\theta^\star$ (Lemme I-2.2.3). La divergence de Kullback-Leibler est ici explicite,

$$\mathrm{KL}\big(\mathcal{B}\mathrm{er}(\theta^\star), \mathcal{B}\mathrm{er}(\vartheta)\big) = M_{\theta^\star}(\theta^\star) - M_{\theta^\star}(\vartheta) = \theta^\star \log\frac{\theta^\star}{\vartheta} + (1-\theta^\star)\log\frac{1-\theta^\star}{1-\vartheta}\ ,$$

strictement positive pour $\vartheta \neq \theta^\star$. L'écart entre $\ell_n(\vartheta)$ et $M_{\theta^\star}(\vartheta)$ se contrôle quantitativement : la variable $\log q_\vartheta(X_1)$ s'écrit $X_1 \log\frac{\vartheta}{1-\vartheta} + \log(1-\vartheta)$, de variance $\theta^\star(1-\theta^\star)\big(\log\frac{\vartheta}{1-\vartheta}\big)^2$, et l'inégalité de Bienaymé-Tchebychev donne, pour tout $\delta > 0$,

$$\mathbb{P}_{\theta^\star}\big(|\,\ell_n(\vartheta) - M_{\theta^\star}(\vartheta)| \geq \delta\big) \leq \frac{\theta^\star(1-\theta^\star)\big(\log\frac{\vartheta}{1-\vartheta}\big)^2}{n\,\delta^2}\ .$$

Pour ϑ fixé, cette probabilité tend vers 0 ; la borne se dégrade lorsque ϑ s'approche de 0 ou de 1, et l'on verra dans la suite du cours que la convergence est en fait uniforme en ϑ sur tout intervalle $[a,b]$ avec $0 < a < b < 1$.



On dispose de 50 *mesures binaires* $(x_1,\dots,x_{50})$*, obtenues comme la réalisation de variables de Bernoulli indépendantes de paramètre* $\theta^\star = 1/3$*, telles que la moyenne empirique* $\bar{x}_n$ *vaut* 0,4 *pour* $n = 5$*,* 0,3 *pour* $n = 10$ *et* 0,32 *pour* $n = 50$*. On trace les fonctions* $\vartheta \mapsto \bar{x}_n \log\vartheta + (1-\bar{x}_n)\log(1-\vartheta)$ *sur* $]0,1[$ *pour* $n = 5$*,* $n = 10$ *et* $n = 50$*, ainsi que la fonction limite* $M_{\theta^\star} : \vartheta \mapsto \theta^\star\log\vartheta + (1-\theta^\star)\log(1-\vartheta)$*, en trait plein noir. Les traits verticaux sont placés en* $\theta^\star$*,* 0,3*,* 0,32 *et* 0,4 *: le maximum de chaque log-vraisemblance est atteint en* $\bar{x}_n$*, et se rapproche de* $\theta^\star$ *lorsque* n *augmente.*

Vraisemblance d'un modèle censuré

Lorsque les observations sont censurées, la loi d'une observation n'a pas de densité par rapport à la mesure de Lebesgue, mais le modèle reste dominé par une mesure bien choisie, et la vraisemblance s'écrit sans difficulté. Cet exemple poursuit celui du contrôle de qualité du chapitre 1.

Exemple I-2.4.5: Estimateur du maximum de vraisemblance pour des durées de vie censurées

Reprenons l'exemple du modèle exponentiel censuré du chapitre 1 (Section I-1.4) : les observations sont des durées de vie censurées à droite,

$$X_i^* := \min(X_i, \tau)\ , \qquad i = 1,\dots,n\ ,$$

où $\tau > 0$ est un instant de censure connu et $(X_1, \dots, X_n)$ est un n-échantillon du modèle exponentiel $\big(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+), \{\mathcal{E}(\theta)\,,\ \theta \in \Theta = \mathbb{R}_+^*\}\big)$. Les mesures dont on dispose sont modélisées par le modèle induit par $(X_1^*, \dots, X_n^*)$, et non par celui des X_i. On a vu au chapitre 1 que $(X_1^*, \dots, X_n^*)$ est un n-échantillon du modèle $\big([0,\tau], \mathcal{B}([0,\tau]), \{q_\theta^* \cdot \mu\,,\ \theta \in \Theta\}\big)$, dominé par la mesure $\mu := \mathrm{Leb} + \delta_\tau$, somme de la mesure de Lebesgue et de la masse de Dirac en τ, avec la densité

$$q_\theta^*(x) := \theta \mathrm{e}^{-\theta x} \mathbb{1}_{\{x<\tau\}} + \mathrm{e}^{-\theta\tau} \mathbb{1}_{\{x=\tau\}}\ .$$

Notons $N_n^- := \{i \le n\ :\ X_i^* < \tau\}$ l'ensemble des indices des pannes observées avant l'instant de censure et $N_n^+ := \{i \le n\ :\ X_i^* = \tau\}$ celui des observations censurées. La vraisemblance s'écrit, pour $\vartheta > 0$,

$$\begin{aligned}\mathrm{L}_n(\vartheta, X_1^*, \dots, X_n^*) &= \vartheta^{\,\mathrm{card}\, N_n^-} \exp\Big(-\vartheta \sum_{i \in N_n^-} X_i^*\Big)\, \mathrm{e}^{-\vartheta\tau\, \mathrm{card}\, N_n^+} \\ &= \vartheta^{\,\mathrm{card}\, N_n^-} \exp\Big(-\vartheta \sum_{i=1}^n X_i^*\Big)\ ,\end{aligned}$$

puisque $X_i^* = \tau$ pour $i \in N_n^+$. Elle est à comparer avec la vraisemblance du modèle sans censure, $\vartheta^n \exp(-\vartheta \sum_{i=1}^n X_i)$. Si $\mathrm{card}\, N_n^- \ge 1$, la log-vraisemblance $\vartheta \mapsto \frac{\mathrm{card}\, N_n^-}{n} \log\vartheta - \vartheta\, \bar{X}_n^*$ est strictement concave et son unique point critique est le maximum global :

$$\hat{\theta}_n^{\mathrm{MV}} = \frac{\mathrm{card}\, N_n^-}{\sum_{i=1}^n X_i^*}\ ,$$

le nombre de pannes observées divisé par la durée totale de fonctionnement observée, censurée ou non. Si $\mathrm{card}\, N_n^- = 0$, c'est-à-dire si toutes les observations sont censurées, événement de probabilité $\mathrm{e}^{-n\theta\tau} > 0$, la vraisemblance $\vartheta \mapsto \mathrm{e}^{-n\vartheta\tau}$ est strictement décroissante sur $\mathbb{R}_+^*$ et l'estimateur du maximum de vraisemblance n'est pas défini.

Estimateur du maximum de vraisemblance du modèle de Cauchy

Dans la plupart des modèles, l'équation de vraisemblance n'a pas de solution explicite, et l'estimateur du maximum de vraisemblance s'obtient par une procédure numérique. Le modèle de Cauchy en donne un exemple simple à écrire.

Exemple I-2.4.6: Équation de vraisemblance du modèle de Cauchy et méthode de gradient

Reprenons le modèle de Cauchy de l'Exemple I-2.4.1. La vraisemblance et la log-vraisemblance normalisée s'écrivent

$$\mathrm{L}_n(\vartheta) = \frac{1}{\pi^n} \prod_{i=1}^n \frac{1}{1 + (X_i - \vartheta)^2}\ , \qquad \ell_n(\vartheta) = -\log\pi - \frac{1}{n}\sum_{i=1}^n \log\big(1 + (X_i - \vartheta)^2\big)\ .$$

La fonction ℓ_n est continue sur $\mathbb{R}$ et tend vers $-\infty$ lorsque $|\vartheta| \to +\infty$: elle atteint son maximum, de sorte qu'un estimateur du maximum de vraisemblance existe. L'équation de vraisemblance,

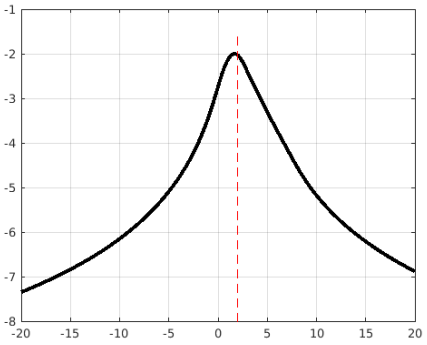
$$\ell_n'(\vartheta) = \frac{2}{n} \sum_{i=1}^n \frac{X_i - \vartheta}{1 + (X_i - \vartheta)^2} = 0\ ,$$

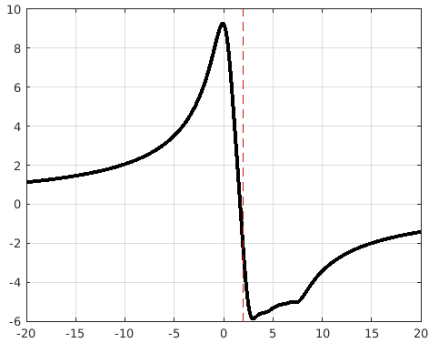
n'admet pas de solution explicite et peut admettre plusieurs solutions, qui ne sont pas toutes des maxima. Pour maximiser la vraisemblance, on a recours à une procédure numérique, par exemple une méthode de gradient : en notant $\vartheta^{(k)}$ la valeur du paramètre à la k-ième itération,

on pose

$$\vartheta^{(k+1)} = \vartheta^{(k)} + \gamma\, \ell_n'\left(\vartheta^{(k)}\right),$$

où $\gamma > 0$ est un pas. La suite $\ell_n(\vartheta^{(k)})$ croît pour un pas assez petit, mais la limite peut n'être qu'un maximum local ; le choix du point de départ, par exemple la médiane empirique, qui est un Z-estimateur consistant de θ, est déterminant.





On dispose de $n = 25$ mesures réelles $x_1, \dots, x_n$, obtenues comme la réalisation de n variables aléatoires i.i.d. de loi de Cauchy de paramètre $\theta^\star = 2$. À gauche, une réalisation de la log-vraisemblance normalisée $\vartheta \mapsto -\log \pi - \frac{1}{n}\sum_{i=1}^n \log(1 + (x_i - \vartheta)^2)$; à droite, une réalisation de la fonction $\vartheta \mapsto \sum_{i=1}^n (x_i - \vartheta)/(1 + (x_i - \vartheta)^2)$, proportionnelle à sa dérivée, dont les zéros sont les racines de l'équation de vraisemblance. Le trait vertical en pointillé est la droite d'équation $\vartheta = \theta^\star$: le maximum de vraisemblance n'est pas $\theta^\star$.

Régression linéaire gaussienne à $p+1$ coefficients

La vraisemblance a été définie pour un n-échantillon, mais sa définition s'étend à tout modèle dominé : c'est la densité de l'observation, évaluée aux données et vue comme fonction du paramètre. Le modèle de régression linéaire gaussienne, dont les observations sont indépendantes mais pas identiquement distribuées, en est l'exemple le plus important.

Exemple I-2.4.7: Régression linéaire gaussienne et moindres carrés

Soient $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$ des variables explicatives connues et $\varphi_0, \dots, \varphi_p : \mathbb{R}^d \to \mathbb{R}$ des fonctions connues. On considère le modèle

$$\left(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \left\{p_\theta \cdot \mathrm{Leb}^{\otimes n},\ \theta \in \Theta\right\}\right),$$

où

$$p_\theta(y_1, \dots, y_n) := \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big(-\frac{1}{2\sigma^2} \sum_{k=1}^n \Big\{ y_k - \sum_{\ell=0}^p \beta_\ell \varphi_\ell(\mathbf{x}_k) \Big\}^2 \Big),$$

de paramètre $\theta = (\beta_0, \dots, \beta_p, \sigma^2) \in \Theta := \mathbb{R}^{p+1} \times \mathbb{R}_+^*$: sous $\mathbb{P}_\theta$, les observations $Y_1, \dots, Y_n$ sont indépendantes, $Y_k \sim \mathcal{N}\big(\sum_{\ell=0}^p \beta_\ell \varphi_\ell(\mathbf{x}_k), \sigma^2\big)$. Notons $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$, $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)^\top$ et Φ la matrice de régression, de taille $n \times (p+1)$, de coefficients $\Phi_{k,\ell} = \varphi_\ell(\mathbf{x}_k)$. La log-vraisemblance normalisée s'écrit, pour $(\boldsymbol{b}, v) \in \mathbb{R}^{p+1} \times \mathbb{R}_+^*$,

$$\ell_n(\boldsymbol{b}, v) = -\frac{1}{2}\log(2\pi v) - \frac{1}{2nv} \|\mathbf{Y} - \Phi \boldsymbol{b}\|^2 .$$

À v fixé, maximiser ℓ_n en $\boldsymbol{b}$ revient à minimiser

$$J_n(\boldsymbol{b}) := \|\mathbf{Y} - \Phi\boldsymbol{b}\|^2 = \sum_{k=1}^n \Big\{ Y_k - \sum_{\ell=0}^p b_\ell \varphi_\ell(\mathbf{x}_k) \Big\}^2 ,$$

quelle que soit la valeur de v : pour le paramètre de régression, l'estimateur du maximum de vraisemblance coïncide avec l'*estimateur des moindres carrés*. La fonction J_n est convexe et différentiable, de gradient $\nabla J_n(\boldsymbol{b}) = -2\Phi^\top(\mathbf{Y} - \Phi\boldsymbol{b})$; ses minima sont les solutions des *équations normales*

$$\Phi^\top\Phi\,\boldsymbol{b} = \Phi^\top\mathbf{Y}\ .$$

Si la matrice $\Phi^\top\Phi$ est inversible, ce qui équivaut à ce que Φ soit de rang $p+1$ et impose $n \geq p+1$, la fonction J_n est strictement convexe et l'estimateur des moindres carrés est unique,

$$\hat{\boldsymbol{\beta}}_n := (\Phi^\top\Phi)^{-1}\Phi^\top\mathbf{Y}\ .$$

Le vecteur $\Phi\hat{\boldsymbol{\beta}}_n$ est la projection orthogonale de $\mathbf{Y}$ sur l'image de Φ. En reportant, $v \mapsto \ell_n(\hat{\boldsymbol{\beta}}_n, v)$ atteint son maximum, par le même argument que dans l'Exemple I-2.1.10, en

$$\hat\sigma_n^2 := \frac{1}{n}\sum_{k=1}^{n}\Big\{Y_k - \sum_{\ell=0}^{p}\hat\beta_{n,\ell}\varphi_\ell(\mathbf{x}_k)\Big\}^2 = \frac{1}{n}\,\|\mathbf{Y} - \Phi\hat{\boldsymbol{\beta}}_n\|^2\ ,$$

pourvu que cette quantité soit strictement positive, ce qui est $\mathbb{P}_\theta$-presque sûr dès que $n > p+1$, puisque $\mathbf{Y} - \Phi\hat{\boldsymbol{\beta}}_n$ est la projection orthogonale de $\mathbf{Y}$ sur l'orthogonal de l'image de Φ, sous-espace de dimension $n-p-1 \geq 1$. Le théorème de Cochran (Théorème I-1.1.1) donne alors les lois : $\hat{\boldsymbol{\beta}}_n \sim \mathcal{N}\big(\boldsymbol{\beta}, \sigma^2(\Phi^\top\Phi)^{-1}\big)$, $n\hat\sigma_n^2/\sigma^2 \sim \chi^2(n-p-1)$, et ces deux estimateurs sont indépendants. Le cas $p=1$, $\varphi_0 = 1$, $\varphi_1(x) = x$ est celui de la tendance linéaire du chapitre 1 (Section I-1.4) et de l'Exercice 1 (PC2).

La divergence de Kullback-Leibler : énoncé complet

Les fondamentaux n'ont utilisé que la positivité de la divergence et son cas d'égalité. On établit ici que l'intégrale qui la définit a toujours un sens, que sa finitude force l'absolue continuité, et que sa valeur ne dépend pas de la mesure dominante choisie.

L'intégrale est bien définie. Reprenons les notations de la Définition I-2.1.7 : $\mathbb{P}_0 = f_0\cdot\mu$ et $\mathbb{P}_1 = f_1\cdot\mu$. La partie négative de l'intégrande, $\big(f_0\log(f_0/f_1)\big)_- := \max\big(-f_0\log(f_0/f_1), 0\big)$, est μ-intégrable sur $\{f_0 > 0\}$. En effet, elle est nulle sauf lorsque $0 < f_0 < f_1$, et, en utilisant $\log y \leq y$ pour $y \geq 1$,

$$\int_{\{f_0>0\}}\big(f_0\log(f_0/f_1)\big)_-\,\mathrm{d}\mu = \int f_0\log\frac{f_1}{f_0}\,\mathbb{1}_{\{0<f_0<f_1\}}\,\mathrm{d}\mu \leq \int f_0\,\frac{f_1}{f_0}\,\mathbb{1}_{\{0<f_0<f_1\}}\,\mathrm{d}\mu \leq \int f_1\,\mathrm{d}\mu = 1\ .$$

L'intégrale $\mathrm{KL}(\mathbb{P}_0,\mathbb{P}_1)$ est donc toujours définie, et $\mathrm{KL}(\mathbb{P}_0,\mathbb{P}_1) > -\infty$; le Théorème I-2.1.1 montre qu'elle est en fait positive.

Théorème I-2.4.1: Propriétés de la divergence de Kullback-Leibler

Soient $(\mathsf{X},\mathcal{X})$ un espace mesurable, $\mathbb{P}_0$ et $\mathbb{P}_1$ deux probabilités sur $(\mathsf{X},\mathcal{X})$, et μ une mesure σ-finie dominant $\mathbb{P}_0$ et $\mathbb{P}_1$, de densités respectives f_0 et f_1 (on peut toujours prendre $\mu = \mathbb{P}_0+\mathbb{P}_1$).

(i) Si $\mathrm{KL}(\mathbb{P}_0,\mathbb{P}_1) < +\infty$, alors $\mathbb{P}_0 \ll \mathbb{P}_1$.

(ii) $\mathrm{KL}(\mathbb{P}_0,\mathbb{P}_1) \in [0,+\infty]$, et $\mathrm{KL}(\mathbb{P}_0,\mathbb{P}_1) = 0$ si et seulement si $\mathbb{P}_0 = \mathbb{P}_1$.

(iii) $\mathrm{KL}(\mathbb{P}_0,\mathbb{P}_1)$ ne dépend pas du choix de la mesure dominante μ.

Démonstration

(i) Supposons qu'il existe $A \in \mathcal{X}$ tel que $\mathbb{P}_1(A) = 0$ et $\mathbb{P}_0(A) > 0$. Alors $f_1 = 0$ μ-presque partout sur A, tandis que $\mu(A\cap\{f_0>0\}) > 0$, puisque $\mathbb{P}_0(A) = \int_A f_0\,\mathrm{d}\mu > 0$. Sur $A\cap\{f_0>0\}$,

l'intégrande $f_0 \log(f_0/f_1)$ vaut $+\infty$ μ-presque partout, par la convention $\log(1/0) = +\infty$: sa partie positive est d'intégrale infinie, et, la partie négative étant intégrable, $\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_1) = +\infty$. Par contraposée, $\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_1) < +\infty$ entraîne que $\mathbb{P}_1(A) = 0$ implique $\mathbb{P}_0(A) = 0$, c'est-à-dire $\mathbb{P}_0 \ll \mathbb{P}_1$.
(ii) C'est le Théorème I-2.1.1.
(iii) Si $\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_1) = +\infty$ pour une mesure dominante, alors, d'après (i), $\mathbb{P}_0$ n'est pas absolument continue par rapport à $\mathbb{P}_1$, propriété qui ne fait pas intervenir μ ; pour toute autre mesure dominante, la divergence est donc encore infinie, par (i) à nouveau. Supposons maintenant $\mathbb{P}_0 \ll \mathbb{P}_1$. Par le théorème de Radon-Nikodym (Section I-1.2), $\mathbb{P}_0$ admet une densité $\tilde{f}_0$ par rapport à $\mathbb{P}_1$, unique à un ensemble $\mathbb{P}_1$-négligeable près. Comme $\mathbb{P}_1 = f_1 \cdot \mu$ et $\mathbb{P}_0 = \tilde{f}_0 \cdot \mathbb{P}_1$, on a $\mathbb{P}_0 = \tilde{f}_0 f_1 \cdot \mu$, et l'unicité de la densité par rapport à μ donne $f_0 = \tilde{f}_0 f_1$ μ-presque partout. Sur l'ensemble $\{f_0 > 0\} = \{\tilde{f}_0 f_1 > 0\}$ (à un ensemble μ-négligeable près), on a $f_1 > 0$ et $f_0/f_1 = \tilde{f}_0$, donc

$$\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_1) = \int_{\{\tilde{f}_0 f_1 > 0\}} \tilde{f}_0 f_1 \log \tilde{f}_0 \, \mathrm{d}\mu = \int_{\{\tilde{f}_0 > 0\}} \tilde{f}_0 \log \tilde{f}_0 \, f_1 \, \mathrm{d}\mu = \int_{\{\tilde{f}_0 > 0\}} \tilde{f}_0 \log \tilde{f}_0 \, \mathbb{P}_1(\mathrm{d}x) \, ,$$

la deuxième égalité venant de ce que l'ensemble $\{f_1 = 0\}$ ne contribue pas à l'intégrale contre $f_1 \, \mathrm{d}\mu$. Le membre de droite ne fait plus intervenir μ : avec la convention $0 \log 0 = 0$,

$$\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_1) = \int_{\mathsf{X}} \tilde{f}_0 \log \tilde{f}_0 \, \mathrm{d}\mathbb{P}_1 \, , \qquad \tilde{f}_0 = \frac{\mathrm{d}\mathbb{P}_0}{\mathrm{d}\mathbb{P}_1} \, ,$$

ce qui montre que la divergence ne dépend pas de la mesure dominante. □

Remarque. Le point (i) éclaire le cas « $\mathbb{P}(Y = 0) > 0$ » de la démonstration du Théorème I-2.1.1 : la divergence vaut $+\infty$ dès que la loi $\mathbb{P}_1$ attribue une probabilité nulle à un événement de $\mathbb{P}_0$-probabilité strictement positive. Dans un modèle statistique, $\mathrm{KL}(\mathbb{Q}_{\theta^\star}, \mathbb{Q}_\vartheta) = +\infty$ signifie que certaines observations possibles sous $\theta^\star$ sont impossibles sous ϑ : la vraisemblance en ϑ est alors nulle avec une probabilité qui tend vers 1 lorsque n croît, et une telle valeur ϑ est éliminée par le maximum de vraisemblance. C'est le cas du modèle uniforme $\mathcal{U}([0, \theta])$ pour $\vartheta < \theta^\star$, où l'hypothèse d'intégrabilité de la Proposition I-2.1.3 n'est pas satisfaite.