Chapitre III

Méthodes générales d'estimation

1 Quelques considérations préliminaires

Nous allons, dans ce chapitre, supposer que nous travaillons sur le *modèle d'échantillonnage*, c'est à dire un modèle stochastique dans lequel on observe une suite de v.a. i.i.d. $X_1, \ldots, X_n, \ldots$ à valeurs dans $(E, \mathcal{E})$ et de loi inconnue P_X. Nous nous intéressons ici à un paramètre inconnu μ fonction de P, par exemple $\mu = \mathbb{E}[X_1]$ et à des suites d'estimateurs $\widehat{\mu}_n(X_1, \ldots, X_n)$ de μ. Pour décrire les propriétés asymptotiques d'une telle suite, on introduit les notions suivantes.

Définition 9 *Soit $X_1, \ldots, X_n, \ldots$ une suite de variables aléatoires i.i.d. de loi P_X et μ un paramètre inconnu fonction de P_X. Soit $\widehat{\mu}_n(X_1, \ldots, X_n)$ un estimateur de μ fonction de $X_1, \ldots, X_n$. La suite d'estimateurs $(\widehat{\mu}_n)_{n\geq 1}$ est dite* convergente (consistante) *si*

$$\widehat{\mu}_n \underset{n\to+\infty}{\xrightarrow{P}} \mu.$$

Si $\mu \in \mathbb{R}$, elle est dite asymptotiquement normale *s'il existe une quantité positive σ^2, telle que*

$$\sqrt{n}\,(\widehat{\mu}_n - \mu) \rightsquigarrow \mathcal{N}\left(0, \sigma^2\right).$$

La normalité asymptotique de la suite $(\widehat{\mu}_n)_{n\geq 1}$ entraîne sa convergence.

Remarque importante : Lorsque l'on considère des suites d'estimateurs asymptotiquement normaux, ils sont d'autant meilleurs (asymptotiquement) que leur variance asymptotique est plus petite. En effet, si $\sqrt{n}\,(\widehat{\mu}_n - \mu) \rightsquigarrow \mathcal{N}\left(0, \sigma^2\right)$, si Z est $\mathcal{N}(0,1)$ et $\mathbb{P}[|Z| \leq t_\alpha] = 1 - \alpha$, alors

$$\mathbb{P}\left[|\widehat{\mu}_n - \mu| \leq n^{-1/2} t_\alpha \sigma\right] \underset{n\to+\infty}{\longrightarrow} 1 - \alpha,$$

de sorte que la variance asymptotique σ^2 donne la précision de l'approximation de μ par $\widehat{\mu}_n$ lorsque n est suffisamment grand. Si σ est indépendant de μ et connu, on en déduit immédiatement des intervalles de confiance asymptotiques pour μ :

$$\mathbb{P}\left[\widehat{\mu}_n - n^{-1/2} t_\alpha \sigma \leq \mu \leq \widehat{\mu}_n + n^{-1/2} t_\alpha \sigma\right] \underset{n\to+\infty}{\longrightarrow} 1 - \alpha.$$

Si σ est inconnu ou dépend de μ, il faut l'estimer. Si l'on dispose d'une suite d'estimateurs convergents $\widehat{\sigma}_n^2$ de σ^2, le Théorème de Slutsky entraîne que $\sqrt{n}\widehat{\sigma}_n^{-1}\,(\widehat{\mu}_n - \mu) \rightsquigarrow \mathcal{N}(0,1)$, de sorte que

$$\mathbb{P}\left[\widehat{\mu}_n - n^{-1/2} t_\alpha \widehat{\sigma}_n \leq \mu \leq \widehat{\mu}_n + n^{-1/2} t_\alpha \widehat{\sigma}_n\right] \underset{n\to+\infty}{\longrightarrow} 1 - \alpha. \qquad \text{(III.1.1)}$$

On obtient évidemment des résultats analogues pour les intervalles de confiance unilatères. Lorsque $\sigma = \sigma(\mu)$ est une fonction continue et connue de μ, on peut l'estimer par $\sigma(\widehat{\mu}_n)$. Attention, les intervalles de confiance (III.1.1) peuvent être vraiment très asymptotiques !

2 Loi empirique et estimateurs empiriques

Nous ferons ici sur P_X des hypothèses minimales, par exemple que P_X a des moments jusqu'à l'ordre 2, ou 4, ..., ou que sa fonction de répartition est continue, dérivable, etc.

2.1 Loi empirique, moments et espérances empiriques

2.1.1 Premier exemple, la moyenne empirique

L'exemple le plus classique d'estimateur *empirique* est la *moyenne empirique* des observations, $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$. Si l'on suppose que P_X a un moment d'ordre 2, c'est à dire que les X_i ont une variance σ^2, on obtient immédiatement que

$$\overline{X}_n \underset{n\to+\infty}{\xrightarrow{P}} \mathbb{E}[X_1] \qquad \text{et} \qquad \sqrt{n}\left(\overline{X}_n - \mathbb{E}[X_1]\right) \rightsquigarrow \mathcal{N}\left(0,\sigma^2\right).$$

Donc la suite des estimateurs $\overline{X}_n$ de l'espérance des X_i est convergente et asymptotiquement normale.

2.1.2 Loi empirique

Si l'on parle de moyenne empirique pour $\overline{X}_n$, c'est parce que $\overline{X}_n$ est effectivement la moyenne (c'est-à-dire l'espérance) d'une certaine loi de probabilité, dite *loi empirique*, associée aux observations.

Définition 10 *Étant donné des points $x_1,\dots,x_n$ dans un ensemble $(E,\mathcal{E})$, on appelle* loi empirique *des x_i la probabilité $\nu_n = n^{-1}\sum_{i=1}^n \delta_{x_i}$ sur E (où δ_x désigne la mesure de Dirac au point x), c'est à dire la probabilité définie par*

$$\nu_n(A) = n^{-1}\sum_{i=1}^n \mathbb{1}_A(x_i) \quad \textit{pour tout } A\in\mathcal{E}. \tag{III.2.1}$$

Les quantités associées à la loi empirique sont également dites empiriques *et l'on parlera ainsi de moyenne empirique des x_i, de variance empirique des x_i, de fonction de répartition empirique des x_i, etc.*

Lorsque les x_i sont des réalisations de variables aléatoires X_i i.i.d., c'est à dire que $x_i = X_i(\omega)$ pour tout i, la loi empirique de l'échantillon $X_1,\dots,X_n$ *est une fonction de $\omega\in\Omega$ qui s'écrit*

$$\nu_n(\omega) = n^{-1}\sum_{i=1}^n \delta_{X_i(\omega)}.$$

Dans ce cas, ν_n est une probabilité aléatoire*, c'est à dire une application (mesurable) de l'ensemble Ω dans l'ensemble des probabilités à support fini sur $(E,\mathcal{E})$, qui est une fonction des X_i, de même que toutes les quantités empiriques associées, en particulier la* fréquence empirique *de l'évènement A : $\nu_n(A) = n^{-1}\sum_{i=1}^n \mathbb{1}_A(X_i)$.*

On notera que si les observations sont toutes distinctes, ν_n est simplement la loi uniforme sur l'ensemble $\{X_1,\dots,X_n\}$. Comme les X_i sont i.i.d. de loi P_X, $n\nu_n(A)\sim\mathcal{B}(n,P_X(A))$.

2.1.3 Espérances empiriques

Si la variable aléatoire Y a la loi ν_n définie par (III.2.1), pour toute fonction h, $\mathbb{E}_{\nu_n}[h(Y)] = n^{-1}\sum_{i=1}^n h(x_i)$. Dans le cas de la loi empirique d'un échantillon $X_1, \ldots, X_n$, ceci s'écrit $\mathbb{E}_{\nu_n}[h(Y)] = n^{-1}\sum_{i=1}^n h(X_i)$ et comme ν_n est aléatoire, les espérances par rapport à ν_n le sont aussi (ce sont des fonctions de ω). En particulier, la moyenne empirique $\overline{X}_n$ des X_i définie plus haut est tout simplement l'espérance (moyenne) $\mathbb{E}_{\nu_n}[Y]$ de la loi empirique ν_n. La loi des grands nombres et le T.L.C. impliquent que si les observations X_i sont i.i.d. de loi P_X,

i) si h est P-intégrable, alors $\mathbb{E}_{\nu_n}[h(Y)] \xrightarrow[n\to+\infty]{P} \mathbb{E}_{P_X}[h(Y)]$;

ii) si $\mathbb{E}_{P_X}\left[h^2(Y)\right] < +\infty$, $\sqrt{n}\left(\mathbb{E}_{\nu_n}[h(Y)] - \mathbb{E}_{P_X}[h(Y)]\right) \rightsquigarrow \mathcal{N}\left(0, \mathrm{Var}_{P_X}(h(Y))\right)$.

2.1.4 Variance empirique

On a souvent besoin d'estimer la variance σ^2 de la loi de X, alors même que la moyenne est inconnue, de sorte qu'il ne suffit pas d'estimer le second moment $\mathbb{E}\left[X^2\right]$ pour y parvenir. On utilise alors la variance empirique $\mathrm{Var}_{\nu_n}(Y)$, c'est à dire la variance $\widehat{\sigma}_n^2$ de la loi empirique qui s'écrit

$$\widehat{\sigma}_n^2 = n^{-1}\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2 = \mathbb{E}_{\nu_n}\left[Y^2\right] - \left(\mathbb{E}_{\nu_n}[Y]\right)^2 = n^{-1}\sum_{i=1}^n X_i^2 - \overline{X}_n^2. \qquad \text{(III.2.2)}$$

L'estimateur obtenu est biaisé d'après l'inégalité de Jensen $\mathbb{E}\left[\overline{X}_n^2\right] > \left(\mathbb{E}\left[\overline{X}_n\right]\right)^2$ (sauf si X_i est presque sûrement constante). On a donc $\mathbb{E}\left[\widehat{\sigma}_n^2\right] < \sigma^2$. Plus précisément, en développant le carré, on trouve que $\mathbb{E}\left[\widehat{\sigma}_n^2\right] = n^{-1}(n-1)\sigma^2$. Il s'ensuit que l'estimateur modifié $\widehat{s}_n^2 = (n-1)^{-1}\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2$ est un estimateur sans biais de σ^2. Asymptotiquement, les deux estimateurs, $\widehat{\sigma}_n^2$ et $\widehat{s}_n^2$ sont équivalents.

Pour calculer la loi limite de l'un ou de l'autre de ces estimateurs, on ne peut utiliser le T.L.C. car les composantes $\left(X_i - \overline{X}_n\right)^2$ de la somme ne sont pas indépendantes. Il faut avoir recours à une astuce. On remarque que le T.L.C. s'applique aux variables $t_n^2 = n^{-1}\sum_{i=1}^n \left(X_i - \mathbb{E}[X_i]\right)^2$ et donne $\sqrt{n}\left(t_n^2 - \sigma^2\right) \rightsquigarrow \mathcal{N}\left(0, v^2\right)$ avec $v^2 = \mathrm{Var}\left((X_i - \mathbb{E}[X_i])^2\right)$. On montre ensuite par un calcul direct de la différence $\widehat{\sigma}_n^2 - t_n^2$ que $\sqrt{n}\left(\widehat{\sigma}_n^2 - t_n^2\right) \to 0$. Le théorème de Slutsky permet alors de conclure que

$$\sqrt{n}\left(\widehat{\sigma}_n^2 - \sigma^2\right) = \sqrt{n}\left(\widehat{\sigma}_n^2 - t_n^2\right) + \sqrt{n}\left(t_n^2 - \sigma^2\right) \rightsquigarrow \mathcal{N}\left(0, v^2\right).$$

Pour traiter le cas de $\widehat{s}_n^2$ il suffit de remarquer à nouveau que $\sqrt{n}\left(\widehat{\sigma}_n^2 - \widehat{s}_n^2\right) \to 0$.

2.2 Fonction de répartition et quantiles empiriques

2.2.1 Fonction de répartition empirique

On rappelle que si X est une v.a.r. de loi P_X, sa fonction de répartition F_X est donnée par

$$F_X(t) = \mathbb{P}[X \le t] = P_X\left[\,]-\infty; t]\right]$$

et que F_X est une fonction croissante (pas nécessairement strictement), de 0 (quand $t \to -\infty$) à 1 (quand $t \to +\infty$), continue à droite et ayant des limites à gauche. Elle a au plus un nombre dénombrable de sauts d'amplitude nécessairement > 0. En particulier la

loi empirique $\nu_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$ a une fonction de répartition que l'on note F_n et que l'on appelle *fonction de répartition empirique des* X_i :

$$F_n(t) = \nu_n\left(]-\infty;t]\right) = n^{-1}\sum_{i=1}^n \mathbb{1}_{]-\infty;t]}(X_i). \tag{III.2.3}$$

Pour la calculer, le plus simple est d'ordonner en croissant l'ensemble des X_i (avec possibilité d'ex-aequos).

Définition 11 *Étant donné un n-échantillon* $X_1,\ldots,X_n$*, on appelle* statistique d'ordre *associée, l'ensemble des valeurs prises par les* X_i *(en conservant les valeurs multiples, s'il y en a, avec leur multiplicité) ordonnées en croissant de la plus petite à la plus grande. On note l'échantillon ordonné* $X_{(1)},\ldots,X_{(n)}$ *avec* $X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$.

On peut alors vérifier, en notant $X_{(0)} = -\infty$ et $X_{(n+1)} = +\infty$, que

$$F_n(t) = i/n \quad \text{si } t \in [X_{(i)}; X_{(i+1)}[, \quad 0 \le i \le n.$$

Une conséquence immédiate de (III.2.3) est que $nF_n(t) \sim \mathcal{B}(n, F_X(t))$, ce qui entraîne que $\lim_{n\to+\infty} F_n(t) = F_X(t)$ quel que soit t. Il n'est pas très difficile de vérifier que, puisque F_X et F_n sont croissantes, cette convergence est en fait uniforme en t, ce qui donne

Théorème 4 (Glivenko-Cantelli) *Si l'on dispose de n v.a.r. i.i.d.* $X_1,\ldots,X_n$ *de fonction de répartion* F_X *et si* F_n *désigne la fonction de répartition empirique,*

$$\sup_{t\in\mathbb{R}} \left|F_n(t) - F_X(t)\right| \underset{n\to+\infty}{\overset{\mathbb{P}}{\longrightarrow}} 0.$$

De plus, pour t fixé, le T.L.C. nous dit que

$$\sqrt{n}\left(F_n(t) - F_X(t)\right) \rightsquigarrow \mathcal{N}\left(0; F_X(t)[1-F_X(t)]\right),$$

et l'inégalité de Hoeffding pour les binomiales que

$$\mathbb{P}\left[\left|F_n(t) - F_X(t)\right| \ge x\right] \le 2\exp\left[-2nx^2\right],$$

ce qui suggère que la différence $|F_n(t) - F_X(t)|$ est d'ordre de grandeur $n^{-1/2}$, même pour n petit. On peut en fait démontrer (mais c'est très difficile) un résultat beaucoup plus fin et non-asymptotique très utile en pratique :

Théorème 5 (Massart) *Supposons que l'on dispose de n v.a.r. i.i.d.* $X_1,\ldots,X_n$ *de fonction de répartion* F_X *et notons* F_n *la fonction de répartition empirique. Alors, pour tout* $n \ge 1$ *et tout* $x > 0$*,*

$$\mathbb{P}\left[\sup_{t\in\mathbb{R}} \left|F_n(t) - F_X(t)\right| \ge x\right] \le 2\exp\left[-2nx^2\right]. \tag{III.2.4}$$

2.2.2 Quantiles et quantiles empiriques

Si l'on considère un n-échantillon de loi $\mathcal{N}(\mu;1)$ où μ désigne un paramètre inconnu, on peut caractériser μ par le fait qu'il est l'espérance de la loi et l'estimer par la moyenne empirique. Si l'on remplace ici la densité gaussienne standard par la densité de Cauchy, $f(x) = \left[\pi\left(1+x^2\right)\right]^{-1}$, qui est elle aussi symétrique, on ne peut plus caractériser son centre de symétrie comme étant l'espérance de la loi car cette espérance n'existe pas. Donc pour estimer le paramètre de symétrie μ dans le modèle $\{P_\mu, \mu \in \mathbb{R}\}$ avec P_μ de densité $\left[\pi\left(1+(x-\mu)^2\right)\right]^{-1}$, on ne peut pas utiliser la moyenne empirique. Il convient donc de caractériser μ autrement. On peut, par exemple, dire que μ est la *médiane* de la loi, c'est-à-dire le point $x_{1/2}$ tel que $F_\mu(x_{1/2}) = 1/2$, où F_μ est la fonction de répartition de P_μ. Plus généralement, on peut considérer le point x_p tel que $F_\mu(x_p) = p \in]0;1[$. La définition de x_p est évidemment très simple lorsque la fonction F est continue et strictement croissante parce qu'elle est alors inversible : quel que soit $p \in]0;1[$, il existe x_p unique tel que $F(x_p) = p$. La situation devient plus complexe lorsque F a des discontinuités ou n'est pas strictement croissante (**dessins**). Pour définir des analogues de x_p dans cette situation, on va utiliser la notion d'*inverse généralisée* de F.

Définition 12 *Étant donné une fonction de répartition F sur $\mathbb{R}$, on appelle* inverse généralisée *de F et l'on note F^{-1} la fonction définie sur $[0;1]$ par*

$$F^{-1}(x) = \inf\{t \in \mathbb{R} \,|\, F(t) \geq x\}, \quad x \in [0;1],$$

avec les conventions habituelles : $\inf \mathbb{R} = -\infty$ *et* $\inf \varnothing = +\infty$.

Les propriétés suivantes de F^{-1} sont immédiates : $F^{-1}(0) = -\infty$, F^{-1} est croissante (au sens large) et lorsque F est continue et strictement croissante, F^{-1} est la fonction réciproque de F sur $]0;1[$ au sens usuel. De plus, la continuité à droite de F implique que $F \circ F^{-1}(x) \geq x$. Par définition,

$$F(t) \geq x \quad \text{si} \quad t \geq F^{-1}(x) \qquad \text{et} \qquad F(t) < x \quad \text{si} \quad t < F^{-1}(x),$$

de sorte que

$$F(t) \geq x \quad \text{équivaut à} \quad t \geq F^{-1}(x). \tag{III.2.5}$$

Attention : Si F n'est pas bijective, il existe $a < b$ tels que $F(b) = F(a)$, donc $F^{-1} \circ F(b) \leq a < b$; si F n'est pas continue, il existe a tel que $F(a_-) < F(a)$, donc $F \circ F^{-1}(t) = F(a) > t$ pour tout $t \in]F(a_-);F(a)[$. Il y a donc équivalence entre le fait que $F \circ F^{-1} = F^{-1} \circ F = Id$ et le fait que F soit inversible au sens usuel.

On peut alors définir les quantiles de n'importe quelle fonction de répartition de la manière suivante :

Définition 13 *Étant donné un nombre $p \in [0;1]$ et une fonction de répartition F sur $\mathbb{R}$, le* p-quantile *$x_p(F)$ est donné par $x_p(F) = F^{-1}(p)$. Les quantiles sont aussi appelés* pourcentiles. *La* médiane *correspond à $p = 1/2$ et les quartiles à $p = 1/4$ et $p = 3/4$.*

On a évidemment toujours $x_0(F) = -\infty$. On peut alors définir les *quantiles empiriques* $x_p(n)$ comme les quantiles de la fonction de répartition empirique, soit

$$x_p(n) := x_p(F_n) = \inf\{t \in \mathbb{R} \,|\, F_n(t) \geq p\} \quad \text{pour } 0 \leq p \leq 1.$$

On vérifie que

$$x_p(n) = X_{(i)} \quad \text{si} \quad \frac{i-1}{n} < p \leq \frac{i}{n} \qquad \text{pour } 1 \leq i \leq n,$$

ce qui permet d'évaluer $x_p(n)$ pour tous les $p \in]0;1]$. De manière équivalente,

$$x_p(n) = X_{(i)} \qquad \text{pour} \qquad np \leq i < np+1. \tag{III.2.6}$$

Dans la mesure où la fonction de répartition empirique converge vers la vraie fonction de répartition, il est facile de prédire que, si la fonction de répartition se comporte bien au voisinage d'un quantile, le p-quantile empirique va converger vers le p-quantile correspondant. Un résultat dans cette direction est le suivant.

Théorème 6 *Soit $X_1, \ldots, X_n$ un n-échantillon de f.r. inconnue F et p un réel dans $]0;1[$. On suppose qu'il existe une solution x_p à l'équation $F(x) = p$ et que F est continue et dérivable au point x_p de dérivée $f(x_p) > 0$. Alors le quantile empirique $x_p(n)$ converge en probabilité vers x_p quand $n \to +\infty$ et*

$$\sqrt{n}\,[x_p(n) - x_p] \rightsquigarrow \mathcal{N}\left(0, \frac{p(1-p)}{f^2(x_p)}\right) \quad \textit{quand } n \to +\infty.$$

Attention : Ce résultat n'est pas valable si F a une tangente verticale au point x_p (ce qui correspond à $f(x_p) = +\infty$) auquel cas la convergence de $x_p(n)$ vers x_p est plus rapide. Si $f(x_p) = 0$ mais F est croissante au voisinage de x_p, la convergence est plus lente. S'il existe $a < b$ tels que $F(b) = F(a) = p$, alors $x_n(p)$ ne converge même pas vers x_p.

3 Applications de la méthode delta

Supposons que l'on sache estimer correctement une certaine fonction strictement monotone $\phi(\theta)$ du paramètre, par exemple un moment ou un quantile de P_θ, par la méthode empirique vue précédemment. Si l'on dispose d'une suite d'estimateurs $\widehat{\phi}_n$ convergents et asymptotiquement normaux de $\phi(\theta)$, et si ϕ^{-1} est une fonction dérivable, la méthode delta consiste à estimer θ par $\widehat{\theta}_n = \phi^{-1}(\widehat{\phi}_n)$ et le Théorème 1 montre que la suite $(\widehat{\theta}_n)_{n\geq 1}$ est asymptotiquement normale. On obtient en particulier le corollaire suivant.

Corollaire 1 *Soit une suite de variables aléatoires $(X_n)_{n\geq 1}$ i.i.d. de loi P_θ avec $\theta \in \Theta$, où Θ est un intervalle ouvert de $\mathbb{R}$ et soit ϕ une bijection de Θ dans $\phi(\Theta)$ telle que ϕ soit continuement dérivable, de dérivée $\phi' \neq 0$. Si, de plus, $\sqrt{n}\left[\widehat{\phi}_n - \phi(\theta)\right] \rightsquigarrow \mathcal{N}\left(0, \sigma^2(\theta)\right)$, alors*

$$\sqrt{n}\left[\phi^{-1}\left(\widehat{\phi}_n\right) - \theta\right] \rightsquigarrow \mathcal{N}\left(0, [\sigma(\theta)/\phi'(\theta)]^2\right).$$

Démonstration : Le Théorème 1 appliqué avec $g = \phi^{-1}$ montre que

$$\sqrt{n}\left[\phi^{-1}\left(\widehat{\phi}_n(X_1, \ldots, X_n)\right) - \theta\right] \rightsquigarrow \left(\phi^{-1}\right)'\left(\phi(\theta)\right)\mathcal{N}\left(0, \sigma^2(\theta)\right)$$

et $\left(\phi^{-1}\right)'\left(\phi(\theta)\right) = 1/\phi'(\theta)$. □

De même, si l'on dispose d'intervalles de confiance $[\widehat{\phi}_n - a_n; \widehat{\phi}_n + b_n]$ pour $\phi(\theta)$ de niveau $1-\alpha$, on en déduit (cas ϕ croissante) des intervalles de confiance de même niveau pour θ puisque

$$\mathbb{P}_\theta\left[\phi^{-1}\left(\widehat{\phi}_n - a_n\right) \leq \theta \leq \phi^{-1}\left(\widehat{\phi}_n + b_n\right)\right] \geq 1 - \alpha.$$

3.1 La méthode des moments

Elle est ainsi appelée parce que l'estimateur initial de $\phi(\theta)$ est un moment de P_θ de la forme $\mathbb{E}_\theta\left[X^k\right]$ ou, plus généralement $\mathbb{E}_\theta[h(X)]$. On peut en particulier utiliser la moyenne empirique $\overline{X}_n$ comme estimateur initial.

Exemple 1 : le modèle uniforme La loi uniforme sur un intervalle $[a;b]$ de $\mathbb{R}$, notée $\mathcal{U}([a;b])$, est la loi qui modélise un tirage "au hasard" dans $[a;b]$, c'est à dire que la probabilité d'un sous-intervalle de $[a;b]$ est proportionnelle à sa longueur. La densité correspondante s'écrit $(b-a)^{-1}\mathbb{1}_{[a;b]}(x)$ et si X a la loi $\mathcal{U}([a;b])$, alors $\mathbb{E}[X] = (a+b)/2$ et $\mathrm{Var}(X) = (b-a)^2/12$.

De cette famille de lois à deux paramètres, on peut déduire de nombreux modèles à un paramètre tels que $\{\mathcal{U}([0;\theta]), \theta > 0\}$, $\{\mathcal{U}([-\theta;\theta]), \theta > 0\}$, $\{\mathcal{U}([\theta;\theta+1]), \theta \in \mathbb{R}\}$ ou $\{\mathcal{U}([\theta-1;\theta+1]), \theta \in \mathbb{R}\}$. À titre d'illustration analysons le premier modèle pour lequel $\mathbb{E}_\theta[X] = \theta/2$ et $\mathrm{Var}_\theta(X) = \theta^2/12$. On peut utiliser $2\overline{X}_n$ comme estimateur (sans biais) de θ et son risque quadratique est $\mathbb{E}_\theta\left[\left(2\overline{X}_n - \theta\right)^2\right] = (3n)^{-1}\theta^2$. De plus $\sqrt{n}\left(2\overline{X}_n - \theta\right) \rightsquigarrow \mathcal{N}\left(0;\theta^2/3\right)$. On a donc

$$\sqrt{3n}\left(2\theta^{-1}\overline{X}_n - 1\right) \rightsquigarrow \mathcal{N}(0;1),$$

ce qui permet de construire des intervalles de confiance asymptotiques pour θ. En effet si Z est $\mathcal{N}(0;1)$ et $\mathbb{P}[|Z| \leq z_\alpha] = 1-\alpha$, alors

$$\mathbb{P}\left[\sqrt{3n}\left|2\theta^{-1}\overline{X}_n - 1\right| \leq z_\alpha\right] = \mathbb{P}\left[\frac{2\overline{X}_n}{1+z_\alpha/\sqrt{3n}} \leq \theta \leq \frac{2\overline{X}_n}{1-z_\alpha/\sqrt{3n}}\right] \underset{n\to+\infty}{\longrightarrow} 1-\alpha.$$

Exemple 2 : les lois gamma On rappelle que, pour tout $t > 0$, la fonction $x \mapsto x^{t-1}e^{-x}$ est intégrable sur $\mathbb{R}_+$ et son intégrale $\Gamma(t) = \int_0^{+\infty} x^{t-1}e^{-x}\,dx$ définit la fonction gamma sur $]0;+\infty[$. On a en particulier

$$\Gamma(1) = 1; \quad \Gamma(t+1) = t\Gamma(t); \quad \Gamma(n+1) = n! \text{ si } n \in \mathbb{N}; \quad \Gamma(1/2) = \sqrt{\pi}.$$

Il s'ensuit que, pour $t > 0, \lambda > 0$, la quantité $[\Gamma(t)]^{-1}\lambda^t x^{t-1}e^{-\lambda x}\mathbb{1}_{\mathbb{R}_+}(x)$ est une densité de probabilité par rapport à la mesure de Lebesgue. La loi associée est dite *loi gamma de paramètres t et λ* et notée $\Gamma(t,\lambda)$. Si X est une variable aléatoire de loi $\Gamma(t,\lambda)$ alors $\mathbb{E}[X] = t/\lambda$, $\mathrm{Var}(X) = t/\lambda^2$ et $\theta X \sim \Gamma(t,\lambda/\theta)$ pour $\theta > 0$. Le cas particulier $t = 1$ correspond à la *loi exponentielle de paramètre* λ, notée $\mathcal{E}(\lambda)$, de densité $\lambda e^{-\lambda x}\mathbb{1}_{\mathbb{R}_+}(x)$. Si X et Y sont deux variables aléatoires indépendantes de lois respectives $\Gamma(t_1,\lambda)$ et $\Gamma(t_2,\lambda)$ alors $X+Y$ a la loi $\Gamma(t_1+t_2,\lambda)$. Enfin la loi $\chi^2(n)$ du chi-carré à n degrés de liberté est la loi $\Gamma(n/2, 1/2)$.

Si $X_1, \ldots, X_n$ est un n-échantillon de loi $\mathcal{E}(\lambda)$, de paramètre inconnu, alors $\overline{X}_n \to \lambda^{-1}$ et $\sqrt{n}\left(\overline{X}_n - \lambda^{-1}\right) \rightsquigarrow \mathcal{N}\left(0;\lambda^{-2}\right)$. L'application de la méthode des moments conduit à estimer λ par $\widehat{\lambda}_n = \overline{X}_n^{-1} \to \lambda$ lorsque $n \to +\infty$. La méthode delta implique que $\sqrt{n}\left(\widehat{\lambda}_n - \lambda\right) \rightsquigarrow \mathcal{N}\left(0;\lambda^2\right)$. On peut aisément obtenir des intervalles de confiance asymptotiques pour λ. Si Z est $\mathcal{N}(0;1)$ et $\mathbb{P}[|Z| \leq z_\alpha] = 1-\alpha$, alors $\mathbb{P}\left[\sqrt{n}\left|\lambda\overline{X}_n - 1\right| \leq z_\alpha\right] \simeq 1-\alpha$ pour n grand, ce qui donne un intervalle de confiance approché de la forme $\left(1 \pm n^{-1/2}z_\alpha\right)/\overline{X}_n$.

Exemple 3 : paramètres de translation et de changement d'échelle Dans un certain nombre de problèmes statistiques, les paramètres sont liés à des transformations affines d'une loi fixe connue. L'exemple le plus classique est celui de la gaussienne. Si X a une loi $\mathcal{N}\left(\mu,\sigma^2\right)$, alors $X=\sigma Y+\mu$ où Y est une gaussienne standard et le paramètre μ correspond à une translation, le paramètre σ à une homothétie. De tels paramètres apparaissent naturellement lorsque l'on fait des changements d'unités. Un autre exemple est celui des lois exponentielles. Si $X\sim\mathcal{E}(\lambda)$ alors $X=Y/\lambda$ avec $Y\sim\mathcal{E}(1)$. Enfin, si $X\sim\mathcal{U}([a,b])$, alors $X=(b-a)Y+a$ avec $Y\sim\mathcal{U}([0,1])$.

Ces trois exemples ne sont que des illustrations d'une situation plus générale, pour laquelle on a une loi de base connue, de fonction de répartition F sur $\mathbb{R}$ et l'on considère la famille des lois des variables $X=aY+b$ où Y a la fonction de répartition F et les paramètres sont $a>0$ et $b\in\mathbb{R}$. La fonction de répartition G de X est donnée par

$$G(t)=\mathbb{P}[aY+b\le t]=\mathbb{P}\left[Y\le a^{-1}(t-b)\right]=F\left(a^{-1}(t-b)\right).$$

Si Y a une densité f, alors X a la densité $g(x)=a^{-1}f\left(a^{-1}(x-b)\right)$. Les paramètres a et b sont appelés respectivement paramètres d'échelle et de translation.

Comme nous venons de le voir, un certain nombre de modèles statistiques sont de ce type, avec des paramètres qui sont fonction de paramètres d'échelle et de translation. Pour $\mathcal{E}(\lambda)$, par exemple, λ est l'inverse du paramètre d'échelle. Dans de tels modèles, les moments de la loi sont des fonctions simples des paramètres d'échelle ou de translation. Pour un modèle de translation avec $X=Y+\mu$ et $\mathbb{E}[Y]=M_1$, constante connue, on a $\mathbb{E}[X]=M_1+\mu$ et la méthode des moments fournit $\widehat{\mu}_n=\overline{X}_n-M_1$. Il est immédiat de voir qu'il s'agit d'estimateurs convergents et, si $\mathbb{E}\left[Y^2\right]<+\infty$, asymptotiquement normaux. Le raisonnement est le même dans le cas d'un paramètre d'échelle θ avec $X=\theta Y$ et $M_1\neq 0$. On peut estimer θ par $\overline{X}_n/M_1$ et vérifier que l'on a encore des estimateurs convergents et asymptotiquement normaux si $\mathbb{E}\left[Y^2\right]<+\infty$.

Pour un modèle à deux paramètres, de la forme $X=\sigma Y+\mu$, en supposant, pour simplifier, que $\mathbb{E}[Y]=0$, $\mathrm{Var}(Y)=1$ et $\mathbb{E}\left[Y^4\right]<+\infty$, on pourra, comme précédemment, estimer μ par $\overline{X}_n$ et $\sigma^2=\mathrm{Var}(X)$ par la variance empirique $\widehat{\sigma}_n^2$ (ou éventuellement $\widehat{s}_n^2$ comme on l'a expliqué précédemment). Dans ce cas, le Théorème de Slutsky implique, une fois de plus, que

$$\sqrt{n}\left(\overline{X}_n-\mu\right)\widehat{\sigma}_n^{-1}\rightsquigarrow\mathcal{N}(0,1),$$

à partir de quoi l'on peut construire des intervalles de confiance asymptotiques pour μ.

3.2 Utilisation des quantiles empiriques

Reprenons d'abord le problème qui a motivé notre introduction des quantiles, à savoir l'estimation du paramètre μ pour la famille de densités $\left[\pi\left(1+(x-\mu)^2\right)\right]^{-1}$ de Cauchy. Il s'agit encore d'un problème d'estimation d'un paramètre de translation mais la méthode des moments ne s'applique pas. On sait que $\mu=x_{1/2}$ est la médiane de la loi. On peut donc l'estimer par la médiane empirique $x_{1/2}(n)$ et l'on déduit du Théorème 6 que

$$\sqrt{n}\left(x_{1/2}(n)-\mu\right)\rightsquigarrow\mathcal{N}\left(0,\pi^2/4\right).$$

On peut utiliser la même méthode pour estimer la moyenne d'une gaussienne $\mathcal{N}(\mu,1)$ et obtenir $\sqrt{n}\left(x_{1/2}(n)-\mu\right)\rightsquigarrow\mathcal{N}(0,\pi/2)$ alors que la méthode des moments donne $\sqrt{n}\left(\overline{X}_n-\mu\right)\rightsquigarrow\mathcal{N}(0,1)$, ce qui est un peu meilleur parce que la variance asymptotique est plus faible.

D'un autre côté, la médiane empirique a des avantages que n'a pas la moyenne comme le montre l'exemple suivant. Supposons que $\mu = 0$ et que l'on ait 100 observations, dont une, la centième, est aberrante, suite à une mauvaise manipulation, et égale à 30. Alors on a $\overline{X}_n = 10^{-2} \sum_{i=1}^{99} X_i + (3/10)$, de loi (conditionnellement au fait que $X_{100} = 30$) $\mathcal{N}(3/10,\, 99.10^{-4})$, ce qui est très différent de la loi $\mathcal{N}(0, 10^{-2})$ attendue. Il est facile de voir qu'alors, avec une grande probabilité, $\overline{X}_n \in [1/10, 5/10]$ car l'écart-type est à peu près 1/10. Au contraire, la médiane empirique ne sera que très peu modifiée par une telle erreur, car remplacer une des observations par une observation aberrante peut, au pire, décaler la médiane d'un rang dans la statistique d'ordre, ce qui aura très peu d'effet, les observations du milieu de l'échantillon étant à une distance de l'ordre de n^{-1} les unes des autres. Donc la médiane est beaucoup plus stable que la moyenne en cas de données erronées. On dit qu'elle est *robuste*.

4 La méthode du maximum de vraisemblance

4.1 Le cas particulier d'un ensemble fini de paramètres

4.1.1 Tests de rapports de vraisemblance entre deux lois de probabilités

On s'intéresse ici au problème suivant : on dispose d'une observation $\boldsymbol{X}$ (éventuellement multidimensionnelle, de la forme $\boldsymbol{X} = (X_1, \ldots, X_n)$) dont la loi est soit P, soit Q mais on ignore laquelle des deux est la vraie et l'on veut tester entre ces deux hypothèses : la loi $P_{\boldsymbol{X}}$ de $\boldsymbol{X}$ est P ou $P_{\boldsymbol{X}} = Q$. Comme on ne considère que deux lois, on peut toujours trouver une mesure dominante μ (par exemple $P + Q$) et décider P ou Q selon que $(dP/d\mu)(\boldsymbol{X}) > (dQ/d\mu)(\boldsymbol{X})$ ou $(dP/d\mu)(\boldsymbol{X}) < (dQ/d\mu)(\boldsymbol{X})$. En cas d'égalité, on fait un choix arbitraire. En récrivant les choses de manière un peu différente, on obtient un test qui choisit P ou Q selon que $\log[dP/d\mu(\boldsymbol{X})] - \log[dQ/d\mu(\boldsymbol{X})]$ est positif ou négatif, c'est à dire celle des deux lois pour laquelle ce que l'on appelle la *log-vraisemblance* de $\boldsymbol{X}$ est la plus grande.

Définition 14 *Étant donné une famille $\mathcal{P}$ de probabilités sur E, dominée par une mesure μ et une variable aléatoire $\boldsymbol{X}$, la* vraisemblance *et la* log-vraisemblance *de $P \in \mathcal{P}$ sont données respectivement par*

$$\frac{dP}{d\mu}(\boldsymbol{X}) \qquad et \qquad \log\left(\frac{dP}{d\mu}(\boldsymbol{X})\right),$$

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

Les tests entre deux lois de probabilités P et Q qui choisissent entre elles selon que la valeur du rapport des vraisemblances $[(dP/d\mu)(\boldsymbol{X})]/[(dQ/d\mu)(\boldsymbol{X})]$ est plus ou moins grande sont dits tests de rapport de vraisemblances.

On remarque que si $\boldsymbol{X} = (X_1, \ldots, X_n)$ où les $X_i \in E$ sont i.i.d. de loi $\overline{P}$, donc $P = \overline{P}^{\otimes n}$ est une loi produit, et que la mesure μ s'écrit $\nu^{\otimes n}$ où la mesure ν sur E domine tout, on aura

$$\frac{dP}{d\mu}(\boldsymbol{X}) = \prod_{i=1}^{n} \frac{d\overline{P}}{d\nu}(X_i) \qquad \text{et} \qquad \log\left(\frac{dP}{d\mu}(\boldsymbol{X})\right) = \sum_{i=1}^{n} \log\left(\frac{d\overline{P}}{d\nu}(X_i)\right).$$

L'intérêt d'utiliser la log-vraisemblance, c'est qu'elle se présente dans ce cas sous la forme d'une somme de variables i.i.d.

Lemme 2 *Si les variables $X_1, \ldots, X_n$ sont i.i.d. de loi P, si $Q \neq P$ et si P et Q sont dominées par ν, alors*

$$\mathbb{P}_P \left[\sum_{i=1}^n \log \left(\frac{dQ}{d\nu}(X_i) \right) \geq \sum_{i=1}^n \log \left(\frac{dP}{d\nu}(X_i) \right) \right] \leq \rho^n(P,Q) \underset{n \to +\infty}{\longrightarrow} 0,$$

où $\rho(P,Q)$ est définie par

$$0 \leq \rho(P,Q) = \int_E \sqrt{\frac{dP}{d\nu}(x) \frac{dQ}{d\nu}(x)}\, d\nu(x) < 1.$$

Démonstration : On remarque d'abord, en remplaçant ν par $\mu = f \cdot \nu$ que $\rho(P,Q)$ a une valeur indépendante du choix de la mesure dominante ν et ne dépend donc que de P et Q, d'où la notation. Par ailleurs, d'après l'Inégalité de Cauchy-Schwarz et le Théorème de Fubini,

$$0 \leq \rho(P,Q) \leq 1 \quad \text{et} \quad \rho(P,Q) < 1 \quad \text{si } P \neq Q; \qquad \rho(P^{\otimes n}, Q^{\otimes n}) = \rho^n(P,Q).$$

Pour montrer la borne de probabilité, on utilise que, quelle que soit la v.a.r. Z,

$$\mathbb{P}[Z \geq 0] \leq \mathbb{E}\left[e^{tZ}\right] \quad \text{pour tout } t \geq 0. \tag{III.4.1}$$

En effet, si $t \geq 0$,

$$\mathbb{E}\left[e^{tZ}\right] \geq \mathbb{E}\left[e^{tZ} \mathbb{1}_{Z \geq 0}\right] \geq \mathbb{E}\left[\mathbb{1}_{Z \geq 0}\right] = \mathbb{P}[Z \geq 0].$$

En appliquant (III.4.1) avec

$$Z = \log \left(\prod_{i=1}^n \frac{dQ}{d\nu}(X_i) \right) - \log \left(\prod_{i=1}^n \frac{dP}{d\nu}(X_i) \right) \quad \text{et} \quad t = 1/2,$$

on obtient la borne $\rho(P^{\otimes n}, Q^{\otimes n}) = \rho^n(P,Q)$ cherchée. □

On a évidemment le même résultat en échangeant les rôles de P et Q, ce qui montre que les erreurs des tests de rapports de vraisemblance entre P et Q tendent vers 0 quand $n \to +\infty$.

4.1.2 Maximum de vraisemblance sur un ensemble fini

On suppose ici que l'ensemble Θ des paramètres est un ensemble fini quelconque et que le modèle est identifiable. Dans ce cas, le modèle est dominé par une mesure positive ν, par exemple par $\sum_{\theta \in \Theta} P_\theta$, avec des densités $f_\theta = dP_\theta/d\nu$. On observe $X_1, \ldots, X_n$ i.i.d. de loi inconnue P_θ. Faire tous les tests de rapport de vraisemblances entre les points de Θ revient à comparer entre elles toutes les vraisemblances et comme on est sur un ensemble fini, il y en a une qui sera plus grande que toutes les autres, donc un $\widehat{\theta}$ tel que les tests de $P_{\widehat{\theta}}$ contre n'importe quel autre P_θ acceptent $P_{\widehat{\theta}}$. Il est alors naturel de choisir $\widehat{\theta}$ comme estimateur de θ. Cet estimateur est dit *estimateur du maximum de vraisemblance* parce qu'il correspond au point pour lequel la vraisemblance est maximale. On peut aisément analyser ses performances lorsque n est grand. En effet, si θ désigne la vraie valeur du paramètre, d'après le Lemme 2 et la finitude de Θ.

$$\begin{aligned}
\mathbb{P}_\theta \left[\widehat{\theta} \neq \theta \right] &\leq \mathbb{P}_\theta \left[\exists \theta' \neq \theta \in \Theta \;\middle|\; \sum_{i=1}^n \log \left(\frac{dP_{\theta'}}{d\nu}(X_i) \right) \geq \sum_{i=1}^n \log \left(\frac{dP_\theta}{d\nu}(X_i) \right) \right] \\
&\leq \sum_{\theta' \in \Theta, \theta' \neq \theta} \mathbb{P}_\theta \left[\sum_{i=1}^n \log \left(\frac{dP_{\theta'}}{d\nu}(X_i) \right) \geq \sum_{i=1}^n \log \left(\frac{dP_\theta}{d\nu}(X_i) \right) \right] \underset{n \to +\infty}{\longrightarrow} 0.
\end{aligned}$$

Donc, sur un ensemble fini, les estimateurs du maximum de vraisemblance sont toujours convergents.

4.2 La méthode générale du maximum de vraisemblance

4.2.1 Définition

Définition 15 *Étant donné un modèle statistique* $\{P_\theta, \theta \in \Theta\}$ *dominé, la famille des densités correspondantes* $dP_\theta/d\mu$*, l'observation* $\boldsymbol{X}$ *et la fonction de vraisemblance* $\theta \mapsto (dP_\theta/d\mu)(\boldsymbol{X})$ *de* Θ *dans* $\mathbb{R}_+$*, on appelle estimateur du maximum de vraisemblance de* θ *tout point* $\widehat{\theta}$ *de* Θ *tel que* $(dP_{\widehat{\theta}}/d\mu)(\boldsymbol{X}) = \sup_{\theta\in\Theta}(dP_\theta/d\mu)(\boldsymbol{X})$.

Pour un modèle d'échantillonnage avec des observations $X_1, \ldots, X_n$ de densités individuelles f_θ, la log-vraisemblance s'écrit $\sum_{i=1}^n \log[f_\theta(X_i)]$ et un estimateur du maximum de vraisemblance $\widehat{\theta}_n$ est caractérisé par le fait que

$$\sum_{i=1}^{n} \log\left(f_{\widehat{\theta}_n}(X_i)\right) = \sup_{\theta\in\Theta} \sum_{i=1}^{n} \log\left(f_\theta(X_i)\right). \tag{III.4.2}$$

Comme on vient de le voir, si Θ est fini un estimateur du maximum de vraisemblance existe et, dans un modèle d'échantillonnage, il est asymptotiquement convergent, donc asymptotiquement unique. Dans le cas général, on peut rencontrer tous les problèmes habituels liés à l'optimisation sur un ensemble infini, à savoir :
— le maximum de vraisemblance n'existe pas ;
— le maximum de vraisemblance n'est pas unique ;
— le maximum de vraisemblance existe mais on ne sait pas le calculer.
Nous verrons ci-après des exemples de telles situations. Néanmoins, dans les bons cas, c'est à dire lorsque Θ n'est pas trop gros (Θ est un compact d'un espace euclidien par exemple) et l'application $\theta \mapsto P_\theta$ est suffisamment régulière, on peut montrer que les choses se passent bien, comme dans le cas où Θ est fini.

4.2.2 Quelques exemples de modèles d'échantillonnage

Le modèle gaussien ordinaire Si l'on dispose de n observations i.i.d. de loi $\mathcal{N}\left(\mu, \sigma^2\right)$, avec deux paramètres μ et σ^2 inconnus, la log-vraisemblance s'écrit

$$-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2. \tag{III.4.3}$$

On remarque d'abord que, indépendamment de la valeur de σ^2, la maximisation de la log-vraisemblance par rapport à μ correspond à la minimisation de $\sum_{i=1}^n (X_i-\mu)^2$, dont la solution est donnée par $\widehat{\mu}_n = \overline{X}_n$. Il suffit alors de maximiser (III.4.3) par rapport à σ^2 avec $\mu = \overline{X}_n$, ce qui conduit à la variance empirique $\widehat{\sigma}_n^2$ donnée par (III.2.2).

La loi de Poisson On rappelle que la loi de Poisson de paramètre $\lambda > 0$, notée $\mathcal{P}(\lambda)$, est une loi sur $\mathbb{N}$ et si N_n est une telle variable, $\mathbb{P}_\lambda[N = k] = e^{-\lambda}\lambda^k/(k!)$ pour $k \in \mathbb{N}$. Comme dans tout modèle statistique correspondant à des v.a. à valeurs dans un espace dénombrable, il y a une mesure dominante "naturelle" qui est la mesure de comptage et la densité correspondante est $f_\lambda(k) = \mathbb{P}_\lambda[N = k]$. L'espace des paramètres est ici $\Lambda =]0, +\infty[$ qui est ouvert. Le modèle de Poisson est obtenu comme limite de modèles

binomiaux. Si $N_n \sim \mathcal{B}(n, \lambda_n/n)$ et $\lambda_n \underset{n\to+\infty}{\longrightarrow} \lambda$, alors $N_n \rightsquigarrow \mathcal{P}(\lambda)$. C'est à dire que la loi de Poisson modélise des phénomènes où l'on compte le nombre de succès dans un très grand nombre d'essais chacun ayant une probabilité de succès très faible, comme pour les phénomènes radioactifs. Les propriétés de cette loi sont classiques : $\mathbb{E}_\lambda[N] = \lambda = \mathrm{Var}_\lambda(N)$ et si les variables N_j, $1 \le j \le J$, sont indépendantes de lois respectives $\mathcal{P}(\lambda_j)$, alors $\sum_{j=1}^J N_j \sim \mathcal{P}\left(\sum_{j=1}^J \lambda_j\right)$.

Si l'on dispose de n v.a. indépendantes, $N_1, \ldots, N_n$ de loi $\mathcal{P}(\lambda)$, la log-vraisemblance s'écrit

$$\sum_{i=1}^n \left[-\lambda + N_i \log\lambda - \log(N_i!)\right] = -n\lambda + (\log\lambda)\sum_{i=1}^n N_i - \sum_{i=1}^n \log(N_i!),$$

dont la maximisation est immédiate et conduit à $\widehat{\lambda}_n = n^{-1}\sum_{i=1}^n N_i$ (qui est aussi l'estimateur de la méthode des moments), sauf si toutes les observations sont nulles, auquel cas $\sum_{i=1}^n N_i = 0$ qui n'est pas une valeur licite puisque $0 \notin \Lambda$. On notera que cet évènement peut très bien se produire puisque il a une probabilité positive $e^{-n\lambda}$ qui peut être grande si n et λ sont petits. Par contre, asymptotiquement, la probabilité de cet évènement tend vers zéro.

Lois uniformes sur $[0, \theta]$ Pour un n-échantillon, la vraisemblance s'écrit

$$\theta^{-n}\prod_{i=1}^n \mathbb{1}_{[0,\theta]}(X_i) = \theta^{-n}\prod_{i=1}^n \mathbb{1}_{[X_i,+\infty]}(\theta) = \theta^{-n}\mathbb{1}_{[X_{(n)},+\infty]}(\theta),$$

où $X_{(n)}$ est la plus grande des observations. Cette fonction de θ est discontinue en $X_{(n)}$, nulle avant puis positive et décroissante et atteint donc son maximum en $X_{(n)}$ qui est l'estimateur du maximum de vraisemblance, lequel ne s'obtient donc pas toujours en annulant une dérivée. Dans ce cas, il est facile de calculer sa loi. Sa fonction de répartition s'écrit $\mathbb{P}_\theta[X_{(n)} \le t] = (t/\theta)^n$ pour $0 \le t \le \theta$ et la densité correspondante est $g_\theta(x) = \theta^{-n}x^{n-1}\mathbb{1}_{[0,\theta]}(x)$, de sorte que $\mathbb{E}_\theta[X_{(n)}] = n\theta/(n+1) < \theta$. C'est un estimateur biaisé, ce qui était prévisible puisque $X_{(n)} < \theta$, $\mathbb{P}_\theta$-p.s.. Comme $\mathbb{E}_\theta\left[X_{(n)}^2\right] = n\theta^2/(n+2)$, le risque quadratique de l'estimateur du maximum de vraisemblance s'écrit

$$\mathbb{E}_\theta\left[(X_{(n)} - \theta)^2\right] = \frac{n\theta^2}{n+2} + \theta^2 - 2\frac{n\theta^2}{n+1} = \frac{2\theta^2}{(n+1)(n+2)}.$$

On peut remarquer que si l'on utilise la méthode des moments à partir de $\mathbb{E}_\theta[X_i] = \theta/2$, on trouve l'estimateur $2\overline{X}_n$ qui est sans biais et de variance $\theta^2/(3n)$. On voit que, dans ce cas, l'estimateur du maximum de vraisemblance, bien que biaisé, est bien meilleur que celui de la méthode des moments puisque son risque quadratique est, asymptotiquement, infiniment plus petit.

Lois uniformes sur $[\theta, \theta+1]$ Dans ce cas, la vraisemblance s'écrit

$$\prod_{i=1}^n \mathbb{1}_{[\theta,\theta+1]}(X_i) = \mathbb{1}_{[\theta,+\infty[}(X_{(1)})\mathbb{1}_{]-\infty,\theta+1]}(X_{(n)}) = \mathbb{1}_{[X_{(n)}-1,X_{(1)}]}(\theta).$$

Cette vraisemblance ne prend que les deux valeurs 0 et 1 et son maximum est atteint en tout point de l'intervalle (p.s. non vide) $[X_{(n)} - 1, X_{(1)}]$. Il n'y a donc pas, dans

ce cas, unicité du maximum de vraisemblance. On peut néanmoins vérifier aisément en calculant, comme on l'a fait déjà plus haut, la loi de $X_{(n)}$ et celle de $X_{(1)}$, que $X_{(1)} \underset{n\to+\infty}{\longrightarrow} \theta$ et $X_{(n)} \underset{n\to+\infty}{\longrightarrow} \theta+1$, $\mathbb{P}_\theta$-p.s., de sorte qu'asymptotiquement tous ces estimateurs sont très proches et tous convergent vers θ. Par contre, lorsque n est petit, on tombe sur une réelle indétermination.

Translatées de la loi de Cauchy La loi de Cauchy a la densité $\pi^{-1}\left(1+x^2\right)^{-1}$. Elle donne lieu à un modèle de translation correspondant à la famille des densités $\pi^{-1}\left(1+(x-\theta)^2\right)^{-1}$ avec $\theta \in \mathbb{R}$. Dans un tel modèle, la log-vraisemblance d'un n-échantillon s'écrit $-n\log\pi - \sum_{i=1}^n \log\left(1+(X_i-\theta)^2\right)$. Il est facile de voir qu'une telle fonction de θ a au moins un maximum sur $\mathbb{R}$, lequel s'obtient en annulant la dérivée mais qu'il est impossible de résoudre analytiquement l'équation correspondante car cela revient à chercher les racines d'un polynôme de degré $2n-1$. La méthode fondée sur la médiane empirique est clairement beaucoup plus simple.

4.2.3 Extension

On est souvent amené, dans un modèle statistique, à considérer plusieurs paramétrages. Un exemple simple est le suivant. Si l'on dispose d'observations de loi exponentielle, les densités peuvent s'écrire $\theta\exp(-\theta x)\mathbb{1}_{\mathbb{R}_+}(x)$ avec $\theta>0$ et l'espérance est θ^{-1}. On peut tout aussi bien paramétrer la même famille de lois par son espérance en considérant les densités $\lambda^{-1}\exp(-x/\lambda)\mathbb{1}_{\mathbb{R}_+}(x)$ avec $\lambda>0$. C'est évidemment la même famille de lois. On a simplement fait un changement de paramètre bijectif $\lambda=\theta^{-1}$. On voit immédiatement que les estimateurs du maximum de vraisemblance construits à partir d'un n-échantillon vérifient $\widehat{\lambda}_n = \widehat{\theta}_n^{-1}$. C'est un phénomène général. Si l'on fait, dans un modèle statistique, un changement de paramètre bijectif $\lambda=g(\theta)$, les estimateurs du maximum de vraisemblance construits à partir d'un n-échantillon sont liés par $\widehat{\lambda}_n = g\left(\widehat{\theta}_n\right)$.

De manière plus générale, si l'on cherche à estimer une fonction $\lambda=g(\theta)$ du paramètre, on appellera estimateur du maximum de vraisemblance de λ l'estimateur $g\left(\widehat{\theta}_n\right)$, même si g n'est pas bijective. Dans tous les cas, pour étudier les propriétés asymptotiques de $g\left(\widehat{\theta}_n\right)$ connaissant celles de $\widehat{\theta}_n$ il suffira d'appliquer la méthode delta.

4.3 Application au modèle exponentiel à un paramètre

4.3.1 Introduction

Les modèles exponentiels constituent une classe générique de modèles statistiques qui inclut un certain nombre d'exemples classiques. Que peut-on trouver de commun entre la famille des lois $\mathcal{N}(\theta,1)$ et celle des lois de Poisson $\mathcal{P}(\theta)$? Si l'on écrit la densité $\mathcal{N}(\theta,1)$ par rapport à la mesure dominante $\mathcal{N}(0,1)$ de densité $(2\pi)^{-1/2}\exp\left[-x^2/2\right]$, on trouve $\exp\left[-\theta^2/2\right]\exp[\theta x]$ et, pour la loi de Poisson et la mesure dominante $(x!)^{-1}d\mu(x)$ où μ est la mesure de comptage sur $\mathbb{N}$, on trouve $e^{-\theta}\exp(-x\log\theta)$. Dans les deux cas, après avoir choisi une mesure dominante convenable ne dépendant que des paramètres connus, la densité se décompose comme un produit de deux facteurs, le premier ne dépendant que des paramètres et le second étant l'exponentielle du produit d'une fonction de la variable par une fonction du paramètre. Comme on va le voir, cette décomposition particulière intervient également dans d'autres modèles classiques, d'où l'intérêt d'étudier de façon systématique cette classe de modèles.

4.3.2 Préliminaires d'analyse

Étant donné, sur l'espace mesurable E, une mesure positive ν et une fonction mesurable T, non presque sûrement constante pour ν, on peut considérer l'ensemble

$$H = \left\{ \eta \in \mathbb{R} \;\middle|\; \int_E \exp[\eta T(x)]\, d\nu(x) < +\infty \right\}. \tag{III.4.4}$$

On remarquera que si T est constante ν p.s., on arrive à des trivialités.

Si cet ensemble n'est pas vide, la convexité de la fonction exponentielle entraîne que H est un intervalle : si η_1 et η_2 appartiennent à H, il en va de même de $\lambda\eta_1 + (1-\lambda)\eta_2$ pour $\lambda \in [0;1]$ puisque $\exp[\lambda\eta_1 T + (1-\lambda)\eta_2 T] \le \lambda \exp[\eta_1 T] + (1-\lambda)\exp[\eta_2 T]$. On supposera dans toute la suite que cet intervalle n'est pas réduit à un seul point, donc d'intérieur $\overset{\circ}{H}$ non vide. Un résultat classique d'analyse concernant les transformées de Laplace dit la chose suivante :

Proposition 4 *La fonction $\eta \mapsto J(\eta) = \int_E \exp[\eta T(x)]\, d\nu(x)$ est indéfiniment différentiable sur $\overset{\circ}{H}$ de dérivées successives*

$$J^{(i)}(\eta) = \int [T(x)]^i \exp[\eta T(x)]\, d\nu(x).$$

4.3.3 Le modèle exponentiel naturel

Étant donné ν et T comme ci-dessus, on peut définir une famille de densités par rapport à la mesure ν, indexée par H, $\{g_\eta, \eta \in H\}$, en posant

$$g_\eta(x) = \exp[\eta T(x) - A(\eta)] \quad \text{avec } A(\eta) = \log\big(J(\eta)\big) = \log\left(\int_E \exp[\eta T(x)]\, d\nu(x)\right). \tag{III.4.5}$$

L'ensemble des lois $R_\eta = g_\eta \cdot \nu$ forme un *modèle exponentiel naturel* (ou *canonique*) indexé par *l'espace des paramètres naturels* H. Par exemple, les lois exponentielles ordinaires de densités $\eta \exp[-\eta x]$ par rapport à la mesure de Lebesgue ν sur $\mathbb{R}_+$ forment un tel modèle avec $H =]0; +\infty[$, $T(x) = -x$ et $A(\eta) = -\log \eta$. Il est important de noter que si la famille des lois R_η est définie de manière unique, sa représentation sous la forme d'un modèle exponentiel ne l'est pas. En effet, on peut remplacer la mesure ν par la mesure $a\nu$ avec $a > 0$ et la densité g_η par $a^{-1} g_\eta$, c'est à dire A par $A + \log a$, sans changer le modèle statistique.

i) Ce modèle est identifiable car si on avait $g_\eta = g_{\eta'}$ pour $\eta \neq \eta'$, on aurait

$$(\eta - \eta')T(x) = A(\eta) - A(\eta') \quad \nu \text{ p.s.},$$

ce qui impliquerait que T est p.s. constante, ce que l'on a exclu.

ii) Les densités g_η sont strictement positives, donc toutes les lois R_η sont mutuellement absolument continues et toutes équivalentes à ν.

iii) Les propriétés de la transformée de Laplace vues ci-dessus impliquent que la fonction A définie par (III.4.5) est indéfiniment différentiable sur $\overset{\circ}{H}$, les dérivées se calculant en dérivant sous le signe $\int$. On obtient ainsi

$$A'(\eta) = \frac{\int_E T(x) \exp[\eta T(x)]\, d\nu(x)}{\int_E \exp[\eta T(x)]\, d\nu(x)} = e^{-A(\eta)} \int_E T(x) \exp[\eta T(x)]\, d\nu(x) = \mathbb{E}_\eta[T(X)]. \tag{III.4.6}$$

Un calcul analogue montre que

$$A''(\eta) = \mathrm{Var}_\eta(T(X)) \geq 0, \tag{III.4.7}$$

les espérances étant prises, dans les deux cas, pour la loi R_η. La variance de $T(X)$ ne peut s'annuler (sinon $T(X)$ serait ν-p.s. constante), donc la fonction A est strictement convexe.

iv) Si l'on dispose d'un n-échantillon $X_1, \ldots, X_n$ de loi R_η du modèle exponentiel précédent, la famille des lois $R_\eta^{\otimes n}$ du vecteur $\boldsymbol{X} = (X_1, \ldots, X_n)$ forme un nouveau modèle exponentiel par rapport à la mesure $\nu^{\otimes n}$, de densités de la forme

$$g_\eta(x_1) \cdots g_\eta(x_n) = \exp\left[\eta \sum_{i=1}^{n} T(x_i) - nA(\eta)\right].$$

Donc lorsque l'on dispose d'un n-échantillon, il suffit de vérifier que le modèle est exponentiel en considérant le cas particulier $n = 1$.

4.3.4 Le modèle exponentiel général

Définition 16 *Le* modèle exponentiel général *à un paramètre est la donnée d'une famille de densités sur E (par rapport à une mesure μ de référence) de la forme*

$$f_\theta(x) = C(\theta) \exp[Q(\theta)T(x)]h(x) \quad \textit{avec } Q(\theta), T(x) \in \mathbb{R}, \quad C(\theta), h(x) \in \mathbb{R}_+, \tag{III.4.8}$$

et $\theta \in \Theta$ où Θ est un intervalle (non dégénéré) de $\mathbb{R}$. La fonction Q est continuement différentiable, de dérivée $Q' \neq 0$ sur Θ et la fonction T n'est pas p.s. constante pour la mesure $\nu = h \cdot \mu$.

Comme il s'agit d'une famille de densités, on a

$$C(\theta) = \left[\int_E \exp[Q(\theta)T(x)]h(x)\, d\mu(x)\right]^{-1}.$$

On notera que Q' a un signe constant, donc que Q est strictement monotone, que l'application $\theta \mapsto Q(\theta)$ est injective et que $Q(\Theta)$ est un intervalle d'intérieur non vide. Par ailleurs, on peut remarquer que la représentation précédente n'est pas unique puisque, quel que soit $\alpha \neq 0$, on peut toujours remplacer $Q(\theta)$ par $\alpha Q(\theta)$ et $T(x)$ par $\alpha^{-1}T(x)$.

On peut toujours associer à ce modèle exponentiel général, par le changement de paramètre $\eta = Q(\theta)$, un modèle naturel, ensemble des densités g_η, $\eta \in H$, données par (III.4.5) avec $\nu = h \cdot \mu$ et $A(\eta) = -\log C(\theta) = -\log\left(Q^{-1}(\eta)\right)$. L'espace H des paramètres naturels donné par (III.4.4) contient $Q(\Theta)$; il est donc d'intérieur non vide. Remarquons qu'il se peut fort bien que H soit plus grand que $Q(\Theta)$. Donc on passe du modèle naturel au modèle général par un changement de paramètre $\theta = Q^{-1}(\eta)$ précédé, éventuellement, d'une restriction de H à $Q(\Theta)$. On notera $\mathbb{E}_\theta$ les espérances évaluées dans le modèle général et $\mathbb{E}_\eta$ celles évaluées dans le modèle canonique. On a évidemment la relation

$$\int k(x) f_\theta(x)\, d\mu(x) = \mathbb{E}_\theta[k(X)] = \mathbb{E}_\eta[k(X)] = \int k(x) g_\eta(x)\, d\nu(x) \quad \text{si } \eta = Q(\theta). \tag{III.4.9}$$

Les propriétés du modèle général se déduisent immédiatement de celles du modèle canonique.

i) Le modèle est identifiable.

ii) Les densités f_θ sont > 0 sur l'ensemble $\{x \,|\, h(x) > 0\}$ et toutes les lois associées $P_\theta = f_\theta \cdot \mu$ sont mutuellement absolument continues de même support que $\nu = h \cdot \mu$.

iii) Le n-échantillon correspondant forme aussi un modèle exponentiel avec la famille des densités produit

$$f_\eta(x_1) \cdots f_\eta(x_n) = [C(\theta)]^n \exp\left[Q(\theta) \sum_{i=1}^n T(x_i)\right] h(x_1) \cdots h(x_n)$$

par rapport à la mesure $\mu^{\otimes n}$.

4.3.5 Exemples de modèles exponentiels

Le modèle binomial : Si X est une variable de loi $\mathcal{B}(n, \theta)$ avec $0 < \theta < 1$, sa densité par rapport à la mesure de comptage μ sur $\mathbb{N}$ s'écrit

$$\binom{n}{x} \theta^x (1-\theta)^{n-x} \mathbb{1}_{\{0;\ldots;n\}}(x) = (1-\theta)^n \exp\left[x \log\left(\frac{\theta}{1-\theta}\right)\right] \binom{n}{x} \mathbb{1}_{\{0;\ldots;n\}}(x),$$

ce qui correspond bien à une décomposition de la forme (III.4.8). Notons que ceci ne fonctionne plus si $\theta = 0$ ou 1 car la loi binomiale dégénère alors en une masse de Dirac qui n'est pas équivalente aux autres lois de la famille.

Les lois de Poisson : Comme on l'a vu, la densité par rapport à la mesure de comptage sur $\mathbb{N}$ de la loi de Poisson $\mathcal{P}(\lambda)$, $\lambda > 0$, s'écrit

$$e^{-\lambda} \lambda^x / x! = e^{-\lambda} \exp(x \log \lambda)(x!)^{-1}.$$

Les densités gaussiennes : Elles s'écrivent

$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right],$$

ce qui donne les décompositions suivantes, selon que c'est μ ou σ^2 qui est le paramètre à estimer. Si σ est connu, on peut écrire la densité sous la forme

$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{\mu^2}{2\sigma^2}\right] \exp\left[\frac{\mu x}{\sigma^2}\right] \exp\left[-\frac{x^2}{2\sigma^2}\right],$$

qui correspond à un modèle exponentiel naturel dominé par la loi $\mathcal{N}\left(0; \sigma^2\right)$ de densité $\left(2\pi\sigma^2\right)^{-1/2} \exp\left[-x^2/\left(2\sigma^2\right)\right] dx$ avec $T(x) = x/\sigma^2$. Si μ est connu, on peut choisir $Q\left(\sigma^2\right) = 1/\sigma^2$ et $T(x) = -(x-\mu)^2/2$, entre autres représentations possibles.

Les lois gamma : La densité $\Gamma(t, \lambda)$ s'écrit $[\Gamma(t)]^{-1} \lambda^t x^{t-1} e^{-\lambda x} \mathbb{1}_{\mathbb{R}_+}(x)$ et peut donc, selon que λ ou t est inconnu, se mettre sous l'une des formes suivantes :

$$[\Gamma(t)]^{-1} \lambda^t \exp(-\lambda x) x^{t-1} \mathbb{1}_{\mathbb{R}_+}(x) \quad \text{ou} \quad [\Gamma(t)]^{-1} \lambda^t \exp(t \log x) e^{-\lambda x} x^{-1} \mathbb{1}_{\mathbb{R}_+}(x).$$

Les modèles uniformes : Une loi uniforme sur un intervalle $[a; b]$ a une densité à support $[a; b]$ par définition. Les paramètres étant les extrémités de l'intervalle, le support de la loi varie avec eux, ce qui démontre que les lois uniformes ne peuvent jamais correspondre à des modèles exponentiels.

4.3.6 Le maximum de vraisemblance dans le modèle exponentiel

Commençons par étudier le modèle naturel. Si l'on dispose de n observations indépendantes $X_1, \ldots, X_n$ de densité g_η, la log-vraisemblance s'écrit alors

$$L_\eta(X_1, \ldots, X_n) = \sum_{i=1}^{n} \eta T(X_i) - nA(\eta).$$

Comme, sur $\overset{\circ}{H}$, $A'' > 0$, L_η est une fonction strictement concave de η. Donc, s'il existe un point $\widehat{\eta}_n$ intérieur à H tel que $L'_{\widehat{\eta}_n}(X_1, \ldots, X_n) = 0$, c'est l'unique estimateur du maximum de vraisemblance de η. Or, si l'on note $\overline{T}_n = n^{-1} \sum_{i=1}^{n} T(X_i)$, la solution de l'équation $L'_{\widehat{\eta}_n}(X_1, \ldots, X_n) = 0$ est donnée par

$$A'(\widehat{\eta}_n) = \mathbb{E}_{\widehat{\eta}_n}[T(X)] = \overline{T}_n \quad \text{ou} \quad \widehat{\eta}_n = \phi^{-1}\left(\overline{T}_n\right) \quad \text{avec } \phi(\eta) = A'(\eta). \tag{III.4.10}$$

En effet, comme $\phi' = A'' > 0$, la fonction ϕ est continue et strictement croissante, donc bijective. En outre, son inverse ϕ^{-1} est différentiable et $(\phi^{-1})'[\phi(\eta)] = 1/A''(\eta)$.

Supposons que la vraie valeur η du paramètre est intérieure à H, ce qui implique que $\phi(\eta)$ est intérieur à $\phi(H)$. La loi des grands nombres et le Théorème limite central impliquent, d'après (III.4.6) et (III.4.7) que

$$\overline{T}_n \underset{n\to+\infty}{\longrightarrow} \mathbb{E}_\eta[T(X)] = \phi(\eta) \in \phi\left(\overset{\circ}{H}\right) \ \mathbb{P}_\eta\text{-p.s.} \quad \text{et} \quad \sqrt{n}\left(\overline{T}_n - \phi(\eta)\right) \rightsquigarrow \mathcal{N}\left(0, \phi'(\eta)\right).$$

Donc, p.s., pour n assez grand, $\overline{T}_n$ est aussi intérieur à $\phi(H)$ et (III.4.10) a une solution unique qui converge vers $\phi^{-1}(\phi(\eta)) = \eta$. Ceci implique que, pour n grand, l'estimateur du maximum de vraisemblance $\widehat{\eta}_n$ défini par (III.4.10) existe et qu'il converge vers la vraie valeur η du paramètre lorsque $n \to +\infty$. De plus, comme ϕ^{-1} est différentiable, par la méthode delta,

$$\sqrt{n}\left(\widehat{\eta}_n - \eta\right) = \sqrt{n}\left[\phi^{-1}\left(\overline{T}_n\right) - \phi^{-1}(\phi(\eta))\right] \rightsquigarrow \mathcal{N}\left(0, \frac{\phi'(\eta)}{[A''(\eta)]^2}\right) = \mathcal{N}\left(0, \frac{1}{A''(\eta)}\right),$$

ou encore,

$$\sqrt{n}\left(\widehat{\eta}_n - \eta\right) \rightsquigarrow \mathcal{N}\left(0, 1/\operatorname{Var}_\eta(T)\right). \tag{III.4.11}$$

Donc l'estimateur du maximum de vraisemblance est convergent et asymptotiquement normal. Remarquons qu'ici l'estimateur du maximum de vraisemblance est un estimateur de la méthode des moments appliqué à la fonction $\phi(\eta) = \mathbb{E}_\eta[T(X)]$.

Pour passer au modèle général, il suffit simplement d'appliquer la méthode delta une seconde fois. En effet, si l'on dispose de n observations indépendantes $X_1, \ldots, X_n$ de densité f_θ donnée par (III.4.8) et l'on suppose que le vrai paramètre θ est à l'intérieur de l'espace Θ des paramètres, dans le modèle naturel associé, le vrai paramètre $\eta = Q(\theta)$ est (compte-tenu des propriétés de Q), intérieur à $Q(\Theta)$, donc à H. D'après ce qui précède $\widehat{\eta}_n \to \eta = Q(\theta)$, $\mathbb{P}_\theta$ p.s. Donc, pour n assez grand, $\widehat{\eta}_n \in Q(\Theta)$ et l'on peut définir $\widehat{\theta}_n = Q^{-1}(\widehat{\eta}_n) \to Q^{-1}(\eta) = \theta$, $\mathbb{P}_\theta$ p.s. C'est l'estimateur du maximum de vraisemblance de θ puisque Q^{-1} correspond à un changement de paramètre. On peut à nouveau appliquer la méthode delta à partir de (III.4.11) en utilisant le fait que $\left(Q^{-1}\right)'(Q(\theta)) = 1/Q'(\theta)$, ce qui donne, compte-tenu de (III.4.9),

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) = \sqrt{n}\left(Q^{-1}(\widehat{\eta}_n) - Q^{-1}(\eta)\right) \rightsquigarrow \mathcal{N}\left(0, \left[\operatorname{Var}_\theta(T)(Q'(\theta))^2\right]^{-1}\right).$$

On peut remarquer que $\widehat{\theta}_n = Q^{-1}(\widehat{\eta}_n)$ est la solution (unique) de l'équation $\mathbb{E}_{\widehat{\theta}_n}[T(X)] = \overline{T}_n$. On a ainsi démontré le théorème suivant :

Théorème 7 *Soit $X_1, \ldots, X_n$ un n-échantillon d'un modèle exponentiel régulier de densités $f_\theta(x) = C(\theta)\exp[Q(\theta)T(x)]h(x)$ par rapport à la loi μ, où θ décrit un intervalle Θ de $\mathbb{R}$. Supposons que $\theta \in \overset{\circ}{\Theta}$. Alors, $\mathbb{P}_\theta$-p.s., pour n assez grand, l'estimateur du maximum de vraisemblance $\widehat{\theta}_n$ existe et c'est l'unique solution de l'équation*

$$\mathbb{E}_{\widehat{\theta}_n}[T(X)] = n^{-1}\sum_{i=1}^{n} T(X_i). \tag{III.4.12}$$

De plus,

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \rightsquigarrow \mathcal{N}\left(0, \left[\mathrm{Var}_\theta(T)(Q'(\theta))^2\right]^{-1}\right).$$

Remarque : Les calculs précédents nous ont permis de démontrer le Théorème 7 mais ne sont pas très commodes. En pratique on écrira simplement la log-vraisemblance que l'on maximisera en θ par un calcul direct plutôt que d'utiliser (III.4.12). On évaluera ensuite la variance asymptotique en appliquant la méthode delta.

4.3.7 Applications

Dans le modèle de Poisson $\mathcal{P}(\theta), \theta > 0$, $\widehat{\theta}_n = \overline{X}_n$ pourvu que $\overline{X}_n > 0$, sinon l'estimateur du maximum de vraisemblance n'existe pas. Mais il existe toujours pour n suffisamment grand puisque $\overline{X}_n \underset{n\to+\infty}{\longrightarrow} \theta > 0$ avec

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \rightsquigarrow \mathcal{N}(0, \theta) \qquad \text{et} \qquad \mathbb{E}_\theta\left[\left(\widehat{\theta}_n - \theta\right)^2\right] = \theta/n.$$

La condition "n assez grand" n'est pas superflue car on peut avoir $\overline{X}_n = 0$ si n est trop petit.

Dans le modèle $\mathcal{N}(\theta, 1)$ avec $\theta \in \Theta = [0; +\infty[$, l'estimateur du maximum de vraisemblance se calcule explicitement et l'on trouve $\widehat{\theta}_n = \max\left\{\overline{X}_n; 0\right\}$. Comme $\overline{X}_n \sim \mathcal{N}(\theta; n^{-1})$, asymptotiquement, si $\theta > 0$, $\widehat{\theta}_n = \overline{X}_n$ qui est la solution de l'équation (III.4.12) et tout se passe comme d'habitude. Par contre, si $\theta = 0$ est sur la frontière de l'espace des paramètres, $\overline{X}_n \sim \mathcal{N}(0; n^{-1})$ et $\widehat{\theta}_n = 0$ avec une probabilité 1/2, ce qui montre bien que l'on n'a ni $\widehat{\theta}_n$ comme solution de l'équation (III.4.12), ni la normalité asymptotique.

4.4 Propriétés asymptotiques des estimateurs du maximum de vraisemblance

4.4.1 Convergence

Sous des hypothèses convenables, mais un peu compliquées à énoncer et que nous ne préciserons pas ici, les estimateurs du maximum de vraisemblance sont convergents. Ceci dit, ces hypothèses ne sont pas toujours satisfaites, loin de là, et l'on peut fabriquer de nombreux contre-exemples. Le plus souvent, d'un point de vue pratique, on pourra calculer les estimateurs du maximum de vraisemblance et montrer, par un calcul direct, qu'ils sont convergents comme c'est le cas dans le modèle exponentiel si le vrai paramètre θ est intérieur à Θ.

4.4.2 Retour sur le modèle exponentiel

Considérons un modèle exponentiel naturel donné par la famille des densités $\{g_\eta, \eta \in H\}$ définies par (III.4.5) et notons $\ell_\eta = \log g_\eta = \eta T - A(\eta)$, les dérivations étant toujours effectuées par rapport à η. Alors, pour $\eta \in \overset{\circ}{H}$, ℓ_η, comme fonction de η, est $\mathcal{C}_\infty$ et, d'après (III.4.6) et (III.4.7),

$$\mathbb{E}_\eta\left[\ell'_\eta(X)\right] = 0; \qquad \mathrm{Var}_\eta\left(\ell'_\eta(X)\right) = A''(\eta) = \mathrm{Var}_\eta(T(X)) = -\mathbb{E}_\eta\left[\ell''_\eta(X)\right].$$

Il s'ensuit que l'on peut définir la fonction $I(\eta)$ par

$$I(\eta) = \int \frac{\left(g'_\eta(x)\right)^2}{g_\eta(x)}\, d\nu(x) = \mathbb{E}_\eta\left[\left(\ell'_\eta(X)\right)^2\right] = \mathrm{Var}_\eta\left(\ell'_\eta(X)\right) = -\mathbb{E}_\eta\left[\ell''_\eta(X)\right],$$

que cette fonction est continue et que, d'après (III.4.11),

$$\sqrt{n}\,(\widehat{\eta}_n - \eta) \rightsquigarrow \mathcal{N}\left(0, [I(\eta)]^{-1}\right).$$

Si l'on se place maintenant dans le cadre du modèle exponentiel général avec une famille de densités $\{f_\theta,\, \theta \in \Theta\}$, données par (III.4.8), si $Q \in \mathcal{C}_2$ et $\theta \in \overset{\circ}{\Theta}$, on peut définir de la même manière $\ell_\theta = \log f_\theta$ et

$$I(\theta) = \int \frac{\left(f'_\theta(x)\right)^2}{f_\theta(x)}\, d\mu(x). \tag{III.4.13}$$

De simples changements de variables et la méthode delta montrent que l'on a encore

$$\mathbb{E}_\theta\left[\ell'_\theta(X)\right] = 0 \qquad \text{et} \qquad I(\theta) = \mathrm{Var}_\theta\left(\ell'_\theta(X)\right) = -\mathbb{E}_\theta\left[\ell''_\theta(X)\right], \tag{III.4.14}$$

ainsi que la normalité asymptotique des estimateurs du maximum de vraisemblance :

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \rightsquigarrow \mathcal{N}\left(0, [I(\theta)]^{-1}\right). \tag{III.4.15}$$

La fonction $I(\theta)$ est dite *Information de Fisher* du modèle statistique et l'on peut conclure que, dans un modèle exponentiel avec $Q \in \mathcal{C}_2$, l'estimateur du maximum de vraisemblance (si $\theta \in \overset{\circ}{\Theta}$) est asymptotiquement normal de variance asymptotique l'inverse de l'Information de Fisher.

4.4.3 Normalité asymptotique

Considérons maintenant un modèle paramétrique dominé par μ et donné par une famille de densités $\{f_\theta,\, \theta \in \Theta \subset \mathbb{R}\}$. Sous des hypothèses convenables de dérivabilité, d'intégrabilité et de domination (on demandera en particulier que $\ell_\theta = \log f_\theta \in \mathcal{C}_2$ et que $\mathbb{E}_\theta\left[\left(\ell'_\theta(X)\right)^2\right] < +\infty$) on peut encore définir l'Information de Fisher $I(\theta)$ du modèle par (III.4.13) et vérifier que (III.4.14) est encore satisfaite. Si ces hypothèses convenables (que je ne préciserai pas ici) sur la famille de densités $\{f_\theta,\, \theta \in \Theta\}$ sont satisfaites, le modèle est dit *régulier*.

Supposons maintenant que la vraie valeur du paramètre soit $\theta_0 \in \overset{\circ}{\Theta}$, qu'une suite d'estimateurs du maximum de vraisemblance $(\widehat{\theta}_n)_{n\geq 1}$ existe et qu'elle converge vers θ_0. La formule des accroissements finis entre θ_0 et $\widehat{\theta}_n$ appliquée à la dérivée de la log-vraisemblance nous donne

$$\sum_{i=1}^n \ell'_{\widehat{\theta}_n}(X_i) - \sum_{i=1}^n \ell'_{\theta_0}(X_i) = \left(\widehat{\theta}_n - \theta_0\right) \sum_{i=1}^n \ell''_{\theta'_n}(X_i),$$

où (θ'_n) est une suite qui converge vers θ_0 en probabilité en même temps que $\widehat{\theta}_n$. Comme la log-vraisemblance $\sum_{i=1}^n \ell_\theta(X_i)$ est dérivable, sa dérivée s'annule en $\widehat{\theta}_n$, ce qui permet d'écrire

$$0 = \frac{1}{\sqrt{n}} \sum_{i=1}^n \ell'_{\widehat{\theta}_n}(X_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \ell'_{\theta_0}(X_i) + \sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \left[\frac{1}{n}\left(\sum_{i=1}^n \ell''_{\theta_0}(X_i)\right) + R_n\right],$$

avec

$$R_n = R_n(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^n \left[\ell''_{\theta'_n}(X_i) - \ell''_{\theta_0}(X_i)\right].$$

Finalement,

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) = \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n \ell'_{\theta_0}(X_i)\right] \left[-\frac{1}{n}\left(\sum_{i=1}^n \ell''_{\theta_0}(X_i)\right) - R_n\right]^{-1}.$$

Or nos hypothèses (III.4.14), le T.L.C. et la loi des grands nombres entraînent que

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \ell'_{\theta_0}(X_i) \rightsquigarrow \mathcal{N}\left(0, I(\theta_0)\right) \qquad \text{et} \qquad \frac{1}{n} \sum_{i=1}^n \ell''_{\theta_0}(X_i) \xrightarrow{P} -I(\theta_0).$$

Comme les hypothèses de régularité que l'on a mis sur le modèle entraînent que $R_n \xrightarrow{P} 0$, le Théorème de Slutsky implique que (III.4.15) est encore satisfaite. On obtient alors le théorème suivant.

Théorème 8 *Sous des hypothèses de régularité convenables, les estimateurs du maximum de vraisemblance $\widehat{\theta}_n$, s'ils convergent vers le vrai paramètre θ_0, sont asymptotiquement normaux de variance asymptotique l'inverse de l'Information de Fisher, c'est à dire que*

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \rightsquigarrow \mathcal{N}\left(0, [I(\theta_0)]^{-1}\right).$$

4.5 Estimateurs dits "à un pas"

4.5.1 Construction

Il se trouve que les estimateurs du maximum de vraisemblance sont souvent difficiles à calculer. Or, dans un modèle régulier, tout estimateur convenablement construit (estimateur empirique construit à partir d'un moment ou d'un quantile empirique par exemple) converge à la vitesse $n^{-1/2}$, il est donc proche à la fois de la vraie valeur du paramètre et de l'estimateur du maximum de vraisemblance. Une idée naturelle dans ce contexte est de chercher à le modifier pour en faire un estimateur équivalent au maximum de vraisemblance, donc asymptotiquement aussi bon.

Le principe de la construction repose sur la *méthode de Newton* que l'on utilise pour résoudre numériquement une équation de la forme $g(x) = 0$. Partant d'une approximation x_1 de la racine x_0, on fait comme si g était linéaire au voisinage de x_1, c'est à dire que l'on remplace g par sa tangente $g_1(x) = g(x_1) + (x - x_1)g'(x_1)$ et que l'on résoud l'équation $g_1(x) = 0$, d'où la solution $x_2 = x_1 - g(x_1)/g'(x_1)$. On itère la procédure qui produit ainsi une suite d'approximations $(x_n)_{n\geq 1}$ qui, sous des hypothèses convenables, converge vers x_0.

Dans notre cas, et en supposant que la log-vraisemblance $L_n(\theta) = \sum_{i=1}^n l_\theta(X_i)$ est deux fois continuement différentiable, on cherche une solution de l'équation $L'_n(t) = 0$.

Partant d'une solution approchée T_n, la méthode de Newton en fournit une nouvelle, $T_n - L'_n(T_n)/L''_n(T_n)$. On utilise alors le fait que $\mathbb{E}_\theta\left[L''_n(\theta)\right] = -nI(\theta)$, I est continue et T_n est proche de θ pour remplacer $L''_n(T_n)$ par $-nI(T_n)$. On obtient ainsi un nouvel estimateur

$$\widetilde{\theta}_n(X_1,\ldots,X_n) = T_n(X_1,\ldots,X_n) + \frac{L'_n(T_n)}{nI(T_n)} = T_n(X_1,\ldots,X_n) + \frac{\sum_{i=1}^n l'_{T_n}(X_i)}{nI(T_n)}. \tag{III.4.16}$$

La méthode classique de Newton est itérative mais ici on ne fait que le premier pas de la méthode. En effet, notre but n'est pas de trouver la racine de l'équation $L'_n(t) = 0$ (c'est à dire l'estimateur du maximum de vraisemblance $\widehat{\theta}_n$) mais d'estimer θ. Comme $\widehat{\theta}_n - \theta$ est de l'ordre de $n^{-1/2}$, il suffit de remplacer $\widehat{\theta}_n$ par une quantité de la forme $\widehat{\theta}_n + o\left(n^{-1/2}\right)$, ce qui est le cas de $\widetilde{\theta}_n$.

4.5.2 Comportement asymptotique

Les propriétés asymptotiques de l'estimateur $\widetilde{\theta}_n$ dépendent évidemment de celles de T_n et, pour que la méthode fonctionne, il faut partir d'un estimateur T_n suffisamment proche de θ_0.

Théorème 9 *Considérons un modèle statistique régulier $\{f_\theta,\, \theta \in \Theta\}$ pour lequel il existe une suite $\left(T_n(X_1,\ldots,X_n)\right)_{n\geq n_0}$ d'estimateurs telle que la suite $\sqrt{n}\,(T_n - \theta_0)$ converge en loi lorsque θ_0 désigne le vrai paramètre avec $I(\theta_0) > 0$. Soit $\widetilde{\theta}_n(X_1,\ldots,X_n)$ l'estimateur défini par (III.4.16). Alors la suite $(\widetilde{\theta}_n)_{n\geq 1}$ est convergente et vérifie*

$$\sqrt{n}\left(\widetilde{\theta}_n - \theta_0\right) \rightsquigarrow \mathcal{N}\left(0, [I(\theta_0)]^{-1}\right).$$

On pourra en particulier utiliser, comme suites d'estimateurs préliminaires, celles qui sont asymptotiquement normales. Des exemples classiques sont les estimateurs de la méthode des moments ou ceux de la méthode des quantiles. Ceci s'applique en particulier à l'estimation du paramètre d'un modèle de translation de Cauchy pour lequel l'estimateur du maximum de vraisemblance est difficile à calculer. On partira de la médiane empirique T_n et l'on construira $\widetilde{\theta}_n$ comme indiqué ci-dessus.

5 Les estimateurs de Bayes

5.1 L'approche Bayésienne

Nous avons vu que, pour évaluer la qualité d'un estimateur $\widehat{\theta}$ de θ, il est commode d'utiliser son risque quadratique qui est la fonction

$$\theta \mapsto R(\theta,\widehat{\theta}) = \mathbb{E}_\theta[(\theta - \widehat{\theta}(\boldsymbol{X}))^2] = \int_{\mathcal{X}} [\theta - \widehat{\theta}(x)]^2\, dP_\theta(x).$$

Mais si cette fonction permet d'évaluer les performances d'un estimateur, elle permet rarement de comparer deux estimateurs $\widehat{\theta}$ et $\widehat{\theta}'$ entre eux. En effet, le plus souvent, leurs fonctions de risque ne sont pas comparables parce que l'on n'a ni $R(\theta,\widehat{\theta}) \leq R(\theta,\widehat{\theta}')$ quel que soit θ, ni l'inégalité inverse. Un exemple très simple dans le modèle binomial avec $N \sim \mathcal{B}(n,\theta)$ est donné par $\widehat{\theta}_1(N) = N/n$ (maximum de vraisemblance) et $\widehat{\theta}_2(N) = 1/2$ (estimateur constant). Leurs fonctions de risque respectives sont $R(\theta,\widehat{\theta}_1) = n^{-1}\theta(1-\theta)$ et

$R(\theta, \widehat{\theta}_2) = (\theta - 1/2)^2$. Elles ne sont pas comparables et, si n est petit, la seconde sera plus petite que la première dans un voisinage non négligeable de 1/2. Cela pose clairement problème si l'on pense que la vraie valeur de θ est proche de 1/2.

5.1.1 Loi a priori et risque de Bayes

Une manière de régler le problème de la comparaison des estimateurs est alors de considérer un risque "moyen". On se donne une probabilité μ sur Θ, dite *loi a priori* et l'on intègre le risque par rapport à μ.

Définition 17 *Le risque de Bayes de l'estimateur $\widehat{\theta}$ pour la* loi a priori *μ sur Θ est donné par*

$$R_B(\mu, \widehat{\theta}) = \int_\Theta R(\theta, \widehat{\theta})\, d\mu(\theta) = \int_\Theta d\mu(\theta) \int_E [\theta - \widehat{\theta}(x)]^2\, dP_\theta(x). \tag{III.5.1}$$

On peut soit voir ce risque comme une moyenne, soit interpréter cela en disant que le paramètre est lui-même une variable aléatoire $\boldsymbol{\theta}$ de loi μ et que P_θ est la loi conditionnelle de $\boldsymbol{X}$ sachant $\boldsymbol{\theta} = \theta$:

$$P_\theta[A] = \mathbb{E}_Q[\mathbb{1}_A(\boldsymbol{X}) \,|\, \boldsymbol{\theta} = \theta],$$

l'espérance étant prise par rapport à la loi jointe Q de $\boldsymbol{\theta}$ et $\boldsymbol{X}$ définie par

$$Q(B \times C) = \mathbb{P}[\boldsymbol{\theta} \in B \ \text{ et } \ \boldsymbol{X} \in C] = \int_B P_\theta[C]\, d\mu(\theta).$$

Dans ce cadre, (III.5.1) s'écrit $R_B(\mu, \widehat{\theta}) = \mathbb{E}_Q[(\boldsymbol{\theta} - \widehat{\theta}(\boldsymbol{X}))^2]$.

5.1.2 Loi a posteriori

Dans tous les exemples que nous considérerons, les lois auront des densités, ce qui facilite les calculs. Supposons que l'on dispose d'une mesure de référence λ sur Θ, typiquement la mesure de Lebesgue, de sorte que l'on puisse écrire $\mu = \overline{g} \cdot \lambda$ et que l'on dispose de même d'une mesure de référence ν sur E pour laquelle $P_\theta = f(\cdot\,|\,\theta) \cdot \nu$ ou $dP_\theta/d\nu(x) = f(x|\,\theta)$. On vérifie alors que Q a la densité $\overline{g}(\theta) f(x|\,\theta)$ par rapport à la mesure produit $\lambda \otimes \nu$ sur $\Theta \times E$. Dans ce cas, $f(x|\,\theta)$ est la densité (de la loi) conditionnelle de $\boldsymbol{X}$ sachant que $\boldsymbol{\theta} = \theta$, ce qui justifie la notation que nous avons adoptée.

Rappel *Étant donné deux variables aléatoires $Y \in \mathcal{Y}$ et $Z \in \mathcal{Z}$ ayant une densité jointe par rapport à une mesure produit sur $\mathcal{Y} \times \mathcal{Z}$, la densité conditionnelle $f(y|\,z)$ de Y sachant Z est le quotient de la densité jointe par la densité marginale de Z.*

Ici la densité jointe est $\overline{g}(\theta) f(x|\,\theta)$ et la densité marginale de $\boldsymbol{\theta}$ est $\overline{g}$ par définition. Mais on peut également calculer la densité conditionnelle de $\boldsymbol{\theta}$ sachant que $\boldsymbol{X} = x$. La densité marginale $\overline{f}$ de X s'obtient par intégration en θ, d'où

$$\overline{f}(x) = \int_\Theta f(x|\,\theta)\overline{g}(\theta)\, d\lambda(\theta) \qquad \text{et} \qquad g(\theta|\,x) = \overline{g}(\theta) f(x|\,\theta)/\overline{f}(x), \tag{III.5.2}$$

où $g(\theta|\,x)$ désigne la densité conditionnelle de $\boldsymbol{\theta}$ sachant que $\boldsymbol{X} = x$, c'est à dire la densité par rapport à λ de la loi conditionnelle Q_x de $\boldsymbol{\theta}$ sachant que $\boldsymbol{X} = x$.

Définition 18 *La loi conditionnelle Q_x de $\boldsymbol{\theta}$ sachant que $\boldsymbol{X} = x$ de densité $g(\theta|\,x)$ par rapport à λ donnée par (III.5.2) est dite* loi a posteriori *de $\boldsymbol{\theta}$ sachant que $\boldsymbol{X} = x$.*

5.1.3 Définition des estimateurs de Bayes

Si l'on prend le risque de Bayes comme référence pour mesurer la qualité d'un estimateur $\widehat{\theta}$ à valeur dans Θ, il est clair qu'un estimateur optimal est un estimateur qui minimise le risque de Bayes. Le Théorème de Fubini et le fait que $\overline{g}(\theta)f(x|\,\theta) = g(\theta|\,x)\overline{f}(x)$ permettent alors de récrire le risque de Bayes (III.5.1) d'un estimateur $\widehat{\theta}$ de la manière suivante :

$$\begin{aligned} R_B(\mu,\widehat{\theta}) &= \int_\Theta \overline{g}(\theta)d\lambda(\theta)\int_E [\theta-\widehat{\theta}(x)]^2 f(x|\,\theta)\,d\nu(x) \\ &= \int_\Theta\int_E [\theta-\widehat{\theta}(x)]^2 g(\theta|\,x)\overline{f}(x)\,d\nu(x)d\lambda(\theta) \\ &= \int_E \overline{f}(x)\,d\nu(x)\int_\Theta [\theta-\widehat{\theta}(x)]^2 g(\theta|\,x)\,d\lambda(\theta) \\ &= \int_E \overline{f}(x)\,d\nu(x)\int_\Theta [\theta-\widehat{\theta}(x)]^2\,dQ_x(\theta). \end{aligned}$$

Pour minimiser cette quantité, il suffit de minimiser, pour tout x fixé,

$$\int_\Theta [\theta-\widehat{\theta}(x)]^2\,dQ_x(\theta) = \mathbb{E}_{Q_x}[(\boldsymbol{\theta}-\widehat{\theta}(x))^2] = \mathbb{E}_Q[(\boldsymbol{\theta}-\widehat{\theta}(x))^2 \mid \boldsymbol{X}=x],$$

c'est à dire l'espérance de $[\boldsymbol{\theta}-\widehat{\theta}(x)]^2$ pour la loi a posteriori de $\boldsymbol{\theta}$. Or on sait bien que le minimum en t de $\mathbb{E}_{Q_x}[(\boldsymbol{\theta}-t)^2]$ est obtenu pour $t=\mathbb{E}_{Q_x}[\boldsymbol{\theta}]$. On obtient ainsi le résultat suivant.

Proposition 5 *L'estimateur $\widetilde{\theta}(\boldsymbol{X})$ défini par $\widetilde{\theta}(x) = \mathbb{E}_{Q_x}[\boldsymbol{\theta}]$, où Q_x désigne la loi a posteriori de $\boldsymbol{\theta}$ quand $\boldsymbol{X}=x$, minimise le risque de Bayes pour la loi a priori μ. Il est appelé* estimateur de Bayes *pour la loi a priori μ.*

5.1.4 Calcul des estimateurs de Bayes

Pour calculer l'estimateur de Bayes, tout revient donc à évaluer la loi a posteriori de $\boldsymbol{\theta}$. Pour ce faire, une remarque sera souvent utile. On veut calculer, pour x fixé, $g(\theta\,|\,x)$ qui est une densité par rapport à λ. Or on sait par (III.5.2) que $g(\theta|\,x) = \overline{g}(\theta)f(x|\,\theta)/\overline{f}(x)$. Si l'on peut montrer que la densité jointe $\overline{g}(\theta)f(x|\,\theta)$ est proportionnelle à une certaine fonction $k_x(\theta)$, c'est à dire que $\overline{g}(\theta)f(x|\,\theta) = \alpha(x)k_x(\theta)$ où $k_x(\theta)$, considérée comme fonction de θ, est une densité de probabilité connue par rapport à λ dont les paramètres dépendent de x, alors nécessairement, puisque $g(\theta\,|\,x)$ est aussi une densité en θ, $g(\theta\,|\,x) = k_x(\theta)$, comme on le voit facilement en intégrant en θ. On procèdera le plus souvent ainsi, en montrant que $f(x|\,\theta)\overline{g}(\theta)$ est proportionnelle à une densité classique avec une constante de proportionalité qui ne dépend que de x pour trouver la loi a posteriori de $\boldsymbol{\theta}$.

Exemple : $X_1,\ldots,X_n$ sont, conditionnellement à $\boldsymbol{\theta}=\theta$, des v.a.r. i.i.d. $\mathcal{N}(\theta,1)$ et $\boldsymbol{\theta}$ est $\mathcal{N}\left(\tau,\sigma^2\right)$. La densité jointe des observations s'écrit donc

$$(2\pi)^{-n/2}\left(2\pi\sigma^2\right)^{-1/2}\exp\left[-\frac{1}{2}\sum_{i=1}^n(x_i-\theta)^2-\frac{(\theta-\tau)^2}{2\sigma^2}\right].$$

Si l'on factorise en séparant ce qui dépend de θ et ce qui n'en dépend pas, on peut récrire cela, après développement des carrés, sous la forme

$$K(x_1,\ldots,x_n)\exp\left[-\frac{1}{2\sigma^2}\left(n\sigma^2\theta^2+\theta^2-2\theta\sigma^2\sum_{i=1}^n x_i-2\theta\tau\right)\right]$$

ou encore en complétant le carré,

$$K'(x_1,\dots,x_n)\exp\left[-\frac{n\sigma^2+1}{2\sigma^2}\left(\theta-\frac{\sigma^2\sum_{i=1}^n x_i+\tau}{n\sigma^2+1}\right)^2\right].$$

On reconnaît là une fonction de θ proportionnelle à la densité gaussienne de moyenne $\left(\sigma^2\sum_{i=1}^n x_i+\tau\right)/\left(n\sigma^2+1\right)$ et de variance $\sigma^2/\left(n\sigma^2+1\right)$, ce qui implique que la loi a posteriori de $\boldsymbol{\theta}$ sachant $\boldsymbol{X}=(X_1,\dots,X_n)$ est la loi gaussienne

$$\mathcal{N}\left(\left[n\sigma^2\overline{X}_n+\tau\right]\left[n\sigma^2+1\right]^{-1},\sigma^2\left[n\sigma^2+1\right]^{-1}\right).$$

L'estimateur de Bayes s'écrit donc

$$\widetilde{\theta}_n=\left[\overline{X}_n+\tau\left(n\sigma^2\right)^{-1}\right]/\left[1+\left(n\sigma^2\right)^{-1}\right].$$

Asymptotiquement (lorsque $n\to+\infty$), on retrouve la moyenne empirique, mais lorsque n est petit, les deux estimateurs peuvent être très différents selon les valeurs des deux paramètres de la loi a priori. Si celle-ci est centrée et de très grande variance, on retrouve à nouveau un estimateur proche de la moyenne empirique. On peut aussi remarquer que cet estimateur est toujours biaisé. Sa fonction de risque s'écrit

$$\begin{aligned}R(\theta,\widetilde{\theta}_n) &= \mathbb{E}_\theta\left[\left(\frac{\overline{X}_n+\tau\left(n\sigma^2\right)^{-1}}{1+\left(n\sigma^2\right)^{-1}}-\theta\right)^2\right]\\ &= \frac{\mathrm{Var}_\theta\left(\overline{X}_n\right)}{\left(1+\left(n\sigma^2\right)^{-1}\right)^2}+\left(\frac{\theta+\tau\left(n\sigma^2\right)^{-1}}{1+\left(n\sigma^2\right)^{-1}}-\theta\right)^2\\ &= \frac{1}{\left(1+\left(n\sigma^2\right)^{-1}\right)^2}\left[\frac{1}{n}+\left(\frac{\tau-\theta}{n\sigma^2}\right)^2\right]\end{aligned}$$

et le risque de Bayes s'écrit alors, puisque $\boldsymbol{\theta}-\tau\sim\mathcal{N}\left(0,\sigma^2\right)$,

$$\begin{aligned}R_B(\mu,\widetilde{\theta}_n) &= \frac{1}{\left(1+\left(n\sigma^2\right)^{-1}\right)^2}\,\mathbb{E}\left[\frac{1}{n}+\left(\frac{\boldsymbol{\theta}-\tau}{n\sigma^2}\right)^2\right]\\ &= \frac{1}{\left(1+\left(n\sigma^2\right)^{-1}\right)^2}\,\frac{n\sigma^4+\sigma^2}{n^2\sigma^4} \;=\; \frac{1}{\left(1+\left(n\sigma^2\right)^{-1}\right)^2}\,\frac{1}{n}.\end{aligned}$$

Comme prévu, le risque de Bayes de l'estimateur de Bayes $\widetilde{\theta}_n$ est plus petit que celui de $\overline{X}_n$ qui est égal à n^{-1}.

On pourra traiter de la même manière à titre d'exercices le cas de n observations $X_1,\dots,X_n$ i.i.d. de loi (conditionnelle par rapport à θ) de Poisson $\mathcal{P}(\theta)$, $\theta>0$, en mettant sur $\boldsymbol{\theta}$ une loi a priori $\Gamma(t,\beta)$ ou le cas de $\boldsymbol{X}$ de loi conditionnelle binomiale $\mathcal{B}(n,\theta)$ avec $\boldsymbol{\theta}$ de loi a priori uniforme sur $[0;1]$.

Quelques remarques utiles

— Les estimateurs de Bayes sont généralement biaisés.

— Pour obtenir de bons estimateurs bayésiens, il convient de prendre une loi a priori qui charge bien tout Θ, c'est à dire tous les points si Θ est dénombrable et tous les ouverts non vides si Θ est une partie d'espace euclidien. Dans le cas i.i.d. et sous des hypothèses convenables de régularité, analogues à celles que l'on utilise pour les estimateurs du maximum de vraisemblance, les estimateurs de Bayes sont convergents et asymptotiquement normaux.

— Il est plus difficile d'utiliser les estimateurs de Bayes dans un cadre cas non-paramétrique parce qu'il est compliqué de construire de "bonnes" lois a priori dans ce cadre.