STT-66972

Statistique computationnelle

Notes de cours par

Jean-Claude MASSÉ

Hiver 2009

Chapitre 1

Le jackknife

Le jackknife est une méthode statistique introduite par Quenouille en 1949 pour estimer le biais d'un estimateur. On doit à John Tukey (1958) l'appellation jackknife de même qu'une extension de la méthode à l'estimation de la variance d'un estimateur.

1.1 Estimation du biais

Pour décrire le jackknife, nous supposerons que l'on observe n variables aléatoires $X_1, X_2, \ldots, X_n$ indépendantes et identiquement distribuées :

$$X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F.$$

[Pour simplifier la notation, on notera souvent un tel échantillon :

$$\underline{X} := (X_1, X_2, \ldots, X_n).]$$

Étant donné $X_i = x_i, i = 1, \ldots, n$, les valeurs observées, on s'intéresse à un estimateur $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n)$ d'un certain paramètre θ. Lorsque cet estimateur a une moyenne, son biais est défini par

$$\text{biais}(\widehat{\theta}) = E(\widehat{\theta}) - \theta.$$

Pour $i = 1, \ldots, n$, on définit les n *échantillons jackknife de taille* $n-1$ par

$$\underline{X}_{(i)} := (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n).$$

Sur ces échantillons, on peut calculer les n *réplications jackknife* de $\widehat{\theta}$:

$$\widehat{\theta}_{(i)} = \widehat{\theta}(\underline{X}_{(i)}) = \widehat{\theta}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n).$$

Quenouille estime alors le biais de $\widehat{\theta}$ par

$$b_{\text{Jack}} = (n-1)(\widehat{\theta}_{(\cdot)} - \widehat{\theta}),$$

où l'on a posé $\widehat{\theta}_{(\cdot)} = \sum_i \widehat{\theta}_{(i)}/n$. Puisque

$$E(\widehat{\theta}) = \theta + \text{biais}(\widehat{\theta}),$$

on en arrive à définir l'*estimateur jackknife corrigé pour le biais* :

$$\begin{aligned} \widehat{\theta}_{\text{Jack}} &= \widehat{\theta} - b_{\text{Jack}} \\ &= n\widehat{\theta} - (n-1)\widehat{\theta}_{(\cdot)}. \end{aligned}$$

Justification

Pour deux suites (a_n) et (b_n) de nombres réels, la notation $a_n = O(b_n)$ (lire : grand ordre de b_n) signifie qu'il existe une constante $c > 0$ telle que

$$\left|\frac{a_n}{b_n}\right| \leq c \quad \text{pour tout } n.$$

Dans le cas particulier où $b_n \to 0$ lorsque $n \to \infty$ comme dans l'exemple ci-dessous, $a_n = O(b_n)$ signifie que $a_n \to 0$ au moins aussi vite que (b_n). Par exemple les suites $-2/n^3, 100/n^4, 1/n^{3.5}$ peuvent toutes s'écrire $O\left(1/n^3\right)$, notation indiquant que ces 3 suites tendent vers 0 à une vitesse au moins aussi grande que $1/n^3$ lorsque $n \to \infty$. Remarquons que toute suite $O(1/n^\alpha)$, $\alpha > 0$, tend nécessairement vers 0 quand $n \to \infty$: cela est une conséquence du fait que $|O(1/n^\alpha)/(1/n^\alpha)| = n^\alpha |O(1/n^\alpha)| \leq c$ pour tout n et que $n^\alpha \to \infty$ lorsque $n \to \infty$.

Plusieurs estimateurs ont une moyenne qui peut s'écrire

$$\begin{aligned} E(\widehat{\theta}) &= \theta + \text{biais}(\widehat{\theta}) \\ &= \theta + \frac{a}{n} + \frac{b}{n^2} + O\left(\frac{1}{n^3}\right), \end{aligned} \tag{1.1}$$

où a, b sont des constantes pouvant dépendre de θ mais indépendantes de n. Lorsque $a \neq 0$, on dira que le biais est d'ordre $1/n$; lorsque $a = 0, b \neq 0$, le biais sera dit d'ordre $1/n^2$. Lorsque n n'est pas trop petit, un biais d'ordre $1/n^2$ est négligeable par rapport à un biais d'ordre $1/n$.

Pour un estimateur vérifiant (1.1), on a

$$\begin{aligned} E(\widehat{\theta}_{\text{Jack}}) &= E(\widehat{\theta}) - E(b_{\text{Jack}}) \\ &= \theta + \frac{a}{n} + \frac{b}{n^2} + O\left(\frac{1}{n^3}\right) - \\ &\quad (n-1)[E(\widehat{\theta}_{(\cdot)}) - E(\widehat{\theta})]. \end{aligned}$$

D'autre part, pour tout i,

$$E(\widehat{\theta}_{(i)}) = \theta + \frac{a}{n-1} + \frac{b}{(n-1)^2} + O\left(\frac{1}{(n-1)^3}\right),$$

où $\text{biais}(\widehat{\theta}_{(i)}) = E(\widehat{\theta}_{(i)}) - \theta$ est indépendant de i, donc est égal à $\text{biais}(\widehat{\theta}_{(\cdot)})$.

Donc,

$$\begin{aligned} (n-1)[E(\widehat{\theta}_{(\cdot)}) - E(\widehat{\theta})] &= (n-1)[\text{biais}(\widehat{\theta}_{(\cdot)}) - \text{biais}(\widehat{\theta})] \\ &= (n-1)[\text{biais}(\widehat{\theta}_{(i)}) - \text{biais}(\widehat{\theta})], \quad \text{quel que soit } i, \\ &= (n-1)\left[\frac{a}{n-1} + \frac{b}{(n-1)^2} + O\left(\frac{1}{(n-1)^3}\right)\right. \\ &\quad \left. -\frac{a}{n} - \frac{b}{n^2} - O\left(\frac{1}{n^3}\right)\right] \\ &= \frac{a}{n} + \frac{b(2n-1)}{n^2(n-1)} + O\left(\frac{1}{n^2}\right), \end{aligned}$$

d'où

$$\begin{aligned} E(\widehat{\theta}_{\text{Jack}}) &= \theta - \frac{b}{n(n-1)} + O\left(\frac{1}{n^2}\right) \\ &= \theta + O\left(\frac{1}{n^2}\right). \end{aligned} \tag{1.2}$$

[Vérifier ici qu'on a bien

$$(n-1)\left[O\left(\frac{1}{(n-1)^3}\right) - O\left(\frac{1}{n^3}\right)\right] = O\left(\frac{1}{n^2}\right)$$

et

$$-\frac{b}{n(n-1)} + O\left(\frac{1}{n^2}\right) = O\left(\frac{1}{n^2}\right).\Big]$$

Selon le raisonnement qui précède, si $\widehat{\theta}$ a un biais d'ordre $1/n$ ($a \neq 0$), alors $\widehat{\theta}_{\text{Jack}}$ a un biais d'ordre $1/n^2$. Lorsque la taille d'échantillon n n'est pas trop petite, $\widehat{\theta}_{\text{Jack}}$ améliore la qualité de l'estimation de θ en diminuant le biais.

1.1.1 Exemple

Prenons le cas de $\widehat{\theta} = \overline{X}$, estimateur sans biais de $\mu = E(X)$. On vérifie que

$$\widehat{\theta}_{(i)} = \frac{\sum_{j\neq i} X_j}{n-1} = \frac{n\overline{X} - X_i}{n-1}.$$

Il est clair que $\widehat{\theta}_{(\cdot)} = \overline{X}$, d'où $b_{\text{Jack}} = (n-1)(\widehat{\theta}_{(\cdot)} - \widehat{\theta}) = 0$ et $\widehat{\theta}_{\text{Jack}} = \overline{X}$. Comme la moyenne arithmétique est déjà sans biais, le jackknife ne la modifie pas. Bien sûr, le contraire serait inacceptable.

1.1.2 Exemple

Soit $\widehat{\theta} = \widehat{\sigma}^2 = \sum_i (X_i - \overline{X})^2/n$, estimateur de la variance σ^2 de la méthode des moments. On sait que $\widehat{\sigma}^2$ a un biais d'ordre $1/n$:

$$E(\widehat{\sigma}^2) = \sigma^2 - \frac{\sigma^2}{n}.$$

Puisque

$$\widehat{\theta}_{(i)} = \frac{\sum_{j\neq i}\left(X_j - \frac{n\overline{X}-X_i}{n-1}\right)^2}{n-1},$$

on vérifie que

$$\begin{aligned} b_{\text{Jack}} = (n-1)(\widehat{\theta}_{(\cdot)} - \widehat{\theta}) &= -\frac{\sum_i (X_i - \overline{X})^2}{n(n-1)} \\ &= -\frac{\widehat{\sigma}^2}{n-1} \end{aligned}$$

et ainsi

$$\widehat{\theta}_{\text{Jack}} = \widehat{\sigma}^2 - b_{\text{Jack}} = \frac{\sum_i (X_i - \overline{X})^2}{n-1},$$

estimateur usuel S^2 de la variance corrigé pour le biais.

1.1.3 Exemple

Considérons $\widehat{\theta} = \overline{X}^2$, estimateur de μ^2 selon la méthode des moments. On sait que

$$E(\overline{X}^2) = \mu^2 + \frac{\alpha_2}{n} = \mu^2 + \frac{\sigma^2}{n},$$

où $\alpha_k = E(X-\mu)^k$ (moment centré d'ordre k), donc $\overline{X}^2$ a un biais d'ordre $1/n$.

En faisant un peu de calculs, on montre que

$$\begin{aligned} b_{\text{Jack}} &= (n-1)(\widehat{\theta}_{(\cdot)} - \overline{X}^2) \\ &= \frac{1}{n(n-1)} \sum_i (X_i - \overline{X})^2 \\ &= \frac{\widehat{\alpha}_2}{n}, \end{aligned}$$

où $\widehat{\alpha}_2 = S^2 = \sum_i (X_i - \overline{X})^2/(n-1)$ (estimateur de σ^2 corrigé pour le biais de l'exemple 1.1.2). L'estimateur de μ^2 corrigé pour le biais est donc

$$\widehat{\theta}_{\text{Jack}} = \overline{X}^2 - \frac{\widehat{\alpha}_2}{n}.$$

On vérifie que cet estimateur est sans biais. De manière générale, le jackknife corrige parfaitement pour le biais tout estimateur ayant un biais exactement de la forme a/n.

Dans les trois exemples qui précèdent, il est possible d'exprimer l'estimateur jackknife $\widehat{\theta}_{\text{Jack}}$ comme fonction explicite des observations. Dans un pareil cas, il n'est donc pas nécessaire en pratique de se donner la peine de calculer les $\widehat{\theta}_{(i)}$. L'utilité du jackknife comme estimateur du biais vient plutôt de la possibilité d'en faire le calcul par un ordinateur en ignorant son expression (parfois extrêmement compliquée) comme fonction des X_i. Nous faisons suite en présentant la deuxième application importante du jackknife.

1.2 Estimation de la variance

Tukey (1958) propose d'estimer la variance d'une statistique $\widehat{\theta}$ par

$$v_{\text{Jack}} = \frac{n-1}{n} \sum_i \left(\widehat{\theta}_{(i)} - \widehat{\theta}_{(\cdot)} \right)^2 .$$

Comme pour l'estimateur de Quenouille b_{Jack}, l'estimateur v_{Jack} est défini pour n'importe quelle statistique $\widehat{\theta}$. Comme les estimateurs de la variance d'un estimateur ont rarement une forme explicite, l'introduction de l'estimateur v_{Jack} a donné une impulsion importante à l'utilisation du jackknife en statistique.

Justification heuristique

On note que

$$\begin{aligned} \widehat{\theta}_{\text{Jack}} &= \widehat{\theta} - (n-1)(\widehat{\theta}_{(\cdot)} - \widehat{\theta}) \\ &= \frac{\sum_i [\widehat{\theta} + (n-1)(\widehat{\theta} - \widehat{\theta}_{(i)})]}{n} \\ &= \frac{\sum_i \widetilde{\theta}_i}{n} \quad (\equiv \widetilde{\theta}). \end{aligned}$$

Tukey nomme les $\widetilde{\theta}_i = \widehat{\theta} + (n-1)(\widehat{\theta} - \widehat{\theta}_{(i)}) = n\widehat{\theta} - (n-1)\widehat{\theta}_{(i)}$ les *pseudo-valeurs* du jackknife. L'estimateur corrigé pour le biais $\widehat{\theta}_{\text{Jack}}$ s'exprime donc comme la moyenne $\widetilde{\theta}$ des pseudo-valeurs. Tukey conjecture qu'on peut traiter celles-ci comme des variables i.i.d (indépendantes identiquement distribuées).

On note que

$$\begin{aligned} \widetilde{\theta}_i - \widetilde{\theta} &= \widehat{\theta} + (n-1)(\widehat{\theta} - \widehat{\theta}_{(i)}) - \widehat{\theta} - (n-1)\widehat{\theta} + (n-1)\widehat{\theta}_{(\cdot)} \\ &= (n-1)(\widehat{\theta}_{(\cdot)} - \widehat{\theta}_{(i)}) \end{aligned}$$

Pour toute statistique $\widehat{\theta}$, on a donc l'expression

$$v_{\text{Jack}} = \frac{\sum_i (\widetilde{\theta}_i - \widetilde{\theta})^2}{n(n-1)}.$$

Pour un échantillon i.i.d. $X_1, \ldots, X_n$, on vérifie plus bas que l'estimateur jackknife de $\mathrm{Var}(\overline{X}) = \sigma^2/n$ est l'estimateur sans biais habituel

$$\frac{\sum_i (X_i - \overline{X})^2}{n(n-1)}.$$

En fait, Tukey remplace l'estimation de $\mathrm{Var}(\widehat{\theta})$ par celle de $\mathrm{Var}(\widehat{\theta}_{\mathrm{Jack}})$. Traitant ensuite les pseudo-valeurs $\widetilde{\theta}_i$ comme des observations i.i.d. (dont la moyenne est $\widehat{\theta}_{\mathrm{Jack}}$), il estime $\mathrm{Var}(\widehat{\theta}_{\mathrm{Jack}})$ par v_{Jack} en suivant l'exemple de l'estimation de $\mathrm{Var}(\overline{X})$ à partir des X_i.

En plus d'être applicable à toute statistique, la méthode a le grand intérêt d'être facile à mettre en pratique quand on dispose d'un ordinateur. On peut se demander si l'estimateur a une forme simple pour les exemples déjà vus.

1.2.1 Suite des exemples 1.1.1, 1.1.2 et 1.1.3

Pour tous ces exemples, nous allons voir que v_{Jack} s'exprime de façon relativement simple en termes des X_i. En effet, pour $\widehat{\theta} = \overline{X}$, on a

$$\widehat{\theta}_{(i)} - \widehat{\theta}_{(\cdot)} = \frac{\overline{X} - X_i}{n-1},$$

ce qui signifie qu'on retrouve l'estimateur sans biais de $\mathrm{Var}(\overline{X})$:

$$v_{\mathrm{Jack}} = \frac{n-1}{n} \sum_i \frac{(X_i - \overline{X})^2}{(n-1)^2} = \frac{\sum_i (X_i - \overline{X})^2}{n(n-1)}.$$

Pour $\widehat{\theta} = \sum_i (X_i - \overline{X})^2/n$, on peut montrer que

$$v_{\mathrm{Jack}} = \frac{n}{(n-1)^2}\widehat{\alpha}_4 - \frac{1}{n-1}\widehat{\alpha}_2,$$

où $\widehat{\alpha}_k = \sum_i (X_i - \overline{X})^k/(n-1), k = 2, 3, \ldots$.

Pour $\widehat{\theta} = \overline{X}^2$, on peut également montrer que

$$v_{\mathrm{Jack}} = \frac{4\overline{X}^2\widehat{\alpha}_2}{n} - \frac{4\overline{X}\widehat{\alpha}_3}{n(n-1)} + \frac{\widehat{\alpha}_4}{n(n-1)^2} - \frac{\widehat{\alpha}_2^2}{n^2(n-1)}.$$

1.3 Conditions d'applications du jackknife

Pour tout estimateur $\widehat{\theta}$ d'un paramètre θ, nous avons vu que la méthode du jackknife associe trois estimateurs :

1. b_{Jack}, estimateur de Quenouille du biais de $\widehat{\theta}$;
2. $\widehat{\theta}_{\text{Jack}}$, estimateur $\widehat{\theta}$ corrigé pour le biais ;
3. v_{Jack}, estimateur de Tukey de la variance de $\widehat{\theta}$.

Dans quelles conditions est-il justifié d'utiliser la méthode ? Pour répondre à cette question, il est nécessaire de se donner des critères pour évaluer les estimateurs. On en retiendra deux : la convergence pour b_{Jack} et v_{Jack}, et l'erreur quadratique moyenne pour $\widehat{\theta}_{\text{Jack}}$. Notons que la convergence est une propriété asymptotique, alors que l'erreur quadratique moyenne est une mesure de variabilité évaluée pour une taille fixe n.

La convergence d'une suite d'estimateurs (δ_n) vers un paramètre δ est dite *faible* (ou en probabilité) si pour tout $\epsilon > 0$:

$$\lim_{n\to\infty} P(|\delta_n - \delta| > \epsilon) = 0,$$

convergence notée $\delta_n \xrightarrow{p} \delta$.

La convergence de (δ_n) vers δ est dite *forte* (ou presque sûre) si

$$P(\lim_{n\to\infty} \delta_n = \delta) = 1,$$

convergence symbolisée par $\delta \to_{p.s.} \delta$.

En gros, la convergence faible de (δ_n) vers δ signifie qu'il est très probable que l'estimateur soit dans le voisinage du paramètre lorsque la taille d'échantillon n est assez grande, et cela, quel que soit le voisinage fixé. Pour les besoins usuels de la statistique, cette convergence est généralement considérée comme satisfaisante. On peut démontrer que la convergence forte implique la convergence faible.

On trouvera des conditions précises de convergence du jackknife dans la documentation de ce chapitre (voir par exemple Shao et Tu). Comme

ces conditions sont relativement complexes, on les énonce ici en omettant quelques détails techniques.

Grosso modo, si $E[X_1|^3 < \infty$ et si $\widehat{\theta}_n = \widehat{\theta}(X_1, \ldots, X_n)$ est une fonction suffisamment lisse des X_i, on montre que, lorsque $n \to \infty$,

$$n(b_{\text{Jack}} - \text{biais}(\widehat{\theta}_n)) \to_{p.s.} 0.$$

Ce résultat signifie que, presque sûrement (donc en probabilité), b_{Jack} et $\text{biais}(\widehat{\theta}_n)$ se rapprochent très vite l'un de l'autre lorsque la taille n augmente.

Pour être un peu plus concret, le terme 'fonction lisse' signifie ici que la fonction $\widehat{\theta}_n$ de n variables vérifie une certaine propriété de différentiabilité.

Les résultats sur la convergence des estimateurs de la variance ne s'expriment pas comme ceux du biais. Si (v_n) est une suite d'estimateurs telle que v_n estime $\text{Var}(\widehat{\theta}_n)$, on dit que (v_n) converge faiblement si

$$v_n/\text{Var}(\widehat{\theta}_n) \xrightarrow{p} 1,$$

et fortement si

$$v_n/\text{Var}(\widehat{\theta}_n) \to_{p.s.} 1.$$

Supposons que $\text{Var}(X_1) < \infty$ et que $n\text{Var}(\widehat{\theta}_n) \to c > 0$. Alors, pourvu que $\widehat{\theta}_n$ soit une fonction suffisamment lisse des observations, on peut montrer que, lorsque $n \to \infty$,

$$v_{\text{Jack}}/\text{Var}(\widehat{\theta}_n) \to_{p.s.} 1.$$

Quant à $\widehat{\theta}_{\text{Jack}}$, on retiendra que cet estimateur n'a pas nécessairement une erreur quadratique moyenne plus petite que celle de $\widehat{\theta}$. Cela signifie que, dans certains cas, la réduction du biais procurée en principe par $\widehat{\theta}_{\text{Jack}}$ n'empêche pas parfois une augmentation de la variance.

1.3.1 Exemple illustrant une limitation du jackknife

Une statistique d'ordre est un exemple de statistique ne s'exprimant pas comme fonction lisse des observations. Supposons, par exemple, que l'on

estime la médiane θ de la distribution F par la médiane échantillonnale

$$\widehat{\theta}_n = \begin{cases} X_{(m)} & \text{si } n = 2m-1, \\ (X_{(m)} + X_{(m+1)}/2 & \text{si } n = 2m, \end{cases},$$

où $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ désignent les statistique d'ordre. Lorsque F possède une densité f positive en θ, on peut montrer que la variance jackknfe de $\widehat{\theta}_n$ est exactement égale à

$$v_{\text{Jack}} = \frac{n-1}{4}\left[X_{(m+1)} - X_{(m)}\right]^2.$$

Soit (X_n) une suite de variables aléatoires réelles et soit X une autre variable aléatoire réelle. Soient F_n et F les fonctions de répartition de X_n et X, respectivement, où l'on suppose que F est continue. On rappelle que la suite (X_n) *tend en loi* (ou en distribution) vers X lorsque $n \to \infty$, si pour tout $x \in \mathbb{R}$: $\lim_{n\to\infty} F_n(x) = F(x)$. On note cette convergence $X_n \xrightarrow{d} X$.

Pour la médiane échantillonnale, on peut montrer (Efron (1982), p. 16) que $nv_{\text{Jack}} \xrightarrow{d} \left(\frac{\chi^2_2}{2}\right)^2 / (4f^2(\theta))$ et que $n\text{Var}(\widehat{\theta}_n) \to 1/(4f^2(\theta))$, où θ est la médiane de la population supposée telle que $f(\theta) > 0$. D'après une propriété de la convergence en loi, il en résulte que

$$\frac{v_{\text{Jack}}}{\text{Var}(\widehat{\theta}_n)} \xrightarrow{d} \left(\frac{\chi^2_2}{2}\right)^2$$

lorsque $n \to \infty$, où χ^2_2 est un khi-deux à 2 degrés de liberté. Puisque le membre droit est une variable aléatoire presque sûrement $\neq 1$, l'estimateur jackknife de la variance ne tend pas dans ce cas vers $\text{Var}(\widehat{\theta}_n)$ quand $n \to \infty$.

1.4 Bibliographie

- Shao, J. et Tu, D. (1995). *The Jackknife and Bootstrap*, Springer-Verlag, New York. Chapitres 1–2. Excellente référence pour la théorie. Niveau de difficulté élevé.
- Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*, SIAM, Philadelphie. Chapitres 1–3. Plutôt théorique. Contient de bons exemples pratiques.

– Efron, B. et Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, Chapman and Hall, New York. Chapitres 10–11. Excellente référence pour la théorie aussi bien que la pratique. Pas trop difficile à lire.

Chapitre 2

Le bootstrap

2.1 Introduction

Étant donné un échantillon aléatoire de taille n, le jackknife fait appel aux n sous-échantillons sans remise de taille $n-1$. Un échantillon de taille n possède exactement $2^n - 1$ sous-ensembles non-vides ; il est tentant de penser qu'on pourra approfondir notre connaissance d'un estimateur en exploitant de manière plus intensive l'information contenue dans plusieurs sous-échantillons de l'échantillon de base.

Bradley Efron a proposé un tel outil en 1979 : le bootstrap. Le grand succès de cette méthode s'explique en partie par le développement rapide de l'informatique à partir du début des années 80. Le bootstrap requiert en effet des moyens de calculs considérables : tous les sous-échantillons de taille n avec remise provenant d'un échantillon de taille n y sont exploités. Lorsque les n observations sont distinctes, il n'est pas difficile de montrer que ce nombre d'échantillons appelés échantillons bootstrap est égal à n^n.

2.2 Le principe de substitution

Supposons que l'on observe n valeurs $X_i = x_i$ de n variables aléatoires indépendantes $X_1, \ldots, X_n$ identiquement distribuées de loi F. On appelle loi empirique associée à l'échantillon $x_1, \ldots, x_n$ la loi de probabilité discrète

associant à chaque x_i la masse $1/n$ (loi uniforme). On désigne par $\widehat{F}(= \widehat{F}_n)$ la *fonction de répartition empirique* correspondante. Lorsque F est une loi univariée, on a donc $\widehat{F}(x) = \#\{x_i : x_i \leq x\}/n$ pour tout $x \in \mathbb{R}$. Lorsque les x_i sont distinctes et $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ représentent les x_i ordonnées, cela signifie que

$$\widehat{F}(x) = \begin{cases} 0 & x < x_{(1)} \\ \frac{i}{n} & x_{(i)} \leq x < x_{(i+1)} \\ 1 & x \geq x_{(n)}. \end{cases}$$

Puisque les x_i sont les valeurs des variables aléatoires X_i, une fonction de répartition empirique est elle-même aléatoire en tout point x. Il est facile de voir que

$$\#\{X_i \leq x\} \sim \text{Bin}(n, F(x)).$$

En tant que variable aléatoire, $\widehat{F}(x) = \#\{X_i \leq x\}/n$, et donc $E[\widehat{F}(x)] = F(x)$ et

$$\widehat{F}(x) \to P(X \leq x) = F(x)$$

lorsque $n \to \infty$, en vertu de la loi des grands nombres. La dernière convergence est vraie au sens fort et, à plus forte raison, au sens faible. Pour tout x, $\widehat{F}(x)$ peut donc être considérée comme un estimateur convergent sans biais de $F(x)$. En ce qui regarde la convergence, le théorème de Glivenko-Cantelli va plus loin en affirmant que, lorsque F est continue, la convergence forte est uniforme en x :

$$P\left(\lim_{n\to\infty} \sup_x |\widehat{F}_n(x) - F(x)| = 0\right) = 1.$$

Comme on n'a pas fait l'hypothèse d'un modèle pour F, l'estimation de F par $\widehat{F} \equiv \widehat{F}_n$ est un exemple d'estimation non paramétrique. Toutes les fois qu'on voudra estimer un paramètre s'exprimant comme fonction de F, $\theta = \theta(F)$, il est naturel de vouloir estimer ce paramètre en appliquant la même fonction θ à $\widehat{F}$, $\widehat{\theta} = \theta(\widehat{F})$. On donne à ce principe le nom de principe de substitution (*plug-in principle*), base d'une méthode générale d'estimation non paramétrique au cœur de la méthode du bootstrap. Nous

donnons ci-dessous quelques exemples d'application du principe de substitution. Sous-jacente à cette méthode est l'idée que l'estimateur $\theta(\widehat{F}) \approx \theta(F)$ lorsque $\widehat{F} \approx F$, pourvu que θ soit continue — par rapport à la convergence en probabilité, par exemple — par rapport à F.

Plusieurs paramètres rencontrés en statistique s'expriment par des intégrales (moyenne, variance, corrélation, etc.). Pour illustrer le principe de substitution, on donnera d'abord un sens à l'intégration par rapport à une loi discrète. La définition suivante est générale.

2.2.1 Une définition générale de l'espérance

Soit X une variable aléatoire discrète prenant les valeurs $x_1, x_2, \ldots$ et soit F sa fonction de répartition. On définit l'intégrale d'une fonction g par rapport à F comme étant

$$\int g(x)dF(x) := \sum_i g(x_i)P(X = x_i),$$

pourvu que la somme du membre droit converge absolument.

Soit X une variable aléatoire possédant une densité f et une fonction de répartition F. On définit l'intégrale d'une fonction g par rapport à F comme étant

$$\int g(x)dF(x) := \int g(x)f(x)dx,$$

pourvu que l'intégrale de $|g(x)|$ par rapport à $f(x)dx$ soit finie.

Les deux définitions précédentes permettent de représenter $E[g(X)]$ par l'intégrale $\int g(x)dF(x)$, tant dans le cas discret que dans le cas continu. Nous nous servons souvent de cette notation dans la suite.

2.2.2 Trois exemples d'applications du principe de substitution

La moyenne d'une variable X de loi F (quand elle existe) peut s'écrire

$$E_F(X) = \theta(F) = \int x dF(x).$$

En appliquant le principe de substitution, on estime la moyenne par

$$\theta(\widehat{F}) = \int x d\widehat{F}(x). \tag{2.1}$$

Selon la définition 2.2.1 appliquée à $H = \widehat{F}$, la dernière expression vaut

$$\theta(\widehat{F}) = \sum_i X_i \cdot \frac{1}{n} = \overline{X},$$

et l'on retrouve ainsi la moyenne échantillonnale.

La variance de X de loi F (si elle existe) peut s'écrire

$$\theta(F) = \int [x - \int y dF(y)]^2 dF(x).$$

Appliquant le principe de substitution, on l'estimera par

$$\theta(\widehat{F}) \quad = \quad \sum_i (X_i - \overline{X})^2 \cdot \frac{1}{n},$$

estimateur dans lequel on reconnaît la forme biaisée de la variance échantillonnale.

Une extension naturelle de cette idée s'obtient en considérant un vecteur $X = (Y, Z)$ de variables positives de loi bivariée F. Un paramètre important dans certaines applications est le rapport des moyennes marginales

$$\theta(F) = \frac{E_F(Y)}{E_F(Z)} = \frac{\int y dF_y(y)}{\int z dF_z(z)},$$

où F_y, F_z désignent les lois marginales. En appliquant le principe de substitution, on estimera ce rapport par l'*estimateur ratio*

$$\theta(\widehat{F}) = \frac{\overline{Y}}{\overline{Z}}.$$

□

L'estimateur $\widehat{\theta}$ obtenu en appliquant le principe de substitution s'appelle l'*estimateur bootstrap idéal* de $\theta(F)$. Dans les 3 exemples qui précèdent, cet estimateur a une forme simple, facile à calculer. On ne peut cependant en

dire autant de la très grande majorité des paramètres de la forme $\theta(F)$ rencontrés en statistique. Nous allons voir qu'il est en effet généralement impossible de calculer la valeur exacte de l'estimateur bootstrap idéal. Dans la pratique, on visera plutôt à obtenir une approximation de l'estimateur bootstrap idéal par une simulation appropriée. Nous examinons d'abord le problème d'estimation d'une variance.

2.3 L'estimation de la variance par le bootstrap

Le bootstrap a d'abord été appliqué à un problème fondamental de la statistique appliquée : l'estimation de la variance d'un estimateur.

Suposons que

$$X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F,$$

et soit $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n)$ un estimateur de variance finie $\text{Var}_F(\widehat{\theta})$. Posant $\underline{X} = (X_1, \ldots, X_n)$, on peut écrire

$$\begin{aligned} \text{Var}_F(\widehat{\theta}) &= E_F[(\widehat{\theta}(\underline{X}) - E_F(\widehat{\theta}))^2] \\ &= \int [\widehat{\theta}(\underline{x}) - \int \widehat{\theta}(\underline{y}) dF(y_1) \cdots dF(y_n)]^2 dF(x_1) \cdots dF(x_n) \\ &= \theta(F) \end{aligned}$$

du fait que les X_i sont indépendantes. On appelle *estimateur bootstrap* de $\text{Var}_F(\widehat{\theta})$ l'estimateur

$$v_{\text{Boot}} := \text{Var}_{\widehat{F}}(\widehat{\theta})$$

obtenu du principe de substitution. [Pour une raison qui sera expliquée dans la suite, cet estimateur sera également noté $v^*(\widehat{\theta}^*)$.]

2.3.1 Exemple

Soit $\widehat{\theta} = \overline{X}$ pour des X_i telles que $\text{Var}(X_i) = \sigma^2$. Alors on sait que

$$\text{Var}_F(\overline{X}) = \theta(F) = \frac{\sigma^2}{n} = \frac{1}{n} \int [x - \int y dF(y)]^2 dF(x).$$

L'estimateur bootstrap de la variance de $\overline{X}$ est alors

$$v_{\text{Boot}} \equiv v^*(\overline{X}^*) \equiv \text{Var}_{\widehat{F}}(\overline{X}) = \frac{1}{n^2}\sum_i (X_i - \overline{X})^2, \tag{2.2}$$

estimateur biaisé de σ^2/n facile à calculer. Dans ce cas, on peut noter que

$$v_{\text{Boot}} = \frac{n-1}{n} v_{\text{Jack}}.$$

□

Dans le cas général, rappelons qu'un échantillon bootstrap est un échantillon aléatoire de taille n tiré *avec remise* de l'échantillon initial $X_1, \ldots, X_n$. Désignons par $X_1^*, \ldots, X_n^*$ les observations définissant un échantillon bootstrap. Par définition, on a $X_1^*, X_2^*, \ldots, X_n^* \overset{\text{iid}}{\sim} \widehat{F}$, autrement dit

$$P(X_i^* = x_j) = 1/n, \quad 1 \le i, j \le n,$$

lorsque $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$ est l'échantillon initial observé. L'estimateur bootstrap de la variance peut donc s'écrire

$$\begin{aligned} v_{\text{Boot}} &= \text{Var}_{\widehat{F}}(\widehat{\theta}) \\ &= E_{\widehat{F}}[\widehat{\theta}(X_1^*, \ldots, X_n^*) - E_{\widehat{F}}(\widehat{\theta}(X_1^*, \ldots, X_n^*)]^2 \\ &= \text{Var}(\widehat{\theta}(X_1^*, \ldots, X_n^*) | X_1, \ldots, X_n). \end{aligned}$$

La dernière notation montre que l'estimateur bootstrap de la variance est une variance conditionnelle : étant donné $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$, on évalue l'espérance définissant v_{Boot} en faisant varier $(X_1^*, X_2^*, \ldots, X_n^*)$ dans l'ensemble fini de tous les échantillons bootstrap possibles. Comme la loi conditionnelle de $(X_1^*, X_2^*, \ldots, X_n^*)$ est discrète, il s'ensuit que v_{Boot} a toujours la forme d'une somme finie.

Combien y a-t-il d'échantillons bootstrap ? Au sens strict, il y en a bien sûr n^n. Pour le calcul de v_{Boot}, nous donnerons cependant ici un sens légèrement différent au terme 'échantillon bootstrap'. Les estimateurs qu'on considère ne dépendent que de $\widehat{F}$, laquelle est indépendante de l'ordre des

X_i. Il est donc naturel d'identifier à un même échantillon bootstrap tous les échantillons ne différant que par une permutation des x_i. Par exemple, si $n = 4$, (x_1, x_1, x_2, x_4) est dans ce sens le même échantillon bootstrap que (x_1, x_4, x_2, x_1), et $\widehat{\theta}$ est le même appliqué à (x_1, x_1, x_2, x_4) ou à (x_1, x_4, x_2, x_1). Lorsque $n = 2$ et $x_1 \neq x_2$ les seuls échantillons bootstrap ainsi définis sont au nombre de 3 : $(x_1, x_2), (x_1, x_1)$ et (x_2, x_2) ; pour $n = 3$ et les x_i distincts, on vérifie en exercice qu'il y en a 10. En général, lorsque les $x_1, \ldots, x_n$ sont distincts, il suit d'un résultat bien connu d'analyse combinatoire que le nombre d'échantillons bootstrap dans ce sens restreint est égal à

$$m = \binom{2n-1}{n}.$$

Le tableau 2.1 présente trois exemples de valeurs de m.

TAB. 2.1 – Quelques nombres d'échantillons bootstrap

n	4	10	15
m	35	92 378	77 558 760

Pour $i = 1, \ldots, n$, posons $A_i = \#\{X_j^* = x_i, 1 \leq j \leq n\}$. Selon le raisonnement du paragraphe précédent, tout échantillon bootstrap est défini par une valeur du vecteur $(A_1, \ldots, A_n)$. Il n'est pas difficile de voir que l'échantillon bootstrap tel que $A_1 = a_1, \ldots, A_n = a_n$ a la probabilité multinomiale :

$$P(A_1 = a_1, \cdots, A_n = a_n) = \frac{n!}{a_1! \cdots a_n!}(1/n)^{a_1} \cdots (1/n)^{a_n},$$

pour $a_1 + \cdots + a_n = n$.

Puisque le nombre d'échantillons bootstrap est fini, l'estimateur bootstrap de la variance de $\widehat{\theta}$ s'écrit

$$v_{\text{Boot}} = \sum_1^m [\widehat{\theta}(z_j) - \widehat{\theta}(\cdot)]^2 w_j = \sum_1^m \widehat{\theta}(z_j)^2 w_j - \widehat{\theta}(\cdot)^2,$$

où z_j est le j^e échantillon bootstrap, w_j est sa probabilité et $\widehat{\theta}(\cdot) = E_{\widehat{F}}(\widehat{\theta}) = \sum_1^m \widehat{\theta}(z_j) w_j$.

Pour la plupart des tailles n rencontrées dans la pratique, le tableau 2.1 montre clairement que m est trop grand pour que le calcul exact de v_{Boot} soit réalisable, même avec un ordinateur. Presque toujours, on devra donc se contenter d'une approximation de v_{Boot} obtenue par simulation. Pour ce faire, on se basera sur la loi des grands nombres énoncée maintenant dans la généralité dont on a besoin.

Loi forte des grands nombres

Soit $X_1, X_2, \ldots$ une suite de vecteurs aléatoires indépendants identiquement distribués et soit g une fonction à valeur réelle telle que $E[g(X_1)]$ existe. Alors

$$\frac{\sum_1^n g(X_i)}{n} \rightarrow_{p.s.} E[g(X_1)].$$

□

Selon la loi des grands nombres, pour calculer une valeur approchée de $E[g(X_1)]$, il suffit de simuler un nombre suffisamment grand de valeurs $X_1, X_2, \ldots, X_n$, puis de calculer $\sum_1^n g(X_i)/n$. De la même façon, pour approcher

$$\text{Var}(g(X)) = E[g(X)^2] - (E[g(X)])^2,$$

il suffira de calculer

$$\frac{\sum_1^n g(X_i)^2}{n} - \left(\frac{\sum_1^n g(X_i)}{n}\right)^2 = \frac{\sum_1^n (g(X_i) - \sum_1^n g(X_i)/n)^2}{n}$$

pour n suffisamment grand. Cette application de la loi des grands nombres est un exemple d'application de la méthode de Monte Carlo au calcul d'une variance.

Pour le calcul approché de $v_{\text{Boot}} = \text{Var}_{\widehat{F}}(\widehat{\theta})$, il suffit donc :

– de tirer un échantillon bootstrap $X_1^*, X_2^*, \ldots, X_n^* \overset{\text{iid}}{\sim} \widehat{F}$, puis de calculer $\widehat{\theta}^* = \widehat{\theta}(X_1^*, X_2^*, \ldots, X_n^*)$ (réplication bootstrap de l'estimateur) ;

– de répéter la première étape un nombre suffisamment grand de fois B pour obtenir les réplications $\widehat{\theta}^*_1, \ldots, \widehat{\theta}^*_B$;
– de calculer

$$v^{(B)}_{\text{Boot}} = \frac{1}{B}\sum_1^B (\widehat{\theta}^*_b - \widehat{\theta}^*(\cdot))^2 = \frac{1}{B}\sum_1^B (\widehat{\theta}^*_b)^2 - (\widehat{\theta}^*(\cdot))^2,$$

où $\widehat{\theta}^*(\cdot) = \sum_b \widehat{\theta}^*_b/B$.

En vertu de la loi des grands nombres, $v^{(B)}_{\text{Boot}} \approx v_{\text{Boot}}$, dès que B est assez grand. En pratique, on donne le nom d'estimateur bootstrap de la variance à $v^{(B)}_{\text{Boot}}$ aussi bien qu'à l'estimateur idéal v_{Boot}. Le premier de ces estimateurs est communément utilisé dans les applications, tandis que le second a surtout un intérêt théorique.

Quelle valeur de B prendre ?

Selon Efron, il est rarement nécessaire d'utiliser plus de $B = 200$ échantillons bootstrap pour estimer une variance ; dans bien des cas, $B = 50$ ou 100 sont suffisants. L'importance des valeurs extrêmes de la statistique $\widehat{\theta}$ étudiée est un facteur important dans la détermination du choix de B : plus ces valeurs sont fréquentes, plus B devrait être grand. On notera cependant que certaines autres applications du bootstrap exigent un B beaucoup plus grand ; ce sera en particulier le cas pour l'application à la construction d'intervalles de confiance.

2.4 Extension à des problèmes plus généraux

La méthode du bootstrap est applicable à d'autres problèmes d'estimation que celui de la variance. Supposons que l'on s'intéresse aux propriétés d'une certaine variable aléatoire $R = R(\underline{X}, F)$, où

$$\underline{X} = (X_1, X_2, \ldots, X_n) \overset{\text{iid}}{\sim} F.$$

Par exemple, R pourrait être $\widehat{\theta} = \widehat{\theta}(\underline{X})$ comme ci-dessus, ou encore $\sqrt{n}(\overline{X} - \mu(F))$. Après avoir observé $\underline{X} = \underline{x}$, on pourra vouloir estimer $E_F(R), P_F(R > 1)$, etc. Pour chaque échantillon bootstrap $\underline{X}^*$, on calcule alors

$$R^* = R(\underline{X}^*, \widehat{F}),$$

puis on évalue les estimateurs bootstrap $E_{\widehat{F}}(R^*), P_{\widehat{F}}(R^* > 1)$, etc. Comme pour l'estimation d'une variance, le nombre d'échantillons bootstrap possible est généralement trop grand pour pouvoir calculer la valeur exacte de ces estimateurs, de sorte qu'on engendrera B échantillons bootstrap pour calculer les approximations

$$E^{(B)}_{\widehat{F}}(R^*) = \frac{1}{B}\sum_1^B R^*_b$$

et

$$\frac{1}{B}\#\{R^*_b > 1; b = 1, \ldots, B\},$$

où $R^*_b = R(\underline{X}^*_b, \widehat{F}), b = 1, \ldots, B$, sont les réplications bootstrap de R.

L'estimation du biais d'un estimateur $\widehat{\theta}$ fournit un exemple intéressant d'application de cette extension de la méthode du bootstrap. Dans ce cas, on estime

$$\text{biais}(\widehat{\theta}) = E_F(\widehat{\theta}) - \theta(F)$$

par

$$\begin{aligned} b_{\text{Boot}} &= E_{\widehat{F}}(\widehat{\theta}^*) - \theta(\widehat{F}) \\ &= E[\widehat{\theta}(\underline{X}^*)|\underline{X}] - \widehat{\theta}. \end{aligned}$$

Encore une fois, le plus souvent, on doit se contenter d'approcher cette valeur en engendrant B échantillons bootstrap $\underline{X}^*_1, \ldots, \underline{X}^*_B$, puis en calculant

$$b^{(B)}_{\text{Boot}} = \widehat{\theta}^*(\cdot) - \widehat{\theta} = \frac{1}{B}\sum_1^B \widehat{\theta}^*_b - \widehat{\theta}$$

Plus généralement, le bootstrap peut servir à estimer la distribution d'une variable aléatoire $R(\underline{X}, F)$, autrement dit à l'estimer

$$H_n(x) = P_F(R(\underline{X}, F) \leq x), \quad x \in \mathbb{R}.$$

L'estimateur bootstrap de la fonction de répartition au point x est donné par

$$H_{\text{Boot}}(x) = P_{\widehat{F}}(R(\underline{X}^*, \widehat{F}) \leq x), \quad x \in \mathbb{R},$$

expression approchée, le plus souvent, en tirant B échantillons bootstrap pour calculer

$$\frac{1}{B}\#\{R(\underline{X}_b^*, \widehat{F}) \leq x; b = 1, \ldots, B\}.$$

Dans le cas particulier de la distribution d'un estimateur $\widehat{\theta} = \widehat{\theta}(\underline{X})$, on estimera ainsi $P(\widehat{\theta}(\underline{X}) \leq x)$ par

$$\frac{1}{B}\#\{\widehat{\theta}(\underline{X}_b^*) \leq x; b = 1, \ldots, B\}.$$

2.5 Comportement asymptotique du bootstrap

Tout estimateur raisonnable doit converger vers le paramètre estimé lorsque la taille d'échantillon tend vers ∞. Dans le cas de l'estimateur bootstrap de la loi d'une variable aléatoire $R(\underline{X}, F)$, on souhaite par exemple que :

$$\begin{aligned} 0 &= \lim_{n\to\infty} \sup_x \left| P_{\widehat{F}}(R(\underline{X}^*, \widehat{F}) \leq x) - P(R(\underline{X}, F) \leq x) \right| \qquad (2.3) \\ &= \lim_{n\to\infty} \sup_x \left| H_{\text{Boot}}(x) - H_n(x) \right|. \end{aligned}$$

L'équation (2.3) exprime que la loi de l'estimateur bootstrap $R(\underline{X}^*, \widehat{F})$ est proche de celle de $R(\underline{X}, F)$ quand n est assez grand. (Comme $P_{\widehat{F}}(R(\underline{X}^*, \widehat{F}) \leq x)$ est une variable aléatoire, on sous-entend ici que la convergence en (2.3) a lieu au sens fort (presque sûre) ou au sens faible (en probabilité).)

Pour un grand nombre de variables aléatoires rencontrées en statistique, on peut montrer que la convergence en (2.3) a lieu faiblement. Citons les exemples suivants :

1. La variable $R(\underline{X}, F) = \sqrt{n}(\overline{X} - \mu)$, où l'on suppose que $\mu = E_F(X)$ et $\sigma^2 = \mathrm{Var}_F(X)$ existent. L'extension au cas multivarié est également vraie. L'estimateur bootstrap de $R(\underline{X}, F)$ est alors $R(\underline{X}^*, \widehat{F}) = \sqrt{n}(\overline{X}^* - \overline{X})$. On sait que le théorème limite central implique que $\sqrt{n}(\overline{X} - \mu)$ tend vers en loi vers la loi $N(0, \sigma^2)$ lorsque $n \to \infty$. Pour cet exemple, cette loi est donc la loi limite de l'estimateur bootstrap.
2. La variable $R(\underline{X}, F) = \sqrt{n}(\tilde{\theta} - \theta)$, où $\theta = F^{-1}(1/2)$ et $\tilde{\theta}$ sont respectivement la médiane de la loi et la médiane échantillonnale ; l'estimateur bootstrap est alors $\sqrt{n}(\tilde{\theta}^* - \tilde{\theta})$. Pour que la convergence faible au sens ci-dessus ait lieu, il suffit par exemple que F possède une densité f telle que $f(\theta) > 0$. Dans ce cas, on montre en statistique mathématique que $\sqrt{n}(\tilde{\theta} - \theta)$ converge en loi vers la loi $N(0, 1/(4f^2(\theta))$, loi limite donc de l'estimateurbootstrap. Ce résultat peut être étendu aux quantiles $F^{-1}(q)$, $0 < q < 1$, et à leurs estimateurs $\widehat{\theta} = \widehat{F}^{-1}(q)$. Rappelons que $F^{-1}(q) = \inf\{x : F(x) \geq q\}$ et que $\widehat{F}^{-1}(q)$ est défini de la même façon.
3. Les variables de la forme $\sqrt{n}[g(\overline{X}) - g(\mu)]$, lorsque $\mathrm{Var}(X) < \infty$ et g a une dérivée continue non nulle en $\mu = E(X)$. D'après la méthode delta, $\sqrt{n}[g(\overline{X}) - g(\mu)]$ converge alors en loi vers une loi normale. Sous ces mêmes conditions, la loi limite de l'estimateur bootstrap $\sqrt{n}[g(\overline{X}^*) - g(\overline{X})]$ est donc celle de $\sqrt{n}[g(\overline{X}) - g(\mu)]$ (au sens fort, donc faible aussi). Par exemple, ce résultat s'appliquera à $g(\overline{X}) = \overline{X}^2$ ou $g(\overline{X}) = 1/\overline{X}$ lorsque $\mu \neq 0$.
4. Sous des conditions peu restrictives, on peut énoncer un résultat semblable au précédent pour les moyennes tronquées.

En règle générale, la convergence en loi n'entraîne pas la convergence des moments : $Y_n \xrightarrow{d} Y$ n'entraîne pas que $E(Y_n^k) \to E(Y^k)$. Par exemple, lorsque les lois de l'estimateur bootstrap $\widehat{\theta}^*$ et de $\widehat{\theta}$ sont de plus en plus proches lorsque $n \to \infty$, il ne s'ensuit pas nécessairement que $v_{\mathrm{Boot}}/\mathrm{Var}(\widehat{\theta}) \to 1$. On a cependant la convergence $v_{\mathrm{Boot}}/\mathrm{Var}(\widehat{\theta}) \to 1$ au sens fort dans les cas suivants :

1. $\widehat{\theta} = g(\overline{X})$ (fonction de la moyenne échantillonnale), avec $\mathrm{Var}(X) < \infty$ et g ayant une dérivée non nulle continue autour de $\mu = E(X)$.
2. $\widehat{\theta} = \widehat{F}^{-1}(q)$, un quantile échantillonnal tel la médiane ou les quartiles, pourvu que F ait une densité f telle que $f(F^{-1}(q)) > 0$ et que $E|X|^c < \infty$ existe pour un certain $c > 0$. Il s'agit donc d'un cas où l'estimateur bootstrap est supérieur à l'estimateur jackknife.
3. $\widehat{\theta}$ est une moyenne tronquée.

2.5.1 Exemple de limitation du bootstrap

Supposons que $X_1, \ldots, X_n \overset{\text{iid}}{\sim}$ pour la loi uniforme sur $(0, \theta)$. L'estimateur du maximum de vraisemblance est alors $\widehat{\theta} = X_{(n)} = \max X_i$. Si $\underline{X}^*$ est un échantillon bootstrap et $\widehat{\theta}^* = \widehat{\theta}(\underline{X}^*) = \max_i X_i^*$, on a alors

$$P_{\widehat{F}}(\widehat{\theta}^* = X_{(n)}) = 1 - P(\max_i X_i^* < X_{(n)}) = 1 - (1 - 1/n)^n \to 1 - e^{-1} \approx .632$$

lorsque $n \to \infty$. Par ailleurs, $\widehat{\theta} = X_{(n)}$ est une variable continue de densité

$$f(x) = n\frac{x^{n-1}}{\theta^n}, \quad 0 < x < \theta.$$

Puisque la loi de $\widehat{\theta} = X_{(n)}$ est continue, la loi de $\widehat{\theta}^*$ et la loi de $\widehat{\theta}$ ne tendent pas à se confondre lorsque $n \to \infty$. On peut le démontrer rigoureusement comme suit.

Posons $H_n(x) = P(n(\theta - X_{(n)}) \leq x)$. Alors

$$\begin{aligned} H_n(x) &= P(X_{(n)} \geq \theta - n^{-1}x) \\ &= 1 - \big(P(X < \theta - n^{-1}x)\big)^n \\ &= 1 - \Big(1 - \frac{x}{n\theta}\Big)^n \\ &\to 1 - e^{-x/\theta}, \quad n \to \infty, \end{aligned}$$

ce qui signifie que $n(\theta - X_{(n)})$ se comporte asymptotiquement comme une variable exponentielle de paramètre $1/\theta$.

Posons maintenant $H_{\text{Boot}}(x) = P_{\widehat{F}}(n(X_{(n)} - X^*_{(n)}) \leq x)$. Alors H_n et H_{Boot} ne tendent pas à se rapprocher lorsque $n \to \infty$ du fait que pour n assez grand

$$|H_n(0) - H_{\text{Boot}}(0)] \approx 1 - e^{-1}$$

et qu'ainsi on ne peut avoir la convergence uniforme en x.

Pour k fixe indépendant de n, on peut montrer que n'importe quelle statistique d'ordre $X_{(k)}$ souffre du même problème. Cependant, lorsque $k = k_n = \lfloor pn \rfloor$ avec $0 < p < 1$ fixe, $X_{(k)}$ est un quantile échantillonnal : médiane, quartile, décile, etc. ; dans ce dernier cas, le bootstrap converge (contrairement au jackknife !).

Voici deux exemples de cas où l'estimateur bootstrap de la loi d'un estimateur $\widehat{\theta}$ peut ne pas converger :

– la variance de $\widehat{\theta}$ n'existe pas (valeurs extrêmes très probables) ;

– $\widehat{\theta}$ n'est pas une fonction suffisamment lisse des observations.

Il n'est pas fréquent de rencontrer une situation où $v_{\text{Boot}}/\text{Var}(\widehat{\theta}_n) \nrightarrow 1$. Soulignons à ce propos que, contrairement au jackknife, $v_{\text{Boot}}/\text{Var}(\widehat{\theta}_n) \to 1$ lorsque $\widehat{\theta}_n$ est la médiane échantillonnale estimant θ, pourvu seulement que $f(\theta) > 0$ et que $E|X_1|^c < \infty$ pour un certain $c > 0$.

Tel que présentée jusqu'ici dans ce chapitre, la méthode du bootstrap est applicable sans faire d'hypothèse sur la loi F des observations i.i.d. On parle dans ce cas de *bootstrap non paramétrique*. La section suivante montre comment le bootstrap peut être appliqué aux situations où la loi F est modélisée à partir d'une famille paramétrique.

2.6 Le bootstrap paramétrique

Supposons que la loi F des observations appartienne à une famille paramétrique (F_ψ), où ψ est un paramètre inconnu. Par exemple, F_ψ pourra être le modèle $N(\mu, \sigma^2)$, avec $\psi = (\mu, \sigma^2)$. Étant donné un estimateur $\widehat{\psi}$ de ψ, on peut considérer le modèle $F_{\widehat{\psi}}$ comme étant un estimateur du vrai

modèle F_ψ. Pour notre exemple, $\widehat{\psi}$ pourra être l'estimateur du maximum de vraisemblance, ce qui veut dire que $F_{\widehat{\psi}} = N(\overline{x}, \widehat{\sigma}^2)$. Dans la situation où l'on souhaite estimer une caractéristique d'un estimateur $\widehat{\theta} = \widehat{\theta}(\underline{X})$ (par exemple la variance) où $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F_\psi$, on pourra tirer un échantilllon aléatoire $X_1^*, \ldots, X_n^*$ à partir de la loi $F_{\widehat{\psi}}$, puis on calculera l'estimateur sur cet échantillon : $\widehat{\theta}^* = \widehat{\theta}(X_1^*, \ldots, X_n^*)$. En répétant cette démarche B fois pour obtenir B réplications $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_B^*$, on obtient un estimateur de la variance en prenant

$$v_{\text{Boot}}^{(B)} = \frac{1}{B} \sum_1^B (\widehat{\theta}_i^* - \overline{\widehat{\theta}^*})^2,$$

où $\overline{\widehat{\theta}^*} = \sum_1^B \widehat{\theta}_i^*/B$. On pourra voir cet estimateur comme étant une approximation de l'estimateur exact ou idéal $v_{\text{Boot}} = \text{Var}_{F_{\widehat{\psi}}}(\theta^*)$.

On donne encore à $X_1^*, \ldots, X_n^* \overset{\text{iid}}{\sim} F_{\widehat{\psi}}$ le nom d'échantillon bootstrap. Notons qu'il peut arriver que $F_{\widehat{\psi}}$ soit une loi continue, alors que, plus haut, le rééchantillonnage se faisait toujours par rapport à la loi discrète $\widehat{F}$. Le bootstrap appliqué dans le cadre d'un rééchantillonnage par rapport à une loi paramétrique s'appelle le *bootstrap paramétrique*. Le dernier estimateur $v_{\text{Boot}}^{(B)}$ porte le nom d'estimateur bootstrap (paramétrique) de $\text{Var}(\widehat{\theta})$. Naturellement, on pourra estimer de la même façon d'autres caractéristiques de l'estimateur (biais, coefficients d'asymétrie ou d'aplatissement, etc.).

Le bootstrap paramétrique est fiable pourvu que le modèle soit adéquat. Dans ce cas, le bootstrap paramétrique permet d'estimer la variance avec une précision aussi bonne, sinon meilleure, que l'estimateur bootstrap non paramétrique ou d'autres estimateurs suggérés par la théorie asymptotique et la loi normale (méthode delta, méthode de la fonction d'influence).

2.6.1 Exemple

Soit $(X_1, Y_1), \ldots, (X_n, Y_n)$ un échantillon aléatoire de la loi normale bivariée de paramètres $\mu_1 = E(X), \mu_2 = E(Y)$ et $\sigma_1^2 = \text{Var}(X), \sigma_2^2 = \text{Var}(Y)$ et $\rho = \text{Cov}(X, Y)/\sqrt{\text{Var}(X)\text{Var}(Y)}$. Alors la méthode du maximum de vrai-

semblance conduit à estimer F par la loi normale bivariée $\widehat{F}_{\text{Norm}}$ de paramètres $\widehat{\mu}_1 = \overline{X}$, $\widehat{\mu}_2 = \overline{Y}$, $\widehat{\sigma}_1^2 = \sum_i (X_i - \overline{X})^2/n$, $\widehat{\sigma}_2^2 = \sum_i (Y_i - \overline{Y})^2/n$ et $\widehat{\text{Cov}}(X,Y) = \widehat{\rho}\widehat{\sigma}_1\widehat{\sigma}_2$, où

$$\widehat{\rho} = \frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_i (X_i - \overline{X})^2 \sum_i (Y_i - \overline{Y})^2}}.$$

En statistique mathématique, on montre que la variance de $\widehat{\rho}$ est approximativement donnée par $(1-\rho^2)^2/n$, expression estimée par $(1-\widehat{\rho}^2)^2/n$. Pour cet exemple, il est donc possible d'évaluer la précision de l'estimation de $\text{Var}(\widehat{\rho})$ par la méthode du bootstrap paramétrique, en comparant v_{Boot} à $(1-\widehat{\rho}^2)^2/n$. L'estimation bootstrap se fera à partir de B réplications bootstrap de $\widehat{\rho}$ obtenues de n observations indépendantes $((X_1, Y_1), \ldots, (X_n, , Y_n))$ provenant de la loi normale bivariée de paramètres $\widehat{\mu}_1$, $\widehat{\mu}_2$, $\widehat{\sigma}_1^2$, $\widehat{\sigma}_2^2$ et $\widehat{\rho}$.

2.7 Bibliographie

- Shao, J. et Tu, D. (1995). *The Jackknife and Bootstrap*, Springer-Verlag, New York. Chapitres 1, 3.
- Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*, SIAM, Philadelphie. Chapitre 5.
- Efron, B. et Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, Chapman and Hall, New York. Chapitres 6, 11.
- Davison, A.C. et Hinkley, D.V. (1997). *Bootstrap Methods and their Applications*, Cambridge University Press, Cambridge. Ouvrage de haut niveau à caractère appliqué. Chapitre 2.

Chapitre 3

Intervalles de confiance basés sur le bootstrap

Un intervalle de confiance (IC) est un outil permettant d'estimer un paramètre et de quantifier ce faisant notre degré de certitude. Étant donné $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$, on rappelle qu'un IC de $(1-2\alpha)\times 100\%$ pour le paramètre $\theta = \theta(F)$ est un intervalle aléatoire $I_\alpha = [T_1, T_2]$ de valeurs possibles de θ tel que T_1 et T_2 sont des statistiques et

$$P(\theta \in I_\alpha) = 1 - 2\alpha.$$

La statistique mathématique propose quelques solutions au problème de construction d'un IC dans le contexte paramétrique. L'exemple classique est celui d'un IC de Student pour la moyenne μ construit à partir d'un échantillon de taille n de la loi $N(\mu, \sigma^2)$:

$$[\overline{X} - t_{n-1}(1-\alpha)\widehat{\sigma}/\sqrt{n-1}, \overline{X} - t_{n-1}(\alpha)\widehat{\sigma}/\sqrt{n-1}],$$

où $\widehat{\sigma}^2 = \sum_1^n (X_i - \overline{X})^2/n$ et $t_{n-1}(\alpha)$, $t_{n-1}(1-\alpha)$ sont les quantiles d'ordres α et $1-\alpha$ de la loi de Student à $n-1$ degrés de liberté. L'intervalle de Student est l'exemple classique d'un intervalle symétrique par rapport à l'estimateur ponctuel $\widehat{\theta} = \overline{X}$, autrement dit un intervalle de la forme

$$\widehat{\theta} \pm k(\alpha) s(\widehat{\theta}),$$

où $s^2(\widehat{\theta}) = \widehat{\mathrm{Var}(\widehat{\theta})}$ est un estimateur sans biais de $\mathrm{Var}(\widehat{\theta}) = \sigma^2/n$.

Dans ce chapitre, on présente trois méthodes de construction d'IC basés sur le bootstrap. Contrairement aux méthodes de la statistique mathématique, les méthodes basées sur le bootstrap n'exigent pas une modélisation paramétrique de la loi F. En outre, les méthodes bootstrap ont l'avantage d'être applicables aux paramètres les plus complexes.

3.1 Méthode du bootstrap-t

En statistique mathématique, l'une des méthodes les plus utiles pour construire un intervalle de confiance est la méthode du pivot. Un pivot exact (resp. approximatif) est une v.a. $R(\underline{X}, F)$ dont la loi exacte (resp. approximative) G_n est indépendante de la loi F. (Les notations $R(\underline{X}, F)$ et G_n sous-entendent ici que $\underline{X} = (X_1, \ldots, X_n)$ est un échantillon aléatoire de taille n de la loi F.)

Par exemple, si $\theta = \mu$ est la moyenne d'une loi normale $N(\mu, \sigma^2)$ et si $\widehat{\sigma}^2 = \sum_1^n (X_i - \overline{X})^2/n$, il est bien connu qu'on a l'égalité en loi

$$R(\underline{X}, F) = \sqrt{n-1}\,\frac{\overline{X} - \mu}{\widehat{\sigma}} \sim t_{n-1}$$

quelles que soient les valeurs de μ et de σ^2, autrement dit $R(\underline{X}, F)$ est un pivot. Pour généraliser cette idée, on introduit ci-après le concept de paramètre de localisation dont μ est un exemple.

Définition 1. Étant donné une loi F, soit $X \sim F$. Faisons l'identification $\theta(F) = \theta(X)$. On dit que θ est un *paramètre de localisation* si les conditions suivantes sont remplies :

i) $\theta(X + b) = \theta(X) + b$, quelle que soit b constante ;

ii) $X \geq 0$ implique que $\theta(X) \geq 0$;

iii) $\theta(aX) = a\theta(X)$, quelle que soit $a > 0$ constante.

Comme exemples de paramètres de localisation $\theta(F)$, citons la moyenne $\mu = E_F(X)$, la médiane $F^{-1}(1/2)$, la moyenne des premier et troisième quar-

tiles $[F^{-1}(1/4) + F^{-1}(3/4)]/2$, la moyenne α-tronquée symétrique $\frac{1}{1-2\alpha}\int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} x dF(x)$, ainsi que les p-quantiles. Pour $0 < p < 1$, rappelons que le *p-quantile* de la loi F est défini par

$$q_p = F^{-1}(p) = \inf\{x : F(x) \geq p\}.$$

(Lorsque F prend toutes les valeurs de $(0,1)$ et possède un inverse (pas de plateau), le p-quantile coïncide avec la solution x_p de l'équation $F(x) = p$.)

Si θ est un paramètre de localisation estimé par $\widehat{\theta}_n$ et si $\widehat{\sigma}_n^2$ est un estimateur de $\sigma^2(\widehat{\theta}_n) = \mathrm{Var}(\widehat{\theta}_n)$, la variable $(\widehat{\theta}_n - \theta)/\widehat{\sigma}_n$ a souvent le comportement d'un pivot exact ou approximatif, en particulier lorsque $\mathrm{Var}(\widehat{\theta}_n)$ est indépendant de θ. Pour rappeler la méthode du pivot, posons G_n comme étant la loi de $(\widehat{\theta}_n - \theta)/\widehat{\sigma}_n$. Si G_n est un pivot exact et si $G_n^{-1}(\alpha)$, $G_n^{-1}(1-\alpha)$ sont les quantiles d'ordre α et $1-\alpha$, on a par définition

$$1 - 2\alpha = P\left(G_n^{-1}(\alpha) \leq \frac{\widehat{\theta}_n - \theta}{\widehat{\sigma}_n} \leq G_n^{-1}(1-\alpha)\right)$$

quel que soit θ, ce qui équivaut à

$$1 - 2\alpha = P\left(\widehat{\theta}_n - G_n^{-1}(1-\alpha)\,\widehat{\sigma}_n \leq \theta \leq \widehat{\theta}_n - G_n^{-1}(\alpha)\,\widehat{\sigma}_n\right). \tag{3.1}$$

(Dans les équations précédentes, l'égalité est remplacée par le symbole $\approx$ lorsque le pivot est approximatif.) Par définition, l'intervalle aléatoire

$$[\widehat{\theta}_n - G_n^{-1}(1-\alpha)\,\widehat{\sigma}_n, \widehat{\theta}_n - G_n^{-1}(\alpha)\,\widehat{\sigma}_n]$$

est un IC de $(1-2\alpha)\times 100\%$ pour θ (exact ou approximatif).

Exemple. Dans certains cas, on ne dispose que d'une approximation de la loi G_n basée sur les propriétés asymptotiques des estimateurs du maximum de vraisemblance. Si G désigne la loi limite des G_n et si cette limite est indépendante de F et connue, on utilise (3.1) pour construire un IC en remplaçant G_n par G. Souvent, $G = G_\psi$ dépend d'un paramètre ψ que l'on estime par $\widehat{\psi}$; on utilise alors (3.1) en remplaçant G_n par $G_{\widehat{\psi}}$. Par exemple,

sous des conditions assez faibles, on sait que la suite $(\widehat{\theta}_n)$ des estimateurs du maximum de vraisemblance de θ converge en loi comme suit :

$$\frac{\widehat{\theta}_n - \theta}{1/\sqrt{nI(\theta)}} = \sqrt{nI(\theta)}(\widehat{\theta}_n - \theta) \xrightarrow{d} N(0,1), \tag{3.2}$$

où $I(\theta)$ est l'information de Fisher associée à une observation, paramètre que l'on peut estimer par $I(\widehat{\theta}_n)$. Pour n assez grand, la variable $\sqrt{nI(\widehat{\theta}_n)}(\widehat{\theta}_n - \theta)$ peut ainsi jouer le rôle de pivot approximatif, la loi d'approximation étant une $N(0,1)$.

La première méthode de construction d'IC basée sur le bootstrap utilise l'idée de pivot approximatif en se servant du bootstrap pour estimer la loi de $(\widehat{\theta} - \theta)/\widehat{\sigma}$. Cette méthode est en fait directement inspirée de la théorie des intervalles de Student. Pour $\widehat{\theta}$ estimateur de $\theta = \theta(F)$ et $\widehat{\sigma}^2$ estimateur de la variance de $\widehat{\theta}$, la méthode fait l'hypothèse que la variable studentisée

$$Z = \frac{\widehat{\theta} - \theta}{\widehat{\sigma}}$$

se comporte comme un pivot approximatif, autrement dit, la loi de Z est approximativement la même quelle que soit θ. À partir de B échantillons bootstrap $\underline{X}_1^*, \ldots, \underline{X}_B^*$ provenant de $\widehat{F}$, on estime la loi de Z à l'aide des valeurs

$$Z_b^* = \frac{\widehat{\theta}_b^* - \widehat{\theta}}{\widehat{\sigma}_b^*}, \quad i = 1, \ldots, B,$$

où $\widehat{\theta}_b^* = \widehat{\theta}(\underline{X}_b^*)$ et $\widehat{\sigma}_b^* = \widehat{\sigma}(\underline{X}_b^*)$ sont des réplications bootstrap calculées à partir de l'échantillon bootstrap $\underline{X}_b^*$.

Pour obtenir un IC pour θ, on calcule ensuite les quantiles échantillonnaux d'ordre α et $1-\alpha$ de la loi empirique des Z_b^*. Pour $k = \lfloor (B+1)\alpha \rfloor$, on définit ceux-ci comme étant les k^e et $(B+1-k)^e$ plus grandes valeurs des Z_b^* ordonnées. Désignons ces valeurs par $\hat{t}^{(\alpha)}$ et $\hat{t}^{(1-\alpha)}$. L'intervalle de confiance pour θ de la méthode du bootstrap-t est défini par

$$[\widehat{\theta} - \hat{t}^{(1-\alpha)}\widehat{\sigma}, \widehat{\theta} - \hat{t}^{(\alpha)}\widehat{\sigma}].$$

Dans cet intervalle, on notera que les quantiles $\hat{t}^{(\alpha)}$ et $\hat{t}^{(1-\alpha)}$ dépendent des observations, ce qui n'est pas le cas pour les intervalles de Student classiques. En outre, contrairement à ces intervalles, on n'a pas nécessairement $\hat{t}^{(\alpha)} = -\hat{t}^{(1-\alpha)}$, et ainsi l'intervalle bootstrap n'est pas nécessairement symétrique par rapport à $\widehat{\theta}$.

Dans la forme présentée ci-dessus, la méthode du bootstrap-t est avant tout applicable à l'estimation des paramètres de localisation. Outre $\overline{X}$, les estimateurs visés sont donc principalement les moyennes tronquées et les quantiles échantillonnaux.

Il nous faut maintenant dire un mot sur le calcul de $\widehat{\sigma}_b^*$. Lorsque $\widehat{\theta} = \overline{X}$, on sait que l'estimateur bootstrap de $\sigma(\widehat{\theta})$ est directement calculable à partir de l'échantillon. En effet, selon la formule (2.2)

$$\widehat{\sigma}_b^* = \widehat{\sigma}(\underline{X}_b^*) = \sqrt{\frac{\sum_1^n (X_{bi}^* - \overline{X}_b^*)^2}{n^2}},$$

où $\overline{X}_b^* = \sum_1^B X_{bi}^*/B$. Lorsqu'il n'existe pas de formule simple pour estimer l'écart type, on pourra toujours se servir du bootstrap à un deuxième niveau, autrement dit en prélevant des échantillons bootstrap sur $\underline{X}_b^*$. En clair, cela signifie :

1. tirer B_1 échantillons bootstrap $\underline{X}_b^*, b = 1, \ldots, B_1$, et calculer $\widehat{\theta}_b^* = \widehat{\theta}(\underline{X}_b^*)$;
2. de chaque échantillon $\underline{X}_b^*$, tirer B_2 échantillons bootstrap pour calculer $\widehat{\theta}_{bj}^*$, $j = 1, \ldots, B_2$, $\overline{\theta}_b^* = \sum_{j=1}^{B_2} \widehat{\theta}_{bj}^*/B_2$ et calculer

$$\widehat{\sigma}_b^{*(B_2)} = \sqrt{\frac{1}{B_2} \sum_1^{B_2} (\widehat{\theta}_{bj}^* - \overline{\theta}_b^*)^2}.$$

Selon Efron, pour estimer un écart type une taille B_2 comprise entre 25 et 200 est en général suffisante ; en revanche, pour estimer des quantiles (dans notre cas, ceux des Z_b^*) il vaut mieux prendre $B_1 = 1000$ au moins. Cela signifie qu'il faut engendrer au minimum 25 000 échantillons bootstrap

pour mettre en œuvre la méthode. Lorsque $\widehat{\theta}$ est compliqué, il peut donc en résulter de grands temps de calcul, défaut principal de la méthode du bootstrap-t.

Une autre faiblesse de la méthode devient apparente lorsque la loi de $\widehat{\theta}$ n'est pas symétrique par rapport à θ. Le bootstrap-t peut alors conduire à des intervalles trop longs ou sortant de l'ensemble des valeurs possibles du paramètre. Par exemple, l'estimation d'un coefficient de corrélation pourra aboutir à un intervalle contenant des valeurs hors de $[-1, 1]$. On pourra souvent contourner ce problème en transformant le paramètre et son estimateur.

Lorsque des observations (X_i, Y_i) proviennent d'une loi normale bivariée, on sait que le coefficient de corrélation de Pearson $r = r_{X,Y}$ a une loi assez fortement asymétrique lorsque $n < 500$. Pour normaliser r, il est bien connu que l'on peut utiliser la transformation de Fisher

$$\widehat{\phi} = T(r) = .5 \log\left(\frac{1+r}{1-r}\right),$$

laquelle produit une variable ayant une loi très proche $N(T(\rho), 1/(n-3))$, où ρ est le coefficient de corrélation de la population. Ainsi, en plus de normaliser, la transformation T a pour effet de produire la variable $T(r)$ dont la variance ne dépend plus de ρ. On dit alors que T *stabilise la variance*. Dans ce cas particulier, pour construire un IC pour ρ, il suffit d'en construire un pour $\phi = T(\rho)$ à partir du pivot approximatif $T(r) - T(\rho)$, puis d'appliquer aux deux bornes la transformation inverse

$$\rho = T^{-1}(\phi) = \frac{e^{2\phi} - 1}{e^{2\phi} + 1}.$$

La même idée de transformation est exploitée pour étendre la méthode du bootstrap-t aux cas où $\widehat{\theta}$ possède une loi asymétrique, cas souvent associés aux situations où $\mathrm{Var}(\widehat{\theta})$ dépend de θ. Dans l'exemple qui précède, la transformation de Fisher stabilise la variance, autrement dit la variance de $\widehat{\phi}$ ne dépend pas de $\phi = T(\rho)$. Des études ont montré que les situations où il

existe une transformation monotone g telle que $\mathrm{Var}(g(\widehat{\theta}))$ soit indépendante de θ sont celles qui conviennent le mieux au bootstrap-t. *On notera qu'une telle transformation normalisante et stabilisante n'existe pas toujours.*

En supposant que l'on puisse stabiliser la variance de $\widehat{\theta}$ par une transformation monotone g, on pourra obtenir cette transformation comme suit. On commence par tirer B_1 échantillons bootstrap $\underline{X}_b^*$, $b = 1, \ldots, B_1$. Pour chaque b, on estime ensuite l'écart type $\sqrt{\mathrm{Var}(\widehat{\theta})}$ par $\widehat{\sigma}_b^*$ à partir de B_2 échantillons bootstrap sur $\underline{X}_b^*$. À l'aide d'une fonction de régression (paramétrique ou non), on ajuste alors une courbe continue $u \mapsto s(u)$ au diagramme de dispersion des points $(\widehat{\theta}_b^*, \widehat{\sigma}_b^*)$, $b = 1, \ldots, B_1$. Cela étant, la transformation stabilisante g s'obtient en appliquant la formule utilisée en analyse de la variance pour stabiliser une variance dépendant de la moyenne :

$$g(z) = \int_a^z \frac{1}{s(u)} du.$$

Finalement, étant donné cette transformation g, on tire B_3 nouveaux échantillons bootstrap par rapport à $\widehat{F}$ pour appliquer la méthode du bootstrap-t à $g(\widehat{\theta}) - g(\theta)$ (puisque la variance de $g(\widehat{\theta})$ ne dépend pas de θ). On obtient ainsi un IC pour $g(\theta)$ que l'on peut inverser par g^{-1} pour obtenir un IC pour θ. On notera que tout cela peut être fait automatiquement dans R (par exemple avec la fonction `boott` du module `bootstrap`).

Selon Efron, la construction d'intervalles de confiance à l'aide du bootstrap devrait normalement se faire en utilisant un nombre d'échantillons bootstrap relativement élevé, de l'ordre de $B \geq 1000$. À la différence de l'estimation du biais et de la variance, l'estimation par intervalle de confiance repose en effet généralement sur une estimation de quantiles d'ordre $0.9, 0.95, 0.975, 0.995$ et leurs symétriques, tous situés dans la région des valeurs extrêmes de la variable utilisée pour la construction.

3.2 Méthode des percentiles

Supposons que le paramètre $\theta = \theta(F)$ soit estimable par $\widehat{\theta} = \widehat{\theta}(\underline{X})$. Étant donné un échantillon bootstrap $X_1^*, \ldots, X_n^* \overset{\text{iid}}{\sim} \widehat{F}$, définissons pour tout x la fonction de répartition bootstrap

$$G_{\text{Boot}}(x) = P_{\widehat{F}}(\widehat{\theta}^* = \widehat{\theta}(X_1^*, \ldots, X_n^*) \leq x).$$

Soit $G_{\text{Boot}}^{-1}(\alpha) \equiv \widehat{\theta}^{*(\alpha)}$ le quantile d'ordre α de la loi de $\widehat{\theta}^*$. L'IC de $(1-2\alpha)\cdot 100\%$ de la méthode des percentiles est défini par

$$[G_{\text{Boot}}^{-1}(\alpha), G_{\text{Boot}}^{-1}(1-\alpha)] \equiv [\widehat{\theta}^{*(\alpha)}, \widehat{\theta}^{*(1-\alpha)}].$$

Dans la pratique, on ne peut énumérer toutes les valeurs possibles de la variable discrète $\widehat{\theta}^*$, ce qui rend impraticable la détermination de la valeur exacte des quantiles $G_{\text{Boot}}^{-1}(\alpha)$. On calcule plutôt une valeur approchée de ces quantiles en simulant des valeurs $\widehat{\theta}_b^*$, $b = 1, \ldots, B$, puis en ordonnant ces valeurs. On obtient alors l'IC approximatif de la méthode des percentiles

$$[\widehat{\theta}_B^{*(\alpha)}, \widehat{\theta}_B^{*(1-\alpha)}] \approx [G_{\text{Boot}}^{-1}(\alpha), G_{\text{Boot}}^{-1}(1-\alpha)]$$

où $\widehat{\theta}_B^{*(\alpha)}$ est le $100\cdot\alpha^e$ percentile de la loi empirique des $\widehat{\theta}_b^*$ (avec la convention déjà vue si $(B+1)\alpha$ n'est pas un entier).

Pour mettre en évidence les qualités de la méthode et la justifier, faisons maintenant l'hypothèse qu'il existe une transformation inversible croissante ϕ telle que la loi de $\phi(\widehat{\theta}) - \phi(\theta)$ soit symétrique continue et indépendante de F, donc un pivot. Si l'on définit la fonction de répartition

$$\Psi(x) = P(\phi(\widehat{\theta}) - \phi(\theta) \leq x), \tag{3.3}$$

on déduit que $\Psi(x) = 1 - \Psi(-x)$ est indépendante de F, donc de θ. [En particulier, lorsqu'elle existe, la variance de $\phi(\widehat{\theta}) - \phi(\theta)$ ne dépend pas de θ, et ainsi ϕ est stabilisante.]

Posons $\widehat{\phi} = \phi(\widehat{\theta})$. Les hypothèses sur ϕ font que la situation est idéale pour construire un IC pour $\phi(\theta)$. En effet, si $z_\alpha = \Psi^{-1}(\alpha)$, la symétrie de

la loi entraîne que $z_{1-\alpha} = \Psi^{-1}(1-\alpha) = -z_\alpha$. Procédant de la manière habituelle, on obtient l'IC de $(1-2\alpha)\cdot 100\%$ pour $\phi(\theta)$:

$$[\widehat{\phi} + z_\alpha, \widehat{\phi} - z_\alpha],$$

d'où par inversion l'IC pour θ :

$$[\phi^{-1}(\widehat{\phi} + z_\alpha), \phi^{-1}(\widehat{\phi} - z_\alpha)].$$

Supposons maintenant qu'on utilise le bootstrap pour estimer la loi de $\phi(\widehat{\theta}) - \phi(\theta)$. On aura

$$\begin{aligned} \alpha &\approx P_{\widehat{F}}(\phi(\widehat{\theta}^*) - \phi(\widehat{\theta}) \leq z_\alpha) \\ &= P(\widehat{\theta}^* \leq \phi^{-1}(\widehat{\phi} + z_\alpha)) \end{aligned}$$

Puisque l'on a également

$$\alpha \approx P_{\widehat{F}}(\widehat{\theta}^* \leq G^{-1}_{\text{Boot}}(\alpha)),$$

on conclut que $G^{-1}_{\text{Boot}}(\alpha)) \approx \phi^{-1}(\widehat{\phi} + z_\alpha)$; de la même façon, on montre que $G^{-1}_{\text{Boot}}(1-\alpha)) \approx \phi^{-1}(\widehat{\phi} - z_\alpha)$.

Le raisonnement précédent fournit une justification théorique de la méthode des percentiles. *Pour que cette méthode soit valide, il suffit que la statistique et le paramètre puissent être transformés par une fonction inversible ϕ telle que $\phi(\widehat{\theta}) - \phi(\theta)$ soit un pivot exact ou approximatif de loi symétrique.* Comme pour la méthode du bootstrap-t, on notera que dans la pratique il n'est pas nécessaire d'identifier une telle transformation ϕ. En effet, la méthode repose totalement sur la loi d'une réplication bootstrap $\widehat{\theta}^*$.

Contrairement à la méthode du bootstrap-t (sans transformation) ou à la méthode dite standard—reposant sur l'approximation normale pour la loi de $(\widehat{\theta} - \theta)/\widehat{\sigma}$, la méthode des percentiles produit toujours des intervalles contenus dans l'ensemble des valeurs possibles du paramètre. D'une manière générale, les méthodes ayant cette propriété ont tendance à être plus précises et plus fiables que les autres.

Si pour toute transformation ϕ l'estimateur $\phi(\widehat{\theta})$ est biaisé pour $\phi(\theta)$ ou possède une variance dépendant de $\phi(\theta)$, la méthode des percentiles peut être insatisfaisante. La méthode suivante a été proposée par Efron en 1987 pour tenir compte de cette possibilité.

3.3 Méthode BC_a

Dans la pratique, la méthode des percentiles produit un IC dont les bornes sont les $100\cdot\alpha^e$ et $100\cdot(1-\alpha)^e$ percentiles des valeurs de $\widehat{\theta}^*$ calculées à partir de B échantillons bootstrap. La méthode BC_a (pour *bias-corrected and accelerated*) produit elle aussi un IC ayant comme bornes des percentiles $\widehat{\theta}^*$, à la différence que ces percentiles dépendent de deux estimations $\widehat{a}$ et $\widehat{z}_0$ de constantes a, z_0 appelées respectivement constantes d'accélération et de correction du biais.

Plus précisément, l'intervalle BC_a de recouvrement $1-2\alpha$ est donné par

$$[\widehat{\theta}^{*(\alpha_1)}, \widehat{\theta}^{*(\alpha_2)}],$$

où α_1, α_2 sont deux probabilités définies par

$$\alpha_1 = \Phi\left(\widehat{z}_0 + \frac{\widehat{z}_0 + z_\alpha}{1-\widehat{a}(\widehat{z}_0+z_\alpha)}\right),$$

$$\alpha_2 = \Phi\left(\widehat{z}_0 + \frac{\widehat{z}_0 + z_{1-\alpha}}{1-\widehat{a}(\widehat{z}_0+z_{1-\alpha})}\right).$$

Dans les formules précédentes, Φ est la fonction de répartition de la loi $N(0,1)$ et z_α le quantile d'ordre α de cette loi. Pour le cas particulier $\widehat{a} = \widehat{z}_0 = 0$, notons que $\alpha_1 = \alpha$ et $\alpha_2 = 1-\alpha$, et l'on constate que l'intervalle BC_a coïncide avec celui de la méthode des percentiles. Malgré la forme relativement complexe des formules du cas général, les calculs ne sont pas difficiles.

La méthode BC_a se propose d'améliorer celle des percentiles en remplaçant l'hypothèse de l'équation (3.3) par l'hypothèse plus générale de

l'existence d'une fonction inversible croissante ϕ et des constantes a, z_0 telles que

$$P\left(\frac{\phi(\widehat{\theta}) - \phi(\theta)}{1 + a\phi(\theta)} + z_0 \leq x\right) = \Phi(x), \tag{3.4}$$

où Φ est la fonction de répartition normale standard. Cela revient à supposer qu'il existe ϕ, a, z_0 telles que

$$U = \frac{\phi(\widehat{\theta}) - \phi(\theta)}{1 + a\phi(\theta)} + z_0 \sim N(0,1).$$

Dans la dernière expression, on peut identifier $1+a\phi(\theta) > 0$ à l'écart type de $\phi(\widehat{\theta})$. La constante a dépend de n et de F et mesure le taux de changement (accélération) de la variance de $\phi(\widehat{\theta})$ comme fonction de $\phi(\theta)$; la constante z_0 également dépendante de n et de F vise à corriger l'influence d'un biais possible dans l'estimation de $\phi(\theta)$ par $\phi(\widehat{\theta})$.

Supposons pour le moment que a et z_0 soient connues. Sous l'hypothèse précédente, des IC de $(1-2\alpha)\cdot 100\%$ pour $\phi(\theta)$ et θ s'obtiennent comme suit :

$$\begin{aligned}
1-2\alpha &= P\left(z_\alpha \leq \frac{\phi(\widehat{\theta}) - \phi(\theta) + z_0[1 + a\phi(\theta)]}{1 + a\phi(\theta)} \leq -z_\alpha\right) \\
&= P\left(\frac{\phi(\widehat{\theta}) + z_0 + z_\alpha}{1 - a(z_0 + z_\alpha)} \leq \phi(\theta) \leq \frac{\phi(\widehat{\theta}) + z_0 - z_\alpha}{1 - a(z_0 - z_\alpha)}\right) \\
&= P\left(\phi^{-1}\left(\frac{\phi(\widehat{\theta}) + z_0 + z_\alpha}{1 - a(z_0 + z_\alpha)}\right) \leq \theta \leq \phi^{-1}\left(\frac{\phi(\widehat{\theta}) + z_0 - z_\alpha}{1 - a(z_0 - z_\alpha)}\right)\right).
\end{aligned}$$

Appelons T_1, T_2 les bornes inférieures et supérieures, respectivement, de l'intervalle de confiance pour $\phi(\theta)$; les bornes de l'IC pour θ sont donc $\phi^{-1}(T_1)$ et $\phi^{-1}(T_2)$.

On peut calculer approximativement la borne inférieure $\phi^{-1}(T_1)$ en ap-

pliquant le principe du bootstrap. On a d'abord

$$\begin{aligned} P_{\widehat{F}}\left(\phi(\widehat{\theta}^*) \leq T_1\right) &= P_{\widehat{F}}\left(\phi(\widehat{\theta}^*) \leq \frac{\phi(\widehat{\theta}) + z_0 + z_\alpha}{1 - a(z_0 + z_\alpha)}\right) \\ &= P_{\widehat{F}}\left(\phi(\widehat{\theta}^*) \leq \phi(\widehat{\theta}) + \frac{z_0 + z_\alpha}{1 - a(z_0 + z_\alpha)}[1 + a\phi(\widehat{\theta})]\right) \\ &= P_{\widehat{F}}\left(\frac{\phi(\widehat{\theta}^*) - \phi(\widehat{\theta})}{1 + a\phi(\widehat{\theta})} + z_0 \leq \frac{z_0 + z_\alpha}{1 - a(z_0 + z_\alpha)} + z_0\right) \\ &\approx \Phi\left(z_0 + \frac{z_0 + z_\alpha}{1 - a(z_0 + z_\alpha)}\right). \end{aligned}$$

Ensuite, puisque $P_{\widehat{F}}\left(\phi(\widehat{\theta}^*) \leq T_1\right) = P_{\widehat{F}}\left(\widehat{\theta}^* \leq \phi^{-1}(T_1)\right)$, le raisonnement précédent montre que la borne inférieure $\phi^{-1}(T_1)$ est le quantile d'ordre α_1 $G^{-1}_{\text{Boot}}(\alpha_1)$ de la loi de $\widehat{\theta}^*$, où

$$\alpha_1 = \Phi\left(z_0 + \frac{z_0 + z_\alpha}{1 - a(z_0 + z_\alpha)}\right).$$

De la même façon, la borne supérieure $\phi^{-1}(T_2)$ correspond au quantile d'ordre α_2 de la loi de $\widehat{\theta}^*$, où

$$\alpha_2 = \Phi\left(z_0 + \frac{z_0 - z_\alpha}{1 - a(z_0 - z_\alpha)}\right).$$

Sans entrer dans les détails de l'estimation de z_0 et a, on estimera la constante z_0 à partir de la proportion de réplications bootstrap de $\widehat{\theta}$ inférieures à $\widehat{\theta}$:

$$\widehat{z}_0 = \Phi^{-1}\left(\frac{\#\{b = 1, \ldots, B : \widehat{\theta}^*_b < \widehat{\theta}\}}{B}\right)$$

où Φ^{-1} est la fonction quantile de la loi normale standard : par exemple, $\Phi^{-1}(.5) = 0, \Phi^{-1}(.95) = 1.645$. En gros, $\widehat{z}_0$ mesure le biais de la médiane de $\widehat{\theta}^*$ par rapport à $\widehat{\theta}$. Pour sa part, la constante a est estimée en termes des réplications jackknife de $\widehat{\theta}$. De manière précise, en posant encore $\widehat{\theta}_{(i)} = \widehat{\theta}(\underline{X}_{(i)})$, $\widehat{\theta}_{(\cdot)} = \sum_i \widehat{\theta}_{(i)}/n$, on prend

$$\widehat{a} = \frac{\sum_1^n (\widehat{\theta}_{(\cdot)} - \widehat{\theta}_{(i)})^3}{6(\sum_1^n (\widehat{\theta}_{(\cdot)} - \widehat{\theta}_{(i)})^2)^{3/2}}.$$

En conclusion, on notera que l'application de la méthode BC_a n'exige pas l'identification de la fonction ϕ dont l'existence est postulée dans l'équation (3.4). Comme on l'a vu ci-dessus, cette application repose plutôt sur trois choses relativement simples : l'estimation de a et z_0, le calcul de α_1 et α_2, et finalement, le calcul des réplications bootstrap de $\widehat{\theta}$.

3.4 Précision des intervalles construits avec le bootstrap

Pour un IC bilatéral exact $[T_1, T_2]$ de $(1-2\alpha)\cdot 100\%$, on a par définition

$$P(\theta < T_1) = \alpha \quad \text{et} \quad P(\theta > T_2) = \alpha. \tag{3.5}$$

On peut mesurer la valeur d'un IC approximatif selon le degré de précision atteint en (3.5). On peut montrer que la méthode BC_a ainsi que la méthode du bootstrap-t ont une précision d'ordre deux, ce qui signifie que

$$P(\theta < T_1) = \alpha + O(1/n) \quad \text{et} \quad P(\theta > T_2) = \alpha + O(1/n).$$

En comparaison, les méthodes standard (basées sur un pivot normal) et des percentiles ont une précision d'ordre un :

$$P(\theta < T_1) = \alpha + O(1/\sqrt{n}) \quad \text{et} \quad P(\theta > T_2) = \alpha + O(1/\sqrt{n}).$$

En général, les suites $O(1/n)$ convergent vers 0 plus vite que les suites $O(1/\sqrt{n})$. En ce sens, on considère qu'une méthode de précision d'ordre deux est supérieure à une méthode de précision d'ordre un. Pour que cette évaluation asymptotique de la précision ait une valeur pratique, il faut bien sûr que n soit suffisamment grand.

3.5 Bibliographie

– Shao, J. et Tu, D. (1995). *The Jackknife and Bootstrap*, Springer-Verlag, New York. Chapitre 4.

– Efron, B. et Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, Chapman and Hall, New York. Chapitres 12–14.
– Davison, A.C. et Hinkley, D.V. (1997). *Bootstrap Methods and their Applications*, Cambridge University Press, Cambridge. Chapitre 5.
– Givens, G. H. et Hoeting, J. A. (2005). *Computational Statistics*, J. Wiley & Sons, New York. Chapitre 9. Ouvrage de bon niveau, assez théorique, couvrant l'ensemble de la matière du cours. Certains sujets sont abordés superficiellement.

Chapitre 4

Estimation de densité

4.1 L'histogramme

Soit X une variable aléatoire ayant la densité f et la fonction de répartition F. En tout point de continuité de f, on sait que

$$f(x) = F'(x) = \lim_{h \to 0} \frac{F(x+h) - F(x)}{h}. \tag{4.1}$$

Supposons que l'on veuille estimer $f(x)$ à partir d'un échantillon $x_1, \ldots,$ x_n de valeurs de X. Pour le faire de manière non paramétrique, il suffit d'utiliser la fonction de répartition empirique. Pour ce faire, on commence par se donner k intervalles (classes)

$$[b_0, b_1], (b_1, b_2], \ldots, \ldots, (b_{k-1}, b_k]$$

de même longueur h et contenant l'ensemble des observations. En vertu de (4.1), il est naturel d'estimer f par la fonction en escalier

$$\begin{aligned} \widehat{f}(x) &= \frac{\widehat{F}(b_{j+1}) - \widehat{F}(b_j)}{h}, \qquad x \in (b_j, b_{j+1}] = (b_j, b_j + h], \\ &= \frac{(\#\{x_i \leq b_{j+1}\} - \#\{x_i \leq b_j\})/n}{h} \\ &= \frac{n_j}{nh}, \end{aligned}$$

où $n_j = \#\{x_i \in (b_j, b_{j+1}]\}$; pour x en dehors de l'intervalle $[b_0, b_k]$, on pose $\widehat{f}(x) = 0$. L'estimateur ainsi défini s'appelle l'*histogramme*.

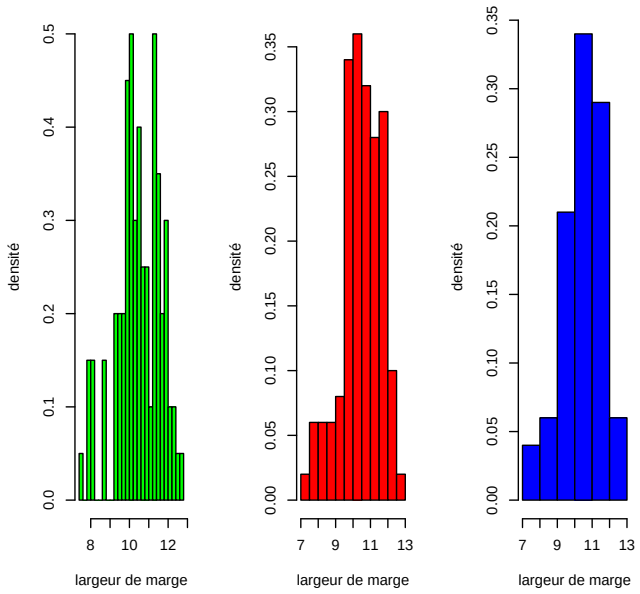

FIG. 4.1 – Histogramme de la variable V1 du jeu `swissmon.dat` utilisant respectivement 28, 12 et 6 classes. La variable V1 représente la largeur de la marge d'un billet de banque suisse contrefait (en mm) et $n = 100$.

On notera qu'en statistique descriptive le terme histogramme désigne la représentation graphique de $\widehat{f}$ dans laquelle, au-dessus de chaque intervalle $(b_j, b_{j+1}] = (b_j, b_j + h]$, on élève le rectangle de hauteur $n_j/(nh)$. Puisque la somme des aires de ces rectangles $\sum_{j=1}^{k} h \times n_j/(nh) = 1$, on voit que l'estimateur $\widehat{f}$ est effectivement une densité.

Il est clair que la construction de $\widehat{f}$ peut être faite même si les données ne proviennent pas d'une loi continue. *Dans la suite, le terme histogramme sera employé dans le sens restreint d'estimateur d'une densité f.*

4.1.1 Propriétés de l'histogramme

Comme le montre la figure 4.1, le nombre de classes a un impact important sur le graphique de l'histogramme des données `swissmon.dat`. Pour

$k = 28$, on détecte trois ou quatre modes sur un histogramme très irrégulier ; pour $k = 12$, deux modes sont encore détectables mais ceux-ci sont moins affirmés ; pour $k = 6$, il n'y a plus qu'un mode, mais on arrive à détecter une asymétrie.

Ces trois histogrammes illustrent une caractéristique commune aux méthodes d'estimation de densité : lorsque la hauteur des rectangles colle bien à la distribution (biais faible), cette hauteur est très variable (lissage faible ou *sous-lissage*, irrégularité du premier graphique) ; lorsque la hauteur des rectangles colle moins bien à la distribution (biais important), la hauteur est peu variable (lissage important ou *surlissage*, régularité du troisième graphique). Un bon histogramme arrive à trouver un juste équilibre entre biais et variabilité de l'estimation, ce qui revient à choisir adéquatement le nombre d'intervalles ou, si l'on préfère, la longueur h de ceux-ci. Dans la suite, on donne à h le nom de *paramètre de lissage*. La détermination d'une valeur adéquate de celui-ci est un problème important que nous examinerons en détail.

Pour mesurer la précision de $\widehat{f}$ en *un point x fixé*, on utilisera l'erreur quadratique aléatoire $SE(x) = (\widehat{f}(x) - f(x))^2$, ou mieux, l'erreur quadratique moyenne

$$MSE_F[\widehat{f}(x)] = E_F[\widehat{f}(x) - f(x)]^2,$$

où la notation E_F signifie que l'espérance est évaluée par rapport à la loi F de la variable X. À ce propos, il est important d'observer que pour tout x fixé la variable aléatoire $\widehat{f}(x)$ est une statistique :

$$\widehat{f}(x) = \frac{N_j}{nh} = \frac{\#\{X_i \in (b_j, b_{j+1}]\}}{nh} = T_x(X_1, \ldots, X_n).$$

Notons en outre que $\widehat{f}(x)$ dépend fortement de h ; lorsque, comme ci-dessous, l'on fait varier h, on désignerait donc plus précisément l'estimateur par $\widehat{f}(x; h)$.

Il est cependant préférable de mesurer la précision de façon globale, autrement dit sur l'ensemble de tous les x. Une telle mesure (aléatoire) est

l'*erreur quadratique intégrée* :

$$ISE = \int [\widehat{f}(x) - f(x)]^2 dx.$$

Plus loin, nous utiliserons plutôt la valeur moyenne de l'ISE :

$$MISE = E_F\left(\int [\widehat{f}(x) - f(x)]^2 dx\right) = \int MSE_F[\widehat{f}(x)]dx.$$

On peut voir que

$$N_j := \#\{X_i \in (b_j, b_{j+1}]\} \sim \text{Bin}(n, p_j),$$

où $p_j = P(b_j < X \leq b_{j+1})$. Pour $x \in (b_j, b_{j+1}]$, on en déduit que

$$E_F[\widehat{f}(x)] = \frac{E_F(N_j)}{nh} = \frac{p_j}{h}$$

et

$$\text{Var}_F[\widehat{f}(x)] = \frac{p_j(1-p_j)}{nh^2}. \tag{4.2}$$

Ces calculs permettent de constater que

$$\begin{aligned} MSE_F[\widehat{f}(x)] &= E_F[\widehat{f}(x) - f(x)]^2 \\ &= \text{Var}_F[\widehat{f}(x)] + \text{Biais}_F^2[\widehat{f}(x)] \end{aligned}$$

dépend de n et h pour la partie variance, et de h pour la partie biais. Cela nous amène plus loin à évaluer la performance de l'histogramme au point x à partir du comportement (asymptotique) de $MSE_F[\widehat{f}(x)]$ lorsque $n \nearrow \infty$ et $h \searrow 0$.

Dans ce qui suit, pour $k > 0$ on désigne par $O(h^k)$ toute fonction $g(h)$ telle $|g(h)|/h^k$ est bornée quand $h \searrow 0$: autrement dit, toute fonction qui en valeur absolue est $\leq Mh^k$ pour une certaine constante $M > 0$ et pour tout $h > 0$ assez petit. Une fonction de la classe $O(h^k)$ tend donc vers 0 à une vitesse au moins aussi grande que h^k lorsque $h \searrow 0$.

Pour f suffisamment lisse, f, f', f'' bornées et $x \in (b_j, b_{j+1}]$, on peut exprimer le biais de $\widehat{f}(x)$ comme fonction de h :

$$\begin{aligned}
\text{Biais}_F[\widehat{f}(x)] &= E_F[\widehat{f}(x)] - f(x) \\
&= \frac{p_j}{h} - f(x) \\
&= \int_{b_j}^{b_{j+1}} f(t)dt/h - f(x) \\
&= \int_{b_j}^{b_{j+1}} \left[f(x) + (t-x)\cdot f'(x) + O(h^2)\right] dt/h - f(x) \\
&= \frac{f'(x)}{2h}\, t^2\Big|_{b_j}^{b_j+h} - xf'(x) + \int_{b_j}^{b_{j+1}} O(h^2)dt/h \\
&= \frac{f'(x)}{2}[h - 2(x-b_j)] + O(h^2). \qquad (4.3)
\end{aligned}$$

En effet, comme f'' est par hypothèse bornée, le reste du développement de Taylor ci-dessus $(t-x)^2 f''(\theta)/2 \leq Mh^2$, où M est indépendante de t, x et n, puisque $\theta \in [b_j, b_{j+1}]$ et $(t-x)^2 \leq h^2$ pour $t \in [b_j, b_{j+1}]$. Cela entraîne que $|\int_{b_j}^{b_{j+1}} O(h^2)dt/h| \leq Mh^2$, et donc que $\int_{b_j}^{b_{j+1}} O(h^2)dt/h = O(h^2) \leq Mh^2$. (Observer que la notation $O(h^2)$ sert ici à désigner deux fonctions différentes.)

À partir de (4.2), on peut ensuite mettre en évidence le comportement asymptotique de $\text{Var}_F[\widehat{f}(x)]$ comme fonction de n et h. Pour $x \in (b_j, b_{j+1}]$, il vient

$$\begin{aligned}
\text{Var}_F[\widehat{f}(x)] &= \frac{p_j(1-p_j)}{nh^2} \\
&= \frac{\int_{b_j}^{b_{j+1}} f(t)dt}{nh^2} - \frac{p_j^2}{nh^2} \\
&= \frac{\int_{b_j}^{b_{j+1}} [f(x) + O(h)]dt}{nh^2} - \frac{p_j^2}{nh^2} \\
&= \frac{f(x)}{nh} + O(1/n) - nh^2\left(\frac{f(x)}{nh} + O(1/n)\right)^2 \\
&= \frac{f(x)}{nh} + O(1/n), \qquad (4.4)
\end{aligned}$$

où $|O(1/n)| \leq K/n$ pour une borne positive K indépendante de x et de h.

De (4.3), on voit que $h \to 0$ entraîne que $\text{Biais}_F[\widehat{f}(x)] \to 0$; de (4.4), il suit que $\text{Var}_F(\widehat{f}(x)) \to 0$ lorsque $nh \to \infty$. En pratique, on prendra $h = h(n)$. Pour que le biais et la variance de l'histogramme en x soient petits, il faut donc que n soit grand et qu'en même temps $h(n)$ soit petit, tout en ayant $nh(n)$ grand.

De ce qui précède, il s'ensuit que

$$\begin{aligned} MSE[\widehat{f}(x)] &= \text{Var}[\widehat{f}(x)] + \text{Biais}^2[\widehat{f}(x)] \\ &= \frac{f(x)}{nh} + O(1/n) + \frac{f'(x)^2}{4}[h - 2(x - b_j)]^2 + O(h^3) + O(h^4) \\ &= \frac{f(x)}{nh} + \frac{f'(x)^2}{4}[h - 2(x - b_j)]^2 + O(h^3) + O(1/n), \end{aligned}$$

où en valeur absolue $O(h^3)$ et $O(1/n)$ sont bornées en x. Finalement, en intégrant par rapport à x on peut montrer que (voir Scott, 54)

$$MISE = \frac{1}{nh} + \frac{h^2 R(f')}{12} + O(h^3) + O(1/n), \tag{4.5}$$

où $R(f') = \int f'(t)^2 dt$.

On interprète $R(f')$ comme une mesure de la régularité de la densité à estimer f. Lorsque f' prend de grandes valeurs (f irrégulière), $R(f')$ est grand ; lorsque f est lisse (valeurs de f' proches de 0), $R(f')$ prend de petites valeurs. Dans (4.5), on voit que le carré du biais intégré est proportionnel au carré du paramètre de lissage h et que la variance intégrée est inversement proportionnelle à ce paramètre. On pourra donc dire qu'un petit h produit un histogramme peu biaisé, tandis qu'un grand nh détermine un histogramme peu variable. Pour atteindre ce double objectif, on devra réconcilier ces contraintes un peu contradictoires.

4.1.2 Choix du paramètre de lissage

En pratique, on choisit h en fonction de n. Nous présentons maintenant trois des règles les plus utilisées.

1) Règle de Sturges (1926)

Choisir le nombre k de classes égal à $\lceil 1+\log_2 n \rceil$ (plus petit entier $\geq$). On prendra alors $h = (X_{(n)} - X_{(1)})/k$ (voir Scott, 48). Cette règle est obtenue à partir du principe que le modèle normal peut servir de référence lorsqu'il est impossible d'identifier un bon modèle.

La règle de Sturges a tendance à produire des histogrammes trop lisses (surlissage). C'est la règle utilisée par défaut dans R et dans la plupart des logiciels statistiques. On la considère insatisfaisante lorsque n est grand.

Nous examinerons deux autres méthodes de sélection qui dans l'ensemble sont plus acceptables. Elles font toutes deux appel à la partie dominante du $MISE$, partie que l'on nomme le $MISE$ *asymptotique* :

$$AMISE = AMISE(h) = \frac{1}{nh} + \frac{h^2 R(f')}{12}.$$

On vérifie facilement que cette mesure globale de précision de l'histogramme vue comme fonction de h est minimisée en prenant

$$h_{opt} = \left[\frac{6}{R(f')}\right]^{1/3} \frac{1}{n^{1/3}}. \tag{4.6}$$

Cette valeur 'optimale' du paramètre de lissage h entraîne que

$$AMISE_{opt} = \left[\frac{9R(f')}{16}\right]^{1/3} \frac{1}{n^{2/3}}.$$

Lorsque $n \to \infty$, l'$AMISE_{opt}$ tend donc vers 0 à la vitesse $n^{-2/3}$, une vitesse plus lente que celle de la variance de la plupart des estimateurs paramétriques laquelle converge généralement à la vitesse n^{-1}.

Pour les applications pratiques, on note toutefois qu'un problème bien concret se pose : comme f n'est pas connue, f' ne l'est pas davantage, ce qui empêche d'utiliser directement l'équation (4.6).

Faute de mieux, on utilise souvent (4.6) en prenant pour f la densité de la loi normale $N(\mu, \sigma^2)$. On peut alors montrer que le paramètre de lissage optimal est

$$\widehat{h}_{opt} = (24\sigma^3\sqrt{\pi})^{1/3} \frac{1}{n^{1/3}} = 3.491\sigma n^{-1/3}.$$

En estimant σ par l'écart type S échantillonnal, on obtient ainsi la règle de Scott.

2) Règle de Scott (1979)

Prendre

$$\boxed{\widehat{h}_{opt} = 3.491 S n^{-1/3}}.$$

Une règle du même type, moins sensible aux données aberrantes, est celle de Freedman-Diaconis.

3) Règle de Freedman-Diaconis (1981)

Prendre

$$\boxed{\widehat{h}_{opt} = 2IQn^{-1/3}}$$

où $IQ = Q_3 - Q_1$ est l'écart interquartile.

Remarque 1. On notera que les deux dernières règles font du surlissage lorsque la densité a plusieurs modes.

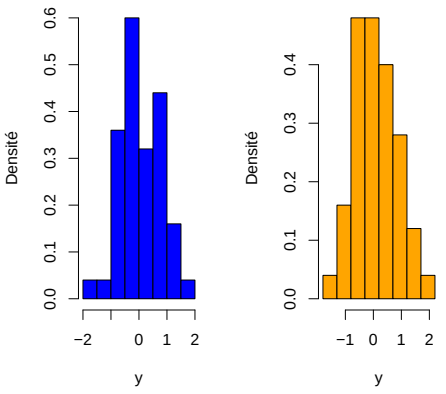

FIG. 4.2 – Influence du point d'ancrage sur un histogramme

La plus grande qualité de l'histogramme est sa simplicité. Parmi ses défauts importants, citons celui d'être dépendant du point d'ancrage, autrement dit de la position du premier intervalle (voir Figure 4.2). On peut aussi reprocher à l'histogramme d'être trop peu sensible aux propriétés locales de f. En outre, alors que la plupart des fonctions de densité ne sont pas des fonctions en escalier, l'histogramme est toujours de cette forme. Nous présentons maintenant un premier exemple d'estimateur de densité ayant la forme d'une fonction continue, le polygone de fréquence.

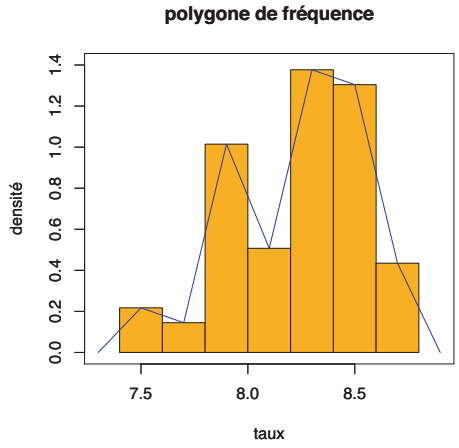

FIG. 4.3 – Polygone de fréquence pour les valeurs de la variable `V1` du tableau `cdrate.dat`.

4.2 Le polygone de fréquence

Même si la fonction de densité f est indéfiniment différentiable, l'histogramme l'estime par une fonction discontinue. Il existe une manière simple de construire un estimateur continu à partir de l'histogramme. Elle consiste à joindre par un segment de droite tous les couples consécutifs de points $(m_i, \widehat{f}(m_i))$, où $\widehat{f}$ est l'histogramme, $m_i = (b_i + b_{i+1})/2$, $i = 0, \ldots, k-1$, $m_{-1} = b_0 - h/2$ et $m_k = b_k + h/2$. On obtient alors l'estimateur appelé

polygone de fréquence. Formellement, celui-ci est défini par

$$\widehat{f}_p(x) = \frac{1}{nh^2}\left[n_{j-1}m_{j+1} - n_j m_j + (n_j - n_{j-1})x\right], \quad x \in [m_j, m_{j+1}],$$

où $n_{-1} = n_k = 0$. La figure 4.3 montre un exemple de polygone de fréquence superposé à un histogramme. Le polygone de fréquence est continu et différentiable partout, sauf en chacun des points milieux des intervalles $[b_j, b_{j+1}]$.

Comme pour l'histogramme, on peut faire une analyse du comportement asymptotique du biais et de la variance du polygone de fréquence. En supposant que f soit suffisamment lisse, Scott a montré que

$$MISE = \frac{2}{3nh} + \frac{49h^4 R(f'')}{2880} + O(n^{-1}) + O(h^6).$$

La variance intégrée est encore inversement proportionnelle à nh. Le carré du biais intégré est proportionnel à h^4 ainsi qu'à $R(f'')$, un terme dépendant de la courbure (dérivée seconde) de la densité f. Le biais du polygone de fréquence est donc d'un ordre de grandeur plus petit que celui de l'histogramme. Le polygone de fréquence parvient à cette augmentation de précision en ne retenant que les valeurs de l'histogramme aux centres des intervalles définissant l'histogramme.

Comme ci-dessus, la minimisation de l'AMISE conduit au paramètre de lissage optimal

$$h_{opt} = 2\left[\frac{15}{49R(f'')}\right]^{1/5} \frac{1}{n^{1/5}}, \tag{4.7}$$

ainsi qu'à la valeur minimale

$$AMISE_{opt} = \frac{5}{12}\left[\frac{49R(f'')}{15}\right]^{1/5} \frac{1}{n^{4/5}}.$$

De ce point de vue, on peut ainsi dire que la vitesse de convergence du polygone de fréquence, proportionnelle à $\frac{1}{n^{4/5}}$, est supérieure à celle de l'histogramme, proportionnelle à $\frac{1}{n^{2/3}}$.

Dans la pratique, la manière la plus simple d'estimer le paramètre de lissage est de supposer que f est gaussienne dans (4.7). On obtient alors

$h_{opt} = 2.15\sigma n^{-1/5}$, ce qui conduit à la règle de Scott

$$\widehat{h}_{opt} = 2.15Sn^{-1/5},$$

où S est un estimateur de l'écart type. Comparé au paramètre de lissage optimal de l'histogramme, le paramètre analogue du polygone de fréquence tend à être plus grand : $n^{-1/5}$ est toujours plus grand que $n^{-1/3}$. Pour que l'histogramme discontinu produise une bonne estimation de f, il faut que son paramètre de lissage soit plus petit. À cause de sa continuité, le polygone de fréquence arrive à estimer une densité continue de façon acceptable à partir d'un paramètre de lissage plus grand.

Selon ce qui précède, le polygone de fréquence est un estimateur plus précis que l'histogramme lorsque n est assez grand et h bien choisi. Des études de Monte Carlo montrent que cette supériorité est observée même pour des tailles n petites. Dans la pratique, on préférera donc le polygone de fréquence à l'histogramme. Comme l'histogramme, le polygone de fréquence a l'inconvénient de dépendre du point d'ancrage. L'estimateur à noyau n'a pas ce défaut, tout en ayant l'avantage d'être plus lisse que l'histogramme ou le polygone de fréquence.

4.3 L'estimateur à noyau

On sait qu'en chacun de ses points de continuité une densité f est liée à sa fonction de répartition F par l'équation :

$$f(x) = \lim_{h \to 0} \frac{F(x+h) - F(x-h)}{2h}. \tag{4.8}$$

En effet, cela est une conséquence de l'identité

$$\frac{F(x+h) - F(x-h)}{2h} = \frac{F(x+h) - F(x)}{2h} + \frac{F(x-h) - F(x)}{-2h},$$

où les deux termes du membre droit tendent vers $f(x)/2$ quand $h \to 0$.

Rappelons que la construction d'un histogramme repose sur le choix d'un point d'ancrage $b_0 \leq \min x_i$, et de k intervalles disjoints recouvrant l'ensemble des observations et ayant la même longueur h, paramètre contrôlant

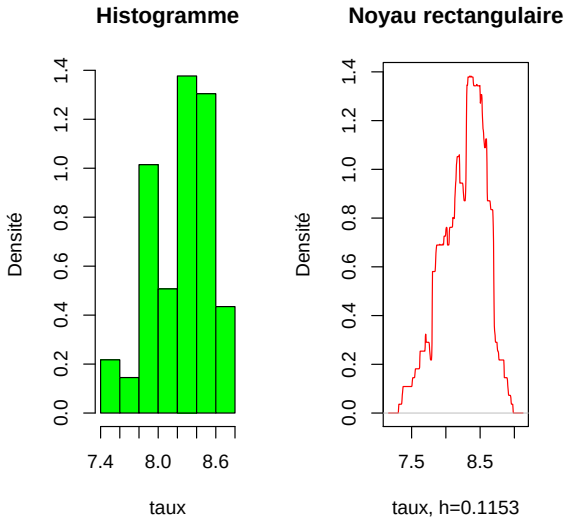

FIG. 4.4 – Histogramme et estimateur à noyau rectangulaire pour la première variable du jeu `cdrate.dat` (69 taux d'intérêt sur des certificats de dépôt)

le lissage. À l'exception du point d'ancrage, les extrémités de ces intervalles sont indépendantes des données.

Pour mesurer la densité des données en un point x sans faire intervenir de point d'ancrage, l'estimateur à noyau propose de s'intéresser aux données contenues dans l'intervalle $(x-h, x+h]$ pour le paramètre h fixé. Cette nouvelle approche est suggérée par (4.8). Selon cette équation, la fonction de répartition empirique permet en effet d'estimer f au point x par

$$\widehat{f}(x) = \frac{\#\{x_i \in (x-h, x+h]\}}{2nh}. \tag{4.9}$$

En posant

$$K(u) = \begin{cases} 1/2, & -1 \leq u < 1 \\ 0, & \text{sinon} \end{cases}$$

on peut réécrire (4.9) sous la forme

$$\widehat{f}(x) = f(x;h) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right). \tag{4.10}$$

Nous venons de définir l'*estimateur à noyau* dit *rectangulaire* (ou uniforme, ou naïf), du nom de la forme prise ici par le noyau K, à savoir la densité uniforme sur $[-1, 1)$. La figure 4.4 ci-dessus compare un histogramme à 7 intervalles avec l'estimateur à noyau rectangulaire de paramètre $h = 0.1153$. L'estimation porte sur la première variable du jeu de données `cdrate.dat`.

L'estimateur à noyau rectangulaire reflète plus fidèlement les propriétés locales de f que l'histogramme. En tant que somme de valeurs de K, cet estimateur a les mêmes propriétés de continuité et de différentiabilité que la densité uniforme, ce qui explique son aspect en dents de scie. Pour obtenir un estimateur plus lisse, il suffit d'utiliser la définition (4.10) avec un noyau K plus régulier. La figure 4.5 ci-après montre le graphique de l'estimateur à noyau pour le *noyau gaussien*

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$$

et $h = 0.1153$.

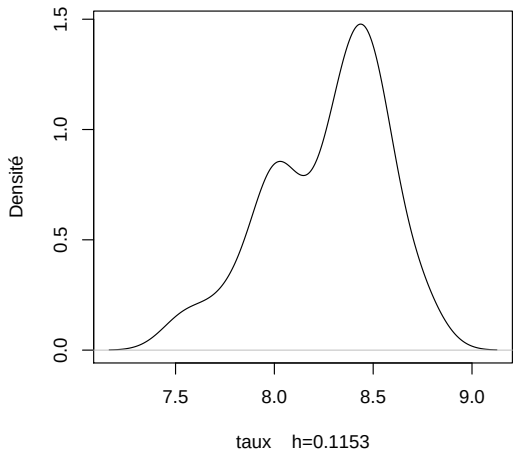

FIG. 4.5 – Estimateur à noyau gaussien pour les données `cdrate.dat`

Comme l'histogramme et le polygone de fréquence, l'estimateur à noyau

dépend fortement du paramètre de lissage h (appelé en anglais *bandwidth* ou *window width*). Comme pour l'histogramme et le polygone de fréquence, pour n fixé on pourra baser le choix d'un h optimal sur l'$AMISE$. Cela revient encore à minimiser à la fois la variance et le carré du biais intégrés, en recherchant un équilibre entre ces deux exigences contradictoires.

Dans le raisonnement qui suit, on suppose que le noyau K est une fonction symétrique par rapport à 0 vérifiant les trois conditions suivantes :

$$\int K(u)du = 1, \int uK(u)du = 0, \int u^2K(u)du = \sigma_K^2 > 0.$$

On notera que ces conditions n'obligent pas K à être positive et, en particulier, à être une densité. On suppose aussi que la densité à estimer f est suffisamment régulière en ce sens que ses dérivées d'ordres deux et trois existent et ont un bon comportement. En outre, on fait l'hypothèse que $f(x) \to 0$ lorsque $|x| \to \infty$.

Compte tenu des propriétés du noyau K, au point x on a

$$\begin{aligned}
\text{Biais}_F[\widehat{f}(x)] &= E_F\left[\frac{1}{nh}\sum_1^n K\left(\frac{x-X_i}{h}\right)\right] - f(x) \\
&= E_F\left[\frac{1}{h}K\left(\frac{x-X}{h}\right)\right] - f(x) \quad \text{(puisque les } X_i \text{ sont i.i.d.)} \\
&= \frac{1}{h}\int K\left(\frac{x-y}{h}\right) f(y)dy - f(x) \\
&= \int K(u)f(x-hu)du - f(x) \\
&= \int K(u)[f(x-hu) - f(x)]du \\
&= \int K(u)[-huf'(x) + \frac{1}{2}h^2u^2f''(x) + \cdots]du \\
&= -hf'(x)\int uK(u)du + \frac{1}{2}h^2f''(x)\int u^2K(u)du + \cdots \\
&= \frac{h^2f''(x)\sigma_K^2}{2} + O(h^4) \\
&= O(h^2)
\end{aligned}$$

lorsque h est petit. De même, lorsque h est petit, le calcul précédent implique que

$$\begin{aligned}
\mathrm{Var}_F[\widehat{f}(x)] &= \frac{1}{n}\mathrm{Var}_F\left(\frac{1}{h}K\left(\frac{x-X}{h}\right)\right) \quad \text{(puisque les } X_i \text{ sont i.i.d.)} \\
&= \frac{1}{n}\left\{E_F\left[\frac{1}{h^2}K\left(\frac{x-X}{h}\right)^2\right] - E_F^2\left[\frac{1}{h}K\left(\frac{x-X}{h}\right)\right]\right\} \\
&= \frac{1}{n}\int \frac{1}{h^2}K\left(\frac{x-y}{h}\right)^2 f(y)dy - \frac{1}{n}\left(f(x) + \mathrm{Biais}_F[\widehat{f}(x)]\right)^2 \\
&= \frac{1}{nh}\int K^2(u)f(x-hu)du - \frac{1}{n}\left(f(x)+O(h^2)\right)^2 \\
&= \frac{1}{nh}\int K^2(u)[f(x) - huf'(x) + \cdots]du + O(1/n) \\
&= \frac{R(K)f(x)}{nh} + O(1/n),
\end{aligned}$$

où R a la même signification que pour l'histogramme et le polygone de fréquence.

L'erreur quadratique moyenne au point x s'obtient en additionnant la variance et le carré du biais :

$$MSE_F[\widehat{f}(x)] = \frac{R(K)f(x)}{nh} + \frac{h^4\sigma_K^4[f''(x)]^2}{4} + O(1/n) + O(h^6).$$

En intégrant par rapport à x, on obtient ensuite le $MISE$, dont la partie dominante, le $MISE$ asymptotique, est donnée par

$$AMISE = AMISE(n,h) = \frac{R(K)}{nh} + \frac{h^4\sigma_K^4 R(f'')}{4}. \tag{4.11}$$

Comme plus haut, l'AMISE est vu comme une mesure globale de précision de l'estimateur à noyau. Pour minimiser la variance et le biais intégrés, on choisira $h = h(n)$ tel que $nh(n) \to \infty$ et $h(n) \to 0$ lorsque $n \to \infty$. Pour n fixé, on vérifie que le h minimisant $AMISE(n,h)$ est

$$h_{opt} = \left[\frac{R(K)}{\sigma_K^4 R(f'')}\right]^{1/5}\frac{1}{n^{1/5}},$$

d'où il suit que l'$AMISE$ minimal est

$$AMISE_{opt} = \frac{5}{4}[\sigma_K R(K)]^{4/5} R(f'')^{1/5} \frac{1}{n^{4/5}}.$$

L'estimateur à noyau et le polygone de fréquence ont donc la même vitesse de convergence mesurée par l'AMISE, soit $n^{-4/5}$.

Parmi tous les noyaux K qui sont des densités, on peut montrer que la valeur minimale de $\sigma_K R(K)$ est atteinte pour le noyau

$$K(u) = \begin{cases} \frac{3}{4}(1-u^2) & |u| \leq 1 \\ 0 & \text{sinon,} \end{cases}$$

fonction appelée *noyau d'Epanechnikov*. Pour ce noyau, la valeur minimale atteinte est $3/(5\sqrt{5})$. Il est alors naturel de mesurer l'efficacité relative d'un noyau K en prenant

$$\frac{\sigma_K R(K)}{3/(5\sqrt{5})}.$$

Le tableau suivant donne l'efficacité relative des noyaux les plus fréquemment utilisés. À l'exception du noyau gaussien, tous ces noyaux valent 0 en dehors de l'intervalle $[-1, 1]$. Comme ces efficacités relatives sont très rapprochées, on est amené à choisir K en fonction de la facilité de calcul plutôt que de l'efficacité relative.

TAB. 4.1 –

Noyau	Forme	Efficacité relative
Epanechnikov	$\frac{3}{4}(1-u^2)$	1
'Biweight'	$\frac{15}{16}(1-u^2)^2$	1.0061
Triangulaire	$1-\lvert t\rvert$	1.0143
Gaussien	$(2\pi)^{-1/2}e^{-u^2/2}$	1.0513
Uniforme	$\frac{1}{2}$	1.0758
Cosinus	$(1+\cos \pi u)/2$	1.0104

4.3.1 Choix du paramètre de lissage en pratique

Comme pour l'histogramme et le polygone de fréquence, une méthode consiste à minimiser en h l'$AMISE$ défini par (4.11). Comme cette expression dépend de l'inconnue f, une approche simple consiste à l'évaluer en prenant pour f la densité de la loi normale $N(\mu, \sigma^2)$. Pour le noyau K gaussien, on obtient ainsi

$$h_{opt} = \left(\frac{4}{3n}\right)^{1/5} \sigma \approx 1.059\,\sigma \frac{1}{n^{1/5}}. \tag{4.12}$$

En estimant σ par l'écart type échantillonnal S, on obtient la règle suivante.

Règle de Scott pour le noyau gaussien

Prendre

$$\boxed{\widehat{h}_{opt} = 1.059 S \frac{1}{n^{1/5}}.}$$

On peut aussi estimer σ avec un estimateur robuste tel que l'écart interquartile $IQ = Q_3 - Q_1$. Pour une loi normale standard, l'écart interquartile

$$R = \Phi^{-1}(.75) - \Phi^{-1}(.25) \approx 1.35.$$

Pour une variable normale X de variance σ^2, on vérifie sans peine que l'écart interquartile vaut

$$F_X^{-1}(.75) - F_X^{-1}(.25) = \sigma R.$$

Cela conduit à estimer σ par IQ/R, d'où la règle

$$\boxed{\widehat{h}_{opt} = 1.059\,\frac{IQ}{R}\frac{1}{n^{1/5}} = 0.79 IQ \frac{1}{n^{1/5}}.} \tag{4.13}$$

Selon Silverman, si (4.13) est satisfaisant pour des densités présentant de l'asymétrie ou des valeurs extrêmes, ce paramètre de lissage a l'inconvénient de provoquer du surlissage pour des densités bimodales. Pour estimer la

densité de manière robuste tout en remédiant au problème de surlissage, Silverman propose de réduire le coefficient 1.059 de l'équation (4.12) à 0.9 et suggère la règle suivante.

Règle de Silverman

Prendre

$$\boxed{\widehat{h}_{opt} = 0.9A\frac{1}{n^{1/5}},}$$

où

$$A = \min\{S, IQ/1.35\}.$$

Il s'agit de la règle utilisée par défaut dans le progiciel R.

4.4 Choix de h par validation croisée

Les idées sous-jacentes à la validation croisée ont été décrites par Stone en 1974. Essentiellement, cette méthode sert à sélectionner un modèle parmi plusieurs, en évaluant un critère d'ajustement approprié à partir de certains sous-ensembles des observations. La technique est fréquemment utilisée entre autres en régression et en classification. Nous appliquerons ici la validation croisée au choix du paramètre de lissage lorsque l'on estime une densité par un estimateur à noyau. Nous verrons que la méthode est étroitement apparentée au jackknife dans sa façon d'utiliser les données pour l'évaluation du critère.

Dans la section précédente, on identifie le h optimal au h obtenu en (4.12) en supposant f gaussienne, une hypothèse rarement réaliste. Plus haut, la moyenne de l'erreur quadratique intégrée (le $MISE$) est présentée comme une mesure globale de précision de $\widehat{f}$. Au lieu de choisir h_{opt} à partir de l'$AMISE(h)$, on est tenté de le faire à partir de

$$MISE(h) = E_F\left[\int [\widehat{f}(t) - f(t)]^2 dt\right].$$

Il s'agit du point de vue adopté par la validation croisée. En fait, comme $MISE(h)$ dépend de f inconnue, la validation croisée visera plutôt à minimiser un estimateur sans biais de $MISE(h)$.

En premier lieu, on peut voir que

$$\begin{aligned} MISE(h) &= E_F\left[\int [\widehat{f}(t) - f(t)]^2 dt\right] \\ &= \int f(t)^2 dt + E_F\left[\int \widehat{f}(t)^2 dt\right] - 2E_F\left[\int \widehat{f}(t) f(t) dt\right] \end{aligned} \quad (4.14)$$

Comme $\int f(t)^2 dt$ est une constante indépendante de h, il suffira de minimiser un estimateur des deux derniers termes.

En tout point t, observons que $\widehat{f}(t)$ est une variable aléatoire fonction de $X_1, \ldots, X_n$. Pour bien le marquer, on notera son espérance $E_F[\widehat{f}(t)] = E_{X_1,\ldots,X_n}[\widehat{f}(t)]$. Suivant la même logique, pour Y indépendante des X_i, on peut voir que

$$\begin{aligned} E_F\left[\int \widehat{f}(t) f(t) dt\right] = \int E_F[\widehat{f}(t)] f(t) dt &= \int E_{X_1,\ldots,X_n}[\widehat{f}(t)] f(t) dt \\ &= E[E_{X_1,\ldots,X_n}[\widehat{f}(Y)|Y]]. \\ &= E_F[\widehat{f}(Y)] \\ &= E\left[E\left[\frac{1}{nh}\sum_i K\left(\frac{Y - X_i}{h}\right)\middle| Y\right]\right] \\ &= E\left[E\left[\frac{1}{h} K\left(\frac{Y - X}{h}\right)\middle| Y\right]\right]. \end{aligned} \quad (4.15)$$

Nous avons ici un exemple d'application de l'identité exprimant l'espérance d'une variable aléatoire W (ici $\widehat{f}(Y)$, fonction de $X_1, \ldots, X_n, Y$) en termes de l'espérance conditionnelle de cette variable par rapport à une autre variable (ici Y) :

$$E[W] = E[E[W|Y]].$$

Le membre gauche de l'équation (4.15) s'obtient donc en calculant successivement une espérance par rapport aux variables $X_1, \ldots, X_n$ pour Y fixée,

ensuite une espérance par rapport à la variable Y, où les variables X_i et Y sont indépendantes et de même loi F.

Nous allons estimer l'espérance du membre gauche de l'équation (4.15) en faisant appel au principe de la validation croisée. Pour $i = 1, \ldots, n$, posons

$$\widehat{f}_{-i}(t) = \frac{1}{(n-1)h} \sum_{j \neq i} K\left(\frac{t - X_j}{h}\right).$$

Noter que $\widehat{f}_{-i}(t)$ est l'estimateur à noyau de f basé sur les observations X_j, $j \neq i$. Il est clair que

$$E[\widehat{f}_{-i}(t)] = E\left[\frac{1}{h} K\left(\frac{t - X}{h}\right)\right] = E\left[\frac{1}{nh} \sum_i K\left(\frac{t - X_i}{h}\right)\right] = E_{X_1,\ldots,X_n}[\widehat{f}(t)],$$

d'où, pour Y indépendante de $X_1, \ldots, X_n$ mais de même loi,

$$\begin{aligned} E[\widehat{f}_{-i}(X_i)] = E[E[\widehat{f}_{-i}(X_i)|X_i]] &= E[E[\widehat{f}_{-i}(Y)|Y]] \\ &= E[E_{X_1,\ldots,X_n}[\widehat{f}(Y)|Y]] \end{aligned}$$

Il en résulte que $\sum_{i=1}^n \widehat{f}_{-i}(X_i)/n$ estime sans biais

$$E\left[\int \widehat{f}(t) f(t) dt\right] = E[E_{X_1,\ldots,X_n}[\widehat{f}(Y)|Y]].$$

Remarque 2. On peut se demander pourquoi on ne pourrait pas estimer $E\left[\int \widehat{f}(t) f(t) dt\right] = E[\widehat{f}(Y)]$ par

$$\int \widehat{f}(y) d\widehat{F}(y) = \frac{1}{n} \sum_i \widehat{f}(X_i).$$

On ne le fait pas parce que cet estimateur est biaisé. En fait, pour tout i,

$$\widehat{f}(X_i) = \frac{1}{nh} K(0) + \frac{1}{nh} \sum_{j \neq i} K\left(\frac{X_i - X_j}{h}\right) \geq \frac{1}{nh} K(0),$$

alors qu'en général $\widehat{f}(Y)$ n'a pas de borne inférieure strictement positive.

Le raisonnement qui précède est un exemple d'application du principe de la validation croisée, dans laquelle chacun des sous-échantillons de $n-1$ observations est utilisé pour obtenir de l'information sur l'observation restante : plus précisément, pour tout i, les observations X_j, $j \neq i$, sont utilisées pour estimer $f(X_i)$.

Le deuxième terme de (4.14) s'estime sans biais par $\int \widehat{f}(t)^2 dt$. On définira donc le h optimal comme étant un point minimum de

$$VC(h) \quad = \quad \int \widehat{f}(t)^2 dt - \frac{2}{n} \sum_{i=1}^{n} \widehat{f}_{-i}(X_i). \tag{4.16}$$

La fonction $VC(h)$ définit le critère de *validation croisée dite sans biais*. Puisque

$$MISE(h) = E[R(f) + VC(h)],$$

on dit aussi que cette validation croisée est celle des moindres carrés.

Pour le calcul de (4.16), on peut utiliser l'approximation

$$VC(h) \approx \frac{\sum_i \sum_j K^* \left(\frac{X_i - X_j}{h} \right)}{n^2 h} + 2 \frac{K(0)}{nh},$$

où $K^*(x) = K^{(2)}(x) - 2K(x)$ et $K^{(2)}(x) = \int K(x-y)K(y)dy$ est le produit de convolution de K avec lui-même. Dans le cas particulier où K est le noyau normal, $K^{(2)}$ est la densité $N(0, 2)$.

Des études ont montré que la méthode de la validation croisée sans biais produit un h très variable. Un autre défaut de cette méthode est qu'elle produit parfois une fonction $VC(h)$ ayant plus d'un minimum local.

4.5 Méthode de Sheather-Jones

On se souvient que l'AMISE de l'estimateur à noyau est minimisé en

$$h_{opt} = \left[\frac{R(K)}{\sigma_K^4 R(f'')} \right]^{1/5} \frac{1}{n^{1/5}}.$$

Comme f'' est inconnue, Sheather et Jones (1991) proposent d'utiliser la méthode du noyau pour estimer $R(f'')$. Pour ce faire, on note que pour un noyau différentiable deux fois L et un paramètre de lissage h_0,

$$\begin{aligned}\widehat{f}''(x) &= \frac{d^2}{dx^2}\left\{\frac{1}{nh_0}\sum_i L\left(\frac{x-X_i}{h_0}\right)\right\} \\ &= \frac{1}{nh_0^3}\sum_i L''\left(\frac{x-X_i}{h_0}\right).\end{aligned}$$

Dans le présent contexte, il est important de noter que l'estimation optimale de f'' ne se fait pas nécessairement avec le même paramètre de lissage optimal que l'estimation de f. Dans le cas où $K = L$ est le noyau gaussien, Sheather et Jones suggèrent une stratégie en deux étapes pour estimer le $\widehat{h}_{opt}$ de $\widehat{f}$. À la première étape, on choisit un $h_0 \propto n^{-1/7}$ (proportionnel) conduisant à $\widehat{f}''$; à la seconde étape, on calcule

$$\widehat{h}_{opt} = \left[\frac{R(K)}{\sigma_K^4 R(\widehat{f}'')}\right]^{1/5} \frac{1}{n^{1/5}}.$$

On trouvera les détails de ces calculs dans la bibliographie de ce chapitre.

Entre toutes les méthodes de sélection du h optimal actuellement disponibles, la méthode de Sheather-Jones est considérée parmi les meilleures. Entre autres avantages, cette méthode produit un h moins variable que la validation croisée.

4.6 Extension à plusieurs dimensions

Soient $x_1, \ldots, x_n$ des observations indépendantes d'un *vecteur aléatoire* de dimension d de densité inconnue f. On définit un estimateur à noyau d-dimensionnel de f par

$$\widehat{f}(y) = \frac{1}{n|H|}\sum_1^n K_d(H^{-1}(y - x_i)),$$

où H est une matrice symétrique $d \times d$ définie positive, $|H|$ son déterminant et K_d un noyau d-dimensionnel prenant, par exemple, la forme d'une densité.

Une manière simple de générer un noyau d-dimensionnel est de prendre un noyau unidimensionnel K et de définir le *noyau produit* :

$$K_d(u_1, \ldots, u_d) = \prod_1^d K(u_i).$$

La matrice H contient en général $d(d+1)/2$ différents paramètres de lissage qu'il nous faut déterminer de façon optimale. Parce qu'il y a généralement trop peu d'observations pour obtenir une estimation assez précise, on suppose la plupart du temps que H est diagonale, ce qui réduit à d le nombre de paramètres à déterminer.

Pour simplifier l'exposé, nous nous limiterons ici au cas $d = 2$. Dans cette situation, le noyau le plus couramment utilisé est le noyau normal produit :

$$K_2(u_1, u_2) = \frac{1}{2\pi} e^{-u_1^2/2} e^{-u_2^2/2}.$$

Lorsque $H = \text{diag}(h_1, h_2)$, l'estimateur à noyau prend alors la forme

$$\widehat{f}(y) = \widehat{f}(y_1, y_2) = \frac{1}{2\pi n h_1 h_2} \sum_1^n \exp\left(-\frac{(y_1 - x_{i1})^2}{2h_1}\right) \exp\left(-\frac{(y_2 - x_{i2})^2}{2h_2}\right).$$

Pour le noyau précédent et $H = \text{diag}(h_1, h_2)$, on pourra encore sélectionner les valeurs optimales de h_1 et h_2 en minimisant l'AMISE correspondant. Une version bidimensionnelle de la règle de Scott en découle par substitution à f de la densité normale à deux composantes indépendantes de variances σ_1^2 et σ_2^2. En conséquence, on peut vérifier que la règle de Scott bidimensionnelle prend la forme

$$\boxed{\widehat{H}_{opt} = \text{diag}\left(n^{-1/6} S_1, n^{-1/6} S_2\right),}$$

où S_i est une estimation de l'écart type σ_i basée sur les i^e composantes de l'échantillon, $i = 1, 2$. Il existe également une règle basée sur la validation croisée mais nous ne la considérerons pas ici.

4.7 Propriétés de convergence de l'estimateur à noyau

4.7.1 Convergence ponctuelle

Rosenblatt et Parzen ont été les initiateurs de la méthode du noyau. Le résultat suivant est dû à Parzen (1962).

Théorème 4.7.1. *Soit K un noyau borné vérifiant les conditions suivantes :*

1. $\int |K(u)|du < \infty$;
2. $\int K(u)du = 1$;
3. $|uK(u)| \to 0$ *lorsque* $|u| \to \infty$.

Soit x un point de continuité de f. Pour que $\widehat{f}(x) \to f(x)$ en probabilité, il suffit que $h_n \to 0$ et $nh_n \to \infty$ lorsque $n \to \infty$.

4.7.2 Convergence uniforme

Le principal résultat est dû à Bertrand-Retali (1978).

Théorème 4.7.2. *Soit K un noyau borné vérifiant les conditions suivantes :*

1. *K est à variation bornée (par exemple une densité unimodale) ;*
2. *K est continu presque partout ;*
3. *K vérifie les conditions 1 et 2 de Parzen.*

Supposons que f soit uniformément continue. Pour que

$$\sup_x |\widehat{f}(x) - f(x)| \to 0, \quad n \to \infty$$

avec probabilité 1 (convergence presque sûre), il suffit que $h_n \to 0$ et que $nh_n / \log n \to \infty$.

4.8 Bibliographie

– Simonoff, J. S. (1996). *Smoothing Methods in Statistics*, Springer-Verlag, New York. Chapitres 1–3. Théorie et pratique. Excellent.

– Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, New York. Chapitres 1–3. Plus théorique que pratique. Bonne référence un peu ancienne.
– Wand, M. P. et Jones, M. C. (1995). *Kernel Smoothing*, Chapman and Hall, New York. Chapitres 1–3. Théorie. Excellente référence.
– Scott, D. W. (1992). *Multivariate Density Estimation*, J. Wiley and Sons, New York. Chapitres 2–4, 6. Théorie avancée. Excellente référence.
– Bowman, A. W. et Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis. The Kernel Approach with S-Plus Illustrations*, Oxford Science Publications, New York. Chapitres 1–2. Accent sur la pratique. Bonne référence.

Chapitre 5

Une introduction à la régression non paramétrique

5.1 La régression à noyau

Les modèles de régression servent à représenter mathématiquement la relation entre une variable aléatoire Y et un ensemble de prédicteurs $\mathbf{X}$. On les utilise souvent pour prédire Y en fonction de certaines valeurs des prédicteurs.

Le modèle de régression le plus élémentaire est le modèle de régression linéaire simple contenant un seul prédicteur X. Pour des données $(x_1, y_1), \ldots, (x_n, y_n)$, ce modèle énonce que

$$y_i \quad = \quad \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n, \tag{5.1}$$

pour deux constantes, β_0, β_1, où l'on suppose généralement que les ϵ_i sont iid de moyenne nulle. À partir des estimateurs des moindres carrés $\widehat{\beta}_0$ et $\widehat{\beta}_1$, on obtient l'équation de prédiction de la réponse Y en $X = x$:

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 x.$$

Lorsque la relation entre Y et X est non linéaire, le modèle (5.1) est insatisfaisant. On s'intéressera naturellement au modèle plus général

$$y_i \quad = \quad m(x_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

où $m(x)$ s'appelle la fonction de régression. Lorsque $E(\epsilon|X=x)=0$, on voit que $m(x)=E(Y|X=x)$. Lorsqu'on ne suppose pas une forme paramétrique pour m, on parlera d'une fonction de *régression non paramétrique*.

Pour construire un estimateur de la fonction de régression m, faisons l'hypothèse que (X,Y) possède une densité jointe f. Si $f(y|x)$ désigne la densité conditionnelle de Y étant donné $X=x$, on peut écrire

$$\begin{aligned} m(x) &= E(Y|X=x) \\ &= \int y f(y|x)dy \\ &= \int y\,\frac{f(x,y)}{f_X(x)}dy, \end{aligned} \tag{5.2}$$

où $f_X(x)$ est la densité marginale de X et $f(x,y)$ est la densité jointe de X et Y.

Supposons maintenant que l'on veuille estimer $m(x_0)$. La méthode du noyau permet d'estimer $f(x_0,y)$ et $f_X(x_0)$, donc (5.2) suggère l'estimateur $\int y\left[\widehat{f}(x_0,y)/\widehat{f}_X(x_0)\right]dy$. Étant donné deux noyaux K_x et K_y, deux estimateurs à noyau sont

$$\widehat{f}(x_0,y)=\frac{1}{nh_xh_y}\sum_1^n K_x\left(\frac{x_0-x_i}{h_x}\right)K_y\left(\frac{y-y_i}{h_y}\right)$$

et

$$\widehat{f}_X(x_0)=\frac{1}{nh_x}\sum_1^n K_x\left(\frac{x_0-x_i}{h_x}\right).$$

Comme plus haut, supposons que $\int K_y(u)du=1$ et $\int uK_y(u)du=0$. En faisant le changement de variable $u_i=(y-y_i)/h_y$, on peut évaluer notre estimateur comme suit :

$$\begin{aligned} \int y\,\frac{\widehat{f}(x_0,y)}{\widehat{f}_X(x_0)}dy &= \frac{\sum_1^n K_x\left(\frac{x_0-x_i}{h_x}\right)\frac{1}{h_y}\int yK_y\left(\frac{y-y_i}{h_y}\right)dy}{\sum_1^n K_x\left(\frac{x_0-x_i}{h_x}\right)} \\ &= \frac{\sum_1^n K_x\left(\frac{x_0-x_i}{h_x}\right)y_i}{\sum_1^n K_x\left(\frac{x_0-x_i}{h_x}\right)}. \end{aligned}$$

L'estimateur obtenu $\widehat{m}_{NW}(x_0)$ s'appelle l'*estimateur à noyau de Nadaraya-Watson* (1964). On voit que celui-ci est une fonction linéaire des y_i, en fait la moyenne pondérée

$$\widehat{m}_{NW}(x_0) = \sum_{i=1}^{n} w_i y_i,$$

où

$$\begin{aligned} w_i &= \frac{K_x\left(\frac{x_0-x_i}{h_x}\right)}{\sum_1^n K_x\left(\frac{x_0-x_i}{h_x}\right)} \\ &= \frac{1}{nh_x}\frac{K_x\left(\frac{x_0-x_i}{h_x}\right)}{\widehat{f}_X(x_0)}. \end{aligned}$$

Pour x_0 fixe et K_x prenant ses valeurs les plus grandes dans le voisinage de 0, cette expression montre que les poids w_i les plus importants sont ceux associés aux x_i proches de de x_0. On se souviendra que l'estimateur à noyau d'une densité possède la même propriété. À l'opposé, on notera que l'ajustement en un point d'une fonction de régression polynomiale (modèle paramétrique) pourrait fortement dépendre d'observations très éloignées dudit point. Comme fonction de x_0, on observe enfin que $\widehat{m}_{NW}(x_0)$ hérite des propriétés de différentiabilité de $K = K_x$.

5.2 La régression polynomiale locale

Il n'est pas difficile de voir que $\widehat{m}_{NW}(x_0)$ est la solution en β_0 du problème de moindres carrés pondérés

$$\min_{\beta_0} \sum_{i=1}^{n} (y_i - \beta_0)^2 \frac{1}{h} K\left(\frac{x_0 - x_i}{h}\right). \tag{5.3}$$

[Exercice : vérifier cette affirmation en dérivant l'expression précédente par rapport à β_0 et en égalant la dérivée à 0.] Au sens des moindres carrés pondérés, on peut donc dire que $\widehat{m}_{NW}(x_0)$ est la constante s'ajustant le

mieux aux y_i lorsque le poids de y_i est $\frac{1}{h}K\left(\frac{x_0-x_i}{h}\right)$, $i=1,\ldots,n$. On voit ainsi que l'estimation de cette constante est donc davantage influencée par les (x_i,y_i) tels que x_i est proche de x_0. En outre, plus h est grand plus l'influence des points x_i éloignés de x_0 se fera sentir, et inversement lorsque h est petit. C'est en ce sens que l'on parle ici de *régression locale*.

Nous verrons que l'estimateur de Nadaraya-Watson peut être fortement biaisé près des extrémités de l'intervalle où se trouvent les x_i. Le problème de minimisation (5.3) suggère qu'il pourrait être préférable d'ajuster localement (i.e. en chaque point x_0) un polynôme de degré $p\geq 1$ plutôt qu'une constante β_0. Au point x_0, on cherchera donc le polynôme $\beta_0+\beta_1(x_0-x)+\cdots+\beta_p(x_0-x)^p$ minimisant en $\beta_0,\beta_1,\ldots,\beta_p$ l'expression

$$\sum_{i=1}^{n}\left[y_i-\beta_0-\beta_1(x_0-x_i)-\cdots-\beta_p(x_0-x_i)^p\right]^2\frac{1}{h}K\left(\frac{x_0-x_i}{h}\right). \qquad (5.4)$$

Le problème de minimisation en est un des moindres carrés pondérés avec la matrice de poids diagonale

$$W=\operatorname{diag}\left[\frac{1}{h}K\left(\frac{x_0-x_1}{h}\right),\ldots,\frac{1}{h}K\left(\frac{x_0-x_n}{h}\right)\right].$$

Si

$$X=\begin{pmatrix}1 & x_0-x_1 & & (x_0-x_1)^p\\ \vdots & \vdots & \cdots & \vdots\\ 1 & x_0-x_n & & (x_0-x_n)^p\end{pmatrix},$$

la somme à minimiser (5.4) peut en effet écrire

$$(\mathbf{y}-X\boldsymbol{\beta})'W(\mathbf{y}-X\boldsymbol{\beta}).$$

Lorsque $X'WX$ est inversible, la solution obtenue par dérivation est

$$\widehat{\boldsymbol{\beta}}=\begin{pmatrix}\widehat{\beta}_0\\ \widehat{\beta}_1\\ \cdots\\ \widehat{\beta}_p\end{pmatrix}=(X'WX)^{-1}X'W\mathbf{y}. \qquad (5.5)$$

On estime alors $m(x_0)$ par $\widehat{m}_p(x_0) = \widehat{\beta}_0$, soit la valeur du polynôme estimé

$$\sum_{i=0}^{p} \widehat{\beta}_i (x_0 - x)^i$$

en $x = x_0$. (Noter que ce $\widehat{\beta}_0$ dépend de x_0, de chacun des (x_i, y_i), de K et de p.) Pour tout x_0, il suit de (5.5) que $\widehat{m}_p(x_0)$ est une combinaison linéaire des y_i. Il en résulte que l'on peut écrire

$$(\widehat{m}_p(x_1), \ldots, \widehat{m}_p(x_n))' = S_h \mathbf{y}, \tag{5.6}$$

pour une matrice de lissage S_h carrée d'ordre n jouant le rôle de la matrice chapeau rencontrée en régression linéaire. (Dans le cas présent, contrairement à la régression linéaire, on notera que S_h n'est pas idempotente : $S_h^2 \neq S_h$.) En raison de la propriété précédente, on dit que le lissage par régression polynomiale locale est linéaire.

Dans la suite, nous nous intéresserons surtout aux cas $p = 0$ (estimateur de Nadaraya-Watson) et $p = 1$ (estimateur dit *linéaire local*). On peut montrer que ce dernier peut s'écrire

$$\widehat{m}_1(x_0) = \frac{1}{nh} \sum_{1}^{n} \frac{[\widehat{s}_2(x_0, h) - \widehat{s}_1(x_0, h)(x_0 - x_i)] K[(x_0 - x_i)/h] y_i}{\widehat{s}_2(x_0, h)\widehat{s}_0(x_0, h) - \widehat{s}_1(x_0, h)^2},$$

où

$$\widehat{s}_r(x_0, h) = \frac{1}{nh} \sum_{1}^{n} (x_0 - x_i)^r K\left(\frac{x_0 - x_i}{h}\right).$$

Le paramètre de lissage h joue le même rôle qu'en estimation de densité. Une valeur appropriée de h fait en sorte que l'estimateur montre bien les caractéristiques importantes de la relation entre x et y tout en n'étant pas trop influencé par les variations locales. La figure 5.1 donne un exemple d'un lissage approprié représenté par la courbe continue ($h = 0.022$) et d'un surlissage appliqué aux mêmes données et représenté par la courbe pointillée ($h = 0.052$).

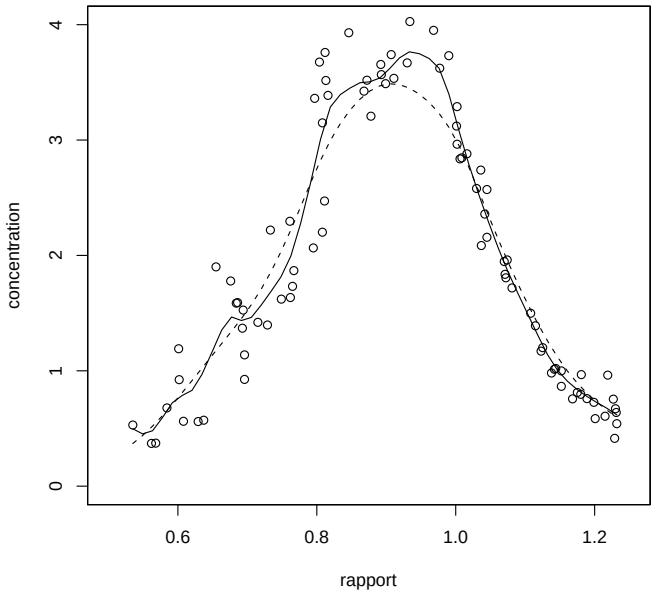

FIG. 5.1 – Effet du paramètre de lissage sur une régression linéaire locale. Courbe continue : $h = 0.022$. Courbe pointillée : $h = 0.052$.

5.3 Choix du paramètre de lissage

Les méthodes sont semblables à celles qu'on utilise en estimation de densité. L'une de celles-ci est basée sur le comportement asymptotique du biais et de la variance mesuré par l'AMISE. *Supposons que les x_i proviennent d'une loi de densité f_X et que* $\mathrm{Var}(Y|X = x) = \sigma^2$, quel que soit x. Lorsque x_0 est éloigné du minimum ou du maximum des x_i, on montre que pour la régression linéaire locale

$$\mathrm{biais}(\widehat{m}_1(x_0)) = \frac{1}{2} m''(x_0)\sigma_K^2 h^2 + o(h^2) = O(h^2)$$

et

$$\mathrm{Var}(\widehat{m}_1(x_0)) = \frac{\sigma^2 R(K)}{nhf_X(x_0)} + o((nh)^{-1}).$$

La notation $o(h^2)$ représente ici une fonction de h telle que $\lim_{h\to 0} o(h^2)/h^2 = 0$; de même $o((nh)^{-1})$ représente une fonction de n et h telle que

$$\lim_{nh\to\infty} \frac{o((nh)^{-1})}{(nh)^{-1}} = \lim_{nh\to\infty} nh\, o((nh)^{-1}) = 0.$$

Des formules tout à fait semblables valent pour $\widehat{m}_0 \equiv \widehat{m}_{NW}$.

Ces résultats montrent que le biais de ces estimateurs est le plus important là où m a une forte courbure ($m''(x_0)$ grand en valeur absolue) ; en outre, ces estimateurs ont leur plus grande variabilité là où $f(x_0)$ est petit (là où les observations sont peu nombreuses).

Lorsque x_0 est proche des x_i extrêmes, on peut montrer que biais$(\widehat{m}_0(x_0))$ $= O(h)$ alors qu'on a encore biais$(\widehat{m}_1(x_0)) = O(h^2)$. Pour leur part, les variances des deux estimateurs restent toutes deux d'ordre $(nh)^{-1}$. Ces résultats font considérer l'estimateur linéaire local comme supérieur à l'estimateur de Nadaraya-Watson. La figure 5.2 illustre ce comportement de $\widehat{m}_0$ et $\widehat{m}_1$ pour le jeu de données `elusage.dat` et $h = 9$.

Comme en estimation de densité, il est possible en principe de sélectionner le paramètre de lissage optimal de $\widehat{m}_1$ en minimisant

$$AMISE(h) \;=\; \left[\frac{\sigma_K^2 h^2}{2}\right]^2 \int m''(u)^2 f_X(u)du + \frac{\sigma^2 R(K)}{nh},$$

où l'on a supposé que les X_i sont iid de densité f_X. Le minimum est atteint en

$$h_{opt} = \left[\frac{\sigma^2 R(K)}{n\sigma_K^4 \int m''(u)^2 f_X(u)du}\right]^{1/5}, \tag{5.7}$$

où σ^2 et m'' doivent être estimées. Le package `KernSmooth` inclut la fonction `dpill` calculant une valeur approximative de h pour le noyau gaussien K et une estimation préalable de σ^2 et $\int m''(u)^2 f_X(u)du$. Nous nommerons cette approche la *méthode de substitution*.

La validation croisée (Stone 1974) offre une autre approche à la sélection du h optimal. Pour celle-ci, on cherche à minimiser

$$VC(h) = \sum_1^n (y_i - \widehat{m}_p^{(i)}(x_i))^2, \tag{5.8}$$

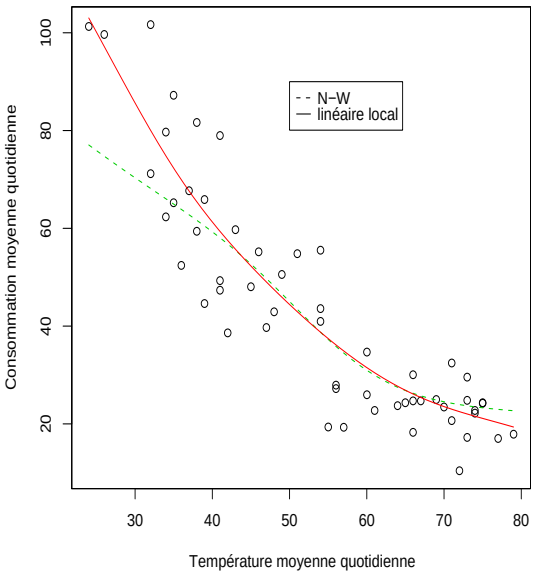

FIG. 5.2 – Comportements différents des estimateurs de Nadaraya-Watson et linéaire local dans la région des valeurs extrêmes de x. Jeu de données `elusage.dat`, $h = 9$.

où $\widehat{m}_p^{(i)}(x_i))$ est l'estimation de $m_p(x_i)$ basée sur toutes les données exceptée (x_i, y_i). De ce point de vue, le paramètre de lissage optimal est $\widehat{h}_{VC}$, point minimal de $VC(h)$. Comme en estimation de densité, $\widehat{h}_{VC}$ a le défaut de posséder une variance relativement grande et tend à produire du sous-lissage. En outre, son calcul est relativement complexe.

À première vue, (5.8) requiert n ajustements. En pratique, à cause de l'identité

$$y_i - \widehat{m}_p^{(i)}(x_i) = \frac{y_i - \widehat{m}_p(x_i)}{1 - S_{ii}(h)}$$

où $S_h = (S_{ij}(h))$ est la matrice de lissage, on pourra calculer plus simplement

$$VC(h) = \sum_1^n \left(\frac{y_i - \widehat{m}_p(x_i)}{1 - S_{ii}(h)} \right)^2 .$$

La vaidation croisée s'avère donc plus facile à appliquer en régression polynomiale locale qu'en estimation de densité.

Appliquées aux données `elusage.dat`, les deux méthodes précédentes donnent $\widehat{h}_{VC} = 7.46$ et $h = 5.07$. La figure 5.3 permet de constater que cette différence a peu d'effet sur l'estimateur local linéaire.

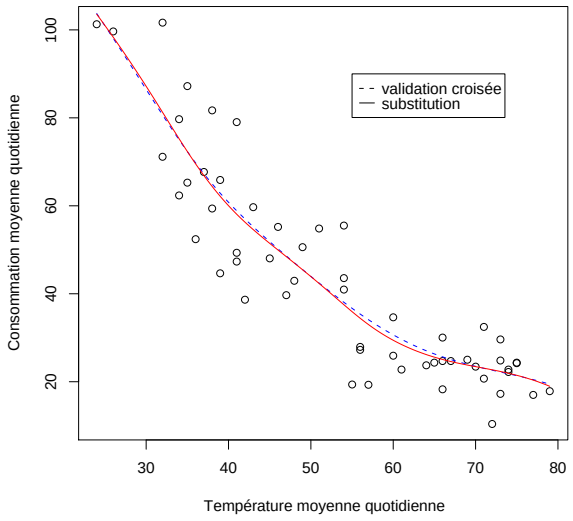

FIG. 5.3 – Optimisation par validation croisée et par substitution ($p = 1$).

5.4 La régression spline

On peut mesurer la qualité de l'ajustement d'une courbe g à des observations (x_i, y_i), $i = 1, \ldots, n$, en calculant la somme de carrés

$$\sum_{1}^{n} (y_i - g(x_i))^2.$$

Il existe évidemment une infinité de fonctions g annulant cette somme : il suffit qu'on ait $g(x_i) = y_i$ pour tout i, une condition facile à réaliser. Une telle fonction s'appelle une *fonction d'interpolation*. Par exemple, on peut

montrer qu'il existe toujours un polynôme de degré $n-1$ passant par chacun des points (x_i, y_i) (polynôme de Lagrange). Ce polynôme est donné par la formule

$$g(x) = \sum_{i=1}^{n} y_i \frac{\prod_{j \neq i}(x - x_j)}{\prod_{j \neq i}(x_i - x_j)}.$$

Une telle solution est en général trop irrégulière pour modéliser la relation entre x et y de façon satisfaisante. On souhaite plutôt un estimateur de régression qui s'ajuste bien aux données tout en étant lisse.

La régression spline ou lissage par splines aborde ce problème en mesurant la qualité de l'ajustement par une expression de la forme

$$L(g) = \sum_{1}^{n} (y_i - g(x_i))^2 + \Phi(g),$$

où Φ est une fonction positive mesurant la "lissité" (régularité) d'une fonction : plus g est lisse, plus $\Phi(g)$ est petit. Pour que la solution au problème de minimisation soit unique, il est nécessaire de restreindre la classe des fonctions g. Celles-ci doivent être suffisamment lisses, en pratique dérivabies au moins deux fois. Comme en estimation de densité, on mesurera la lissité d'une fonction g dérivable deux fois à l'aide de l'intégrale $\Phi(g) = \int g''(x)^2 dx$.

Supposons que $a = x_0 < x_1 < x_2 < \cdots < x_n < x_{n+1} = b$. On quantifiera la qualité de l'ajustement et la régularité de g sur $[a, b]$ au moyen de la somme

$$L_\lambda(g) = \sum_{1}^{n} (y_i - g(x_i))^2 + \lambda \int_a^b g''(x)^2 dx,$$

où $\lambda \geq 0$ jouera le rôle d'un paramètre de lissage. Dans la dernière expression, la somme pénalise l'inadéquation de g, l'intégrale pénalise l'irrégularité de g et λ pondère l'importance de ces deux pénalités. Dans la classe des fonctions polynomiales g, deux cas limites méritent d'être soulignés. Lorsque $\lambda = 0$, L_λ atteint la valeur minimale 0 pour n'importe quelle interpolation exacte, autrement dit pour tous les polynômes g tels que $g(x_i) = y_i$, $i = 1, \ldots, n$. Lorsque $\lambda = \infty$, L_λ est infinie, sauf si l'intégrale s'annule, ce qui

ne peut avoir lieu que pour les polynômes g tels que $g''(x) = 0$ dans $[a, b]$, autrement dit les polynômes linéaires sur $[a, b]$; ainsi, $L_\infty(g) = \sum_1^n (y_i - g(x_i))^2$ est minimisée lorsque g est la régression linéaire des moindres carrés.

Pour qu'une fonction lisse minimisant $L_\lambda(g)$ soit unique, il est nécessaire d'imposer des conditions sur les valeurs des dérivées de g. Pour les conditions ci-dessous et λ fixé positif, on peut montrer en résolvant un système d'équations linéaires qu'il existe un minimum unique de $L_\lambda(g)$ dans la classe des fonctions g sur $[a, b]$ *différentiables deux fois telles que* $(g'')^2$ *est intégrable et* g, g' *sont absolument continue* (i.e. exprimables sous forme d'une intégrale de a à b). Le minimum unique est alors un spline cubique appelé *spline de lissage*. Un tel spline S_λ est une fonction coïncidant avec un polynôme cubique p_i sur l'intervalle $[x_i, x_{i+1}]$ et telle qu'aux *nœuds* x_i, $i = 2, \ldots n-1$, on a les relations

$$\begin{aligned} p_i(x_i) &= p_{i-1}(x_i) \quad \text{(continuité de } S_\lambda \text{ aux nœuds)} \\ p_i'(x_i) &= p_{i-1}'(x_i) \quad \text{(continuité de } S_\lambda' \text{ aux nœuds)} \\ p_i''(x_i) &= p_{i-1}''(x_i) \quad \text{(continuité de } S_\lambda'' \text{ aux nœuds),} \end{aligned}$$

avec les égalités $p_1''(x_1) = p_1'''(x_1) = p_n''(x_n) = p_n'''(x_n) = 0$. On exige en outre que $S_\lambda''(x) = S_\lambda'''(x) = 0$ sur $[a, x_1]$ et $[x_n, b]$ S_λ, de sorte que S_λ est linéaire sur ces intervalles. Le spline cubique ainsi défini a une dérivée seconde continue sur $[a, b]$, mais peut ne pas avoir de dérivée troisième aux nœuds. Ce type de spline est connu plus précisément sous le nom de *spline cubique naturel*.

Exemple 1. Une interpolation par spline.

Les données suivantes sont obtenues de l'équation $g(x) = 10/(1 + x^2)$.

j	1	2	3	4	5	6	7
x_j	-3	-2	-1	0	1	2	3
$y_j = g(x_j)$	1	2	5	10	5	2	1

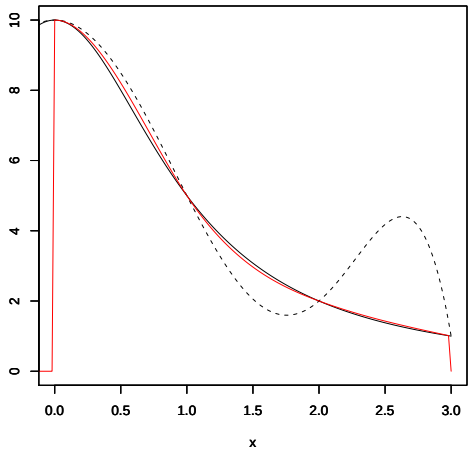

FIG. 5.4 – Comparaison des graphiques de $g(x) = 10/(1+x^2)$ (trait continu noir), du polynôme de Lagrange (en pointillé) et du spline cubique d'interpolation (en rouge).

On peut vérifier que le polynôme d'interpolation de Lagrange est égal à

$$P_L(x) = 10 - 6.4x^2 + 1.5x^4 - 0.1x^6.$$

On utilise en analyse numérique un autre type de fonction d'interpolation : le *spline cubique d'interpolation*. Dans le cas présent, ce spline passant par les points (x_i, y_i) est symétrique par rapport à 0 et s'exprime sur $[0, 3]$ comme suit :

$$S(x) = \begin{cases} p_4(x), & 0 \leq x < 1 \\ p_5(x), & 1 \leq x < 2 \\ p_6(x), & 2 \leq x < 3 \end{cases}$$

où

$$\begin{aligned} p_4(x) &= 5 - 5.65385(x-1) + 3.69231(x-1)^2 + 4.34615(x-1)^3 \\ p_5(x) &= 2 - 1.38462(x-2) + 0.57692(x-2)^2 - 1.03846(x-2)^3 \\ p_6(x) &= 1 - 0.80769(x-3) - 0.19231(x-3)^3 \end{aligned}$$

Sur le graphique de la figure 5.4, on a comparé sur $[0, 3]$ les fonctions g, le polynôme d'interpolation de Lagrance et le spline cubique d'interpolation. On constate que le polynôme d'interpolation s'écarte assez largement de $g(x)$, tandis que le spline cubique d'interpolation en est une excellente approximation.

Les splines de lissage de la régression non paramétrique ne doivent pas être confondus avec les splines d'interpolation utilisés en analyse numérique pour calculer des intégrales ou des dérivées de façon approximative. L'exemple précédent a pour but de montrer qu'il est possible de bien approximer une courbe relativement complexe à partir d'un objet relativement simple à manipuler, le spline cubique. Sans être des splines d'interpolation, les splines de lissage poursuivent le même objectif en régression non paramétrique.

5.4.1 Choix du paramètre de lissage λ

La méthode la plus populaire est sans doute celle de la validation croisée. Pour celle-ci, on recherche la valeur de λ minimisant

$$VC(\lambda) = \sum_{1}^{n} (y_i - S_\lambda^{(i)}(x_i))^2,$$

où $S_\lambda^{(i)}(x_i)$ est le spline cubique minimal évalué en x_i, obtenu en utilisant toutes les observations exceptée x_i.

On peut montrer que le spline de lissage est un estimateur linéaire. On veut dire par là que, si l'on définit le vecteur des valeurs ajustées en posant $\widehat{y}_i = S_\lambda(x_i)$, ce vecteur s'écrit

$$\widehat{\mathbf{y}} = A(\lambda)\mathbf{y},$$

où l'on donne à $A(\lambda)$ le nom de matrice chapeau (par analogie avec la régression paramétrique). On peut alors vérifier que

$$VC(\lambda) = \sum_{1}^{n} \left(\frac{y_i - S_\lambda(x_i)}{1 - A_{ii}(\lambda)} \right)^2,$$

où $A_{ii}(\lambda)$ est le i^e élément diagonal de la matrice chapeau. Par analogie avec la régression paramétrique, la valeur $A_{ii}(\lambda)$ est appelée valeur de levier ; cette valeur mesure l'influence du point (x_i, y_i) sur l'ajustement.

Pour réduire l'influence des points à effet de levier important, on peut aussi choisir le paramètre de lissage en minimisant plutôt la *validation croisée généralisée* (Craven et Wahba 1979). Celle-ci est définie par

$$VCG(\lambda) = \frac{\sum_1^n (y_i - S_\lambda(x_i))^2}{(1 - n^{-1}\mathrm{tr}(A(\lambda)))^2},$$

où $\mathrm{tr}(A(\lambda))$ est la trace de la matrice $A(\lambda)$ (somme des éléments diagonaux). Pour bien comprendre la réduction d'influence des points à effet de levier, on notera que la validation croisée généralisée remplace les poids $1 - A_{ii}(\lambda)$ par leur valeur moyenne $\mathrm{tr}(A(\lambda))/n$. Les valeurs minimales des deux types de validation croisée seront notées $\widehat{\lambda}_{VC}$ et $\widehat{\lambda}_{VCG}$.

5.5 Bibliographie

- Simonoff, J. S. (1996). *Smoothing Methods in Statistics*, Springer-Verlag, New York. Chapitre 5. Excellente référence, présentant le sujet de façon très abordable.
- Ruppert, D., Wand, M. P. et Carroll, R. J. (2003). *Semiparametric Regression*, Cambridge University Press, Cambridge.
- Bowman, A. W. et Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis. The Kernel Approach with S-Plus Illustrations*, Oxford Science Publications, New York. Chapitres 3, 4.

Chapitre 6

Résolution d'équations non linéaires et optimisation

L'application de la méthode du maximum de vraisemblance conduit fréquemment à des équations dont il est impossible de déterminer explicitement la solution. Ce chapitre présente quelques méthodes d'analyse numérique utiles pour résoudre ce type d'équations. Comme la maximisation d'une fonction de vraisemblance équivaut souvent à déterminer la racine d'une équation, on présente d'abord quelques méthodes de résolution d'équations non linéaires.

6.1 Problèmes univariés

6.1.1 Méthode de la bissection ou de Bolzano

On cherche une racine de l'équation $g(x) = 0$, où g est une fonction continue définie sur l'intervalle $[a_0, b_0]$. Supposons que $g(a_0)$ et $g(b_0)$ soient de signes opposés ; par exemple, sans perte de généralité, supposons que $g(a_0) < 0 < g(b_0)$. Un théorème bien connu d'analyse garantit qu'il existe alors un point de $[a_0, b_0]$ où g s'annule.

À la première étape de l'algorithme de la bissection, on calcule $g((a_0 + b_0)/2)$. Lorsque $g((a_0 + b_0)/2) = 0$, $(a_0 + b_0)/2$ est une racine, et la recherche

est terminée. Si l'on a plutôt $g((a_0+b_0)/2) < 0$, il existe comme ci-dessus une racine dans $((a_0 + b_0)/2, b_0)$; dans le cas contraire, il existe une racine dans l'intervalle $(a_0, (a_0 + b_0)/2)$. À l'étape suivante, on poursuit la recherche en coupant en deux l'intervalle de longueur $(b_0 - a_0)/2$ identifié comme contenant une racine ; on vérifie si le point milieu de cet intervalle est une racine, sinon on identifie un sous-intervalle de longueur $(b_0 - a_0)/4$ contenant une racine. On continue d'appliquer ce raisonnement jusqu'à ce qu'on ait identifé un sous-intervalle contenant une racine et de longueur correspondant au degré de précision souhaité.

Formellement, le point de départ de l'algorithme est le calcul de g au point $x_0 = (a_0 + b_0)/2$. À l'étape n, lorsque $g((a_{n-1} + b_{n-1})/2) = 0$, $(a_{n-1} + b_{n-1})/2$ est une racine ; sinon, on retient l'intervalle

$$[a_n, b_n] = \begin{cases} [a_{n-1}, x_{n-1}] & \text{si} \quad g(a_{n-1})g(x_{n-1}) < 0 \\ [x_{n-1}, b_{n-1}] & \text{si} \quad g(a_{n-1})g(x_{n-1}) > 0 \end{cases},$$

et on pose $x_n = (a_n + b_n)/2$. Après n étapes, si une racine n'est pas encore identifiée, on se retrouve avec l'intervalle $[a_n, b_n]$ contenant une racine et ayant la longueur $(b_0 - a_0)/2^n$. Si cette longueur est suffisamment petite, il est naturel d'estimer la racine par le point x_n. Puisque $a_{n-1} \leq a_n \leq b_n \leq b_{n-1}$ pour tout n, on a $\lim a_n = \lim b_n = x_\infty$. Comme $\lim g(a_n)g(b_n) = [g(x_\infty)]^2 \leq 0$, on conclut que $g(x_\infty) = 0$ et le point limite x_∞ est une racine.

Dans la pratique, lorsqu'une itération générale ne réussit pas à identifier la racine exacte, il est nécessaire de se donner une règle d'arrêt. Une première règle est celle de la *convergence absolue* : on s'arrête à l'étape n pour le plus petit n tel que

$$|x_n - x_{n-1}| < \epsilon,$$

où $\epsilon > 0$ mesure le degré de précision souhaité. Une autre règle d'arrêt est celle de la *convergence relative* pour laquelle on s'arrête dès que

$$\frac{|x_n - x_{n-1}|}{|x_{n-1}|} < \epsilon.$$

Ce deuxième critère a l'avantage d'atténuer l'importance des unités dans l'atteinte de la précision souhaitée. Il peut cependant s'avérer instable lorsque x_n est trop proche de 0. (Une solution à ce problème est d'utiliser le critère $\frac{|x_{n+1}-x_n|}{|x_n|+\epsilon} < \epsilon$.) Notons enfin que pour toute itération de ce type il est souhaitable d'imposer une limite au nombre de répétitions.

La méthode de la bissection a l'avantage de pouvoir s'appliquer à toutes les fonctions continues g changeant de signe sur un intervalle $[a, b]$. La méthode a cependant l'inconvénient d'exiger un nombre d'itérations relativement grand (convergence lente).

6.1.2 Méthode de Newton

Faisons l'hypothèse que g est dérivable, et notons par x_∞ une racine de $g(x) = 0$. Supposons que x_0 soit une première approximation de cette racine et que $g'(x_0) \neq 0$. Alors dans le voisinage de x_0, l'approximation linéaire de Taylor

$$0 = g(x_\infty) \approx g(x_0) + (x_\infty - x_0)g'(x_0) \tag{6.1}$$

entraîne que

$$x_\infty \approx x_0 - \frac{g(x_0)}{g'(x_0)}.$$

Cela conduit à penser que

$$x_1 = x_0 - \frac{g(x_0)}{g'(x_0)} \tag{6.2}$$

a des chances d'être plus proche de x_∞ que ne l'est x_0. Notons que l'on peut aussi obtenir x_1 comme solution d'un problème d'optimisation d'une fonction quadratique. En effet, soit G une fonction telle que $G' = g$. Il est alors facile de vérifier que x_1 est l'unique point x annulant la dérivée de l'approximation quadratique de Taylor de $G(x)$ autour de x_0 :

$$G_q(x) = G(x_0) + (x - x_0)g(x_0) + \frac{(x - x_0)^2 g'(x_0)}{2}.$$

En effet $G_q'(x_1) = 0$ et $G_q''(x_1) = g'(x_0)$, d'où x_1 est le minimum de G_q si $g'(x_0) > 0$ et le maximum si $g'(x_0) < 0$.

En substituant x_1 à x_0 dans (6.1) et (6.2), on peut de la même façon obtenir une deuxième approximation x_2 qui, on l'espère, est meilleure que x_1. Plus généralement, dès que $g'(x_n) \neq 0$, on peut définir

$$x_{n+1} = x_n - \frac{g(x_n)}{g'(x_n)},$$

itération définissant la méthode de Newton (ou encore de Newton-Raphson)[1] pour résoudre l'équation $g(x) = 0$. Sous certaines hypothèses, on peut montrer que $x_n \to x_\infty$.

Pour bien comprendre la méthode, il est important de noter que $h(x) = g(x_0) + (x - x_0)g'(x_0)$ est l'équation de la droite de pente $g'(x_0)$ passant par le point $(x_0, g(x_0))$ et que cette droite coupe l'abscisse en x_1. L'exemple 2 ci-dessous en fournit une illustration.

Remarque 3. En statistique, les équations non linéaires à résoudre apparaissent le plus souvent en théorie de l'estimation par la méthode de la vraisemblance maximales où elles prennent la forme $l'(\theta) = 0$, $l(\theta)$ désignant la log-vraisemblance. Lorsque la solution de l'équation précédente est un point maximum global, la racine n'est autre que l'estimateur du maximum de vraisemblance. Dans ce contexte, les équations de l'itération s'écrivent

$$\theta_{n+1} = \theta_n - \frac{l'(\theta_n)}{l''(\theta_n)}.$$

Exemple 2. Les figures 6.1 et 6.2 illustrent une application de la méthode de Newton. La fonction $G(x) = \log x/(1 + x)$ possède un point maximum unique coïncidant avec la racine de

$$g(x) \equiv G'(x) = \frac{x^{-1} + 1 - \log x}{(1 + x)^2} = 0.$$

On peut vérifier que $g(3) = 0.015$ et que $g(4) = -0.005$, et l'on pourra donc chercher la racine entre 3 et 4. Prenant $x_0 = 3$, on peut vérifier que la

[1] Isaac Newton (1642-1727) et Joseph Raphson (1648-1715) étaient deux mathématiciens anglais.

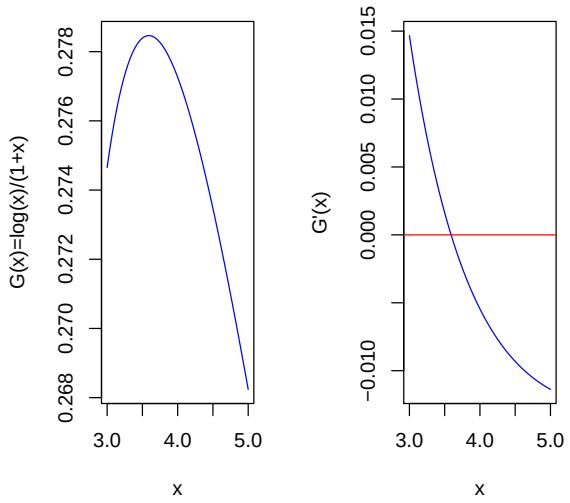

FIG. 6.1 – Graphiques de $G(x) = \log(x)/(1+x)$ et de sa dérivée $g(x) \equiv G'(x) = (x^{-1} + 1 - \log(x))/(1+x)^2$.

racine est identifiée à 6 décimales près après 4 étapes : $x_1 = 3.417798, x_2 = 3.574045, x_3 = 3.590946, x_4 = x_5 = 3.591121$. En comparaison, ce niveau de précision n'est atteint qu'à la 19^e itération avec la méthode de la bissection.

La méthode de Newton ne converge pas toujours. Pour analyser cette convergence, considérons l'erreur de la n^e itération $e_n = x_n - x_\infty$. Supposons que g soit dérivable deux fois et que $g'(x) \neq 0$ dans le voisinage de x_∞. Le développement de Taylor d'ordre deux donne

$$g(x_\infty) = 0 = g(x_n) + (x_\infty - x_n)g'(x_n) + (x_\infty - x_n)^2 g''(t)/2$$

pour un certain t entre x_n et x_∞. Pour x_n près de x_∞, la dernière équation équivaut à

$$x_n - \frac{g(x_n)}{g'(x_n)} - x_\infty = \frac{(x_n - x_\infty)^2 g''(t)}{2g'(x_n)},$$

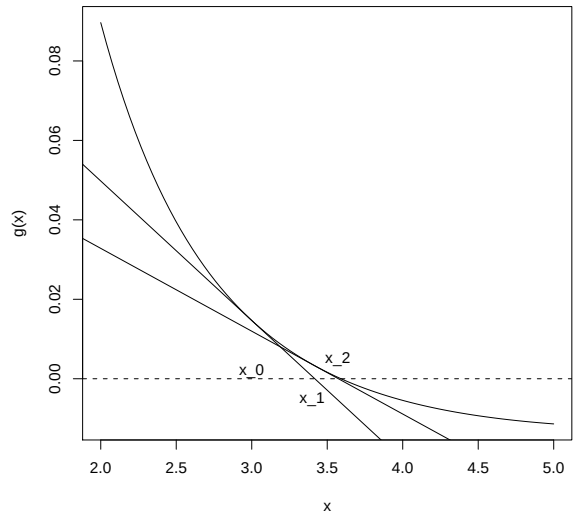

FIG. 6.2 – Méthode de Newton pour la dérivée de $\log(x)/(1+x)$.

laquelle peut s'exprimer sous la forme

$$e_{n+1} = e_n^2 \cdot \frac{g''(t)}{2g'(x_n)}. \tag{6.3}$$

La dernière expression révèle le caractère quadratique de la convergence de la méthode de Newton. (En comparaison, pour la méthode de la bissection on a une convergence linéaire et il se peut que l'on ait $|e_{n+1}] = |e_n|$.) Lorsque $x_n \to x_\infty$ et g'' est continue, on notera que (6.3) implique que

$$\lim_n \frac{e_{n+1}}{e_n^2} = \frac{g''(x_\infty)}{2g'(x_\infty)}.$$

Le taux de convergence dépend de l'importance du terme $g''(t)/2g'(x_n)$. Lorsque la pente g' est grande par rapport à g'' dans le voisinage de x_∞, ce terme est petit et la convergence plus facile. Lorsque la pente g' est proche de 0, l'approximation linéaire de la fonction g en x_n pourra couper l'axe des x en un point très éloigné de x_n, ce qui risque d'entraîner la divergence.

En général, la convergence de la méthode de Newton vers une racine dépend de la position du point de départ x_0. Chaque racine possède en effet son propre domaine d'attraction ; lorsqu'il y a plusieurs racines, cela signifie que l'itération partant de x_0 mène à une racine contenant x_0 dans son domaine d'attraction. En définitive, les nombreux facteurs impliqués dans la convergence font que la méthode de Newton doit être appliquée avec précaution.

Citons sans démonstration le résultat théorique suivant.

Théorème 6.1.1. *Supposons que g ait exactement une racine. Pour que la méthode de Newton converge à partir de n'importe quel point de départ, il suffit que g soit continûment dérivable deux fois et convexe* ($g''(x) > 0$ partout). Un résultat analogue vaut pour g concave ($g''(x) < 0$ partout).

6.2 Problèmes multivariés

6.2.1 Méthode d'optimisation de Nelder-Mead

Dans l'espace $\mathbb{R}^d$, on appelle d-simplexe de sommets $x_0, \ldots, x_d$ l'ensemble des points

$$\left\{ \sum_0^d t_i x_i : \sum_0^d t_i = 1, t_1, \ldots, t_d \geq 0 \right\},$$

où les x_i sont des points distincts de $\mathbb{R}^d$ tels que les différences $x_1 - x_0, \ldots, x_d - x_0$ sont linéairement indépendantes. En dimension 1, un 1-simplexe est un segment de droite aux bornes x_0 et x_1 ; en dimension 2, un 2-simplexe est un triangle aux sommets x_0, x_1 et x_2, tandis qu'en dimension 3 un 3-simplexe est un trièdre aux sommets x_0, x_1, x_2 et x_3. La méthode de Nelder-Mead (1965), ou méthode du simplexe, optimise une fonction f en ne faisant appel qu'aux valeurs de f sur les sommets d'une suite de simplexes. Par rapport à la plupart des méthodes d'optimisation en contexte multivarié elle a l'avantage de ne pas exiger que l'on calcule des dérivées. Nous la décrivons ici en dimension $d = 2$.

Supposons que l'on veuille minimiser f. On commence par se donner les trois sommets d'un triangle $S_i = (x_i, y_i)$, $i = 1, 2, 3$. On calcule ensuite les valeurs $z_i = f(S_i)$ où l'on suppose que $z_1 \leq z_2 \leq z_3$, quitte à redéfinir les indices. On introduit la notation

$$M = (x_1, y_1), B = (x_2, y_2), P = (x_3, y_3),$$

où M, B, P désignent respectivement le meilleur sommet, un bon sommet et le pire sommet.

À l'étape suivante, on souhaite remplacer P par un "meilleur" sommet. On calcule d'abord le point milieu du segment de droite joignant M et B :

$$C = \frac{M + B}{2} = \left(\frac{x_1 + y_1}{2}, \frac{x_2 + y_2}{2} \right).$$

Comme f diminue en passant de P à M ou de P à B, il est plausible que f prenne de plus petites valeurs du côté du segment MB opposé à P. Comme troisième sommet, on peut alors examiner le point R, réflexion du point P par rapport à MB :

$$R = C + (C - P) = 2C - P.$$

(Plus généralement, on pourra examiner $R = C + \alpha(C - P) = (1 + \alpha)C - \alpha P$, où $\alpha > 0$ s'appelle le *facteur de réflexion.*)

I) Si $f(R) < f(M)$, un nouveau minimum a été trouvé. Avant de le retenir, on peut penser que la direction choisie nous rapproche du minimum, et peut-être celui-ci est-il un peu plus loin dans cette même direction. On prolonge donc le segment PR jusqu'au point

$$E = R + (R - C) = 2R - C.$$

(Plus généralement, on prolonge jusqu'au point $E = R + (\gamma - 1)(R - C) = \gamma R + (1 - \gamma)C$, où $\gamma > 1$ s'appelle le *facteur d'expansion.*) Lorsque $f(E) < f(R)$, on répète le raisonnement précédent avec le triangle EMB ; sinon, on le fait avec le triangle RMB. Voir ci-dessus la figure 6.3.

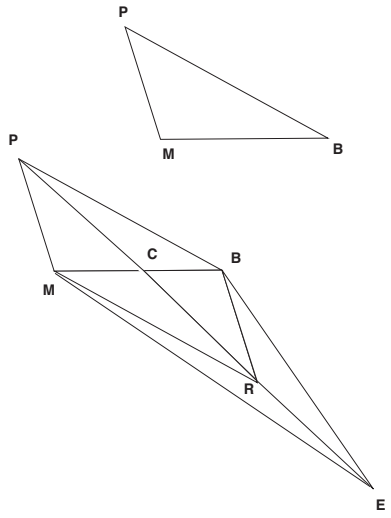

FIG. 6.3 – Première phase de la méthode du simplexe

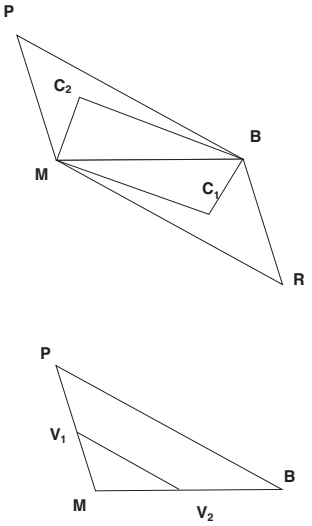

FIG. 6.4 – Deuxième phase de la méthode du simplexe

II) Si $f(M) \leq f(R) < f(B)$, on poursuit le raisonnement avec le triangle MRB.

III) Reste le cas $f(R) \geq f(B)$ que l'on va traiter en réduisant la taille de l'un des deux triangles MBR ou MBP. Une première possiblité (a) est que $f(B) \leq f(R) < f(P)$; dans ce cas, on retient les deux sommets M et B, et l'on choisit un troisième sommet C_1 entre C et R (réduction du côté des petites valeurs). L'autre possibilité (b) est que $f(R) \geq f(P)$, auquel cas on retient les sommets M, B, et on choisit un troisième sommet C_2 entre C et P (réduction du côté des grandes valeurs). Pour ces deux possibilités, la réduction est contrôlée par un facteur $0 < \beta < 1$ appelé *facteur de contraction*. Dans le cas (a), on choisit comme troisième sommet $S = C + \beta(R - C)$, et dans le cas (b) $S = C + \beta(P - C)$. Le triangle retenu est alors MSB, à moins que $f(S) > f(B)$. Dans cette dernière situation, on choisit le triangle en opérant une contraction par rapport à M ; pour ce faire, si V désigne l'un des sommets B, P, on définit deux autres sommets par

$$V' = M + \beta'(V - M),$$

où $0 < \beta' < 1$ est un facteur de contraction. Dans la figure 6.4, on désigne les sommets ainsi construits par V_1 et V_2, et on poursuit avec le triangle MV_1V_2.

Dans R, la méthode du simplexe est une option de la fonction `optim`. Par défaut, on y prend $\alpha = 1, \beta = \beta' = 0.5$ et $\gamma = 2$.

6.2.2 Méthodes de type Newton

Partout dans cette sous-section, nous nous plaçons dans le contexte où l'on désire optimiser une fonction de log-vraisemblance $l(\theta)$, où $\theta = (\theta_1, \ldots, \theta_k)$. La dérivée première de l (ou gradient) est souvent appelée *fonction score*, dérivée que l'on écrira ci-après sous la forme d'un vecteur ligne :

$$l'(\theta) = \left(\frac{\partial l}{\partial \theta_1}(\theta), \ldots, \frac{\partial l}{\partial \theta_k}(\theta) \right).$$

La dérivée seconde de l est définie comme étant la *matrice hessienne*, matrice $k \times k$ formée des dérivées partielles d'ordre deux de l :

$$l''(\theta) = \begin{pmatrix} \frac{\partial^2 l}{\partial\theta_1\partial\theta_1}(\theta) & \cdots & \frac{\partial^2 l}{\partial\theta_1\partial\theta_k}(\theta) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 l}{\partial\theta_k\partial\theta_1}(\theta) & \cdots & \frac{\partial^2 l}{\partial\theta_k\partial\theta_k}(\theta) \end{pmatrix}.$$

Dans la suite, nous supposerons que les dérivées partielles secondes de l sont continues, ce qui fera de la matrice hessienne une matrice symétrique. On notera qu'en théorie de l'estimation la matrice $-l''(\theta)$ s'appelle l'information (de Fisher) observée.

On sait qu'optimiser la log-vraisemblance équivaut à résoudre les k équations

$$\frac{\partial l}{\partial\theta_i}(\theta) = 0, i = 1, \ldots, k,$$

système que l'on écrira de manière plus compacte à partir du vecteur $\mathbf{0}$:

$$l'(\theta) = \mathbf{0}^T.$$

Pour reprendre la deuxième approche conduisant à la méthode de Newton univariée (voir p. 83), on considère l'approximation quadratique de Taylor de $l(\theta)$ autour de la valeur initiale θ_0 :

$$l_q(\theta) = l(\theta_0) + l'(\theta_0) \cdot (\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)^T l''(\theta_0)(\theta - \theta_0). \tag{6.4}$$

Selon la méthode de Newton, le maximum de la log-vraisemblance s'obtient en procédant à une itération dont le premier terme θ_1 est le point maximum en θ du membre droit de (6.4). Lorsque ce point est unique, on vérifie qu'il est solution de l'équation

$$l'(\theta_0) + (\theta - \theta_0)^T l''(\theta_0) = \mathbf{0}^T,$$

équation obtenue en égalant à $\mathbf{0}^T$ la dérivée première de (6.4). (Pour la justification de cette dérivation par rapport à un argument vectoriel, voir

l'exercice 4 a) de la série 9.) Si l'on suppose en outre que la matrice hessienne est inversible en θ_0, l'équation précédente a pour solution

$$\theta_1 = \theta_0 - (l''(\theta_0))^{-1} l'(\theta_0)^T,$$

équation analogue à (6.2). Dans le cas multivarié, l'itération de Newton pour la maximisation de la log-vraisemblance prend ainsi la forme

$$\theta_{n+1} = \theta_n - (l''(\theta_n))^{-1} l'(\theta_n)^T.$$

Le terme $(l''(\theta_n))^{-1} l'(\theta_n)^T$ peut être vue comme la direction ou le pas que l'on imprime à θ_n pour obtenir l'itéré suivant. Pour éviter d'avoir à inverser la matrice hessienne, notons qu'on peut aussi résoudre en θ_{n+1} l'équation équivalente

$$l''(\theta_n)(\theta_{n+1} - \theta_n) = -l'(\theta_n)^T.$$

Utiliser l'itération précédente ne va pas de soi. Du point de vue pratique, le principal problème rencontré est l'évaluation des dérivées. Il faut voir que si θ a k composantes il est nécessaire de calculer à l'étape n les k dérivées partielles d'ordre un de $l'(\theta_n)$ et les $k(k+1)/2$ dérivées partielles d'ordre deux de $l''(\theta_n)$. En outre, rien ne garantit que l'itération converge vers l'estimateur de vraisemblance maximale $\widehat{\theta}$ ou même que $l(\theta_{n+1}) > l(\theta_n)$.

On contourne souvent le problème du calcul de la matrice hessienne $l''(\theta_n)$ en substituant à celle-ci une approximation de forme plus simple. Dans cette veine, on se sert parfois d'une itération de la forme

$$\theta_{n+1} = \theta_n - M^{-1} l'(\theta_n)^T,$$

où M est une matrice $k \times k$ inversible *fixée*. Selon la théorie de l'optimisation des fonctions de plusieurs variables, dans le voisinage de son point maximum $\widehat{\theta}$ la log-vraisemblance l possède une matrice hessienne $l''(\theta)$ définie négative. De ce point de vue, les choix les plus simples pour l'approximation M sont des matrices diagonales des types $-cI_k$ $(c > 0)$ ou $-\mathrm{diag}(\alpha_1, \ldots, \alpha_k)$ $(\min_i \alpha_i > 0)$; un autre bon choix peut être $l''(\theta_0)$ si θ_0 est suffisamment

proche de $\widehat{\theta}$. Quel que soit ce choix, l'algorithme convergera à condition que M^{-1} soit suffisamment proche de $(l''(\widehat{\theta}))^{-1}$.

Le choix $M = -I_k$ mérite un peu plus d'attention. L'itération prend alors la forme

$$\theta_{n+1} = \theta_n + l'(\theta_n)^T,$$

où $l'(\theta_n)^T$ est le gradient de la log-vraisemblance. On sait que le gradient d'une fonction de plusieurs variables est un vecteur indiquant la direction où la fonction augmente le plus rapidement. La méthode d'optimisation correspondante s'appelle la méthode du gradient (*steepest ascent*). Notons que pour ne pas s'exposer à une divergence en s'éloignant indûment de θ_n, il peut être souhaitable de prendre $M = -\alpha I_k$ pour un $\alpha > 0$ petit.

Pour mieux comprendre le choix de M, on notera que si A_n est une matrice $k \times k$ définie positive et $\theta_{n+1} = \theta_n + \alpha_n A_n^{-1} l'(\theta_n)^T$, on aura

$$l(\theta_{n+1}) > l(\theta_n)$$

pourvu que $\alpha_n > 0$ soit assez petit. En effet, puisque $l'(\theta)$ est défini comme un vecteur ligne, il découle du développement de Taylor autour de θ_n que

$$l(\theta_{n+1}) - l(\theta_n) = l(\theta_n + \alpha_n A_n^{-1} l'(\theta_n)^T) - l(\theta_n) = \alpha_n l'(\theta_n) A_n^{-1} l'(\theta_n)^T + o(\alpha_n),$$

où $o(\alpha_n)$ est un multiple de α_n^2 et le terme $\alpha_n l'(\theta_n) A_n^{-1} l'(\theta_n)^T > 0$ du fait que A_n^{-1} est définie positive. Finalement, lorsque $\alpha_n \to 0$, $\alpha_n l'(\theta_n) A_n^{-1} l'(\theta_n)^T$ domine $o(\alpha_n)$, car par définition celui-ci tend plus vite vers 0 que α_n .

Nous présenterons maintenant un peu plus en détail deux autres méthodes de type Newton.

Méthode de Fisher ou du scoring

Rappelons que l'information de Fisher associée aux observations i.i.d. X_1, $\ldots, X_n$ est définie par

$$I(\theta) = -E[l''(\theta)] = -E[l''(\theta|X_1, \ldots, X_n)] = -nE[l''(\theta|X)],$$

où X a la même loi que les X_i. L'entrée (i,j) de la matrice hessienne $l''(\theta)$ est donnée par

$$\begin{aligned} l''_{ij}(\theta) &= \frac{\partial^2}{\partial\theta_i\partial\theta_j} l(\theta|X_1,\ldots,X_n) \\ &= \sum_{m=1}^{n} \frac{\partial^2}{\partial\theta_i\partial\theta_j} l(\theta|X_m), \end{aligned}$$

Or, selon la loi des grands nombres,

$$\frac{l''_{ij}(\theta)}{n} = \frac{1}{n}\sum_{m=1}^{n} \frac{\partial^2}{\partial\theta_i\partial\theta_j} l(\theta|X_m) \to_p E\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j} l(\theta|X)\right) = E(l''_{ij}(\theta|X))$$

lorsque $n \to \infty$. Au niveau matriciel, cela revient à dire que

$$\frac{l''(\theta)}{n} \to_p E(l''(\theta|X)$$

lorsque $n \to \infty$, ou encore, pour n assez grand, $l''(\theta) \approx nE(l''(\theta|X) = -I(\theta)$ en probabilité.

Dans les équations définissant la méthode de Newton, il est donc naturel de ce point de vue d'estimer $l''(\theta_n)$ par $-I(\theta_n)$. La méthode d'optimisation qui en découle s'appelle la méthode de Fisher ou méthode du scoring. Formellement, celle-ci est définie par l'itération

$$\theta_{n+1} = \theta_n + (I(\theta_n))^{-1}(l'(\theta_n)^T.$$

En pratique, on constate qu'il est souvent plus facile de calculer $I(\theta)$ que la matrice hessienne $l''(\theta)$. Un autre avantage de cette approche est que $(I(\widehat{\theta}))^{-1}$ estime la matrice de covariance asymptotique de l'estimateur du maximum de vraisemblance $\widehat{\theta}$. Sous certaines conditions de régularité, si $\theta = (\theta_1,\ldots,\theta_k)$ est le paramètre estimé, on sait en effet que $\widehat{\theta}$ est asymptotiquement $N_k(\theta,(I(\theta))^{-1})$.

Méthodes de type quasi-Newton

Les méthodes de type quasi-Newton reposent toutes sur une itération de la forme

$$\theta_{n+1} = \theta_n - M_n^{-1} l'(\theta_n)^T,$$

où M_n est une approximation de la matrice hessienne $l''(\theta_n)$. Elles se distinguent dans leur manière de mettre à jour les M_n d'une étape à l'autre. Une particularité de cette mise à jour est qu'elle incorpore de manière relativement simple de l'information sur le comportement de la fonction score l' au voisinage de θ_n dans la direction $-M_n^{-1}l'(\theta_n)^T$.

Plus précisément, cela se fait en imposant à la matrice M_{n+1} de vérifier une condition du type

$$l'(\theta_{n+1}) - l'(\theta_n) = M_{n+1}(\theta_{n+1} - \theta_n). \tag{6.5}$$

Il existe plusieurs façons de définir M_{n+1} comme fonction de M_n tout en satisfaisant à l'équation (6.5). Nous n'en retiendrons qu'une, considérée par plusieurs comme supérieure aux autres et pour laquelle les matrices M_n sont symétriques : la méthode BFGS (pour Broyden, Fletcher, Goldfarb et Shanno). Pour celle-ci, la mise à jour est faite à partir de l'équation

$$M_{n+1} = M_n - \frac{M_n z_n z_n^T M_n}{z_n^T M_n z_n} + \frac{y_n y_n^T}{z_n^T y_n},$$

où $z_n = \theta_{n+1} - \theta_n$ et $y_n = l'(\theta_{n+1}) - l'(\theta_n)$.

6.3 L'algorithme EM

Lorsqu'une fonction de vraisemblance est difficile à maximiser, il s'avère parfois avantageux d'insérer les variables observées dans un ensemble plus vaste de variables dont la loi est paramétrisée de la même façon, tout en ayant la particularité de faciliter la maximisation de la nouvelle fonction de vraisemblance obtenue par cet ajout de variables. En pratique, les variables additionnelles prennent alors la forme de variables non observées ou partiellement observées. Pour évaluer la contribution des variables additionnelles à la fonction de vraisemblance, on se sert alors de la loi conditionnelle de ces variables étant donné les variables observées. Cette idée est à la base d'un algorithme très général de maximisation des fonctions de vraisemblance connu

sous le nom d'algorithme EM, où la lettre E fait référence à une espérance mathématique et la lettre M à une maximisation. La méthode a été présentée pour la première fois dans un article de Dempster, Laird et Rubin (1977).

Désignons par Y le vecteur des variables observées et par U celui des variables non observées ou partiellement observées. Nous dirons que $X = (Y, U)$ est le vecteur des *données augmentées ou complètes*. On suppose que X et Y ont une densité (ou fonction de masse) dépendant du même paramètre θ. En outre, nous sommes dans la situation où, relativement à l'observation $Y = y$, la vraisemblance

$$L(\theta) = f(y;\theta) = \int f(y,u;\theta)du = \int f(y|u;\theta)f(u;\theta)du$$

est difficile à calculer ou à maximiser.

De son côté, la log-vraisemblance en θ relative aux données augmentées (y, u) s'écrit

$$\log f(y,u;\theta) = l(\theta) + \log f(u|y;\theta), \tag{6.6}$$

où $l(\theta) = \log L(\theta)$ est la log-vraisemblance relative à l'observation y. Comme u n'est pas complètement observée, l'équation (6.6) ne sera utile qu'après avoir calculé la moyenne en u de ses termes par rapport à la loi conditionnelle $f(u|y;\theta')$ pour une valeur appropriée θ' du paramètre. Plus précisément, à l'étape m de l'algorithme θ' sera le m^e élément θ_m d'une suite de valeurs de θ déterminées par l'algorithme et devant converger vers l'estimateur du maximum de vraisemblance $\widehat{\theta}$.

Chaque étape de l'algorithme EM est divisée en une phase 'Espérance' suivie d'une phase 'Maximisation'. À l'étape m, la première phase consiste à calculer l'espérance conditionnelle des deux membres de (6.6) pour obtenir

$$E[\log f(Y,U;\theta)|Y = y;\theta_m] = l(\theta) + E[\log f(U|Y;\theta)|Y = y;\theta_m],$$

identité qu'on récrit plus simplement

$$Q(\theta;\theta_m) = l(\theta) + C(\theta;\theta_m). \tag{6.7}$$

La deuxième phase est la maximisation en θ du membre gauche de (6.7). Nous allons exploiter cette maximisation au moyen de l'inégalité dite de Jensen. Pour énoncer celle-ci, rappelons qu'une fonction réelle définie sur un intervalle est dite convexe si $h(\alpha x + (1-\alpha)y) \leq \alpha h(x) + (1-\alpha)h(y)$, quels que soient x, y dans l'intervalle et $0 \leq \alpha \leq 1$. On parle de stricte convexité lorsque l'inégalité précédente est stricte lorsque $x \neq y$ et $\alpha \neq 0, 1$. Lorsque h est différentiable deux fois, une condition suffisante pour qu'elle soit strictement convexe est que $h''(x) > 0$ pour tout x.

Théorème 6.3.1. (Inégalité de Jensen) *Étant donné une variable aléatoire X, soit h une fonction convexe définie sur un intervalle contenant l'ensemble des valeurs de X et telle que $E(X)$ et $E(h(X))$ existent. Alors*

$$h(E(X)) \leq E(h(X)).$$

De plus, lorsque h est strictement convexe, l'égalité ne peut avoir lieu que si X est constante avec probabilité 1.

On trouvera une preuve de l'inégalité de Jensen dans tout bon livre de théorie des probabilités. Dans l'algorithme EM, on applique l'inégalité à la fonction strictement convexe $h(x) = -\log x$, $x > 0$, ce qui donne

$$h(E(X)) = -\log E(X) < -E(\log X) = E(h(X)),$$

pourvu que $X > 0$ ne soit pas constante avec probabilité 1.

Posons

$$\theta_{m+1} = \text{argmax}_\theta \, Q(\theta; \theta_m),$$

et montrons que $l(\theta_{m+1}) > l(\theta_m)$. De (6.7), on peut voir que

$$\begin{aligned}
l(\theta_{m+1}) - l(\theta_m) &= Q(\theta_{m+1};\theta_m) - Q(\theta_m;\theta_m) - E\left[\log \frac{f(U|Y;\theta_{m+1})}{f(U|Y;\theta_m)}\middle| Y = y;\theta_m\right] \\
&\geq -E\left[\log \frac{f(U|Y;\theta_{m+1})}{f(U|Y;\theta_m)}\middle| Y = y;\theta_m\right] \\
&= -\int \log\left[\frac{f(u|y;\theta_{m+1}}{f(u|y;\theta_m)}\right] f(u|y;\theta_m)du \\
&\geq -\log E\left[\frac{f(U|Y;\theta_{m+1})}{f(U|Y;\theta_m)}\middle| Y = y;\theta_m\right] \qquad \text{(Jensen)} \\
&= -\log \int \frac{f(u|y;\theta_{m+1})}{f(u|y;\theta_m)} f(u|y;\theta_m)du \\
&= -\log \int f(u|y;\theta_{m+1})du \\
&= 0.
\end{aligned}$$

Dans tous les cas où deux valeurs différentes de θ déterminent des densités conditionnelles $f(u|y;\theta)$ différentes, on a $f(U|y;\theta_{m+1})/f(U|y;\theta_m) \neq$ d'une constante avec probabilité 1, et donc l'inégalité de Jensen s'applique strictement. On a alors $l(\theta_{m+1}) > l(\theta_m)$ dès que $\theta_{m+1} \neq \theta_m$, ce qui fait de $(l(\theta_m))$ une suite croissante. Soulignons que la méthode de Newton-Raphson n'a pas toujours cette propriété. Notons finalement que, sous des conditions de régularité assez faibles, la suite (θ_m) converge vers un point maximum local de $l(\theta)$.

En résumé, l'algorithme EM procède comme suit.

Algorithme EM

1. Se donner une valeur initiale θ_0. Aux itérations $m = 1, 2, \ldots$, passer par les deux phases suivantes dans l'ordre.
2. La phase E (espérance mathématique) : calculer

$$Q(\theta;\theta_m) = E[\log f(Y,U;\theta)|Y = y;\theta_m] = \int \log f(y,u;\theta) f(u|y;\theta_m)du.$$

3. La phase M (maximisation) : calculer

$$\theta_{m+1} = \text{argmax}_\theta \, Q(\theta; \theta_m).$$

L'application de l'algorithme prend fin lorsque $|Q(\theta_{m+1}; \theta_m) - Q(\theta_m; \theta_m)|$ ou $|\theta_{m+1} - \theta_m|$ sont suffisamment petits. Lorsque la log-vraisemblance possède un maximum unique, la suite (θ_m) converge généralement vers l'estimateur du maximum de vraisemblance cherché. On trouvera dans les références des conditions garantissant la convergence.

Exemple 3. Cet exemple est emprunté à la statistique génétique. Avant de le présenter, on donne d'abord quelques définitions empruntées au *Nouveau Petit Robert* et au *Grand Robert*.

allèles : gènes de même fonction et d'action dissemblable occupant la même place dans une même paire de chromosomes.

génotype : patrimoine génétique d'un individu dépendant des gènes hérités de ses parents, qu'ils soient exprimés ou non.

phénotype : ensemble des caractères individuels correspondant à la réalisation du génotype, déterminée par l'action de facteurs du milieu au cours du développement de l'organisme.

L'appartenance d'un individu à un groupe sanguin est déterminée par la présence de certains gènes dans le bagage génétique. Les trois allèles impliqués sont désignés par A, B et O. Chaque individu est porteur d'exactement deux allèles, l'un hérité du père, l'autre de la mère. Chaque combinaison de deux allèles définit un génotype ; pour le groupe sanguin, on désigne ces génotypes par $A/A, A/B, A/O$, etc., où l'ordre des lettres est sans importance. Les allèles A et B étant dominants par rapport à l'allèle O, la paire A/O s'exprime de la même façon que la paire A/A, et la paire B/O comme la paire B/B. Pour leur part, les paires A/B et O/O ont chacune une expression particulière. Ces quatre formes d'expression définissent autant de phénotypes que l'on représente simplement par A, B, AB et O,

les groupes sanguins observables bien connus. Comme certaines combinaisons d'allèles ne sont pas distinguables, les proportions d'allèles A, B et O dans une population ne sont pas directement estimables. Dans ce premier exemple, l'algorithme EM est utilisé pour estimer ces proportions à partir des proportions observées des 4 phénotypes.

Les données proviennent de $n = 521$ individus se répartissant entre les phénotypes selon les fréquences $n_A = 186, n_B = 38, n_{AB} = 13$ et $n_O = 284$. Dans l'application de l'algorithme EM qui suit, ces fréquences définissent le vecteur des variables observées Y, tandis que le vecteur U des variables non observées est défini par les fréquences des génotypes $n_{A/A}$ et $n_{B/B}$. Noter que $n_A = n_{A/A} + n_{A|O}$ et $n_B = n_{B/B} + n_{B|O}$. Le problème statistique consiste à estimer les proportions des trois allèles p_A, p_B et p_O, proportions positives vérifiant l'identité $p_A + p_B + p_O = 1$.

Pour obtenir l'expression de la log-vraisemblance, on se sert de la loi de Hardy-Weinberg pour les populations en équilibre génétique ; selon cette loi, la probabilité d'un génotype s'obtient en considérant comme indépendants les deux allèles hérités des parents. Par exemple, cela donne p_A^2 pour la probabilité du génotype A/A et $2p_Ap_O$ pour celle du génotype A/O, puisque, dans ce dernier cas, chacun des deux parents peut être porteur de l'allèle A.

Il est facile de voir que les valeurs du vecteur $X = (Y, U)$ sont en correspondance bijective avec les valeurs de la multinomiale $(n_{A/A}, n_{A/O}, n_{B/B}, n_{B/O}, n_{A/B}, n_O)$. Comme fonction de $p = (p_A, p_B, p_O)$, la log-vraisemblance des données augmentées s'écrit donc

$$\begin{aligned} \log f(y, u; p) &= n_{A/A} \log(p_A^2) + n_{A/O} \log(2p_Ap_O) + n_{B/B} \log(p_B^2) \\ &\quad + n_{B/O} \log(2p_Bp_O) + n_{AB} \log(2p_Ap_B) + n_O \log(p_O^2) \\ &\quad + \log \left(\begin{array}{c} n \\ n_{A/A}, n_{A/O}, n_{B/B}, n_{B/O}, n_{AB}, n_O \end{array} \right). \end{aligned} \tag{6.8}$$

En comparaison, la log-vraisemblance relative aux données observées est

égale à

$$\begin{aligned}\log f(y;p) &= n_A \log(p_A^2 + 2p_A p_O) + n_B \log(p_B^2 + 2p_B p_O) \\ &\quad + n_{AB} \log(2p_A p_B) + n_O p_O^2 \\ &\quad + \log \begin{pmatrix} n \\ n_A, n_B, n_{AB}, n_O \end{pmatrix}.\end{aligned}$$

La dernière fonction est plus difficile à maximiser en p que la première.

À l'étape m de l'algorithme EM, il nous faut calculer l'espérance conditionnelle de $\log f(Y, U; p)$ étant donné le vecteur des fréquences observées $Y = y = (n_A, n_B, n_{AB}, n_O)$ lorsque la loi conditionnelle est paramétrisée par le m^e itéré $p_m = (p_{mA}, p_{mB}, p_{mO})$. Pour commencer, notons qu'au membre droit de (6.8) les six premiers termes logarithmiques sont constants. Pour tous ces termes, le calcul de l'espérance conditionnelle se ramène ainsi à celui des fréquences $n_{A/A}, n_{A/O}$, etc. Il n'est pas difficile de voir que les lois conditionnelles de $n_{A/A}, n_{A/O}, n_{B/B}, n_{B/O}$ sont binomiales. Prenons par exemple $n_{A/A}$ dont les valeurs possibles sont $0, 1, \ldots, n_A$. La probabilité de succès est celle qu'un individu appartienne au génotype A/A sachant qu'il appartient au phénotype A. Lorsque les probabilités des allèles sont (p_{mA}, p_{mB}, p_{mO}), la formule de Bayes implique que cette probabilité de succès vaut $p_{mA}^2/(p_{mA}^2 + 2p_{mA}p_{mO})$, ce qui entraîne que

$$n_{mA/A} := E(n_{A/A}|Y = y; p_m) = n_A \frac{p_{mA}^2}{p_{mA}^2 + 2p_{mA}p_{mO}}.$$

On peut obtenir pareillement les espérances conditionnelles de $n_{A/O}, n_{B/B}$ et $n_{B/O}$. Par ailleurs, il est clair que $E(n_{AB}|Y = y; p_m) = n_{AB}$ et $E(n_O|Y = y; p_m) = n_O$. Quant à l'espérance conditionnelle du logarithme du coefficient multinomial, il n'est pas indispensable de la calculer car $p = (p_A, p_B, p_O)$ est absent de ce terme et la maximisation de Q est faite sur p.

À la deuxième phase de l'algorithme, nous devons maximiser $Q(p|p_m)$ obtenue de (6.8) en remplaçant $n_{A/A}$ par $n_{mA/A}$, $n_{A/O}$ par $n_{mA/O}$, etc. Comme la maximisation est faite sous contrainte, on utilise un multiplicateur

de Lagrange λ pour maximiser plutôt en p la fonction

$$H(p,\lambda) = Q(p|p_m) + \lambda(p_A + p_B + p_O - 1).$$

Il suffit pour cela de résoudre les équations obtenues en égalant à 0 les dérivées partielles premières de H. Ces équations sont :

$$\begin{aligned}
\frac{\partial}{\partial p_A}H(p,\lambda) &= \frac{2n_{mA/A}}{p_A} + \frac{n_{mA/O}}{p_A} + \frac{n_{AB}}{p_A} + \lambda = 0 \\
\frac{\partial}{\partial p_B}H(p,\lambda) &= \frac{2n_{mB/B}}{p_B} + \frac{n_{mB/O}}{p_B} + \frac{n_{AB}}{p_B} + \lambda = 0 \\
\frac{\partial}{\partial p_O}H(p,\lambda) &= \frac{n_{mA/O}}{p_O} + \frac{n_{mB/O}}{p_O} + \frac{2n_O}{p_O} + \lambda = 0 \\
\frac{\partial}{\partial \lambda}H(p,\lambda) &= p_A + p_B + p_O - 1 = 0.
\end{aligned}$$

Il n'est pas difficile de voir $\lambda = -2n$, et la solution unique est donc

$$\begin{aligned}
p_{m+1,A} &= \frac{2n_{mA/A} + n_{mA/O} + n_{AB}}{2n} \\
p_{m+1,B} &= \frac{2n_{mB/B} + n_{mB/O} + n_{AB}}{2n} \\
p_{m+1,O} &= \frac{n_{mA/O} + n_{mB/O} + 2n_O}{2n}.
\end{aligned}$$

Pour comprendre de façon intuitive les itérés obtenus, considérons par exemple $p_{m+1,A}$. Comme estimation de p_A, cet itéré prend la somme des fréquences observées ou estimées à l'étape m de chacune des trois sources d'allèles A, soient les génotypes $A/A, A/O$ et A/B, puis divise cette somme par $2n$, le nombre total d'allèles des n individus.

Dans cet exemple, l'algorithme est relativement facile à programmer. Pour une valeur initiale p_0 bien choisie, la convergence sera rapide. Comparé à la méthode de Newton-Raphson, l'algorithme EM est en général plus lent. Un résultat général établit que sa vitesse de convergence a un caractère linéaire plutôt que quadratique.

Exemple 4. Estimation de la densité d'un mélange

Un mélange de lois est un modèle utilisé pour représenter une observation pouvant provenir de plusieurs sous-populations distinctes, sans que l'on sache de laquelle. Lorsque le nombre k de sous-populations est fini, une telle observation Y possède une densité de la forme

$$f(y;\theta) = \sum_1^k \pi_i f_i(y;\phi_i), \quad 0 \le \pi_i \le 1, \sum_1^k \pi_i = 1,$$

où π_i est la probabilité que y provienne de la i^e sous-population, $f_i(y;\phi_i)$ est la densité conditionnelle à cet événement, supposée connue dans sa forme. On a ici $\theta = (\pi_1, \ldots, \pi_k, \phi_1, \ldots, \phi_k)$, où les ϕ_i peuvent eux-mêmes être des vecteurs. En raison du grand nombre de paramètres, estimer la loi d'un mélange est en général un problème particulièrement compliqué.

Aux données observées, on ajoutera ici l'indicatrice U égale à i avec probabilité π_i, une variable non observée. Une donnée augmentée (y, i) détermine la fonction de vraisemblance $\pi_i f_i(y;\phi_i)$. Lorsqu'on a n observations d'indices $j = 1, \ldots, n$, la fonction de vraisemblance relative à une donnée (y_j, u_j) peut donc s'écrire $\prod_{i=1}^k [\pi_i f_i(y_j;\phi_i)]^{1(u_j=i)}$, où $1(u_j = i) = 1$ ou 0, selon que $u_j = i$ ou non. La log-vraisemblance correspondante est donc

$$\log f(y_j, u_j;\theta) = \sum_{i=1}^k 1(u_j = i)[\log \pi_i + \log f_i(y_j;\phi_i)].$$

En sommant ensuite sur $j = 1, \ldots, n$, on obtient la log-vraisemblance relative aux n données augmentées.

À l'étape m de l'algorithme EM, il nous faut calculer l'espérance de $\log f(Y, U;\theta)$ par rapport à la loi conditionnelle

$$P(U_j = i|Y = y;\theta_m) = \frac{\pi_{mi} f_i(y_j;\phi_{mi})}{\sum_{l=1}^k \pi_{ml} f_l(y_j;\phi_{ml})} := w_i(y_j;\theta_m),$$

expression découlant de la formule de Bayes. Comme

$$E(1(U_j = i)|Y = y;\theta_m) = P(U_j = i|Y = y;\theta_m) = w_i(y_j;\theta_m),$$

la valeur espérée de la log-vraisemblance pour l'ensemble des données augmentées $(y_j, u_j), j = 1, \ldots, n$, est

$$\begin{aligned} Q(\theta;\theta_m) &= \sum_{j=1}^{n}\sum_{i=1}^{k} w_i(y_j;\theta_m)[\log \pi_i + \log f_i(y_j;\phi_i)] \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n} w_i(y_j;\theta_m)\log \pi_i + \sum_{i=1}^{k}\sum_{j=1}^{n} w_i(y_j;\theta_m)\log f_i(y_j;\phi_i) \quad (6.9) \end{aligned}$$

Comme la maximisation en θ est faite sous la contrainte $\sum_1^k \pi_i = 1$, on pourra utiliser un multiplicateur de Lagrange. On remarque cependant que la maximisation en π_i peut être faite en ne considérant que le premier terme du membre droit de (6.9). Or, à une constante près, ce terme a la forme d'une log-vraisemblance multinomiale. Puisque $\sum_{i=1}^k w_i(y_j;\theta_m) = 1$ pour tout j, le maximum en π_i est donc

$$\pi_{m+1,i} = \frac{\sum_{j=1}^n w_i(y_j;\theta_m)}{\sum_{i=1}^k \sum_{j=1}^n w_i(y_j;\theta_m)} = \frac{\sum_{j=1}^n w_i(y_j;\theta_m)}{n},$$

soit le poids moyen pour la composante i du mélange.

Le maximum sur les autres paramètres est déterminé par le deuxième terme de (6.9). *Prenons le cas particulier où les f_i sont des densités normales de paramètres μ_i, σ_i^2* ($2k$ paramètres). Il nous suffit alors d'égaler à 0 les $2k$ dérivées partielles :

$$\begin{aligned} \frac{\partial}{\partial \mu_i}\sum_{j=1}^{n} w_i(y_j;\theta_m)\log f_i(y_j;\mu_i,\sigma_i^2) &= -\frac{\sum_{j=1}^n w_i(y_j;\theta_m)(y_j-\mu_i)}{\sigma_i^2}, \\ \frac{\partial}{\partial \sigma_i^2}\sum_{j=1}^{n} w_i(y_j;\theta_m)\log f_i(y_j;\mu_i,\sigma_i^2) &= \sum_{j=1}^{n} w_i(y_j;\theta_m)\left[-\frac{1}{2\sigma_i^2} + \frac{(y_j-\mu_i)^2}{2\sigma_i^4}\right], \end{aligned}$$

et résoudre pour $i = 1, \ldots, k$. Il n'est pas difficile de voir que les solutions sont

$$\begin{aligned} \mu_{m+1,i} &= \frac{\sum_{j=1}^n w_i(y_j;\theta_m) y_j}{\sum_{j=1}^n w_i(y_j;\theta_m)} \\ \sigma^2_{m+1,i} &= \frac{\sum_{j=1}^n w_i(y_j;\theta_m)(y_j-\mu_{m+1,i})^2}{\sum_{j=1}^n w_i(y_j;\theta_m)} \end{aligned}$$

pour $i = 1, \ldots, k$.

6.4 Bibliographie

- Lange, K. (1999). *Numerical Analysis for Statisticians*, Springer-Verlag, New York. Chapitres 5, 10 et 11. Bon exposé, surtout pour la théorie. Couvre un grand nombre de sujets.
- Monahan, J. F. (2001). *Numerical Methods of Statistics*, Cambridge University Press, Cambridge. Chapitre 8 et 9. Forte teinte informatique. Valable surtout pour les algorithmes. Intérêt inégal.
- Givens, G. H. et Hoeting, J. A. (2005). *Computational Statistics*, J. Wiley & Sons, New York. Chapitres 2 et 4. Couvre un grand nombre de sujets, parfois de manière assez superficielle. La partie optimisation est bien faite.
- Thisted, R. A. (1988). *Elements of Statistical Computing*, Chapman & Hall, Londres. Chapitre 4. Très bon livre. Mériterait une mise à jour.
- Nash, J. C. (1990). *Compact Numerical Methods for Computers*, 2^e éd., Adam Hilger, New York. Chapitres 12–15. Très bon pour les algorithmes.
- Davison, A. (2003). *Statistical Models*, Cambridge University Press, Cambrige. Chapitres 4 et 5. Livre excellent. D'un très bon niveau.
- Venables, W. N. et Ripley, B. (2002). *Modern Applied Statistics with S*, Springer, New York. La meilleure référence pour R et S-Plus.
- Bickel, P. J. et Doksum, K. A. (2001). *Mathematical Statistics*, 2^e éd., vol. 1, Prentice Hall, New York. Chapitre 2. Bon livre. De niveau élevé.