Un algorithme stochastique en statistique semi-paramétrique

Dans les études médicales ou pharmacologiques, on recueille souvent des données longitudinales, c'est-à-dire que pour chaque patient, on relève une série de mesures dans le temps. Par exemple on mesure un taux d'hormone dans le sang toutes les heures, pendant une journée. On peut alors modéliser ces mesures de la façon suivante :

$$Y_{ij} = f_j(x_{ij}) + \varepsilon_{ij}, \quad i = 1, \dots, n, j = 1, \dots, J$$

où J est le nombre de patients et n le nombre de mesures par patients, prises aux temps x_{ij} . Les variables aléatoires ε_{ij} représentent un bruit de mesure. L'objet d'étude est l'estimation de la courbe f_j pour chaque patient, par exemple pour connaître sa réaction particulière à un traitement. On considère souvent que ces courbes sont issues d'une même réponse générique f au traitement, modifiées par un paramètre θ_j propre à chaque patient : $f_j(x_{ij}) = f(x_{ij}, \theta_j)$. Ce type de modèle peut être utilisé pour de nombreuses autres données : vitesse moyenne sur une route pour différents jours, reconnaissance vocale, courbes de dépenses en économétrie, *etc.* Ici on va se restreindre au cas où θ_j est un paramètre de translation, c'est-à-dire que

$$Y_{ij} = f(x_{ij} - \theta_j) + \varepsilon_{ij}, \quad i = 1, \dots, n, j = 1, \dots, J.$$

C'est le cas par exemple en reconnaissance vocale lorsque J personnes prononcent le même mot mais en commençant à des instants différents. On cherche alors à retrouver f et les θ_j . Ici on se restreint également au cas $J = 1$. Il sera alors nécessaire d'ajouter des hypothèses pour que le modèle soit identifiable. Pour estimer $\theta = \theta_1$, nous allons utiliser un algorithme stochastique, construit initialement pour résoudre des équations. Cet algorithme est présenté en première partie.

1 Algorithme de Robbins-Monro

L'algorithme de Robbins-Monro a été introduit en 1951 par Herbert Robbins et Sutton Monro pour résoudre numériquement une équation du type $U(x) = a$ où $a \in \mathbb{R}^d$ et U est une fonction de $\mathbb{R}^d$ dans $\mathbb{R}^d$.

Si U est une fonction réelle croissante, une idée naturelle pour trouver une solution de l'équation est d'utiliser la suite suivante :

$$x_{n+1} = x_n + \gamma_n(a - U(x_n))$$

où γ_n est une suite de réels positifs. Selon que $U(x_n)$ est inférieur ou supérieur à a , on corrige pour trouver l'approximation suivante x_{n+1} . On choisit γ_n tendant vers 0 de façon à stopper l'évolution de (x_n) . On espère que de cette façon la suite x_{n+1} converge vers un zéro de $U - a$. C'est effectivement le cas sous hypothèses sur U et γ_n .

L'algorithme de Robbins-Monro reprend cette idée mais utilise des variables aléatoires : c'est un algorithme stochastique¹. On l'utilise quand la fonction n'est pas accessible directement. Par exemple, la valeur de $U(x)$ est obtenue par un appareil de mesure qui commet de

1. En règle général les algorithmes stochastiques fournissent une meilleure exploration de l'espace d'état et peuvent ainsi améliorer leurs homologues déterministes.

petites erreurs. On a alors seulement accès à $Z = U(x) + \varepsilon$, où ε est un bruit de mesure, que l'on modélise généralement comme une variable aléatoire centrée. De façon plus générale, on peut supposer que U s'écrit

$$U(x) = \int \mathcal{V}(x, y) \mu(dy) \quad (1)$$

où $\mathcal{V}$ est une fonction de $\mathbb{R}^d \times E$ dans $\mathbb{R}^d$, E est un espace mesurable et μ une mesure de probabilité sur E . Il est courant que $\mathcal{V}$ s'obtienne facilement et qu'on sache simuler des variables aléatoires de loi μ mais que U soit difficile à calculer. Alors, si l'on a accès à une variable Z de loi μ , puisque $U(x) = \mathbb{E}(\mathcal{V}(x, Z))$, on peut essayer de remplacer $U(x)$ dans l'algorithme par $T = \mathcal{V}(x, Z)$.

Finalement l'algorithme de Robbins-Monro est le suivant. Soit γ_n une suite de réels positifs. A partir d'une variable aléatoire quelconque V_0 , on définit V_n récursivement par

$$V_{n+1} = V_n + \gamma_n(a - T_{n+1}). \quad (2)$$

Ici $(T_n)_{n \geq 1}$ est une suite de variables aléatoires telles que $\mathbb{E}(T_{n+1} | \mathcal{F}_n) = U(V_n)$ où on a noté $\mathcal{F}_n = \sigma(V_0, \dots, V_n, T_1, \dots, T_n)$. Le théorème suivant montre que cet algorithme converge bien vers une solution de $U(x) = a$ sous certaines conditions. On note $\langle \cdot, \cdot \rangle$ le produit scalaire usuel de $\mathbb{R}^d$ et $\|\cdot\|$ la norme associée.

Théorème 1 *Soit V_n la suite définie par l'algorithme (2). On suppose que l'équation $U(x) = a$ admet une unique solution $x = x_a$ et que :*

1. U est continue et pour tout $x \neq x_a$, $\langle a - U(x), x_a - x \rangle > 0$,
2. la suite de réels positifs (γ_n) est telle que $\sum_n \gamma_n = \infty$ et $\sum_n \gamma_n^2 < \infty$,
3. On a $\mathbb{E}\|V_0\| < \infty$ et $\sup_n \|T_n\| < \infty$ presque sûrement.

Alors la suite $(V_n)_{n \geq 0}$ converge presque sûrement vers x_a .

Une application de l'algorithme est l'estimation récursive de quantile. En effet si U est la fonction de répartition d'une variable aléatoire Z admettant une densité strictement positive, elle vérifie les hypothèses et le quantile d'ordre a est la solution de l'équation $U(x) = a$. Si on observe un échantillon $Z_1, \dots, Z_n$, on peut alors utiliser $T_{n+1} = \mathbb{1}_{Z_{n+1} \leq V_n}$, ce qui correspond à $\mathcal{V}(x, y) = \mathbb{1}_{y \leq x}$ dans (1). Nous allons voir une deuxième application en statistiques de cet algorithme dans la deuxième partie de ce texte, mais démontrons d'abord le théorème.

Démonstration du théorème

On cherche à montrer que $\lim_{n \rightarrow \infty} \|V_n - x_a\| = 0$. Par définition de l'algorithme

$$\mathbb{E}(\|V_{n+1} - x_a\|^2 | \mathcal{F}_n) = \|V_n - x_a\|^2 + \gamma_n^2 \mathbb{E}(\|a - T_{n+1}\|^2 | \mathcal{F}_n) - 2\gamma_n \langle V_n - x_a, U(V_n) - a \rangle.$$

Notons $\sigma(\gamma) = \sum_{n=0}^{\infty} \gamma_n^2$ et $b = \sup_n \|T_n\|$. On remarque que

$$\sum_n \gamma_n^2 \mathbb{E}(\|a - T_{n+1}\|^2 | \mathcal{F}_n) \leq \sigma(\gamma) (\|a\| + b)^2.$$

On pose alors pour tout $n \geq 0$

$$W_n = \|V_n - x_a\|^2 + \sigma(\gamma) (\|a\| + b)^2 - \sum_{k=0}^{n-1} \gamma_k^2 \mathbb{E}(\|a - T_{k+1}\|^2 | \mathcal{F}_k)$$

avec la convention $\sum_{k=0}^{-1} = 0$. Un calcul permet de montrer que $(W_n)_{n \geq 0}$ est une surmartingale, qui est de surcroît positive par construction. Par conséquent, elle converge presque

sûrement vers une variable aléatoire W_∞ . De plus $\mathbb{E}(W_0) < \infty$ donc W_∞ est finie presque sûrement. En reprenant la définition de W_n , on en déduit que la suite $(\|V_n - x_a\|^2)$ converge presque sûrement vers une limite finie p.s.

Il reste à montrer que cette limite est nulle. Pour tout $\epsilon > 0$, posons

$$\Omega_\epsilon = \{\epsilon < \liminf_n \|V_n - x_a\| < 1/\epsilon\}.$$

Sur Ω_ϵ , il existe N_ϵ tel que, pour tout $n \geq N_\epsilon$, $\|V_n - x_a\| \in]\epsilon, 1/\epsilon[$. Etant données les hypothèses faites sur la fonction U ,

$$\inf_{\epsilon < \|x - x_a\| < 1/\epsilon} \langle a - U(x), x_a - x \rangle > 0.$$

D'où, sur Ω_ϵ ,

$$\sum_{k \geq 0} \gamma_k \langle V_k - x_a, U(V_k) - a \rangle \geq \sum_{k \geq N_\epsilon} \gamma_k \inf_{\epsilon < \|x - x_a\| < 1/\epsilon} \langle a - U(x), x_a - x \rangle = \infty$$

Par ailleurs

$$\mathbb{E}(W_{n+1}) = \mathbb{E}(W_n) - 2\gamma_n \mathbb{E}\langle V_n - x_a, U(V_n) - a \rangle = \mathbb{E}(W_0) - 2 \sum_{k=0}^n \gamma_k \mathbb{E}\langle V_k - x_a, U(V_k) - a \rangle$$

ce qui implique pour tout $n \geq 0$, $2 \sum_{k=0}^n \gamma_k \mathbb{E}\langle V_k - x_a, U(V_k) - a \rangle \leq \mathbb{E}(W_0) < \infty$ et donc $\mathbb{P}(\Omega_\epsilon) = 0$ pour tout $\epsilon > 0$. Ainsi, avec probabilité 1,

$$\lim_{n \rightarrow \infty} \|V_n - x_a\| = 0.$$

□

2 Estimation semi-paramétrique

On s'intéresse maintenant au modèle statistique suivant

$$Y_i = f(X_i - \theta) + \varepsilon_i, \quad i = 1, \dots, n \quad (3)$$

où (X_i) et (ε_i) sont des suites indépendantes de variables indépendantes et identiquement distribuées. Dans ce modèle, on suppose que θ et f sont inconnues et on cherche à estimer ces deux quantités à partir des observations $(X_i, Y_i)_{1 \leq i \leq n}$. On parle de modèle semi-paramétrique car la famille de lois de probabilités du modèle est indexée à la fois par un paramètre réel θ et un élément non-paramétrique f .

Des hypothèses sont nécessaires pour rendre ce modèle identifiable. En effet, en toute généralité il peut exister plusieurs fonctions f et plusieurs paramètres θ qui engendrent la même fonction $x \mapsto f(x - \theta)$. On va supposer ici que la fonction f est paire, périodique de période 1 et que $\theta \in \Theta =]-1/4, 1/4[$. Cela restreint les applications possibles, cependant un grand nombre de phénomènes sont représentées par des fonctions périodiques : cycles de températures en météorologie, électrocardiogrammes, sons, courants électriques, etc..

Pour estimer θ à l'aide de l'algorithme de Robbins-Monro, on va chercher une fonction U qui s'annule en θ . Notons g la densité de X_1 . On pose

$$\Phi(t) = \mathbb{E} \left(\frac{\sin(2\pi(X_1 - t))Y_1}{g(X_1)} \right).$$

Proposition 1 *On suppose que les ε_i sont centrées et que la densité g des X_i est à support dans $[-1/2, 1/2]$. Alors la fonction Φ peut s'écrire de la manière suivante :*

$$\Phi(t) = \sin(2\pi(\theta - t))f_1$$

où f_1 est le premier coefficient de Fourier de f : $f_1 = \int_{-1/2}^{1/2} \cos(2\pi x)f(x)dx$.

Démonstration de la proposition 1

Puisque ε_1 est centré, la fonction Φ se réécrit

$$\Phi(t) = \mathbb{E} \left(\frac{\sin(2\pi(X_1 - t))f(X_1 - \theta)}{g(X_1)} \right) = \int_{-1/2}^{1/2} \sin(2\pi(x - t))f(x - \theta)dx.$$

Par un changement de variable, et en utilisant la périodicité de f ,

$$\Phi(t) = \int_{-1/2}^{1/2} \sin(2\pi(x + \theta - t))f(x)dx.$$

On peut alors décomposer

$$\Phi(t) = \cos(2\pi(\theta - t)) \int_{-1/2}^{1/2} \sin(2\pi x)f(x)dx + \sin(2\pi(\theta - t)) \int_{-1/2}^{1/2} \cos(2\pi x)f(x)dx.$$

La parité de f donne finalement le résultat. □

Grâce à cette proposition, on voit facilement que Φ s'annule en θ . Pour pouvoir appliquer le théorème 1 avec $a = 0$, il faut trouver une fonction U qui s'annule en θ mais aussi telle que $U(x)(x - \theta) > 0$ pour tout x . Or $\Phi(x)(x - \theta) = f_1 \sin(2\pi(\theta - x))(x - \theta)$ est toujours du signe opposé à f_1 si $|x - \theta| < 1/2$. On va donc utiliser

$$U = -\text{signe}(f_1)\Phi.$$

On va également modifier un peu l'algorithme de façon à assurer $|x - \theta| < 1/2$. On note K le compact $[-1/4, 1/4]$ et on définit π_K la projection sur K par, pour tout $x \in \mathbb{R}$,

$$\pi_K(x) = \begin{cases} -1/4 & \text{si } x < -1/4 \\ x & \text{si } |x| \leq 1/4 \\ 1/4 & \text{si } x > 1/4 \end{cases}$$

Finalement on peut estimer le paramètre θ par un algorithme de Robbins-Monro projeté :

$$\hat{\theta}_{n+1} = \pi_K \left(\hat{\theta}_n - \gamma_n T_{n+1} \right)$$

où

$$T_{n+1} = -\text{signe}(f_1) \frac{\sin(2\pi(X_{n+1} - \hat{\theta}_n))Y_{n+1}}{g(X_{n+1})}$$

On peut montrer que cet estimateur de θ est consistant :

Théorème 2 *On fait les hypothèses suivantes :*

1. les X_i possèdent une densité g continue à support dans $[-1/2, 1/2]$,
2. les ε_i sont centrées et de variance finie,
3. $|\theta| < 1/4$,

4. f est paire, périodique de période 1, et bornée,

5. $\sum_n \gamma_n = \infty$ et $\sum_n \gamma_n^2 < \infty$.

Alors l'estimateur $\hat{\theta}_n$ converge presque sûrement vers θ . De plus le nombre de fois où la variable aléatoire $\hat{\theta}_n - \gamma_n T_{n+1}$ sort du compact K est fini presque sûrement.

On peut aussi prouver la normalité asymptotique de cet estimateur ce qui permet de construire des intervalles de confiance asymptotiques.

En ce qui concerne la fonction f , des méthodes classiques d'estimation d'une fonction de régression peuvent être appliquées. Dans un modèle de la forme

$$Y_i = b(X_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

une idée pour estimer b en un point x est la suivante : lorsque X_i est proche de x alors Y_i est proche de $b(x)$. Ainsi on pose

$$\hat{b}_n(x) = \frac{\sum_{i=1}^n Y_i \mathbb{1}_{|X_i - x| \leq h_n}}{\sum_{i=1}^n \mathbb{1}_{|X_i - x| \leq h_n}}$$

où (h_n) est une suite de réels positifs tendant vers 0. Cet estimateur est consistant :

Proposition 2 *On suppose que (ε_i) est une suite de variables centrées, indépendante de (X_i) . On suppose que b est continue, ainsi que la densité g des X_i . Alors, si $h_n \rightarrow 0$ et $nh_n \rightarrow \infty$, $\hat{b}_n(x)$ converge presque sûrement vers $b(x)$.*

Démonstration de la proposition 2

Soit $A_n(h) = (2nh)^{-1} \sum_{i=1}^n Y_i \mathbb{1}_{|X_i - x| \leq h}$, $a(h) = (2h)^{-1} \int_{x-h}^{x+h} b(u)g(u)du$, $C_n(h) = (2nh)^{-1} \sum_{i=1}^n \mathbb{1}_{|X_i - x| \leq h}$, $c(h) = (2h)^{-1} \int_{x-h}^{x+h} g(u)du$. Il suffit d'écrire

$$\frac{\hat{b}_n(x)}{b(x)} = \frac{A_n(h_n)}{C_n(h_n)b(x)} = \frac{A_n(h_n)}{a(h_n)} \frac{a(h_n)}{b(x)g(x)} \frac{g(x)}{c(h_n)} \frac{c(h_n)}{C_n(h_n)}$$

et d'appliquer la loi des grands nombres. □

Pour estimer f , on va donc poser

$$\hat{f}_n(x) = \frac{\sum_{i=1}^n Y_i \mathbb{1}_{|X_i - \hat{\theta}_{i-1} - x| \leq h_n}}{\sum_{i=1}^n \mathbb{1}_{|X_i - \hat{\theta}_{i-1} - x| \leq h_n}}.$$

En ajoutant quelques conditions sur f et sur la suite (ε_n) , on peut démontrer que cet estimateur est consistant et asymptotiquement normal.

3 Suggestions

- Implémenter l'algorithme de Robbins-Monro pour l'estimation de la médiane.
- Donner d'autres façons d'estimer un quantile.
- Dans le modèle (3), si on suppose que $f(x) = ax$ et que ε est un bruit gaussien, on retrouve la régression linéaire simple. Comment peut-on estimer θ dans ce cas ?
- Critiquer le choix du modèle (3).
- Détailler la démonstration de la proposition 2.