Agrégation externe de mathématiques
Épreuve de modélisation, option A : Probabilités et Statistiques

(*Texte public*)

Résumé : L'objet de ce texte est de présenter une méthode non paramétrique pour estimer un signal entaché d'erreurs dans le cadre d'un modèle de régression avec dispositif expérimental régulier. Un exemple en vibrométrie laser est donné.

Mots clefs : Régression, estimation, projection.

► *Il est rappelé que le jury n'exige pas une compréhension exhaustive du texte. Vous êtes laissé(e) libre d'organiser votre discussion comme vous l'entendez. Des suggestions de développement, largement indépendantes les unes des autres, vous sont proposées en fin de texte. Vous n'êtes pas tenu(e) de les suivre. Il vous est conseillé de mettre en lumière vos connaissances à partir du fil conducteur constitué par le texte. Le jury appréciera que la discussion soit accompagnée d'exemples traités sur ordinateur.*

Introduction

En vibrométrie laser, le signal observé après émission d'une onde laser continue, réflection sur un objet visé, réception et démodulation, peut s' écrire :

$$X_j = h(j/n) + \varepsilon_{1,j} + i\varepsilon_{2,j}\,, \quad j = 1,\dots,n\,, \tag{1}$$

où $i^2 = -1$, les $\varepsilon_{1,j}$ et $\varepsilon_{2,j}$ sont des variables centrées indépendantes, h est une fonction complexe périodique inconnue. Pour pouvoir mieux décrire le modèle et faire par exemple des prédictions, on désire estimer la fonction h. Il s'agit d'un modèle de régression non paramétrique pour le dispositif expérimental régulier j/n, $j = 1,\dots,n$. L'étude de modèles non paramétrique est d'un intérêt considérable, puisque ces modèles ne supposent pas de connaissance quant à la forme explicite de la fonction de régression h.

1. Simplification du modèle

Etudier le modèle (1) revient en fait à étudier deux modèles réels (donnés par la partie imaginaire et la partie réelle) de la forme

$$Y_j = f(j/n) + \varepsilon_j\,,\ j = 1,\dots,n\,, \tag{2}$$

où les ε_j sont des variables centrées indépendantes telles que $\mathbb{E}(\varepsilon_i^2) = \sigma^2 < \infty$ pour $i = 1,\dots,n$ et f est une fonction réelle périodique.

Dans la suite, on suppose que $f : [0,1] \to \mathbb{R}$ est une fonction de $\mathbb{L}_2[0,1]$ i.e. $\int_0^1 f^2(x)dx < +\infty$. On cherche alors à estimer cette fonction de régression f.

2. Estimateurs par projection

On choisit maintenant pour base orthonormée $\{\varphi_j\}_{j=1}^{\infty}$ de $\mathbb{L}_2[0,1]$ la base trigonométrique définie pour tout $x \in [0,1]$ par :

$$\varphi_1(x) \equiv 1,\ \varphi_{2k}(x) = \sqrt{2}\cos(2\pi kx) \text{ et } \varphi_{2k+1}(x) = \sqrt{2}\sin(2\pi kx)\,, k = 1,2,\ldots.$$

On introduit alors les coefficients de Fourier de f :

$$\theta_j = \int_0^1 f(x)\varphi_j(x)dx\,.$$

Dans un premier temps, on suppose que l'on peut écrire

$$f(x) = \sum_{j=1}^{\infty} \theta_j \varphi_j(x)\,, \tag{3}$$

où la série converge pour tout x dans $[0,1]$. Pour estimer f, nous allons prendre l'approximation de f par sa projection sur les N premières fonctions de la base $\varphi_1,\ldots,\varphi_N$ et remplacer les coefficients de Fourier θ_j par leurs estimateurs. Notons que les coefficients θ_j sont bien approchés par les sommes $n^{-1}\sum_{i=1}^{n} f(i/n)\varphi_j(i/n)$. Ainsi en remplaçant dans ces sommes les inconnues $f(i/n)$ par les observations Y_i, on obtient les estimateurs $\hat{\theta}_j$ des coefficients θ_j donnés par

$$\hat{\theta}_j = \frac{1}{n}\sum_{i=1}^{n} Y_i \varphi_j(\frac{i}{n})\,.$$

La statistique

$$\hat{f}_{n,N}(x) = \sum_{j=1}^{N} \hat{\theta}_j \varphi_j(x)$$

est alors appelée estimateur par projection de $f(x)$. Le paramètre N a un rôle important. Comme nous le verrons à la section 3, son choix est crucial pour établir l'équilibre entre le biais et la variance de notre estimateur.

Pour commencer, on donne des résultats préliminaires dans un cas simple :

Proposition 1. *On suppose ici que f est une fonction de classe $\mathscr{C}^\infty$ sur $[0,1]$. On montre alors que pour tout $j \in \mathbb{N}^*$:*

$$\sqrt{n}\big(\hat{\theta}_j - \theta_j\big) \xrightarrow[n\to\infty]{\mathscr{L}} \mathcal{N}(0,\sigma^2).$$

De plus, pour tout $x \in [0,1]$ avec $N = \lfloor n^c \rfloor$ où $0 < c < 1$ et $\lfloor y \rfloor$ désignant la partie entière de y,

$$\hat{f}_{n,N}(x) \xrightarrow[n\to\infty]{\mathscr{P}} f(x).$$

Preuve de la proposition 1. On a

$$\mathbb{E}[(\hat{\theta}_j - \theta_j)^2] = \mathbb{E}[(\hat{\theta}_j - \mathbb{E}(\hat{\theta}_j))^2] + (\mathbb{E}(\hat{\theta}_j) - \theta_j)^2 := V(\hat{\theta}_j) + b^2(\hat{\theta}_j)\,.$$

Clairement $\mathbb{E}(\hat{\theta}_j) = \alpha_{j,n} + \theta_j$, avec

$$\alpha_{j,n} = \frac{1}{n}\sum_{i=1}^{n} f(i/n)\varphi_j(i/n) - \int_0^1 f(x)\varphi_j(x)dx. \tag{4}$$

Par ailleurs,

$$V(\hat{\theta}_j) = \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\varphi_j(i/n)\Big)^2 = \sigma^2\Big(\frac{1}{n^2}\sum_{i=1}^{n}\varphi_j^2(i/n)\Big).$$

Comme $\{\varphi_j\}_{j=1}^{\infty}$ est la base trigonométrique, on peut vérifier que

$$\frac{1}{n}\sum_{i=1}^{n}\varphi_j(i/n)\varphi_k(i/n) = \delta_{jk} \text{ pour } 1 \le j,k \le n, \tag{5}$$

où δ_{jk} est le symbole de Kronecker. Ainsi $V(\hat{\theta}_j) = \dfrac{\sigma^2}{n}$. Par ailleurs, comme f est $\mathscr{C}^\infty$ sur $[0,1]$, on a $\alpha_{j,n} = o(1/n)$ pour tout j. En utilisant la fonction caractéristique de $\hat{\theta}_j$ et l'indépendance des ϵ_j, on montre le théorème de la limite centrale.
La convergence de l'estimateur $\hat{f}_{n,N}(x)$ s'obtient également grâce à la régularité de la fonction f induisant le comportement asymptotique de ses coefficients de Fourier. □

Comme on désire estimer la fonction f sur tout l'intervalle $[0,1]$ et non seulement localement, il paraît naturel d'étudier les propriétés de l'estimateur par projection $\hat{f}_{n,N}$ pour la norme $\mathbb{L}_2$. On rappelle que si f est dans $\mathbb{L}_2[0,1]$, sa norme $\mathbb{L}_2$ est définie par $\|f\|_2 := \sqrt{\int_0^1 f^2(x)dx}$. Le risque $\mathbb{L}_2$, encore appelé risque quadratique intégré, est défini par

$$MISE = \mathbb{E}\|\hat{f}_{n,N} - f\|_2^2 = \mathbb{E}\int_0^1 \big(\hat{f}_{n,N}(x) - f(x)\big)^2 dx.$$

Pour l'étude du risque quadratique intégré, nous allons nous placer sous l'hypothèse suivante.

Hypothèse (A) : les coefficients de Fourier $\theta_j = \int_0^1 f(x)\varphi_j(x)dx$ de f vérifient

$$\sum_{j=1}^{\infty}|\theta_j| < \infty.$$

Cette hypothèse garantit que la série $\sum_{j=1}^{\infty}\theta_j\varphi_j(x)$ converge absolument pour tout x dans $[0,1]$: on a donc bien la représentation (3).

Proposition 2. *Sous l'hypothèse (A), avec $\alpha_{j,n}$ défini en (4),*

$$MISE = \mathbb{E}\|\hat{f}_{n,N} - f\|_2^2 = \mathscr{A}_{n,N} + \sum_{j=1}^{N}\alpha_{j,n}^2,$$

où

$$\mathscr{A}_{n,N} = \frac{\sigma^2 N}{n} + \rho_N \text{ pour } \rho_N = \sum_{j=N+1}^{\infty}\theta_j^2.$$

Preuve de la proposition 2. On peut écrire que :

$$\begin{aligned} MISE &= \mathbb{E}\int_0^1 \Big(\sum_{j=1}^{N}(\hat{\theta}_j-\theta_j)\varphi_j(x)-\sum_{j=N+1}^{\infty}\theta_j\varphi_j(x)\Big)^2 dx \\ &= \sum_{j=1}^{N}\mathbb{E}[(\hat{\theta}_j-\theta_j)^2]+\sum_{j=N+1}^{\infty}\theta_j^2. \end{aligned}$$

On achève la preuve en utilisant la preuve de la proposition 1. □

Dans la proposition 2, on peut remarquer que les valeurs $\alpha_{j,n}$ sont les résidus issus de l'approximation des sommes par les intégrales. On verra dans la section 3 que pour certaines classes de fonctions (voir Définition 1) ces résidus sont négligeables par rapport à $\mathscr{A}_{n,N}$ quand n est grand. La quantité $\mathscr{A}_{n,N}$ représente alors la partie principale du risque intégré de l'estimateur par projection $\hat{f}_{n,N}$. Les termes $\frac{\sigma^2 N}{n}$ et ρ_N sont respectivement interprétés comme le terme de variance et le terme de biais de l'estimateur $\hat{f}_{n,N}$. Le paramètre N est donc choisi pour permettre le meilleur compromis entre ces deux termes.

3. Espace fonctionnel et vitesse de convergence

Afin d'établir des vitesses de convergence pour notre estimateur, on sera amené à supposer que f appartient à des classes de fonctions régulières, au sens de la définition ci-dessous.

Définition 1. *Soit $\beta \in \{1,2,\ldots\}$, $L>0$. Notons $f^{(i)}$ la dérivée $i^{ème}$ de f. La classe fonctionnelle de Sobolev $W(\beta,L)$ est définie par*

$$W(\beta,L)=\Big\{f\in[0,1]\to\mathbb{R}:\ f^{(\beta-1)}\ \textit{est absolument continue et}\ \int_0^1 (f^{(\beta)}(x))^2dx\le L^2\Big\}.$$

La classe de Sobolev périodique $W^{per}(\beta,L)$ est définie par

$$W^{per}(\beta,L)=\Big\{f\in W(\beta,L):\ f^{(j)}(0)=f^{(j)}(1),\ j=0,1,\ldots,\beta-1\Big\}.$$

Notons que toute fonction $f\in W^{\mathrm{per}}(\beta,L)$ admet la représentaion (3) où la suite $\theta=\{\theta_j\}_{j=1}^{\infty}$ appartient à l'espace $\ell^2(\mathbb{N})=\{\theta:\ \sum_{j=1}^{\infty}\theta_j^2<\infty\}$. Par ailleurs, si l'on définit les nombres

$$a_j=\begin{cases} j^{\beta} & \text{pour } j \text{ pair} \\ (j-1)^{\beta} & \text{pour } j \text{ impair,}\end{cases}$$

et l'ensemble

$$\Theta(\beta,Q)=\{\theta\in\ell^2(\mathbb{N}):\quad \sum_{j=1}^{\infty}a_j^2\theta_j^2\le Q\},$$

où $Q=L^2/\pi^{2\beta}$, il existe un isomorphisme entre $\Theta(\beta,Q)$ et $W^{\mathrm{per}}(\beta,L)$ pour $\beta\in\mathbb{N}^*$ et $L>0$.

En effet, pour montrer que si $f\in W^{\mathrm{per}}(\beta,L)$ alors $\theta\in\Theta(\beta,Q)$, on définit tout d'abord les coefficients de Fourier de $f^{(j)}$, pour $j=1,\ldots,\beta$, par rapport à la base trigonométrique. Notons : $s_1(j)=0$, $s_{2k}(j):=\sqrt{2}\int_0^1 f^{(j)}(t)\cos(2\pi kt)dt$ et $s_{2k+1}(j):=\sqrt{2}\int_0^1 f^{(j)}(t)\sin(2\pi kt)dt$

pour $k = 1, 2, \ldots$. On peut alors démontrer (par récurrence) que

$$s_{2k}^2(\beta) + s_{2k+1}^2(\beta) = (2\pi k)^{2\beta}\big(s_{2k}^2(0) + s_{2k+1}^2(0)\big) = (2\pi k)^{2\beta}\big(\theta_{2k}^2 + \theta_{2k+1}^2\big), \text{ pour } k = 1, 2, \ldots.$$

Ainsi via l'égalité de Parseval,

$$\int_0^1 (f^{(\beta)}(t))^2 \, dt = \pi^{2\beta} \sum_{j=1}^{\infty} a_j^2 \theta_j^2,$$

ce qui démontre l'implication. La réciproque est un peu plus technique. □

Le théorème suivant indique que pour N de l'ordre de $n^{1/(2\beta+1)}$, la vitesse de convergence en norme $\mathbb{L}_2$ de l'estimateur par projection $\hat{f}_{n,N}$ sur la classe $W^{\text{per}}(\beta, L)$ vaut : $n^{-\beta/(2\beta+1)}$.

Théorème 1. *Supposons que l'hypothèse (A) est vérifée, $\beta \in \{1, 2, \ldots\}$, $L > 0$, et définissons pour $\alpha > 0$, un entier*

$$N = \lfloor \alpha n^{1/(2\beta+1)} \rfloor$$

Alors, avec une constante $C < \infty$, l'estimateur par projection $\hat{f}_{n,N}$ vérifie :

$$\limsup_{n\to\infty} \sup_{f \in W^{per}(\beta,L)} \mathbb{E}\big[n^{\frac{2\beta}{2\beta+1}} \| \hat{f}_{n,N} - f \|_2^2 \big] \le C. \tag{6}$$

Il est possible d'obtenir un résultat analogue si l'on remplace $W^{\text{per}}(\beta, L)$ par $W(\beta, L)$ et si l'on choisit une base $\{\varphi_j\}$ autre que la base trigonométrique.

Preuve du théorème 1. D'après la proposition 2, on a

$$\mathbb{E}\| \hat{f}_{n,N} - f \|_2^2 = \mathscr{A}_{n,N} + \sum_{j=1}^{N} \alpha_{j,n}^2,$$

avec pour $N = \lfloor \alpha n^{1/(2\beta+1)} \rfloor$

$$\mathscr{A}_{n,N} \le \sigma^2 \alpha n^{\frac{-2\beta}{2\beta+1}} + \rho_N.$$

En utilisant la monotonie de la suite a_j, on obtient que

$$\rho_N \le \frac{1}{a_{N+1}^2} \sum_{j=1}^{\infty} \theta_j^2 a_j^2 = O\big(n^{\frac{-2\beta}{2\beta+1}}\big),$$

où $O(\cdot)$ est uniforme en $f \in W^{\text{per}}(\beta, L)$. Pour montrer (6), on prouve que $\sum_{j=1}^{N} \alpha_{j,n}^2 = O\big(n^{\frac{-2\beta}{2\beta+1}}\big)$, où $O(\cdot)$ est uniforme en $f \in W^{\text{per}}(\beta, L)$. Celà découle du lemme suivant :

Lemme 1. *Pour la base trigonométrique $\{\varphi_j\}_{j=1}^{\infty}$, les résidus $\alpha_{j,n}$ définis en (4) vérifient*

(i): *Si $\sum_{j=1}^{\infty} |\theta_j| < \infty$, alors $\max_{1\le j\le n} |\alpha_{j,n}| \le 2\sum_{m=n+1}^{\infty} |\theta_m|$.*

(ii): *Si $\theta = \{\theta_j\}_{j=1}^{\infty} \in \Theta(\beta, Q)$, $\beta > 1/2$, alors $\max_{1\le j\le n} |\alpha_{j,n}| \le C_{\beta,Q} n^{-\beta+1/2}$ pour une constante $C_{\beta,Q} < \infty$ qui ne dépend que de β et Q.*

Pour démontrer le point (i), on écrit tout d'abord que

$$\alpha_{j,n} = \Big(\frac{1}{n}\sum_{i,k=1}^{n} \theta_k \varphi_k(i/n)\varphi_j(i/n) - \theta_j\Big) + \frac{1}{n}\sum_{i=1}^{n}\sum_{k=n+1}^{\infty} \theta_k \varphi_k(i/n)\varphi_j(i/n).$$

On en déduit alors en utilisant (5) que pour tout $1 \le j \le n$,

$$|\alpha_{j,n}| = \Big|\sum_{k=n+1}^{\infty} \theta_k\big(\frac{1}{n}\sum_{i=1}^{n} \varphi_k(i/n)\varphi_j(i/n)\big)\Big| \le 2\sum_{k=n+1}^{\infty} |\theta_k|.$$

Le point (ii) se démontre en partant de la majoration obtenue en (i) suivie d'une utilisation judicieuse de l'inégalité de Cauchy-Schwarz. □

Ce résultat donne une vitesse de convergence pour l'estimateur. Dans le cas où β tend vers l'infini par exemple, on obtient une vitesse en $1/\sqrt{n}$. Ceci-dit, cette vitesse est obtenue avec un choix de N dépendant de la régularité a priori inconnue de la fonction f. On dit alors que la procédure d'estimation n'est pas adaptative. En pratique, comme $\mathscr{A}_{n,N} \to \infty$ lorsque $N \to \infty$, pour tout n fixé, il existe toujours un point qui minimise $\mathscr{A}_{n,N}$. On définit alors $\tilde{N}_n = \arg\min_{N\ge 1} \mathscr{A}_{n,N}$ et on prend pour estimateur de f la quantité $\hat{f}_{n,\tilde{N}_n}$.

Suggestions pour le développement

► *Soulignons qu'il s'agit d'un menu à la carte et que vous pouvez choisir d'étudier certains points, pas tous, pas nécessairement dans l'ordre, et de façon plus ou moins fouillée. Vous pouvez aussi vous poser d'autres questions que celles indiquées plus bas. Il est très vivement souhaité que vos investigations comportent une partie traitée sur ordinateur et, si possible, des représentations graphiques de vos résultats.*

— *Modélisation.*
  — Quelles sont les limites du modèle proposé ? Que pourriez vous proposer comme modèle pour prendre en compte plus de phénomènes.
  — Que feriez vous dans un cadre paramétrique ?
  — Comment expliquez-vous les variations de N et de la vitesse de convergence en fonction du paramètre β ?
  — Expliquez pourquoi $\hat{f}_{n,\tilde{N}_n}$ est un bon candidat pour estimer f.

— *Développements mathématiques.* Vous pouvez compléter les preuves des résultats énoncés dans ce texte. La base trigonométrique vous semble t-elle adaptée pour tout type de fonctions f ? Est-il possible de travailler avec d'autres bases ?

— *Etude numérique.* En vibrométrie laser, lorsque la cible est composée de plusieurs réflecteurs vibrant de façon sinusoïdale, h est de la forme :

$$h(t) = \sum_{m=1}^{M} a_m \exp\left[\frac{4i\pi\gamma_m}{\lambda}\cos(2\pi F_f t)\right].$$

Construire un estimateur par projection pour un échantillon bruité en prenant par exemple :

$$M = 1\,,\ a_1 = 0.06\,,\ \gamma_1 = 35 \times 10^{-6}\ \text{et}\ F_f = 48Hz\,.$$