M2 recherche fiche 8: Estimation d'une fonction de régression par projection

Emeline Schmisser, `emeline.schmisser@math.univ-lille1.fr`, bureau 314 (bâtiment M3).

On considère une suite de variables (x_i, y_i) i variant de 1 à n tels que :
- les x_i soient indépendants et identiquement distribués suivant une loi h connue.
- $y_i = f(x_i) + \varepsilon_i$, les variables ε_i étant centrées et iid indépendants des x_i et de variance σ^2.

Nous voulons estimer la fonction f sur un compact $[a, b]$.

1 Estimation sur un espace de dimension fixé

Soit V un espace vectoriel de dimension finie de $L^2([a, b])$ engendré par les fonctions $g_1, \ldots, g_D$ orthonormales. Supposons dans un premier temps que f appartient à V. Alors f peut s'écrire :

$$f = \sum_{k=1}^{D} a_k g_k \quad \text{avec} \quad a_k = \int f(x) g_k(x)$$

et estimer f revient à estimer les coefficients a_k. Posons

$$\hat{a}_k = \frac{1}{n} \sum_{i=1}^{n} \frac{g_k(x_i) y_i}{h(x_i)} \mathbf{1}_{x_i \in [a,b]}$$

Exercice 1 (Estimation de a_k)

1. Montrer que $\hat{a}_k$ est un bon estimateur de a_k (Calculer sa moyenne et sa variance).
2. Est-ce que l'estimateur obtenu est une densité ?
3. En déduire un estimateur de a_k si les x_i sont distribués suivant une densité h inconnue. On admettra que cet estimateur est convergent.

Exercice 2 (Estimation de f)
Notons $\hat{f}_V = \sum_{k=1}^{D} \hat{a}_k g_k$.

1. Si f appartient à V, montrer que $\hat{f}_V(x)$ est un estimateur sans biais de $f(x)$ pour tout x. Calculer le risque L^2 de $\hat{f}_V(x)$.
2. Si f n'appartient pas à V, notons f_V son projeté sur l'espace V. Montrer que le risque L^2 de $\hat{f}_V$ se décompose de la façon suivante :

$$\mathcal{R}(\hat{f}_V) = \underbrace{\|f - f_V\|_{L^2}^2}_{biais} + \underbrace{\mathbb{E}\left(\left\|\hat{f}_V - f_V\right\|_{L^2}^2\right)}_{variance}$$

Majorer le terme de variance.

Si V vérifie "les conditions usuelles" et que f appartient à l'espace de Besov $\mathcal{B}^{\alpha}_{2,\infty}$, alors $\|f - f_V\|_{L^2}^2 \leq D^{-2\alpha}$.
En général, on se donne une suite d'espaces vectoriels V_m de dimension finie D_m. Sur chacun d'eux, on calcule un estimateur $\hat{f}_m = \hat{f}_{V_m}$.

1. Si la régularité α de f est connue, quelle valeur doit-on choisir pour m ? Quelle est dans ce cas le risque de l'estimateur ?

Exemple de suites d'espaces S_m sur $[0,1]$: (vérifiant ces conditions usuelles)

- les polynômes trigonométriques : $V_m = \text{Vect}(sin(2\pi kx), cos(2\pi kx), 0 \leq k \leq m)$ et $D_m = 2m+1$
- les polynômes de degré $\leq r$: $V_m = \text{Vect}(x^j \mathbf{1}_{x \in [k/2^m,(k+1)/2^m]}, 0 \leq j \leq r, 0 \leq k \leq 2^m - 1)$, $D_m = r2^m$.
- les polynômes de degré $\leq r$ $\mathcal{C}^{r-1}$ (les fonctions splines). $D_m = 2^m + r$.
- les ondelettes (et en particulier la base de Haar). $D_m = 2^m$

Exercice 3 (avec R)

1. Simuler 1000 variables x_i uniformes sur $[-2,2]$ et 1000 variables ε_i normales centrées réduites.
2. Créer une fonction f dépendant de x tel que $f(x) = cos(x)$. Il est important de créer une fonction spéciale, cela nous simplifiera les choses quand on voudra considérer une autre fonction (on n'aura pas besoin de tout changer).
3. Construire les variables $y_i = f(x_i) + \varepsilon_i$. Représenter les points x_i, y_i sur le graphe et tracer la fonction f.
   Maintenant, on oublie f et on veut l'estimer sur $[-2,2]$.
4. Donner une base de fonctions trigonométrique sur $[-2,2]$.
5. On pose $\delta = 0.01$ et on va construire un estimateur pour $z = -2, -2+\delta, \ldots, 2-\delta, 2$.
6. Construire une fonction *est_trigo_m* qui étant donné m, x et y, calcule $\hat{f}_m$ sur la base des polynômes trigonométriques et superpose cette fonction au graphe existant. Pour cela, on utilise une boucle (utiliser une fonction récursive nous compliquerait la tâche par la suite). Attention : la boucle n'est pas très longue, elle dépend uniquement du nombre de coefficients à calculer.
7. Faire varier m pour obtenir différents estimateurs de f.
8. Remplacer maintenant f par x^2 ou une autre fonction non trigonométrique. Faire de nouveau varier m pour obtenir différents estimateurs.

2 Estimation adaptative

On est maintenant capable de construire une collection d'estimateurs de f. Il reste à choisir m. Dans le cas des ondelettes, les méthodes de seuillage sont très utilisées : on ne garde que les coefficients supérieurs à un certain seuil.
Ici, on va utiliser une autre méthode : la pénalisation. Considérons la quantité

$$R_n(\hat{f}_m) = \frac{1}{n} \sum_{i=1}^{n} \frac{\left(y_i - \hat{f}_m(x_i)\right)^2}{h(x_i)}$$

Exercice 4 ()

1. Montrer que $\mathbb{E}(\mathbb{R}_n(\hat{f})) = biais$. Pour obtenir un estimateur sans biais du risque, il suffit donc de compenser le terme de variance qui n'est pas pris en compte. On considère donc la quantité $\gamma(m) = R_n(\hat{f}_m) + pen(m) = R_n(\hat{f}_m) + 2 * variance$
2. Créer une fonction *est_trigo_2* en modifiant la fonction est_trigo_m. Cette fonction calcule $\gamma(m)$.
3. Créer une fonction *est_trigo* qui dépend de x et de y et qui calcule le meilleur estimateur possible (celui qui minimise $\gamma(m)$). On fait varier m de 0 à 20, Pour chaque m, on calcule $\gamma(m)$ et on garde en mémoire la valeur minimale de $\gamma(m)$ et la valeur de m pour laquelle elle a été obtenue. À la fin, on calcule l'estimateur de f pour cette valeur et on le trace.
4. Appliquer la fonction *est_trigo* aux données. Modifier f et voir ce que cela donne.

3 Ondelettes et base de Haar

3.1 Définitions

Je ne donne pas ici de définition rigoureuse, on peut aller voir Meyer (Ondelettes et opérateurs, tome 1, chapitre 2) pour tous les théorèmes sur les ondelettes.

Une base d'ondelettes est donnée par deux fonctions : l'ondelette père ϕ et l'ondelette mère ψ. Notons

$$\phi_k = \phi(. - k) \quad \text{et} \quad \psi_{k,m} = 2^{m/2}\psi(2^m. - k)$$

et V_m l'espace engendré par ces fonctions. Les fonctions ϕ et ψ vérifient les propriétés suivantes :

- Les fonctions $\phi_k, k \in \mathbb{Z}$ sont orthonormées.
- Les fonctions $\psi_{k,0}, k \in \mathbb{Z}$) sont orthornormées. Cela induit en particulier que pour tout m, les fonctions $\psi_{m,k}, k \in \mathbb{Z}$ sont orthonormées.
- Pour tous k, l, m, les fonctions $\phi_k, \psi_{j,m}$ sont orthogonales.
- Pour tous k, l, m, p, $m \neq p$, les fonctions $\psi_{k,m}, \psi_{j,p}$ sont orthogonales.
- La fonction ψ est d'intégrale 0.

Pour tout m, les fonctions $(\phi_k, \psi_{j,k}, k \in \mathbb{Z}, 0 \leq j \leq m)$ sont orthonormées. On note V_m l'espace engendré par ces fonctions et f_m le projeté orthogonal de f sur V_m :

$$f_m = \sum_{k\in\mathbb{Z}} a_k\phi_k + \sum_{j=0}^{m}\sum_{k\in\mathbb{Z}} b_{k,j}\psi_{k,j}$$

Intuitivement : les fonctions ϕ calculent une "moyenne" de f sur des plages $[k, k+1]$ et les fonctions ψ rajoutent des oscillations. Décomposer une fonction en ondelettes nous permet de voir où et à quelles fréquences sont les oscillations. Elles permettent une analyse plus fine des fonctions que les polynômes trigonométriques qui ne nous donnent des informations que sur la fréquence des oscillations et pas sur leur localisation.

Il existe des méthodes pour trouver des bases d'ondelettes (encore une fois, cf Meyer) une fois qu'on a une fonction qui a de "bonnes propriétés" (on peut créer des ondelettes à partir des fonctions splines, par exemple).

Plus ϕ et ψ sont régulières, plus on a de bonnes propriétés sur les estimateurs (mieux on sait contrôler le terme de biais).

3.2 Base de Haar

La base de Haar n'est pas très régulière (on dit qu'elle est 0-régulière), mais elle est très facile à construire.

$$\phi(x) = \mathbf{1}_{x\in[0,1]} \quad \psi(x) = -\mathbf{1}_{x\in[0,1/2]} + \mathbf{1}_{x\in[1/2,1]}$$

Montrer que la base de Haar vérifie bien les propriétés demandées.

Exercice 5 (avec R)

1. Simuler 1000 variables aléatoires x_i de loi uniforme sur $[0, 1]$, et 1000 variables ε_i normales centrées réduites.
2. Construire les variables $y_i = f(x_i) + \varepsilon_i$, et tracer le graphe de f sur $[0, 1]$
3. On veut estimer f sur $[0, 1]$. Entre quoi et quoi fait-on varier k (en fonction de m) ?
4. Créer une fonction *haar* qui étant donné x, k, m, calcule $\psi(x, k, m)$.
5. Créer une fonction $coef\_haar$ qui étant donné x, y, k, m, calcule $\hat{a}_{k,m}$
6. Créer une fonction est_haar_m qui superpose $\hat{f}_m$ (calculé avec la base de Haar) à la fonction f.