Université Paris-Sud 2014-2015
Projet en modélisation L3 MINT

Estimation non-paramétrique pour l'étude de la pollution atmosphérique

On s'intéresse ici à la quantité de monoxyde d'azote dans l'air à Boston en 1978. On dispose de 506 mesures relevées dans différentes parties de la ville. Calculer la moyenne ou la variance de cette variable est trop sommaire, on cherche à avoir une information plus complète sur sa distribution.

On modélise alors cette quantité de monoxyde d'azote comme une variable aléatoire réelle X, et on note f la densité de X. On dispose ainsi d'observations $X_1, \ldots, X_n$ que l'on considère comme des variables aléatoires indépendantes et identiquement distribuées de densité f.

Le but de ce projet est d'estimer cette densité f. On dit qu'il s'agit d'estimation non-paramétrique car on ne cherche pas à estimer un paramètre $\theta \in \mathbb{R}^d$ mais une fonction. Pour estimer f, on va utiliser des séries de Fourier. On suppose que la densité recherchée est à support compact dans un intervalle. Grâce à une renormalisation affine, on peut supposer que f est définie sur $[-\pi, \pi]$.

Pour toute fonction $g : [-\pi, \pi] \to \mathbb{R}$, on notera

$$\|g\|^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |g(x)|^2 dx$$

On supposera que $\|f\| < \infty$.

1 Approximation d'une fonction

L'idée pour estimer une fonction est de se ramener à un problème d'estimation plus simple. Pour cela, on réalise une approsimation de f grâce à son développement en série de Fourier. On est alors ramené à l'estimation des coefficients de Fourier (on a remplacé une fonction par des réels ce qui simplifie le problème). Les coefficients de Fourier de f seront notés :

$$a_k(f) = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos(kx), \qquad b_k(f) = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin(kx).$$

Pour tout $D \geq 2$ entier, on note aussi

$$S_D = \mathrm{Vect}\{1, \cos(kx), \sin(kx), 1 \leq k \leq D-1\}.$$

Question 1. Quel est le développement de f en série de Fourier ? Quel est celui-ci de f_D la projection orthogonale de f sur S_D ? Pourquoi peut-on dire que f_D est une approximation de f ?

Si on remplace l'estimation de f par f_D, on commet nécessairement une erreur. Pour quantifier celle-ci, va va chercher à évaluer l'erreur d'approximation $\|f - f_D\|$.

Question 2. Que peut-on attendre de $\|f - f_D\|$ quand D augmente ? Montrer que

$$\|f - f_D\|^2 = \frac{1}{2}\sum_{k=D}^{\infty} |a_k(f)|^2 + \frac{1}{2}\sum_{k=D}^{\infty} |b_k(f)|^2.$$

Pour avoir une idée plus précise, il faut faire des hypothèses sur f. On suppose donc maintenant que f est de classe C^p avec $p \geq 1$.

Question 3. Calculer les coefficients de Fourier $a_k(f^{(p)})$ et $b_k(f^{(p)})$ de $f^{(p)}$ en fonction de $a_k(f)$ et $b_k(f)$. En utilisant le théorème de Parseval pour $f^{(p)}$, montrer qu'il existe une constante K tel que

$$\|f - f_D\|^2 \leq KD^{-2p}.$$

2 Estimation des coefficients

Maintenant qu'on dispose d'une bonne approximation de f, on va estimer f_D à partir des données $X_1, \ldots, X_n$.

Question 4. Montrer que pour tout $k \geq 0$, $a_k(f) = \dfrac{1}{\pi}\mathbb{E}[\cos(kX_1)]$. En déduire un estimateur $\hat{a}_k$ de $a_k(f)$. Proposer de même un estimateur $\hat{b}_k$ de $b_k(f)$.

Il est alors naturel de poser

$$\forall x \in [-\pi, \pi] \qquad \hat{f}_D(x) = \frac{\hat{a}_0}{2} + \sum_{k=1}^{D-1} \hat{a}_k \cos(kx) + \sum_{k=1}^{D-1} \hat{b}_k \sin(kx).$$

A nouveau, on commet une erreur en remplaçant f_D par $\hat{f}_D$. On appelle $\|f_D - \hat{f}_D\|$ l'erreur stochastique.

Question 5. Que peut-on attendre de $\|f_D - \hat{f}_D\|$ quand D augmente ? Montrer que

$$\|f_D - \hat{f}_D\|^2 = \frac{|\hat{a}_0 - a_0(f)|^2}{4} + \frac{1}{2}\sum_{k=1}^{D-1} |\hat{a}_k - a_k(f)|^2 + \frac{1}{2}\sum_{k=1}^{D-1} |\hat{b}_k - b_k(f)|^2.$$

Question 6. Montrer que

$$\forall k \in \{0, \ldots, D-1\} \qquad \mathbb{E}\left[|\hat{a}_k - a_k(f)|^2\right] \leq \frac{1}{\pi^2 n}.$$

et en déduire une majoration de $\mathbb{E}\|f_D - \hat{f}_D\|^2$.

3 Simulation numérique

Pour évaluer numériquement notre estimateur, on peut simuler un échantillon $X_1, \dots, X_n$ de densité f connue, et observer si $\hat{f}_D$ estime bien f. On pourra ensuite le calculer pour des vraies données de densité f inconnue.

On considère la densité Beta $B(2,3)$ renormalisée sur $[-\pi, \pi]$:

$$f : x \mapsto \frac{3}{4\pi}\left(1+\frac{x}{\pi}\right)\left(1-\frac{x}{\pi}\right)^2$$

Si Y suit une loi Beta $B(2,3)$ alors $X = 2\pi Y - \pi$ est de densité f.

Question 7. Montrer que si Y suit une loi Beta $B(2,3)$ alors $X = 2\pi Y - \pi$ est de densité f. Simuler 100 observations de densité f.

Question 8. Programmer le calcul de $\hat{f}_D$ pour différentes valeurs de D. Comparer avec f. On pourra éventuellement tracer aussi $\|f - f_D\|^2$ et $\mathbb{E}\|f_D - \hat{f}_D\|^2$ en fonction de D.

Question 9. Adapter le programme pour des données à valeurs dans $[a, b]$ au lieu de $[-\pi, \pi]$. Estimer la densité pour la distribution du monoxyde de carbone à Boston (fichier `Boston2.mat`).

4 Risque de l'estimateur et choix de D

On appelle risque de l'estimateur l'erreur moyenne d'estimation suivante : $\mathbb{E}\|f - \hat{f}_D\|^2$. L'estimateur est considéré d'autant plus performant que cette quantité est petite.

Question 10. En utilisant le théorème de Pythagore, montrer que

$$\mathbb{E}\|f - \hat{f}_D\|^2 = \|f - f_D\|^2 + \mathbb{E}\|f_D - \hat{f}_D\|^2$$

Pourquoi parle-t-on de compromis biais-variance ? En utilisant les parties précédentes, déterminer pour quel choix de D (en fonction de n et p) le risque est le plus faible, et calculer le risque pour cette valeur.

Question 11. En pratique, peut-on choisir cette valeur optimale de D ? Commenter.