Approximation d'une densité inconnue

Mots-clés : Analyse, Convergences de variables aléatoires, Statistique,...

Ce travail est une introduction à la statistique non paramétrique[1]. Le problème est le suivant : on observe un échantillon

$$X_1, X_2, \ldots, X_n$$

de variables aléatoires i.i.d. dont la loi est supposée avoir une densité f qui est inconnue. On suppose simplement que f est continue par morceaux et on cherche à "deviner" f, c'est-à-dire trouver une fonction $\hat{f}_n$ (qui dépend des données et est donc aléatoire) qui est une bonne approximation de f, au moins quand n est grand.

On va chercher à étudier de façon théorique et expérimentale deux méthodes : une méthode utilisant la fonction de répartition empirique et une méthode *par régularisation*.

Notations

On note F_n la fonction de répartition empirique associée aux X_k :

$$\begin{array}{rccc} F_n : & \mathbb{R} & \to & [0,1] \\ & x & \mapsto & \frac{1}{n}\text{card}\left\{k \leq n,\ X_k \leq x\right\} \end{array}$$

(F_n est donc aléatoire). Pour pouvoir quantifier une notion de "bonne" approximation, nous allons munir l'ensemble des densités de la distance L^p :

$$\| f - g \|_{L^p} = \left(\int_{\mathbb{R}} |f(x) - g(x)|^p dx \right)^{1/p}.$$

Utilisation de la fonction de répartition empirique

Pour cette première méthode, nous allons approcher f par une sorte de dérivée de F_n. On choisit pour tout n une "taille" de fenêtre h_n telle que la suite $(h_n)_{n\geq 1}$ vérifie

$$(h_n) \to 0, \qquad (nh_n) \to +\infty. \tag{1}$$

On pose alors pour tout x

$$\hat{f}_n(x) = \frac{F_n(x + h_n) - F_n(x - h_n)}{2h_n}.$$

Remarquons que $\hat{f}_n$ est une fonction aléatoire, car elle dépend des données à travers F_n.

1. Vérifier que $\hat{f}_n$ est bien une densité. Quelle est la régularité de $\hat{f}_n$?

[1] On parle de statistique *non* paramétrique par opposition à la situation où l'on suppose que f est dans une famille décrite par un ou plusieurs paramètres, par exemple la famille des lois exponentielles.

Solution: $\hat{f}_n$ est bien sûr ≥ 0, on a

$$\hat{f}_n(x) = \frac{1}{2nh_n} \sum_{k=1}^{n} \mathbf{1}_{X_k \in [x \pm h_n]} = \frac{1}{2h_n} \sum_{k=1}^{n} \mathbf{1}_{x \in [X_k \pm h_n]}$$

$$\begin{aligned} \int \hat{f}_n(x)dx &= \frac{1}{2nh_n} \sum_{k=1}^{n} \int \mathbf{1}_{x \in [X_k \pm h_n]} dx \\ &= \frac{1}{2nh_n} \sum_{k=1}^{n} 2h_n = 1 \end{aligned}$$

2. **Expérimental.** Pour différentes densités f (penser à varier la régularité), simuler un échantillon de n v.a. de densité f et présenter un tracé de $\hat{f}_n$.
3. Soit x fixé, quelle est la loi de $2nh_n \times \hat{f}_n(x)$?

Solution: $2nh_n \times \hat{f}_n(x)$ représente le nombre de réalisations qui sont tombées dans $[x \pm h_n]$, on a donc

$$2nh_n \times \hat{f}_n(x) \sim \text{Binom}\left(n, \int_{[x \pm h_n]} f(t)dt\right).$$

4. Pour x fixé, on définit l'erreur quadratique $\text{EQ}(x)$ par

$$\text{EQ}(x) = \mathbb{E}\left[\left(\hat{f}_n(x) - f(x)\right)^2\right].$$

Vérifier que

$$\text{EQ}(x) = \left(\mathbb{E}\left[\hat{f}_n(x)\right] - f(x)\right)^2 + \text{Var}\left(\hat{f}_n(x)\right)$$

et démontrer que si f est continue en x on a $\mathbb{E}\left[\hat{f}_n(x)\right] \overset{n \to +\infty}{\to} f(x)$.

Solution:

$$\mathbb{E}\left[\hat{f}_n(x)\right] = \frac{1}{2h_n} \int_{x-h_n}^{x+h_n} f(t)dt \to f(x)$$

car f est continue en x (théorème de la moyenne par exemple).

5. En déduire que $\hat{f}_n$ est une bonne approximation de f dans le sens suivant : pour presque tout x

$$\hat{f}_n(x) \overset{\text{prob.}}{\to} f(x).$$

Solution: On a une binomiale :

$$\begin{aligned}\mathrm{Var}\left(\hat{f}_n(x)\right) &= \frac{1}{4nh_n^2}\int_{x-h_n}^{x+h_n} f(t)dt\left(1-\int_{x-h_n}^{x+h_n} f(t)dt\right)\\ &\sim \frac{1}{4nh_n^2}2h_nf(x)\left(1-2h_nf(x)\right) \sim f(x)\frac{1}{2nh_n}.\end{aligned}$$

Finalement, d'après la décomposition biais/variance, $\hat{f}_n(x) \xrightarrow{L^2} f(x)$, donc en probabilité.

6. **Expérimental.** Choisir arbitrairement une densité f et un réel x et essayer de mettre en évidence avec une simulation le fait que la suite de variables aléatoires $\hat{f}_n(x)$ vérifie un TCL (par exemple en traçant un histogramme d'un grand nombre de réalisations de v.a. $\hat{f}_n(x)$).

Méthode par régularisation

On va maintenant construire une autre approximation $\hat{g}_n$ en régularisant les données à l'aide d'une suite de fonctions (K_n). Pour simplifier les preuves on suppose dans cette partie que f est continue à support compact, mais la méthode marche pour des fonctions continues par morceaux.

On pose

$$K(x) = \begin{cases} \frac{15}{16}(x-1)^2(x+1)^2 & \text{si } -1 \le x \le 1,\\ 0 & \text{sinon.}\end{cases}$$
$$K_n(x) = n^\alpha K(n^\alpha x),$$

où α est un paramètre tel que $0 < \alpha < 1$. On vérifie facilement que K et K_n sont des densités. On définit alors

$$\hat{g}_n(x) = \frac{1}{n}\sum_{k=1}^{n} K_n(x - X_k).$$

7. Vérifier que $\hat{g}_n$ est bien une densité. Que peut-on dire de la régularité de $\hat{g}_n$?

Solution: On vérifie facilement que chaque fonction $x \mapsto K_n(x - X_k)$ est $\mathcal{C}^1$ (y compris en $|x - X_k| = 1$), donc $\hat{g}_n$ est de classe $\mathcal{C}^1$.

8. **Expérimental.** Choisir une densité f et présenter différents tracés de $\hat{g}_n$ (en faisant varier n et α).
9. L'erreur quadratique $\mathrm{EQ}(x)$ est définie comme précédemment, on a donc également l'égalité

$$\mathrm{EQ}(x) = \left(\mathbb{E}\left[\hat{g}_n(x)\right] - f(x)\right)^2 + \mathrm{Var}\left(\hat{g}_n(x)\right).$$

Démontrer que pour tout x

$$\mathbb{E}\left[\hat{g}_n(x)\right] \to f(x).$$

(on pourra utiliser le fait que pour tout $\delta > 0$, $\int_{|y|>\delta} K_n(y)dy \to 0$)

Solution:

$$\begin{aligned}\mathbb{E}\left[\hat{g}_n(x)\right] &= \mathbb{E}[K_n(x-X_1)] \\ &= \int K_n(x-t)f(t)dt = \int_{x-1/n^\alpha}^{x+1/n^\alpha} K_n(x-t)f(t)dt \\ &= \int_{-1/n^\alpha}^{1/n^\alpha} n^\alpha K(n^\alpha u)f(x-u)du \\ &= \int_{-1}^{1} K(s)f(x-\tfrac{s}{n^\alpha})ds \to \int_{-1}^{1} K(s)f(x)ds = f(x)\end{aligned}$$

par convergence dominée.

10. Démontrer que pour tout x

$$\hat{g}_n(x) \overset{\text{prob.}}{\to} f(x).$$

Solution: On cherche à contrôler la variance.

$$\begin{aligned}\mathrm{Var}\left(\hat{g}_n(x)\right) &= \mathrm{Var}\left(\frac{1}{n}\sum_{k=1}^{n} K_n(x-X_k)\right) \\ &= \frac{1}{n}\mathrm{Var}\left(K_n(x-X_1)\right) \leq \frac{1}{n}\mathbb{E}[K_n(x-X_1)^2].\end{aligned}$$

Maintenant,

$$\mathbb{E}[K_n(x-X_1)^2] = \int K_n(x-t)^2 f(t)dt = [\dots] \sim n^\alpha f(x)\int K^2.$$

On conclut avec la décomposition biais/variance.

11. **Expérimental.** Présenter brièvement, à partir de l'étude théorique et des simulations, les avantages/inconvénients de α petit (proche de zéro) ou grand (proche de 1) dans le construction de $\hat{g}_n$.

Comparaison expérimentale des deux méthodes

On cherche à comparer les performances de $\hat{f}_n$ et $\hat{g}_n$ sur une même densité f, précisément on souhaite comparer les erreurs (aléatoires)

$$\| \hat{f}_n - f \|_{L^1} \qquad \text{et} \qquad \| \hat{g}_n - f \|_{L^1} .$$

Ces deux quantités sont en général difficiles à calculer, même si f est connue. Nous allons donc les estimer par simulation.

12. Choisir une densité f et simuler un échantillon de n v.a. de densité f (on prendra n de l'ordre de quelques centaines). Tracer sur le même graphique f, une réalisation de $\hat{f}_n$ et de $\hat{g}_n$.

13. Simuler maintenant un grand nombre d'échantillons de n v.a. de densité f, et en déduire une approximation de $\| \hat{f}_n - f \|_{L^1}$ et $\| \hat{g}_n - f \|_{L^1}$.