

(public 2008)

Mots clefs : Loi des grands nombres, espace des polynômes, estimation non-paramétrique

- *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

Lorsque l'on cherche à étudier une suite de mesures provenant de la répétition d'une expérience, une méthode de modélisation consiste à supposer que ces mesures sont des réalisations de variables aléatoires indépendantes équi-distribuées. Comprendre ces mesures et la façon dont elles sont distribuées revient à étudier la loi de probabilité de la variable aléatoire sous-jacente. Par exemple, en médecine, on cherche à étudier l'assimilation d'un traitement antibiotique administré par voie orale. Pour cela, on mesure, pour chaque patient $i = 1, \dots, n$, la concentration x_i de l'antibiotique qui est passée dans le sang du patient après 5 heures (temps moyen de digestion). On modélise le phénomène de la manière suivante : x_1, \dots, x_n sont les réalisations de n variables aléatoires indépendantes X_1, \dots, X_n ayant même densité f . Dans ce contexte médical, comprendre le processus d'assimilation de l'antibiotique dans le sang revient à connaître f .

Lorsque l'on n'a pas d'idée a priori sur la forme particulière que peut prendre la densité f , construire un estimateur de f ne se résume pas à l'estimation d'une moyenne et d'une variance, comme c'est le cas pour des lois gaussiennes. Il s'agit de reconstruire une fonction. Le problème est alors dit non-paramétrique.

1. Estimation non-paramétrique d'une densité

1.1. De la fonction de répartition à la densité

Supposons que nous observons n variables aléatoires indépendantes et identiquement distribués X_1, \dots, X_n de densité de probabilité par rapport à la mesure de Lebesgue une fonction inconnue f de \mathbb{R} dans $[0, +\infty[$. L'objectif de notre étude est la construction d'un estimateur de f , c'est-à-dire une fonction $\hat{f}_n(x) = f_n(x, X_1, \dots, X_n)$ mesurable par rapport à la tribu engendrée par (X_1, \dots, X_n) .

Notons $F(x) = \mathbf{P}(X_1 \leq x)$ la fonction de répartition de la loi de X_1 et considérons la fonction de répartition empirique

$$(1) \quad \hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{X_i \leq x\}}, \forall x \in \mathbb{R}.$$

La loi forte des grands nombres permet d'affirmer que \hat{F}_n est un estimateur de F . Il est même possible d'obtenir des intervalles de confiance et de tester l'adéquation des données à différentes lois. Néanmoins, il n'est pas évident d'utiliser \hat{F}_n pour estimer f . Une des premières idées intuitives est de considérer pour $h > 0$ petit

$$\hat{f}_n(x) = \frac{\hat{F}_n(x+h) - \hat{F}_n(x-h)}{2h} = \frac{1}{2nh} \sum_{i=1}^n \mathbf{1}_{\{-h \leq X_i - x \leq h\}}.$$

Cet estimateur, appelé estimateur de Rosenblatt (1956), est le premier exemple d'estimateur à noyau construit à l'aide du noyau $K(u) = \frac{1}{2} \mathbf{1}_{\{-1 < u \leq 1\}}$, notion que nous allons étudier maintenant.

1.2. Noyaux

Définissons maintenant plus généralement la notion d'estimateur à noyau :

Définition 1. Soit $K : \mathbb{R} \rightarrow \mathbb{R}$ une fonction intégrable telle que $\int K(u) du = 1$. K est appelé noyau. Pour tout $n \in \mathbb{N}^*$, on appelle $h_n > 0$ la fenêtre et \hat{f}_n l'estimateur à noyau de f , défini pour tout $x \in \mathbb{R}$ par

$$\hat{f}_n(x) = \frac{1}{nh_n} \sum_{i=1}^n K\left(\frac{X_i - x}{h_n}\right).$$

Un noyau est dit *positif* si $K \geq 0$: l'estimateur à noyau est alors une densité quelles que soient les valeurs des observations X_1, \dots, X_n . Un noyau est dit *symétrique* si, pour tout u dans son ensemble de définition, $K(u) = K(-u)$.

Exemples de noyaux : Voici quelques exemples de noyaux les plus communément utilisés :

- $K(u) = \frac{1}{2} \mathbf{1}_{\{|u| \leq 1\}}$ (noyau rectangulaire) ;
- $K(u) = \frac{3}{4} (1 - u^2) \mathbf{1}_{\{|u| \leq 1\}}$ (noyau d'Epanechnikov) ;
- $K(u) = \frac{1}{\sqrt{2\pi}} \exp(-u^2/2)$ (noyau Gaussien) ;

Définition 2. Soit $r \geq 1$ un entier. On dit qu'un noyau K est d'ordre r si :

$$\forall j = 1, \dots, r, \quad \int u^j K(u) du = 0 \quad \text{et} \quad \int u^{r+1} K(u) du \neq 0.$$

Existence de noyaux d'ordre donné : il est possible de construire explicitement des noyaux

d'ordre r . Par exemple, considérons les polynômes de Legendre

$$P_0(x) = \frac{1}{\sqrt{2}},$$
$$P_m(x) = \sqrt{\frac{2m+1}{2}} \frac{1}{2^m m!} \frac{d^m}{dx^m} [(x^2 - 1)^m].$$

Les polynômes $(P_m)_{m \geq 0}$ constituent une base orthonormée de $\mathbb{L}^2([-1, 1])$. Dès lors, on peut voir que la fonction K définie par

$$K(u) = \sum_{k=0}^r P_k(0) P_k(u) \mathbf{1}_{\{|u| \leq 1\}}$$

est un noyau d'ordre r .

2. Propriétés des estimateurs à noyaux

2.1. Estimation de la densité

Afin d'évaluer la performance de l'estimateur à noyau défini précédemment, nous calculerons son écart quadratique moyen en un point x_0 donné :

$$EQM(x_0) = \mathbf{E}[\hat{f}_n(x_0) - f(x_0)]^2.$$

La vitesse de décroissance vers 0 de cette quantité, appelée vitesse de convergence, mesurera la qualité de l'estimateur au point x_0 : plus la perte quadratique est petite et plus l'estimateur sera un "bon" estimateur. La vitesse de convergence dépend de la régularité de la densité que l'on cherche à estimer. Ainsi, estimer une fonction régulière est plus facile qu'estimer une fonction qui fluctue beaucoup et rapidement. Nous nous limiterons donc aux densités appartenant à la classe de Hölder définie de la manière suivante :

Définition 3. Soient $[a, b]$ un intervalle de \mathbb{R} , $\alpha > 0$, s la partie entière de α et $L > 0$. La classe de Hölder $H(\alpha, L)([a, b])$ est formée de toutes les fonctions $f : [a, b] \rightarrow \mathbb{R}$ telles que la dérivée $f^{(s)}$ existe (par convention $f^{(0)} = f$) et vérifie

$$|f^{(s)}(x) - f^{(s)}(y)| \leq L|x - y|^{\alpha - s}, \quad \forall (x, y) \in [a, b]^2.$$

On peut décomposer l'erreur quadratique en 2 termes, respectivement le biais $b(\cdot)$ et la variance $\sigma^2(\cdot)$ de l'estimateur au point x_0 :

$$(2) \quad EQM(x_0) = \underbrace{\mathbf{E}[\hat{f}_n(x_0) - f(x_0)]^2}_{b^2(x_0)} + \underbrace{\mathbf{E}[\hat{f}_n(x_0) - \mathbf{E}(\hat{f}_n(x_0))]^2}_{\sigma^2(x_0)},$$

en notant

$$b(x_0) = \mathbf{E}\hat{f}_n(x_0) - f(x_0) \quad \text{et} \quad \sigma^2(x_0) = \mathbf{E}[\hat{f}_n(x_0) - \mathbf{E}(\hat{f}_n(x_0))]^2$$

Proposition 1. Supposons que $f \in H(\alpha, L)([a, b])$ soit bornée, c'est-à-dire qu'il existe $M \in]0, +\infty[$ tel que $f(x) \leq M$ pour tout $x \in \mathbb{R}$. Soit K un noyau d'ordre $s \in \mathbb{N}$ tel que

$$\int K^2(u)du < \infty, \quad \int |u|^\alpha |K(u)|du < \infty.$$

Alors il existe deux constantes C_1 et C_2 telles que

$$(3) \quad \sigma^2(x_0) \leq \frac{C_1}{nh_n}$$

$$(4) \quad |b(x_0)| \leq C_2 h_n^\alpha.$$

Démonstration. La partie variance est facile à étudier, en considérant les variables aléatoires

$$K\left(\frac{X_i - x_0}{h_n}\right) - \mathbf{E}\left[K\left(\frac{X_i - x_0}{h_n}\right)\right], \quad \forall i = 1, \dots, n.$$

De plus, après changement de variable, le biais s'écrit

$$b(x_0) = \int K(u)[f(x_0 + uh_n) - f(x_0)]du.$$

Un développement de Taylor à l'ordre s permet d'obtenir le résultat annoncé. \square

Ainsi, pour le choix optimal de $h_n = O\left(n^{-\frac{1}{2\alpha+1}}\right)$, nous obtenons, quand $n \rightarrow +\infty$ et uniformément en x_0 ,

$$(5) \quad EQM(x_0) = O\left(n^{-\frac{2\alpha}{2\alpha+1}}\right).$$

De façon similaire, on définit l'écart quadratique moyen intégré

$$(6) \quad EQMI = \mathbf{E}\left[\int (\hat{f}_n(x) - f(x))^2 dx\right].$$

Alors, on peut obtenir [mais la preuve est plus technique] l'estimation

$$\begin{aligned} EQMI &= O\left(\left[\frac{1}{s!} \int |u|^\alpha |K(u)|du\right]^2 h_n^{2\alpha} + \frac{\int K^2(u)du}{nh_n}\right) \\ &= O\left(n^{-\frac{2\alpha}{2\alpha+1}}\right) \text{ pour le choix optimal de } h_n. \end{aligned}$$

3. Risque optimal en pratique

3.1. Choix du paramètre de lissage

Nous avons vu que lorsque K est choisi, nous pouvons calculer en fonction du choix de la fenêtre h_n la valeur de l'écart quadratique moyen intégré $EQMI$. Nous écrivons donc dorénavant $EQMI(h_n)$. Le meilleur choix théorique obtenu pour h_n dépend de la régularité de la densité. Or, cette régularité étant représentée par un paramètre inconnu, ce choix théorique n'est donc pas

utilisable en pratique. Cependant, cet inconvénient peut être contourné en utilisant la technique dite de *validation croisée*. Remarquons en effet qu'on a l'égalité

$$\arg \min_{h_n > 0} EQMI(h_n) = \arg \min_{h_n > 0} \mathbf{E} \left[\int \hat{f}_n^2(x) dx - 2 \int f(x) \hat{f}_n(x) dx \right].$$

Il suffit donc d'estimer sans biais les deux quantités

$$G_n^{(1)} = \mathbf{E} \left(\int \hat{f}_n^2(x) dx \right) \text{ et } G_n^{(2)} = \mathbf{E} \left(\int f(x) \hat{f}_n(x) dx \right) = \mathbf{E} \hat{f}_n(X),$$

en désignant par X une variable aléatoire de densité f , indépendante de X_1, \dots, X_n .

On définit

$$(7) \quad \hat{f}_{n,-i}(x) = \frac{1}{(n-1)h_n} \sum_{j \neq i} K \left(\frac{X_j - x}{h_n} \right).$$

On a alors le théorème suivant.

Théorème 1. Soit \hat{f}_n un estimateur à noyau K d'une densité f telle que $\int f^2(x) dx < +\infty$ et

$$\iint \left| K \left(\frac{x-y}{h_n} \right) \right| f(x) f(y) dx dy < \infty.$$

Alors, $\hat{G}_n^{(1)} = \int \hat{f}_n^2(x) dx$ et $\hat{G}_n^{(2)} = \frac{1}{n} \sum_{i=1}^n \hat{f}_{n,-i}(X_i)$ sont des estimateurs sans biais respectivement de $G^{(1)}$ et $G^{(2)}$. Il s'ensuit, en posant

$$VC(h_n) = \int \hat{f}_n^2(x) dx - \frac{2}{n} \sum_{i=1}^n \hat{f}_{n,-i}(X_i),$$

qu'on a, pour tout $h_n > 0$, l'égalité

$$(8) \quad \mathbf{E}[VC(h_n)] = EQMI(h_n) - \int f^2(x) dx.$$

Finalement, on peut calculer

$$h_n^* = \arg \min_{h_n > 0} VC(h_n)$$

et définir l'estimateur optimal, obtenu par validation croisée, comme étant égal à

$$\hat{f}^*(x) = \frac{1}{nh_n^*} \sum_{i=1}^n K \left(\frac{X_i - x}{h_n^*} \right).$$

Le calcul de l'estimateur $G_n^{(1)}$ peut se faire explicitement, en utilisant l'expression analytique de $K(u)$. En outre, on peut démontrer (*il n'est pas demandé pas de le faire ici*) que l'écart quadratique moyen intégré de cet estimateur est proche asymptotiquement de celui de l'estimateur théorique idéal.

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.*

A quoi peut servir concrètement une estimation de la densité ? Quelles sont les limites du modèle proposé ? Que feriez-vous dans un cadre paramétrique ? Quelle autre méthode proposer dans le cadre non paramétrique ?
 - *Développements mathématiques.*

Quelles sont les propriétés de l'estimateur de Rosenblatt ? Quel est l'ordre d'un noyau symétrique ? Complétez les preuves de la proposition 1 et du théorème 1. Que pensez-vous du rôle de la fenêtre ? Expliquez l'idée à la base de la validation croisée. Que pensez-vous du choix des critères EQM et EQMI ?
 - *Etude numérique.*

Construire un estimateur par noyau pour différents échantillons que vous simulerez : avec une loi gaussienne, un mélange de deux loi gaussiennes (c'est-à-dire une densité constituée d'une combinaison convexe de deux densités gaussiennes), une loi non-gaussienne. Le choix de la fenêtre est-il important ? Que pensez-vous du choix du noyau ?