Université de Sidi Bel Abbès.
Faculté des Sciences Exactes
Département de Proba.Stat

Année : 2021/2022
Master : STAT-APP
Module Stat-Nonparam

Estimation nonparamétrique

Problème 1. *Soit X une variable aléatoire définie sur l'espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$, à valeur dans $\mathbb{R}$. pour tout $x \in \mathbb{R}$, on désigne par F la fonction de répartition de X, qu'on suppose qu'elle est $(k+1)$-fois continûment dérivable et par f la fonction de densité, qu'on suppose qu'elle est strictement positive, et de classe C^k au voisinage de x.*

Etant donné $X_1, X_2, \ldots, X_n$ une suite de variable aléatoire réelle de même loi que X, l'estimateur de la fonction de répartition par la méthode du noyau noté $F_n(x)$, défini par :

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right), \forall x \in \mathbb{R}$$

où K est noyau et h_n est une suite de réels positifs, vérifiants

(1) *Le noyau K est supposé d'ordre k intégrable, d'intégrale égale à 1, borné et positif et à support compact $(0,1)$, vérifiant :*
   (i) $\int t^j K(t)dt = 0 \ \forall j = 1, \ldots, k-1$, *et* $0 < |\int t^k K(t)dt| < \infty$.
   (ii) $\exists A < \infty, \ \forall x_1, x_2 \in \mathbb{R}$ *on a :* $|K^{(i)}(x_1) - K^{(i)}(x_2)| \leq A|x_1 - x_2|$, *où* $i = 0, 1$.

(2) $\lim_{n\to+\infty} h_n = 0$ *et* $\lim_{n\to+\infty} n^{\beta} h_n = \infty \ \ \forall \ \beta > 0, \quad j = 0, 1.$

On déduit de F_n un estimateur de la densité, noté f_n, défini par

$$f_n(x) = F_n^{(1)}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_n} K^{(1)}\left(\frac{x - X_i}{h_n}\right) = \frac{1}{nh_n}\sum_{i=1}^{n} K^{(1)}\left(\frac{x - X_i}{h_n}\right)$$

Montrer qu'on a

(1) $|f_n(x) - f(x)| = \mathcal{O}(h_n^k) + \mathcal{O}\left(\sqrt{\frac{\log n}{nh_n}}\right)$, *p.co.* *et* $\exists \delta > 0 : \ \mathbb{P}\left(f_n(x) < \delta\right) < \infty$.

(2) $|F_n(x) - F(x)| = \mathcal{O}(h_n^k) + \mathcal{O}\left(\sqrt{\frac{\log n}{n}}\right)$, *p.co.*

(3) $\exists \delta > 0$ *tel que* $\sum_{n=1}^{\infty} \mathbb{P}\left\{\inf_{x\in\mathbb{R}} |1 - F_n(x)| < \delta\right\} < \infty$.

Problème 2. *Soit* $K : \mathbb{R} \longrightarrow \mathbb{R}$ *une fonction quelconque et soit* h *un réel positif. On appelle estimateur à noyau la fonction*

$$f_n = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$$

où K *est le noyau de cet estimateur et* h *est la fenêtre. Montrer que si* K *est positive et* $\int_{\mathbb{R}} K(u)du = 1$*, alors* $f_n(\cdot)$ *est une densité de probabilité. De plus,* f_n *est continue si* K *est continue. Lorsqu'on définit un estimateur à noyau, on a non-seulement le choix de la fenêtre* $h > 0$ *mais aussi celui du noyau* K*. Il y a un certain nombre de conditions qui sont considérées comme usuelles pour les noyaux et qui permettent d'analyser le risque de l'estimateur à noyau qui en résulte.*

HYPOTHÈSE K *: On suppose que* K *vérifie les* 4 *conditions suivantes :*

(1) $\int_{\mathbb{R}} K(u)du = 1$,

(2) K *est une fonction paire ou, plus généralement,* $\int_{\mathbb{R}} uK(u)du = 0$,

(3) $\int_{\mathbb{R}} u^2|K(u)|du < \infty$,

(4) $\int_{\mathbb{R}} K(u)^2du < \infty$,

(i) Si les trois premières conditions de l'hypothèse K *sont remplies et* f *est une densité bornée dont la dérivée seconde est bornée, montrer que*

$$|Biais\,(f_n(x))| \leq C_1 h^2,$$

où $C_1 = 1/2 \sup_{z\in\mathbb{R}} |f''(z)| \int_{\mathbb{R}} u^2\,|K(u)|\,du.$

(ii) Si, de plus, la condition 4 *de l'hypothèse* K *est satisfaite, montrer que*

$$Var\,(f_n(x)) \leq \frac{C_2}{nh}$$

avec $C_2 = \sup_{z\in\mathbb{R}} f(z) \int_{\mathbb{R}} K(u)^2du.$

Problème 3. *Soit* $X_1, X_2, \ldots X_n$ *un* n*-échantillon de* X *de fonction de répartition* F *et de densité* f *et soit* $x \in \mathbb{R}$ *tel que* $F(x) < 1$*. Admettant que que la fonction de hasard* $\lambda(x) = \dfrac{f(x)}{1 - F(x)}$ *est un paramètre fonctionnel.*

(1) *Déduire un estimateur* $\lambda_n(x)$ *pour* $\lambda(x)$ *par la méthode du noyau.*

(2) *Montrer que*

$$\lambda_n(x) - \lambda(x) = \frac{f_n(x) - f(x)}{1 - f_n(x)} + (F_n(x) - F(x))\,\frac{\lambda(x)}{1 - F_n(x)}$$

(3) *En utilisant la convergence presque complète de $F_n(x)$ vers $F(x)$ montrer qu'il existe $\delta > 0$ tel que :*

$$\sum_n \mathbb{P}\left[(1 - F_n(x)) < \delta\right] < \infty$$

(4) *Etudier la convergence presque complète de l'estimateur $\lambda_n(x)$.*

Problème 4. *Soit $(X_1, Y_1) \dots (X_n, Y_n)$ un n-échantillon de (X, Y) dans $\mathbb{R}^2$. On considère la fonction de répartition conditionnelle de Y, sachant X, définie par*

$$\forall x \in \mathbb{R} \qquad F(x, y) = \mathbb{E}\left(\psi_y(Y)|X = x\right)$$

où $\psi_y(Y) = \mathcal{I}_{Y \leq y}$ avec $\mathcal{I}$ est la fonction indicatrice. On suppose que la densité de la variable explicative X et la fonction $\mathbb{E}\left(\psi_y(Y)|X = x\right)$ vérifient la condition suivante :

$$\exists k > 0,\ \exists C < \infty,\ \forall z \in]x - \varepsilon, x + \varepsilon[,\ |\phi(x) - \phi(z)| \leq C|x - z|^k,$$

où ϕ désigne indifféremment f ou $\mathbb{E}\left(\psi_y(Y)|X = x\right)$.

(1) *Estimer la fonction $F(x, y) = \mathbb{E}\left(\psi_y(Y)|X = x\right)$ par la méthode du noyau.*

(2) *Montrer que, si :*

(a) $\lim_{n \longrightarrow \infty} h_n = 0$ et $\lim_{n \longrightarrow \infty} \frac{nh_n}{\log n} = \infty$,

(b) Le noyau K est borné, intégrable et à support compact,

(c) La fonction f est telle que $f(x) > 0$, alors, l'estimateur construit converge presque complètement et que sa vitesse de convergence est

$$\mathcal{O}(h_n^k) + \mathcal{O}\left(\sqrt{\frac{\log n}{nh_n}}\right) \quad \text{en p.co.}$$

Problème 5. *On va s'intéresser au modèle de régression, où*

$$Y_i = f(x_i) + \varepsilon_i, \quad i = 1, \dots, n$$

où $x_i \in [0, 1]$ sont connus et les ε_i sont i.i.d centrés de même variances σ^2, et on cherche à estimer f, fonction de $[0, 1]$ à valeurs dans $\mathbb{R}$. Supposant à présent $\widehat{f}$ un estimateur linéaire de f tel que :

$$\forall x \in [0, 1],\ \widehat{f}(x) = \sum_{i=1}^{n} W_{n,i}(x) Y_i, \qquad \text{où} \qquad W_{n,i}(x) = \frac{K\left(\frac{x_i - x}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h_n}\right)}$$

.

(i) Soient $Z_i, \dots, Z_n$ des v.a.r telles que $\exists \alpha > 0$ et $C > 0$ tels que pour tout $i = 1, \dots, n$ on a : $\mathbb{E}\left(\exp(\alpha Z_i)\right) \leq C$. Montrer que

$$\mathbb{E}\left(\max_{1 \leq i \leq n} Z_i\right) \leq \frac{1}{\alpha} \ln(Cn).$$

(ii) Soit $x \in [0,1]$, f continue, et qu'il esxiste $(h_n)_{n\geq 1}$ où $h_n \underset{n\to\infty}{\longrightarrow} 0$ telle que les deux conditions suivantes sont vérifiées :

(1) $\lim_{n\to\infty} \sum_{i=1}^{n} W_{n,i}^2(x) = 0.$

(2) *Pour tout $\delta > 0$,* $\lim_{n\to\infty} \sum_{i=1}^{n} \mathbf{1}_{|x-x_i|>\delta} W_{n,i}(x) = 0.$

Montrez que $\lim_{n\to\infty} \mathbb{E}\left[\left(\widehat{f}(x) - f(x)\right)^2\right] = 0.$

(iii) *Supposons f continue, et que les deux conditions suivantes sont vérifiées :*

(3) $\lim_{n\to\infty} \int_0^1 \sum_{i=1}^{n} W_{n,i}^2(x)dx = 0.$

(4) *Pour tout $\delta > 0$,* $\lim_{n\to\infty} \sum_{i=1}^{n} \int_{|x-x_i|>\delta} W_{n,i}(x)dx = 0.$

Vérifie qu'on a : $\lim_{n\to\infty} \mathbb{E}\left[\int_0^1 \left(\widehat{f}(x) - f(x)\right)^2 dx\right] = 0.$

Problème 6. *Soit $(X_1, Y_1) \ldots (X_n, Y_n)$ un n-échantillon de (X, Y) dans $\mathbb{R}^d \times \mathbb{R}$. On considère la moyenne conditionnelle de Y, sachant X, définie par*

$$\forall x \in \mathbb{R}^d \qquad r(x) = \mathbb{E}\left(Y|X = x\right).$$

On suppose que la densité de la variable explicative X et la fonction $r(x)$vérifient la condition suivante :

$$\exists a > 0,\ \exists C < \infty,\ \forall z \in]x-\varepsilon, x+\varepsilon[,\ |\phi(x) - \phi(z)| \leq C\|x - z\|^a,$$

où ϕ désigne indifféremment f ou $\mathbb{E}\left(Y|X = x\right)$.

(1) *Estimer la fonction $r(x) = \mathbb{E}\left(Y|X = x\right)$ par la méthode du noyau.*

Montrer que si

(i) $f > 0$ et r sont k fois continûment différentiables autour de x.

(ii) $|Y| < M < \infty$.

(iii) $\lim_{n\longrightarrow\infty} h_n = 0$ *et* $\lim_{n\longrightarrow\infty} \frac{nh_n^d}{\log n} = \infty.$

(iv) Le noyau K est borné, intégrable, d'ordre k et à support compact.

(2) *L'estimateur construit converge presque complètement et que sa vitesse de convergence est $\mathcal{O}(h_n^k) + \mathcal{O}\left(\sqrt{\frac{\log n}{nh_n^d}}\right)$ en p.co.*

Supposons à présent que

(a) r et f sont continues autour de x et f est strictement positive.

(b) $|Y| < M < \infty$.

(c) $\lim_{n\longrightarrow\infty} h_n = 0$ *et* $\lim_{n\longrightarrow\infty} nh_n^d = \infty$.

(d) K *est borné, intégrable, positive, symétrique et à support compact.*

Montrer que

(3) $\mathbb{E}\left(\widehat{r}(x)\right) \longrightarrow r(x)$ *où* $\widehat{r}(x)$ *est l'estimateur de* $r(x)$ *et* $Var(\widehat{r}(x)) = \mathcal{O}\left(\frac{1}{nh_n^d}\right)$ *p.o.*

(4) *Endéduire que* $\mathbb{E}\left(\widehat{r}(x) - r(x)\right)^2 \longrightarrow 0$

Problème 7. *Soient* X *et* Y *deux variables aléatoires définies sur l'espace de probabilité* $(\Omega, \mathcal{A}, \mathbb{P})$ *à valeurs dans* $\mathbb{R}\times\mathbb{R}$. *Pour tout* $x \in \mathbb{R}$, *on désigne par* F^x *la fonction de répartition conditionnelle de* Y *sachant* $X = x$, *on suppose que* F^x *est absolument continue par rapport à la mesure de Lebesgue de densité* f^x.

Étant donné $(X_1, Y_1), \ldots, (X_n, Y_n)$ *une suite des observations de même loi que* (X, Y), *on estime par la méthode à noyau la fonction de répartition conditionnelle* F^x *par l'estimateur, noté* $\widehat{F}^x$, *défini par :*

$$\widehat{F}^x(y) = \frac{\sum_{i=1}^n K(h_K^{-1}(x - X_i))H(h_H^{-1}(y - Y_i))}{\sum_{i=1}^n K(h_K^{-1}(x - X_i))}, \quad \forall y \in \mathbb{R}$$

où K *est un noyau,* H *est une fonction de répartition et* $h_K = h_{K,n}$ *(resp.* $h_H = h_{H,n}$*) est une suite de réels positifs. On déduit de* $\widehat{F}^x$ *un estimateur de la densité conditionnelle, noté* $\widehat{f}^x$, *défini par*

$$\widehat{f}^x(y) = \frac{h_H^{-1}\sum_{i=1}^n K(h_K^{-1}(x - X_i))H^{(1)}(h_H^{-1}(y - Y_i))}{\sum_{i=1}^n K(h_K^{-1}(x - X_i))}, \quad \forall y \in \mathbb{R}.$$

Pour simplifier la notation, on pose $K_i(x) = K(h_K^{-1}(x - X_i))$, $H_i(y) = H(\frac{y-Y_i}{h_H})$ *et* $f_n(x) = \frac{1}{nh_K}\sum_{i=1}^n K_i(x)$, $g_n^{(j)}(x, y) = \frac{1}{nh_H^j h_K}\sum_{i=1}^n K_i(x)H_i^{(j)}(y), \quad j = 0, 1.$

Supposons que

(a) La f.d.r conditionnelle est $k+1$*-fois continûment dérivable autour de* $\mathcal{S}$.

(b) La densité de la variable explicative est strictement positive, et de classe $\mathcal{C}^k$.

(c) La fonction H *est strictement croissante, de dérivée bornée et d'ordre* k *:*

$$\forall (y_1, y_2) \in \mathbb{R}^2, \quad |H^{(j)}(y_1) - H^{(j)}(y_2)| \leq A|y_1 - y_2|; \qquad j = 0, 1.$$

(d) K *est supposé d'ordre* k *intégrable, d'intégrale égale à* 1, *borné et positif.*

Montrer que

$$\sup_{y\in\mathcal{S}} |F^x(y)f(x) - \mathbb{E}g_n(x, y)| = \mathcal{O}(h_K^k + h_H^k).$$

$$\sup_{y\in\mathcal{S}} \left|f^x(y)f(x) - \mathbb{E}g_n^{(1)}(x, y)\right| = \mathcal{O}(h_K^k + h_H^k).$$

où $\mathcal{S}$ *est un compact de* $\mathbb{R}$