ESTIMATION DE LA DENSITÉ VIA LA MÉTHODE DES POLYNÔMES LOCAUX

par

Ulysse Gazin

Résumé. —
Tout les documents sont autorisées, hormis l'article [**CJM20**].
Les parties de ce devoir sont indépendantes entres elles, hormis la question 6 de la partie 3 qui dépend des autres parties. Par contre, des résultats et des notations des parties précédentes sont nécessaires afin de faire les parties suivantes. Vous pourrez admettre les résultats non démontrées.
Ce devoir est fondé sur l'article [**CJM20**].

Une des méthodes utilisées en régression non-paramétrique est la méthode des polynômes locaux. Une de ses particularités est que pour estimer la fonction de régression en un point on estime en fait les coefficients du développement de Taylor du régresseur ce qui nous donne accès à ses dérivées successives en ce point.

Le but de ce devoir sera donc d'adapter la méthode des polynômes locaux à l'estimation non paramétrique de la densité en utilisant cette particularité. Ce devoir est la transformation de l'article [**CJM20**] en un sujet d'examen. Si la majorités des résultats sont issues de l'article, d'autres en sont des résultats immédiats ou des prémisses nécessaires enfin de formellement le comprendre et de comprendre l'intuition derrière l'utilisation de la méthode des polynômes locaux dans ce cadre.

On se donne $(X_i)_{i\in[\![1;n]\!]}$ un n-échantillon de variables aléatoires à valeurs dans $[x_L;x_U]\subset\mathbb{R}$ avec $-\infty\leq x_L<x_U\leq+\infty$. On suppose que la loi de X_1 a une densité f par rapport à la mesure de Lebesgue, et que f est de classe $\mathcal{C}^p$ sur $[x_L;x_U]$ avec $p\geq 1$. On suppose de plus que pour tout $x\in[x_L;x_U]$ on a $f(x)>0$.

On note F la fonction de répartition de X_1, avec $F(x)=\mathbb{P}(X_1\leq x)$

Si $\mathbb{A}$ est une matrice, on note $\mathbb{A}^\star$ sa transposée.

1. Adaptation de la méthode des polynômes locaux.

On se donne $h>0$ une fenêtre et $K:\mathbb{R}\to\mathbb{R}$ un noyau. On suppose que K est positif, continu, symétrique, à support dans $[-1;1]$ et d'intégrale 1. On note $K_h(u)=\frac{1}{h}K\left(\frac{u}{h}\right)$.

On définit le vecteur colonne $\mathbb{U}(u)=\begin{pmatrix}1\\u\\\vdots\\u^p\end{pmatrix}$.

On définit les matrices diagonales $\mathbb{K}_h(x)\in\mathcal{M}_{n,n}(\mathbb{R})$ et $\mathbb{H}\in\mathcal{M}_{p+1;p+1}$ avec $[\mathbb{K}_h(x)]_{i;i}=K_h\left(X_i-x\right)$ si $i\in[\![1;n]\!]$ et $[\mathbb{H}]_{j;j}=h^{j-1}$ si $j\in[\![1;p+1]\!]$.

On définit la matrice $\mathbb{X}_h(x)=\left[\left(\frac{X_i-x}{h}\right)^{j-1}\right]_{1\leq i\leq n;\ 1\leq j\leq p+1}\in\mathcal{M}_{n,p+1}(\mathbb{R})$

1.1. Question de cours : méthode des polynômes locaux et estimateur empirique. —

On observe $(Y_i;X_i)_{i\in[\![1;n]\!]}$ un n-échantillon dans un modèle de régression avec Y_i admettant un moment d'ordre 1 et $Y_i=g(X_i)+\varepsilon_i$ où $g(x)=\mathbb{E}[Y_1|X_1=x]$ est de classe $\mathcal{C}^{p+1}$.

1. Donner les hypothèses classique sur le bruit ε dans le cadre d'un modèle de régression.

2. On note $\mathbb{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$. On cherche à estimer, par la méthode des polynômes locaux le vecteur $\beta(x) = \begin{pmatrix} \frac{g(x)}{0!} \\ \frac{g^{(1)}(x)}{1!} \\ \vdots \\ \frac{g^{(p)}(x)}{p!} \end{pmatrix}$ par le vecteur $\widehat{\beta}$.

Donner la définition de l'estimateur $\widehat{\beta}$ en terme de minimiseur, puis donner une expression de $\widehat{\beta}$ à l'aide des matrices $\mathbb{Y}$, $\mathbb{H}, \mathbb{K}_h(x)$ et de $\mathbb{X}_h(x)$.

3. On définit $\widetilde{F}$ la fonction de répartition empirique de F définit comme suit pour tout $x \in \mathbb{R}$: $\widetilde{F}(x) = \frac{1}{n} \sum_{i \in [\![1;n]\!]} \mathbb{1}_{X_i \leq x}$.

$\widetilde{F}$ est il un estimateur fortement consistant de F pour la norme infinie ? Quelle est sa vitesse de convergence ?

1.2. Adaptation à l'estimation de la densité. —

4. Quel est la régularité sur $[x_L; x_U]$ de F la fonction de répartition de X_1 en fonction de p ? Quel est la dérivée de F sur $]x_L; x_U[$? Et sur $\mathbb{R}$?

Nous allons appliquer la méthode des polynômes locaux sur la famille $\left(\widetilde{F}(X_i), X_i\right)_{i \in [\![1;n]\!]}$.

On note désormais $\mathbb{Y} = \begin{pmatrix} \widetilde{F}(X_1) \\ \widetilde{F}(X_2) \\ \vdots \\ \widetilde{F}(X_n) \end{pmatrix}$.

5. Calculer $\mathbb{E}\left[\widetilde{F}(X_1)\middle|X_1\right]$. On définit $g : x \in [x_L; x_U] \mapsto \mathbb{E}\left[\widetilde{F}(X_1)\middle|X_1 = x\right] \in \mathbb{R}$. Quel est la régularité de g ? Quel est sa dérivée ? Montrer que $g'(x)$ converge vers $f(x)$.

6. On note $\varepsilon_i = \widetilde{F}(X_i) - \mathbb{E}\left[\widetilde{F}(X_i)\middle|X_i\right]$. On suppose que ε_i admet un moment d'ordre 2 pour tout $i \in [\![1; n]\!]$. Calculer son moment d'ordre 1. En déduire que

pour tout i, ε_i est asymptotiquement centrée. Montrer que si $j \neq i$, on a ε_i et ε_j asymptotiquement décorrélés.

Dans la suite, on estimera $\beta(x) = \begin{pmatrix} \frac{F(x)}{0!} \\ \frac{F^{(1)}(x)}{1!} \\ \vdots \\ \frac{F^{(p)}(x)}{p!} \end{pmatrix}$ par le vecteur $\widehat{\beta}(x)$ obtenu en appliquant la méthode des polynômes locaux à la famille $\left(\widetilde{F}(X_i), X_i\right)_{i\in[\![1;n]\!]}$.

7. En utilisant la question 2., montrer que :

$$\widehat{\beta}(x) - \beta(x) = \mathbb{H}^{-1}\left(\frac{1}{n}\mathbb{X}_h(x)^\star\mathbb{K}_h(x)\mathbb{X}_h(x)\right)^{-1}\left(\frac{1}{n}\mathbb{X}_h(x)^\star\mathbb{K}_h(x)\left[\mathbb{Y} - \mathbb{X}_1(x)\beta(x)\right]\right)$$

2. Première étude de $\widehat{\beta}(x)$

On se fixe un x dans $]x_L; x_U[$. Dans la suite du sujet on omettra l'évaluation en x ce que vous pouvez faire sur votre copie

On rappelle que dans notre asymptotique on a h qui tend vers 0 et n vers $+\infty$.

2.1. Étude du dénominateur. —

On étudie le comportement asymptotique de $\frac{1}{n}\mathbb{X}_h^\star\mathbb{K}_h\mathbb{X}_h$.

1. Soit $(i,j) \in [\![1; p+1]\!]^2$. Calculer $\left[\frac{1}{n}\mathbb{X}_h^\star\mathbb{K}_h\mathbb{X}_h\right]_{i;j}$.

2. En déduire que $\mathbb{E}\left(\left[\frac{1}{n}\mathbb{X}_h^\star\mathbb{K}_h\mathbb{X}_h\right]_{i;j}\right) = \int_{\frac{x_L-x}{h}}^{\frac{x_U-x}{h}} v^{i+j-2}K(v)f(x+vh)\,\mathrm{d}v$ puis que $\mathbb{E}\left(\left[\frac{1}{n}\mathbb{X}_h^\star\mathbb{K}_h\mathbb{X}_h\right]_{i;j}\right) = f(x)\int_{\mathbb{R}} v^{i+j-2}K(v)\,\mathrm{d}v + o(1)$.

3. Montrer que $\mathbb{V}\left(\left[\frac{1}{n}\mathbb{X}_h^\star\mathbb{K}_h\mathbb{X}_h\right]_{i;j}\right) \leq \mathcal{O}_{\mathbb{P}}\left(\frac{1}{nh}\right)$.

4. En déduire à l'aide de l'inégalité de Bienaymé-Tchebychev que $\left[\frac{1}{n}\mathbb{X}_h^\star\mathbb{K}_h\mathbb{X}_h\right]_{i;j} = \mathbb{E}\left(\left[\frac{1}{n}\mathbb{X}_h^\star\mathbb{K}_h\mathbb{X}_h\right]_{i;j}\right) + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{nh}}\right)$

5. On pose $\mathbb{S}_x = \int_{\mathbb{R}} \mathbb{U}(u)\mathbb{U}(u)^\star K(u)\,\mathrm{d}u \in \mathcal{M}_{p+1;p+1}(\mathbb{R})$.

Conclure que :

$$\frac{1}{n}\mathbb{X}_h^\star \mathbb{K}_h \mathbb{X}_h = f(x)\mathbb{S}_x + o(1) + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{nh}}\right)$$

2.2. Décomposition du numérateur. —

On étudie le comportement asymptotique de $\frac{1}{n}\mathbb{X}_h^\star \mathbb{K}_h (\mathbb{Y} - \mathbb{X}_1\beta)$ en le décomposant en 4 termes de type biais/variances.

6. Montrer que :

$$\frac{1}{n}\mathbb{X}_h^\star \mathbb{K}_h (\mathbb{Y} - \mathbb{X}_1\beta) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{U}\left(\frac{X_i - x}{h}\right)\left[\widetilde{F}(X_i) - \mathbb{U}(X_i - x)^\star \beta\right] K_h(X_i - x)$$

On définit :

$$\widehat{L} = \int_{\frac{x_L - x}{h}}^{\frac{x_L - x}{h}} \mathbb{U}(u)\left(\widetilde{F}(x + hu) - F(x + hu)\right) K(u) f(x + hu)\,\mathrm{d}u.$$

7. Exprimer $\frac{1}{n^2}\sum\limits_{\substack{i,j\in[\![1;n]\!]\\ j\neq i}} \mathbb{E}\left[\mathbb{U}\left(\frac{X_i - x}{h}\right)\left[\mathbb{1}_{X_j \leq X_i} - F(X_i)\right] K_h(X_i - x)\middle| X_j\right]$ en fonction de $\widehat{L}$.

On définit les quantités suivantes :

$$\widehat{B_S} = \frac{1}{n}\sum_{i\in[\![1;n]\!]} \mathbb{U}\left(\frac{X_i - x}{h}\right)\left[F(X_i) - \mathbb{U}(X_i - x)^\star \beta\right] K_h(X_i - x)$$

$$\widehat{B_{LI}} = \frac{1}{n^2}\sum_{i\in[\![1;n]\!]} \mathbb{U}\left(\frac{X_i - x}{h}\right)\left[1 - F(X_i)\right] K_h(X_i - x)$$

$$\widehat{R} = \frac{1}{n^2}\sum_{\substack{i,j\in[\![1;n]\!]\\ j\neq i}} \left(\mathbb{U}\left(\frac{X_i - x}{h}\right)\left[\mathbb{1}_{X_j \leq X_i} - F(X_i)\right] K_h(X_i - x) - \mathbb{E}\left[\mathbb{U}\left(\frac{X_i - x}{h}\right)\left[\mathbb{1}_{X_j \leq X_i} - F(X_i)\right] K_h(X_i - x)\middle| X_j\right]\right)$$

8. Décomposer $\frac{1}{n}\mathbb{X}_h^\star \mathbb{K}_h (\mathbb{Y} - \mathbb{X}_1\beta)$ en terme de $\widehat{L}$, $\widehat{B_S}$, $\widehat{B_{LI}}$ et de $\widehat{R}$.

3. Étude des termes dans le numérateur

Ici, nous allons faire tendre n vers $+\infty$, h vers 0 et nh vers $+\infty$.

3.1. De simples développements asymptotiques. —

1. En vous inspirant de la méthode de la section 2.1, montrer que $\widehat{B_{LI}} = \mathcal{O}_{\mathbb{P}}\left(n^{-1}\right)$.

2. On pose $\mathbb{A} = \int_{\mathbb{R}} \mathbb{U}(u) u^{p+1} K(u)^2 \, \mathrm{d}u \in \mathbb{R}^{p+1}$. En utilisant un devellopement de Taylor de F et en utilisant la méthode de la question 2.1, montrer que $\widehat{B_S} = h^{p+1} \frac{F^{(p+1)}(x) f(x)}{(p+1)!} \mathbb{A} + o_{\mathbb{P}}\left(h^{p+1}\right)$

3.2. Conclusion. —

On suppose désormais que $h \to 0$, $n \to +\infty$, $nh \to +\infty$, $nh^2 \to +\infty$ et $nh^{2p+1} \sim 1$.

3. Déduire des résultats des sections 3.1 et 2.1 que :

$$\sqrt{\frac{n}{h}} \left[\frac{1}{n} \mathbb{X}_h^\star \mathbb{K}_h \mathbb{X}_h \right]^{-1} \widehat{B_{LI}} \xrightarrow{\mathbb{P}} 0.$$

4. Déduire des résultats des sections 3.1 et 2.1 que :

$$\sqrt{\frac{n}{h}} \left[\frac{1}{n} \mathbb{X}_h^\star \mathbb{K}_h \mathbb{X}_h \right]^{-1} \widehat{B_S} \xrightarrow{\mathbb{P}} \frac{F^{(p+1)}(x)}{(p+1)!} \mathbb{S}_x^{-1} \mathbb{A}.$$

On suppose de plus que :

$$\sqrt{\frac{n}{h}} \left[f(x) \mathbb{S}_x \right]^{-1} \widehat{L} \xrightarrow{\mathcal{L}} \mathcal{N}_{p+1}\left(0, \Sigma_x\right)$$

et que

$$\sqrt{\frac{n}{h}} \left[\frac{1}{n} \mathbb{X}_h^\star \mathbb{K}_h \mathbb{X}_h \right]^{-1} \widehat{B_S} \xrightarrow{\mathbb{P}} 0.$$

5. Montrer le résultat suivant :

$$\sqrt{\frac{n}{h}} \mathbb{H} \left[\widehat{\beta}(x) - \beta(x) \right] - \frac{F^{(p+1)}(x)}{(p+1)!} \mathbb{S}_x^{-1} \mathbb{A} \xrightarrow{\mathcal{L}} \mathcal{N}_{p+1}\left(0, \Sigma_x\right).$$

3.3. Un test pour conclure. —

On suppose désormais que f est définit sur $\mathbb{R}$ et que f est de classe $\mathcal{C}^1$ (i.e. $p = 1$). Soit $r > 0$. Soit $x_0 \in \mathbb{R}$. On cherche à tester si la densité f atteint en maximum ou un minimum local en x_0 avec $f(x_0) = r$. Pour faire cela, on cherche donc à tester la chose suivante :

(H_0) : $f(x_0) = r, f'(x_0) = 0$ contre l'hypothèse fantôme.

On suppose connue Σ_{x_0} que l'on suppose définie positive.

6. A l'aide du théorème précédent, mettre en place un test asymptotiquement de taille $\alpha \in]0; 1[$. Ce test est il consistant sous H_0 et sous H_1 ?

7. On suppose désormais inconnue Σ_{x_0} ; mais on sait que cette matrice est définie positive et on possède un estimateur $\widetilde{\Sigma_{x_0}}$ de Σ_{x_0} consistant.

Pouvez vous modifier le test obtenu à la question précédente pour obtenir de nouveau un test asymptotiquement consistant sous toutes hypothèses et asymptotiquement de taille α ?

4. Devoir Maison

Implémenter, sur R ou python, la méthode décrite ici. Pouvez vous illustrer la consistance ? Peut on voir l'effet de p sur une loi simple (une loi uniforme par exemple) ?

Références

[CJM20] Matias D. Cattaneo, Michael Jansson et Xinwei Ma : Simple local polynomial density estimators. *Journal of the American Statistical Association*, 115(531):1449–1455, 2020.

15 décembre 2022

U.Gazin, Université Rennes 1