Sur une famille paramétrique d'estimateurs séquentiels de la densité pour un processus fortement mélangeant

Aboubacar Amiri

Sur une famille paramétrique d'estimateurs séquentiels de la densité pour un processus fortement mélangeant

Aboubacar Amiri

Université d'Avignon et des Pays de Vaucluse,
Laboratoire d'Analyse Non Linéaire et Géométrie
(EA 2151), F-84018 Avignon
aboubacar.amiri@univ-avignon.fr

Abstract Let $(X_t,\ \ t \in \mathbb{N})$ be a $\mathbb{R}^d$-valued α-mixing process, where the X_t's have the same unknown density f. We suggest to estimate f, recursively, from the data $X_1, \ldots, X_n$. So, we introduce a subfamily of the general recursive kernel estimators initiated by Deheuvels (1974), including the most popular recursive estimators. For this subfamily, we establish the exact asymptotic square error and then we introduce criteria for comparison that allow us to make a choice among our estimators.

Key words non parametric estimation, recursive kernel estimators, strong-mixing process.

Résumé Soit $(X_t,\ \ t \in \mathbb{N})$ un processus α-mélangeant, où les X_t sont des vecteurs de $\mathbb{R}^d$ de même loi, de densité de probabilité inconnue f. Nous nous proposons d'estimer f de manière récursive à l'aide des observations $X_1, \ldots, X_n$. Pour cela, nous considérons une sous-famille des estimateurs récursifs généraux initiés par Deheuvels (1974), incluant les estimateurs récursifs les plus utilisés. Pour cette sous-famille, nous obtenons l'erreur quadratique asymptotique exacte, ensuite, nous introduisons des critères de comparaison qui nous permettent de classifier et comparer nos estimateurs.

Mots clés estimation non paramétrique , estimateurs récursifs à noyaux, processus mélangeants.

1 Introduction

Soit $(X_t,\ \ t \in \mathbb{N})$ un processus à valeurs dans $\mathbb{R}^d$, α-mélangeant. Les X_t sont supposés équidistribués de densité f inconnue. Parmi les estimateurs les plus utilisés pour estimer f à partir des observations $X_1, \ldots, X_n$, il y a les histogrammes et les polygones de fréquences. L'histogramme mobile est un cas particulier du célèbre estimateur à noyau introduit par Rosenblatt (1956) et Parzen (1962) défini par

$$f_n^{\mathrm{PR}}(x) := \frac{1}{nh_n^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right).$$

L'étude de cet estimateur a donné lieu à une vaste littérature statistique. Nous nous intéressons aux estimateurs récursifs introduits pour la première fois par Wolverton Wagner (1969) et Yamato (1972) sous la forme

$$f_n^{\mathrm{WW}}(x) := \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_i^d}K\left(\frac{x-X_i}{h_i}\right).$$

De nombreuses variantes récursives ont également été proposées et étudiées depuis. En particulier, Deheuvels (1973, 1974) s'est intéressé à la famille

$$f_n^H(x) := \left(\sum_{i=1}^{n} h_i H(h_i)\right)^{-1}\sum_{i=1}^{n} H(h_i)K\left(\frac{x-X_i}{h_i}\right).$$

2 Présentation de l'estimateur

Nous proposons la sous-famille paramétrique d'estimateurs récursifs à noyau définie par:

$$f_n^l(x) := \frac{1}{\sum_{i=1}^n h_i^{d(1-l)}}\sum_{i=1}^{n}\frac{1}{h_i^{dl}}K\left(\frac{x-X_i}{h_i}\right), \quad l\in[0,1]$$

qui correspond pour $d=1$ au cas $H(u)=u^{-l}$. Pour tout $l\in[0,1]$, $(f_n^l(x))$ peut se calculer de manière récursive par

$$f_{n+1}^l(x) = \frac{\sum_{i=1}^n h_i^{d(1-l)}}{\sum_{i=1}^{n+1} h_i^{d(1-l)}}f_n^l(x) + K_{n+1}^l\left(x-X_{n+1}\right) \text{ avec } K_i^l(.) := \frac{1}{h_i^{dl}\sum_{j=1}^{i} h_j^{d(1-l)}}K\left(\frac{\cdot}{h_i}\right).$$

Nous donnons ici, les biais, variance et erreur quadratique asymptotiques exacts de $f_n^l(x)$, en fonction de l, ensuite nous introduisons trois critères de comparaison qui nous permettent de préférer ou non notre sous-famille à l'estimateur à noyau habituel et aussi de classer nos estimateurs en fonction de la valeur de l. Mais pour cela nous avons besoin des hypothèses suivantes.

Hypothèse $\mathcal{K}$:

(i): $K : \mathbb{R}^d \longmapsto \mathbb{R}$ est une densité de probabilité, strictement positive , symétrique et bornée;

(ii): $\lim_{||x||\to\pm\infty} ||x||^d K(x) = 0, \quad \forall x\in\mathbb{R}^d$;

(iii): $\int_{\mathbb{R}^d} |v_i v_j| K(v)dv < \infty, \quad i,\ j = 1,\ldots,d.$

Hypothèse $\mathcal{H}$:

(i): h_n est une suite réelle qui décroît vers 0 et $nh_n^{d+2} \longrightarrow \infty$ lorsque $n \longrightarrow \infty$;

(ii): $B_{n,r} := \frac{1}{n}\sum_{i=1}^{n}\left(\frac{h_i}{h_n}\right)^r \longrightarrow \beta_r < \infty,\ n \to \infty\ \forall\ r \in]-\infty,\ \ d+2]$.

Hypothèse $\mathcal{P}$:

$f \in C_d^2(b)$, où $C_d^2(b)$ désigne l'ensemble des fonctions $\psi : \mathbb{R}^d \mapsto \mathbb{R}$ telles que $\psi^{(2)}$ existe pour toute dérivée partielle d'ordre 2, continue et bornée.

Hypothèse $\mathcal{Q}$:

(i): Le processus (X_t) est $2-\alpha$-mélangeant avec: $\alpha^{(2)}(k) \leq \gamma k^{-\rho}, \quad k \geq 1$ pour deux constantes strictement positives γ et ρ.

(ii): Pour chaque couple $(s,t), \quad s \neq t$, le vecteur aléatoire (X_s, X_t) admet une densité $f_{(X_s,X_t)}$ telle que $\sup_{|s-t|\geq 1} \parallel g_{s,t} \parallel_\infty < \infty$ où $g_{s,t} := f_{(X_s,X_t)} - f \otimes f$.

Les hypothèses $\mathcal{P}$ et $\mathcal{Q}$ sont classiques dans ce domaine: en particulier, il n'y a pas d'hypothèse de stationnarité de second ordre statuée sur le processus. On les rencontre par exemple dans Bosq-Blanke (2007). L'hypothèse $\mathcal{H}(ii)$ est très utile dans nos calculs, et est propre à la récursivité.

3 Résultats

Nous pouvons maintenant déterminer les biais, variance et erreur quadratique asymptotiques de notre famille d'estimateurs.

Théorème 3.1 *Sous les hypothèse $\mathcal{K}$, $\mathcal{H}$, $\mathcal{P}$, et $\mathcal{Q}$:*

(a): $h_n^{-4}\left(Ef_n^l(x) - f(x)\right)^2 \to_{n\to\infty} \left(\frac{\beta_{d(1-l)+2}}{\beta_{d(1-l)}}\right)^2 b_2^2(x)$,
avec $b_2^2(x) := \frac{1}{4}\left(\sum_{1\leq i,j\leq d}\frac{\partial^2 f}{\partial x_i \partial x_j}(x)\int_{\mathbb{R}^d} v_i v_j K(v)dv\right)^2$.

(b): *Pour tout* $l \in \left[\left(\frac{d-2}{2d}\right)^+,\ 1\right]$, $nh_n^d Var f_n^l(x) \to \frac{\beta_{d(1-2l)}}{\beta_{d(1-l)}^2} f(x)\int_{\mathbb{R}^d} K^2(u)du,\ n \to \infty,$ *si* $\rho > 2$, *où:* $x^+ = \max(x, 0)$.

(c): *Si* $d \geq 3$ *et* $l \in \left[0, \frac{d-2}{2d}\right[$ *, la conclusion du (b) reste encore vraie si* $\rho > \frac{d+2}{2}$.

(d): *Sous les conditions du (b) (avec $\rho > 2$) ou du (c)(avec $\rho > \frac{d+2}{2}$), le choix $h_n = C_n n^{-\frac{1}{d+4}}$, $C_n \to c > 0$, entraîne que*

$$n^{\frac{4}{d+4}} E\left(f_n^l(x) - f(x)\right)^2 \longrightarrow c^4 \left(\frac{4+dl}{2+dl}\right)^2 b_2^2(x) + \frac{(4+dl)^2 f(x)\|K\|_2^2}{2c^d(4+d)(2+dl)},\ n \to \infty,$$

pour les valeurs respectives de l en tout point où $f(x) > 0$.

Notons que si l'on précise la forme de h_n, le résultat (c) du théorème 3.1 se réécrit sous la forme:

(c'): *Si $d \geq 3$ et $l \in \left[\left(1 - \frac{1}{2\nu d}\right)^+, \frac{d-2}{2d}\right[$, le choix $h_n = C_n n^{-\nu}$, $C_n \to c > 0$, avec $0 < \nu < \frac{1}{d+2}$, entraîne:*

$$nh_n^d Var f_n^l(x) \to \frac{(1-\nu d(1-l))^2}{1-\nu d(1-2l)} f(x) \int_{\mathbb{R}^d} K^2(u)du,\ n \to \infty,\ si\ \rho > 2.$$

4 Comparaison d'estimateurs

Définition 4.1 *(**Critères de comparaison**): Soient $f_n(x)$ et $g_n(x)$ deux estimateurs à noyau de f.*

(i): *On dira que $f_n(x)$ est préférable à $g_n(x)$ au sens de la variance si:*

$$0 \leq \lim_{n\to\infty} \frac{Var\left(f_n(x)\right)}{Var\left(g_n(x)\right)} < 1.$$

(ii): *On dira que $f_n(x)$ est préférable à $g_n(x)$ au sens du biais si:*

$$0 \leq \lim_{n\to\infty} \frac{\left(Ef_n(x) - f(x)\right)^2}{\left(Eg_n(x) - f(x)\right)^2} < 1.$$

(iii): *On suppose que $f(x) > 0$, on choisit: $h_n = C_n n^{-\frac{1}{d+4}}$, $C_n \to c > 0$, avec $c = c_{min}(f_n(x))$ (resp. $c = c_{min}(g_n(x))$) pour l'estimateur$f_n(x)$(resp.$g_n(x)$).*
$c_{min}(\diamond)$ désigne la constante qui minimise l'erreur quadratique asymptotique de l'estimateur $\diamond$.
Sous ces conditions, on dira que $f_n(x)$ est préférable à $g_n(x)$ au sens de la moyenne quadratique (ou du MSE) si : $0 \leq \lim_{n\to\infty} \frac{E(f_n(x)-f(x))^2}{E(g_n(x)-f(x))^2} < 1$.

Le critère (i) a été introduit par Bannon (1976). Notre premier résultat de cette partie permet de classifier nos estimateurs selon les valeurs de l par les critères précédents.

Théorème 4.2 *On suppose que les hypothèses $\mathcal{K}-\mathcal{Q}$ sont vérifiées avec $\rho > \max\left(2, \frac{d+2}{2}\right)$. On choisit $h_n = C_n n^{-\nu},\ C_n \to c > 0\ , 0 < \nu < \frac{1}{d+2}$. Alors:*

(a): *l'efficacité de $(f_n^l(x))$ est décroissante (resp. croissante) selon le critère de la variance (resp. du biais)*

(b): *si $f(x) > 0$, et $\nu = \frac{1}{d+4}$, l'efficacité de $(f_n^l(x))$ est croissante selon le critère du MSE.*

Notre dernier résultat compare notre famille d'estimateurs à l'estimateur à noyau usuel $f_n^{\mathrm{PR}}(x)$.

Théorème 4.3 *On se place sous les hypothèses du théorème 4.2. Alors :*

(a): *tous les estimateurs $f_n^l(x),\ l \in [0,1]$ sont préférable à $f_n^{PR}(x)$ au sens de la variance.*

(b): *aucun estimateur $f_n^l(x),\ l \in [0,1]$ n'est préférable à $f_n^{PR}(x)$ au sens du biais.*

(c): *pour $d = 1$, Si $f(x) > 0$, et $\nu = \frac{1}{d+4}$, $f_n^{PR}(x)$ est préférable à tous les estimateurs $f_n^l(x),\ l \in [0,1]$ au sens du M.S.E, pour les choix "optimaux" respectifs de c.*

References

[1] G. **Banon.** Sur un estimateur non paramétrique de la densité de probabilité . Revue de statistique appliquée, tome 24, no. 4 (1976), p. 61- 73

[2] D. **Bosq.** Nonparametric statistics for Stochastic Processes lecture. Notes in statistics (1998)

[3] D. **Bosq**, D. **Blanke**. Inference and Prediction in large dimensions. Wiley Series in Probability and Statistics 2007 ISBN 978-0-470-08147-1

[4] Paul **Deheuvels**. Conditions Nécessaires et Suffisantes de Convergence Ponctuelle Presque Sûre et Uniforme Presque Sûre des Estimateurs de la Densité. Comptes rendues de l'Academie des Sciences de Paris vol. 278, 1973 p. 1217-1220

[5] Paul **Deheuvels**. Sur l'Estimation séquentielle de la densité. Compte Rendues de l'Academie des Sciences de Paris Serie A, 276 :1119-1121, 1973.

[6] Paul **Deheuvels**. Sur une famille d'estimateurs de la densité d'une variable aléatoire. Comptes Rendues de l'Academie des Sciences de Paris Serie A, 276 :1013-1015, 1974.

[7] Elias **Masry**. Recursive Probability Density Estimation for Weakly Dependent Stationary Processes. IEEE, 1986.

[8] E. **Parzen**. On the estimation of a probability density function and the mode. Ann. Math. Stat. 33, pp. 1065-1076

[9] F. **Rozenblatt**. Remarks on some nonparametric estimates of a density function. Ann. Math. Stat. 38, pp. 482-493.

[10] Edward J. **Wegman** and H.I. **Davis**. Remarks on Some Recursive Estimators of a Probability Density. The Annals of Statistics 1979, Vol 7, No. 2, 316-327.

[11] C. T. **Wolverton**, Terry J. **Wagner**. Recursive Estimates of Probability Densities. IEEE Transactions on Systemes Sciences and Cybernetics Vol 5, 1969 p. 307

[12] H. **Yamato**. Sequential estimation of a continuous probability density function and mode. Bull. Math. Statist. Jap., 1972, Vol. 14, p 1-12