République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

جامعة بجاية
Tasdawit n Bgayet
Université de Béjaïa

Université A. Mira de Béjaia
Faculté des Sciences Exactes
Département de Recherche Opérationnelle

Mémoire de fin de cycle
en vue de l'obtention du diplôme de master
en Mathématiques appliquées

Spécialité : Science des données et aide à la décision

Thème :

Estimateurs récursifs par noyaux : théorie et applications

Présenté par :

BENACHOUR Sonia
BOUTEGRABET Lydia

Soutenu devant le jury composé de :

DJABRI Rabah	Président	Université de Béjaia
AMROUN Sonia	Examinatrice	Université de Béjaia
DJERROUD Lamia	Examinatrice	Université de Béjaia
ZOUGAB Nabil	Encadrant	Université de Béjaia

Année Universitaire : 2023 – 2024

Remerciments

Nous débutons nos remerciements en exprimant notre gratitude envers Dieu le Tout-Puissant pour nous avoir accordé la santé, la volonté, la force et le courage nécessaires pour surmonter les obstacles et mener à bien ce projet de recherche académique.

Nous adressons nos remerciements les plus chaleureux à Monsieur Zougab Nabil, pour son encadrement expert, ses conseils éclairés et son soutien constant tout au long de cette aventure intellectuelle. Votre dévouement et votre expertise ont grandement contribué à l'enrichissement de notre mémoire.

Nous souhaitons également remercier chaleureusement Monsieur Khemici Mohamed pour son aide précieuse, ses conseils avisés et son soutien tout au long de ce projet. Votre contribution a été déterminante pour la réussite de notre mémoire.

Nous adressons également nos remerciements respectueux à l'ensemble des membres du jury pour leur évaluation rigoureuse, leurs commentaires constructifs et leurs recommandations éclairées. Votre expertise a été cruciale pour la qualité de notre mémoire.

Nos remerciements s'étendent également à nos familles, nos amis et toutes les personnes qui nous ont soutenues et encouragées durant cette période exigeante. Votre soutien moral et votre compréhension ont été des piliers essentiels de notre réussite.

Nous exprimons notre profonde gratitude envers tous les enseignants et les professeurs du département de Recherche Opérationnelle de la Faculté des Sciences exacte de Université A-Mira de Béjaia . Enfin, nous adressons nos remerciements à toutes les personnes qui ont participé de près ou de loin à la réalisation de ce travail. Votre contribution a été précieuse et nous vous en sommes sincèrement reconnaissants.

Dédicaces

À nos parents,

Pour leur amour inconditionnel, leur soutien constant et leurs sacrifices innombrables. Sans leur encouragement et leur foi en nous, ce travail n'aurait jamais été possible.

À nos professeurs et encadrants,

Pour leur guidance précieuse, leurs conseils avisés et leur dévouement à notre éducation. Leur passion pour l'apprentissage et la recherche a été une source d'inspiration constante.

À nos amis et camarades de la promotion 2023/2024,

Pour leur amitié, leur aide précieuse et leur soutien moral. Leur présence à nos côtés a rendu ce parcours académique plus enrichissant et agréable.

Nous tenons à leur dédier ce modeste travail.

Table des matières

Table des figures

Liste des tableaux

Liste d'abriviations et notations

i.i.d	Indépendantes et identiquement distribuées
MSE	L'erreur quadratique moyenne
MISE	L'erreur quadratique moyenne intégrée
AMISE	L'erreur quadratique moyenne intégrée asymptotique
ISE	L'erreur quadratique intégré
CV	Cross-Validation (Validation croisée)
UCV	Unbiased Cross-Validation (Validation croisée non biaisée)
IG	Gaussien inverse
RIG	Gaussien inverse réciproque
$\mathbb{1}_A(.)$	Indicatrice sur A
$\widehat{f_n}(.)$	Estimateur non récursif de la fonction de densité
$\widetilde{f_n}(.)$	Estimateur récursif de la fonction de densité
$X_1, X_2, \ldots, X_n$	n-échantillion
$\mathbf{E}(.)$	Espérance
$\mathbf{V}(.)$	Variance
$\mathcal{GS}(.)$	Un ensemble de suites réelles positives dans $\mathbb{R}$

Introduction générale

La théorie de l'estimation est un domaine majeur d'étude en statistique. Parmi les différentes approches pour estimer la densité de probabilité, on distingue principalement l'approche paramétrique et l'approche non paramétrique.

L'approche paramétrique repose sur l'hypothèse que les données suivent une distribution de probabilité spécifique, dont seuls les paramètres sont inconnus. Son objectif est de déterminer la valeur réelle des paramètres ou une fonction de ces paramètres. Cependant, cette approche nécessite une connaissance préalable de la forme de la distribution sous-jacente, ce qui peut être contraignant.

En revanche, l'approche non paramétrique n'impose pas de structure spécifique à la distribution des données. Elle vise plutôt à estimer la densité de probabilité directement à partir des données, sans faire d'hypothèses sur la forme de la distribution. Bien que moins contraignante en termes de connaissances préalables, l'approche non paramétrique peut nécessiter davantage de données pour produire des estimations précises. Différentes méthodes d'estimation non paramétrique de la fonction densité sont disponibles. On retrouve l'estimation par histogramme, les séries orthogonales (basées sur le développement en séries de Fourier), l'interpolation par les fonctions Splines, et la méthode du noyau. Parmi ces méthodes, la plus répandue et la plus efficace est la méthode du noyau, qui a été introduite par Rosenblatt [2] et Parzen [3]. L'estimateur à noyau a été largement adopté par les utilisateurs en raison de sa formulation théorique simple, de sa convergence dans différents contextes, et de sa flexibilité en ce qui concerne le paramètre de lissage.

Ce projet de fin d'études se focalise de manière spécifique sur l'analyse des estimateurs récursifs par la méthode du noyau. Cette approche, qui se base sur des principes non paramétriques, est employée pour estimer la fonction de densité de probabilité. Sa nature récursive revêt une importance cruciale dans le contexte de données évolutives, où des mises à jour périodiques des estimations sont nécessaires pour refléter les évolutions temporelles des données. L'estimation récursive de la densité de probabilité par la méthode du noyau est largement répandue dans divers domaines comme l'économie, la finance, l'écologie, ... etc. Plusieurs auteurs ont également étudié de manière approfondie les estimateurs récursifs par noyau, pour le cas des données indépendantes dans la littérature statistique. L'estimateur récursif de densité a été initialement introduit par Wolverton et Wagner [4] et a depuis été largement étudié, notamment par Yamato [5], Davies [7], Devroye [11], Wegman et Davies [10], ainsi que Roussas [18]. Plus tard, Mokkadem et ses collègues [29] ont proposé une variété d'estimateurs symétriques à noyau récursif en utilisant une approche d'approximation stochastique. Les travaux de Slaoui [2019] et ceux d'Amiri et Jemai fournissent également des informations complémentaires sur le sujet. Une contribution récente, attribuée à Kakizawa [46], porte sur l'estimateur récursif de correction du biais de frontière utilisant des noyaux asymétriques, spécifiquement conçu pour des données indépendantes non négatives.

L'objectif premier de ce mémoire est de définir et d'analyser en profondeur les estimateurs récursifs par noyaux. Nous examinerons ensuite de près les propriétés statistiques et asymptotiques de ces estimateurs afin de comprendre leur comportement et leur performance dans divers contextes. Enfin, nous illustrerons l'efficacité de ces estimateurs en mettant en œuvre des études de simulations et des applications sur des données réelles issues de différents domaines.

Dans le contexte actuel de la statistique et de l'analyse de données, plusieurs défis majeurs se posent. Il est nécessaire de développer des techniques statistiques robustes pour la collecte de données, d'élaborer des méthodes statistiques efficaces permettant de tirer des inférences à partir de ces données, et d'améliorer les modèles statistiques en vue de leur utilisation future basée sur ces expériences.

Ce projet de recherche vise à relever ces défis en se concentrant sur l'estimation de la fonction de densité de probabilité à l'aide de techniques récursives basées sur les noyaux. La récursivité de ces estimateurs est particulièrement avantageuse dans le cas de flux de données continus et évolutifs, car elle permet une mise à jour efficace des estimations sans nécessiter une refonte complète du processus d'estimation.

La recherche se distingue par l'étude des propriétés statistiques et asymptotiques des estimateurs récursifs par noyaux, ainsi que leur application sur des données réelles. Ces travaux contribueront à l'avancement des connaissances en estimation de densité de probabilité et auront des implications pratiques importantes dans divers domaines d'application. Le mémoire est structuré comme suit :

- Une introduction générale pour situer notre étude.
- Le premier chapitre sera consacré à un rappel sur les estimateurs non-récursifs par noyaux dans l'estimation de la densité de probabilité.
- Le deuxième chapitre sera dédié à la présentation des estimateurs récursifs par noyaux de la densité de probabilité. Les propriétés statistiques et asymptotiques des estimateurs seront examinées. Le choix optimal des paramètres des estimateurs présentés sera étudié.
- Le dernier chapitre sera cansacré aux applications. Afin de valider la performance des estimateurs récursifs, nous nous intéressons à l'analyse des flux de données continues qui évoluent dans le temps (data stream).
- Nous terminerons par une conclusion générale.

1 Estimateurs non récursifs par noyaux

1.1 Introduction

L'estimateur à noyau non récursif est une méthode d'estimation de la densité de probabilité d'une variable aléatoire. Il s'agit d'une approche non paramétrique, ce qui signifie qu'elle ne repose pas sur des hypothèses spécifiques concernant la distribution des données et permet d'estimer la densité en tout point du support. Il existe deux types de noyau continu, les noyaux symétriques et les noyaux asymétriques.

Dans ce chapitre, nous abordons l'étude de l'estimateur de la fonction de densité en utilisant la méthode du noyau dans les cas symétriques et asymétriques. Nous commençons par définir les estimateurs à noyaux symétriques ainsi que les estimateurs à noyaux asymétriques associés. Ensuite, nous présentons les différentes propriétés des estimateurs à noyaux associés, qu'ils soient symétriques ou asymétriques. Nous donnons également quelques exemples de noyaux symétriques et asymétriques, puis nous discutons du paramètre de lissage h.

1.2 Estimateur de Rosanblat-Parzen

En 1956, Rosenblatt [2] a proposé le premier estimateur à noyau pour la densité de probabilité $f(x)$. Cet estimateur est de type histogramme obtenu à l'aide de noyau uniform : $K(\mu) = \frac{1}{2}\mathbb{1}\{-1 < \mu \leq 1\}$ puis Parzen [3] a proposé une généralisation de l'idée de Rosenblatt [2], cette derniére consiste à remplacer la fonction de densité K par une fonction de densité continue pour que l'estimateur à noyau soit continu aussi.

1.3 Définition de l'estimateur à noyau

Soit $X_1, X_2, \ldots, X_n$ un échantillon de variables aléatoires de fonction de densité inconnue f. L'estimateur non paramétrique de la fonction densité f est l'estimateur à noyau classique

introduit par Rosenblatt [2] et Parzen [3] est défini comme suit :

$$\boxed{\widehat{f_n}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big),} \tag{1.1}$$

où $h > 0$ est le paramètre de lissage qui tend vers 0 lorsque n tend vers l'infini, il détermine le voisinage de x et K est une fonction de densité de probabilité appelé noyau elle détermine la forme autour de x. Il existe deux types de noyau continu

— Noyau continu symétrique
— Noyau continu asymétrique

1.4 Noyau continu symétrique

Un noyau continu est dit symétrique si

$$K(x) = K(-x), \quad \int_{\mathbb{R}} K(x)\,dx = 1, \quad \int_{\mathbb{R}} xK(x)\,dx = 0, \quad \text{et} \quad \int_{\mathbb{R}} x^2K(x)\,dx < +\infty.$$

1.4.1 Quelque noyaux usuelles symétriques

Dans le tableau suivant 1.1 nous allons présenter quelques noyaux usuelles symétriques.

Noyau	Densité	Support
Epanechnikov	$\frac{3}{4}(1 - x^2)$	$[-1, 1]$
Uniforme	$\frac{1}{2}$	$[-1, 1]$
Triangulaire	$\lvert 1 - x\rvert$	$[-1, 1]$
Gaussien	$\frac{1}{\sqrt{2\pi}} \exp(\frac{-x^2}{2})$	$\mathbb{R}$
Biweight	$\frac{15}{16}(1 - x^2)^2$	$[-1, 1]$

TABLE 1.1 – Exemple de quelques noyaux continus symétriques

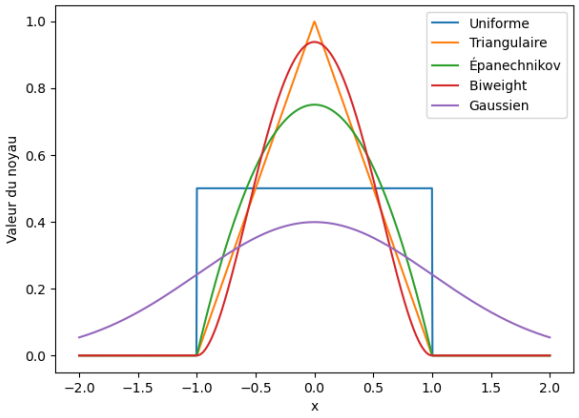

FIGURE 1.1 – Exemples de quelques noyaux continus symétriques

1.4.2 Espérence, biais et variance de l'estimateur

Dans cette partie, nous présentons les propriétés fondamentales de l'estimateur et les critères d'erreurs usuels.

Proposition 1.4.1 (Parzen [3]). *Soit x fixé dans $\mathbb{R}$. L'espérance mathématique de l'estimateur à noyau est donné par :*

$$\boxed{\mathbf{E}[\widehat{f_n}(x)] = f(x) + \frac{h^2}{2} f''(x)\,\sigma_K^2 + o(h^2),} \tag{1.2}$$

avec $\sigma_K^2 = \int_{\mathbb{R}} u^2 K(u)\,du$.

Proposition 1.4.2 (Parzen [3]). *L'estimateur (1.1) est un estimateur biaisé* $\mathbf{E}[\widehat{f_n}(x)] \neq f(x)$ *mais asymptotiquement sans biais, quand* $h \to 0$. *Le biais de l'estimateur à noyau est :*

$$\boxed{\text{Biais}[\widehat{f_n}(x)] = \frac{h^2}{2} f''(x)\,\sigma_K^2 + o(h^2).} \tag{1.3}$$

Proposition 1.4.3 (Parzen [3]). *Pour un x fixé dans $\mathbb{R}$. La variance de l'estimateur à noyau (1.1) est donné par :*

$$\boxed{\mathbf{V}[\widehat{f_n}(x)] = \frac{f(x)}{nh} R(K) + o(\frac{1}{nh}),} \tag{1.4}$$

avec $R(K) = \int_{\mathbb{R}} K^2(u)\,du$.

1.4.3 MSE et MISE de l'estimateur

L'évaluation des performances d'un estimateur à noyau implique la définition d'un critère d'erreur approprié pour mesurer l'erreur d'estimation à un point unique ainsi que sur l'ensemble des points. Pour cela, nous commencerons par examiner la proximité entre notre estimateur $\widehat{f_n}(x)$ et la véritable densité f. Cet estimateur $\widehat{f_n}(x)$ est influencé par le choix du noyau K et du paramètre de lissage h.

MSE de l'estimateur

L'erreur quadratique moyenne (MSE) est définit par :

$$\boxed{\mathrm{MSE}[\widehat{f_n}(x)] = \mathbf{V}[\widehat{f_n}(x)] + Biais^2[\widehat{f_n}(x)],}$$

Cela nous conduit à obtenir :

$$\boxed{\mathrm{MSE}[\widehat{f_n}(x)] = \frac{f(x)}{nh}R(K) + \frac{h^4}{4}\{f''(x)\}^2\sigma_K^4 + o\Big(h^4 + \frac{1}{nh}\Big).} \tag{1.5}$$

MISE de l'estimateur

L'erreur quadratique moyenne intégré (MISE) est définit par :

$$\boxed{\mathrm{MISE}[\widehat{f_n}(x)] = \int_{\mathbb{R}} \mathrm{MSE}[\widehat{f_n}(x)],}$$

Cela nous permet d'obtenir :

$$\boxed{\mathrm{MISE}[\widehat{f_n}(x)] = \frac{R(K)}{nh} + \frac{h^4}{4}\sigma_K^4 R(f''(x)) + o\Big(h^4 + \frac{1}{nh}\Big),} \tag{1.6}$$

avec $R(f''(x)) = \int_{\mathbb{R}}\{f''(x)\}^2\,dx$.

1.5 Noyau continu asymétrique

Si le support de f est $[0, +\infty[$ ou $[0,1]$, les noyaux symétriques donnent une estimation de plus en plus biaisée lorsque x est proche de 0. D'où le problème du biais aux bords ce qui réduit la qualité de l'estimateur. Dans la section suivante nous présentons l'estimateur à noyau asymétrique.

Définition 1.5.1. *Étant donné $x \in \mathbb{T}_1$ et $h > 0$, on appelle un noyau continu asymétrique $K_{x,h}$ toute fonction densité liée à la variable aléatoire $\mathcal{K}_{x,h}$ de support $\mathbb{S}_{x,h} \subseteq \mathbb{R}$ (pouvant ne pas dépendre de x et (ou) h) contenant au moins x, vérifiant les quatres conditions suivantes :*

$$x \in \mathbb{S}_{x,h},$$

$$\mathbf{E}[\mathcal{K}_{x,h}] = x + a(x,h),$$

$$\mathbf{V}[\mathcal{K}_{x,h}] < \infty,$$

$$\mathbf{V}[\mathcal{K}_{x,h}] = b(x,h),$$

où les quantités $a(x,h)$ et $b(x,h)$ sont en fonction de h et x.

Définition 1.5.2. *Soit $X_1, X_2, \ldots, X_n$ un n-échantillon (i.i.d), issu d'une variable aléatoire X de la fonction de densité inconnue f sur l'ensemble $\mathbb{T}_1 \subseteq \mathbb{R}$ borné ou borné d'un seul côté. L'estimateur à noyau continu asymétrique est donné par :*

$$\boxed{\widehat{f_n}(x) = \frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i),} \tag{1.7}$$

où h est le paramètre de lissage et $K_{x,h}$ est le noyau continu asymétrique. Le passage de (1.7) à (1.1) n'est pas possible comme dans le cas symétrique où l'on a $K_{x,h}(*) = (1/h)K\{(x-*)/h\}$.

1.5.1 Exemple des noyaux asymétriques

Dans la littérature une multitude de noyau asymétrique liés aux densité de probabilité : Beta, Gamma, gaussien inverse ,gaussien inverse réciproque ont été introduit, en effet les noyaux : Beta et Gamma ont été proposé par Chen [24] [25] et les noyaux : gaussien inverse et gaussien inverse réciproque par Scaillet [27] mais ils n'ont jamais montré dans leur études comment ils ont construit ces noyaux.

Noyaux	Support	Densité
Bêta	$[0,1]$	$K(y) = \frac{y^{x/h}(1-y)^{(1-x)/h}}{\beta(1+x/h,1+(1-x)/h)}$
Gamma	$[0,\infty[$	$K(y) = \frac{1}{\Gamma(\frac{x}{h}+1)}\frac{y^{\frac{x}{h}}\exp(-y/h)}{h^{\frac{x}{h}+1}}$
Gaussien inverse réciproque	$[0,\infty[$	$K(y) = \frac{1}{\sqrt{2\pi hy}}\exp\left(-\frac{x-h}{2h}\left[\frac{y}{x-h} - 2 + \frac{x-h}{y}\right]\right)$
Lognormal	$[0,\infty[$	$K(y) = \frac{1}{\sqrt{2\pi hy}}\exp\left(-\frac{1}{2h}\left[\log(y) - \log(x) + \sigma^2\right]^2\right)$
Gaussien inverse	$[0,\infty[$	$K(y) = \frac{1}{\sqrt{2\pi hy^3}}\exp\left(-\frac{1}{2hx}\left[\frac{y}{x} - 2 + \frac{x}{y}\right]\right)$
Birnbaum-Saunders	$[0,\infty[$	$K(y) = \frac{1}{2\sqrt{h}}\left(\sqrt{\frac{1}{xy}} + \sqrt{\frac{x}{y^3}}\right)\sqrt{2\pi}\exp\left(\frac{-1}{2\alpha^2}\left(\frac{y}{x} + \frac{x}{y}\right) - 2\right)$
Gamma généralisé	$[0,\infty[$	$k(y) = \frac{\gamma y^{\alpha-1}\exp\left(-\{\frac{y}{\beta\Gamma(\alpha/\gamma)/\Gamma(\alpha+1)/\gamma}\}^{\gamma}\right)}{\{\beta\Gamma(\alpha/\gamma)/\Gamma(\alpha+1)/\gamma\}^{\alpha}\Gamma(\alpha/\gamma)}$
Gamma inverse	$[0,\infty[$	$K(y) = \frac{(\frac{x(h+x)}{h})^{1+\frac{x}{h}}}{\Gamma(1+\frac{x}{h})}(\frac{1}{y})^{2+\frac{x}{h}}\exp(\frac{-x(h+x)}{hy})$

Table 1.2 – Quelques noyaux continus asymétriques

La Table (1.2) présente quelques noyaux continus asymétriques. On peut se référer aux travaux de Libengué [31], Hirukawa and Sakudo [35], Mousa et al [38] et Harfouche [41] pour plus de détails sur ces noyaux ainsi que leurs applications.

1.5.2 Propriétés de l'estimateur à noyau asymétrique

La convergence en moyenne quadratique et en moyenne quadratique intégrée

Nous donnons ici la convergence au sens de l'erreur quadratique moyenne intégrée. Ces résultats sont donnés dans Senga Kiessé [28]. Nous rappelons d'abord le développement du

biais et de la variance de l'estimateur a noyau asymétrique $\widehat{f_n}$ (1.7), par la suite nous déduisons les expressions de l'erreur quadratique moyenne intégrée.

Le Biais de l'estimateur $\widehat{f_n}$

Le Biais de $\widehat{f_n}$ est donné par :

$$\begin{aligned}\text{Biais}[\widehat{f_n}(x)] &= \mathbf{E}[[\widehat{f_n}(x)] - f(x) \\ &= \int_{\mathbb{T}_1} f(t)K_{x,h}(t)\,dt - f(x) \\ &= \mathbf{E}[f(\mathcal{K}_{x,h})] - f(x).\end{aligned}$$

En utilisant un développement limité de Taylor de $f(\mathcal{K}_{x,h})$ au point moyen $\mathbf{E}[\mathcal{K}_{x,h}] = \mu_{x,h}$ on obtient :

$$f(\mathcal{K}_{x,h}) = f(\mu_{x,h}) + (\mathcal{K}_{x,h} - \mu_{x,h})f'(\mu_{x,h}) + \frac{1}{2}(\mathcal{K}_{x,h} - \mu_{x,h})^2 f''(\mu_{x,h}).$$

En calculant l'espérance $\mathbf{E}[f(\mathcal{K}_{x,h})]$, on obtient

$$\begin{aligned}\text{Biais}[\widehat{f_n}(x)] &= f(\mu_{x,h}) + \frac{1}{2}\mathbf{E}[(\mathcal{K}_{x,h} - \mu_{x,h})^2]f''(x) - f(x) \\ &= f(\mathbf{E}[\mathcal{K}_{x,h}]) + \frac{1}{2}\mathbf{V}[\mathcal{K}_{x,h}]f''(x) - f(x).\end{aligned}$$

A travers la premiere définition , le biais tend vers 0 quand $h = h(n) \to 0$ et $n \to \infty$.

La Variance de l'estimateur $\widehat{f_n}$

La Variance de $\widehat{f_n}$ est donné par :

$$\begin{aligned}\mathbf{V}[\widehat{f_n}(x)] &= \frac{1}{n}\int_{\mathbb{T}_1} K_{x,h}^2(t)f(t)\,dt - \frac{1}{n}\Big(\int_{\mathbb{T}_1} K_{x,h}(t)f(t)\,dt\Big)^2 \\ &= \frac{1}{n}\int_{\mathbb{T}_1} K_{x,h}^2(t)f(t)\,dt - \frac{1}{n}\Big(Biais[\widehat{f_n}(x)] + f(x)\Big)^2.\end{aligned}$$

La variance de $\widehat{f_n}(x)$ tend vers 0 quand $\frac{1}{n}\int_{\mathbb{T}_1} K_{x,h}^2(t)f(t)\,dt$ tend vers 0.

L'expression du biais et de la variance de l'estimateur (1.7) avec les différents noyaux définis dans le tableau (1.2) prend respectivement la forme suivante :

$$\boxed{\text{Biais}[\widehat{f_n}(x)] = q(x,f)\,h + o(h),} \tag{1.8}$$

$$\boxed{\mathbf{V}[\widehat{f_n}(x)] = p(x,h)\frac{f(x)}{n} + o\Big(\frac{1}{n}\Big),} \tag{1.9}$$

où la forme explicite de $q(x,f)$ et $p(x,h)$ pour chaque noyau est donnée par le tableau (1.3)

Noyaux	$q(x, f)$	$p(x, h)$
Bêta	$(1-2x)f'(x) + \frac{1}{2}(1-x)f''(x)$	$\frac{1}{2\sqrt{\pi x(x-1)h}}$
Gamma	$f'(x) + \frac{1}{2}f''(x)$	$\frac{1}{2\sqrt{\pi xh}}$
Gaussien inverse réciproque	$\frac{1}{2}xf''(x)$	$\frac{1}{2\sqrt{\pi xh}}$
Lognormal	$2xf'(x) + \frac{1}{2}x^2f''(x)$	$\frac{x^{-1}}{4\sqrt{\pi h}}$
Gaussien inverse	$\frac{1}{2}x^3f''(x)$	$\frac{x^{-3/2}}{2\sqrt{\pi h}}$
Birnbaum-Saunders	$\frac{1}{2}xf'(x) + 2x^2f''(x)$	$\frac{x^{-1}}{\sqrt{2\pi h}}$
Gamma généralisé	$\frac{C}{2}xf''(x)$	$\frac{1}{\sqrt{xh}}V(2)$
Gamma inverse	$\frac{1}{2}xf''(x)$	$\frac{1}{2\sqrt{2\pi xh}}$

TABLE 1.3 – La forme explicite de $q(x, f)$ et $p(x, h)$.

où $0 < |C| < \infty$.

L'expression finale de l'erreur quadratique moyenne intégrée est donnée par :

$$\begin{aligned}\mathrm{MISE}[\widehat{f_n}(x)] &= \int_{\mathbb{T}_1} \mathrm{MSE}[\widehat{f_n}(x)]\,dx \\ &= \int_{\mathbb{T}_1} \left[f(\mathbf{E}[\mathcal{K}_{x,h}]) + \frac{1}{2}\mathbf{V}[\mathcal{K}_{x,h}]f''(x) - f(x)\right]^2 dx \\ &\qquad + \frac{1}{n}\int_{\mathbb{T}_1}\left(\int_{\mathbb{T}_1} K^2_{x,h}(t)f(t)\,dt - \frac{1}{n}\left[\mathrm{Biais}[\widehat{f_n}(x)] + f(x)\right]^2\right)dx.\end{aligned}$$

1.5.3 Choix du noyau asymétrique

En pratique, le choix du noyau dépend du support de la distribution des données que l'on cherche à estimer. Cependant, ce critère seul n'est pas très informatif, sinon il ne serait pas nécessaire de proposer d'autres types de noyaux après ceux initialement proposés par Chen [24], notamment pour les distributions dont le support est $[0, +\infty[$. En réalité, le type de données peut grandement influencer le choix du noyau, en tenant compte des queues, des pôles et d'autres caractéristiques spécifiques des distributions de ces données.

1.6 Choix du paramètre de lissage

L'estimation de la fonction de densité de probabilité par la méthode des noyaux dépend largement du paramètre de lissage h, est un élément clé dans cette méthode.

Ce paramètre agit comme une fenêtre qui définit le niveau de lissage de l'estimation d'une fonction de densité.

Un faible h entraîne un lissage minimal et produit une fonction de densité irrégulière. En revanche, une valeur élevée de h conduit à une estimation plus lisse.
Ainsi, le choix de h influence directement la qualité et la régularité de l'estimation obtenue par la méthode des noyaux.

Il est important de garder à l'esprit que le choix du paramètre de lissage dépend de l'objectif de l'estimation de la densité. Plusieurs méthodes ont été développées dans la littérature pour choisir ce paramètre, et des études comparatives ont été menées pour évaluer ces méthodes.

Deux études notables sont celles de Berlinet et Devroye [21] ainsi que celle de Cao et al. Elles comparent différentes méthodes de sélection du paramètre de lissage pour diverses distributions. Toutes ces méthodes visent à fournir un paramètre de lissage optimal adapté à la distribution à estimer, chacune se distinguant par le critère d'optimisation utilisé. Ce chapitre abordera différentes méthodes permettant de calculer ce paramètre.

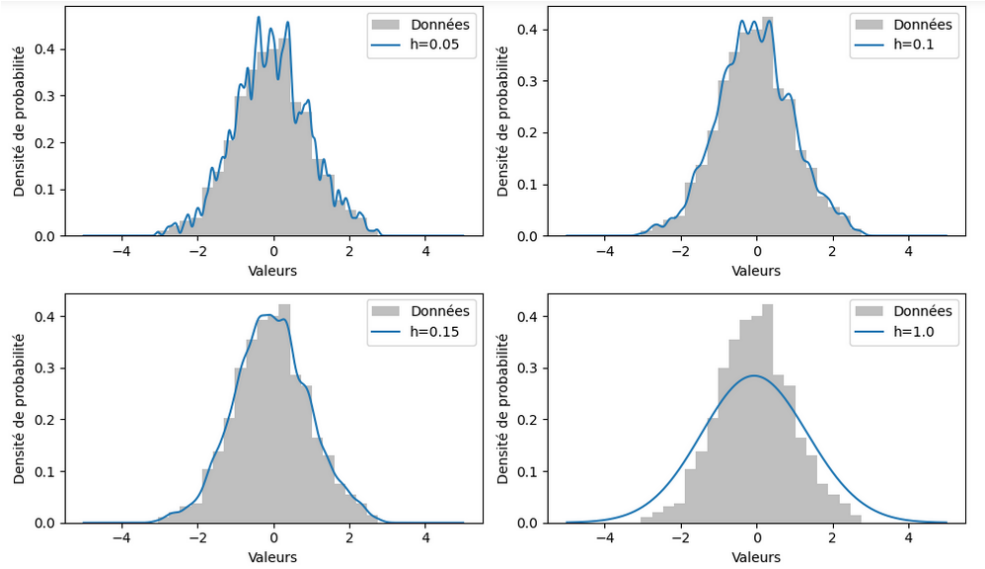

FIGURE 1.2 – Influence du paramètre de lissage h sur la qualité de l'estimation

1.6.1 Méthodes classiques

Il existe de nombreuses méthodes proposées dans cette catégorie (pour plus de détails, voir la monographie de Wand et Jones [20], Silverman [14] et Scott [19]). Voici quelques unes de ces approches.

Méthodes plug-in : cas de noyau symétrique

Pour calculer le h optimale, on dérive l'AMISE (MISE Asymptotique) par rapport à h.
L'erreur quadratique moyenne intégrée asymptotique (AMISE) est :

$$\boxed{\text{AMISE}[h] = \frac{h^4\sigma_K^4}{4} R(f''(x)) + \frac{R(K)}{nh}.} \tag{1.10}$$

avec $R(f''(x)) = \int_{\mathbb{R}} \{f''(x)\}^2\, dx, \quad R(K) = \int_{\mathbb{R}} K^2(u)\, du.$

Le paramètre de lissage optimal qui minimise l'erreur quadratique moyenne intégrée asymptotique (AMISE) est de la forme suivante

$$\boxed{h^* = \left[\frac{R(K)}{\sigma_K^4 R(f''(x))}\right]^{\frac{1}{5}} n^{-\frac{1}{5}}} \tag{1.11}$$

Le paramètre de lissage optimal h^* dépend de la densité inconnue f à travers la fonction $R(f''(x))$. Étant donné que ce paramètre optimal n'est pas calculable directement, une approche courante consiste à remplacer $R(f''(x))$ par un estimateur adapté. Diverses solutions ont été proposées dans la littérature. Par exemple, la méthode de Rule of Thumb choisit f comme une distribution normale avec une moyenne de 0 et une variance de σ_f^2.

Le paramètre σ_f^2 est estimé par la variance empirique $\sigma_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ où $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ Lorsque le noyau K est gaussien, le paramètre de lissage (1.11) a la forme suivante.

$$\boxed{h_{rot} = 1.06\,\sigma_n\, n^{-\frac{1}{5}}.} \tag{1.12}$$

Méthodes plug-in : cas de noyau asymétrique

Dans le cas d'utilisation des noyaux asymétriques, Scaillet [27] et Hirukawa [43] propose d'estimer les quantités inconnues en remplaçant la fonction de densité f par un modèle de référence log-normal en utilisant les noyaux *IG* et *RIG*, voir aussi Hirukawa [43] .

1.6.2 Méthodes de validation croisée

En plus des méthodes de sélection classique, d'autres approches se sont avérées efficaces. Les méthodes de validation croisée (cross validation) ont une idée de base différente : elles cherchent à trouver une fonction de score CV(h) qui possède la même structure que le MISE mais dont le calcul est plus simple. On cherche ensuite la fenêtre h qui minimise ce critère, en anticipant un comportement asymptotique similaire à h^*. Contrairement à la sélection plug-in, la fenêtre h n'est pas déterministe ici ; elle est influencée par les données observées.

Parmi les trois approches principales de la validation croisée - non biaisée, biaisée et lissée, l'approche non biaisée par validation croisée UCV est la plus utilisée.

La validation croisée non biaisée

La validation croisée non biaisée pour la sélection du paramètre de lissage dans l'estimation à noyau de densite a été initialement introduite par Rudemo [12] et Bowman [13]. La méthode consiste à choisir le paramètre de lissage qui minimise un estimateur optimal

$$\boxed{\mathrm{ISE}(h) = \int_{\mathbb{R}} (f(x) - \widehat{f_n}(x))^2\, dx = \int_{\mathbb{R}} f^2(x)\, dx + \int_{\mathbb{R}} \widehat{f_n^2}(x)\, dx - 2\int_{\mathbb{R}} f(x)\widehat{f_n}(x)\, dx.} \tag{1.13}$$

Le paramètre h est choisi de sorte à ce qu'il minimise un estimateur de

$$\mathrm{UCV}(h) = \int_{\mathbb{R}} \widehat{f_n^2}(x)\,dx - 2\int_{\mathbb{R}} f(x)\widehat{f_n}(x)\,dx. \tag{1.14}$$

Le critère de la validation croisée est donné par :

$$\widehat{\mathrm{UCV}} = \int_0^\infty \left(\frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i^2)\right)^2 dx - \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{i\neq j} K_{X_i,h}(X_j). \tag{1.15}$$

Cette approche comporte deux problèmes majeurs : premièrement, elle manque de stabilité lorsqu'elle est confrontée à des variations de taille de l'échantillon, ce qui peut entraîner une grande variabilité dans les résultats de la simulation selon l'échantillon considéré. Deuxièmement, la fonction à minimiser présente fréquemment plusieurs minimums locaux, comme cela a été discuté dans les travaux de Zougab [33].

1.7 Conclusion

Dans ce chapitre, nous avons présenté la méthode du noyau non récursif pour l'estimation non paramètrique de la densité de probabilité. Les deux types de noyau les noyaux symétriques et asymétriques. Nous avons donné la forme des noyaux cités dans la littérature, ainsi que les propriétés des estimateurs basés sur ces noyaux. Le choix du noyau K et du paramètre de lissage h influence directement l'estimateur à noyau. Lorsqu'on estime des densités asymétriques, le choix du noyau prend une importance cruciale, ce qui diffère du cas des densités symétriques.Il reste des difficultés à pallier, notamment en ce qui concerne l'estimation au bord.

Le chapitre suivant explore en profondeur le fonctionnement et les avantages de l'estimateur récursif par noyau. Nous examinerons comment cet estimateur utilise la récursivité pour améliorer progressivement ses estimations de densité, en s'adaptant aux caractéristiques locales des données.

2

Estimateurs récursifs par noyaux

2.1 Introduction

Les estimateurs récursifs offrent un avantage significatif par rapport à leurs équivalents non récursifs, car la mise à jour d'un échantillon de taille n à un échantillon de taille $(n+1)$ nécessite beaucoup moins de calculs. Cette caractéristique est cruciale dans le contexte de l'estimation de densité, où la fonction doit souvent être estimée à de nombreux points. La première version récursive de l'estimateur de densité à noyau de Rosenblatt [2], particulièrement renommée, a été développée par Wolverton et Wagner [4] et a suscité de nombreuses études, notamment celles menées par Yamato [5], Davies [7], Devroye [11], Wegman et Davies [10], ainsi que Roussas [25]. Plus tard, Mokkadem et al [29] ont proposé une large gamme d'estimateurs symétriques à noyau récursif en utilisant l'approche d'approximation stochastique. On peut aussi consulter les travaux de Slaoui [44], Amiri [30] et Jemai [40] pour plus d'informations sur le sujet. Un travail récent attribué à Kakizawa [46] concerne l'estimateur récursif de correction du biais de frontière utilisant des noyaux asymétriques, conçu spécifiquement pour des données indépendantes non négatives.

Ce chapitre vise principalement à présenter ces différents estimateurs récursifs à noyau, tant symétriques qu'asymétriques, dans le contexte de données indépendantes, en mettant en évidence leurs propriétés statistiques.

2.2 Estimateurs récursifs symétrique

Les techniques d'approximation stochastique sont utilisées pour trouver le zéro θ^* d'une fonction inconnue $\rho : \mathbb{R} \to \mathbb{R}$, qui peut être difficile à calculer directement. L'algorithme le plus célèbre pour cela est celui de Robbins et Monro [1], qui fonctionne comme suit :

1. On choisit $\theta_0 \in \mathbb{R}$;

2. Pour $n \geq 1$, on construit la suite θ_n par la relation récursive suivante :

$$\boxed{\theta_n = \theta_{n-1} + \gamma_n W_n,} \tag{2.1}$$

où W_n est une observation de la fonction ρ au point θ_{n-1} et γ_n est une suite de réels positifs qui tend vers zéro appelée "pas" de l'algorithme.

Soit $X_1, X_2, \ldots, X_n$ un échantillon (i.i.d) à valeur dans $\mathbb{R}$ de densité de probabilité f. Pour construire un estimateur récursif de f en un point x par la méthode des algorithmes stochastiques, Mokkadem et al. [29] ont défini un algorithme de recherche du zéro de la fonction ρ : $y \to f(x) - y$

1. Soit $f_0(x) \in \mathbb{R}$ fixé ;
2. Pour $n \geq 1$, on a :

$$\boxed{\widetilde{f_n}(x) = f_{n-1}(x) + \gamma_n W_n(x)} \tag{2.2}$$

Où $W_n(x)$ est une observation de la fonction ρ au point $f_{n-1}(x)$. On peut estimer $f(x)$ par $Z_n(x) = h_n^{-1} K\left(\frac{x-X_n}{h_n}\right)$, ce qui permet de poser $W_n(x) = Z_n(x) - f_{n-1}(x)$.

D'aprés l'algorithme d'approximation stochastique introduit par Mokkadem et al. [29] Pour estimer récursivement la densité f au point x, la formule (2.2) peut s'écrire de la manière suivante :

$$\boxed{\widetilde{f_n}(x) = (1 - \gamma_n) f_{n-1}(x) + \gamma_n h_n^{-1} K\left(\frac{x - X_n}{h_n}\right).} \tag{2.3}$$

Supposons que W_n est une suite positive avec $\sum_{i=1}^{n} W_i = \infty$. Si nous sélectionnons le pas de taille (γ_n) comme étant égal à $W_n(\sum_{i=1}^{n} W_i)^{-1}$, alors l'estimateur (2.3) peut être reformulé de la manière suivante :

$$\boxed{\widetilde{f_n}(x) = \frac{1}{\sum_{i=1}^{n} W_i} \sum_{i=1}^{n} W_i h_i^{-1} K\left(\frac{x - X_i}{h_i}\right).} \tag{2.4}$$

La classe d'estimateurs définie par l'algorithme d'approximation stochastique (2.3) englobe ainsi la catégorie générale des estimateurs récursifs présentés sous la forme (2.4) et introduits dans l'étude de Hall et Patil [22].

Plus précisément, si l'on choisit $W_n = 1$, on obtient l'estimateur proposé par Wolverton et Wagner [4]

$$\boxed{\widetilde{f_n}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_i} K\left(\frac{x - X_i}{h_i}\right).} \tag{2.5}$$

Tandis que le choix de $W_n = h_n^{\frac{1}{2}}$, conduit à l'estimateur considéré par Wegman et Davies [10]

$$\widetilde{f_n}(x) = \frac{1}{n\sqrt{h_n}} \sum_{i=1}^{n} h_i^{\frac{1}{2}} K\Big(\frac{x - X_i}{h_i}\Big). \tag{2.6}$$

Enfin le choix de $W_n = h_n$ aboutit à l'estimateur examiné par Deheuvels [8] et Duflo [23] donné par :

$$\widetilde{f_n}(x) = \frac{1}{\sum_{i=1}^{n} h_n} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{h_i}\Big). \tag{2.7}$$

Pour donner les caractéristiques de l'estimateur $\widetilde{f_n}$ donné par l'equation (2.4), on introduit une classe de suites à variations régulières que nous utiliserons dans nos hypothèses tout au long des chapitres suivants.

La dernière versions de l'estimateur récursive à noyau symétrique a été introduite par Amiri [30] est donnée par : :

$$\widetilde{f_n^l}(x) = \frac{1}{\sum_{i=1}^{n} h_i^{(1-l)}} \sum_{i=1}^{n} \frac{1}{h_i^l} K\Big(\frac{x - X_i}{h_i}\Big), \forall x \in \mathbb{R}\, (l \in [0,1]) \tag{2.8}$$

2.2.1 Résultats principaux

Définition 2.2.1. *Soit $\gamma_n \in \mathbb{R}$ et $(v_n)_{n\geq 1}$ une suite réelle positive. On dit que $v_n \in \mathcal{GS}(\gamma)$ si*

$$\lim_{n\to\infty} n[1 - \frac{v_{n-1}}{v_n}] = \gamma \tag{2.9}$$

La condition (2.9) a été introduite par Galambos et Seneta [9] pour définir les suites à variations régulières.

Des exemples classiques de suites dans $\mathcal{GS}(\gamma)$ sont, $\forall b \in \mathbb{R}$, n^γ, $n^\gamma[\log(n)]^b$, $n^\gamma[\log\log(n)]^b$, et ainsi de suite.

D'abord, nous présentons le biais et la variance de l'estimateur $\widetilde{f_n}$, suivis de l'erreur quadratique intégrée (MISE) ainsi que le choix optimal du pas et de la fenêtre qui la minimise.

Biais et variance de l'estimateur $\widetilde{f_n}$

On se plaçe sous les hypothèses suivantes :

(H1) : $K : \mathbb{R} \to \mathbb{R}$ est la fonction noyau continue, bornée et vérifiant les propriétés de la définition (2.2.1)

(H2) :

i) $(\gamma_n) \in \mathcal{GS}(-\alpha)$ avec $\alpha \in]\frac{1}{2}, 1]$.

ii) $(h_n) \in \mathcal{GS}(-a)$ avec $a \in]0, \frac{\alpha}{2}]$.

iii) $\lim_{n\to\infty}(n\gamma_n) \in]\min\{2a, \frac{1-a}{2}\}, \infty]$.

(H3) : f est bornée, deux fois différentiables et f'' est bornée. L'hypothèse **(H2)** implique que la limite de $(n\gamma_n)^{-1}$ est finie. On note alors $\xi = \lim_{n\to\infty}(n\gamma_n)^{-1}$

Proposition 2.2.1 (Mokkadem et al [29]). *Supposons que les hypothèses **(H1) – (H3)** soient vérifiées et que f'' soit continue au point x, alors.*

1. *Si $0 < a \leq \frac{\alpha}{5}$, alors*

$$\boxed{\mathbf{E}[\widetilde{f_n}(x)] - f(x) = \frac{1}{2(1-2a\xi)} h_n^2 \sigma_K^2 f''(x) + o(h^2),} \tag{2.10}$$

Si $a > \frac{\alpha}{5}$, alors

$$\boxed{\mathbf{E}[\widetilde{f_n}(x)] - f(x) = o(\sqrt{\gamma_n h_n^{-1}}).}$$

2. *Si $a \geq \frac{\alpha}{5}$, alors*

$$\boxed{\mathbf{V}[\widetilde{f_n}(x)] = \frac{1}{2-(1-a)\xi} \frac{\gamma_n}{h_n} f(x) \int_{\mathbb{R}} K^2(u)\, du + o(\frac{\gamma_n}{h_n}),} \tag{2.11}$$

Si $0 < a \leq \frac{\alpha}{5}$, alors

$$\boxed{\mathbf{V}[\widetilde{f_n}(x)] = o(h_n^4).}$$

3. *Si $\lim_{n\to\infty}(n\gamma_n) > \max\{2a, \frac{(1-a)}{2}\}$ alors les deux énoncés (2.10) et (2.11) sont simultanément validés.*

2.2.2 MSE et MISE de l'estimateur

Nous pésentons les choix de (γ_n) et (h_n), qui minimisent le MSE et MISE de l'estimateur récursif défini par l'algorithme d'approximation stochastique (2.3).

Corollaire 2.2.1 (Mokkadem et al [29]). *Supposons que les hypothèses **(H1)**-**(H3)** soient vérifiées et que f'' soit continue sur $\mathbb{R}$. Pour minimiser MSE de $\widetilde{f_n}$ au point x, le pas (γ_n) doit être choisi dans $\mathcal{GS}(-1)$ tel que $\lim_{n\to\infty}(n\gamma_n) = 1$.*

$$\boxed{h_n = \left(\left[\frac{3f(x)\int_{\mathbb{R}} K^2(u)\, du}{10\sigma_K^4 \int_{\mathbb{R}} (f''(x))^2\, dx}\right]^{\frac{1}{5}}\right)\gamma_n^{\frac{1}{5}},} \tag{2.12}$$

d'Où le MSE de $\widetilde{f_n}$ est :

$$\boxed{\mathrm{MSE}[\widetilde{f_n}(x)] = n^{-\frac{4}{5}} \frac{5^{\frac{11}{5}}}{4^{\frac{7}{5}} 3^{\frac{6}{5}}} \left[\sigma_K^2 \int_{\mathbb{R}} (f''(x))^2\, dx\right]^{\frac{2}{5}} \left[f(x) \int_{\mathbb{R}} K^2(u)\, du\right]^{\frac{4}{5}} + o\left(n^{-\frac{4}{5}}\right).} \tag{2.13}$$

Le choix le plus simple de pas appartenant à $\mathcal{GS}(-1)$ tel que $\lim_{n\to\infty}(n\gamma_n) = 1$ est $(\gamma_n) = (n^{-1})$. Pour ce choix de pas, l'estimateur $\widetilde{f_n}$ défini par (2.3) est égal à l'estimateur à noyau récursif introduit par Wolverton et Wagner [4] (2.5) . Cet estimateur le plus récent appartient donc à la sous-classe des estimateurs à noyau récursifs, qui, grâce à un choix adéquat de la largeur de bande, ont une MSE minimale.

MISE de l'estimateur

Dans la proposition suivante, on donne la MISE de $\widetilde{f_n}$.

Proposition 2.2.2 (Mokkadem et al [29])**.** *Supposons que les hypothèses **(H1)-(H3)** soient vérifiées, et que f'' soit continue sur $\mathbb{R}$.*

1. *Si $0 < a < \frac{\alpha}{5}$, alors*

$$\boxed{\mathrm{MISE}[\widetilde{f_n}(x)] = \frac{1}{2(1-2a\xi)} h_n^4 \sigma_K^4 \int_{\mathbb{R}} f''(x)\,dx + o(h_n^4).} \tag{2.14}$$

2. *Si $a = \frac{\alpha}{5}$, alors*

$$\boxed{\mathrm{MISE}[\widetilde{f_n}(x)] = \frac{1}{2(1-2a\xi)} h_n^4 \sigma_K^4 \int_{\mathbb{R}} f''(x)\,dx + \frac{1}{2-(1-a)\xi} \frac{\gamma_n}{h_n} \int_{\mathbb{R}} f(x)\,dx \int_{\mathbb{R}} K^2(u)\,du + o\Big(h_n^4 + \frac{\gamma_n}{h_n}\Big).} \tag{2.15}$$

3. *Si $a > \frac{\alpha}{5}$, alors*

$$\boxed{\mathrm{MISE}[\widetilde{f_n}(x)] = \frac{1}{2-(1-a)\xi} \frac{\gamma_n}{h_n} \int_{\mathbb{R}} f(x)\,dx \int_{\mathbb{R}} K^2(u)\,du + o\Big(\frac{\gamma_n}{h_n}\Big).} \tag{2.16}$$

De cette proposition, on déduit le choix optimal du pas qui permet de minimiser le MISE de l'estimateur.

Corollaire 2.2.2 (Mokkadem et al [29])**.** *Supposons que les hypothèses **(H1)-(H3)** soient vérifiées et que f'' soit continue sur $\mathbb{R}$.*

Pour minimiser MSE *de $\widetilde{f_n}$ au point x, le pas (γ_n) doit être choisi dans $\mathcal{GS}(-1)$ tel que $\lim_{n\to\infty}(n\gamma_n) = 1$. D'après la formule* (2.12)

$$\boxed{\mathrm{MISE}[\widetilde{f_n}(x)] = \frac{5^{\frac{11}{5}}}{4^{\frac{7}{5}} 3^{\frac{6}{5}}} \Big[\sigma_K^2 \int_{\mathbb{R}} (f''(x))^2\,dx\Big]^{\frac{2}{5}} \Big[\int_{\mathbb{R}} f(x)\,dx \int_{\mathbb{R}} K^2(u)\,du\Big]^{\frac{4}{5}} n^{-\frac{4}{5}} + o\Big(n^{-\frac{4}{5}}\Big).} \tag{2.17}$$

L'estimateur à noyau récursif présente une MISE plus élevée que l'estimateur à noyau non récursif de Rosenblatt [2]. Par conséquent, il est préférable d'utiliser l'estimateur non récursif.

Comme dans le cas non récursif, le choix de la fenêtre h_n doit se faire à l'aide de méthodes pratiques. À cet égard, le lecteur peut se référer au travail de Slaoui et al [34], qui a utilisé la méthode de plug-in pour estimer le paramètre de lissage h_n.

Vitesse de Convergence de l'estimateur $\widetilde{f_n}$

On donne dans le Théorème suivant la vitesse de convergence faible de l'estimateur $\widetilde{f_n}$

Théorème 2.2.1 (Mokkadem et al [29]). *Supposons que les hypothèses **(H1)-(H3)** soient vérifiées. Pour $x \in \mathbb{R}$ tel que $f(x) > 0$ et f'' soit continue au pointx, on a*

1. *S'il existe $c \geq 0$ tel que $\gamma_n^{-1} h_n^5 \underset{n\to\infty}{\longrightarrow} c$, alors*

$$\boxed{\sqrt{\gamma_n^{-1}}[\widetilde{f_n}(x) - f(x)] \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\Big(\frac{c^{1/2}}{2(1-2a\xi)}\sigma_K^2 f''(x), \frac{1}{2-(1-a)\xi} f(x)\int_{\mathbb{R}} K^2(u)\,du\Big).} \tag{2.18}$$

2. *Si $\gamma_n^{-1} h_n^5 \underset{n\to\infty}{\longrightarrow} \infty$, alors*

$$\boxed{\frac{1}{h_n^2}[\widetilde{f_n}(x) - f(x)] \xrightarrow[n\to\infty]{\mathbb{P}} \frac{1}{2(1-2a\xi)}\sigma_K^2 f''(x).} \tag{2.19}$$

2.3 Estimateur récursif à noyau asymétrique

Considérons une suite de variables aléatoires indépendants et identiquement distribués $X_1, X_2 \ldots, X_n$, où les valeurs sont dans l'intervalle $[0, \infty[$ de densité commune $f(x)$ inconnue. Pour estimer cette densité de manière récursive pour chaque valeur de x, Kakizawa [46] a proposé d'adapter l'estimateur récursif (2.3) pour le cas asymétrique :

$$\boxed{\widetilde{f_n}(x) = (1-\gamma_n) f_{n-1} + \gamma_n K_{x,h_n}(X_n),} \quad \textit{pour} \quad n = 1, 2, \ldots \tag{2.20}$$

où h_n est le paramètre de lissage qui tendant vers zéro lorsque $n \to \infty$, γ_n est le pas et $K_{x,h}$ est le noyau continu asymétrique.

2.3.1 Propriété de l'estimateur récursif à noyau asymétrique

Les théorèmes suivant, établis par Kakizawa [46], s'appliquent dans le cas univarié pour l'estimateur récursif à noyau asymétrique. Pour simplifier, nous posons

$$B(x) = \xi_{1,1} f'(x) + \frac{\xi_{2,1}}{2} f''(x), \qquad V(x) = \xi \frac{f(x)}{x^{1/2}},$$

Où les constantes $\xi_{1,1} > 0$, $\xi_{2,1} > 0$ $\xi > 0$. De plus, nous utilisons les notations I^{B^2}, I^V, $I^{|B|}$, et $I^{V^{1/2}}$, où $I^* = \int_0^\infty *(x)\,dx$.

Nous commençons par donner les hypothèses suivants :

Hypothèses

(H4) : $\forall y \geq 0$, le j^{eme} moment $\mu_j(y) = \int_0^\infty \mu^j p(u,y)\,du$, existe pour $j = 1,2,4$, avec $\sup_{y\geq 0}(\widehat{\mu_j'})/(1+y)^k < \infty$, pour $k = 1,2$.

(H5) : Pour $\sup_{x\geq 0} u_h(x) \leq L_k h^{-1}$ pour une constante $L_k > 0$ (indépendante de h), $\forall x > 0$, $u_h(x) \leq L_k'(hx)^{\frac{-1}{2}}$ pour une constante $L_k' > 0$ (indépendante de h et x).

(H6) : étant donné $\tau \in (0,1)$, et pour tout suffisamment petit $h > 0$,

$$\int_0^\infty \int_{h^{-\tau}}^\infty K(s,h,x)f(s) = o(h^{\tau(k+1)})\,ds\,dx,$$

à condition que $\int_0^\infty s^{K+1}f(s)$ existe pour une certaine constante $k > 0$.

(H7) : f est bornée et deux fois continument différentiable, où f , f' et f'' sont bornés.

(H8) :$(\gamma_n) \in \mathcal{GS}(-\gamma)$, $(h_n) \in \mathcal{GS}(-c)$ et $\lim_{n\to\infty}(n\gamma_n)^-1 = \xi$ pour une certaine constante $0 < a \leq 1$, $c > 0$ et $\xi \geq 0$ avec

$$\begin{cases} 0 < c < 2\gamma & \text{et } 0 \leq \xi < \max(\frac{1}{c}, \frac{2}{(\gamma - c/2)}), \\ 0 < c < 2\gamma & \text{et } 0 \leq \xi < \frac{2}{(\gamma-c/2)}, \\ 0 < c < 2\gamma & \text{et } 0 \leq \xi < \frac{2}{\gamma}. \end{cases} \tag{2.21}$$

(H9) : f'' satisfait la condition de Holder d'ordre $\eta \in (0,1)$, c'est-à-dire qu'il existe une constante $L > 0$ telle que $|f''(u) - f''(v)| \leq L|u-v|^\eta$ pour tout $u, v \geq 0$.

Théorème 2.3.1 (Kakizawa [46]). *Soit $\widetilde{f_n}(x)$ l'estimateur récursif à noyau asymétrique défini par (2.20). Pour une cible donnée x, si les hypothèses **(H4)-(H5)-(H7)-(H8)**, et **(H9)** sont valables, Le biais de l'estimateur récursif à noyau asymétrique prend l'approximation suivante*

$$\text{Biais}[\widetilde{f_n}(x)] = \begin{cases} \frac{1}{(1-c\xi)}(h_n B(x)) + o(h_n), & si\ 0 < c \leq \frac{2}{5}\gamma, \\ o((\gamma_n h_n^{-\frac{1}{2}})^{\frac{1}{2}}), & si\ \frac{2}{5}\gamma < c \leq 2\gamma. \end{cases} \tag{2.22}$$

Théorème 2.3.2 (Kakizawa [46]). *Soit $\widetilde{f_n}(x)$ l'estimateur récursif à noyau asymétrique défini par (2.20) . Pour une cible donnée x, si les hypothèses **(H4)-(H5)-(H7)-(H8)**, et **(H9)** sont valables, La variance de l'estimateur récursif à noyau asymétrique prend l'approximation suivante :*

$$\mathbf{V}[\widetilde{f_n}(x)] = \begin{cases} o(h_n^2), & si\ 0 < c \leq \frac{2}{5}\gamma, \\ \frac{\gamma_n h_n^{-1/2} V(x)}{2-(\gamma-c/2)\xi} + o\Big(\gamma_n h_n^{-1/2}\Big), & si\ \frac{2}{5}\gamma < c \leq 2\gamma. \end{cases} \tag{2.23}$$

Théorème 2.3.3 (Kakizawa [46]). *Soit $\widetilde{f_n}(x)$ l'estimateur récursif à noyau asymétrique défini par (2.20) . Pour une cible donnée x, si les hypothèses **(H4)-(H5)-(H7)-(H8)**, et **(H9)** sont valables, La normalité asymptotique de l'estimateur récursif à noyau asymétrique prend l'approximation suivante :*

1. Si $\frac{2}{5}\gamma < c < 2\gamma$, $\lim_{n\to\infty}(\gamma_n^{-1}h_n^{\frac{5}{2}}) = 0$, *alors*

$$\boxed{\frac{\widetilde{f_n}(x) - \mathbf{E}(\widetilde{f_n}(x))}{\sqrt{\mathbf{V}(\widetilde{f_n}(x))}} \to \mathcal{N}\Big(0, \Big[2 - (\gamma - c/2)\xi\Big]^{-1} V(x)\Big).} \tag{2.24}$$

2. Si $c = \frac{2}{5}$, $\lim_{n\to\infty}(\gamma_n^{-1}h_n^{\frac{5}{2}}) = \omega \geq 0$, *alors*

$$\boxed{\frac{\widetilde{f_n}(x) - \mathbf{E}(\widetilde{f_n}(x))}{\sqrt{\mathbf{V}(\widetilde{f_n}(x))}} \to \mathcal{N}\Big(\omega^{\frac{1}{2}}\Big[2 - (1 - \frac{2}{5}\gamma)\xi\Big]^{-1} B(x), \Big[2 - (1 - \frac{2}{5}\gamma)\xi\Big]^{-1} V(x)\Big).} \tag{2.25}$$

Théorème 2.3.4 (Kakizawa [46]). *Si les hyphothèses* ***(H4), (H6), (H7)*** *et* ***(H8)*** *sont vérifiées, si* $\int_0^\infty \{f'(x)\}^2\,dx$ *existent,* $\int_0^\infty \{xf''(x)\}^2\,dx$, *et* $\int_0^\infty x^{k+1}f(x)\,dx$ *existent pour une certaine constante* $k > (6+\eta)/\eta$, *où* $\eta \in (0,1)$, *alors le MISE de* $\widetilde{f_n}$ *est donné :*

$$\mathrm{MISE}[\widetilde{f_n}(x)] = \begin{cases} \frac{h_n^2 I^{B^2}}{(1-c\xi)^2} + o\Big(h_n^2\Big), & si\ 0 < c < \frac{2}{5}\gamma, \\ \frac{h_n^2 I^{B^2}}{(1-(2/5)\gamma\xi)^2} + \frac{\gamma_n h_n^{-1/2} I^V}{2-(4/5)\gamma\xi} + o(h_n^2 + \gamma_n h_n^{-1/2}), & si\ c = \frac{2}{5}\gamma, \\ \frac{\gamma_n h^{-1/2} I^V}{2-(\gamma - c/2)\xi} + o(\gamma_n h^{-1/2}), & si\ \frac{2}{5}\gamma < c < 2\gamma. \end{cases} \tag{2.26}$$

Remarque 2.3.1. *Le terme principal du AMISE est donné par*

$$\mathrm{AMISE}[\widetilde{f_n}(x)] \in \begin{cases} \mathcal{GS}(-2c), & si\ 0 < c < \frac{2}{5}\gamma, \\ \mathcal{GS}(-\frac{4}{5}\gamma, & si\ c = \frac{2}{5}\gamma, \\ \mathcal{GS}(-(\gamma - c/2)), & si\ \frac{2}{5}\gamma < c < 2\gamma. \end{cases}$$

Par conséquent, la valeur de $c = \frac{2}{5}\gamma$ doit être imposée pour minimiser $AMISE[\widetilde{f_n}]$ dans ce cas, nous notons que $\mathrm{AMISE}[\widetilde{f_n}] \in \mathcal{GS}(-(4/5)\gamma)$ où $\gamma \in]0,1]$ (ainsi, $\gamma = 1$ doit être imposé). En particulier, avec $\gamma = 1$ et $c = \frac{2}{5}, \gamma = \frac{2}{5}$, nous avons

(En supposant que $B(x) \neq 0$)

$$\mathrm{AMISE}[\widetilde{f_n}] = \frac{h_n^2 I^{B^2}}{(1-(2/5)\xi)^2} + \frac{\gamma_n h_n^{-1/2} I^V}{2-(4/5)\xi} \geq \frac{5}{4^{4/5}}\Big[\frac{\gamma_n^2}{4(1-(2/5)\xi)^3}\Big]^{2/5} (I^{B^2})^{1/5}(I^V)^{4/5}. \tag{2.27}$$

2.4 Comparaison de MISE récursif et non récursif

La Remarque 1 indique que, avec $(\gamma_n) = (1/n) \in \mathcal{GS}(-1)$, l'estimateur récursif $\widetilde{f_n}$, en utilisant $(h_n) \in \mathcal{GS}(-2/5)$, atteint le minimum de AMISE.

$$\boxed{\mathrm{AMISE}[\widetilde{f_n}] = \frac{5}{4^{4/5}}\frac{1}{4(3/5)^3}^{2/5} (I^{B^2})^{1/5}(I^V)^{4/5} n^{-4/5},} \quad si \quad B(x) \neq 0. \tag{2.28}$$

Pour mettre en contexte, nous avons des références, telles que les travaux d'Igarashi et Kakizawa [45], ainsi que celui de Kakizawa [42], qui démontrent que l'estimateur non récursif $\widehat{f_n}$ donne

$$\boxed{\text{AMISE}[\widehat{f_n}] = \beta_n^2 I^{B^2} + \beta^{-1/2} I^V \geq \frac{5}{4^{4/5}} (I^{B^2})^{1/5} (I^V)^{4/5} n^{-4/5},} \qquad B(x) \neq 0. \tag{2.29}$$

L'efficacité asymptotique (en termes de l'AMISE) de l'estimateur récursif par rapport à l'estimateur non récursif est égale à $\{4(3/5)^3\}^{-2/5} \approx 1,06$, c'est à dire que la perte d'efficacité est assez faible.

2.5 Conclusion

Dans ce chapitre, nous avons exposé l'estimateur récursif à noyau, aussi bien symétriques qu'asymétriques, pour la densité de probabilité. Nous avons également prèsenté les propriétés de ces estimateurs. Dans le chapitre suivant, nous nous consacrerons à la mise en pratique de tout ce qui a été présenté jusqu'à présent. Nous réaliserons des simulations et des études pratiques pour approfondir notre compréhension et évaluer l'efficacité des différents estimateurs abordés.

3
Simulation

3.1 Introduction

Nous présentons dans ce chapitre une étude de simulation effectuée à l'aide du logiciel Python, pour essayer d'illustrer les différents aspects théoriques abordés dans les chapitres précédents. Cette illustration numérique nous servira à étudier la performance de deux types d'estimateurs de densité à noyau : récursif et non récursif pour des densités symétriques et asymétriques dans le cas continu sur des écgantillion de taille finie.

L'outil statistique Python est un langage de programmation très populaire pour l'analyse de données, créé par Guido van Rossum [17]. Il est largement adopté par les professionnels des statistiques et les data scientists grâce à sa syntaxe simple, sa flexibilité et son écosystème riche en bibliothèques puissantes comme Pandas, NumPy, SciPy, Statsmodels, Scikit-learn, Matplotlib et Seaborn. Ces outils permettent de manipuler, analyser et visualiser les données avec une grande efficacité.

3.2 Plan de simulation

Nous utilisons pour la simulation des échantillons de variables aléatoires indépendantes et identiquement distribuées (iid) de taille 50, 100, 300 et 500 pour chaque densité avec un nombre de réplications $Rep = 50$. Notons aussi que dans la programmation, nous avons utilisé le noyau gaussien dans le cas symétrique et le noyau gamma modifier dans le cas asymétrique pour la correction des effets de bords dans l'estimation d'une densité bornée. Pour le choix du paramètre de lissage, nous avons utilisé la methode classique de Rule of Thumb, notée h_{rot}. Ainsi, le paramètre de lissage est choisi comme $h_n = h_{rot}$.

3.2.1 Scénario de simulation 1 : cas symétrique

Cette étude permet de déterminer l'efficacité des estimateurs de densité à noyau récursif et non récursif, dans des conditions de distribution symétrique. Nous utilisons trois densités de test, chacune présentant des caractéristiques différentes.

— $\mathbf{D}_1$ une densité de loi normale centrée et réduite : $f_1 \sim \mathcal{N}(0,1)$.

— $\mathbf{D}_2$ Le mélange de deux densités de loi normale : $f_2 \sim \frac{1}{2}\mathcal{N}(-1,0.5) + \frac{1}{2}\mathcal{N}(1,0.5)$.

— $\mathbf{D}_3$ Le mélange de trois densités de loi normale : $f_3 \sim \frac{1}{3}\mathcal{N}(-1,0.5) + \frac{1}{3}\mathcal{N}(0.5,0.5) + \frac{1}{3}\mathcal{N}(2,0.5)$.

3.2.2 Scénario de simulation 2 : cas asymétrique

Les distributions asymétriques, où les données ne sont pas uniformément réparties autour de leur centre, présentent des défis supplémentaires pour l'estimation de densité, nécessitant des techniques plus sophistiquées pour obtenir des estimations précises. Pour garantir une évaluation complète et variée, nous avons sélectionné trois différentes densités.

— $\mathbf{D}_4$ une densité de la loi exponentielle avec $\lambda = 3$: $f_4 \sim \frac{1}{3}\exp^{-x/3}$
— $\mathbf{D}_5$ une densité de la loi gamma avec $k = 2, \theta = 2$: $f_5 \sim \frac{x}{4}\exp^{(-x/2)}$
— $\mathbf{D}_6$ une densité de la loi lognormal avec $\mu = 0$, $\sigma = 1$: $f_6 \sim \frac{1}{x\sqrt{2\pi}}\exp^{-(\ln(x))^2/2}$

3.3 Critère de performance

Les performances des estimateurs sont comparées en utilisant l'erreur quadratique intégrée (ISE) , définie par :

$$ISE(h) = \int_{\mathbb{R}} \left(\widehat{f_n}(x) - f(x)\right)^2 dx \tag{3.1}$$

L'ISE mesure la différence entre la densité estimée $\widehat{f_n}(x)$ ou $(\widetilde{f_n}(x))$ et la densité vraie $f(x)$ sur tout l'intervalle réel. Une valeur d'ISE plus faible indique une meilleure performance de l'estimateur.

3.4 Résultats de la simulation

Les résultats de la simulation sont présentés sous forme de tableaux et de graphiques, ce qui permet une compréhension approfondie et une visualisation claire des données. Les tableaux offrent une vue structurée des valeurs numériques et des mesures clés, simplifiant la comparaison et l'analyse détaillée des résultats. Les graphiques, quant à eux, permettent de visualiser les tendances et les distributions de manière intuitive, rendant les informations plus accessibles. Cette approche combinée permet aux lecteurs d'interpréter facilement les résultats et de tirer des conclusions significatives à partir des données de simulation.

3.4.1 Résultats de simulation 1 : cas symétrique

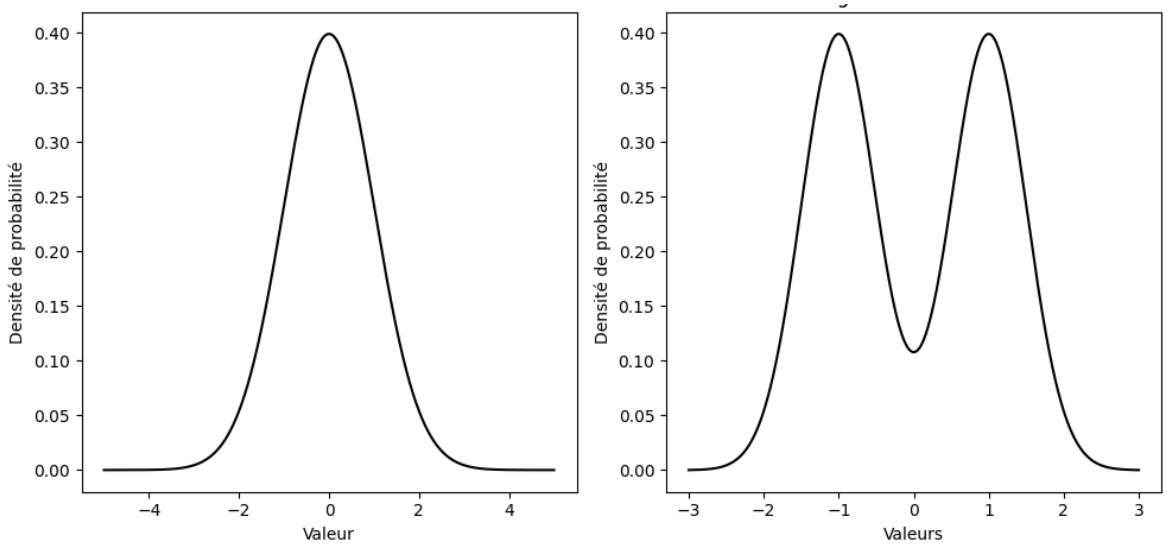

FIGURE 3.1 – La vraie densité de $\mathbf{D}_1$ et $\mathbf{D}_2$

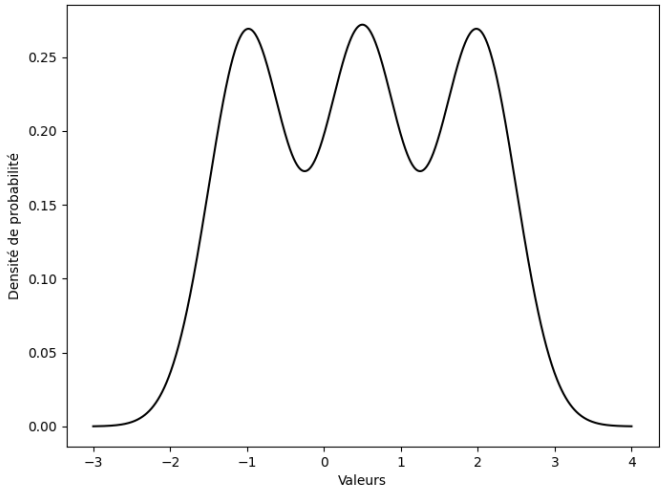

FIGURE 3.2 – La vraie densité $\mathbf{D}_3$

La figure (3.1) présente deux graphiques illustrant des distributions de probabilité selon la loi normale.

— Graphique de gauche : Ce graphique montre une distribution normale classique. La courbe représente la densité de probabilité théorique, qui est une courbe symétrique en forme de cloche centrée sur zéro, avec un écart-type de 1. Cette distribution est souvent utilisée comme référence pour comparer d'autres distributions en raison de sa symétrie et de son pic unique.

— Graphique de droite : Ce graphique illustre un mélange de deux lois normales distinctes. La présence de deux pics distincts indique que les données proviennent de deux sous-populations différentes, chacune ayant sa propre distribution normale.

La figure (3.2) présente un graphique montrant un mélange de trois lois normales, identifiable par trois pics distincts. Chaque pic représente une sous-population différente.

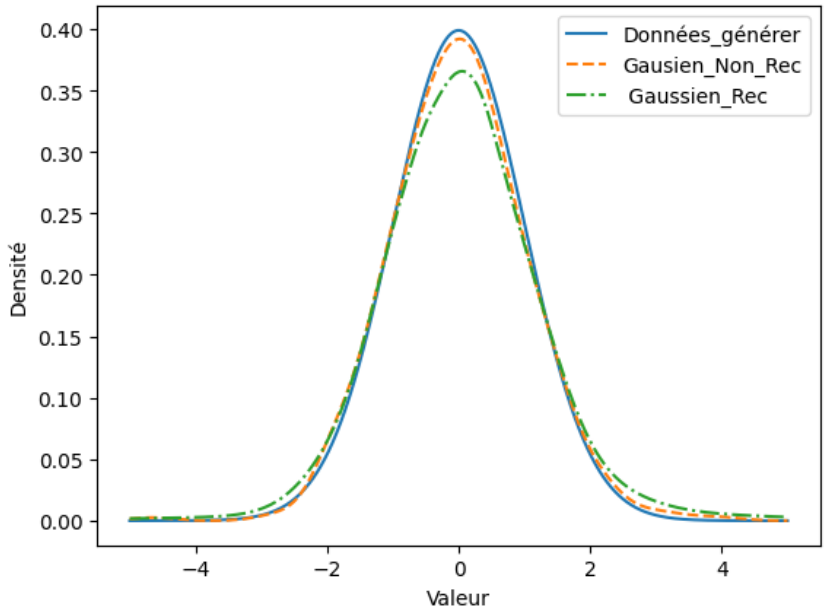

FIGURE 3.3 – Comparaision de l'estimateur récursif et non récursif $\mathbf{D}_1$

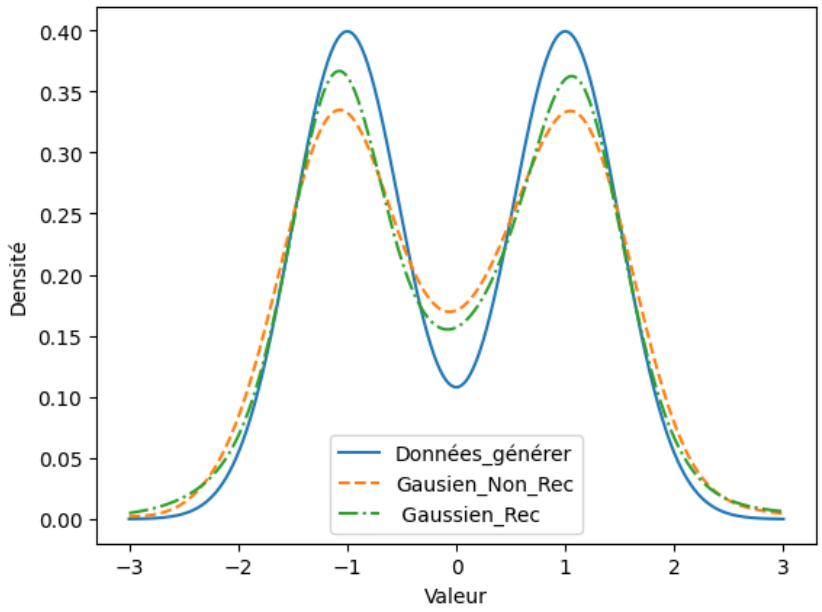

FIGURE 3.4 – Comparaision de l'estimateur récursif et non récursif $\mathbf{D}_2$

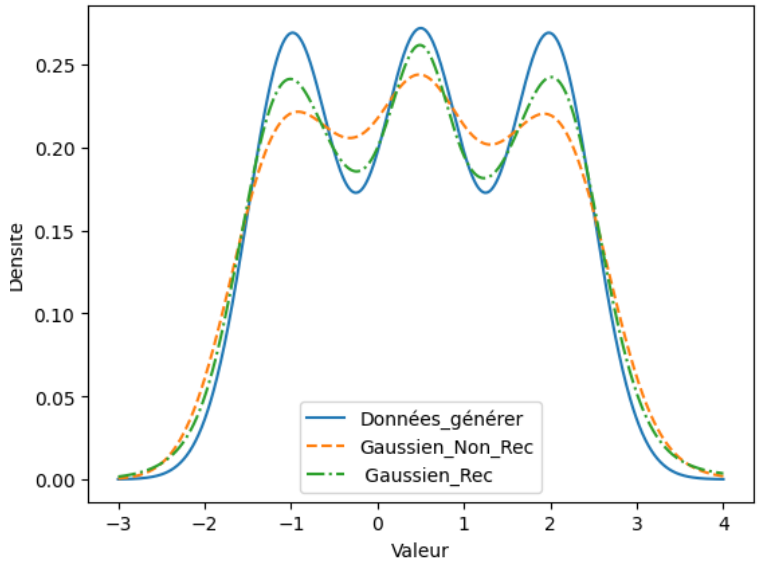

FIGURE 3.5 – Comparaision de l'estimateur récursif et non récursif $\mathbf{D}_3$

Les figures (3.3) (3.4) (3.5) comparent les estimations de densité de probabilité des données générées à la densité réelle en utilisant deux méthodes : l'estimateur de Parzen (1.1) (méthode classique) et l'estimateur d'Amiri (2.8) (méthode récursive) avec un noyau gaussien. La densité réelle des données générées est représentée par une ligne bleue continue. L'estimation gaussienne classique est illustrée par une ligne en pointillés orange, tandis que l'estimation gaussienne récursive est représentée par une ligne en tirets-pointillés verts. Les résultats sont basés sur un échantillon de taille $n = 500$ observations pour chaque densité de probabilité.

Discussion 1

Les résultats de l'estimation de la densité $\mathbf{D}_1$ montrent que les deux méthodes fournissent des estimations de densité proches de la réalité, l'estimateur classique semble offrir une meilleure correspondance avec les caractéristiques principales de la distribution , notamment le pic. L'estimateur récursif, bien qu'il soit très précis, présente des ajustements plus lissés, ce qui peut entraîner de légères déviations par rapport à la densité réelle, notamment dans les régions où la densité varie rapidement.

Contrairement aux résultats de l'estimation des densités $\mathbf{D}_2$ e t$\mathbf{D}_3$ Les deux figures représentées par (3.4) et (3.5) illustrent que l'estimateur récursif offre une estimation légèrement plus précise que l'estimateur classique en ce qui concerne la densité réelle. Bien que les deux estimateurs soient proches de la vraie densité, l'estimateur récursif semble mieux suivre les contours des pics et des creux de cette densité réelle.

Dans l'estimation par noyau, le critère de performance (précision) est généralement mesuré en évaluant la capacité de l'estimateur à suivre fidèlement la forme réelle de la distribution. La précision des deux estimateurs, récursif et non récursif, est affichée dans le tableau (3.2).

f	Rep	n	$\widehat{f_n}$	$\widetilde{f_n}$
$\mathbf{D}_1$	50	50	0.01115	0.01187
		100	0.00551	0.00644
		300	0.00262	0.00404
		500	0.00177	0.00359
$\mathbf{D}_2$	50	50	0.02888	0.02112
		100	0.02100	0.01297
		300	0.01097	0.00617
		500	0.00853	0.00476
$\mathbf{D}_3$	50	50	0.01262	0.01136
		100	0.01023	0.00798
		300	0.00623	0.00355
		500	0.00518	0.00276

TABLE 3.1 – Les valeurs de l'ISE moyen (ISE) basées sur 50 réplications

L'analyse des résultats montre une tendance significative : l'erreur quadratique intégrée (ISE) moyenne décroît notablement avec l'augmentation de la taille de l'échantillon, un constat valable pour les deux estimateurs étudiés. Cependant, lorsque l'échantillon devient considérablement grand, l'estimateur récursif se distingue par sa supériorité en termes de performance par rapport à l'estimateur non récursif.

Cette observation se précise davantage lorsqu'on considère les différentes densités utilisées : l'estimateur récursif se montre particulièrement efficace pour les densités $\mathbf{D}_2$ et $\mathbf{D}_3$, bien qu'il affiche des résultats légèrement moins satisfaisants pour la densité $\mathbf{D}_1$ comparé à l'estimateur non récursif.

3.4.2 Résultats de simulation 2 : cas asymétrique

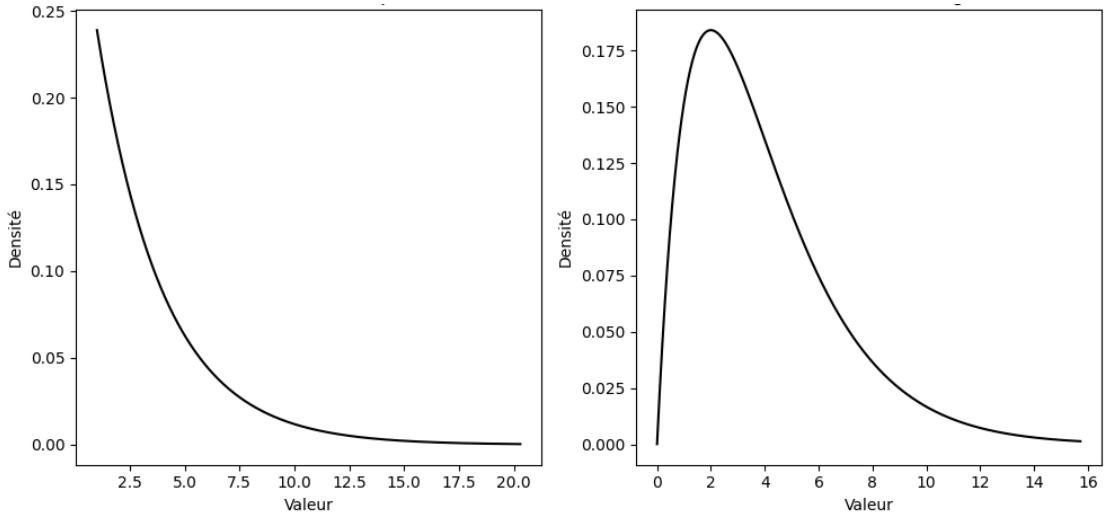

FIGURE 3.6 – La vraie densité de probabilité $\mathbf{D}_4$ et $\mathbf{D}_5$

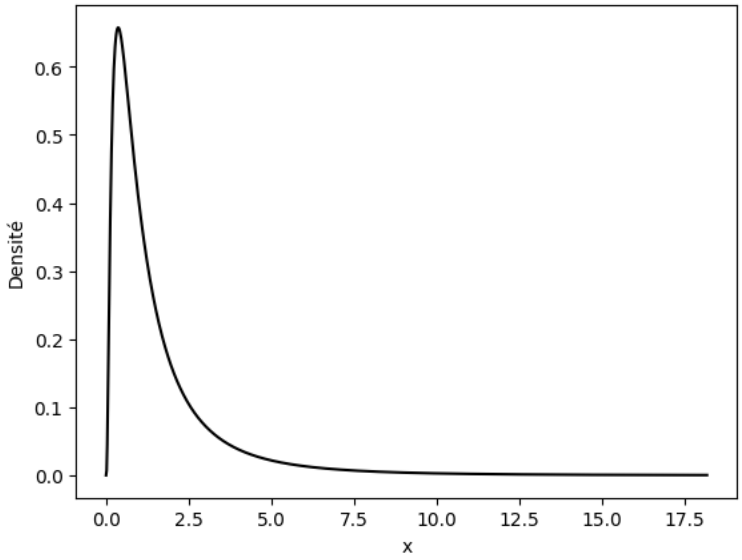

FIGURE 3.7 – La vraie densité de probabilité $\mathbf{D}_6$

La figure (3.6) présente deux graphiques qui décrivent les distributions de probabilité exponentielle et gamma.

La figure (3.7) représente une distribution log-normale, caractérisée par son asymétrie positive et sa longue queue s'étendant vers la droite.

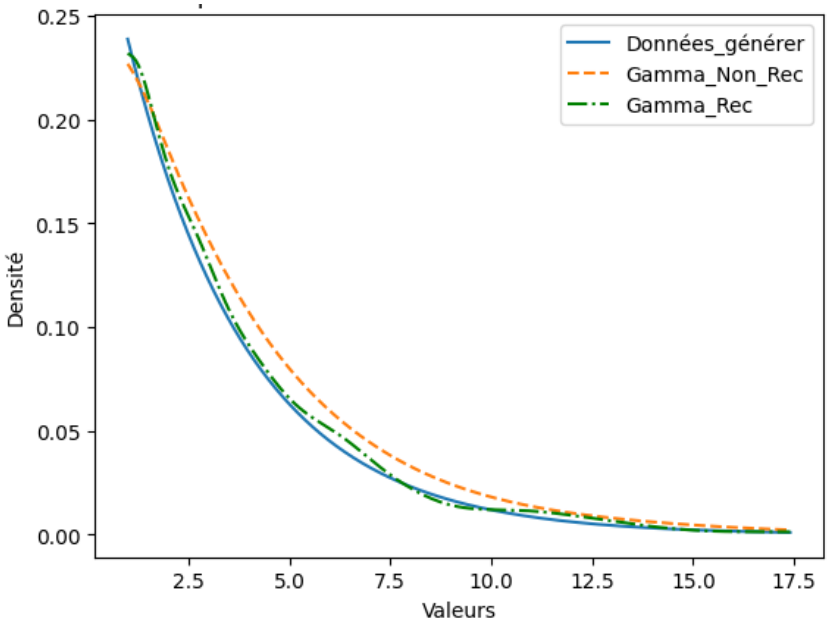

FIGURE 3.8 – Comparaision de l'estimateur récursif et non récursif $\mathbf{D}_4$

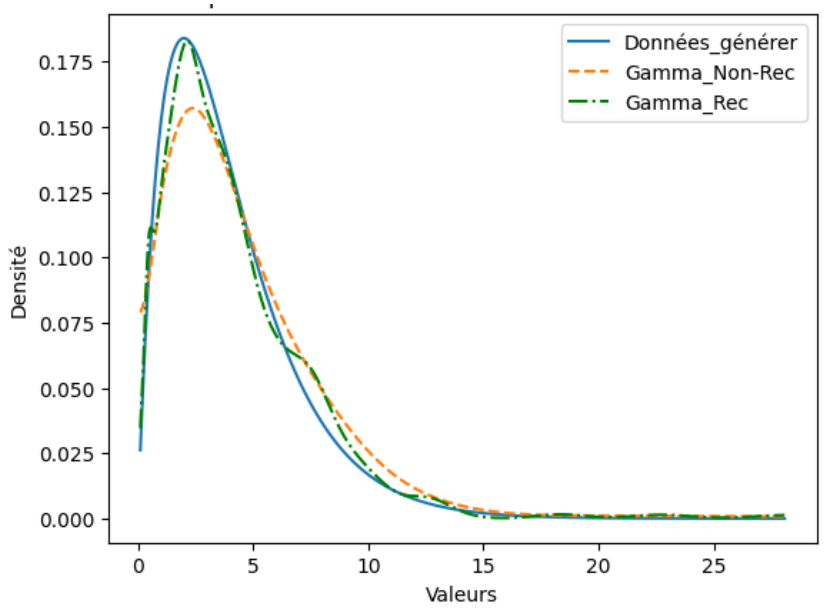

FIGURE 3.9 – Comparaision de l'estimateur récursif et non récursif $\mathbf{D}_5$

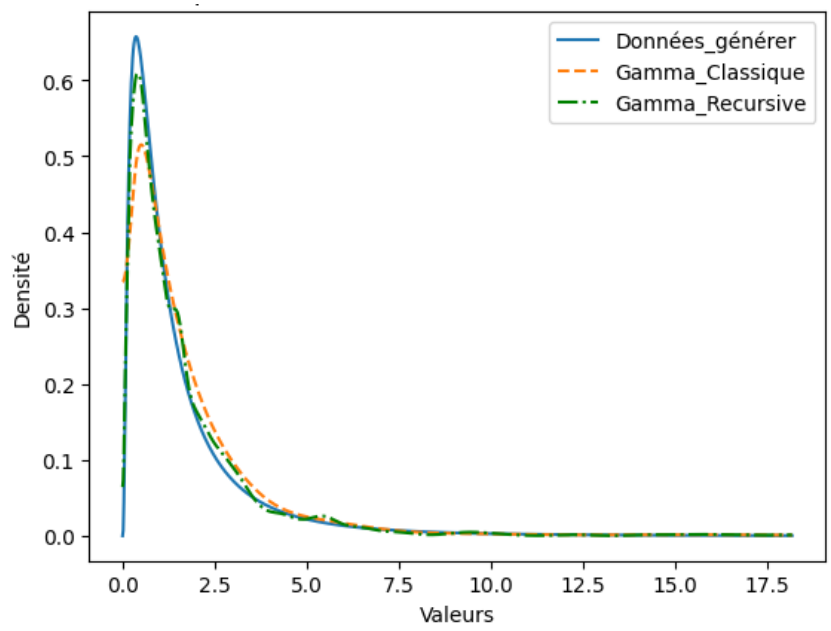

FIGURE 3.10 – Comparaision de l'estimateur récursif et non récursif $\mathbf{D}_6$

Les figure (3.8) (3.9) (3.10) comparent deux méthodes d'estimation de densité de probabilité pour les données générées par rapport à la densité réelle. D'une part, l'estimateur non récursif (1.7) est utilisé, et d'autre part, l'estimateur récursif (Kakizawa pour le cas asymétrique) (2.20) en utilisant le noyau gamma modifié.

Dans le cas de l'estimateur récursif, la taille du pas est fixée à $\gamma_n = \frac{c}{n}$ avec $c = 0.8$ comme paramètre spécifique. Le paramètre de lissage $h_n = h_{rot}$, avec $h_{rot} = 1.06\,\sigma_n\, n^{-\frac{2}{5}}$. Pour l'estimateur non récursif, nous avons également choisi le paramètre de lissage h_n en utilisant la méthode

classique (plugin, rule of thumb) avec $h_n = h_{rot}$ (1.12)

Discussion 2

La courbe bleu des données générées représente la distribution d'une loi exponentielle, gamma et lognormal dans les figures (3.8) (3.9) et (3.10), tandis que les courbes orange (non récursif) et verte (récursif) représentent les estimations des densités par noyau Gamma modifié.

On observe que, dans les trois cas l'estimateur récursif (courbe verte) s'ajuste mieux aux données générées que l'estimateur non récursif ou classique (courbe orange), particulièrement pour les valeurs élevées et proches de zéro. L'estimateur récursif offre une meilleure estimation en restant plus proche des données générées sur toute la plage de valeurs, indiquant une précision supérieure par rapport aux estimateurs non récursif et classique, qui tendent à sous-estimer la densité.

f	Rep	n	$\widehat{f_n}$	$\widetilde{f_n}$
$\mathbf{D}_4$	50	50	0.01012	0.00741
		100	0.00722	0.00421
		300	0.00429	0.00212
		500	0.00334	0.00133
$\mathbf{D}_5$	50	50	0.00465	0.00571
		100	0.00277	0.00351
		300	0.00142	0.00265
		500	0.00085	0.00154
$\mathbf{D}_6$	50	50	0.03763	0.02228
		100	0.02852	0.01281
		300	0.02104	0.00536
		500	0.01712	0.00381

TABLE 3.2 – Les valeurs de l'ISE moyen (ISE) basées sur 50 réplications de l'estimateur asymétrique

L'analyse de l'erreur quadratique intégrée (ISE) dans le cas asymétrique présente des similitudes avec celle du cas symétrique. Globalement, on observe une décroissance significative de l'ISE moyenne avec l'augmentation de la taille de l'échantillon pour les deux cas. Cependant, lorsque l'échantillon devient considérablement grand, des nuances apparaissent.

Dans le contexte asymétrique, l'estimateur récursif se démarque également par sa performance supérieure à celle de l'estimateur non récursif, tout comme dans le cas symétrique. Cette distinction devient particulièrement notable lorsque l'on considère les différentes densités. Par exemple, pour les densités $\mathbf{D}_4$ et $\mathbf{D}_6$, l'estimateur récursif montre une efficacité remarquable, tandis que pour la densité $\mathbf{D}_5$, ses résultats peuvent être légèrement inférieurs à ceux de l'estimateur non récursif.

Ainsi, bien que les tendances générales soient similaires entre les cas asymétrique et symétrique, il est essentiel de noter que les performances spécifiques des estimateurs peuvent varier selon la distribution sous-jacente, avec l'estimateur récursif montrant souvent un avantage significatif dans des conditions d'échantillonnage plus importantes.

3.5 Temps d'exécution

Nous avons mené une étude comparative des temps de calcul nécessaire pour évaluer les estimateurs récursifs et non récursifs, en utilisant l'approche "plugin rule of thumb" pour la sélection des paramètres de lissage. Pour commencer, nous avons utilisé un échantillon de 300 observations tirées du la loi normale centrée réduite , noté $X_1, \ldots, X_{300}$. Ensuite, en ajoutant N observations supplémentaires, notées $X_{n+1}, \ldots, X_{n+N}$, nous avons calculé les estimateurs à noyau gaussien en utilisant des sous-échantillons $X_1, \ldots, X_{n+J}$ pour chaque valeur de $J = 1, \ldots, N$. Dans notre cas spécifique avec $(n, N) = (300, 200)$, cela signifie que nous avons étudié les estimateurs en utilisant différentes tailles de sous-échantillons, en augmentant progressivement le nombre d'observations pour évaluer la performance des estimateurs dans des conditions variables.

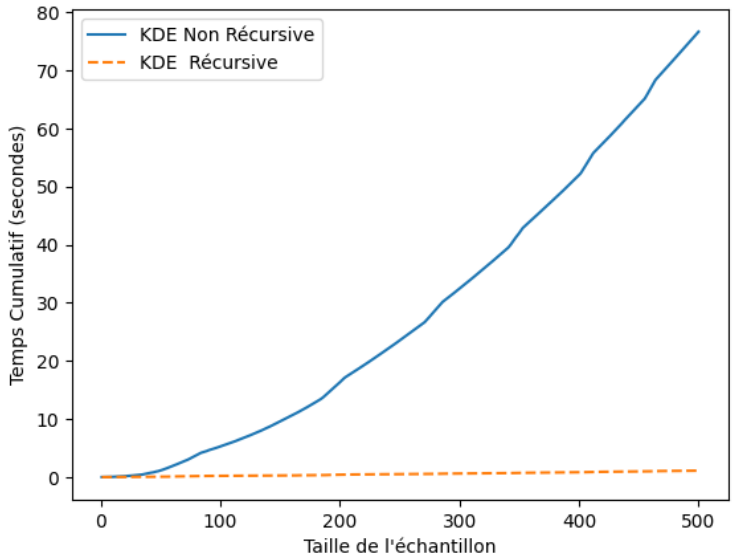

FIGURE 3.11 – Temps d'exécution (en secondes) de l'estimateur à noyau gaussien non récursif et récursif

La figure (3.11) illustre les temps de calcul cumulatifs (en secondes) nécessaires pour calculer les estimateurs non récursifs (ligne continue bleu) et récursifs (ligne pointillée orange) en fonction de la taille de l'échantillon. On observe que le temps de calcul de l'estimateur non récursif augmente lorsque la taille de l'échantillon augmente, contrairement à celui de l'estimateur récursif, qui reste stable.

3.6 Conclusion

Dans ce chapitre, nous avons examiné et comparé les performances des estimateurs récursifs et non récursifs dans des contextes de cas symétriques et asymétriques. Cette analyse nous a permis d'évaluer leur précision et leur efficacité lorsqu'ils sont appliqués à des distributions de densité réelle. Les estimateurs récursifs se démarquent en mettant à jour les estimations avec les nouvelles données sans revoir tout le processus. Cela présente plusieurs avantages par rapport aux estimateurs non récursifs.

4
Application sur données réelles

4.1 Introduction

Dans ce chapitre, nous appliquons les estimateurs récursifs et non récursifs à noyau sur deux jeux de données réelles : Old Faithful et Air pollution. Pour l'estimation de la densité des données, nous utiliserons le noyau Gaussien pour les cas symétriques et le noyau Gamma pour les cas asymétriques. En adoptant la méthode classique du "rule of thumb" pour le choix du paramètre de lissage, nous évaluerons l'efficacité et la précision des deux approches. Cette analyse comparative mettra en lumière les performances respectives des estimateurs sur des ensembles de données divers, démontrant leurs avantages et limites dans des contextes variés.

4.2 Applications sur des données Old Faithful

Dans cette section, nous présentons le premier jeu de données, "Old Faithful", qui provient de mesures effectuées sur le geyser Old Faithful dans le parc national de Yellowstone. Ce jeu de données contient deux variables principales : "eruptions", qui représente la durée des éruptions (en minutes), et "waiting", qui indique le temps d'attente jusqu'à la prochaine éruption (en minutes). Il comprend généralement 272 observations.

Nous nous concentrons particulièrement sur la variable "eruptions", car elle offre un aperçu détaillé de la dynamique temporelle du geyser.

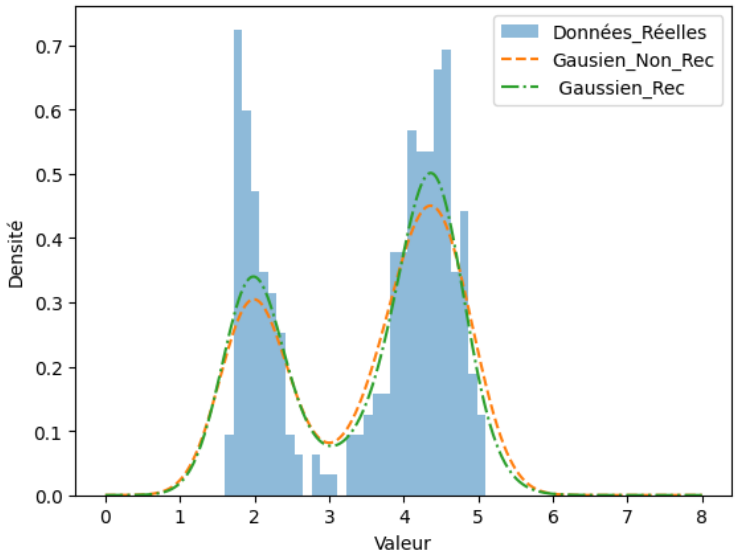

FIGURE 4.1 – Comparaision de l'estimateur récursif et non récursif par un noyau gaussien

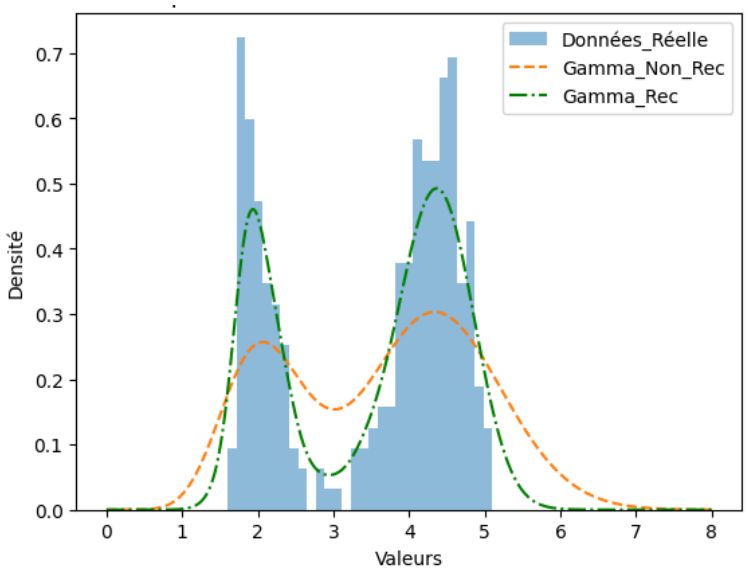

FIGURE 4.2 – Comparaision de l'estimateur récursif et non récursif par un noyau gamma

Les figures (4.1) et (4.2) comparent les estimations de densité de probabilité des données "Old Faithful" à la vraie densité en utilisant deux méthodes : l'estimateur de Parzen (1.1) (méthode classique), représenté par une ligne en pointillés orange, et l'estimateur d'Amiri (2.8) (méthode récursive), représenté par une ligne en tirets-pointillés verts. Pour le cas symétrique, nous avons utilisé le noyau gaussien, tandis que pour le cas asymétrique, nous avons utilisé le noyau gamma modifié.

L'estimateur récursif est particulièrement efficace pour la distribution bimodale réelle observée dans les données "Old Faithful". Il offre une flexibilité accrue et une capacité d'ajustement fine, ce qui lui permet de modéliser avec précision en capturant les caractéristiques clés telles que le nombre de modes, l'asymétrie et les irrégularités de forme.

4.3 Applications sur des données Air pollution

Les données quotidiennes d'ozone à New York entre mai et septembre 1973 fournissent un aperçu détaillé des variations saisonnières de la pollution atmosphérique, permettant ainsi d'identifier les tendances météorologiques et les événements exceptionnels qui influencent la qualité de l'air.

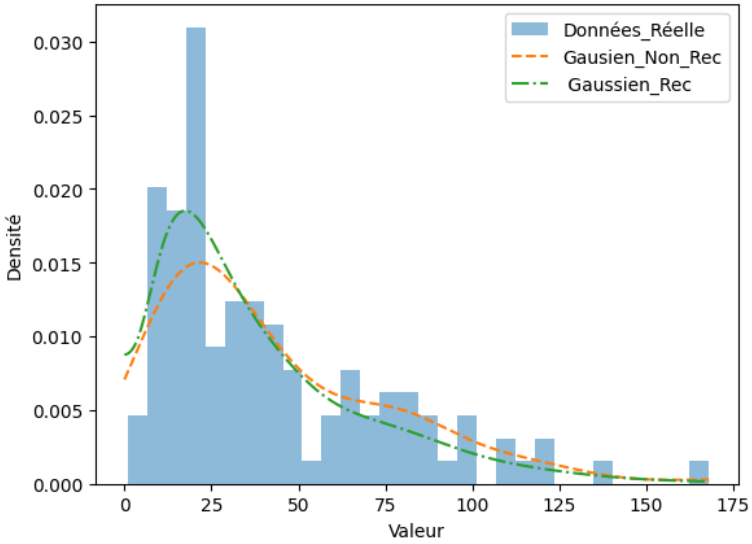

FIGURE 4.3 – Comparaision de l'estimateur récursif et non récursif par un noyau gaussien

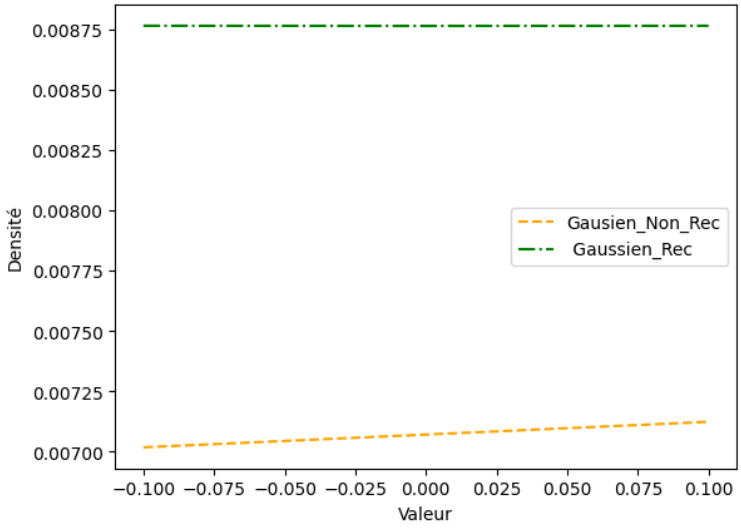

FIGURE 4.4 – Effet du bord pour le noyau gaussien

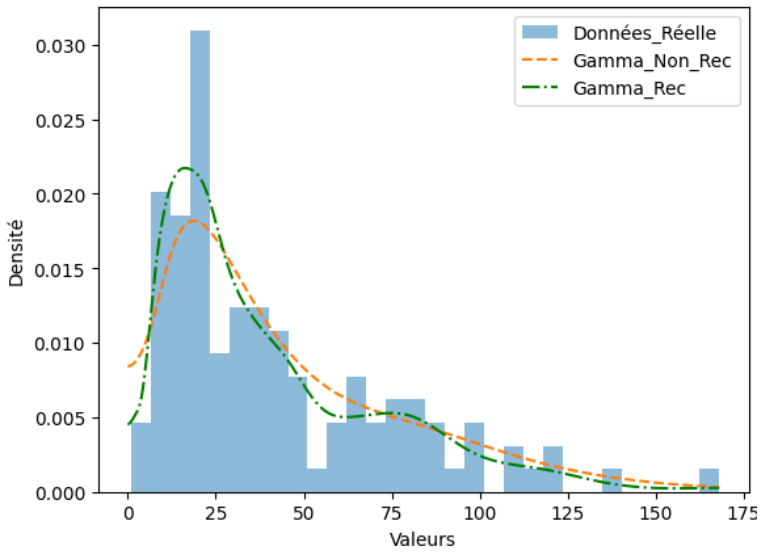

FIGURE 4.5 – Comparaision de l'estimateur récursif et non récursif par un noyau gamma

Discusion :

Les figures (4.3) (4.4) (4.5) comparent les estimations de densité de probabilité des données "Air pollution " à la vraie densité en utilisant deux méthodes : l'estimateur (1.7) (méthode classique) est représentée par une ligne en pointillés orange et l'estimateur (2.20) (méthode récursive) est représenté par une ligne en tirets-pointillés verts. pour le cas symétrique, asymétrique on a utilisé le noyau gaussien et gamma (respectivement).

Les estimateurs récursifs offrent un ajustement supérieur par rapport aux non récursifs pour les noyaux Gaussien et Gamma. Les figures (4.3) et (4.5) montrent clairement que l'estimateur récursif fournit la meilleure estimation des données réelles, particulièrement pour les petites valeurs et les extrémités de la distribution.

La figure (4.4) zoome sur une petite plage de valeurs près de zéro, permettant de voir plus clairement les détails de l'estimation de la figure (4.3) près du bord. L'estimateurs récursif et non récursif présentent toutes les deux des effets de bord lorsqu'un noyau gaussien est utilisé. Pour corriger ces effets, un noyau gamma modifié a été employé, offrant une meilleure précision (3.4)

4.4 Conclusion

Notre analyse comparative des estimateurs récursifs et non récursifs, appliqués aux jeux de données réelles d'Old Faithful et de la pollution de l'air, a révélé des différences significatives en termes de précision et d'efficacité. En utilisant les noyaux gaussien et gamma modifié pour les cas symétriques et asymétriques respectivement, et en adoptant la méthode classique du "rule of thumb" pour le choix du paramètre de lissage, nous avons pu réaliser une évaluation rigoureuse des deux approches. Les résultats démontrent que, même si les estimateurs non récursifs peuvent être performants dans certains cas, les estimateurs récursifs se révèlent plus appropriés pour les applications qui exigent des mises à jour régulières et rapides des estimations.

Conclusion générale et travaux futures

Dans ce mémoire, nous avons introduit une méthode d'estimation de la densité de probabilité dans le cas continu en utilisant la méthode du noyau. L'applicabilité de cette méthode nécessite, préalablement, le choix du noyau et du paramètre de lissage. Dans un premier temps, nous avons brièvement présenté un rappel sur l'estimation non récursive par noyaux dans l'estimation de la densité de probabilité, plus précisément celles des estimateurs à noyaux continus (classiques) de Rosenblatt [2] et Parzen [3], Nous avons aussi présenté les différents noyaux : symétriques et asymétriques pour l'estimation de la densité dans le contexte des données indépendantes. Ensuite, nous avons présenté des estimateurs récursifs à noyaux symétrique et asymétriques (continus) et leurs propriétés. Nous avons établi les propriétés asymptotiques, y compris le biais et la variance, ainsi que l'erreur quadratique intégrée moyenne *MISE* en tant que mesure globale. Dans un second temps, nous avons étudions le choix optimal du paramètre de lissage h qui est d'importance capitale dans l'estimation de la fonction densité par la méthode du noyau (méthode plug-in, validation croisée non biaisée *UCV*). Une étude de simulation a comparé les performances des estimateurs à noyau continu récursif à ceux des estimateurs non récursifs en utilisant le critère *ISE*, en utilisant des données générées, ensuite on a appliquer les deux estimateurs sur des données réelles.

Les résultats obtenus ont clairement mis en évidence la supériorité des estimateurs récursifs par rapport aux estimateurs non récursifs dans le cas continu. les estimateurs récursifs par noyaux de densité de probabilité offrent une mise à jour progressive des estimations, nécessitant moins de calculs, ce qui les rend particulièrement adaptés pour des échantillons de grande taille. De plus, ces estimateurs se distinguent par leur efficacité en termes de temps de calcul et de mémoire, leur adaptabilité aux environnements dynamiques et leur capacité à intégrer progressivement les nouvelles informations, les rendant ainsi supérieurs aux estimateurs non récursifs. En conclusion, l'utilisation d'estimateurs récursifs se révèle être une stratégie optimale pour une estimation précise, efficiente et robuste de la densité de probabilité dans le cadre de la simulation et de l'analyse des données Réelles.

Les travaux ainsi réalisés offrent des perspectives intéressantes on peut citer :

- Estimation récursive de la fonction densité dans le cas multidimensionnel.
- Étude comparative visant à évaluer différentes méthodes de sélection des paramètres de lissage dans l'estimation récursive
- Une étude comparative entre les données indépendantes et des données dépendantes dans le contexte de l'estimation récursive.

Bibliographie

[1] H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[2] M. Rosenblatt. Remarks in some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 1956.

[3] E. Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3) :1065–1076, 1962.

[4] C. T. Wolverton and T. J. Wagner. Recursive estimates of probability densities. *IEEE Transactions on Systems Science and Cybernetics*, 5(3) :246–247, 1969.

[5] H. Yamato. Sequential estimation of a continuous probability density function and mode. 1971.

[6] P. Révész. Robbins-monro procedure in a hilbert space and its application in the theory of learning processes i. 1973.

[7] H. Davies. Strong consistency of a sequential estimator of a probability density function. 1973.

[8] P. Deheuvels. Sur l'estimation séquentielle de la densité. 1973.

[9] J. Galambos and E. Seneta. Regularly varying sequences. *Proceedings of the American Mathematical Society*, 41(1) :110–116, 1973.

[10] E. Wegman and H. Davies. Remarks on some recursive estimators of a probability density. *The Annals of Statistics*, pages 316–327, 1979.

[11] L. Devroye. The pointwise and the integral convergence of recursive kernel estimates of probability densities. *Utilitas Math.*, 15 :113–128, 1979.

[12] M. Rudemo. Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, pages 65–78, 1982.

[13] A. W. Bowman. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71(2) :353–360, 1984.

[14] B. W. Silverman. *Density estimation for statistics and data analysis*. Routledge, 1986.

[15] D. W. Scott and G. R. Terrell. Biased and unbiased cross-validation in density estimation. *Journal of the american Statistical association*, 82(400) :1131–1146, 1987.

[16] A. B. Tsybakov. Recursive estimation of the mode of a multivariate distribution. *Problemy Peredachi Informatsii*, 26(1) :38–45, 1990.

[17] G. Van Rossum and J. De Boer. Interactively testing remote servers using the python programming language. *CWI quarterly*, 4(4) :283–303, 1991.

[18] G. G. Roussas. Exact rates of almost sure convergence of a recursive kernel estimate of a probability densiy function : Application to regression and hazard rate estimation. *Journal of Nonparametric Statistics*, 1(3) :171–195, 1992.

[19] D. W. Scott. *Multivariate density estimation : theory, practice, and visualization*. John Wiley & Sons, 1992.

[20] M. P. Wand and M. C. Jones. Comparison of smoothing parameterizations in bivariate kernel density estimation. *Journal of the american Statistical Association*, 88(422) :520–528, 1993.

[21] A. Berlinet and L. Devroye. A comparison of kernel density estimates. In *Annales de l'ISUP*, volume 38, pages 3–59, 1994.

[22] P. Hall and P. Patil. On the efficiency of on-line density estimators. *IEEE Transactions on Information Theory*, 40(5) :1504–1512, 1994.

[23] M. Duflo. *Random iterative models*, volume 34. Springer Science & Business Media, 1997.

[24] S. X. Chen. Beta kernel estimators for density functions. *Computational Statistics & Data Analysis*, 31(2) :131–145, 1999.

[25] S. X. Chen. Gamma kernel estimators for density functions. *Annals of the Institute of Statistical Mathematics*, 52(3) :471–80, 2000.

[26] X. Jin, J. Kawczak, et al. Birnbaum-saunders and lognormal kernel estimators for modelling durations in high frequency financial data. *Annals of Economics and Finance*, 4 :103–124, 2003.

[27] O. Scaillet. Density estimation using inverse and reciprocal inverse gaussian kernels. *Nonparametric statistics*, 16(1-2) :217–226, 2004.

[28] T. Senga Kiessé. *Approche non-paramétrique par noyaux associés discrets des données de dénombrement*. PhD thesis, Université de Pau et des Pays de l'Adour, 2008.

[29] A. Mokkadem, M. Pelletier, and Y. Slaoui. The stochastic approximation method for the estimation of a multivariate probability density. *Journal of Statistical Planning and Inference*, 139(7) :2459–2478, 2009.

[30] A. Amiri. Estimateurs fonctionnels récursifs et leurs applications à la prévision. Avignon, 2010.

[31] F. G. Libengue. *Méthode non-paramétrique des noyaux associés mixtes et applications*. PhD thesis, Besançon, 2013.

[32] M. Duflo. *Random iterative models*, volume 34. Springer Science & Business Media, 2013.

[33] N. Zougab. *Approche bayésienne dans l'estimation non paramétrique de la densité de probabilité et la courbe de régression de la moyenne*. PhD thesis, Université de Béjaia, 2013.

[34] Y. Slaoui et al. Bandwidth selection for recursive kernel density estimators defined by stochastic approximation method. *Journal of Probability and Statistics*, 2014, 2014.

[35] M. Hirukawa and M. Sakudo. Family of the generalised gamma kernels : a generator of asymmetric kernels for nonnegative data. *Journal of Nonparametric Statistics*, 27(1) :41–63, 2015.

[36] S. Semmar et al. *Sur l'estimation récursive de la fonction densité conditionnelle pour des variables censurées*. PhD thesis, Université Djillali Liabès de Sidi Bel Abbès, 2016.

[37] Y. Ziane. *Sur l'estimation non paramétrique de l'indice de variabilité et la distribution des densités Heavy Tailed.* PhD thesis, Université Abderahmane MIRA de Bejaia, 2016.

[38] A. M. Mousa, M. Kh. Hassan, and A. Fathi. A new non parametric estimator for pdf based on inverse gamma distribution. *Communications in Statistics-Theory and Methods*, 45(23) :7002–7010, 2016.

[39] I. Belahcene. *Estimation non paramétrique de la fonction densité de probabilité avec un noyau*. PhD thesis, Université Kasdi Merbah Ouargla, 2017.

[40] A. Jemai. *Estimation fonctionnelle non paramétrique au voisinage du bord*. PhD thesis, Poitiers, 2018.

[41] L. Harfouche. *Technique de réduction du biais et approche bayésienne dans l'estimation non paramétrique de la densité par noyaux associés*. PhD thesis, Université de Béjaia-Abderrahmane Mira, 2018.

[42] Y. Kakizawa. Nonparametric density estimation for nonnegative data, using symmetrical-based inverse and reciprocal inverse gaussian kernels through dual transformation. *Journal of Statistical Planning and Inference*, 193 :117–135, 2018.

[43] M. Hirukawa. *Asymmetric Kernel Smoothing : Theory and Applications in Economics and Finance*. Springer, 2018.

[44] Y. Slaoui and A. Jmaei. Recursive density estimators based on robbins-monro's scheme and using bernstein polynomials. *arXiv preprint arXiv :1904.06675*, 2019.

[45] G. Igarashi and Y. Kakizawa. Multiplicative bias correction for asymmetric kernel density estimators revisited. *Computational statistics & data analysis*, 141 :40–61, 2020.

[46] Y. Kakizawa. Recursive asymmetric kernel density estimation for nonnegative data. *Journal of Nonparametric Statistics*, 33(2) :197–224, 2021.

Résumé

Dans ce travail, nous avons étudié la fonction de densité de probabilité en utilisant à la fois les estimateurs récursifs et non récursifs par la méthode du noyau pour des données indépendantes. Nous avons également présenté les propriétés statistiques des deux types d'estimateurs pour les cas symétriques et asymétriques (esperance, biais, variance, erreur quadratique moyenne et erreur quadratique moyenne intégrée). La qualité de l'estimation dépend du choix de paramètre de lissage h et du noyau K, pour lesquels nous avons utilisé des noyaux gaussien et gamma modifier dans nos estimateurs. La méthode classique du plugin (rule of thumb) a été employée pour sélectionner le paramètre de lissage optimal h. Ensuite, nous avons évalué les performances de nos estimateurs à travers des simulations et des données réelles, en utilisant comme critère d'évaluation l'erreur quadratique intégrée *ISE*. Les conclusions tirées des résultats indiquent que les estimateurs récursifs sont supérieurs aux estimateurs non récursifs en termes de performance pour les échantillons de grandes tailles.

Mots clés : Estimation non paramétrique, méthode du noyau, paramètre de lissage, densité de probabilité, estimation récursifs, erreur quadratique intégrée moyenne.

Abstract

In this work, we investigated the probability density function using both recursive and non-recursive estimators through the kernel method for independent data. We also presented the statistical properties of both types of estimators for symmetric and asymmetric cases (Expectation, bias, variance, mean squared error, and integrated mean squared error). The quality of estimation depends on the choice of smoothing parameter h and kernel K, for which we used modified Gaussian and gamma kernels in our estimators. The classical plugin method (rule of thumb) was used to select the optimal smoothing parameter h. Subsequently, we evaluated the performance of our estimators through simulations and real data, using integrated squared error (ISE) as the evaluation criterion. The conclusions drawn from the results indicate that recursive estimators outperform non-recursive estimators in terms of performance for large sample sizes.

Key words : Non parametric estimation, kernel method, smoothing parameter, probability density, recursive estimation, integrated mean squared error.