République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université A. Mira de Béjaïa

Faculté des Sciences Exactes

Département de Recherche Opérationnelle

Mémoire de Master

en

Recherche Opérationnelle

Option : Modélisation Mathématique et Evaluation de Performance des Réseaux

Thème

Estimation de la fonction de densité discrète par la méthode du noyau associé discret

Présenté par :

M^r **MENACEUR Oussama** M^r **AMARA Abdeljalil**

Devant le Jury :

Rapporteur	M^r	**N.ZOUGAB**	**Professeur**	**Université de Béjaia**
Co-Rapporteur	M^{me}	**L.DJERROUD**	**M.C.B**	**Université de Béjaia**
Examinatrice	M^{me}	**Y.ZIANE**	**M.C.A**	**Université de Béjaia**
Examinatrice	M^{me}	**L.HARFOUCHE**	**Docteur**	**Université de Béjaia**

Béjaia, 2021.

Remerciements

Nous remercions le Dieu pour le courage, la patience et la volonté qui nous ont été utiles tout au long de notre parcours.

Nous tenons à remercier Mr ZOUGAB Nabil et Meme DJERROUD Lamia pour la proposition du thème, l'encadrement de ce travail, pour leurs précieux conseils et orientations.

Nous remercions également les membres du jury pour avoir accepté d'examiner et d'évaluer notre travail.

Nos sincères remerciements s'adressent enfin à tous ceux qui nous ont soutenu de près ou de loin.

Nous remercions également tous les enseignants du département de Recherche Opérationnelle.

Dedicaces

Du profond de mon cœur, je dédie ce travail à tous ceux qui me sont chers.

À LA MÉMOIRE DE MON PERE LAICHE.

Ce travail est dédié à mon père (RAHIMAHO ALLAH), décédé trop tôt, qui m'a toujours poussé et motivé dans mes études et dans toute ma vie. Que Dieu lui fasse miséricorde et lui accorde le plus au paradis

À MA CHÈRE MÈRE

Aucune dédicace ne saurait exprimer mon respect, mon amour éternel et ma considération pour les sacrifices que vous avez consenti pour mon instruction et mon bien-être. Je vous remercie pour tout le soutien et l'amour que vous me portez depuis mon enfance et Puisse Dieu, vous accorder santé, bonheur et longue vie.

À MON FRÈRE ABdELHAQ

Je ne saurai traduire sur du papier l'affection que j'ai pour toi, je n'oublierai jamais ces merveilleux moments passés ensemble. Tu es le plus intelligent que j'aie

jamais connu.

À mes chères sœurs

Vous êtes l'une des bénédictions de Dieu. Je vous ai toujours trouvés à mes côtés. La vie sans vous est insipide. YA ALLAh, rends mes sœurs heureuses et rends-les toujours bonnes.

À TOUTE MA FAMILLE

À mes chèrs amis

En souvenir des moments heureux passés ensemble, avec mes vœux sincères de réussite, bonheur, santé et de prospérité. Particulièrement à mon binôme JALIL.

sincèrement :

Merci !

Ossama MENACEUR.

Dedicaces

Celle qui m'a transmis la vie, l'amour, le courage, à toi chère maman toutes mes joies, mon amour et ma reconnaissance.

À mon père pour l'éducation qu'il m'a prodigué ; avec tous les moyens et au prix de toutes les sacrifices qu'il a consentis à mon égard et à mes études depuis mon enfance.

À mes chers frères Abderrahmen, Mohammed et Adel
À ma chère soeur chaimaa

À mes amis Bouzid, Khayer, Abdelhak et Djalil
À ma petite amie Lamia
À mes amies

À mes condisciples et surtout à Oussama
À toutes ces personnes,
sincèrement :
Merci !
Jalil AMARA

Table des matières

Liste des tableaux

Table des figures

Introduction générale

L'estimation d'une densité f ou d'une fonction de répartition F à partir d'un échantillon de variables aléatoires réelles indépendantes et de loi inconnue tient une place importante dans l'étude de nombreux phénomènes de nature aléatoire. Car elle nous donne un estimateur convenable permet de décrire la loi de probabilité des observations et résoudre beaucoup des problèmes statistiques. La densité de probabilité à plusieurs avantages, en citant l'aperçu plus visible des caractéristiques principales de la distribution mieux que la fonction de répartition, en plus est beaucoup plus facile à interpréter que la fonction de répartition.

On trouve dans la littérature deux types d'approches d'estimation : l'approche paramétrique et l'approche non-paramétrique. L'approche paramétrique consiste à supposer que f appartient à une famille de densités continues ou discrètes qui peuvent être décrites par un certain nombre de paramètres réels. Pour pallier les insuffisances et les défauts de cette première approche, on fait appel à l'approche non-paramétrique, qui permet d'estimer la densité de probabilité directement à partir de l'information disponible sur l'ensemble d'observations. On dit souvent que dans cette approche les données parlent d'elles mêmes. Nous nous intéressons dans ce mémoire à l'approches non-paramétriques par noyau, en utilisant la fonction noyau K et le paramètre de lissage h. L'estimateur à noyau a été proposé initialement par Rosenblatt [1956] et Parzen [1962] pour estimer la fonction de densité f à support non borné. Le choix du noyau dans ce cas est peu important, car il influe peu sur l'estimation. Les noyaux les plus utilisés pour les densités à support non borné sont les noyaux symétriques dits aussi classiques : noyau gaussien, noyau Epanechnicov (Epanechnikov [1969a]), noyau triangulaire, noyau biweight et noyau uniforme, etc.

Pour estimer une fonction de densité discrète (dite généralement, fonction de masse de probabilité (fmp)) par l'approche non-paramètrique, l'estimateur empirique appelé aussi estimateur à noyau du type Dirac est souvent utilisé. Deux autre classes de noyaux discrets : les noyaux discrets standards (Noyau Binomial, Binomial négatif et Poisson) et les noyaux associés discrets de deuxième ordre (Dirac uniforme et Triangulaires discrets), voir par exemple Kokonendji et al. [2007b], Kiesse [2008] et Aitchison and Aitken [1976]. Ces derniers ont introduit

la notion de l'estimateur à noyau associé, en donnant la définition d'un noyau associé $K_{x,h}$ de cible x et de paramètre de lissage h à partir d'une loi de probabilité discrète.

Ce mémoire est composée d'une introduction générale, de quatre chapitres, d'une conclusion générale et d'une bibliographie. Le chapitre 1, comportera une présentation de l'approche non-paramétrique par l'estimateur de noyau dans le cas univarié et les noyaux symétriques usuels. Le deuxième chapitre est consacré à la présentation de l'estimation non-paramétrique de la fonction de masse de probabilitée par le noyau associé discret ainsi que ses propriétés statistiques. Ensuite, on cite les différents types de noyaux et les différents méthodes d'estimation du paramètre de lissage h. Dans le troisième chapitre, nous présentons une étude de simulation pour évaluer l'utilisation de différents noyaux discrets du premier et deuxième ordre de l'estimation non paramétrique de la fonction de masse de probabilitée f selon le critére *ISE* sur des fonctions discretes, ainsi que la méthode validation croisée utilisée pour estimer le paramètre de lissage h. Tous les résultats numériques et graphiques sont effectués à l'aide du logiciel R. Et pour Le dernier chapitre, trois applications numériques sont réalisée sur des données réelles où on a choisi le domaine de finance. Ce travail se termine par une conclusion générale, une bibliographie et une annexe dans laquelle nous présentons les différents programmes informatiques mis en place sous R, qui permettent d'estimer la densité de probabilité.

Chapitre 1

Estimation non-paramétrique de la fonction de densité par la méthode du noyau

Introduction

L'approche non paramétrique par noyau prend son sens lorsqu'on ne possède aucune information précise sur la forme et la classe de la vraie densité. Dans cette approche, ce sont les observations qui vont nous permettre de déterminer l'estimation de la densité f.

L'estimateur à noyau a été proposé par Rosenblatt [1956] et Parzen [1962] pour estimer des fonctions de densités, il est en fonction d'un paramètre appelé paramètre de lissage ou fenêtre et d'un noyau K.

Dans ce chapitre, nous présentons l'approche non paramétrique par l'estimateur de noyau dans le cas univarié, en utilisant les noyaux symétriques dits aussi classiques, par example : noyau gaussien, noyau Epanechnikov [1969a], noyau triangulaire, noyau biweight et noyau uniforme. Ensuite nous présentons également les différentes propriétés relatives à cet estimateur : biais, variance, l'erreur quadratique moyenne (MSE), et l'erreur quadratique moyenne intégrée (MISE). Nous rappelons par la suite les méthodes du choix du paramètre de lissage : "Plug-in", Validation croisée par moindres carrés et Validation croisée par maximum de vraisemblance.

1.1 Notion de noyaux

L'estimation par noyau est une méthode non paramétrique d'estimation de la densité d'une variable aléatoire. Cette méthode permet d'obtenir une densité continue. En effet, la fonction indicatrice utilisée pour l'histogramme est remplacée par une fonction continue (noyau).

En 1962, Parzen [1962] a étudié les propriétés fondamentales de l'estimateur à noyau de la densité, juste aprés son introduction par Rosenblatt [1956]. A partir de ce moment, cet estimateur à noyau de la densité est devenu un objet classique étudié par les statisticiens. L'estimateur de la densité de probabilité par la méthode du noyau est le plus utilisé, car il possède de bonnes propriétés. L'idée consiste à évaluer la densité f au point x en comptant le nombre d'observations tombées dans un certain voisinage de x sur $\mathbb{R}$.

Supposons que l'on dispose d'un échantillon d'observations $X_1..., X_n$, issu d'une $v.a.X$ poss édant pour fonction de densité la fonction f que l'on désire estimer. L'estimateur à noyau continu symétrique de la fonction f est défini par Parzen [1962] :

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - X_i}{h}), \tag{1.1}$$

où K est la fonction noyau telle que $K(t) \geq 0$ et $\int_{\mathbb{R}} K(t)dt = 1$ et $h > 0$ est le paramètre de lissage. Dans l'expression de l'estimateur à noyau continu (1.1), la fonction noyau K est une densité de probabilité sur $\mathbb{R} \to \mathbb{R}^+$ est symetrique par rapport a zero :

$$K(-u) = K(u),$$

ce qui implique l'égalité suivante

$$\int_{\mathbb{R}} uK(u)du = 0$$

1.2 Exemples de noyaux

Voici quelques exemples de noyaux les plus communément utilisés :

Noyau	Fonction noyau	Domaine de définition
Cauchy	$K(u) = [\pi(1 + u^2)]^{-1}$	$\mathbb{R}$
Biweight	$K(u) = (15/16) * (1 - u^2)^2$	[-1, 1]
Triangulaire continu	$K(u) = 1 - \|u\|$	[-1, 1]
Epanechnikov	$K(u) = (3/4) * (1 - u^2)$	[-1, 1]
Gaussien	$K(u) = (1/\sqrt{2\pi}) * exp(u^2/2)$	$\mathbb{R}$

TABLE 1.1 – Exemples de noyaux continues symétriques.

Pour plus de détails sur les types des noyaux, on peut se référer à Epanechnikov [1969a] et Tsybakov [2003]. L'expression de K détermine la forme du noyau et h est un paramètre qui détermine le niveau de lissage de l'estimation. Dans l'estimation à noyau continu symétrique, le choix du paramètre de lissage est prépondérant par rapport à celui du noyau K.

1.3 Propriétés de l'estimateur à noyau

Dans cette partie, nous allons maintenant donner les propriétés fondamentales de l'estimateur et les critères d'erreurs usuels. Nous présentons d'abord le biais et la variance de l'estimateur $\hat{f}_h(x)$. Ensuite, nous exprimons le risque quadratique exact en un point x fixé, puis le risque intégré. Enfin les convergences uniforme et en loi.

1.3.1 Espérance, Biais et Variance de l'estimateur

L'espérance mathématique de $f_h(x)$ est :

$$\begin{aligned}\mathbf{E}f_h(x) &= \frac{1}{nh}\mathbf{E}\sum_{i=1}^{n}K(\frac{x-X_i}{h}) \\ &= \frac{1}{h}\int_{-\infty}^{\infty}K(\frac{x-u}{h})f(u)du\end{aligned}$$

En posant $t=\frac{x-u}{h} \Longrightarrow dt=-\frac{du}{h}$

$$\mathbf{E}f_h(x)=\int_{-\infty}^{\infty}K(t)f(x-ht)dt \tag{1.2}$$

•Le biais de $f_h(x)$:

$$\mathbf{Biais}f_h(x)=\mathbf{E}(f_h(x)-f(x))=\int_{-\infty}^{\infty}K(t)f(x-ht)dt-f(x) \tag{1.3}$$

•La variance de $f_h(x)$ est :

$$\begin{aligned}\mathbf{Var}f_h(x) &= \mathbf{var}\sum_{i=1}^{n}\frac{1}{nh}K(\frac{x-X_i}{h}) \\ &= \frac{1}{n^2h^2}\sum_{i=1}^{n}\mathbf{var}K(\frac{x-X_i}{h}) \\ &= \frac{1}{n^2h^2}\sum_{i=1}^{n}\Big[\mathbf{E}\Big(K(\frac{x-X_i}{h})\Big)^2\Big]-\frac{1}{n^2h^2}\sum_{i=1}^{n}\Big[\mathbf{E}K(\frac{x-X_i}{h})\Big]^2 \\ &= \frac{1}{nh^2}\int_{-\infty}^{\infty}\Big[K(\frac{x-t}{h})\Big]^2f(t)dt-\frac{1}{n}\Big(\frac{1}{h}\int_{-\infty}^{\infty}K(\frac{x-t}{h})f(t)dt\Big)^2\end{aligned}$$

Avec le changement de variable, $t = \frac{x-u}{h}$, on obtient :

$$\mathbf{Var} f_h(x) = \frac{1}{nh}\int_{-\infty}^{\infty}(K(t))^2 f(x-ht)dt - \frac{1}{n}\Big(\int_{-\infty}^{\infty} K(t)f(x-ht)dt\Big)^2 \tag{1.4}$$

En faisant le développement de Taylor à l'ordre *2* au point *t=0* de *f(x-ht)*. On obtient :

$$f(x-ht) = f(x) - \frac{ht}{1!}f'(x) + \frac{h^2t^2}{2!}f"(x) + o(h^2).$$

Si le noyau K est une fonction symétrique par rapport à 0 c'est-à-dire :

$$\int_{-\infty}^{\infty} tK(t)dt = 0 \quad et \quad \int_{-\infty}^{\infty} t^2K(t)dt < \infty$$

Alors les expressions finales sont données par :

$$\mathbf{E} f_h(x) = f(x) + \frac{h^2}{2}f"(x)\int_{-\infty}^{\infty} t^2K(t)dt + o(h^2). \tag{1.5}$$

$$\mathbf{Biais} f_h(x) = \frac{h^2}{2}f"(x)\int_{-\infty}^{\infty} t^2K(t)dt + o(h^2). \tag{1.6}$$

$$\mathbf{Var} f_h(x) = \frac{1}{nh}f(x)\int_{-\infty}^{\infty} K^2(t)dt + o(\frac{1}{nh}). \tag{1.7}$$

Discussion du comportement du biais et de la variance

- Le biais décroit si h diminue mais la variance augmente.
- La variance diminue si h augmente mais le biais augmente.
- Pour que la variance tende vers zéro, il faut que $nh \to \infty$.
- Plus la courbure de la densité est haute en x, plus le biais est grand.
- La variance est plus grande pour des valeurs plus grandes de la densité.

1.3.2 Convergence de l'estimateur à noyau

Dans ce qui suit, nous allons énoncer quelques résultats qui nous indiquent les diférents types de convergence de l'estimateur à noyau.

Erreur quadratique moyenne (Mean-Squared Error :MSE)

L'analyse de la performance de l'estimateur à noyau exige la spécification d'un critère d'erreur approprié afin de mesurer l'erreur d'estimation aussi bien qu'en un point que sur l'ensemble des points. Nous étudierons dans un premier temps la proximité de notre estimateur de la densité $\hat{f}_h$ de la vraie densité *f*. L'estimateur $\hat{f}_h$ dépend des données, du noyau K et du paraètre de lissage h. Cette dépendance n'est généralement pas exprimée explicitement. Pour chaque x,

$\hat{f}_h(x)$ peut être considérée comme une variable aléatoire. Lorsque nous considérons l'estimation en un point, En utilisant les expressions finales des deux termes : le biais et la variance, l'erreur quadratique moyenne *(MSE)* en un point x est donnée par

$$MSE(f_h(x)) = \textbf{Biais}^2(\hat{f}_h(x)) + \textbf{Var}(\hat{f}_h(x))$$

En remplaçant le biais (1.6) et la variance (1.7) par leurs valeurs respectives, on trouve :

$$MSE(f_h(x)) = \frac{1}{nh} f(x) \int K^2(t)dt + \frac{h^4}{4}\{f"(x)\}^2\Big[\int t^2K(t)dt\Big]^2 + o(\frac{1}{nh} + h^4). \tag{1.8}$$

Theorem 1.3.1 (Parzen ((1962))). *Si,* $\lim\limits_{n\to\infty} h = 0$ *et* $\lim\limits_{n\to\infty} nh = \infty$ *et* K *satisfait aux conditions suivantes :*

$$\begin{array}{ll} \text{-}sup K(y) < \infty & et \quad \lim\limits_{y\to\infty} \mid yK(y) \mid = 0, \\ \text{-}\int_{-\infty}^{\infty} \mid K(y) \mid dy < \infty & et \quad \int_{-\infty}^{\infty} K(y)dy = 1, \end{array}$$

Alors l'estimateur $\hat{f}_h(x)$ *est consistant en moyenne quadratique en tout point* x *pour lequel la densité* f *est continue, c'est à dire :*

$$\lim_{n\to\infty} MSE\hat{f}_h(x) = 0.$$

Erreur quadratique moyenne intégrée (Mean-Integrated Squared Error :MISE)

Cependant, Il peut être intéeressant d'avoir une mesure globale de la précision de $\hat{f}_h$ comme estimateur de f au lieu d'avoir une mesure de précision en un point donné. Cette mesure globale qu'on note *MISE* est définie par :

$$MISE(f_h) = \int_{-\infty}^{\infty} MSE(f_h)dx$$

En remplaçant le *MSE* 1.8 par leur valeur , on trouve :

$$MISE(f_h) = \frac{1}{nh}\int_{-\infty}^{\infty} K^2(t)dt + \frac{h^4}{4}\int_{-\infty}^{\infty}\{f"(x)\}^2\Big[\int t^2K(t)dt\Big]^2 + o(\frac{1}{nh} + h^4).$$

Theorem 1.3.2 (De Oliveira ((1963))). *Si* K *est un noyau de Parzen-Rosenblatt, c'est-à-dire 1.3.1* K *verifé :*

- $\int_{\mathbb{R}} K(x)dx = 1.$
- $\int_{\mathbb{R}} \mid K(x) \mid dx < \infty.$
- $sup \parallel K(x) \parallel dx < \infty.$

- $\lim_{|x|} K(x) = 0$.

Et $\lim_{n\to\infty} h = 0$ *et* $\lim_{n\to\infty} nh = \infty$, *alors*

$$\forall f \in L^p, \lim_{n\to\infty} MISE\hat{f}_h = 0$$

où L^p *est l'ensemble des fonctions reelles de puissance* p^{ieme} *integrable, c'est à dire* $\int \mid f(x) \mid^p < \infty$.

Convergence uniforme en probabilité et presque complète

La convergence uniforme en probabilité et la convergence presque complète ont été obtenues respectivement par Parzen[1962] et Nadaraya[1965].

Theorem 1.3.3 (Parzen ((1962))). *Si* $\lim_{n\to\infty} nh^2 = 0$, *si le noyau* K *satisfait les conditions du théorème 1 et si la transformé de Fourier*$\tau F(z) = \int exp(-izy)K(y)dy$ *est absolument intégrable, alors* $\hat{f}_h(x)$ *est un estimateur uniformément consistant en probabilité, c'est à dire*

$$\forall \varsigma > 0, \lim_{n\to\infty} pr(sup_x \mid \hat{f}_h(x) - f(x) \mid < \varsigma) = 1.$$

Theorem 1.3.4 (Nadaraya ((1965))). *Si* K *est un noyau positif à variation bornée et* f *est uniformément continue, si* $\lim_{n\to\infty} h = 0$ *et* $\sum_{k=1}^{\infty} exp(-\nu nh^2 < \infty), \forall \nu$, *alors*

$$Pr(sup_x \mid \hat{f}_h(x) - f(x) \mid = 0) = 1$$

Silverman [1986] a donné le même resultat sur la convergence presque complete en remplacant la condition $\sum_{k=1}^{\infty} exp(-\nu nh^2 < \infty, \forall \nu$, *par les deux conditions suivantes*

$$\lim_{n\to\infty} h = 0 \quad et \quad \lim_{n\to\infty} \frac{\log n}{nh} = 0.$$

Convergence en loi

La convergence en loi a été établie par Parzen [1962].

Theorem 1.3.5. *Si* $\lim_{n\to\infty} h = 0, \lim_{n\to\infty} nh = \infty$ *et* K *satisfait les conditions du théorème 1, alors* $\hat{f}_h(x)$ *est un estimateur asymptotiquement normal en tout point* x *pour lequel la densité* f *est continue, c'est à dire*

$$\frac{\hat{f}_h(x) - \boldsymbol{E}(\hat{f}_h(x))}{\sqrt{\boldsymbol{Var}(\hat{f}_h(x))}} \underrightarrow{\mathcal{L}} \mathcal{N}(0,1),$$

où $\underrightarrow{\mathcal{L}}$ *désigne la convergence en loi et* $\mathcal{N}(0,1)$ *et la loi normale standarde.*

1.4 Choix du noyau

Le premier choix porte sur la nature de la densité noyau que nous utilisons. Pour mesurer l'efficacité de chacun des noyaux continus symétriques présentés dans le tableau 1, nous utilisons une mesure commune qui consiste à calculer le rapport du critère *AMISE* (*MISE* asymptotique) des deux noyaux mis en évidence donné par

$$AMISE = \frac{h^4\sigma_K^4}{4}\int f"(x)^2dx + \frac{\int K^2(y)dy}{nh}, \tag{1.9}$$

où $\sigma_K^4 = \left(\int y^2K(y)dy\right)^2$

Pour minimiser *AMISE*, il suffit de minimiser l'expression $\int K^2(y)dy$ sur

$$\mathcal{K} = \left(K_0 : K_0 \geq 0, K_0(y) = K_0(-y), \int K_0(y)dy = 1et \int y^2K_0(y) < \infty\right)$$

Theorem 1.4.1 (Epanechnikov ((1969a))). *Si* $\lim_{n\to\infty} h = 0$, $\lim_{n\to\infty} nh = \infty$, $f \in \mathbf{L}^2$, $\int f"(x)^2dx \neq 0$ *et* $\int f(x)^2dx < \infty$,
alors le noyau d'Epanechnikov K_E *défini par*

$$K_E(u) = \frac{3}{4}(1-u)^2\mathbf{1}_{[-1,1]}$$

est de AMISE *minimum.*

Pour le noyau d'Epanechnikov, la valeur minimale atteinte est $3/(5\sqrt{5})$. *L'efficacite relative d'un noyau* K *se mesure alors en prenant*

$$\begin{aligned} Eff(K) &= \frac{\int K^2(u)du}{\int K_E^2(u)du} \\ &= \frac{\int K^2(u)du}{\frac{3}{5\sqrt{5}}} \end{aligned}$$

$$Eff(K) = \frac{\int K^2(u)du}{\int K_E^2(u)du} = \frac{\int K^2(u)du}{\frac{3}{5\sqrt{5}}}$$

Le choix de K dépend seulement de la nature de f et nous admettons qu'en pratique le choix du noyau d'Epanechnikov est le plus staisfaisant. Nous donnons la table récapitulatif 1.2 qui présente la valeur d'efficacité des différents noyaux continus symétriques.

Noyau	Efficacité
Epanechnikov	1.0000
Uniforme	1.0758
Triangulaire	1.0143
Gaussien	1.0513
Biweight	1.0061

TABLE 1.2 – Efficacité des noyaux continus symétriques.

1.5 Choix du paramètre de lissage

D'aprés la formule (1.1) on constate que l'estimateur $\hat{f}_h(x)$ de *f(x)* ne dépend pas seulement du noyau K mais aussi du paramètre h, appelé paramètre de lissage ou fenêtre (bandwidth or window). Une petite perturbation de ce dernier est suffisante pour que $\hat{f}_h(x)$ change complètement ses caractéristiques, ce qui signifie $\hat{f}_h(x)$ est fortement lié à ce paramètre. C'est pour cette raison que plusieurs travaux ont été consacrés au choix de ce paramétre.

Il existe plusieurs méthodes de sélection de ce paramètre. Dans cette section on va présenter trois méthodes pour la détermination du paramètre de lissage optimal h_{opt} pour approcher la valeur idéale de la fenêtre h.

La valeur idéale h_{id} du paramètre h qui minimise l'erreur quadratique moyenne intégrée (*MISE*) a été obtenue par Parzen [1962]. Pour une taille d'échantillon n donnée et un noyau K fixé nous avons

$$\frac{\partial AMISE(h)}{\partial h} = 0$$

Ce qui est équivalent à

$$h^5 = \frac{\int_{\mathbb{R}} K(t)^2 dt}{n\mathbf{Var}(K^2) \int_{\mathbb{R}} f"(x)^2 dx}$$

alors la valeur idéale de la fenêtre h définie par

$$h_{id} = \frac{1}{\sqrt[5]{n}} \left\{ \frac{\int_{\mathbb{R}} K(t)^2 dt}{\mathbf{Var}(K^2) \int_{\mathbb{R}} f"(x)^2 dx} \right\}^{1/5}. \tag{1.10}$$

On remarque que le h_{id} dépend du noyau K, de la variance de K et de la deuxième dérivée de f, ce qui la rend impossible à son utilisation dans la pratique.

Méthode Rule of thumb [Silverman ((1986))]

Dans la procedure de Plug-in, l'idee de base est d'estimer dans l'expression 1.10 la quantité inconnue : $\int_{\mathbb{R}} f"(x)^2 dx$. En effet, il y a deux approches possibles pour le faire : soit on suppose que la densité f appartient a une famille de distributions paramétriques et là on estime les paramètres et on retrouvera facilement cette quantité, soit on l'estime par l'approche non-paramétrique et donc faire appel à un estimateur a noyau (par exemple). Ceci va compliquer davantage les calculs parce que on trouvera une fonction qui dépend elle même de h. Donc, en gros, la méthode Plug-in consiste a "injecter" une estimation de f en adoptant une méthode commode et pratique.

Méthode de validation croisée par moindres carrés

Cette méthode a été introduite par Bowman [1984], pour un noyau fixé K, le principe de la validation croisée est la minimisation d'estimateur de risque intégré (*MISE*) par rapport à h. En effet, Le *MISE* dépend de la fonction inconnue f et ne peut donc pas être calculé. On remplace la fonction *MISE* par une fonction de h, mesurable par rapport à l'échantillon et dont la valeur pour chaque $h > 0$ est un estimateur sans biais de MISE(h).

$$\begin{aligned} MISE(h) &= \mathbf{E}\int_{-\infty}^{\infty}\{\hat{f}_h(x) - f(x)\}^2 dx \\ &= \mathbf{E}\int_{-\infty}^{\infty}\hat{f}_h^2(x)dx - 2\mathbf{E}\int_{-\infty}^{\infty}\hat{f}_h(x)f(x)dx + \int_{-\infty}^{\infty}f^2(x)dx. \end{aligned}$$

Le dernier terme ne dépend pas de h, pour minimiser *MISE(h)* il suffit de minimiser l'expression :

$$w(h) = \mathbf{E}\int_{-\infty}^{\infty}\hat{f}_h^2(x)dx - 2\mathbf{E}\int_{-\infty}^{\infty}\hat{f}_h(x)f(x)dx.$$

Le deuxième terme de *w(h)* dépend de f qui peut être estimer par :

$$\hat{W} = \frac{1}{n}\sum_{i=1}^{n}\hat{f}_{h,-i}(X_i),$$

avec

$$\hat{f}_{h,-i}(X_i) = \frac{1}{n-1}\frac{1}{h}\sum_{j=1,n,j\neq i}K(\frac{X_i - X_j}{h})$$

Finalement, l'estimateur sans biais de w(h) est donné par

$$CV(h) = \int_{-\infty}^{\infty}\hat{f}_h^2(x)dx - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{h,-i}(X_i)$$

Et la fenêtre optimale est telle que :

$$h_{CV} = \arg\min_{h>0} CV(h).$$

Méthode de validation croisée par maximum de vraisemblance

La distance de Kullback-Leibler est une distance entropique qui mesure la différence entre deux densités de probabilité. Dans ce cas, la distance entre la densité à estimer f et l'estimateur à noyau $\hat{f}_h(x)$ s'écrit :

$$\begin{aligned} D(f, \hat{f}_h(x)) &= \int_{\mathbb{R}} f(x) \log\Big\{\frac{f(x)}{\hat{f}_h(x)}\Big\} dx \\ &= \int_{\mathbb{R}} f(x) \log\{f(x)\} dx - \int_{\mathbb{R}} f(x) \log\{\hat{f}_h(x)\} dx \\ &= \mathbf{E}\{\log\{f(X)\}\} - \mathbf{E}\Big\{\log\{\hat{f}_h(X)\}\Big\}. \end{aligned}$$

L'idée de la validation croisée par vraisemblance est de minimiser $D(f, \hat{f}_h(x))$. Toutefois, cette distance n'est pas métrique et les critères définis en la minimisant ne sont pas appropriés pour obtenir un lissage adéquat. Donc, minimiser $D(f, \hat{f}_h(x))$ revient à maximiser $\boldsymbol{E}\Big\{\log\{\hat{f}_h(X)\}\Big\}$. Ainsi, la fenêtre optimale est (Voir Habbema et al. ((1974)) et Duin ((1976))).

$$h_{LCV} = \arg\max_{h>0} LCV(h),$$

où

$$LCV(h) = \mathbf{E}\Big\{\log\{\hat{f}_h(X_i)\}\Big\}$$

L'estimateur sans biais est

$$w(h) = \frac{1}{n}\sum_{i=1}^{n,-i} \log\{\hat{f}_{h,-i}(X_i \mid h)\},$$

où

$$\hat{f}_{h,-i}(X_i) = \frac{1}{n-1}\frac{1}{h}\sum_{j=1,n,j\neq i} K(\frac{X_i - X_j}{h})$$

Enfin, la fenêtre optimale obtenue par la méthode de validation croisée par vraisemblance se calcule à partir de :

$$h_{LCV} = \arg\max_{h>0} \Big\{\frac{1}{n-1}\frac{1}{h}\sum_{j=1,n,j\neq i} K(\frac{X_i - X_j}{h})\Big\}$$

Cependant, cet estimateur est très sensible aux valeurs aberrantes. Sa difficulté apparaît lorsque la méthode est appliquée à des observations dont la distribution présente de grandes queues. Les points situés dans les queues de la distribution à estimer ont des valeurs faibles, ce qui implique de faibles valeurs des estimations correspondantes. La présence de l'opérateur log dans l'expression de l'estimateur pose un problème de convergence pour les valeurs de densités

aux queues. Par conséquent, il est difficile dans ce cas de choisir h_{LCV} de façon optimale, puisque l'on risque soit le sur-lissage soit une trop grande erreur sur les queues.

1.6 Conclusion

Dans ce chapitre, nous avons rappelé la notation de l'estimateur à noyau symétrique (cas :univarié). Ainsi nous avons présenté les propriétés statistiques (biais, variance, erreur quadratique moyenne et erreur quadratique moyenne intégrée). Et pour le choix de paramètre de lissage on a défini trois méthodes classiques (Plug-in, Validation croisée par moindres carrés et Validation croisée par maximum de vraisemblance).

Chapitre 2

Estimation de la fonction de densité discrète par noyau associé discret

Introduction

Soit $X_1, X_2, ..., X_n$ un n-échantillon aléatoire indépendant et identiquement distribué issu d'une variable aléatoire X de fonction de densité discrète inconnue f sur $\aleph \subseteq \mathbb{R}$. L'estimateur à noyau associé $\hat{f}_h$ de f utilisant $K_{x,h}$ est défini par :

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i), \qquad x \in \aleph, \tag{2.1}$$

où $K_{x,h}$ est le noyau associé discret de cible x et de fenêtre h sur $\aleph_x$.

Ce chapitre est consacrée à la notion du noyau associé discret $K_{x,h}$ de cible x, et du paramètre de lissage h. Cette notion a été introduite par Kokonendji et al. [2007b] et Kiesse [2008] dans le cas discret. Nous présentons ensuite la définition de estimateur à noyau discret $\hat{f}$ et on cite des exemples des noyaux discrets standards (Poisson, Binomial et Binomial négatif) et le deuxième cas de deuxième ordre(type Dirac, Dirac uniforme,Triangulaire). Nous présontons les comportements du biais, de la variance et par conséquent du risque quadratique de $\hat{f}$. Puis, on y donne les propriétés globales de $\hat{f}$. Enfin, on se concentrerait sur le choix du paramètre de lissage h (la méthode de validation croisée) dans l'estimation de la fonction de densité f par la méthode des noyaux associés discrets.

2.1 Noyaux associé discrets

Les définitions suivantes présentent les notions du noyau associé discret, et de l'estimateur à noyau associé discret pour la fonction de densité f inconnue sur le support $\aleph$.

Définition 1. [Kiesse ((2008))]

Soit $\aleph_x$ le support d'une fonction de masse de probabilitée f à estimer. Etant donnée $x \in \aleph$ et $h > 0$, on appelle "noyau associée discret" $K_{x,h}(.)$ toute fonction de masse de probabilitée liée à la variable aléatoire discrète $\mathcal{K}_{x,h}$ de support $\aleph_x$ contenant au moins x et indépendant de h, vérifiant les quatre conditions :

$$\bigcup_x \aleph_x \supseteq \aleph \tag{2.2}$$

$$\mathbf{E}(\mathcal{K}_{x,h}) \sim x \quad lorsque \quad h \to 0 \tag{2.3}$$

$$\mathbf{Var}(\mathcal{K}_{x,h}) < +\infty \tag{2.4}$$

$$\mathbf{Var}(\mathcal{K}_{x,h}) \to 0 \quad lorsque \quad h \to 0 \tag{2.5}$$

Définition 2. [Kokonendji et al. ((2007b))]

Soit $x \in \mathbb{N}$ et $h > 0$. Etant donnée une loi de probabilité discrète paramétrique K_θ, $\theta \in \Theta \subset \mathbb{R}^d$, de support $\aleph_\theta \subseteq \mathbb{N}$. On appelle " noyau discret associé " $K_{x,h}$ à K_θ, de cible x et de paramètre de lissage discret h, s'il existe une correspondance entre θ et (x,h) telle que $K_{x,h}$ est une loi de probabilité sur le support $\aleph_{x,h}$ de même famille que K_θ :

$$K_{x,h} \geq 0 \quad et \quad \sum_{y \in \aleph_{x,h}} K_{x,h}(y) = 1 \tag{2.6}$$

avec

$$\sum_{y \in \aleph_{x,h}} y K_{x,h}(y) \sim x \quad lorsque \quad h \to 0. \tag{2.7}$$

Remarque 1. La relation (2.7) permet d'assurer la convergence ponctuelle de l'estimateur à noyau discret. Elle traduit la prise en compte d'un maximum d'information autour de la cible et dans son entourage immédiat, de telle sorte qu' asymptotiquement nous retrouvons l'estimateur naïf. On peut remplacer (2.7) par :

$$\sum_{y \in \aleph_{x,h}} y \mathcal{K}_{x,h}(y) = x + h + o(h). \tag{2.8}$$

La condition (2.7) (ou (2.8)) est fondamentale pour l'estimateur à noyau discret. La qualité de lissage obtenue en appliquant ces noyaux discrets associés change selon le comportement de leur variance par rapport à la cible x. Ce qui nous amène à distinguer trois types de noyaux discrets associés : sousdispersés $(\boldsymbol{var}[\mathcal{K}_{x,h}] < \boldsymbol{E}(\mathcal{K}_{x,h}))$, equidispersés $(\boldsymbol{var}[\mathcal{K}_{x,h}] = \boldsymbol{E}(\mathcal{K}_{x,h}))$ et surdispersés $(\boldsymbol{var}[\mathcal{K}_{x,h}] > \boldsymbol{E}(\mathcal{K}_{x,h}))$.

2.1.1 Noyaux associés discrets standards (première ordre)

Nous présentons dans cette section la première classe des noyaux associés discrets, dite, classe des noyaux discrets standards ou de premier ordre proposés par Kiesse [2008]. Ce type de noyaux ne vérifient pas la condition (2.5).

Nous présentons trois exemples de noyaux discrets standards.

• Noyau Poissonien

Pour un type de noyau Poissonnien $\mathcal{P}(\lambda)$, on considère le noyau discret associé $P_{x,h}$ de loi $\mathcal{P}(x,h)$ sur $\aleph_{x,h} = \mathbb{N}$ avec $x \in \mathbb{N}$ et $h > 0$, tels que :

$$P_{x,h}(y) = \frac{(x+h)^y e^{-(x+h)}}{y!}, y \in \mathbb{N}. \tag{2.9}$$

On signale que le noyau discret associé proposé dans Marsh and Mukhopadhyay [1999] interéchange x en y dans (2.9). Notons que pour une cible $x \in \mathbb{N}$ et pour tout $h > 0$, le noyau associé $P_{x,h}$ est de support $\mathbb{N}$, équidispersé de moyenne égale à la variance $x + h$, et de mode compris entre $x + h$ - *1* et $x + h$.

La figure suivante donne l'allure de noyau de Poisson de premier ordre pour x fixé $y = x = 5$ et $h > 0$.

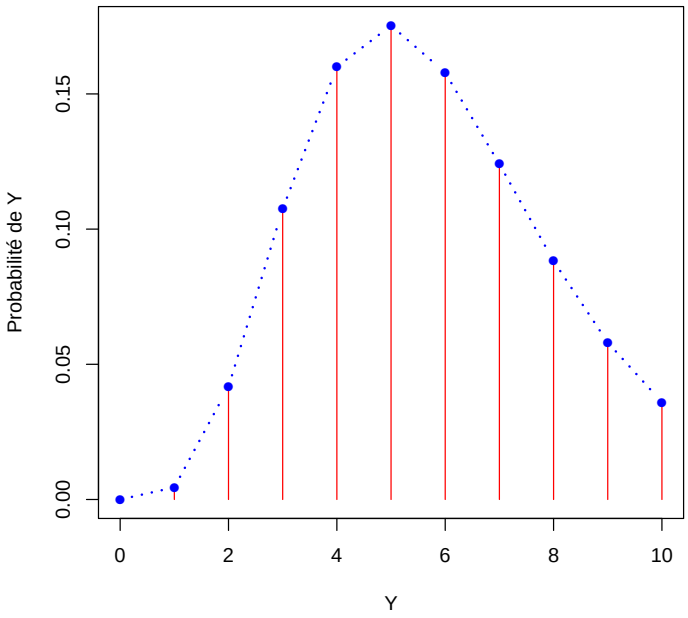

FIGURE 2.1 – Noyau de Poisson pour $y = 5$ et $h = 0.1$.

• **Noyau Binomial**

Si on considère un type de noyau Binomial $\mathcal{B}(N, p)$, on lui associe $B_{x,h}$ de loi $\mathcal{B}(x+1,(x+h)/(x+1))$ sur $\aleph_{x,h}$= { 0, 1, . . . , $x+1$} pour tout $x \in \mathbb{N}$ et $h \in]0, 1]$ avec $\cup_x \aleph_{x,h} = \mathbb{N}$, de telle sorte :

$$B_{x,h}(y) = \frac{(x+1)!}{y!(x+1-y)!}(\frac{x+h}{x+1})^y(\frac{1-h}{x+1})^{x+1-y}, \quad y \in \mathbb{N} \tag{2.10}$$

Ce noyau discret associé Binomial $B_{x,h}$ est à support {0, 1, . . . , $x+1$} (dépendant uniquement de x), sousdispersé de moyenne $x+h$ et de variance $(x+h)$ (1 - h) / $(x+1) < x+h$, et de mode autour de $x+h$.

La figure suivante donne l'allure de noyau Binomial de premier ordre pour x fixé $y = x = 5$ et $h > 0$.

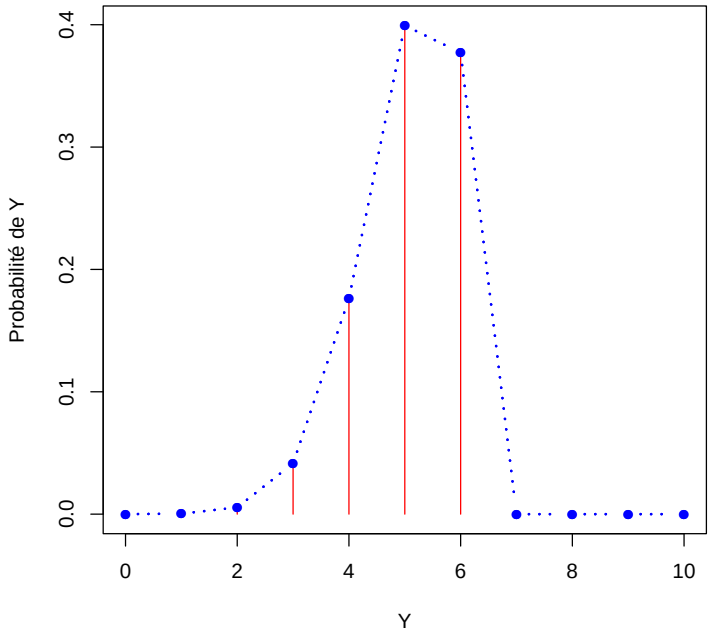

FIGURE 2.2 – Noyau Binomial pour $y = 5$ et $h = 0.1$.

• **Noyau Binomial Négatif**

Dans le cas d'un type de noyau Binomial Négatif $\mathcal{BN}(\lambda, p)$, on considère le noyau discret associé $BN_{x,h}$ de loi $\mathcal{BN}(x+1,(x+1)/(2x+1+h))$ sur $\aleph_{x,h} = \mathbb{N}$ pour tout $x \in \mathbb{N}$ et $h > 0$, tels que :

$$BN_{x,h}(y) = \frac{(x+y)!}{x!y!}\left(\frac{x+h}{2x+1+h}\right)^{y}\left(\frac{x+1}{2x+1+y}\right)^{x+1}, \quad y \in \mathbb{N}. \tag{2.11}$$

Ce noyau discret associé $BN_{x,h}$ est de support $\mathbb{N}$, surdispersé de moyenne $x+h$ et de variance $(x+h)[1+(x+h)/(x+1)] > x+h$, et de mode autour de $x+h$.

2.1.2 Noyaux associés discrets de deuxième ordre

Nous présentons ici la deuxième classe des noyaux associés discrets, dite, classe des noyaux associés discrets de deuxième ordre, c'est à dire, les condition (2.2) et (2.5) sont vérifiées.

Nous présentons trois exemples de noyaux de deuxième ordre.

• **Noyau de type Dirac**

Soit le noyau du type Dirac $D_{x,0}$ lié à la variable aléatoire $\mathcal{D}_{x,0}$ pour $x \in \mathbb{N}$ et $h = 0$ donné par

$$D_{x,0}(y) = \mathbf{1}_{y=x}\begin{cases} 1 & si\ y = x \\ 0 & sinon, \end{cases} \tag{2.12}$$

où $\boldsymbol{1}$ est la fonction indicatrice de A.

Le noyau du type Dirac vérifie les conditions (2.1)-(2.4), car $\aleph_x = \{x\}$, $\boldsymbol{E}(\mathcal{D}_{x,0}) = x$ et $\boldsymbol{Var}(\mathcal{D}_{x,0}) = 0$.

La figure suivante donne l'allure de noyau de type Dirac de deuxième ordre pour x fixé y=x=5 et h = 0.

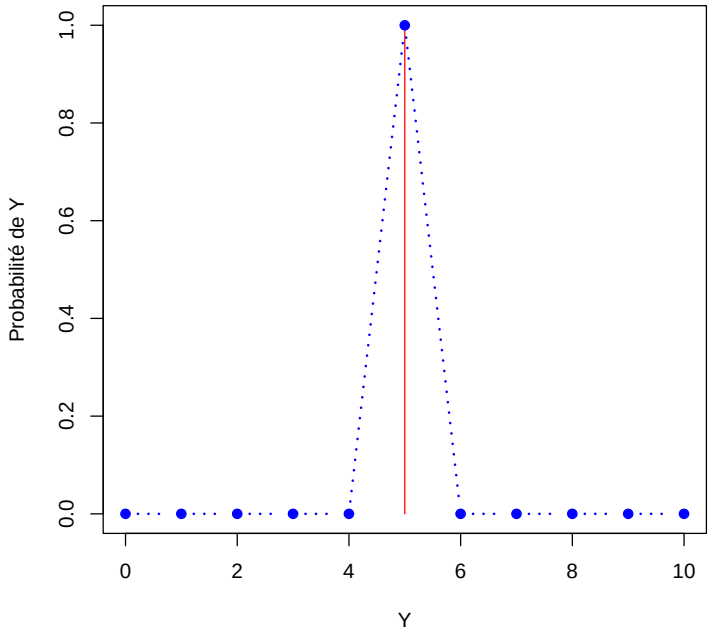

FIGURE 2.3 – noyau de type Dirac pour y=5 et h=0.

• Dirac uniforme discret

En [1976], Aitchison and Aitken ont introduit un estimateur à noyau discret pour une distribution discrète catégorique ou finie (voir aussi, Li and Racine [Li and Racine]). Par conséquent, nous en déduire son noyau associé discret asymétrique que nous présentons comme suit. D'abord, le support $\aleph$ de la fonction f à estimer, est fini de taille fixe $c \in \mathbb{N} - \{0, 1\}$. Si la variable aléatoire X étudiée prend c valeurs différentes, c'est-à-dire $\aleph = \{0, 1, ..., c1\}$, alors, le noyau discret dans (2.1) pourrait être

$$K_{x,h}(y) = (1-h)\mathbf{1}_{y=x} + \frac{h}{c-1}\mathbf{1}_{y\neq x}, \quad \forall y \in \aleph \tag{2.13}$$

où h appartient à [0, 1]. De plus, la cible x peut être considérée comme point de référence de X et le paramètre de lissage h est tel que $1 - h$ est la probabilité de succès du point de référence. Son esperance et sa variance sont donnée, comme :

$$\mathbf{E}(\mathcal{A}_{x,h}) = x + h\Big(1 - x - \frac{x}{c-1} + \frac{hc}{2}\Big),$$

$$\mathbf{Var}(\mathcal{A}_{x,h}) = -\left\{\frac{c^2(-2x+c-1)^2}{4(c-1)^2}\right\}h^2 + \left\{\frac{c(6x^2+2c^2-3c+1-6xc+6x)}{6(c-1)}\right\}h.$$

La figure suivante donne l'allure de noyau associé Dirac uniforme discrète de deuxième ordre pour x fixé $y = x = 5$, $h > 0$ et $c = 4$.

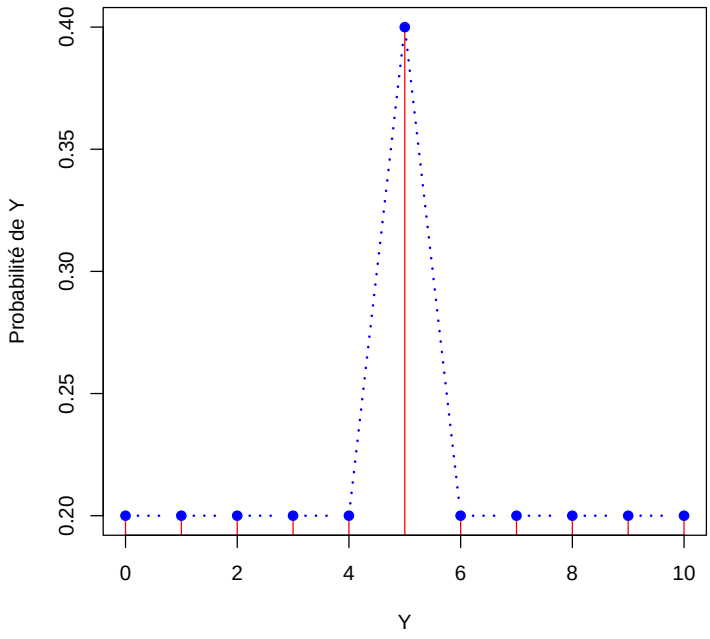

FIGURE 2.4 – noyau de type Dirac uniforme discret pour $y = 5$, $h = 0.6$ et $c = 4$.

• Noyau associé discret Triangulaire

Les noyaux associés discrets Triangulaires ont été proposés par Kokonendji et al. [2007b] et Kokonendji and Zocchi [2010]. Nous donnons maintenent la définition d'un noyau associé discret Triangulaire.

Définition 3. Soit f une fonction de masse de probabilité sur $\aleph$ donneé dans la définition . Soit $h > 0$ le paramètre de lissage et a $\in \mathbb{N}$ un entier fixé. Le noyau discret Triangulaire $T_{a,x,h}$ associé à la variable aléatoire $\mathcal{T}_{a,x,h}$ d'ordre h, de centre x et de bras a définie sur $\aleph_x = \{x, x\pm1, ..., x\pm a\}$ est donné par :

$$T_{x,a,h}(y) = P(\mathcal{T}_{a,x,h} = y) = \frac{(a+1)^h - |\, y-h\,|^h}{(2a+1)(a+1)^h - 2\sum_{k=0}^{a} k^h}, \qquad \forall y \in \aleph_x. \tag{2.14}$$

Le noyau Traingulaire discret vérifie les conditions (2.2) - (2.5) d'un noyau associé discret. La figure suivante donne l'allure de noyau associé discrète Triangulaire de deuxième ordre pour x fixé $y = x = 5$, $h > 0$ et $a = 1$.

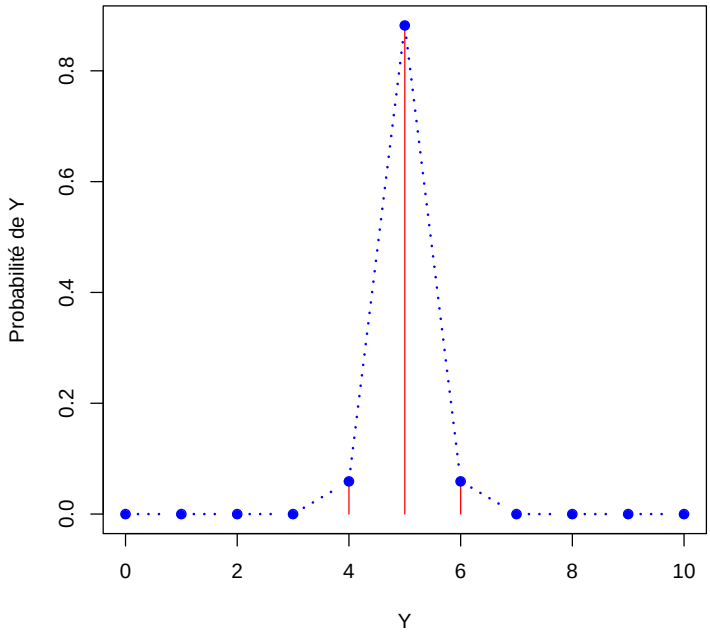

FIGURE 2.5 – Noyau Triangulaire pour $y = 5$, $h = 0.1$ et $a = 1$.

Remarque 2. Il existe des lois discrètes qui ne peuvent être associées à aucun noyau discret. En effet, si on considère la loi uniforme discrète $\mathcal{U}(c, a)$ centrée en $c \in \mathbb{N}$ et de bras $a \in \mathbb{N}^*$, le noyau discret associé serait *U(x, a)*, de loi $\mathcal{U}(x, a)$ sur $\aleph_{x,a} = \{x, x \pm 1, ..., x \pm a\}$ avec $\cup_x \aleph_{x,a} = \{-a, ..., -1\} \cup \mathbb{N} \supseteq \mathbb{N}$. Le noyau discret associé correspondant s'écrirait :

$$U_{x,a}(y) = \tfrac{1}{2a+1} \mathbf{1}_{\{x,x\pm1,...,x\pm a\}}(y), \qquad y \in \mathbb{N}.$$

D'après la Définition 2 du noyau discret associé, il apparaît qu'on ne peut pas établir de correspondance entre *(x,a)* et *(x,h)*. Le paramètre de lissage habituel $h > 0$ ne peut se substituer ici à $a \in \mathbb{N}^*$. Cette remarque est aussi valable pour une loi Triangulaire discrète malgré sa propriété de symétrie autour de x. Cependant si $a = 0$, la loi discrète uniforme $\mathcal{U}(x, 0)$ correspond à une loi de Dirac $\mathcal{D}(x)$ en x.

2.2 Propriétés de l'estimateur

La proposition suivante est nécessaire pour l'étude des estimateurs à noyaux discrets.

Proposition 1.

Soit $X_1, ..., X_n$ un n-échantillon *i.i.d* issu d'une variable aléatoire X de la fonction de masse de probabilité inconnue f sur $\aleph$. Soit $\hat{f}_h$ un estimateur à noyau associé discret de f. Alors, pour $x \in \aleph$ et $h > 0$ on a :

$$\mathbf{E}\{\hat{f}_h(x)\} = \mathbf{E}\{f(\mathcal{K}_{x,h})\}; \tag{2.15}$$

où $\mathcal{K}_{x,h}$ est la variable aléatoire de loi $K_{x,h}$ sur $\aleph_x$. De plus, on a $\hat{f}_h(x) \in [0,\, 1]$ pour $x \in \aleph$ et

$$\sum_x \hat{f}_h(x) = \mathcal{C}, \tag{2.16}$$

où C est une constante strictement positive et finie (Voir Kiesse ((2008))).

2.2.1 Biais et variance

Soit $\hat{f}(x)$ l'estimateur d'une loi discrete construit par les noyaux discrèts sont biais et sa variance et donnée par :

• **noyau de Poisson**

$$\mathbf{Biais}(\hat{f}_h(x)) = hf^{(1)}(x) + \frac{1}{2}(x+h)f^{(2)}(x) + o(h) \tag{2.17}$$

$$\mathbf{Var}(\hat{f}_h(x)) = \frac{1}{n}f(x)\frac{(x+h)^x}{x!}. \tag{2.18}$$

où $f^{(1)}$ et $f^{(2)}$ est la différence finie donnée comme suit :

$$f^{(1)}(x) = \begin{cases} \{f(x+1) - f(x-1)\}/2 & \text{si} x \in \mathbb{N}\backslash\{0\}; \\ \{f(1) - f(0)\} & \text{si } x = 0; \end{cases} \tag{2.19}$$

et

$$f^{(2)}(x) = \begin{cases} \{f(x+2) - 2f(x) + f(x-2)\}/4 & \text{si} x \in \mathbb{N}\backslash\{0,1\}; \\ \{f(3) - 3f(1) + 2f(0)\}/4 & \text{si } x = 1; \\ \{f(2) - 2f(1) + f(0)\}/2 & \text{si } x = 0. \end{cases} \tag{2.20}$$

- **Noyau Binomial**

$$\textbf{Biais}(\hat{f}_h(x)) = hf^{(1)}(x) + \frac{1}{2}(x+h)(\frac{1-h}{x+1})f^{(2)}(x) + o(h) \quad (2.21)$$

$$\textbf{Var}(\hat{f}_h(x)) = \frac{1-h}{n}(\frac{x+h}{x+1})^x f(x). \quad (2.22)$$

où $f^{(1)}(x)$ et $f^{(2)}(x)$ sont données dans (2.19) et (2.20), respectivement.

- **Noyau Binomial Négatif**

$$\textbf{Biais}(\hat{f}_h(x)) = hf^{(1)}(x) + \frac{1}{2}(x+h)(\frac{2x+1+h}{x+1})f^{(2)}(x) + o(h) \quad (2.23)$$

$$\textbf{Var}(\hat{f}_h(x)) = \frac{1}{n}\frac{2}{x!}(\frac{x+h}{2x+1+h})^x(\frac{x+1}{2x+1+h})^{x+1} f(x). \quad (2.24)$$

où $f^{(1)}(x)$ et $f^{(2)}(x)$ sont données dans (2.19) et (2.20), respectivement.

Remarque 3. L'estimateur avec les différents noyaux standarde ne converge pas au sens de l'erreur quadratique moyenne intégrée *(MISE)*, car pour x fixé, la limite du biais de $\hat{f}_h(x)$ quand $h \to 0$ est donnée par :

$$lim_{h\to 0}\textbf{Biais}(\hat{f}_h(x)) \neq 0 \quad (2.25)$$

- **Noyau Dirac**

$$\textbf{Biais}(\hat{f}_h(x)) = hf^{(1)}(x) + \frac{1}{2}(x+h)(\frac{2x+1+h}{x+1})f^{(2)}(x) + o(h) \quad (2.26)$$

$$\textbf{Var}(\hat{f}_h(x)) = \frac{1}{n}\frac{2}{x!}(\frac{x+h}{2x+1+h})^x(\frac{x+1}{2x+1+h})^{x+1} f(x). \quad (2.27)$$

où $f^{(1)}(x)$ et $f^{(2)}(x)$ sont données dans (2.19) et (2.20), respectivement.

- **Noyau Dirac uniforme discret**

$$\textbf{Biais}(\hat{f}_h(x)) = \frac{-hc}{c-1}f(x) + \frac{h}{c-1}\sum_{i=0}^{c-1} f(i), \quad (2.28)$$

$$\textbf{Var}(\hat{f}_h(x)) = \frac{1}{n}\Big[f(x)(1-h)^2 + \frac{h^2}{(c-1)^2}\Big\{\sum_{i=0}^{c-1} f(i) - f(x)\Big\}\Big] - \frac{1}{n}\Big[f(x)(1-h) + \frac{h}{c-1}\Big\{\sum_{i=0}^{c-1} f(i) - f(x)\Big\}\Big]^2 \quad (2.29)$$

- **Noyau associé discret Triangulaire**

$$\textbf{Biais}(\hat{f}_h(x)) = \left(\frac{\log(a+1)}{2}S(a) - 2\sum_{k=1}^{a} k^2\log(k)\right)h\frac{f^{(2)}}{2} + o(h) \tag{2.30}$$

$$\textbf{Var}(\hat{f}_h(x)) = \frac{1}{n}\left\{f(x)\frac{(a+1)^{2h}}{\imath^2(a,h)} + \sum_{y\in\aleph_x\setminus\{x\}} f(y)\mathbb{T}^2_{a,x,h}(y)\right\} - \frac{1}{n}\left\{f(x)\frac{(a+1)^h}{\imath(a,h)} + \sum_{y\in\aleph_x\setminus\{x\}} f(y)\mathbb{T}_{a,x,h}(y)\right\}^2. \tag{2.31}$$

2.2.2 Convergence ponctuelle

Le résultat suivant garantit que l'estimateur à noyau discret est asymptotiquement sans biais en tout point x.

Proposition 2.

Soit $f : \aleph \in \mathbb{R}$ une fonction de masse de probabilité et soit $x \in \aleph$ fixé. Si $\hat{f}_h(x)$ est l'estimateur à noyau discret de f alors :

$$\boldsymbol{E}\{\hat{f}_h(x)\} = \sum_{y\in\aleph\cap\aleph_x} f(y)K_{x,h}(y) \to f(x) \;\; quand \;\; h \to 0 \; et \; n \to \infty$$

2.2.3 Moyenne quadratique et moyenne quadratique intégrée

Les expressions de l'erreur quadratique moyenne et l'erreur quadratique moyenne intégrée (en englais "Mean Integrated Squared Error") s'expriment en fonction du biais et de la variance comme suit

$$\begin{aligned} MSE(\hat{f}_h(x)) &= \mathbf{E}\{\hat{f}_h(x) - f(x)\}^2 && (2.32)\\ &= \textbf{Var}\{\hat{f}_h(x)\} + \textbf{Biais}^2[\hat{f}_h(x)] && (2.33)\end{aligned}$$

et

$$\begin{aligned} MISE(\hat{f}_h(x)) &= \mathbf{E}\sum_{x\in\aleph}\{\hat{f}_h(x) - f(x)\}^2 && (2.34)\\ &= \textbf{Var}\sum_{x\in\aleph}\{\hat{f}_h(x)\} + \sum_{x\in\aleph}\textbf{Biais}^2[\hat{f}_h(x)] && (2.35)\end{aligned}$$

Nous présentons maintenant les résultats de convergence en moyenne quadratique et en moyenne quadratique intégrée de l'estimateur à noyau discret.

Proposition 3.

Soit $f : \aleph \in \mathbb{R}$ une fonction de masse de probabilité et soit $x \in \aleph$ fixé. Si $\hat{f}_h(x)$ est l'estimateur à noyau discret de f alors $\boldsymbol{E}\{\hat{f}_h(x) - f(x)\}^2 \to 0$ *quand* $h \to 0$ *et* $n \to \infty$.

Proposition 4. [Kokonendji and Kiessé ((2011))]

Soit f une fonction de masse de probabilité sur $\aleph$. Alors l'estimateur à noyau discret $\hat{f}_h(x)$ de f est tel que pour $n \to \infty$ et $h \to 0$, on a

$$MISE(\hat{f}_h(x)) = \frac{1}{n}\sum_{x\in\aleph} f(x)[\{Pr(\mathcal{K}_{x,h} = x)\}^2 - f(x)] + \sum_{x\in\aleph}[f\{\mathbf{E}(\mathcal{K}_{x,h})\} - f(x) + \frac{1}{2}\mathbf{Var}(\mathcal{K}_{x,h})f^{(2)}(x)]_2 + o(\frac{1}{n} + h^2), \quad (2.36)$$

où $f^{(2)}(x)$ est donnée dans (2.20).

Pour l'estimateur à noyau du type Dirac f_0 (estimateur empirique), l'erreur quadratique moyenne intégrée est données comme suit[]

$$MISE(f_0(x)) = \frac{1}{n}\sum_{x\in\aleph} f(x)\{1 - f(x)\} = \frac{1}{n}\{1 - \sum_{x\in\aleph} f^{(2)}(x)\}. \quad (2.37)$$

L'estimateur à noyau du type de Dirac converge en moyenne quadratique intégrée, c'est à dire,$MISE(f_0(x)) \to 0$ quand $n \to \infty$ car $0 \leq \sum_{x\in\aleph} f^2(x) < 1$.

Même pour l'estimateur à noyau Traingulaire discret, L'erreur quadratique moyenne intégrée MISE est donnée comme suit

$$MISE(\hat{f}_h(x)) = \frac{1}{n}\sum_{x\in\aleph} f(x)[\{\frac{(a+1)^h}{p(a,h)}\}^2 - f(x)] + \frac{1}{4}\{\mathbf{Var}(a,h)\}^2\sum_{x\in\aleph}\{f^{(2)}(x)\}^2 + o(\frac{1}{n} + h^2). \quad (2.38)$$

L'estimateur à noyau Triangulaire converge en moyenne quadratique intégrée, c'est à dire $MISE(\hat{f}_h(x)) \to 0$ quand $n \to \infty$ et $h \to 0$, car $lim_{h\to 0}\frac{(a+1)^h}{p(a,h)} = 1$, $0 \leq \sum_{x\in\aleph} f(x)\{1 - f(x)\} < 1$, $lim_{h\to 0}\boldsymbol{Var}(a,h) = 0$ et $\sum_{x\in\aleph}\{f^2(x)\}^2$ est finie.

2.2.4 Convergence en loi

Proposition 1. Abdous and Kokonendji ((2009))

Soit $f : \aleph_1 \to \mathbb{R}$ une fonction de masse de probabilité et soit $x \in \aleph_1$ fixé et $\hat{f}_h(x)$ est l'estimateur à noyau discrète de f, donc

$$\frac{\hat{f}_h(x) - \mathbf{E}(\hat{f}_h(x))}{\sqrt{\mathbf{Var}(\hat{f}_h(x))}} \underset{\rightarrow}{\mathcal{L}} \mathcal{N}(0,1) \quad quand \quad n \to \infty$$

où $\underset{\rightarrow}{\mathcal{L}}$ désigne la convergence en loi et $\mathcal{N}(0,1)$ et la loi normale standarde.

2.3 Choix du paramètre de lissage

Dans cette section nous présentons quelques méthodes classiques pour le choix du paramètre de lissage dans l'estimation des fonctions discrètes pour approcher la valeur idéale de la fenêtre h définie par (Voir Kokonendji and Kiessé ((2006)))

$$h_{id} = \arg\min_{h>0} MISE(n,h,K,f) = h_{id}(n,K,f) \tag{2.39}$$

La première approche consiste à minimiser le *MISE*, ou encore les approximations *AMISE* du *MISE* obtenues en utilisant les différences finies de f. En effet, l'existence d'un minimum par rapport à h est garantie par la décroissance de la variance intégrée et la croissance du carré du biais intégré dans le risque quadratique global (2.35). Pour une petite valeur de h, le biais est également petit mais la variance est grande. A l'inverse, si h est grand, la variance qui devient petite et le biais plus grand. Pour trouver la fenêtre optimale, on doit balancer les approximations du carré du biais et de la variance. Autrement dit, il existe $\epsilon > 0$ telle que la fonction $h \to AMISE(n,h,K,f)$ soit décroissante sur $]0,\epsilon[$ est croissante sur $]\epsilon,+\infty[$ pour tout $h > 0$ (On peut se référer aux travaux de Kiesse ((2008))).

La deuxième méthode est la validation croisée qui est bien connue, elle consiste à minimiser par rapport à h un estimateur de *MISE* pour trouver le paramètre de lissage optimal (Voir Kiesse ((2008))).

2.3.1 Minimisation de l'erreur quadratique moyenne intégrée

Cette méthode consiste à minimiser l'erreur quadratique moyenne intégrée *MISE* ou asymptotique(AMISE). Nous rappelons que le *MISE* est donnée par

$$MISE = \boldsymbol{E}\sum_{x\in\mathbb{N}}[\hat{f}_h(x) - f(x)]^2 = \sum_{x\in\mathbb{N}} \boldsymbol{Var}[\hat{f}_h(x)] + \sum_{x\in\mathbb{N}}\{\boldsymbol{Biais}^2[\hat{f}_h(x)]\}.$$

La variance peut être approximée comme suit

$$\begin{aligned}
\mathbf{Var}[\hat{f}_h(x)] &= \frac{1}{n}\mathbf{Var}[K_{x,h}(X)] && (2.40)\\
&= \frac{1}{n}\mathbf{E}[K_{x,h}]^2 - \mathbf{E}\{[K_{x,h}(X)]\}^2 && (2.41)\\
&= \frac{1}{n}\{\sum_{y\in\mathbb{N}} f(y)[Pr(\mathcal{K}_{x,h}=y)]^2 - [\sum_{y\in\mathbb{N}} f(y)Pr(\mathcal{K}_{x,h}=y)]^2\} && (2.42)\\
&= \frac{1}{n}[f(x)\sum_{x\in\mathbb{N}}\mathcal{K}_{x,h}^2 - f^2(x)] + o(n^{-1}h). && (2.43)
\end{aligned}$$

Sous la condition $\sum_{y\in\mathbb{N}} yK_{x,h}(y) \to x$ quand $h \to 0$, on a l'approximation suivante

$$\boldsymbol{Var}[\hat{f}_h(x)] = \tfrac{1}{n}f(x)Pr(\mathcal{K}_{x,h}=x),$$

où $\mathcal{K}_{x,h}$ est la variable aléatoire discrète de densité $K_{x,h}$. On approxime le biais de emph$\hat{f}_h$ en utilisant le développement discret de Taylor à l'ordre 2.

$$\begin{aligned}
\mathbf{Biais}[\hat{f}_h(x)] &= \mathbf{E}[\hat{f}_h(x) - f(x)]\\
&= f(x)\mathbf{E}[K_{x,h}] - f(x) + \frac{1}{2}\mathbf{Var}[\hat{f}_h]f^2(x) + o(h),
\end{aligned}$$

où $f^{(k)}(x)$ est la différence finie d'ordre $K \in \mathbb{N}\backslash\{0\}$: $[f^{(k)}(x) = f^{(k-1)}(x)]^{(1)}$ avec

$$f^{(1)}(x) = \begin{cases} [f(x+1) - f(x-1)]/2 & \text{si } x \in \mathbb{N}^*;\\ f(1) - f(0) & \text{si } x = 0. \end{cases}$$

Finalement, le *MISE* peut être approximer par

$$AMISE(h) = \frac{1}{n}\sum_{x\in\mathbb{N}} f(x)Pr(\mathcal{K}_{x,h}=x) + \sum_{x\in\mathbb{N}}\{f(x)\mathbf{E}[K_{x,h}] - f(x) + \frac{1}{2}\mathbf{Var}[\hat{f}_h]f^2(x)\}^2.$$

Le paramètre de lissage h_{amise} dans ce cas peut être obtenu de la manière suivante

$$h_{amise} = \arg\min_h AMISE(h).$$

Le paramètre de lissage h_{amise} n'est pas utilisable directement en pratique, car *AMISE(h)* dépend de la densité discrète inconnue *f*. Cependant, on utilise l'estimateur du type de Dirac f_0 comme estimateur de f et calculer le paramètre de lissage optimal par la suite.

2.3.2 Validation croisée

La méthode classique de validation croisée (en anglais "Cross-Validation") ne faisant pas usage des approximations des dérivées de f est toujours applicable dans le contexte des estimateurs à noyau discret pour mieux estimer la valeur idéale (2.39) de h. La fenêtre optimale s'obtient par (Voir Kiesse ((2008)))

$$h_{cv} = \arg\min_{h>0} CV(h), \tag{2.44}$$

où

$$\begin{aligned} CV(h) &= \sum_{x\in\mathbb{N}} \hat{f}_h^2(x) - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{h,-i}(X_i) \\ &= \sum_{x\in\mathbb{N}} [\frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i)]^2 - \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} K_{X_{i,h}}(X_j) \end{aligned}$$

où $\hat{f}_{h,-i}$ est calculé sans l'observation X_i.

Le principe de cette méthode est de minimiser par rapport à h un estimateur de *MISE* pour trouver le paramètre de lissage optimal. Pour cela, *MISE* (2.34) peut être dévelopé comme suit :

$$MISE = \mathbf{E}\{\sum_{x\in\mathbb{N}} \hat{f}_h^2(x)\} - 2\mathbf{E}\{\sum_{x\in\mathbb{N}} \hat{f}_h^2(x) f(x)\} + \sum_{x\in\mathbb{N}} f^2(x)\}.$$

Le terme $\sum_{x\in\mathbb{N}} f^2(x)$ n'est pas aléatoire, et ne dépend pas du paramètre de lissage h. On note alors

$$MISE_{cv} = \mathbf{E}\{\sum_{x\in\mathbb{N}} \hat{f}_h^2(x)\} - 2\mathbf{E}\{\sum_{x\in\mathbb{N}} \hat{f}_h^2(x) f(x)\} = MISE_{cv}(h)$$

le terme de *MISE* qui dépend de h. Dans la suite, nous déterminerons un estimateur *CV(h)* de $MISE_{cv}$.
D'abord, on a évidemment $\sum_{x\in\mathbb{N}} \hat{f}_h^2(x)$ qui est un estimateur sans biais de $\boldsymbol{E}\{\sum_{x\in\mathbb{N}} \hat{f}_h^2(x)\}$.
Ensuite, soit

$$\hat{f}_{h,i-1}(x) = \frac{1}{n-1}\sum_{j\neq i} K_{x,h}(X_j).$$

Par construction,

$$\begin{aligned}\hat{G} &= \frac{1}{n}\sum_{i=1}^{n}\hat{f}_{h,i-1}(X_i)\\ &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}K_{X_{i,h}}(X_j)\end{aligned}$$

est un estimateur de $\boldsymbol{E}\{\sum_{x\in\mathbb{N}}\hat{f}_h^2(x)f(x)\}$ et on vérifie de plus qu'il est sans biais. En effet, d'une part, comme les $v.a.X_1,...,X_n$ sont i.i.d, on a

$$\begin{aligned}\mathbf{E}\{\hat{G}\} &= \mathbf{E}\{\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}K_{X_{1,h}}(X_j)\}\\ &= \frac{1}{(n-1)}\sum_{j\neq i}K_{X_{i,h}}(X_j)\\ &= \mathbf{E}\{K_{X_{1,h}}(X_2)\}\end{aligned}$$

D'autre part, on a successivement :

$$\begin{aligned}\mathbf{E}\sum_{x\in\mathbb{N}}\hat{f}_{n,h,K}(x)f(x) &= \mathbf{E}\{\sum_{x\in\mathbb{N}}f(x)\frac{1}{n}\sum_{i=1}^{n}K_{x,h}(X_i)\}\\ &= \mathbf{E}\{\frac{1}{n}\sum_{i=1}^{n}K_{X_{1,h}}(X_i)\}\\ &= \mathbf{E}\{K_{X_{1,h}}(X_2)\}\end{aligned}$$

Finalement, on vient de montrer que *CV(h)* est un estimateur sans biais de $MISE_{cv}$.

2.4 Conclusion

Durant ce chapitre, nous avons présenté la méthode d'estimation de la fonction de masse de probabilité par le noyau associé discret. Ensuite, nous avons rappelé les propriétés statistiques de cet estimateur ainsi la méthode de validation croisée pour le choix du paramètre de lissage. Dans le chapitre suivant, nous allons faire une comparaison par simulation des estimateurs obtenus par différents noyaux.

Chapitre 3

Applications sur des données sumulés

Introduction

Nous présentons dans ce chapitre, une étude de simulation pour comparer les différents noyaux discrets du premier ordre et deuxième ordre de l'estimation non paramétrique de la fonction de masse de probabilitée *f*, selon le critère ISE, à l'aide de package DISAKE (Discrete associated kernel estimators) sous logiciel R. Le package DISAKE est réalisé par Wansouwé, Kokonendji, Kolyang, and Wansouwé ((2015)), le lissage discret de la fonction de densité de probabilité est effectué à l'aide de trois noyaux discrets associés : Dirac Discrete Uniform (DiracDU), Binomial et Discrete Triangulare.

3.1 Etude de simulation

Dans cette section, nous présentons une étude de simulation pour évaluer l'utilisation de différents noyaux discrets pour l'estimation non paramétrique de la fonction de masse de probabilitée f sur des données de comptage simulées selon les quatre fonctions discrètes :

•D_1 une distribution binomiale avec les paramètres $n_1 = 15 \ \ et \ \ p = 0.4$:

$$f(x) = \frac{15!}{x!(15-x)!} 0.4^x.0.1^{15-x}, \quad x \in \{0, 1, ..., 15\}.$$

•D_2 une distribution de Poisson avec le paramètre $\lambda = 6$:

$$f(x) = e^{-6}\frac{6^x}{x!}, \quad x \in \mathbb{N}.$$

•D_3 une distribution géométrique avec le paramètre $p = 0.2$:

$$f(x) = 0.2.(0.8)^{x-1}, \quad x \in \mathbb{N}.$$

•D_4 un mélange de loi de Poisson et de la loi géométrique de paramètres $\mu = 8$ et $p = 0.2$:

$$f(x) = \frac{3}{5}.e^{-8}\frac{8^x}{x!} + \frac{2}{5}.0.2.(0.8)^{x-1}, \quad x \in \mathbb{N}.$$

D'abord, nous utilisons la méthode de validation croisée pour le choix de paramètre de lissage. Ensuite, Pour faire une comparaison sur la performance des éstimateurs à noyaux, nous utilisons le critère de l'erreur quadratique intégrée (ISE) tq :

$$ISE = \sum_{x\in\mathbb{N}}[\hat{f}(x) - f(x)]^2$$

Résultats de simulation

Notons que pour les fonctions de masse de probabilitée considérées, 500 replications de taille d'échantillon n = *50, 100* sont générés, et *200* replications pour *n=500*, et 30 fois pour *n=1000*. Les estimateurs à noyaux discret Binomial (Bino), Triangulaire (Triang) et dirac uniforme discret (DirDU) sont appliqués pour estimer ces fmps. Les paramètre a et c sont fixé à (a=2 et 4) dans le cas de l'utilisation de noyau Triangulaire et (c=2 et 5) pour le noyau Dirac uniforme. Nous calculons :

$$\hat{ISE} = \frac{1}{nr}\sum_{i=1}^{nr}(ISE)_i$$

et

$$sd(ISE) = \sqrt{\mathbf{Var}(ISE)}$$

Les résultats de ($\hat{ISE}$ et *sd(ISE)*) sont mentionnés dans les Tables 3.1, 3.2, 3.3 et 3.4 tout en variant la taille de l'échantillon $n \in \{50, 100, 500, 1000\}$ pour les différentes fonctions discrètes D_1, D_2, D_3 et D_4. (*Note : Les valeurs en gras indiquent les meilleurs résultats*).

Noyau	ISE	$n = 50$	$n = 100$	$n = 500$	$n = 1000$
Bino	$I\hat{S}E$	0.09314	0.05816	**0.01672**	0.00968
	sd(ISE)	0.05288	0.03703	0.01063	0.00767
DirDU	$I\hat{S}E$	**0.08554**	**0.05653**	0.02699	0.13572
c=2	sd(ISE)	0.03579	0.02261	0.00896	0.01073
DirDU	$I\hat{S}E$	0.10392	0.06213	0.02055	0.13909
c=5	sd(ISE)	0.05093	0.03518	0.01382	0.01164
Triang	$I\hat{S}E$	0.08837	0.05889	0.01939	0.01297
a=2	sd(ISE)	0.04635	0.03807	0.01368	0.00981
Triang	$I\hat{S}E$	0.08714	0.06077	0.02104	**0.00797**
a=4	sd(ISE)	0.04993	0.04023	0.01898	0.00775

TABLE 3.1 – La moyenne et ecart type de *ISE* pour la distribution Binomiale avec les paramètres $n_1 = 15$ *et* $p = 0.4$.

Noyau	ISE	$n = 50$	$n = 100$	$n = 500$	$n = 1000$
Bino	$I\hat{S}E$	0.04244	**0.01840**	0.00494	0.00257
	sd(ISE)	0.02984	0.15262	0.00442	0.00347
DirDU	$I\hat{S}E$	0.04441	0.03043	0.01930	0.11460
c=2	sd(ISE)	0.01884	0.01403	0.00507	0.00725
DirDU	$I\hat{S}E$	0.05260	0.03006	0.00678	0.00380
c=5	sd(ISE)	0.03125	0.01718	0.00542	0.00197
Triang	$I\hat{S}E$	0.04385	0.02333	0.00539	0.00802
a=2	sd(ISE)	0.02565	0.01864	0.00572	0.00755
Triang	$I\hat{S}E$	**0.04181**	0.02337	**0.00419**	**0.00075**
a=4	sd(ISE)	0.02721	0.01749	0.00540	0.00015

TABLE 3.2 – La moyenne et ecart type de *ISE* pour la distribution de Poisson avec le paramètre $\lambda = 6$.

Noyau	ISE	$n = 50$	$n = 100$	$n = 500$	$n = 1000$
Bino	$I\hat{S}E$	0.01029	0.00480	**0.00077**	**0.00035**
	sd(ISE)	0.00403	0.00202	0.00049	0.00019
DirDU	$I\hat{S}E$	0.03851	0.04642	0.06110	0.06289
c=2	sd(ISE)	0.01760	0.01843	0.01252	0.00933
DirDU	$I\hat{S}E$	0.16028	0.15065	0.02803	0.02982
c=5	sd(ISE)	0.02422	0.01624	0.01115	0.00713
Triang	$I\hat{S}E$	0.00951	**0.00336**	0.00364	0.00122
a=2	sd(ISE)	0.00309	0.00095	0.00089	0.00035
triang	$I\hat{S}E$	**0.00898**	0.00446	0.00364	0.00378
a=4	sd(ISE)	0.00260	0.00191	0.00089	0.00099

TABLE 3.3 – La moyenne et ecart type de *ISE* pour la distribution Géométrique avec le paramètre $p = 0.2$.

Noyau	ISE	$n = 50$	$n = 100$	$n = 500$	$n = 1000$
Bino	$I\hat{S}E$	0.01257	0.00589	0.00087	0.00098
	sd(ISE)	0.00574	0.00261	0.00036	0.00102
DirDU	$I\hat{S}E$	0.02008	0.01910	0.02600	0.00433
c=2	sd(ISE)	0.01312	0.01091	0.01050	0.00074
DirDU	$I\hat{S}E$	**0.00865**	**0.00470**	0.00689	0.00712
c=5	sd(ISE)	0.01067	0.00455	0.00143	0.00173
Triang	$I\hat{S}E$	0.01426	0.00556	0.00086	0.00095
a=2	sd(ISE)	0.00950	0.00152	0.00025	0.00080
Triang	$I\hat{S}E$	0.01357	0.00573	**0.00082**	**0.00091**
a=4	sd(ISE)	0.00620	0.00151	0.00023	0.00239

TABLE 3.4 – La moyenne et ecart type de *ISE* pour la mélange de loi de Poisson et Géométrique de paramètres $\mu = 8$ et $p = 0.2$.

3.2 Interprétation des résultats

Daprés les résultats de simulations obtenus dans le cadre de ce travail dans le cas de l'estimation non paramétrique de la fonction de masse de probabilitée avec l'utilisation des différents noyaux discrèts et la technique de validation croisée pour le choix de la fenêtre h, qui sont mentionnés dans les Tables 3.1, 3.2, 3.3 et 3.4 on peut observée que :

• Pour la distribution D_1, les meilleures performances au sens de *ISE* sont obtenus par le noyau Dirac uniform discret (c=2) pour un échantillon de petite de taille. Pour un échantillon de moyenne taille, le noyau Binomial fournit la meilleure performance de *ISE*. Pour un échantillon de grande taille, le noyau Triangulaire (a=4) donne le meilleur estimateur de la distribution.

• Pour la distribution D_2, les meilleures performances au sens de *ISE* sont obtenus par le noyau Triangulaire (a=4) pour un échantillon de petite ou grand de taille. Et pour un échantillon de moyenne taille, le noyau Binomial fournit la meilleure performance de *ISE*.

• Pour la distribution D_3, les meilleures performances au sens de *ISE* sont obtenus par le noyau Triangulaire (a=4) pour un échantillon de petite de taille. Pour un échantillon de moyenne taille, le noyau Triangulaire (a=2) fournit la meilleure performance de *ISE*. Et pour un échantillon de grande taille, le noyau Binomial donne le meilleur estimateur de la distribution.

• Pour la distribution D_4, les meilleures performances au sens de *ISE* sont obtenus par le noyau Dirac uniform discret (c=5) pour un échantillon de petite de taille. Et pour un échantillon de moyenne ou grande taille, le noyau Triangulaire (a=4) donne le meilleur estimateur de la distribution.

3.3 Comparaison graphique

• L'illustration graphique suivante 3.1 présente le lissage discret des données simulées par le noyau naïf (h = 0). pour la fonction discrète étudier "poisson $\lambda = 6$", ont varie la taille des échantillons. Pour les tailles d'échantillons $n \in \{50, 100, 500\}$, l'ajustement discret de f n'est pas satisfaisant. Ce n'est que dans le cas $n = 1000$ que le lissage discret à l'aide du noyau naïf est entièrement convenable.

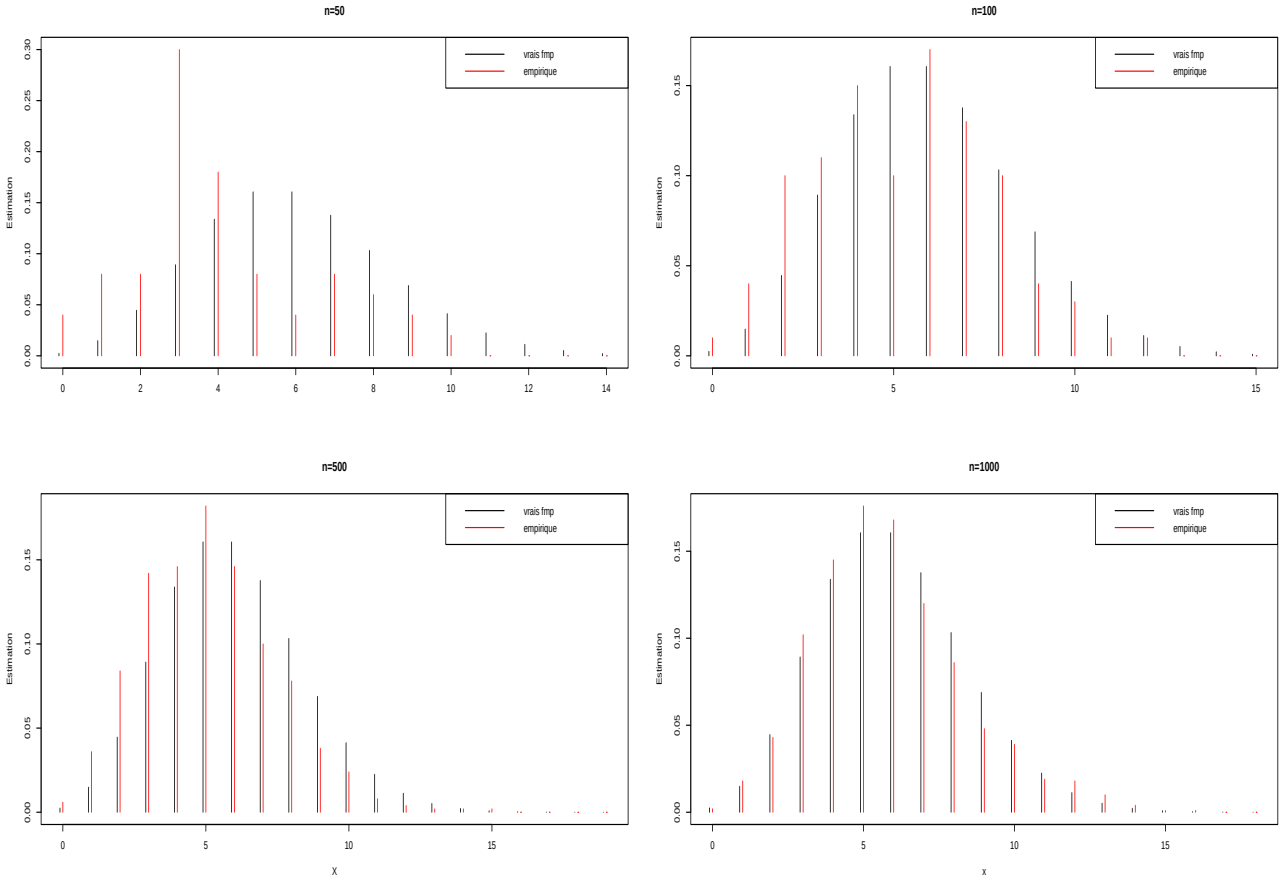

FIGURE 3.1 – Lissages discrets par un estimateur empirique ("naïf") des données simulées pour $n \in \{50, 100, 500, 1000\}$ de la distribution du Poisson $f = \mathcal{P}(6)$.

• Les illustrations graphiques suivants 3.2, 3.3, 3.4 et 3.5 présentes les lissages discrets de f à travers les noyaux discrets naïf, standards et de deuxiem ordre. Pour $n = 1000$, de manière générale les lissages discrets sont réguliers en suivant l'allure de f_0 bien que le noyau Binomial ne soit pas très performant par rapport aux autres. Dans ce cas, l'estimateur empirique ou naïf réalise le meilleur ajustement. Pour des échantillons de petites et moyennes tailles ($n \in \{50, 100\}$), l'estimateur à noyau Binomial devient le plus approprié. Cette dernière constatation est nettement visible pour une petite taille d'échantillon (n = 50), situation dans laquelle le noyau Dirac uniforme n'est plus adéquat. Pour des échantillons de grande de tailles $(n = 500)$ l'estimateur à noyau Triangulaire devient le plus approprié.

• L'illustration graphique suivante 3.2 présente le lissage discret des données simulées par les noyau dirac (h = 0), Binomiale ($h_{CV} = 0.02$), Triangulaire ($h_{CV} = 0.25$) et Dirac uniforme discret ($h_{CV} = 0.05$) pour la fonction discrète étudier "Poisson $\lambda = 6$". Pour la taille d'échantillon *n=50*.

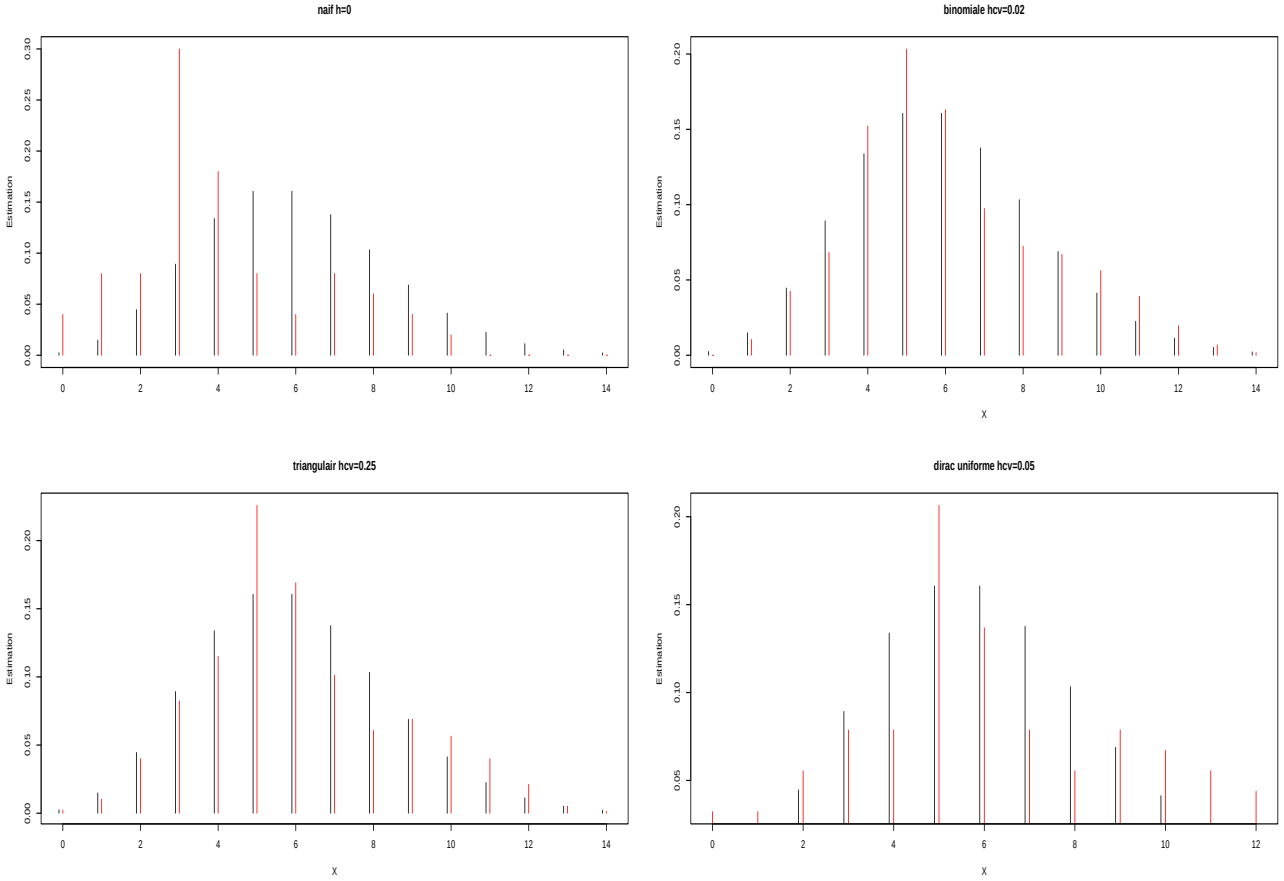

FIGURE 3.2 – Lissages discrets par les noyaux de type naïf, Binomial, Triangulaire (a=3) et Dirac uniforme discret (c=2) des données simulées *(n = 50)* de la distribution du Poisson $f = \mathcal{P}(6)$.

• L'illustration graphique suivante 3.3 présente le lissage discret des données simulées par les noyau dirac (h = 0), Binomiale ($h_{CV} = 0.026$), Triangulaire ($h_{CV} = 0.975$) et Dirac uniforme discret ($h_{CV} = 0.065$) pour la fonction discrète étudier "Poisson $\lambda = 6$". Pour la taille d'échantillon *n=100*.

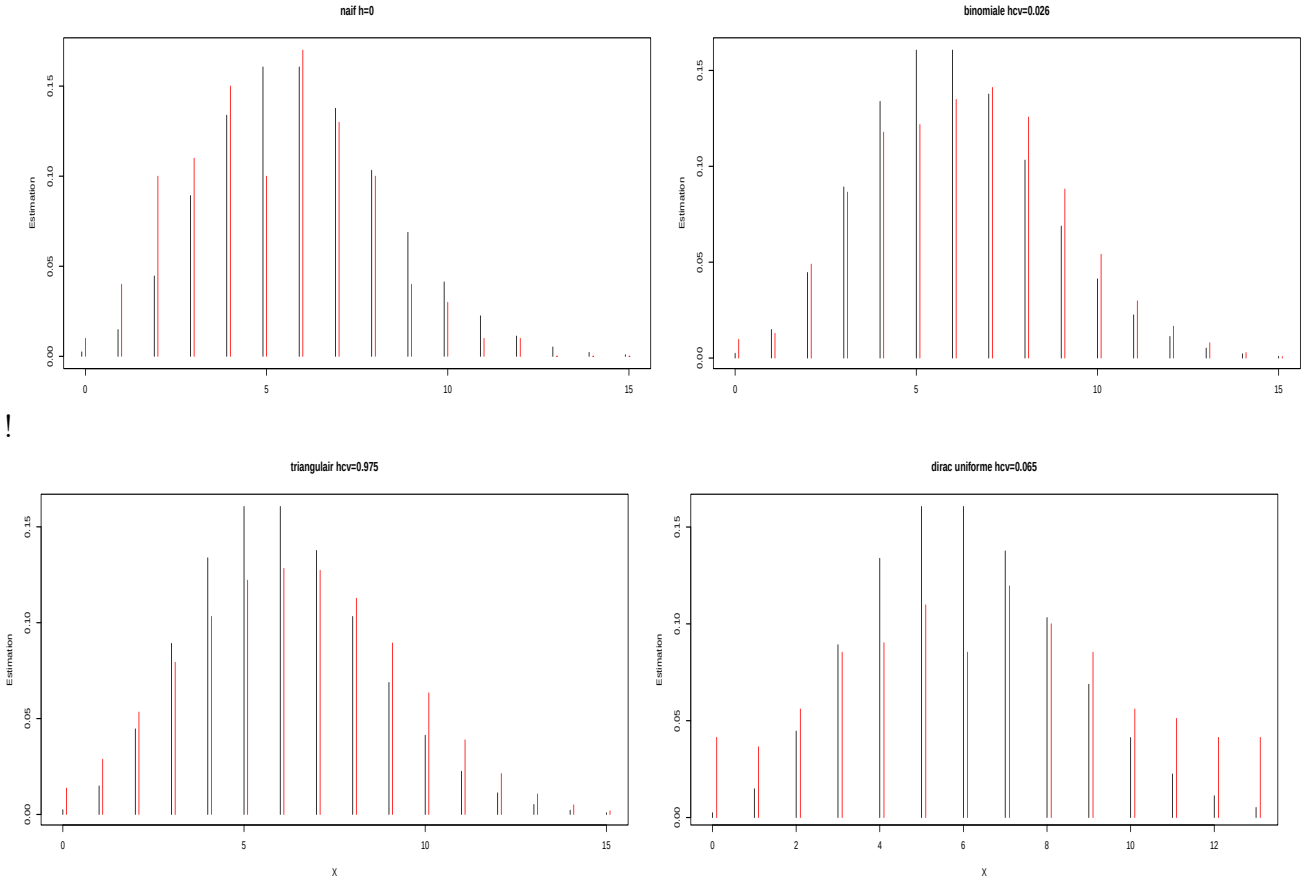

FIGURE 3.3 – Lissages discrets par les noyaux de type naïf, Binomial, Triangulaire (a=3) et Dirac uniforme discret (c=2) des données simulées *(n = 100)* de la distribution du Poisson $f = \mathcal{P}(6)$.

• L'illustration graphique suivante 3.4 présente le lissage discret des données simulées par les noyau Dirac (h = 0), Binomiale ($h_{CV} = 0.032$), Triangulaire ($h_{CV} = 0.08$) et dirac uniforme discret ($h_{CV} = 0.08$) pour la fonction discrète étudier "Poisson $\lambda = 6$". Pour la taille d'échantillon *n=500*.

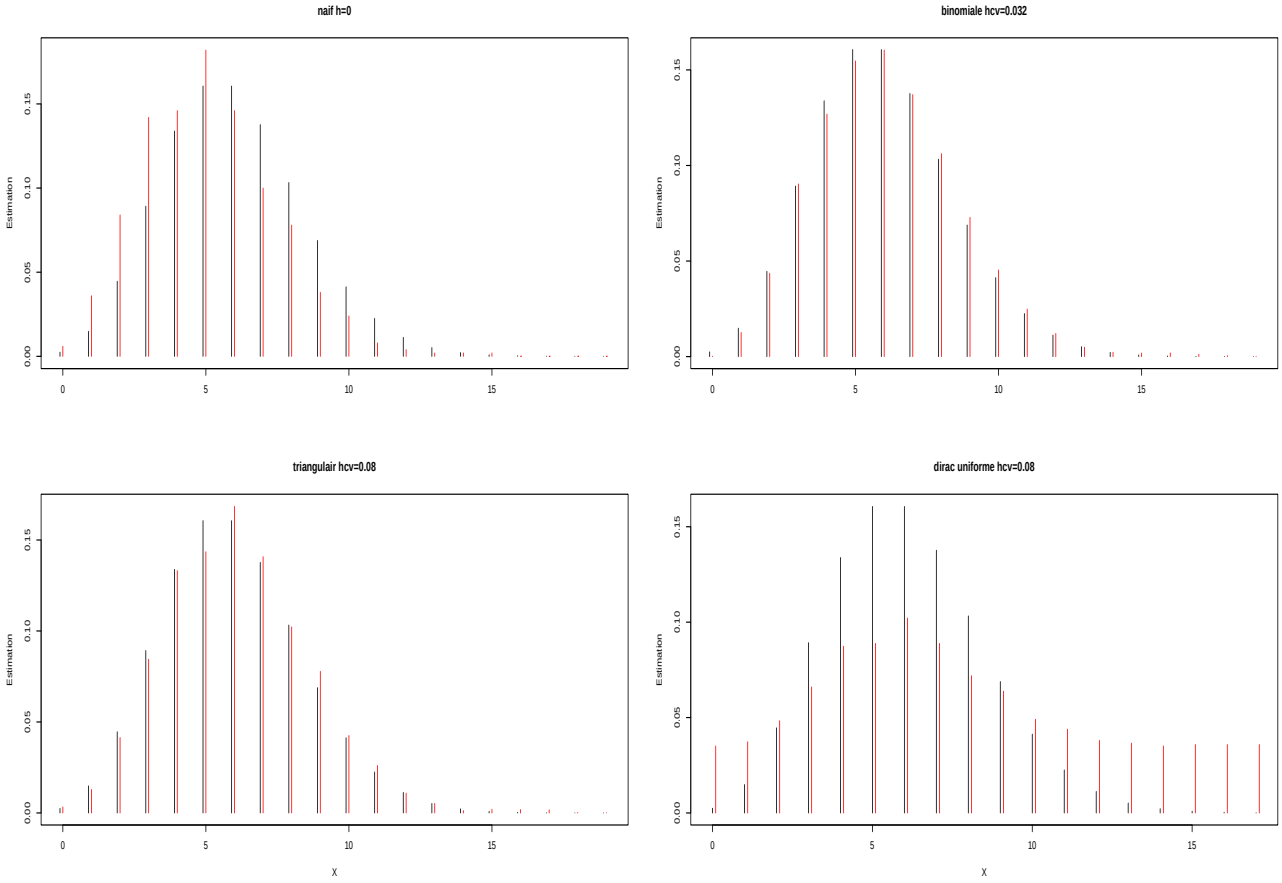

FIGURE 3.4 – Lissages discrets par les noyaux de type naïf, binomial, triangulaire (a=3) et dirac uniforme discret (c=2) des données simulées *(n = 500)* de la distribution du Poisson $f = \mathcal{P}(6)$.

• L'illustration graphique suivante 3.5 présente le lissage discret des données simulées par les noyau dirac (h = 0), binomiale ($h_{CV} = 0.032$), triangulaire ($h_{CV} = 0.08$) et dirac uniforme discret ($h_{CV} = 0.08$) pour la fonction discrète étudier "Poisson $\lambda = 6$". Pour la taille d'échantillon *n=1000*.

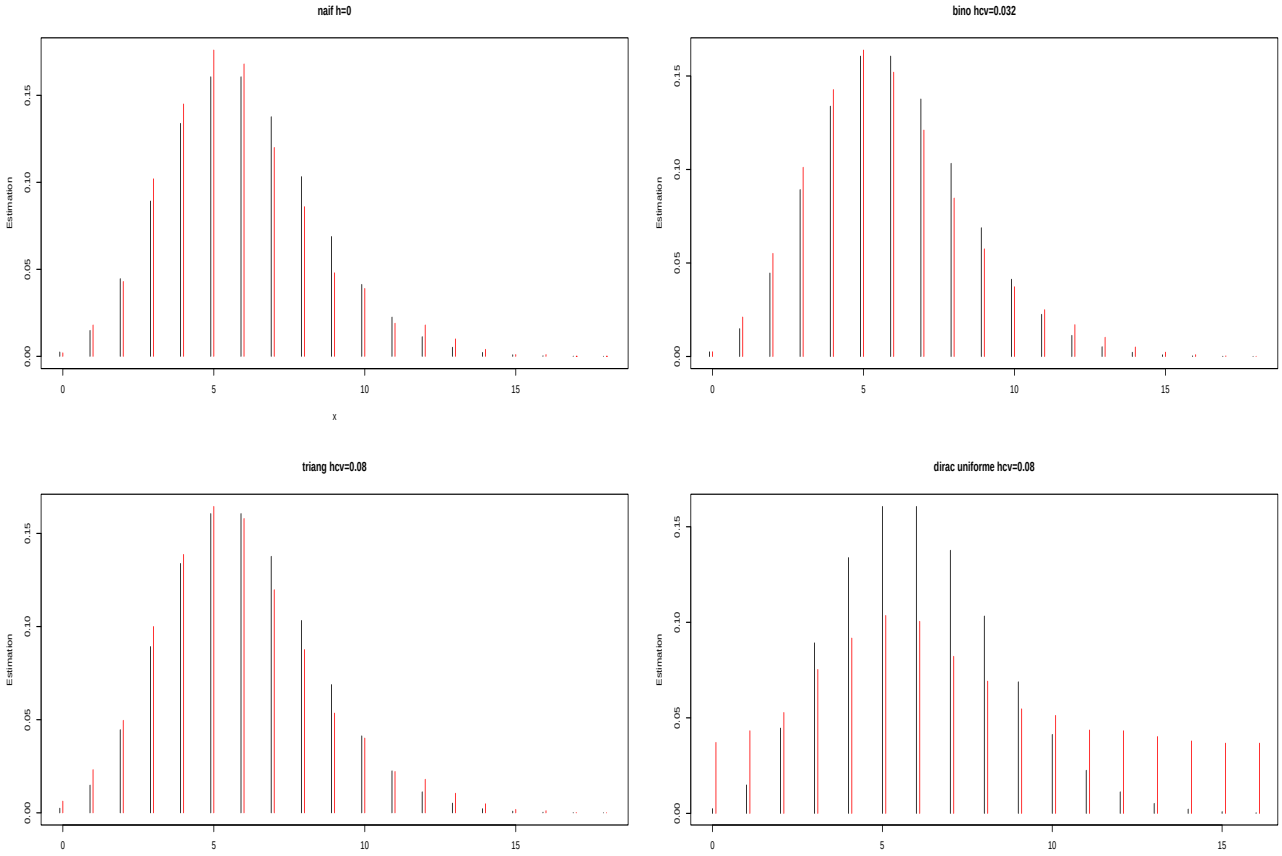

FIGURE 3.5 – Lissages discrets par les noyaux de type naïf, binomial, triangulaire (a=3) et dirac uniforme discret (c=2) des données simulées *(n = 1000)* de la distribution du Poisson $f = \mathcal{P}(6)$.

3.4 Conclusion

Durant ce chapitre, nous avons comparé les différents noyaux discrets du premier et deuxième ordre de l'estimation non paramétrique de la fonction de masse de probabilitée f, selon le critére ISE sur des données simulées, ainsi que la méthode validation croisée utilisée pour estimer le paramètre de lissage h. Les résultats obtenus montrent que l'estimation de la fonction de masse de probanilitée f avec les différents noyaux varient d'un noyau à un autre et d'un échantillon à un autre. D'aprés notre étude, nous concluons que le noyau binomial est approprié pour les données de comptage avec des échantillons de petite ou moyenne taille, le noyau triangulaire discret est recommandé pour les données de comptage avec des échantillons de grande taille.

Chapitre 4

Application sur des données réelles

Introduction

Le domaine de finance est l'un des nombreux domaines où on peut réaliser notre étude en utilison la méthode des noyaux discrets. Dans ce chapitre, nous présentons trois série statistiques réelles concernant le trading et les investissements dans les cryptomonnaies pour évaluer l'utilisation de différents noyaux discrets, du premier ordre et deuxième ordre pour l'estimation non paramétrique de la fonction de masse de probabilitée f inconnue selon le critère *ISE*.

4.1 Présentation de crypto-monnaie

Une crypto-monnaie est une devise numérique décentralisée (une monnaie virtuelle qui opère indépendamment des banques et des gouvernements). Elle peut être échangée et négociée, comme n'importe quelle devise physique, elle utilise des algorithmes cryptographiques et un protocole nommé blockchain pour assurer la fiabilité et la traçabilité des transactions. Les cryptomonnaies peuvent être stockées dans un portefeuille numérique protégé par un code secret appartenant à son propriétaire. Des plateformes d'échanges (Binance, Coinbase, Bitstamp, etc.) servent à acheter et revendre de la cryptomonnaie en ligne.

La première cryptomonnaie à avoir vu le jour, et sans doute la plus célèbre d'entre elles, est le Bitcoin. Créée en 2009 par Satoshi Nakamoto, elle a propulsé le principe de blockchain et a entrainé la création de nombreuses autres devises numériques cryptées, nommées *Altcoins* par les cryptotraders.

4.2 Application 1

Le marché des cryptomonnaies consiste plus de 12 000 genres des pièces de monnaie nommé (cryptocoins) avec une capitalisation boursière de 2,32 trillions de dollars.(Voir le site CoinMarketCap)

On a construit une série statistique de 21 éléments où on considère un élément comme suit 1 milliard de dollars investit dans un cryptocoin. Les 20 premiers cryptocoins sont les plus dominants selon la capitalisation boursière investie de dans. Ces 20 cryptocoins avec leurs abréviations sont : (Bitcoin(BTC), Ethereum(ETH), Binance coin(BNB), Cardano(ADA), Tether(USDT), Ripel(XRP), Solana(SOL), Polkadot(DOT), USD Coin(USDC), Dogecoin(DOGE), Uniswap(UNI), Terra(LUNA), Wrapped Bitcoin(WBT), Binance USD(BUSD), Litecoin(LTC), Avalanche(AVAX), Chainlink(LINK), Bitcoin Cash(BCH), Algorand(ALGO), SHIBA INU(SHIB)). Et l'élément 21 c'est la somme des unités inventée dans tout le reste des cryptocoins.

Le premier élément, c'est 1 milliard de dollars investis en BTC avec un effectif de 1163. C'est-à-dire 1163 milliards de dollars investis en BTC. La Table suivante 4.1 représente les cryptocoins ou les unités son inventés avec l'effectif des unités.

BTC	1163	DOT	42	LTC	13
ETH	455	USDC	32	AVAX	12
BNB	79	DOGE	31	LINK	12
ADA	72	UNI	15	BCH	11
USDT	68	LUNA	14	ALGO	10
XRP	53	WBT	13	SHIB	10
SOL	48	BUSD	13	Le reste	156

TABLE 4.1 – cryptomonie

Pour évaluer l'utilisation de différents noyaux discrets pour l'estimation non paramétrique de la fonction de masse de probabilitée f sur les données fournit dans la Table 4.1, nous utilisons l'erreur quadratique intégrée définie par

$$ISE^0 = \sum_{x\in\mathbb{N}} [\hat{f}(x) - f^0(x)]^2$$

où $f^0(x)$ est l'estimateur empirique (naïf) de la fmp. Les variables indépendantes catégorielles peuvent être utilisées dans une estimation fmp non paramétrique, mais elles doivent être codées. Dans notre étude, nous utilisons le code suivant : 1="investis en BTC"; 2="investis en ETH" . . .; 21="le reste" respectivement. Le paramètre de lissage est obtenu avec la méthode

validation croisée. Les résultats sont donnés dans la Table 4.2. (*Note : La valeur en gras indique la meilleur résultat*).

Noyau	bino	triang a=3	triang a=4	dirDU c=2	dirDU c=5
h_{cv}	0.69	0.1	0.1	0.1	0.1
ISE^0	0.27565	0.28729	**0.24865**	0.33252	0.27022

TABLE 4.2 – Résultats pour la première application.

• L'illustration graphique suivante 4.1 présente les résultats graphiques de différents lissages discrets par les noyaux discrets naif, standards et de dexième ordre de la distribution empirique pour la taille d'échantillon $n = 2322$.

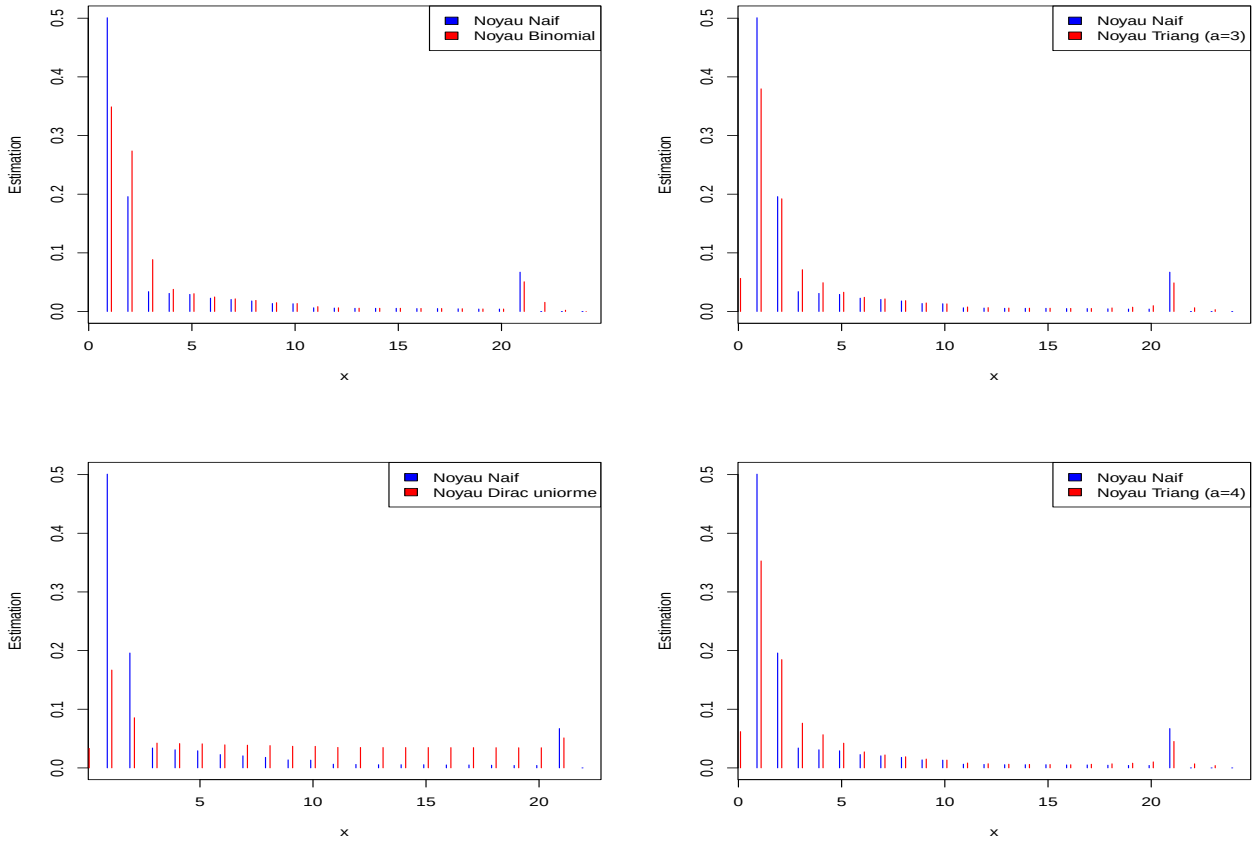

FIGURE 4.1 – Lissages discrets par les noyaux de type de dirac, binomial , triangulaire (a=3 et 4) et dirac uniforme (c=2) pour les données réelles d'investisement de crypto-monaies n = 2322.

4.2.1 Discussion

D'aprés la Table 4.2 on peut observer que le meilleur résultat au sens de minimisation de ISE^0 est obtenu avec le noyau associé discret triangulaire (a=4). La valeur de ISE^0 diminué avec l'augmentation de bras a de noyau triangulaire. Et d'aprés la figure 4.1 on observe que, le meilleur ajustement est obtenu par le noyau triangulaire (a=4).

4.3 Application 2

Dans cet exemple, on a extrait une série statistique des valeurs atteintes par BTC dans les jours entre (03/08/2021) et (04/10/2021) ou les éléments sont les valeurs de BTC et l'effectif de chaque élément c'est le nombre des jour ou le BTC attint cette valeur.

Le BTC atteint la valeur 38 000 dollars pendant un seul jour dans cette période. Alors l'élément, c'est 38 000 et son effectif c'est 1. Le diagramme en baton donné dans la figure 4.2 représente la variation de prix de BTC (en 1000$) durant les 64 jours. (Voir le site CoinMarketCap)

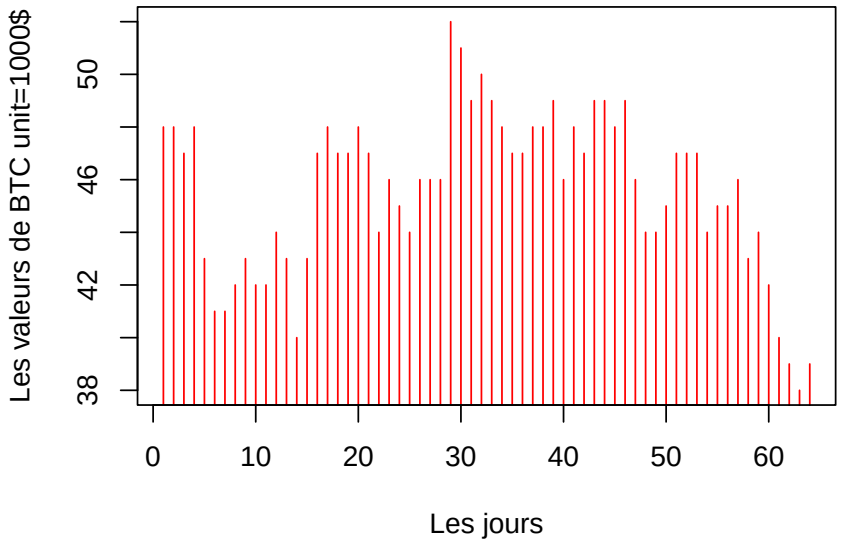

FIGURE 4.2 – Données de la variation de prix de BTC (en 1000$) par jour.

La Table suivante 4.3 fournit le résumé statistique de la variation de prix de BTC (en 1000$) durant les jours avec $n = 64$.

38	1	43	5	48	10
39	2	44	7	49	6
40	2	45	4	50	1
41	2	46	7	51	1
42	4	47	11	52	1

TABLE 4.3 – Données de prix de BTC (en 1000$) observées pendant les 64 jours.

Les résultats de (ISE^0 et h_{CV}) sont mentionnés dans la Table 4.4 pour la taille de d'échantillon $n = 64$ pour les différentes noyaux discrètes. (*Note : La valeur en gras indique la meilleur résultat*).

Noyau	bino	triang a=3	triang a=4	dirDU c=2	dirDU c=5
h_{cv}	0.028	0.84	0.56	0.07	0.07
ISE^0	0.19671	0.18801	0.18445	**0.09575**	0.12539

TABLE 4.4 – Résultats pour la deuxième application.

• L'illustration graphique suivante 4.3 présente les résultats graphiques de différents lissages discrets par les noyaux discrets naif, standards et de dexième ordre de la distribution empirique pour la taille d'échantillon n=*64*.

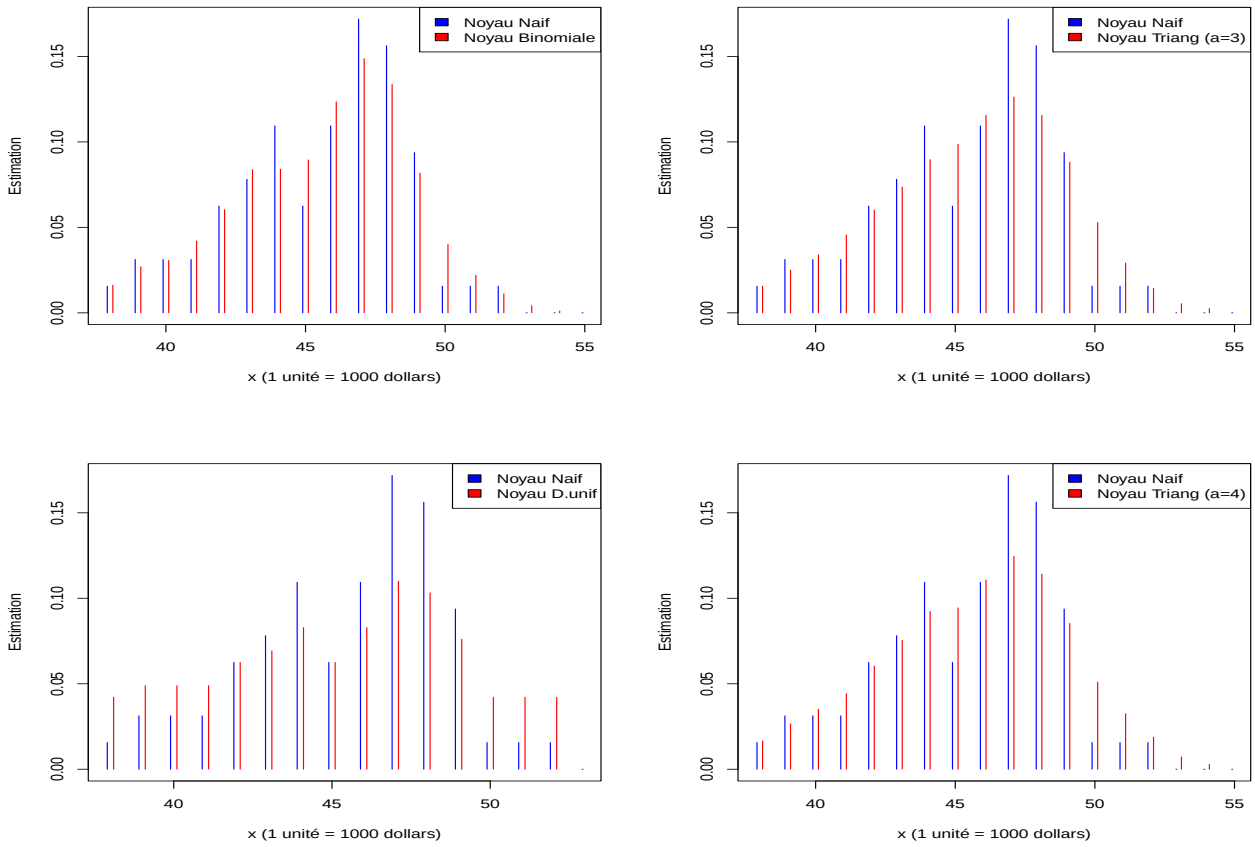

FIGURE 4.3 – Lissages discrets par les noyaux de type de dirac, binomial, triangulaire (a=3 et 4) et dirac uniforme (c=2) pour les données réelles sur la variation de prix de BTC (en 1000$) durant les 64 jours $n = 64$.

4.3.1 Discussion

D'aprés la Table 4.4 on peut observer que le meilleur résultat au sens de minimisation de ISE^0 est obtenu avec le noyau associé discret triangulaire (a=4). La valeur de ISE^0 diminué avec l'augmentation de bras a de noyau triangulaire et augmente avec l'augmentation de paramètre c pour le noyau dirac uniforme discret. Et d'aprés la figure 4.3 on observe que, le meilleur ajustement est obtenu par le noyau dirac uniforme (c=2).

4.4 Application 3

Pour garder les cryptomonnaies, on doit avoir des portefeuilles numériques, le contenu de ces portefeuilles reste toujours visible par tout le monde, mais seulement leurs propriétaires peuvent le changer. Les investisseurs s'intéressent des mouvements et de la possession des cryptomonnaies avant d'engager et comme on peut voir toujours ce contenu on peut déduire le volume de ces mouvements et de possession de ces cryptomonnaies.

Pour ça, on a construit un échantillon composé de 300 portefeuilles regroupés selon la possession de BTC entre 100 BTC à 1500 BTC avec une pas de (100 BTC) (Voir le site BitcoinExchangeFlows). On a les regroupés comme ça, car le BTC a une valeure bourcierre plus de 50 % du marché des cryptomonnées. La Table suivante 4.5 fournit le résumé statistique de regroupement des portefeuilles selon la possession de BTC (1 unité = 100 BTC)

1	41	6	17	11	5
2	64	7	15	12	9
3	20	8	9	13	8
4	22	9	11	14	8
5	30	10	35	15	6

TABLE 4.5 – Données de regroupement des portefeuilles selon la possession de BTC (1 unité = 100 BTC).

Les résultats de (ISE^0 et h_{CV}) sont mentionnés dans la Table 4.6 pour la taille de l'échantillon $n = 300$ pour les différentes noyaux discrètes. (*Note : La valeur en gras indique la meilleur résultat*).

Noyau	bino	triang a=3	triang a=4	dirDU c=2	dirDU c=5
h_{cv}	0.03781	0.07	0.07	0.07	0.07
ISE^0	**0.03505**	0.04349	0.04113	0.03797	0.04590

TABLE 4.6 – Résultats pour la troixième application.

• L'illustration graphique suivante 4.4 présente les résultats graphiques de différents lissages discrets par les noyaux discrets naif, standards et de dexième ordre de la distribution empirique pour la taille d'échantillon *n=300*.

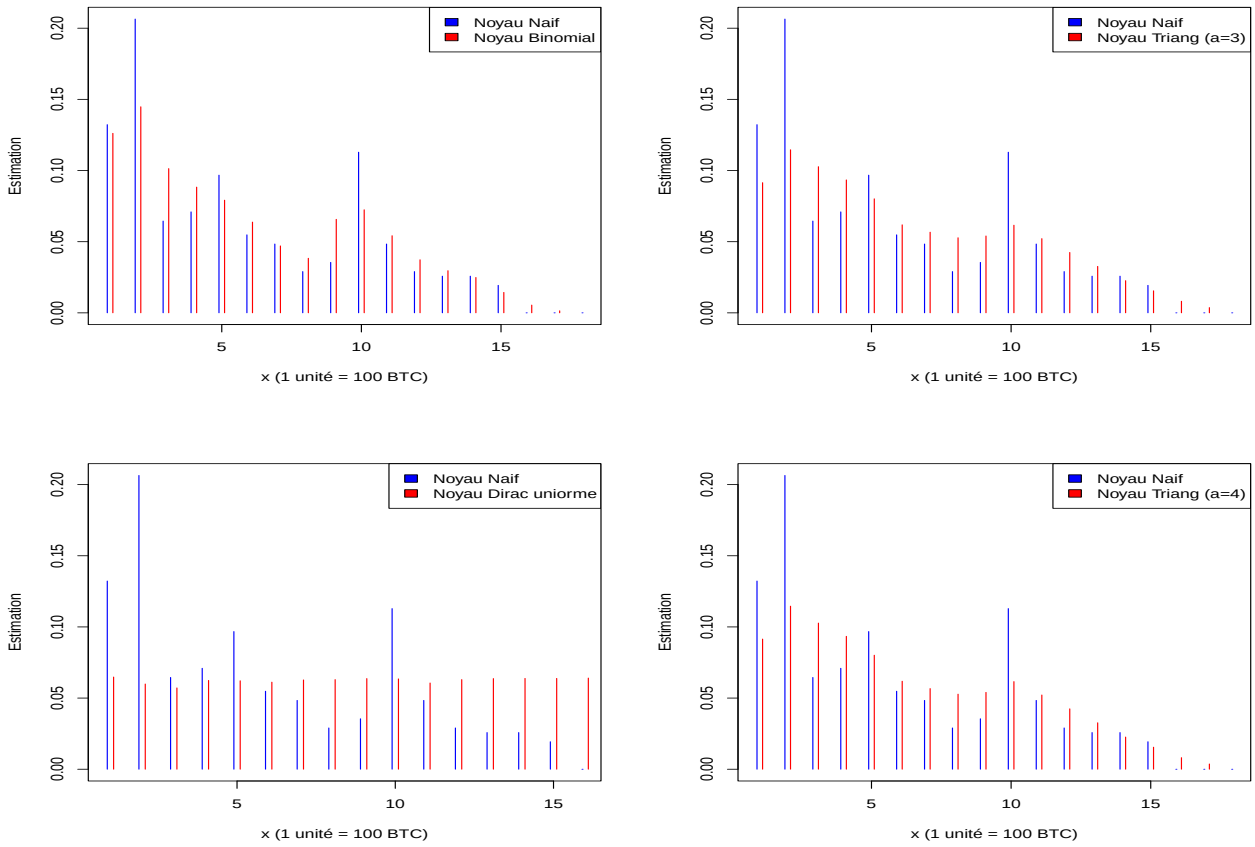

FIGURE 4.4 – Lissages discrets par les noyaux de type de dirac, binomial , triangulaire (a=3 et 4) et dirac uniforme (c=2) pour les données réelles sur le regroupement des portefeuilles selon la possession de BTC (1 unité = 100 BTC).

4.4.1 Discussion

D'aprés la Table 4.6 on peut observer que le meilleur résultat au sens de minimisation de ISE^0 est obtenu avec le noyau associé discret binomiale. La valeur de ISE^0 augmente avec l'augmentation de paramètre c pour le noyau dirac uniforme discret. Et d'aprés la figure 4.4 on observe que, le meilleur ajustement est obtenu par le noyau binomiale.

4.5 Conclusion

Dans ce dernier chapitre, nous avons discuté la performance d'utilisation des différents noyaux discrets dans l'estimation non paramétrique de la fonction de masse de probabilité sur des données réelles on utilisant le critère (ISE^0), ainsi que la méthode de validation caroisée pour le choix de paramètre de lissage. D'aprés notre étude sur les trois applications, nous avons constaté que le noyau Binomial est approprié pour les données de comptage avec des échantillons de petite ou moyenne taille, le noyau Triangulaire discret est recommandé pour les données de comptage avec des échantillons de grande taille et que le noyau Dirac uniforme discret est performant pour les données catégorielles.

Conclusion générale

Ce mémoire est une étude comparative de l'estimation de la fonction de masse de probabilité par différents noyaux discrets. L'avantage de cette méthode est qu'elle possède de bonnes propriétés asymptotiques.

Dans la première partie de ce mémoire, nous avons cité un rappel sur l'estimation de la densité à noyau symétrique (univarié) et les différentes méthodes de sélection du paramètre de lissage ainsi les différentes propriétés relatives à cet estimateur.

Dans la deuxième partie, nous avons présenté la méthode d'estimation de la fonction de densité par noyau associé dans le cas discret. Cette notion a été introduite par Kokonendji et al. [2007b] et Kiesse [2008], en donnant la définition unifiée d'un noyau associé $K_{x,h}$ de cible x et du paramètre de lissage h. Nous avons indiqué le principe de cette méthode ainsi que les différentes propriétés statistiques associées et deux méthodes pour estimé le paramètre de lissage.

La dernière partie est consacrée à l'application de cette méthode d'estimation de la densité discrète à savoir l'estimation de la loi de Poisson, d'une loi Binomial, la loi Géométrique et un mélange de Poisson et Géométrique, par les différents noyaux discrets tels que le noyau Binomial, noyau Dirac uniforme discret (c=2 et 5) et le noyau Triangulaire (a=2 et 4), et la méthode de validation croisée pour le choix de paramètre de lissage. Le but principal de cette partie réside dans la comparaison de ces dernières tout en variant la taille d'échantillon selon le critère de minimisation d'erreur quadratique intégrée (ISE) pour des données simulées et réelles.

D'aprés notre étude nous avons constaté que le noyau Binomial est approprié pour les données de comptage avec des échantillons de petite ou moyenne taille, le noyau Triangulaire discret est recommandé pour les données de comptage avec des échantillons de grande taille, le noyau Dirac uniforme discret est performant pour les données catégorielles.

Bibliographie

B. Abdous and C. C. Kokonendji. Consistency and asymptotic normality for discrete associated-kernel estimator. *African Diaspora Journal of Mathematics*, 8(2) :63–70, 2009.

J. Aitchison and C. G. Aitken. Multivariate binary discrimination by the kernel method. *Biometrika*, 63(3) :413–420, 1976.

BitcoinExchangeFlows. Bitcoin exchange flows. URL https://www.cryptoquant.com/overview/btc-miner-flows.html.

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

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

S. X. Chen. Probability density function estimation using gamma kernels. *Annals of the Institute of Statistical Mathematics*, 52(3) :471–480, 2000.

CoinMarketCap. cryptocurrency prices charts and market capitalizations. URL https://www.coinmarketcap.com/~uno/abcde.html.

J. T. De Oliveira. *Estatística de densidades : resultados assintóticos*. Universidade de Lisboa. Faculdade de Ciencias de Lisboa, 1963.

R. Duin. On the choice of smoothing parameters for parzen estimators of probability density functions' ieee trans, 1976.

V. A. Epanechnikov. Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*, 14(1) :153–158, 1969a.

J. Habbema, J. Hermans, and K. Van der Broek. A stepwise discrimination program using density estimation. In *Compstat*, volume 1974, pages 100–110. Physica Verlag Vienna, 1974.

T. S. Kiesse. *Approche non-paramétrique par noyaux associés discrets des données de dénombrement*. PhD thesis, Université de Pau et des Pays de l'Adour, 2008.

C. Kokonendji and T. S. Kiessé. Estimateur à noyau discret standard pour une densité de probabilité discrète. 2006.

C. C. Kokonendji and T. S. Kiessé. Discrete associated kernels method and extensions. *Statistical Methodology*, 8(6) :497–516, 2011.

C. C. Kokonendji and S. S. Zocchi. Extensions of discrete triangular distributions and boundary bias in kernel estimation for discrete functions. *Statistics & probability letters*, 80(21-22) : 1655–1662, 2010.

C. C. Kokonendji, D. Mizere, and N. Balakrishnan. Connections of the poisson weight function to overdispersion and underdispersion. *Journal of Statistical Planning and Inference*, 138(5) : 1287–1296, 2007b.

Q. Li and J. S. Racine. *Nonparametric econometrics : theory and practice*. Princeton University Press, 2007.

L. C. Marsh and K. Mukhopadhyay. Discrete poisson kernel density estimation-with an application to wildcat coal strikes. *Applied Economics Letters*, 6(6) :393–396, 1999.

E. Nadaraya. On non-parametric estimates of density functions and regression curves. *Theory of Probability & Its Applications*, 10(1) :186–190, 1965.

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

M. Rosenblatt. Remarks on some nonparametric estimates of a density function. annals of mathematical statistics. 1956.

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

T. Senga Kiessé. On finite sample properties of nonparametric discrete asymmetric kernel estimators. *Statistics*, 51(5) :1046–1060, 2017.

B. Silverman. Nonparametric density estimation for statistic and data analysis, 1986.

A. B. Tsybakov. *Introduction à l'estimation non paramétrique*, volume 41. Springer Science & Business Media, 2003.

W. Wansouwé, C. Kokonendji, D. Kolyang, and M. W. Wansouwé. Package 'disake'. 2015.

Résumé

Ce mémoire traite le choix du noyau pour estimer la densité de probabilité discrète par la méthode du noyau associé discret. La méthode d'estimation à noyau de la densité de probabilité est une technique très importante dans l'analyse statistique des données. Cette estimation par la méthode du noyau à partir d'un échantillon nécessite le choix du noyau discret K et du paramètre de lissage h. La performance de l'estimateur est examinée et comparée, tout en combinant les différents noyaux et des échantillons de petite, moyenne et grande taille sur des données simulées et réelles. Les résultats obtenus montrent qu'il existe une préférence selon le noyau discret et la taille des données.

Mots clés : Estimation non-paramétrique, fonction de densité, loi discrète, $MISE$, noyau associé, paramètre de lissage, ISE, validation croiséé, biais.

Abstract

This work treat the problem of the choice of the type of kernel to estimate the density of probabilities by the method of the associated kernel in the discrete case. The kernel probability density estimation method is a very important technique in statistical analysis of data. This estimation by the kernel method from a sample requires the choice of the kernel K and the smoothing parameter h. This tool has become very popular today and is used much more. This is due to its simple interpretation and its asymptotic properties. The performance of the estimator is examined and compared, while combining the different kernels and variant of small, medium and large samples on simulated and real data. The results obtained depending on whether there is a preference for the class and size of the data to be studied.

Key words : Associated kernel, bias, cross-validation, density estimation, discrete distribution, MISE, nonparametric estimation, ISE, parameter of smoothing.