République Tunisienne
Ministère de l'Enseignement Supérieur, de la Recherche Scientifique et de la Technologie
Université du 7 Novembere à Carthage

Ecole **S**upérieure de la **S**tatistique et de l'**A**nalyse de l'**I**nformation

Projet de **F**in d'**E**tude

Estimation Non-paramétrique par Noyaux Associés et Données de Panel en Marketing

---

Présenté par :
Imen Ben Khalifa

Sous la direction de :

Célestin C. Kokonendji, HDR
Université de Pau et des Pays de l'Adour
Laboratoire de Mathématiques Appliquées - UMR 5142 CNRS
E-mail : celestin.kokonendji@univ-pau.fr

Dhafer Malouche, MA
Ecole Spérieure de la Statistique et de l'Ananlyse de l'Information
E-mail : dhafer.malouche@essai.rnu.tn

Année Universitaire : 2007 - 2008

Résumé

Dans ce rapport, nous nous intéressons à la notion d'estimation non-paramétrique d'une densité (fonction de masse) inconnue sur $\aleph \subseteq \mathbb{R}$ par la méthode des noyaux associés. Pour ce faire, nous présentons d'abord une définition (unifiée) d'un noyau associé à une loi de probabilité quelconque (continue ou discrète). Nous étudions de manière détaillée quelques exemples des noyaux continus symétriques (*e.g.,* normal, Epanechnikov, etc.), continus asymétriques (*e.g.,* bêta, gamma, gaussien-inverse et gaussien-inverse-réciproque), discret catégoriel (Aitchison & Aitken, 1976) et discret de dénombrement (*e.g.,* triangulaires symétriques, standards asymétriques d'ordre 1 tels que Poisson, binomial et binomial négatif). Ensuite, nous donnons la définition de l'estimateur à noyau associé. Nous montrons la convergence ponctuelle de cet estimateur. Nous vérifions si cet estimateur est bien de masse totale égale à l'unité. D'autres propriétés (globales) sont étudiées, telles que biais, variance et erreur quadratique moyenne intégrée. Nous proposons une extension dans le cas multivarié ($\aleph \subseteq \mathbb{R}^d$) pour des fonctions de densité (fonction de masse) et de régression. Enfin, nous illustrons une partie de la méthode sur des données de panel en marketing, lesquelles données sont de comptage et parsemées.

Mots clés : noyau associé, fenêtre de lissage, biais, variance, erreur quadratique moyenne intégrée, données parsemées.

Abstract

In this report, we are interested in the notion of nonparametric estimation of an unknown density (mass function) in $\aleph \subseteq \mathbb{R}$ by using associated-kernel method. First, we present an unified definition of a kernel associated to a probability law that might be either continuous or discrete. Then, we provide a thorough treatment of some symmetric and continuous kernels (*e.g.,* Gaussian, Epanechnikov, etc.) ; asymmetric and continuous (*e.g.,* beta, gamma, inverse gaussian, reciprocal inverse gaussian) ; discrete categorical (Aitchison & Aitken, 1976) and discrete count data (*e.g.,* symmetric triangular, or some known asymmetric distributions such as : Poisson, binomial and negative binomial). Furthermore, we define the associated kernel estimators and investigate some of their finite and asymptotic properties. More precisely, we verify if the proposed estimators are bona fide densities or mass functions (i.e. functions which are simultaneously nonnegative and integrate/sum up to one). Their pointwise consistency, bias, variance and mean integrated squared error are tackled as well. Moreover, we extend these estimators to the multivariate setting $\aleph \subseteq \mathbb{R}^d$ for both density/mass and regression functions. Finally, the practical usefulness of this approach is illustrated by a case study based on some sparse count data obtained from a marketing panel research survey.

Key words : associated kernel, smoothing bandwidth, bias, variance, mean integrated squared error, sparse data.

Remerciements

Joindre l'utile à l'agréable :
Je suis extrêmement heureuse d'avoir accompli mon devoir, ça ne peut que me satisfaire davantage. La recherche du plaisir n'implique-t-elle pas la nécessité des moments difficiles pour que je puisse savoir ce que je cherche? Ce qui m'est une source de plaisir ici c'est la recherche, ce n'étais pas facile au début mais mon vrai défi était d'utiliser les moments déplaisants comme des occasions pour retrouver au plus vite mon état de joie.

Pour les moments innoubliables :
Je tiens tout d'abord à remercier Célestin C. Kokonendji d'avoir encadré ce travail avec beaucoup de compétence, d'enthousiasme et de disponibilité ; Merci Celestin pour ton acceuil chaleureux, pour tous les moyens que tu as mis à ma disposition, pour m'avoir fait confiance, d'avoir toujours su orienter mon travail, de m'avoir prodigué des conseils avisés ; enfin, je te remercie pour tes vifs encouragements et tes mots du jour.

Je suis particulièrement reconnaissante à Dhafer Malouche pour toute l'aide qu'il m'a approté cette année ; Merci Dhafer pour tout. Je n'oublie pas aussi de remercier chaleureusement tout mes enseignants.

J'exprime mes remerciements les plus vifs à Belkacem Abdous pour sa sympathie, son talent de pédagogue et sa touche exceptionnelle ; Merci Belkacem pour ta confiance et tes encouragements. Je remercie M. Herbert Castéran, professeur de marketing à l'Ecole Supérieure de Commerce de Pau, de nous avoir fourni et expliqué en détails les données.

Mes remerciements vont également à :

Tristan Senga Kiessé, pour son indéfectible soutien. Merci Tristan pour les moments du déjeuner, les longues discussions, les encouragements et la complicité.
Mariem Zouch, une personne que j'ai eu la chance de considérer comme amie aujourd'hui. Merci Mariem pour tes conseils, ta succulente cuisine et les balades photos.
Lena Griguoroscuta, Layal Lizaik, Ali Salami, Daniel Gaspar et à tous ceux que j'ai eu le plaisir de côtoyer durant quatres mois.

Merci Papa & Maman pour millions de raisons, vous me donnez jour après jour autant d'amour et de confiance.

Merci mon cher frère Amine pour tout l'humour que tu m'offres, pour les rigolades, les petites surprises et pour tes calins.

Je ne sais pas comment exprimer simplement ce que je dois à mes plus chères tantes ; Tata Hasna, Tata Najet et à mon adorable Ramoul. Mille Merci pour mille et une raisons.

Un Merci très cher pour mes supers amis Rima, Myriam & Omar.
Un grand Merci aussi pour tous ceux et celles qui m'aiment !

Table des matières

Table des figures

Liste des tableaux

Présentation générale du stage

J'ai effectué mon stage de fin d'étude au sein du laboratoire de mathématiques appliquées, département Statistique et Traitement Informatique des Données (STID). Ce département regroupe 7 enseignants et enseignantes, tous sous la direction du chef de département. Le STID offre un enseignement théorique assurant une solide formation de base ainsi qu'une initiation à la recherche.

Mon stage s'est déroulé du 1^{er} février jusqu'à 29 mai, sous la direction de M. Kokonendji. Dés mon arrivée, mon encadreur direct m'a expliqué le sujet, m'a fait visité le département, m'a accordé un bureau et m'a éclairci le régime du travail.

Tout au long de ce stage, j'ai assisté au séminaire hebdomadaire de l'équipe probabilités et statistique où j'ai eu l'occasion de présenter mon travail.

Durant ces quatres mois passés au sein du STID, j'ai eu l'opportunité de découvrir le métier du chercheur sous toute ses formes et de comprendre les difficultés qui peuvent être rencontrées.

Chapitre 1

Introduction à l'estimation non-paramétrique

L'objet principal de la statistique est de faire, à partir d'observations d'un phénomène aléatoire, une inférence au sujet de la loi générant ces observations en vue d'analyser le phénomène ou de prévoir un événement futur. Pour réduire la complexité du phénomène étudié, nous pouvons utilisé deux approches statistiques : non-paramétrique et paramétrique.

Dans le premier cas, nous considérons que l'inférence statistique doit prendre en compte la complexité autant que possible et donc cherche à estimer la distribution du phénomène dans son intégralité, mettant en oeuvre l'estimation des fonctionnelles (densités, régression, etc.). En opposition, l'approche paramétrique cherche à représenter la distribution des observations par une fonction densité $f(x|\theta)$ où le paramètre θ est la seule inconnue. Dans plusieurs cas, l'approche non paramétrique est préférable ; nous pouvons mettre en oeuvre des lois de probabilité sur des espaces fonctionnels.

Les estimateurs non-paramétriques classiques ont été introduit par Rosemblatt pour estimer des densités de probabilité, par Parzen pour estimer le mode d'une densité de probabilité et par Nadaraya & Watson pour estimer une fonction de régression. Le comportement asymptotique de ces estimateurs a été étudié par de nombreux auteurs tel que Tsybakov (2004). Ainsi, le but de ce travail est de définir les estimateurs à noyau associé et d'établir les propriétés relatives.

Avant de présenter les résultats de façon détaillée, nous en donnons tout d'abord les grandes lignes.

Dans le deuxième chapitre, nous nous intéressons à l'estimateur à noyau continu symétrique $\widehat{f_n}$ d'une densité de probabilité f inconnue sur $\mathbb{R}$. Plus précisémment, nous présentons cet estimateur et nous prouvons les propriétés fondamentales telle que biais, variance, erreur quadratique moyenne, etc. Nous donnons les méthodes de sélection du noyau ainsi de la fenêtre en s'appuyant sur des exemples de données simulées. Il s'avère que le choix du paramètre de lissage est beaucoup plus important que celui du noyau dans l'estimation des densités inconnues à support symétrique. Nous étudions également

le cas multivarié en fin de chapitre.

Dans le troisième chapitre, nous donnons la définition d'un noyau associé. Nous présentons, à partir de cette définition, les estimateurs de Chen (1999, 2000) et de Scaillet (2004) et leurs propriétés en rendant les calculs moins sombre et plus compréhensible. Toujours dans le cas du noyau associé asymétrique, nous généralisons ces résultats au cas multidimensionnel.

Dans le quatrième chapitre, nous représentons pareillement la définition du noyau associé discret en s'appuyant sur les travaux de Kokonendji & Senga Kiessé (2006). Nous définissons l'estimateur à noyau associé discret. Nous étudions les propriétés fondamentales de cet estimateur d'une manière générale, ensuite nous les appliquons dans deux sections ; la première se base sur les données discrètes catégorielles où nous allons étudier l'estimateur à noyau associé d'Aitchison & Aitken et la seconde partie repose sur les données de comptage où nous allons traiter des exemples des noyaux associés symétriques et standards asymétriques. Nous donnons un critére de choix des fenêtres de lissages. Nous généralisons cet estimateur dans une version multidimensionnelle.

Dans le cinquième chapitre, nous étudions la régression multiple à noyaux associés mixtes. Nous nous focalisons sur le fameux estimateur de Nadaraya-Watson.

Dans le sixième chapitre, nous appliquons une partie de ces estimateurs à noyaux associés sur des données parsemées de panel en marketing.

Nous terminons ce rapport par une conclusion générale et des idées de recherches futures.

Nous présentons maintenant de manière plus développée le contenu des six chapitres de ce rapport.

Chapitre 2

Noyau continu symétrique

Dans cette partie, nous présentons l'estimateur à noyau continu symétrique. Nous développons cet estimateur dans le cas univarié, ensuite nous le traitons dans le cas multivarié. Nous étudions également les différentes propriétés élémentaires relatives à cet estimateur telle que biais, variance, erreur quadratique moyenne et erreur quadratique moyenne intégrée. Nous détaillons par la suite les méthodes de choix des fenêtres et des noyaux en se focalisant sur l'importance du choix du paramètre de lissage. Nous expliquons explicitement 3 méthodes d'estimation de la fenêtre. Enfin, nous concluons par un exemple de données simulées.

2.1 Cas univarié

Considérons un échantillon de variables aléatoires X_1,X_2,...,X_n, indépendant et identiquement distribué (i.i.d.), de densité de probabilité continue inconnue f sur $\aleph = \mathbb{R}$. L'estimateur à noyau continu symétrique de f est défini par :

$$\begin{aligned} \widehat{f}_n(x) &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}\,K\left(\frac{x-X_i}{h}\right) \\ &= \widehat{f}_{n,h,K}(x), \end{aligned} \tag{2.1}$$

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 ou la fenêtre. L'expression (2.1) découle des travaux des pionniers de l'estimation non-paramétrique Rosemblatt (1956) puis Parzen (1962). Dans l'expression de l'estimateur à noyau continu (2.1), la fonction noyau K est une densité de probabilité sur $\mathbb{R} \to \mathbb{R}_+$ et est symétrique par rapport à zéro :

$$K(-x) = K(x), \tag{2.2}$$

ce qui implique l'égalité suivante

$$\int_{\mathbb{R}} tK(t)dt = 0. \tag{2.3}$$

De plus, elle est de carré intégrable

$$\int_{\mathbb{R}} K^2(t)dt < +\infty \tag{2.4}$$

et nous avons aussi la variance de K finie

$$\int_{\mathbb{R}} t^2 K(t)dt < +\infty. \tag{2.5}$$

Enfin, le noyau K peut être écrit sous plusieurs formes dont la plus connue est

$$K_h(x - X_i) = \frac{1}{h} K\left(\frac{x - X_i}{h}\right).$$

Le tableau 2.1 donne un récapitulatif des fonctions noyaux continues classiques dont les graphiques sont présentés dans la figure 2.1. Nous rappelons qu'une loi de Cauchy n'admet aucun moment fini.

TAB. 2.1 – *Exemples de noyaux continus symétriques*

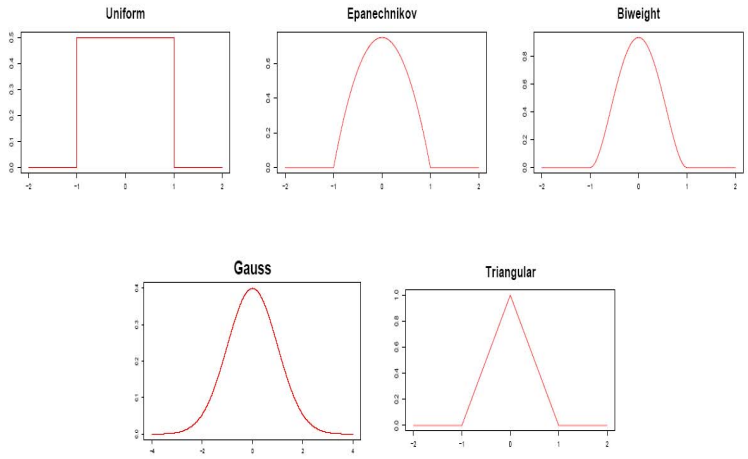
Noyau	Fonction noyau	Domaine de définition
Cauchy	$[\pi(1+u^2)]^{-1}$	$\mathbb{R}$
Biweight	$(15/16)(1-u^2)^2$	[-1,1]
Triangulaire	$1-\lvert u\rvert$	[-1,1]
Epanechnikov	$(3/4)(1-u^2)$	[-1,1]
Gaussien	$(1/\sqrt{2\pi})\exp\left(-u^2/2\right)$	$\mathbb{R}$

Pour plus de détails sur les types des noyaux, nous pouvons se référer à l'article d'Epanechnikov (1969) et le livre de Tsybakov (2004).
L'expression de K détermine la forme du noyau et h est un paramètre d'échelle qui détermine le niveau de lissage de l'estimation. Dans l'estimation à noyau continu symétrique, le choix de la fenêtre de lissage est prépondérant à celui du noyau K. De plus, la contribution de chaque point de l'échantillon est additionnée pour obtenir l'estimation totale. Ceci est illustré dans la figure 2.2.

2.1.1 Propriétés élémentaires

Dans cette partie, nous présentons les propriétés fondamentales de l'estimateur et les critères d'erreurs usuels. Nous calculons d'abord le biais et la variance de l'estimateur $\widehat{f}_n(x)$. Ensuite, nous exprimons le risque quadratique exact en un point x fixé, puis le risque intégré. Enfin, nous approximons ces résultats. Dans ce qui suit, nous supposons que les dérivées première et seconde de f existent et admettent une intégrale finie sur le support de la densité $\aleph$.

FIG. 2.1 – *Illustration des noyaux continus symétriques*



Propriété 1 : La fonction $x \longmapsto \widehat{f}_n(x)$ est une densité de probabilité.
Démonstration : La somme continue de $\widehat{f}_n(x)$ sur le support $\aleph = \mathbb{R}$ est

$$\begin{aligned}\int_{\mathbb{R}} \widehat{f}_n(x)dx &= \int_{\mathbb{R}} \sum_{i=1}^{n} \frac{1}{nh} K\left(\frac{x - X_i}{h}\right) dx \\ &= \int_{\mathbb{R}} \frac{1}{h} K\left(\frac{x - X_1}{h}\right) dx,\end{aligned}$$

en posant $t = (x - X_1)/h$ et donc $dx = hdt$, nous trouvons

$$\int_{\mathbb{R}} \widehat{f}_n(x)dx = \int_{\mathbb{R}} K(t)dt = 1.$$

De plus, le noyau K est défini positif. La somme sur tout l'échantillon reste aussi positive. Par conséquent, l'hypothèse de positivité est vérifiée.

2.1.2 Biais ponctuel

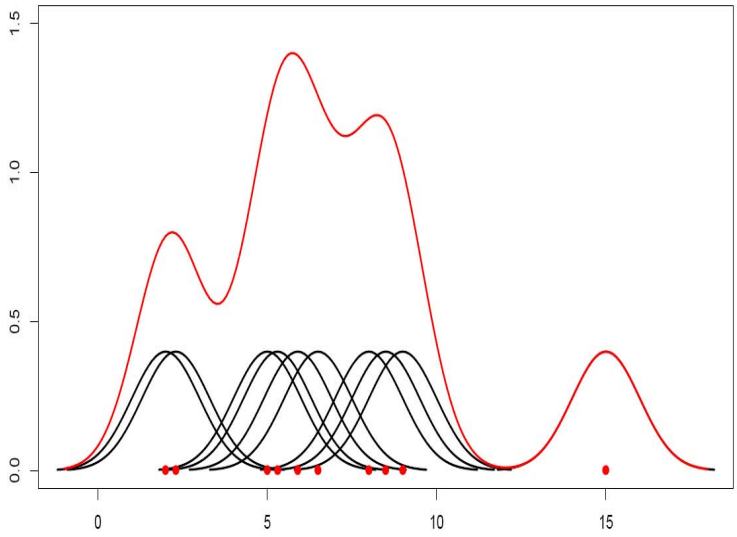
Le biais ponctuel mesure la différence entre la valeur moyenne de l'estimateur $\widehat{f}_n$ et la valeur de la fonction inconnue f en un point x.

$$Biais\left\{\widehat{f}_n(x)\right\} = \mathbb{E}\left\{\widehat{f}_n(x)\right\} - f(x). \tag{2.6}$$

Propriété 2 : Soit x fixé dans $\mathbb{R}$.
Le biais de l'estimateur à noyau présenté dans (2.1) est

$$Biais\left\{\widehat{f}_n(x)\right\} \doteq \frac{1}{2}h^2 f''(x) \int_{\mathbb{R}} t^2 K(t)\, dt. \tag{2.7}$$

FIG. 2.2 – *Estimation totale à noyau gaussien*



Le signe "$\doteq$" indique que la quantité à gauche est équivalente à la quantité à droite. Démonstration : Comme les variables aléatoires X_1,X_2,...,X_n sont i.i.d., nous avons successivement

$$\begin{aligned}
\mathbb{E}\left\{\widehat{f}_n(x)\right\} &= \mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}\,K\left(\frac{x-X_i}{h}\right)\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\{\frac{1}{h}\,K\left(\frac{x-X_i}{h}\right)\right\} \\
&= \mathbb{E}\left\{\frac{1}{h}K\left(\frac{x-X_1}{h}\right)\right\} \\
&= \int_{\mathbb{R}}\frac{1}{h}K\left(\frac{x-x_1}{h}\right)f\left(x_1\right)dx_1.
\end{aligned}$$

Nous effectuons le changement de variables suivant : $-t = (x-x_1)/h$, d'où $x_1 = ht+x$. De là, en utilisant l'hypothèse (2.2), le biais de $\widehat{f}_n(x)$ s'exprime ainsi par

$$\begin{aligned}
Biais\left\{\widehat{f}_n(x)\right\} &= \int_{\mathbb{R}}K\left(-t\right)f\left(x+ht\right)dt - f(x) \\
&= \int_{\mathbb{R}}K\left(t\right)f\left(x+ht\right)dt - f(x).
\end{aligned}$$

Dans le but d'avoir une forme plus simple, qui ne dépend que du paramètre h, nous approximons la formule du biais en utilisant la formule de Taylor-Lagrange :

$$f\left(x+ht\right) = f(x) + htf^{'}\left(x\right) + \frac{h^2t^2}{2}f^{''}\left(x\right) + o\left(h^2t^2\right).$$

Ainsi, nous obtenons

$$\begin{aligned} Biais\left\{\widehat{f}_n(x)\right\} &= f(x)\int_{\mathbb{R}} K(t)dt + hf'(x)\int_{\mathbb{R}} tK(t)dt \\ &\quad +\frac{1}{2}h^2 f''(x)\int_{\mathbb{R}} t^2K(t)dt - f(x) + o(h^2). \end{aligned}$$

D'après les hypothèses (2.3), (2.4) et (2.5) nous avons finalement

$$Biais\left\{\widehat{f}_n(x)\right\} \doteq \frac{1}{2}h^2 f''(x)\int_{\mathbb{R}} t^2K(t)\,dt.$$

2.1.3 Variance ponctuelle

Propriété 3 : Soit x fixé dans $\mathbb{R}$. La variance de l'estimateur $\widehat{f}_n$ est

$$Var\left\{\widehat{f}_n(x)\right\} \doteq \frac{1}{nh} f(x)\int_{\mathbb{R}} K(t)^2\,dt. \tag{2.8}$$

Démonstration : Partant de l'hypothèse d'indépendance entre les X_i, nous avons

$$\begin{aligned} Var\left\{\widehat{f}_n(x)\right\} &= Var\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}\,K\left(\frac{x-Xi}{h}\right)\right\} \\ &= \frac{1}{n}Var\left\{\frac{1}{h}\,K\left(\frac{x-X_1}{h}\right)\right\} \\ &= \frac{1}{n}\mathbb{E}\left[\left\{\frac{1}{h}\,K\left(\frac{x-X_1}{h}\right)\right\}^2\right] - \frac{1}{n}\left[\mathbb{E}\left\{\frac{1}{h}\,K\left(\frac{x-X_1}{h}\right)\right\}\right]^2 \\ &= \frac{1}{n}\int_{\mathbb{R}}\frac{1}{h^2}K^2\left(\frac{x-x_1}{h}\right)f(x_1)\,dx_1 \\ &\quad -\frac{1}{n}\left\{\int_{\mathbb{R}}\frac{1}{h}K\left(\frac{x-x_1}{h}\right)f(x_1)\,dx_1\right\}^2. \end{aligned}$$

Nous effectuons le changement de variable $-t = (x-x_1)/h$. Nous trouvons

$$\begin{aligned} Var\left\{\widehat{f}_n(x)\right\} &= \frac{1}{nh^2}\int_{\mathbb{R}} K(-t)^2 f(ht+x)\,hdt - \frac{1}{n}\left\{\int_{\mathbb{R}} K(-t)\,f(ht+x)\,hdt\right\}^2 \\ &= \frac{1}{nh}\int_{\mathbb{R}} K(t)^2 f(ht+x)\,dt - \frac{1}{n}\left[Biais\left\{\widehat{f}_n(x)\right\} + f(x)\right]^2 \\ &= \frac{1}{nh}\int_{\mathbb{R}} K(t)^2 f(ht+x)\,dt - \frac{1}{n}\left\{O\left(h^2\right) + f(x)\right\}^2. \end{aligned}$$

Finalement, sous la condition d'avoir $\int K(t)^2dt < +\infty$ et pour n grand, nous avons

$$Var\left\{\widehat{f}_n(x)\right\} \doteq \frac{1}{nh} f(x)\int_{\mathbb{R}} K(t)^2\,dt.$$

2.1.4 Erreur quadratique moyenne (MSE)

Propriété 4 : L'erreur quadratique moyenne (en anglais "Mean squared Error") en un point x fixé s'exprime par

$$MSE\left(x\right) = Var\left\{\widehat{f}_n(x)\right\} + Biais^2\left\{\widehat{f}_n(x)\right\} \tag{2.9}$$

Démonstration : Nous obtenons par succession

$$\begin{aligned} MSE\left(x\right) &= \mathbb{E}\left[\left\{\widehat{f}_n(x) - f\left(x\right)\right\}^2\right] \\ &= \mathbb{E}\left(\left[\widehat{f}_n(x) - \mathbb{E}\left\{\widehat{f}_n(x)\right\} + \mathbb{E}\left\{\widehat{f}_n(x)\right\} - f(x)\right]^2\right) \\ &= \mathbb{E}\left(\left[\widehat{f}_n(x) - \mathbb{E}\left\{\widehat{f}_n(x)\right\}\right]^2\right) + 2\mathbb{E}\left[\widehat{f}_n(x) - \mathbb{E}\left\{\widehat{f}_n(x)\right\}\right] \\ &\quad \left[\mathbb{E}\left\{\widehat{f}_n(x)\right\} - f(x)\right] + \left[\mathbb{E}\left\{\widehat{f}_n(x)\right\} - f(x)\right]^2 \\ &= Var\left\{\widehat{f}_n(x)\right\} + Biais^2\left\{\widehat{f}_n(x)\right\} \\ &= MSE(x;n,h,K,f). \end{aligned}$$

D'après les résultats (2.7) et (2.8), l'approximation du critère MSE en un point x fixé est

$$AMSE\left(x\right) = \frac{1}{nh} f\left(x\right) \int_{\mathbb{R}} K\left(t\right)^2 dt + \left\{\frac{1}{2} h^2 f''\left(x\right) \int_{\mathbb{R}} t^2 K\left(t\right) dt\right\}^2. \tag{2.10}$$

2.1.5 Erreur quadratique moyenne intégrée (MISE)

Propriété 5 : L'erreur quadratique moyenne intégrée (en anglais "Mean Integrated Squared Error ") est la mesure théorique commune la plus utilisée pour évaluer l'erreur entre la fonction f et $\widehat{f}_n$. Nous avons étudié dans la partie précédente le comportement de $\widehat{f}_n(x)$ en un point fixe. Il est également convenable d'évaluer l'erreur globale sur le support $\mathbb{R}$ de cet estimateur.

$$\begin{aligned} MISE(n,h,K,f) &= \int_{\mathbb{R}} MSE(x)dx \qquad (2.11) \\ &= \int_{\mathbb{R}} Var\left\{\widehat{f}_n(x)\right\} dx + \int_{\mathbb{R}} Biais^2\left\{\widehat{f}_n(x)\right\} dx. \end{aligned}$$

En utilisant l'expression approchée du critère MSE (2.10), nous avons successivement

$$\begin{aligned} AMISE(n,h,K,f) &= \frac{1}{nh}\int_{\mathbb{R}} K(t)^2 dt \int_{\mathbb{R}} f(x)dx \\ &\quad + \frac{1}{4}h^4 \left\{\int_{\mathbb{R}} t^2 K(t)dt\right\}^2 \int_{\mathbb{R}} f''(x)^2 dx \\ &= \frac{1}{nh}\int_{\mathbb{R}} K(t)^2 dt + \frac{1}{4}h^4 \left\{\int_{\mathbb{R}} t^2 K(t)dt\right\}^2 \int_{\mathbb{R}} f''(x)^2 dx \\ &= \frac{1}{nh}\int_{\mathbb{R}} K(t)^2 dt + \frac{1}{4}h^4 V(K)^2 \int_{\mathbb{R}} f''(x)^2 dx, \qquad (2.12) \end{aligned}$$

avec

$$V(K) = \int_{\mathbb{R}} t^2 K(t)dt = Var(K).$$

2.1.6 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 présenté dans le tableau 2.1, nous utilisons une mesure commune qui consiste à calculer le rapport du critère AMISE des deux noyaux mis en évidence.

$$eff(K_1,K_2) \quad = \quad \frac{AMISE(K_1)}{AMISE(K_2)}$$

Nous supposons que K_1 est le noyau d'Epanechnikov. Ce noyau est considéré comme une référence par rapport à tous les autres noyaux continus classiques. Il est largement apprécié pour ses performances (au sens où sa forme répond bien à la plupart des questions soulevées par le problème de l'estimation non paramétrique de densité) et il est considéré comme optimal au sens des mesures d'erreur. Il offre la valeur d'éfficacité maximale. Nous nous sommes appuis sur les travaux de Tsybakov (2004). Ainsi, après avoir fait les calculs nécessaires, l'efficacité d'un noyau K para rapport au noyau d'Epanechnikov se mesure par

$$eff(K) = \frac{3}{5\sqrt{5}} \frac{1}{\sqrt{\int_{\mathbb{R}} t^2 K(t)dt} \int_{\mathbb{R}} K(t)^2 dt} \leq 1.$$

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 le tableau récapitulatif (Tab. 2.2) qui présente la valeur d'efficacité des différents noyaux continus symétriques.

TAB. 2.2 – *Efficacité des noyaux continus symétriques*

Noyau	Efficacité
Epanechnikov	1.000
Biweight	0.994
Triangular	0.986
Normal	0.951
Uniform	0.930

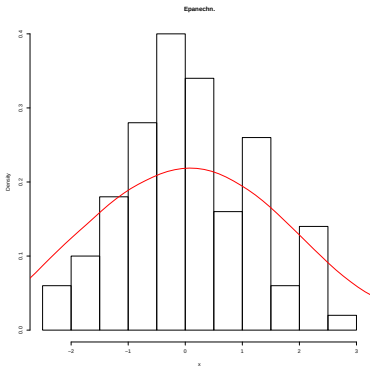
Commentaire : Dans le cas des noyaux continus symétriques, nous remarquons que les valeurs d'efficacité des noyaux tels que le noyau biweight, triangulaire ou Epanechnikov sont très proches. Par conséquent, Le choix du noyau n'est pas très important.

2.1.7 Choix de fenêtres

a. Importance du choix de h

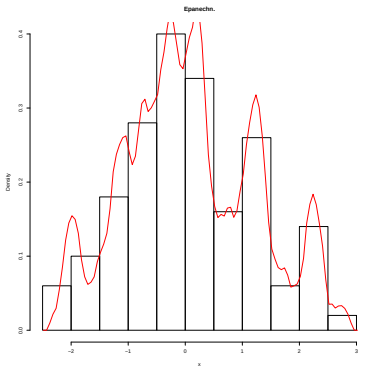
Le paramètre de lissage h est un réel positif dont le choix est prépondérant sur celui du noyau continu symétrique K. Le choix d'une valeur de h trop grande conduit à une courbe trop lisse. La courbe estimée ne traduit pas suffisament les variations de la vraie distribution (voir figure 2.3).

FIG. 2.3 – *Illustration d'un phénomène de sous-lissage lors de l'estimation d'une densité*



Par contre, en choisissant un paramètre de lissage très petit que celui adopté précédemment, l'allure de la distribution change. Il s'agit d'une distribution surestimé (figure 2.4).

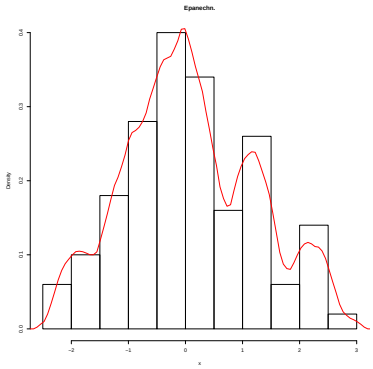
FIG. 2.4 – *Illustration d'un phénomène de sur-lissage lors de l'estimation d'une densité*



L'estimation de densité nécessite également le choix adéquat de la fenêtre h, et pour cette valeur idéale du paramètre de lissage, nous obtenons une allure qui suit parfaite-

ment la distribution de départ (figure 2.5). Les courbes obtenues illustrent à quel point

FIG. 2.5 – *Illustration d'une estimation idéale*



les formes estimées sont différentes en fonction de l'ordre de grandeur du paramètre de lissage. La principale difficulté repose sur le choix optimal de la fenêtre h. La valeur idéale h_{id} du paramètre h est celle qui minimise l'erreur quadratique moyenne intégrée (MISE). Pour une taille d'échantillon n donnée et un noyau K fixé, nous avons

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

Ce qui est équivalent à

$$h^3 V(K)^2 \int_{\mathbb{R}} f''(x)^2 dx - \frac{1}{nh^2} \int_{\mathbb{R}} K(t)^2 dt = 0.$$

Ainsi, nous obtenons successivement

$$\begin{aligned} nh^5 V(K)^2 \int_{\mathbb{R}} f''(x)^2 dx &= \int_{\mathbb{R}} K(t)^2 dt \\ h^5 &= \frac{\int_{\mathbb{R}} K(t)^2 dt}{nV(K)^2 \int_{\mathbb{R}} f''(x)^2 dx} \\ h_{id} &= \frac{1}{\sqrt[5]{n}} \left\{ \frac{\int_{\mathbb{R}} K(t)^2 dt}{V(K)^2 \int_{\mathbb{R}} f''(x)^2 dx} \right\}^{1/5}. \end{aligned} \tag{2.13}$$

En particulier pour $K = K_{Epanechn.}$, nous avons

$$h_{id}(K_{Epanechn.}) = \left(\frac{15}{n \int_{\mathbb{R}} f''(x)^2 dx} \right)^{1/5}.$$

En définitive, à partir de (2.13), nous obtenons

$$\begin{aligned} AMISE(h_{id}) &= \frac{5}{4} \frac{1}{n^{4/5}} \left\{ \int_{\mathbb{R}} t^2 K(t) dt \right\}^{2/5} \left\{ \int_{\mathbb{R}} K(t)^2 dt \right\}^{4/5} \left\{ \int_{\mathbb{R}} f''(x)^2 dx \right\}^{1/5} \\ &= \frac{5}{4n^{4/5}} I(K) \left\{ \int_{\mathbb{R}} f''(x)^2 dx \right\}^{1/5}, \end{aligned}$$

avec

$$I(K) = \left\{ \int_{\mathbb{R}} t^2 K(t) dt \right\}^{2/5} \left\{ \int_{\mathbb{R}} K(t)^2 dt \right\}^{4/5}.$$

Conséquences : Quand n est grand, h_{id} tend vers 0. Le paramètre de lissage h idéal dépend en fait de la densité à travers f''. Ainsi pour un h petit, nous avons un petit biais et une variance plus grande. Le noyau optimal est obtenu en minimisant $\int_{\mathbb{R}} K(t)^2 dt$, ceci en admettant les hypothèses (2.4) et (2.5).

b. Méthodes de choix de fenêtres

Nous considérons donc avec plus d'intérêt la question de selection du paramètre de lissage h. Comme fenêtre optimale, nous choisissons la valeur qui minimise le $MISE$. Nous étudions trois méthodes dans la détermination du paramètre de lissage optimal h_{opt} : le "Plug-in", la validation croisée par moindres carrés et la validation croisée par maximum de vraisemblance.

b.1. Méthode Plug-in

Dans la procédure de Plug-in, l'idée de base est d'estimer dans l'expression (2.13) la quantité inconnue : $\int_{\mathbb{R}} f''(x)^2 dx$. En effet, il y a deux approches possibles pour le faire : soit nous supposons que la densité f appartient à une famille de distributions paramétriques et là nous estimons les paramètres et nous retrouvons facilement cette quantité, soit nous l'estimons par l'approche non-paramétrique et donc faire appel à un estimateur à noyau (par exemple). Ceci va compliquer davantage les calculs parceque nous trouvons une fonction qui dépend elle même de h. Donc, en gros, la méthode Plug-in réside à "injecter" une estimation de f en adoptant une méthode commode et pratique. Dans notre étude, nous supposons que $f(x)$ appartient à une famille de distribution normale centrée et de variance σ^2.
Sous cette hypothèse :

$$\int_{\mathbb{R}} f''(x)^2 dx = \frac{3}{8\sqrt{\pi}} \sigma^{-1/5} \approx 0.212 \sigma^{-1/5}.$$

Il reste alors à remplacer le paramètre inconnu σ par la valeur estimée $\widehat{\sigma}$. Nous choisissons la valeur empirique comme valeur optimale définie comme suit :

$$\widehat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2},$$

tel que $\bar{X} = n^{-1} \left(X_1 + X_2 + \ldots + X_n\right)$.
Le résultat obtenu sera remplacé dans la formule de h_{id} et nous avons

$$\begin{aligned} h_{opt} &= (4\pi)^{-1/10} \left(\frac{1}{n^{1/5}}\right) \left(\frac{3}{8}\pi^{-1/2}\right)^{-1/5} \widehat{\sigma} \\ &= \left(\frac{4\widehat{\sigma}^5}{3n}\right)^{1/5} \\ &= 1.06 \left(\frac{\widehat{\sigma}}{n^{1/5}}\right). \end{aligned}$$

Ce que nous avons accompli en travaillant sous la supposition de la normalité est une formule explicite applicable pour la selection de la fenêtre h. En réalité, cette méthode donne des résultats raisonnables pour toute les distributions symétriques, unimodales et ne possédant pas des queues trop lourdes. Le problème donc avec cette méthode est qu'elle est très sensible aux valeurs aberrantes. Un estimateur plus robuste dans ce cas est obtenu à partir de l'intervalle interquartile : $R = X_{[0.75n]} - X_{[0.25n]}$ où X_p désigne le quantile d'ordre p d'une $N\left(\mu,\sigma^2\right)$. La différence entre ces deux quartiles donne 50% de l'ensemble des observations. En supposant toujours que X suit une normale $N(\mu,\sigma^2)$, nous posons $Z = (X-\mu)/\sigma$ qui suit une $N(0,1)$. Ainsi, nous montrons que $\left(X_{[0.75n]} - X_{[0.25n]}\right) = 1.34\sigma$ Par conséquent, un estimateur puissant de σ serait $\widehat{\sigma} = R/(1.34)$. Dans ce cas, le paramètre de lissage optimal est donné par

$$h_{opt} = 1.06\left(\frac{R}{1.34}\right)n^{-1/5} \approx 0.79\widehat{\sigma}n^{-1/5}.$$

Enfin, la fenêtre optimale est

$$h_{opt} = 1.06\min\left\{\widehat{\sigma},\frac{R}{1.34}\right\}n^{-1/5}.$$

Cette méthode présente des inconvénients : si la vraie densité f devie substantiellement de la forme d'une distribution normale (en étant multimodal par exemple) nous pouvons être trompés considérablement et nous aurons soit un sur-lissage soit un sous-lissage.

b.2. Méthode de validation croisée par moindres carrés

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. Nous allons essayer de remplacer la $MISE$ 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 $MISE(h)$. Pour cela, notons que :

$$\begin{aligned} MISE(h) &= \mathbb{E}\int_{\mathbb{R}}\left\{\widehat{f}_n(x) - f(x)\right\}^2 dx \\ &= \mathbb{E}\int_{\mathbb{R}}\widehat{f}_n(x)^2dx - 2\mathbb{E}\int_{\mathbb{R}}\widehat{f}_n(x)f(x)dx + \int_{\mathbb{R}}f^2(x)dx. \end{aligned}$$

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

$$J(h) = \mathbb{E}\int_{\mathbb{R}}\widehat{f}_n(x)^2dx - 2\mathbb{E}\int_{\mathbb{R}}\widehat{f}_n(x)f(x)dx.$$

Pour cela, nous déterminons un estimateur des deux termes de $J(h)$. Le premier terme admet l'estimateur $\int_{\mathbb{R}}\widehat{f}_n(x)^2dx$ comme estimateur trivial (d'après la propriété des estimateurs sans biais : $\mathbb{E}(\widehat{\beta}) = \beta$).
Il reste à trouver un estimateur sans biais du second terme. Pour cela, nous admettons par construction l'estimateur sans biais $\widehat{G}$ défini en tout points du support sauf en X_i :

$$\widehat{G} = \frac{1}{n}\sum_{i=1}^{n}\widehat{f}_{n,-i}(X_i),$$

avec

$$\widehat{f}_{n,-i}(x) = \frac{1}{n-1}\frac{1}{h}\sum_{i\neq j} K\left(\frac{x-Xi}{h}\right).$$

Montrons que $\mathbb{E}(\widehat{G}) = \mathbb{E}\left\{\int_{\mathbb{R}} \widehat{f}_n(x)f(x)dx\right\}$.
Comme les X_i sont i.i.d., d'une part nous avons

$$\begin{aligned}
\mathbb{E}\left\{\int_{\mathbb{R}} \widehat{f}_n(x)f(x)dx\right\} &= \mathbb{E}\int_{\mathbb{R}} \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) f(x)dx \\
&= \frac{1}{h}\mathbb{E}\int_{\mathbb{R}} K\left(\frac{x-X_1}{h}\right) f(x)dx \\
&= \frac{1}{h}\int_{\mathbb{R}} f(x)\int_{\mathbb{R}} K\left(\frac{x-x_1}{h}\right) f(x_1)dx_1dx.
\end{aligned}$$

D'autre part, nous avons

$$\begin{aligned}
\mathbb{E}(\widehat{G}) &= \mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\widehat{f}_{n,-i}(X_i)\right\} \\
&= \mathbb{E}\left\{\widehat{f}_{n,-1}(X_1)\right\} \\
&= \mathbb{E}\left\{\frac{1}{(n-1)h}\sum_{j\neq 1} K\left(\frac{X_j-X_1}{h}\right)\right\} \\
&= \mathbb{E}\left\{\frac{1}{h}K\left(\frac{X-X_1}{h}\right)\right\} \\
&= \frac{1}{h}\int_{\mathbb{R}} f(x)\int_{\mathbb{R}} K\left(\frac{x-x_1}{h}\right) f(x_1)dx_1dx \\
&= \mathbb{E}\int_{\mathbb{R}} \widehat{f}_n(x)f(x)dx.
\end{aligned}$$

Donc, $\widehat{G}$ est un estimateur sans biais de $\int_{\mathbb{R}} \widehat{f}_n(x)f(x)dx$. Finalement, l'estimateur sans biais de $J(h)$ est donné par

$$CV(h) = \int_{\mathbb{R}} \widehat{f}_n(x)^2dx - \frac{2}{n}\sum_{i=1}^{n}\widehat{f}_{n,-i}(X_i).$$

Et la fenêtre optimale est telle que :

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

b.3. 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 $\widehat{f}_n$ s'écrit :

$$\begin{aligned} D(f,\widehat{f}_n) &= \int_{\mathbb{R}} f(x) \log \left\{ \frac{f(x)}{\widehat{f}_n(x)} \right\} dx \\ &= \int_{\mathbb{R}} f(x) \log f(x) dx - \int_{\mathbb{R}} f(x) \log \left\{ \widehat{f}_n(x) \right\} dx \\ &= \mathbb{E}\left[\log\left\{f(X)\right\}\right] - \mathbb{E}\left[\log\left\{\widehat{f}_n(X)\right\}\right]. \end{aligned}$$

L'idée de la validation croisée par vraisemblance est de minimiser $D(f,\widehat{f}_n)$. 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 $D(f,\widehat{f}_n)$ revient à maximiser $\mathbb{E}\left[\log\left\{\widehat{f}_n(X)\right\}\right]$. Ainsi, la fenêtre optimale est

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

où

$$LCV(h) = \mathbb{E}\left[\log\left\{\widehat{f}_n(X)\right\}\right].$$

Par construction, nous avons l'estimateur sans biais de $LCV(h)$:

$$J_n = \frac{1}{n}\sum_{i=1}^{n} \log\left\{\widehat{f}_{n,-i}(X_i|h)\right\},$$

où

$$\widehat{f}_{n,-i}(X_i|h) = \frac{1}{(n-1)h}\sum_{i\neq j} K\left(\frac{X_i - X_j}{h}\right).$$

Montrons que $\mathbb{E}(J_n) = \mathbb{E}\left[\log\left\{\widehat{f}_n(X)\right\}\right]$.
Comme les variables aléatoires $X_1,X_2,\ldots,X_n$ sont i.i.d., d'une part nous obtenons

$$\begin{aligned} \mathbb{E}(J_n) &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \log\left\{\widehat{f}_{n,-i}(X_i|h)\right\}\right] \\ &= \mathbb{E}\left[\log\left\{\widehat{f}_{n,-1}(X_1|h)\right\}\right] \\ &= \mathbb{E}\left[\log\left\{\frac{1}{(n-1)h}\sum_{j\neq 1} K\left(\frac{X_1 - X_j}{h}\right)\right\}\right] \\ &= \mathbb{E}\left[\log\left\{\frac{1}{h}K\left(\frac{X_1 - X_2}{h}\right)\right\}\right]. \end{aligned}$$

D'autre part, nous trouvons

$$\begin{aligned} \mathbb{E}\left[\log\left\{\widehat{f}_n(X)\right\}\right] &= \mathbb{E}\left[\log\left\{\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X - X_i}{h}\right)\right\}\right] \\ &= \mathbb{E}\left[\log\left\{\frac{1}{h}K\left(\frac{X - X_1}{h}\right)\right\}\right] \\ &= \mathbb{E}(J_n). \end{aligned}$$

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

$$h_{LCV} = \arg\max_{h>0} \left[\frac{1}{n} \sum_{i=1}^{n} \log \left\{ \widehat{f}_{n,-i}(X_i|h) \right\} \right].$$

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.

2.1.8 Simulation des données

Dans cette partie, nous illustrons certains estimateurs à noyaux continus symétriques à savoir le noyau d'Epanechnikov, le noyau gaussien, le noyau biweight et le noyau triangulaire. Nous simulons un échantillon de taille $n = 100$ de la loi normale centrée et réduite. Pour chaque noyau fixé, la fenêtre optimale est choisie par les méthodes de validation croisée par moindre carrés et de Plug-in.

a. Choix de fenêtre par Plug-in

Cette méthode suppose que la densité suit une loi normale dans l'expression de la fenêtre h optimale. La valeur du paramètre de lissage est la même pour un échantillon donné. Nous obtenons pratiquement des estimations similaires pour chaque noyau continu utilisé ; ceci s'explique par le fait que les noyaux continus symétriques possèdent tous des efficacités proches l'une de l'autre (Figure 2.6).

b. Choix de fenêtre par validation croisée par maximum de vraisemblance

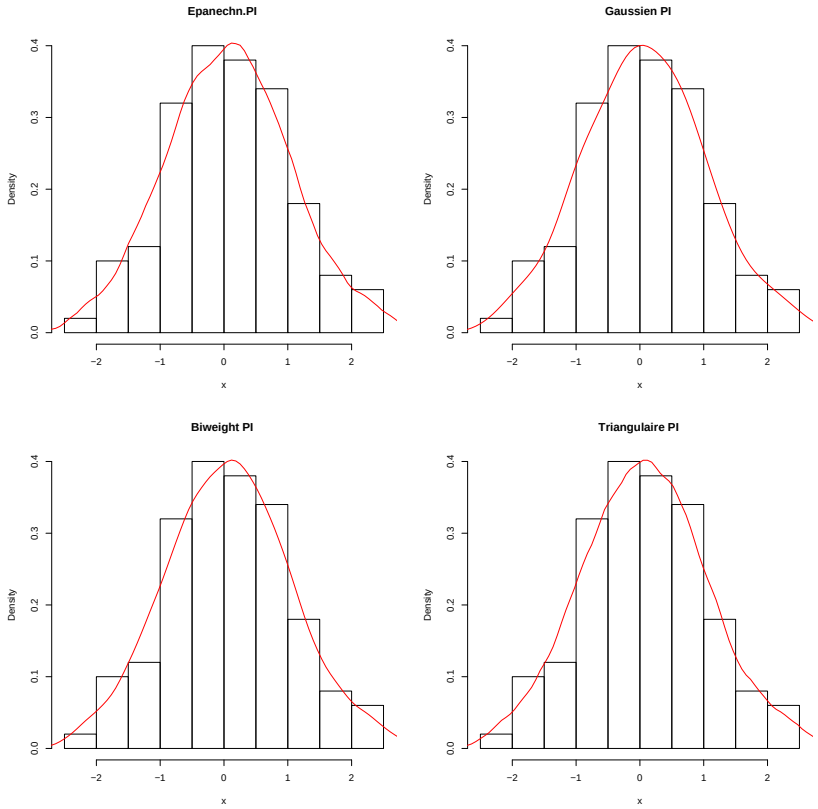
Le choix de la fenêtre optimale h_{opt} se fait en fixant au préalable le noyau continu. Le noyau par défaut dans le logiciel R est le noyau gaussien. Le choix des noyaux continus symétriques n'est pas important car ils ont quasiment les mêmes propriétés, c'est pourquoi le choix de la fenêtre optimale se fait sous l'hypothèse gaussienne (et aussi pour des raisons techniques imposées sous R). Pour chaque noyau continu symétrique fixé, la figure (2.7) présente la fenêtre optimale $h_{CV} = 0.1636$. Pour cette valeur de h, les estimations des différentes densités sont pratiquement similaires.

c. Effets de choix de fenêtres

Nous comparons différentes estimations en faisant varier la valeur de la fenêtre pour le même noyau continu. Nous choisissons le noyau optimal d'Epanechnikov.

Les simulations effectuées dans la figure (2.8) mettent en lumière que les performances pratiques des estimateurs à noyaux continus symétriques considérés dépendent fortement du choix de la fenêtre h. Par conséquent, ce choix est plus crucial que le choix

FIG. 2.6 – *Lissages par des estimateurs à noyaux continus de la distribution d'un échantillon de loi normale centrée réduite,* $n = 100$ *et* $h_{PI} = 0.338$



du noyau. Les valeurs de h sont celles choisies par plug-in ($h_{pi} = 0.338$), validation croisée par vraissemblance ($h_{CV} = 0.429$) et deux autres valeurs arbitraires tel que $h = 0.05$ et $h = 1$.

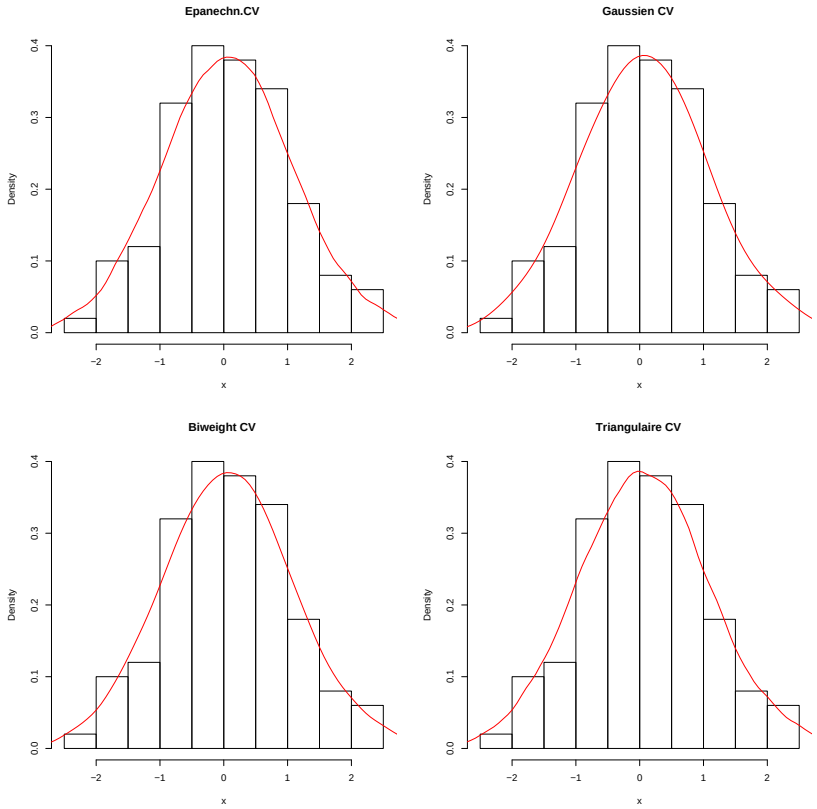
2.2 Cas multivarié

Les principales techniques d'estimation non-paramétriques de densité dans le cas génèral d'observation de dimension quelconque restent des variantes d'estimateurs à noyau. Nous pouvons choisir d'estimer toutes les composantes des observations simultanément ou selon chaque composante séparément (en faisant le produit des noyaux univariés).

Nous considérons ainsi les observations (X_{ij}) i.i.d. avec $i = 1, \ldots, n$ et $j = 1, \ldots, d$. Cette échantillon est de densité de probabilité f continue et inconnue sur $\aleph = \mathbb{R}^d$. L'estimateur à noyau continu symétrique $\widehat{f}_n$ de f admet une version multidimensionnelle et se présente de manière générale par :

$$\widehat{f}_n(\underline{x}) = \frac{1}{n} \sum_{i=1}^{n} K_H \left(X_i - \underline{x} \right). \tag{2.14}$$

FIG. 2.7 – *Lissages par des estimateurs à noyaux continus de la distribution d'un échantillon de loi normale centrée réduite,* $n = 100$ *et* $h_{CV} = 0.429$



où $\underline{x} =^t (x_1, \ldots, x_d) \in \mathbb{R}^d$, $X_i =^t (X_{i1}, \ldots, X_{id})$ et $H =$ est la matrice de variance-covariance de la fenêtre h, de dimension $d \times d$, donnée par

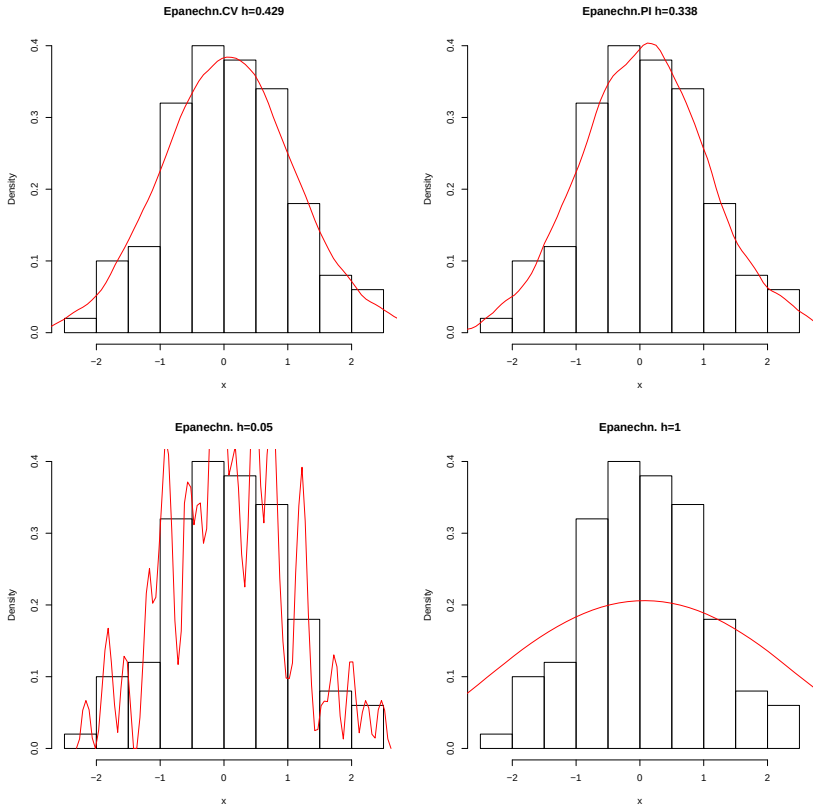
$$H = \begin{bmatrix} h_1^2 & \ldots & h_{1d} \\ h_{21} & \ldots & h_{2d} \\ \ldots & h_j^2 & \ldots \\ h_{d1} & \ldots & h_d^2 \end{bmatrix}.$$

La fonction K_H est la fonction noyau définie sur $\aleph_{x,h} = \mathbb{R}^d$ et reliée avec le noyau univarié (que nous avons présenté précédemment) par la relation suivante :

$$K_H(\underline{x}) = \{det(H)\}^{-1/2} K\left(H^{-1/2}\underline{x}\right). \tag{2.15}$$

En effet, comme nous pouvons le remarquer dans l'expression de H, il peut y avoir des termes de corrélation entre les différents paramètres de lissage. Ces coefficients de corrélation vont compliquer davantage les calculs. Nous proposons ainsi une expression plus simple qui fait appel à un produit des noyaux univariés et qui néglige l'effet des

FIG. 2.8 – *Comparaison des lissages par l'estimateur à noyau continu d'Epanechnikov en faisant varier la fenêtre h*



corrélations. Dans ce cas, l'estimateur est

$$\widehat{f}_n(\underline{x}) = \frac{1}{nh_1 \dots h_d} \sum_{i=1}^{n} \left\{ \prod_{j=1}^{d} K_j \left(\frac{X_{ij} - \underline{x}_j}{h_j} \right) \right\}, \tag{2.16}$$

avec $\underline{x} =^t (x_1, \dots, x_d) \in \mathbb{R}^d$, $h_j > 0$, $\sum_{j=1}^{d} h_j \to 0$ et $n \prod_{j=1}^{d} h_j \to \infty$ et K_j est la fonction noyau univarié présentée antérieurement. En pratique, les noyaux-produits sont recommandés. Les estimateurs à noyau généralisés sont importants pour les études numériques, mais ils restent cependant utiles pour des considérations théoriques et dans certains cas particuliers.

Note : De manière plus simple, nous prenons le noyau $K_j = K$, c'est à dire que nous utilisons ce même noyau pour toutes les observations. Cependant, nous pouvons faire un mélange de différents types de noyaux tels que le noyau d'Epanechnikov avec le gaussien, le biweight, etc.

Chapitre 3

Noyau associé continu asymétrique

Dans ce chapitre nous commençons par donner la définition d'un noyau associé continu. A partir de cette définition, nous présentons l'estimateur à noyau associé continu asymétrique dans le cas univarié puis multivarié. Nous étudions les propriétés élémentaires de cet estimateur. Différents exemples seront traités en guise de conclusion.

3.1 Cas univarié

Dans cette première partie, nous présentons l'estimateur à noyau associé continu asymétrique dans le cas univarié. Cet estimateur est approprié pour estimer des densités à support compact ou bornées d'un côté. Nous allons traiter quatres noyaux différents : gamma, bêta, gaussien inverse (IG) et gaussien inverse réciproque (RIG). Pour de récentes références nous pouvons consulter Chen (1999, 2000) et Scaillet (2004). Nous montrons les propriétés élémentaires telles que biais, variance et MISE. Ensuite nous déterminons les fenêtres optimales pour chaque noyau associé considéré et l'erreur en fonction de ces valeurs.

Soit X_1,X_2,...,X_n un échantillon de variables aléatoires i.i.d. de densité de probabilité continue inconnue f à support $\aleph = [a,b]$, avec $a \in \mathbb{R}$ et $b \in \bar{\mathbb{R}}$ ($\aleph$ est par exemple le support [0,1] ou $[0, +\infty[$). De manière génèrale, l'estimateur à noyau continu est de la forme suivante :

$$\begin{aligned} \widehat{f}_n(x) &= \frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i) \\ &= \widehat{f}_{n,h,K}(x), \end{aligned} \tag{3.1}$$

où x est fixé dans $\aleph$, $K_{x,h}$ est la fonction "noyau associé" et h est un réel strictement positif appelé paramètre de lissage.
Dans le cas où $K_{x,h}$ est associé à un noyau continu symétrique, il vérifie :

$$K_{x,h}(.) = \frac{1}{h} K\left(\frac{x - .}{h}\right).$$

Dans le cas purement asymétrique, $K_{x,h}$ est un noyau variable en fonction de la cible x (point d'estimation). Il change de forme chaque fois que x varie dans $\aleph$.

3.1.1 Définition

Définition 1 : *Soit $x \in \aleph$ et $h > 0$. Nous appelons "noyau associé continu " $K_{x,h}$ toute densité de probabilité d'une variable aléatoire $\mathcal{K}_{x,h}$ sur le support $\aleph_{x,h}$ tels que :*

$$\aleph_{x,h} \cap \aleph \neq \varnothing \tag{3.2}$$

$$\cup_x \aleph_{x,h} \supseteq \aleph \tag{3.3}$$

$$\mathbb{E}(\mathcal{K}_{x,h}) \sim x \text{ quand } h \to 0 \tag{3.4}$$

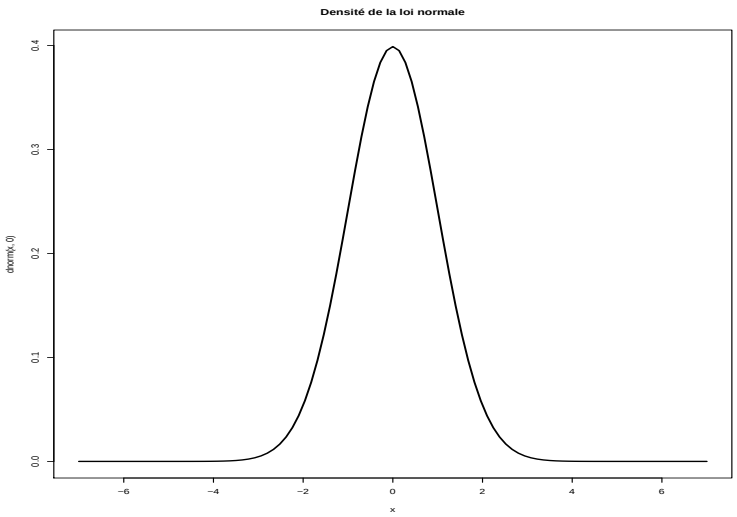
$$Var(\mathcal{K}_{x,h}) < \infty \tag{3.5}$$

$$Var(\mathcal{K}_{x,h}) \to 0 \text{ quand } h \to 0. \tag{3.6}$$

Commentaires :
a. La relation (3.2) traduit le fait que l'intersection entre le support des observations et le support du noyau associé continu asymétrique doit contenir au moins un élement. Pour un h fixé, quand x parcourt $\aleph$, le support $\aleph_{x,h}$ change, l'expression (3.3) suppose que $\aleph$ doit être toujours contenu dans la réunion des $\aleph_{x,h}$. La condition (3.4) permet d'assurer la convergence ponctuelle de l'estimateur ; elle met en évidence que le noyau $K_{x,h}$ est un noyau variable ou adaptif à la cible x. Par analogie au cas continu symétrique, la relation (3.5) n'est que la formule annoncée dans (2.5) du chapitre précédent. Enfin, la relation (3.6) assure la convergence de la variance de la variable aléatoire du noyau associé et va nous servir dans les calculs suivants.

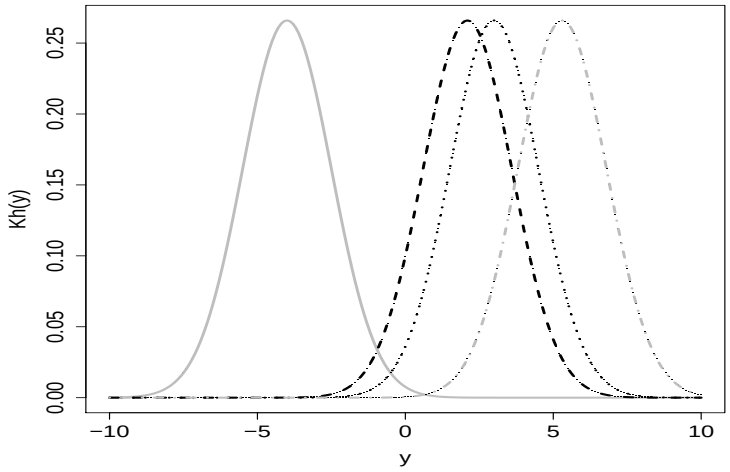
b. Avant que nous passons à l'étude des des noyaux continus asymétriques, nous reve-

FIG. 3.1 – *Densité de loi normale centrée*



nons aux noyaux symétriques pour vérifier la définition du noyau associé. Nous prenons le cas de la loi normale de moyenne μ et de variance σ^2. Nous rappelons qu'une loi nor-

FIG. 3.2 – *Illustration de la densité normale pour $h = 1.5$ et $x = y$ varié*



male $N(\mu,\sigma^2)$ est une loi continue définie sur $\aleph = \mathbb{R}$ de densité de probabilité $g_{N(\mu,\sigma^2)}$ telle que

$$g_{N(\mu,\sigma^2)}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\}.$$

Si X est une variable aléatoire qui suit la loi normale, alors l'espérance et la variance sont respectivement

$$\mathbb{E}(X) = \mu \text{ et } Var(X) = \sigma^2.$$

La figure 3.1 donne l'allure génèrale d'une densité normale centrée. Soit $K_{N(x,h^2)}$ le noyau associé à la variable aléatoire $\mathcal{K}_{N(x,h^2)}$ de loi normale $N(x,h^2)$ défini sur $\aleph_{x,h} = \mathbb{R}$. Nous vérifions ainsi chacune des hypothéses de la définition 1. En effet, la relation (3.2) se traduit par l'intersection de $\aleph = \mathbb{R}$ avec $\aleph_{x,h} = \mathbb{R}$ qui n'est que $\mathbb{R}$. En plus, d'après (3.3), la réunion sur x de $\mathbb{R}$ reste inchangée puisque le support ne dépend pas de x. A partir de (3.4), l'espérance est exactement égale à x ;

$$\mathbb{E}(\mathcal{K}_{N(x,h^2)}) = x.$$

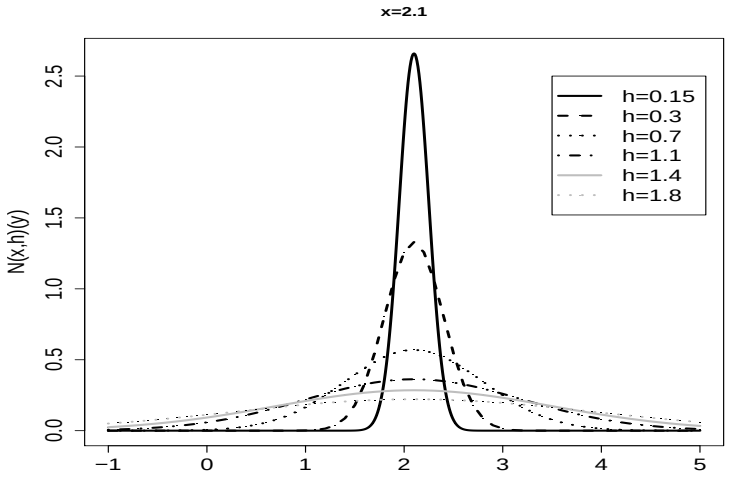
Finalement, la variance est finie et égal exactement à 0 quand $h \to 0$;

$$Var(\mathcal{K}_{N(x,h^2)}) = h^2 < \infty.$$

A ce niveau, nous donnons l'estimateur à noyau associé normal défini sur $\aleph = \mathbb{R}$. Soit $X_1,\ldots,X_n$ un échantillon de variables aléatoires i.i.d. de densité de probabilité f continue et inconnue sur $\mathbb{R}$. L'estimateur à noyau associé gaussien est

$$\begin{aligned}\widehat{f}_n(x) &= \frac{1}{n}\sum_{i=1}^{n} K_{N(x,h^2)}(X_i) \\ &= \frac{1}{nh\sqrt{2\pi}} \sum_{i=1}^{n} \exp\left\{-\frac{1}{2}\frac{(X_i-x)^2}{h^2}\right\}.\end{aligned}$$

FIG. 3.3 – *Illustration de la densité normale pour $x = 2.1$ et h varié*



Cet estimateur est-il une densité de probabilité? Oui, en effet

$$\begin{aligned}\int_{\mathbb{R}} \widehat{f}_n(x)dx &= \int_{\mathbb{R}} \frac{1}{h\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{t-x}{h}\right)^2\right\} dx \\ &\overset{(a)}{=} \int_{\mathbb{R}} \frac{1}{h\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-t}{h}\right)^2\right\} dx \\ &= 1.\end{aligned}$$

(a) : La loi gaussienne est symétrique, le fait que nous intégrons par rapport à x (la cible qui est aussi la moyenne) ou à t (la variable aléatoire) ne change rien ; nous nous permettons ainsi de permuter entre la cible x et t et nous trouvons que c'est une densité de probabilité (voir figure 3.2 et 3.3). Bien que la vérification dans le cas d'un noyau associé continu symétrique paraît simple, la question reste valable pour chacun des noyaux asymétriques.

Nous présentons maintenant les densités continus asymétriques classiques que nous allons utiliser dans la suite de cette section (Tab 3.1).
Soient a et b deux réels strictement positifs qui vérifient :

$$\Gamma(a) = \int_0^{+\infty} e^{-t} t^{a-1} dt$$

et

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1} dt.$$

3.1.2 Propriétés élémentaires

Nous donnons dans cette partie les différentes propriétés fondamentales de l'estimateur à noyau associé.

TAB. 3.1 – *Tableau récapitulatif des lois de probablité continues asymétriques*

Loi de probabilité	Support	Densité
Gamma(a,b)	$\mathbb{R}_+$	$\frac{1}{\Gamma(a)b^a}t^{a-1}\exp(-t/b)$
Bêta(a,b)	$[0,\,1]$	$\frac{1}{B(a,b)}t^{a-1}(1-t)^{b-1}$
IG(a,b)	$\mathbb{R}_+$	$\frac{\sqrt{b}}{\sqrt{2\pi t^3}}\exp\left\{-\frac{b}{2a}\left(\frac{t}{a}-2+\frac{a}{t}\right)\right\}$
RIG(a,b)	$\mathbb{R}_+$	$\frac{\sqrt{b}}{\sqrt{2\pi t}}\exp\left\{-\frac{b}{2a}\left(at-2+\frac{1}{at}\right)\right\}$

Loi de probabilité	Espérance	Variance
Gamma(a,b)	ab	ab^2
Bêta(a,b)	$a/(a+b)$	$ab/\left\{(a+b)^2(a+b+1)\right\}$
IG(a,b)	a	a^3/b
RIG(a,b)	$1/a+1/b$	$1/ab+2/b^2$

Propriétés 1 : *Soit X_1,X_2,...,X_n un échantillon de variables aléatoires i.i.d. de densité de probabilité continue inconnue f à support $\aleph$. Soit $\widehat{f}_n$ un estimateur de f à noyau associé continu asymétrique défini sur $\aleph$. Alors, la fonction $x \longmapsto \widehat{f}_n(x)$ n'est pas nécessairement une densité de probabilité sur $\aleph$. En posant*

$$c = \int_{\aleph} \widehat{f}_n(x)dx = c(h,K) \geq 0,$$

nous considérons désormais l'estimateur $\widehat{f}_n$ tel que

$$\widehat{f}_n(x) = \frac{1}{nc}\sum_{i=1}^{n} K_{x,h}(X_i). \tag{3.7}$$

Dans la suite, nous supposons que $\widehat{f}_n(x)$ est une densité de probabilité. Nous illustrons cette hypothèse dans la partie exemple.

Propriété 2 : *Soit X_1,X_2,...,X_n un échantillon de variables aléatoires i.i.d. d'une*

densité de probabilité continue inconnue f de support $\aleph$. Soit $\widehat{f}_n$ l'estimateur de f à noyau associé continu asymétrique $K_{x,h}$ de variable aléatoire $\mathcal{K}_{x,h}$ sur le support $\aleph_{x,h}$. Alors, $\forall x \in \aleph$ et $h > 0$, nous avons

$$\mathbb{E}\left\{\widehat{f}_n(x)\right\} = \mathbb{E}\left\{f(\mathcal{K}_{x,h})\right\}. \tag{3.8}$$

Démonstration : Soit $x \in \aleph$. Nous avons successivement

$$\begin{aligned}
\mathbb{E}\left\{\widehat{f}_n(x)\right\} &= \mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i)\right\} \\
&= \mathbb{E}\left\{K_{x,h}(X_1)\right\} \\
&\overset{(a)}{=} \int_{\aleph\cap\aleph_{x,h}} K_{x,h}(t)f(t)dt \\
&= \mathbb{E}\left\{f(\mathcal{K}_{x,h})\right\}.
\end{aligned}$$

(a) ; les X_i sont dans $\aleph$ et le noyau associé est défini sur $\aleph_{x,h}$. D'où l'intégrale se fait sur l'intersection des deux supports.■

Dans le but d'assurer la convergence ponctuelle de l'estimateur, nous avons adapté le lemme présenté par Hille (1948) et dont une démonstration à été donnée par Feller (1966) dans le lemme 1, page 219. Nous signalons que ce lemme était énoncé dans le travail récent de Chaubey *et al.* (2007) [1]. Ainsi, nous le formulons dans la propriété suivante.

Propriété 3 : Soient f une fonction continue et bornée sur $\aleph$ et x est fixée sur ce support. Soit $\widehat{f}_n$ l'estimateur à noyau associé continu $K_{x,h}$ sur $\aleph_{x,h}$. Nous supposons que $\forall x \in \aleph$, $\aleph_{x,h} \subseteq \aleph$. Alors nous avons

$$\mathbb{E}\left\{\widehat{f}_n(x)\right\} = \int_{\aleph\cap\aleph_{x,h}} f(t)K_{x,h}(t)dt \to f(x) \text{ quand } n \to \infty.$$

La convergence est uniforme en toute subdivision de $\aleph$ dans laquelle $Var(\mathcal{K}_{x,h}) \to 0$ quand $h \to 0$ et la fonction f est uniformément continue.

Démonstration : Nous partons de l'expression de l'estimateur dans (3.1) et nous calculons son espérance

$$\mathbb{E}\left\{\widehat{f}_n(x)\right\} = \int_{\aleph_{x,h}\cap\aleph} K_{x,h}(z)f(z)dz.$$

Comme $\aleph_{x,h} \subseteq \aleph$, nous pouvons écrire $f(x) = f(x)\int_{\aleph_{x,h}} K_{x,h}(z)dz$. Ainsi, il existe $\delta > 0$ tel que

$$\begin{aligned}
\left|\mathbb{E}\left\{\widehat{f}_n(x)\right\} - f(x)\right| &= \left|\int_{\aleph_{x,h}} \left\{f(z) - f(x)\right\} K_{x,h}(z)dz\right| \\
&\leq \int_{|z-x|<\delta} \left|f(z) - f(x)\right| K_{x,h}(z)dz + \int_{|z-x|>\delta} \left|f(z) - f(x)\right| K_{x,h}(z)dz.
\end{aligned}$$

1. Nous étions orientés dans ce résultat par le Professeur Belkacem Abdous à l'université Laval à Quebec et en visite au Laboratoire de Mathématiques Appliquées de Pau.

Pour calculer la première quantité, nous utilisons directement la définition de la continuité :

$$\forall \epsilon > 0, \exists\ \delta > 0, \forall z : |z - x| < \delta \Rightarrow |f(z) - f(x)| < \epsilon.$$

D'où nous obtenons,

$$\begin{aligned} \int_{|z-x|<\delta} |f(z) - f(x)| K_{x,h}(z)dz &\leq \epsilon \int_{|z-x|<\delta} K_{x,h}(z)dz \\ &\leq \epsilon. \end{aligned}$$

Pour calculer la deuxième quantité, nous utilisons l'inégalité de Tchebychev-Markov. Comme f est bornée $\Rightarrow \exists\ M > 0$ tel que $f \leq M$. Ainsi, nous avons

$$\begin{aligned} \int_{|z-x|>\delta} |f(z) - f(x)| K_{x,h}(z)dz &\leq 2M \int_{|z-x|>\delta} K_{x,h}(z)dz \\ &= \frac{2M}{\delta^2} \int_{|z-x|>\delta} \delta^2 K_{x,h}(z)dz \\ &\leq \frac{2M}{\delta^2} \int_{\aleph_{x,h}} (z-x)^2 K_{x,h}(z)dz \\ &= \frac{2M}{\delta^2} \mathbb{E}\left\{(\mathcal{K}_{x,h} - x)^2\right\} \\ &\overset{(a)}{=} \frac{2M}{\delta^2} Var\left(\mathcal{K}_{x,h}\right) + \frac{2M}{\delta^2} \left\{\mathbb{E}\left(\mathcal{K}_{x,h}\right) - x\right\}^2 \end{aligned}$$

(a) : nous appliquons directement la formule $\mathbb{E}(X^2) = Var(X) + \{\mathbb{E}(X)\}^2$. Or, d'après les deux hypothèses (3.4) et (3.6) du noyau associé, la dernière inégalité tend vers 0. Nous concluons enfin que

$$\mathbb{E}\left\{\widehat{f}_n(x)\right\} - f(x) \to 0 \text{ quand } n \to +\infty. \blacksquare$$

Remarque : *La propriété que nous avons présenté est valable dans le cas des noyaux continus symétriques et asymétriques.*

Propriétés 4 : *Nous présentons le développement limité de Taylor-Lagrange à l'ordre 2 et au point moyen de la variable aléatoire $\mathbb{E}(\mathcal{K}_{x,h}) = m_{x,h}$ tel que*

$$f(\mathcal{K}_{x,h}) \doteq f(m_{x,h}) + (\mathcal{K}_{x,h} - m_{x,h}) f'(x) + \frac{1}{2}(\mathcal{K}_{x,h} - m_{x,h})^2 f''(x). \tag{3.9}$$

En calculant l'espérance de cette quantité, nous obtenons :

$$\mathbb{E}\left\{f(\mathcal{K}_{x,h})\right\} \doteq f\left\{\mathbb{E}(\mathcal{K}_{x,h})\right\} + \frac{1}{2} Var(\mathcal{K}_{x,h}) f''(x). \tag{3.10}$$

3.1.3 Biais ponctuel

Pour un x fixé, nous calculons le biais de l'estimateur à noyau associé continu asymétrique de manière générale. Ce résultat sera exploité dans la partie exemple.

Propriétés 5 : *Soit x fixé dans $\aleph$. Nous avons*

$$\begin{aligned} Biais\left\{\widehat{f}_n(x)\right\} &= \mathbb{E}\left\{\widehat{f}_n(x)\right\} - f(x) \\ &\doteq [f\{\mathbb{E}(\mathcal{K}_{x,h})\} - f(x)] + \frac{1}{2}Var(\mathcal{K}_{x,h})f''(x). \end{aligned} \tag{3.11}$$

Démontration : En effet, d'après le résultat de (3.8) et les deux expressions d'approximation de Taylor-Lagrange (3.9) et (3.10), le biais s'obtient facilement en retranchant $f(x)$.■

Remarque : *Nous remarquons que le biais ne dépend pas de n et tend vers 0 quand h est très petit.*

3.1.4 Variance ponctuelle

Pour un x fixé, nous généralisons l'expression de la variance de $\widehat{f}_n$. Nous précisons que ce résultat sera utilisé dans la partie exemple.

Propriétés 6 : *Soit x fixé dans $\aleph$. Nous avons*

$$Var\left\{\widehat{f}_n(x)\right\} \doteq \frac{1}{n}\int_{\aleph_{x,h}\cap\aleph} K_{x,h}^2(t)f(t)dt - \frac{1}{n}\left[Biais\left\{\widehat{f}_n(x)\right\} + f(x)\right]^2. \tag{3.12}$$

Démonstration : Comme les X_i sont i.i.d., nous obtenons successivement

$$\begin{aligned} Var\left\{\widehat{f}_n(x)\right\} &= Var\left\{\frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i)\right\} \\ &= \frac{1}{n}\left[Var\left\{K_{x,h}(X_1)\right\}\right] \\ &= \frac{1}{n}\left[\mathbb{E}\left\{K_{x,h}(X_1)\right\}^2\right] - \frac{1}{n}\left[\mathbb{E}\left\{K_{x,h}(X_1)\right\}\right]^2 \\ &= \frac{1}{n}\left\{\int_{\aleph_{x,h}\cap\aleph} K_{x,h}^2(t)f(t)dt\right\} - \frac{1}{n}\left\{\int_{\aleph_{x,h}\cap\aleph} K_{x,h}(t)f(t)dt\right\}^2. \end{aligned}$$

Par analogie avec le noyau continu symétrique, nous avons

$$\left\{\int_{\aleph_{x,h}\cap\aleph} K_{x,h}(t)f(t)dt\right\}^2 \doteq \left[Biais\left\{\widehat{f}_n(x)\right\} + f(x)\right]^2$$

et sous la condition $\int_{\aleph_{x,h}\cap\aleph} K_{x,h}^2(t)f(t)dt$ est finie, la variance de $\widehat{f}_n$ est

$$Var\left\{\widehat{f}_n(x)\right\} \doteq \frac{1}{n}\left\{\int_{\aleph_{x,h}\cap\aleph} K_{x,h}^2(t)f(t)dt\right\} - \frac{1}{n}\left[Biais\left\{\widehat{f}_n(x)\right\} + f(x)\right]^2.\blacksquare$$

Remarque : *Nous remarquons que la variance tend vers 0 quand n tend vers $+\infty$, ceci pour tout x fixé dans $\aleph$ et $h > 0$.*

3.1.5 MISE

L'erreur globale de $\widehat{f}_n$ s'obtient en sommant le carré de l'expression (3.11) avec le resultat obtenu dans (3.12).

Propriétés 7 : *En sommant sur l'intersection des deux supports, le MISE est*

$$\begin{aligned} MISE &= \int_{\aleph_{x,h}\cap\aleph} \mathbb{E}\left\{\widehat{f}_n(x) - f(x)\right\} dx \\ &= \int_{\aleph_{x,h}\cap\aleph} Var\left\{\widehat{f}_n(x)\right\} dx + \int_{\aleph_{x,h}\cap\aleph} Biais^2\left\{\widehat{f}_n(x)\right\} dx. \end{aligned}$$

3.1.6 Exemples

Nous supposons dans toute la suite que f admet une dérivée seconde continue sur le support $\aleph$ et que les termes suivants sont finis : $\int_\aleph \left\{f'(x)\right\}^2 dx$, $\int_\aleph \left\{xf''(x)\right\}^2 dx$ et $\int_\aleph \left\{x^3 f''(x)\right\}^2 dx$.

a. Cas d'un noyau associé gamma

Chen (2000) était le premier à introduire l'estimateur à noyau asymétrique. Il présentait au début un premier estimateur $\widehat{f}_n$ à noyau gamma de paramètres $a = x/h + 1$ et $b = h$, il calculait ensuite les propriétés ponctuelles et globales liées à cet estimateur. Puis, à cause des problèmes du biais au bord qu'avait cet estimateur, Chen effectuait une légère modification au niveau des paramètres du noyau gamma pour réduire l'erreur et il représentait un deuxième estimateur que nous notons $\widehat{\widehat{f}}_n(x)$.
Nous rappelons qu'une loi gamma est une loi continue asymétrique définie sur $\aleph = \mathbb{R}_+$ de densité de probablité $g_{G(a,b)}$ telle que :

$$g_{G(a,b)}(t) = \frac{t^{a-1}e^{-t/b}}{\Gamma(a)b^a},$$

avec

$$\Gamma(a) = \int_{\mathbb{R}_+} e^{-t}t^{a-1}dt.$$

Si X une variable aléatoire qui suit la loi gamma, alors

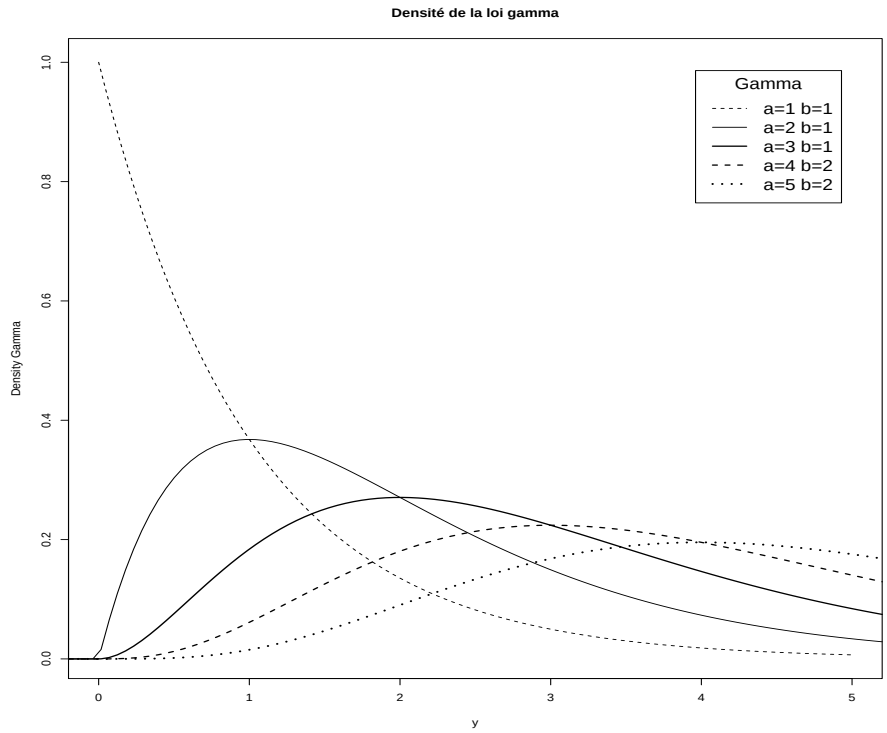
$$\mathbb{E}(X) = ab \text{ et } Var(X) = ab^2.$$

D'après la figure 3.4, nous remarquons que selon les valeurs que prennent a et b, l'allure de la courbe change. Dans le cas particulier où $a = 1$ nous retrouvons la loi exponentionnelle.

Soit $K_{G(x/h+1;h)}$ le noyau associé à la variable aléatoire $\mathcal{K}_{G(x/h+1;h)}$ de loi gamma et de support $\aleph_{x,h} = \mathbb{R}_+$. Il est donné par

$$K_{G(x/h+1;h)}(t) = \frac{1}{h^{x/h+1}\Gamma(x/h+1)}t^{x/h}e^{-t/h}.$$

FIG. 3.4 – *Allure générale d'une densité gamma*



Les deux figures 3.5 et 3.6 donnent l'allure du noyau associé gamma qui dépend des paramètres x et h. Nous donnons en premier lieu la représentation du noyau gamma pour un h fixé, nous remarquons qu'en changeant x la courbe change légérement de forme et se déplace principalement sur l'axe des abscisses. Cependant, si nous varions h comme indiqué dans le graphique 3.6, l'allure de cette densité change complétement.

Nous révisons d'abord les différentes hypothèses du noyau associé $K_{G(x/h+1;h)}$.

i. $\mathbb{R}_+ \cap \mathbb{R}_+ = \mathbb{R}_+ \neq \varnothing$.

ii. $\cup_x \mathbb{R}_+ = \mathbb{R}_+$.

iii. $\mathbb{E}(\mathcal{K}_{G(x/h+1,h)}) = (x/h+1)h = x+h \sim x$ quand $h \to 0$.

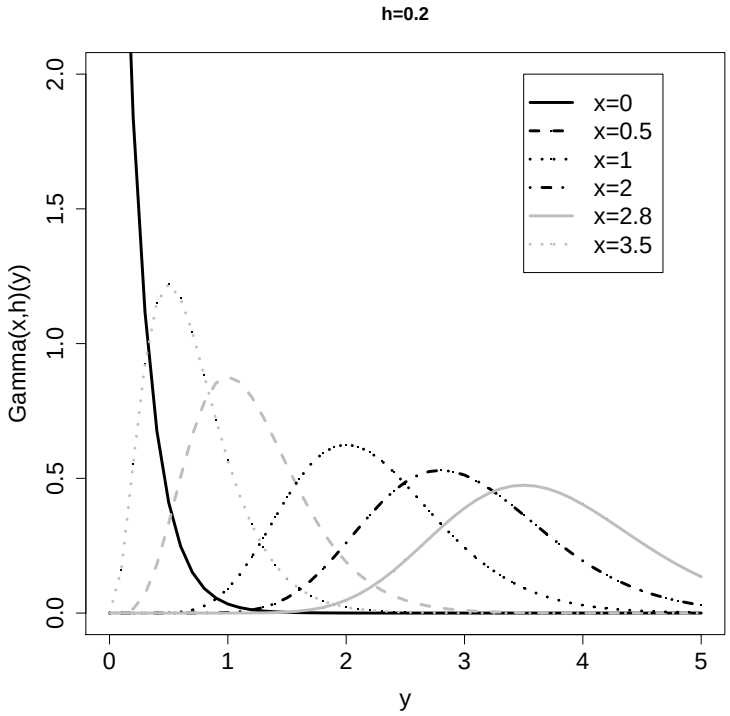
iv. $Var(\mathcal{K}_{G(x/h+1,h)}) = (x/h+1)h^2 = xh+h^2 < \infty$.

v. $h \to 0 \Rightarrow Var(\mathcal{K}_{G(x/h+1,h)}) = 0$.

Soit $X_1, X_2, \ldots, X_n$ un échantillon de variables aléatoires i.i.d. à support $\aleph = \mathbb{R}$, de densité de probabilité continue inconnue f. Nous considérons l'estimateur $\widehat{f_n}$ à noyau associé gamma tel que

$$\begin{aligned} \widehat{f_n}(x) &= \frac{1}{n}\sum_{i=1}^{n} K_{G(x/h+1;h)}(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\Gamma(x/h+1)} \frac{X_i^{x/h} e^{-X_i/h}}{h^{x/h+1}}, \end{aligned}$$

FIG. 3.5 – *Allure du noyau associé gamma pour $h = 0.2$ et x varié*



où $h > 0$ est le paramètre de lissage et K est le noyau associé à une variable aléatoire de loi gamma de paramètres $x/h + 1$ et h. D'après (3.11), nous avons

$$Biais\left\{\widehat{f}_n(x)\right\} = hf^{'}(x) + \frac{1}{2}hxf^{''}(x) + o(h). \tag{3.13}$$

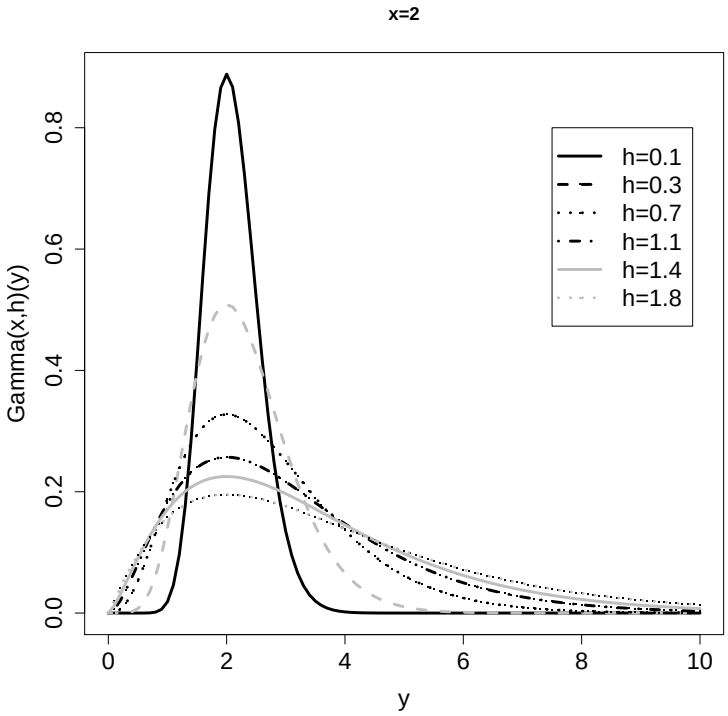
Dans le calcul du biais, nous nous arrêtons à l'ordre 1 pour avoir une homogénéité des puissances avec la variance dans le calcul de l'erreur quadratique moyenne intégrée MISE (Le biais sera élevé au carré). D'après cette expression, nous remarquons que le biais tend vers 0 quand h tend aussi vers 0. Le fait que $f^{'}$ et $f^{''}$ figurent dans la même équation, n'est pas très favorable dans le calcul du biais puisque ça augmente l'erreur. La complicité de la dérivée première avec la dérivée seconde est dûe au fait que x n'est pas la cible mais elle est plutôt le mode.
Nous calculons la variance de cet estimateur. D'après (3.12), nous avons

$$Var\left\{\widehat{f}_n(x)\right\} = \frac{1}{n}\left[\mathbb{E}\left\{K_{G(x/h+1;h)}(X_1)\right\}^2\right] - \frac{1}{n}\left[\mathbb{E}\left\{K_{G(x/h+1;h)}(X_1)\right\}\right]^2.$$

Nous calculons chacun des deux termes. En effet, nous avons

$$\mathbb{E}\left\{K_{G(x/h+1;h)}(X_1)\right\}^2 = \mathbb{E}\left(\frac{X_1^{2x/h}e^{-2X_1}/h}{h^{2(x/h+1)}\Gamma 2x/h+1}\right).$$

FIG. 3.6 – *Allure du noyau associé gamma pour $x = y = 2$ et h varié*



Soit $K_{G(2x/h+1;h)}$ un noyau associé gamma de paramètres $2x/h+1$ et h ;

$$\left\{K_{G(2x/h+1;h)}(X_1)\right\}^2 = \frac{X_1^{2x/h} e^{-2X_1}/h}{h^{2x/h+1}\Gamma(2x/h+1)}.$$

Ce qui implique

$$X_1^{2x/h} e^{-2X_1/h} = h^{2x/h+1}\Gamma(2x/h+1) K_{G(2x/h+1;h)}(X_1).$$

Ainsi, nous trouvons finalement

$$\begin{aligned}
\mathbb{E}\left\{K_{G(x/h+1;h)}(X_1)\right\}^2 &= \mathbb{E}\left\{\frac{h^{2x/h+1}}{h^{2(x/h+1)}}\frac{\Gamma(2x/h+1)}{\Gamma^2(x/h+1)} K_{G(2x/h+1;h)}(X_1)\right\} \\
&= h^{-1}\frac{\Gamma(2x/h+1)}{\Gamma^2(x/h+1)}\mathbb{E}\left\{K_{G(2x/h+1;h)}(X_1)\right\}.
\end{aligned}$$

Nous examinons les différentes conditions du noyau associé $K_{G(2x/h+1;h)}$.

i. $\mathbb{R}_+ \cap \mathbb{R}_+ = \mathbb{R}_+ \neq \varnothing$.
ii. $\cup_x \mathbb{R}_+ = \mathbb{R}_+$.
iii. $\mathbb{E}(\mathcal{K}_{G(2x/h+1;h)}) = (2x/h+1)h = 2x+h$.
iv. $Var(\mathcal{K}_{G(2x/h+1;h)}) = (2x/h+1)h^2 = 2xh + h^2 < \infty$.
v. $h \to 0 \Rightarrow Var(\mathcal{K}_{G(2x/h+1;h)}) = 0$.

Nous avons ainsi

$$\mathbb{E}\left\{K_{G(2x/h+1;h)}(X_1)\right\} = f(x) + \frac{h}{2}f^{'}(x) + o(h).$$

Soit l'expression de $A_h(x)$ telle que

$$A_h(x) = h^{-1}\frac{\Gamma(2x/h+1)}{\Gamma^2(x/h+1)}.$$

Nous considérons la fonction $R(z)$ monotone, croissante et converge vers 1 quand z tend vers l'infini (i.e: $\forall z > 0$, $R(z) < 1$). Elle est donnée par

$$R(z) = \frac{\sqrt{2\pi}}{\Gamma(z+1)}e^{-z}z^{z+1}. \tag{3.14}$$

En prenant $z = 2x/h$ et $z = x/h$, nous obtenons

$$R(2x/h) = \frac{\sqrt{2\pi}}{\Gamma(2x/h+1)}e^{-2x/h}(2x/h)^{2x/h+1/2}.$$

$$R^2(x/h) = \frac{2\pi}{\Gamma^2(x/h+1)}e^{-2x/h}(2x/h)^{2(x/h+1/2)}.$$

Ainsi, $A_h(x)$ peut être exprimée en fonction de $R(x/h)$ et $R(2x/h)$.

$$\begin{aligned} A_h(x) &= h^{-1}\frac{\sqrt{2\pi}}{2\pi}\frac{R^2(x/h)}{R(2x/h)}\frac{e^{-2x/h}}{e^{-2x/h}}\left(\frac{2x}{h}\right)^{2x/h+1/2}\left(\frac{x}{h}\right)^{-2(x/h+1/2)} \\ &= \frac{h^{1/2}}{\sqrt{2\pi}}\frac{R^2(x/h)}{R(2x/h)}x^{-1/2}2^{2x/h+1}. \end{aligned}$$

Comme $R(z) < 1$ alors $R^2(z)$ reste encore inférieur à 1. Par conséquent, le rapport $\frac{R^2(x/h)}{R(2x/h)} < 1$ et nous trouvons

$$\begin{aligned} A_h(x) &= \frac{h^{1/2}}{\sqrt{2\pi}}\frac{R^2(x/h)}{R(2x/h)}x^{-1/2}2^{2x/h+1} \\ &\leq \frac{h^{1/2}x^{-1/2}}{2\sqrt{\pi}} \\ &\leq \frac{\sqrt{h}}{2\sqrt{\pi x}}. \end{aligned}$$

Pour un h suffisamment petit,

$$A_h(x) \sim \begin{cases} \frac{1}{2\sqrt{h\pi}}x^{-1/2} & \text{si } x/h \to \infty \\ \frac{\Gamma(2k+1)}{2^{1+2k}\Gamma^2(k+1)}\frac{1}{h} & \text{si } x/h \to k, \end{cases}$$

où k est une constante positive.
Nous calculons à ce niveau le deuxième terme de la variance ;

$$\left[\mathbb{E}\left\{K_{G(x/h+1;h)}(X_1)\right\}\right]^2 = \left[\mathbb{E}\left\{\int_{\mathbb{R}} \widehat{f}_{n,h,K}(x)dx\right\}\right]^2 =^{(a)} 1.$$

(a): D'après la propriété (3.7).
En conclusion, la variance est donnée par

$$Var\left\{\widehat{f}_n(x)\right\} \sim \begin{cases} \frac{1}{2n\sqrt{h\pi}} x^{-1/2} f(x) + O(n^{-1}) & \text{si } x/h \to \infty \\ \frac{\Gamma(2k+1)}{2^{1+2k}\Gamma^2(k+1)} \frac{1}{hn} f(x) & \text{si } x/h \to k. \end{cases}$$

L'impact de la variance au bord est négligeable dans la calcul de son intégrale, nous ne tenons compte que du terme qui se trouve à l'intérieur de notre support, ceci se démontre par le calcul suivant :
Soit $\delta = h^{1-\epsilon}$, où $0 < \epsilon < 1$.

$$\begin{aligned} \int_0^\infty Var\left\{\widehat{f}_n(x)\right\} dx &= \int_0^\delta Var\left\{\widehat{f}_n(x)\right\} dx + \int_\delta^\infty Var\left\{\widehat{f}_n(x)\right\} dx \\ &= \int_\delta^\infty \frac{1}{2n\sqrt{h\pi}} x^{-1/2} f(x)dx + O(n^{-1}h^{-\epsilon}) \\ &= \frac{1}{2n\sqrt{h\pi}} \int_0^\infty x^{-1/2} f(x)dx + o(n^{-1}h^{-\epsilon}). \end{aligned}$$

La valeur de la variance dans la petite boule de centre 0 et de rayon $h^{1-\epsilon}$ dispose d'une valeur dérisoire ce qui fait que la quantité qui pése le plus est celle qui se trouve au milieu de $]0, +\infty[$.
Nous mesurons ainsi l'erreur quadratique moyenne intégrée MISE :

$$\begin{aligned} MISE(n,h,K,f) &= \int_0^\infty Biais\left\{\widehat{f}_n(x)\right\}^2 + \int_0^\infty Var\left\{\widehat{f}_n(x)\right\} \\ &= h^2 \int_0^\infty \left\{f'(x) + \frac{1}{2} x f''(x)\right\}^2 dx \\ &\quad + \frac{1}{2n\sqrt{h\pi}} \int_0^\infty x^{-1/2} f(x)dx + o(\frac{1}{n\sqrt{h}}). \end{aligned}$$

En minimisant le MISE par rapport à h, nous avons

$$2h \int_0^\infty \left\{f'(x) + \frac{1}{2} x f''(x)\right\}^2 dx - \frac{1}{2n} \frac{1}{2h^2\sqrt{\pi}} \int_0^\infty x^{-1/2} f(x)dx = 0.$$

En essayant de déterminer la fenêtre optimale, nous regroupons les termes en h de même côté ;

$$2h \int_0^\infty \left\{f'(x) + \frac{1}{2} x f''(x)\right\}^2 dx = \frac{1}{2n} \frac{1}{2h^2\sqrt{\pi}} \int_0^\infty x^{-1/2} f(x)dx.$$

C'est-à-dire

$$h^{5/2} = \frac{1/2\sqrt{\pi}\int_0^\infty x^{-1/2}f(x)dx}{4\int_0^\infty \left\{f^{'}(x) + \frac{1}{2}xf^{''}(x)\right\}^2 dx} n^{-1}.$$

Enfin, la fenêtre optimale est

$$h_{opt} = \frac{(1/2\sqrt{\pi})^{2/5}\left\{\int_0^\infty x^{-1/2}f^{'}(x)dx\right\}^{2/5}}{\left[\int_0^\infty \left\{f^{'}(x) + \frac{1}{2}xf^{''}(x)\right\}^2 dx\right]^{2/5}} n^{-2/5}.$$

La fenêtre optimale dans le cas asymétrique est d'ordre $O(n^{-2/5})$ inférieur que dans le cas symétrique $O(n^{-1/5})$. En remplaçant cette valeur optimale dans l'expression du MISE, nous avons successivement

$$\begin{aligned} MISE_{opt}(h_{opt}) &\doteq h_{opt}^2 \int_0^\infty \left\{f^{'}(x) + \frac{1}{2}xf^{''}(x)\right\}^2 dx \\ &\quad + \frac{1}{2\sqrt{\pi}}\frac{1}{n}h_{opt}^{-1/2}\int_0^\infty x^{-1/2}f(x)dx \\ &\doteq \frac{n^{-4/5}}{4^{4/5}}\left\{\frac{1}{2\sqrt{\pi}}\int_0^\infty x^{-1/2}f(x)dx\right\}^{4/5}\left[\int_0^\infty \left\{f^{'}(x) + \frac{1}{2}xf^{''}(x)\right\}^2 dx\right]^{1/5}. \end{aligned}$$

Dans le but de réduire le biais et par la suite l'erreur entre $\widehat{f}_n$ et f, nous présontons le deuxième estimateur qu'a introduit Chen (2000) ; la modification s'est faite au niveau de la cible de sorte qu'elle devient la moyenne de la variable aléatoire du noyau associé. Pour cela, soit

$$\widehat{\widehat{f}}_n(x) = \frac{1}{n}\sum_{i=1}^n K_{G(\rho_h(x);h)}(X_i), \tag{3.15}$$

où

$$\rho_h(x) \sim \begin{cases} x & \text{si } x \geq 2h \\ \frac{x^2}{4h^2+1} & \text{si } x \in [0,2h[. \end{cases}$$

De la même manière, nous calculons toute les propriétés de cet estimateur. Le biais est tel que :

$$Biais\left\{\widehat{\widehat{f}}_n(x)\right\} \sim \begin{cases} \frac{1}{2}xf^{''}(x)h + o(h) & \text{si } x \geq 2h \\ \xi_h(x)hf^{'}(x) + o(h) & \text{si } x \in [0,2h[. \end{cases}$$

La variable ξ dépend de h et change de valeur en fonction de x, elle est égale à :

$$\xi_h(x) = (1-x)\left\{\rho_h(x) - x/h\right\} / \left\{1 + h\rho_h(x) - x\right\}.$$

Clairement, le biais est plus petit dans ce cas ; quand x tend vers l'infini, nous obtenons une expression qui ne dépend que de la dérivée seconde $f^{''}$, ce qui est plus faible par rapport au biais de $\widehat{f}_n$.

La variance de $\widehat{\widehat{f}}$ est équivalente à celle de $\widehat{f}_n$ pour x/h tend vers l'infini. Nous distinguons une légère différence dans le cas où x/h s'approche de la constante k. En effet, la variance est égale à :

$$Var\left\{\widehat{\widehat{f}}_n(x)\right\} \sim \begin{cases} \frac{1}{2\sqrt{h\pi}}\frac{1}{n}x^{-1/2}f(x) + O(n^{-1}) & \text{si } x/h \to \infty \\ a(k)\frac{1}{nh}f(x) & \text{si } x/h \to k, \end{cases}$$

avec $a(k)$ un coefficient qui dépend seulement de k.
La somme du biais au carré et de la variance nous amène à déterminer le MISE de cet estimateur ;

$$MISE(\widehat{\widehat{f}_n}) \doteq \frac{1}{4}h^2 \int_0^\infty \left\{xf''(x)\right\}^2 dx + \frac{1}{2\sqrt{h\pi}}\frac{1}{n}\int_0^\infty x^{-1/2}f(x)dx.$$

Ainsi, la fenêtre optimale est

$$h_{opt} = \frac{(1/2\sqrt{\pi})^{2/5}\left\{\int_0^\infty x^{-1/2}f(x)dx\right\}^{2/5}}{\left[\int_0^\infty \left\{xf''(x)\right\}^2 dx\right]^{2/5}} n^{-2/5}.$$

En substituant cette valeur dans l'expression du MISE, l'erreur quadratique moyenne intégrée optimale est :

$$MISE_{opt}(h_{opt}) \doteq \frac{1}{4^{4/5}}\left[\frac{1}{2\sqrt{\pi}}\int_0^\infty x^{-1/2}f(x)dx\right]^{4/5}\left[\int_0^\infty \left\{xf''(x)\right\}^2 dx\right]^{1/5} n^{-4/5}.$$

Nous pouvons être tenter que les deux estimateurs $\widehat{f}$ et $\widehat{\widehat{f}}$ atteignent la vitesse de convergence optimale. Nous montrons que pour toute densité f continue :

$$\int_0^\infty \left\{f'(x) + \frac{1}{2}xf''(x)\right\}^2 dx \geq \int_0^\infty \left\{\frac{1}{2}xf''(x)\right\}^2 dx.$$

Ceci implique systématiquement

$$MISE_{opt}(\widehat{f}) \geq MISE_{opt}(\widehat{\widehat{f}}).$$

Enfin, du point de vue purement théorique, il est clair que le deuxième estimateur $\widehat{\widehat{f}}$ donne de meilleure performance en utilisant une fenêtre plus faible par rapport au premier estimateur $\widehat{f}$.

b. Cas d'un noyau associé bêta

Tout comme les noyaux gamma, Chen (1999) applique le même principe pour les noyaux bêta. Il introduit pour cela un premier estimateur où il remarque que les paramètres choisis ne sont pas les plus adéquats, donc, il essaye de les harmoniser et les

arranger pour aboutir à de meilleures estimations et par conséquent de meilleures performances. L'idée est strictement la même, nous commençons ainsi par rappeler la loi bêta. La densité de probabilité d'une loi bêta est définie continue sur [0,1] telle que :

$$g_{Be(a,b)}(t) = \frac{1}{B(a,b)} t^{a-1}(1-t)^{b-1} 1_{[0,1]}(t),$$

où $a > 0$, $b > 0$ et vérifiant

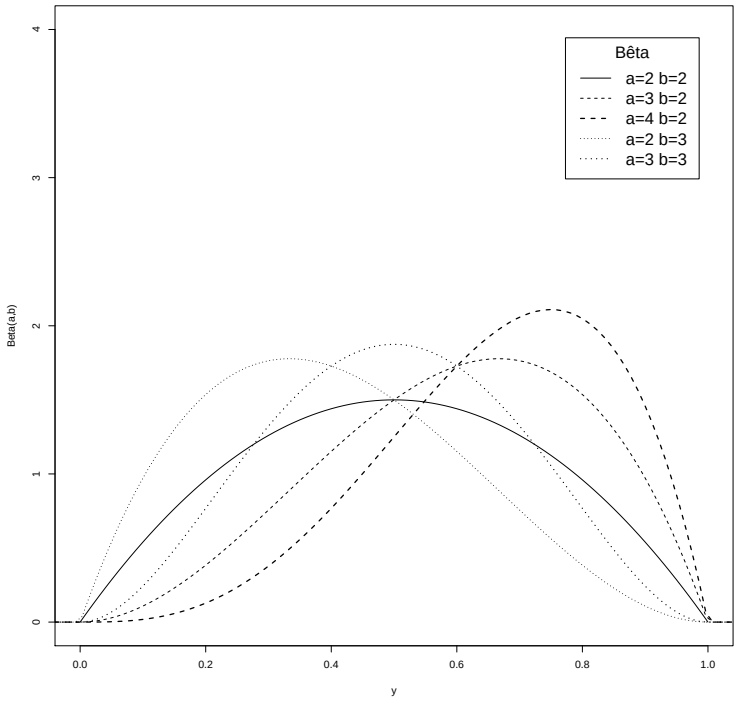
$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1} dt.$$

Si X est une variable aléatoire qui suit la loi bêta, alors

$$\mathbb{E}(X) = \frac{a}{a+b} \text{ et } Var(X) = \frac{ab}{(a+b)^2(a+b+1)}.$$

La figure 3.7 donne l'allure de la fonction bêta d'une manière générale.

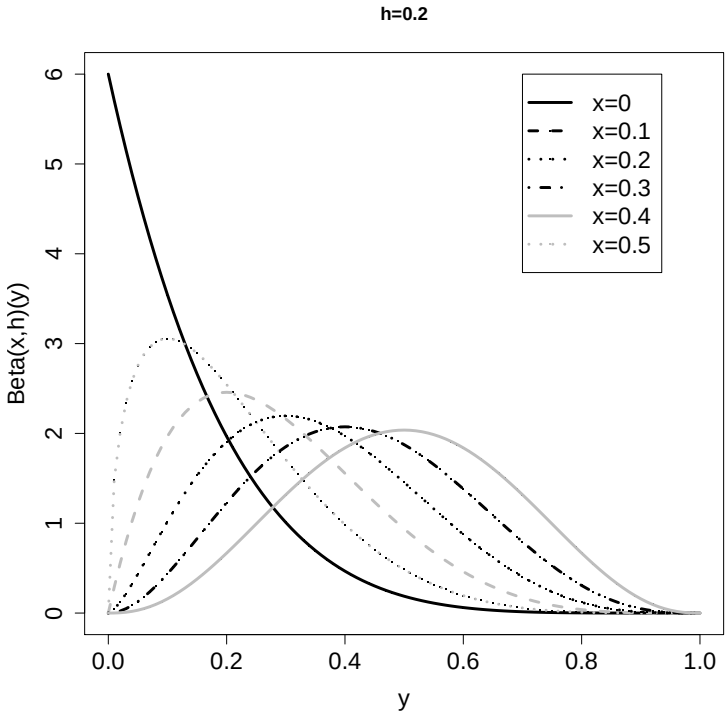
FIG. 3.7 – *Allure générale de la densité bêta*



Le noyau $K_{Be(x/h+1;(1-x)/h+1))}$ est le noyau associé à une variable aléatoire $\mathcal{K}_{Be(x/h+1;(1-x)/h+1))}$ de loi bêta et de support $\aleph_{x,h} = [0,1]$ tel que :

$$K_{Be(x/h+1;(1-x)/h+1)}(t) = \frac{1}{B(x/h+1,(1-x)/h+1)} t^{x/h}(1-t)^{(1-x)/h}.$$

FIG. 3.8 – *Allure du noyau associé bêta pour $h = 0.2$ et x varié*



Les figures 3.8 et 3.9 donnent la variation du noyau bêta chaque fois que nous changeons les paramètres x et h.

Nous nous assurons que ce noyau est bel et bien un noyau associé :

i. $[0,1] \cap [0,1] = [0,1] \neq \varnothing$.

ii. $\cup_x [0,1] = [0,1]$.

iii. $\mathbb{E}(\mathcal{K}_{Be(x/h+1;(1-x)/h+1)}) = \frac{(x+h)}{(1+2h)} \sim x$ quand $h \to 0$.

iv. $Var(\mathcal{K}_{Be(x/h+1;(1-x)/h+1)}) = \frac{x(1-x)h+h^2+h^3}{(1+2h)^2(1+3h)} < \infty$.

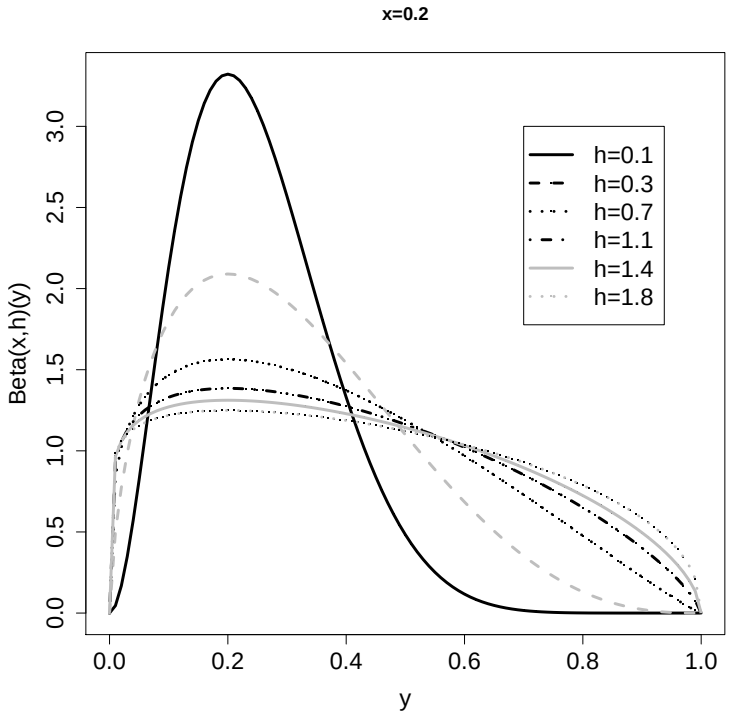
v. $h \to 0 \Rightarrow Var(\mathcal{K}_{Be(x/h+1;(1-x)/h+1)}) = 0$.

Soit $X_1, X_2, \ldots, X_n$ un échantillon de variables aléatoires i.i.d. sur $\aleph = [0,1]$, de densité de probabilité continue asymétrique inconnue f. Nous considérons l'estimateur $\widehat{f}_n$ de f à noyau bêta tel que

$$\begin{aligned} \widehat{f}_n(x) &= \frac{1}{n}\sum_{i=1}^{n} K_{Be(x/h+1;(1-x)/h+1)}(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{B(x/h+1,(1-x)/h+1)} X_i^{x/h}(1-X_i)^{(1-x)/h}, \end{aligned}$$

avec $x \in [0,1]$ et $h > 0$ est le paramètre de lissage.

FIG. 3.9 – *Allure du noyau associé bêta pour $x = y = 2$ et h varié*



En se bénéficiant des calculs antérieurs, nous avons

$$Biais\left\{\widehat{f}_n(x)\right\} = h(1-2x)f^{'}(x) + \frac{1}{2}x(1-x)hf^{''}(x) + o(h),$$

et

$$Var\left\{\widehat{f}_n(x)\right\} = \frac{1}{n}\left[\mathbb{E}\left\{K_{Be(x/h+1;(1-x)/h+1)}(X_1)\right\}^2\right] + O(n^{-1}),$$

où

$$\left\{K_{Be(x/h+1;(1-x)/h+1)}(X_1)\right\}^2 = \frac{1}{B^2(x/h+1;(1-x)/h+1)}X_i^{2x/h}(1-X_i)^{2(1-x)/h}.$$

Soit $K_{Be(x/h+1;(1-x)/h+1)}$ le noyau associé de loi bêta défini par

$$K_{Be(2x/h+1;2(1-x)/h+1)}(X_i) = \frac{1}{B(2x/h+1;2(1-x)/h+1)}X_i^{2x/h}(1-X_i)^{2(1-x)/h}.$$

Ce qui fait que

$$X_i^{2x/h}(1-X_i)^{2(1-x)/h} = B(2x/h+1;2(1-x)/h+1)K_{Be(2x/h+1;2(1-x)/h+1)}(X_i).$$

Ainsi :

$$\left\{K_{Be(x/h+1;(1-x)/h+1)}(X_1)\right\}^2 = \frac{B(2x/h+1;2(1-x)/h+1)}{B^2(x/h+1;(1-x)/h+1)} K_{Be(2x/h+1;2(1-x)/h+1)}(X_i).$$

Tout compte fait, nous avons

$$\begin{aligned}\mathbb{E}\left\{K_{Be(x/h+1;(1-x)/h+1)}(X_1)\right\}^2 &= \frac{B(2x/h+1;2(1-x)/h+1)}{B^2(x/h+1;(1-x)/h+1)} \\ &\quad \mathbb{E}\left\{K_{Be(2x/h+1;2(1-x)/h+1)}(X_i)\right\}.\end{aligned}$$

Nous appellons $A_h(x) = \frac{B(2x/h+1;2(1-x)/h+1)}{B^2(x/h+1;(1-x)/h+1)}$ et nous rappellons que $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$. Nous vérifions les conditions du noyau associé $K_{Be(2x/h+1;2(1-x)/h+1)}(X_i)$.

i. $[0,1] \cap [0,1] = [0,1] \neq \varnothing$.
ii. $\cup_x [0,1] = [0,1]$.
iii. $\mathbb{E}(\mathcal{K}_{Be(2x/h+1;2(1-x)/h+1)}) = \frac{x+h/2}{1+h} \sim x$ quand $h \to 0$.
iv. $Var(\mathcal{K}_{Be(2x/h+1;2(1-x)/h+1)}) = \frac{4x(1-x)h+2h+h^2}{(2+2h)^2(2+3h)} < \infty$.
v. $h \to 0 \Rightarrow Var(\mathcal{K}_{Be(2x/h+1;2(1-x)/h+1)}) = 0$.

En exploitant la fonction (3.14), nous avons

$$\begin{aligned}R(2x/h) &= \frac{\sqrt{2\pi}}{\tau(2x/h+1)} e^{-2x/h} \left(\frac{2x}{h}\right)^{2x/h+1/2} \\ R(2(1-x)/h) &= \frac{\sqrt{2\pi}}{\tau(2(1-x)/h+1)} e^{-2(1-x)/h} \left(\frac{2(1-x)}{h}\right)^{2(1-x)/h+1/2} \\ R(2/h+1) &= \frac{\sqrt{2\pi}}{\tau(2/h+2)} e^{-2/h+1} \left(\frac{2}{h}+1\right)^{2/h+1+1/2}\end{aligned}$$

De même, nous avons

$$\begin{aligned}R^2(x/h) &= \frac{2\pi}{\tau^2(x/h+1)} e^{-2x/h} \left(\frac{x}{h}\right)^{2x/h+1/2} \\ R^2((1-x)/h) &= \frac{2\pi}{\tau^2(2(1-x)/h+1)} e^{-2(1-x)/h} \left(\frac{1-x}{h}\right)^{2(1-x)/h+1/2} \\ R^2(1/h+1) &= \frac{2\pi}{\tau^2(1/h+2)} e^{-2(1/h+1)} \left(\frac{1+h}{h}\right)^{2(1/h+1)}\end{aligned}$$

Ainsi, nous trouvons

$$A_h(x) = \frac{1}{2\sqrt{\pi}} \{x(1-x)\}^{-1/2} h^{-1/2} \frac{R(2/h+1)R^2(x/h)R^2((1-x)/h)}{R(2x/h)R(2(1-x)/h)R^2(1/h+1)}.$$

En majorant cette expression par 1, $A_h(x)$ prend deux valeurs différentes selon la convergence du rapport x/h et $(1-x)/h$.

$$A_h(x) \sim \begin{cases} \frac{1}{2\sqrt{\pi}} \{x(1-x)\}^{-1/2} h^{-1/2} & \text{si } x/h \text{ et } (1-x)/h \to \infty \\ \frac{\Gamma(2k+1)}{2^{2k+1}\Gamma^2(k+1)} h^{-1} & \text{si } x/h \text{ ou } (1-x)/h \to k. \end{cases}$$

Enfin, la variance est égale à

$$Var\left\{\widehat{f}_n(x)\right\} \sim \begin{cases} \frac{1}{2\sqrt{\pi}} \{x(x-1)\}^{-1/2} \frac{1}{nh^{1/2}} f(x) + O(n^{-1}) & \text{si } x/h \text{ et } (1-x)/h \to \infty \\ \frac{\tau(2k+1)}{2^{2k+1}\tau^2(k+1)} \frac{1}{nh} f(x) + O(n^{-1}) & \text{si } x/h \text{ ou } (1-x)/h \to k. \end{cases}$$

Nous évaluons l'erreur quadratique moyenne intégrée de cet estimateur.

$$MISE\left\{\widehat{f}_n(x)\right\} \doteq h^2\left\{\int_o^1 (1-2x)f^{'}(x)+\frac{1}{2}x(1-x)f^{''}(x)\right\}^2 dx + \frac{1}{2n\sqrt{h\pi}}\int_o^\infty \{x(x-1)\}^{-1/2} f(x)dx.$$

Nous minimisons le MISE par rapport h et nous déterminons la fenêtre optimale h_{opt}.

$$h_{opt} = \frac{1}{4^{2/5}} \frac{\left[\frac{1}{2\sqrt{\pi}}\int_o^1 \{x(x-1)\}^{-1/2} f(x)dx\right]^{2/5}}{\left[\left\{\int_o^1 (1-2x)f^{'}(x)+\frac{1}{2}x(1-x)f^{''}(x)\right\}^2 dx\right]^{2/5}} n^{-2/5}.$$

De manière similaire au noyau associé gamma et en considérant les mêmes raisons pour les quelles nous avons introduit le second estimateur à noyau associé gamma qui corrige le biais au bord, nous présentons à ce niveau le second estimateur à noyau associé bêta défini sur [0,1] :

$$\widehat{\widehat{f}}_n(x) = \frac{1}{n}\sum_{i=1}^n K_{Be(x;h)}(X_i), \tag{3.16}$$

avec

$$K_{Be(x;h)}(X_i) \sim \begin{cases} K_{Be(x/h;(1-x)/h)}(X_i) & \text{si } x \in [2h,1-2h] \\ K_{Be(\rho_h(x);(1-x)/h)}(X_i) & \text{si } x \in [0,2h[\\ K_{Be(x/h;\rho_h(1-x))}(X_i) & \text{si } x \in]1-2h,1] \end{cases}$$

où $\rho_h(x) = 2h^2 + 2.5 - \sqrt{4h^4+6h^2-x^2-x/h}$. $\forall h$ fixé, $\rho_h(x)$ est croissante sur [0,2h]. Nous faisons tendre h vers 0 et vers 1, les quantités au bord deviennent faibles. Ainsi, nous récupérons juste l'expression qui se trouve à l'intérieur de l'intervalle. Nous révisons les hypothèses mis sur le noyau associé :

i. $[0,1]\cap[0,1] = [0,1] \neq \varnothing$.
ii. $\cup_x[0,1] = [0,1]$.
iii. $\mathbb{E}(\mathcal{K}_{Be(x/h;(1-x)/h)}) = x$.
iv. $Var(\mathcal{K}_{Be(x/h;1-x/h)}) = \frac{x(1-x)h}{1+h} < \infty$.
v. $h \to 0 \Rightarrow Var(\mathcal{K}_{Be(x/h;1-x/h)}) = 0$.

Le biais est égal à

$$Biais\left\{\widehat{\widehat{f}}_n(x)\right\} \sim \begin{cases} \frac{1}{2}hx(1-x)f^{''}(x)+o(h) & \text{si } x \in [2h,1-2h] \\ \zeta_h(x)hf^{'}(x)+o(h) & \text{si } x \in [0,2h] \\ -\zeta_h(x)hf^{'}(x)+o(h) & \text{si } x \in [1-2h,1] \end{cases}$$

avec $\zeta_h(x) = (1-x)\{\rho_h(x)-x/h\}\{1+h\rho_h(x)-x\}$.
La variance de ce deuxième estimateur est similaire au premier quand x/h et $(1-x)/h$

tendent vers l'infini.

$$Var\left\{\widehat{\widehat{f}}_n(x)\right\} = \frac{1}{2n\sqrt{h\pi}}\left\{x(x-1)\right\}^{-1/2} f(x) + O(n^{-1}).$$

Enfin, la fenêtre optimale est

$$h_{opt} = \frac{\left[\frac{1}{2\sqrt{\pi}}\int_o^1 \left\{x(x-1)\right\}^{-1/2} f(x)dx\right]^{2/5}}{\left[\left\{\int_o^1 x(1-x)f''(x)\right\}^2 dx\right]^{2/5}} n^{-2/5}.$$

Comme

$$\left\{\int_o^1 (1-2x)f'(x) + \frac{1}{2}x(1-x)f''(x)\right\}^2 dx \geq \int_o^1 \left\{x(1-x)f''(x)\right\}^2 dx,$$

alors la fenêtre optimale du premier estimateur est plus grande que celle du second. En remplaçant la valeur optimale de h dans l'expression du MISE, nous constatons que l'erreur quadratique moyenne intégrée trouvée dans le cas de $\widehat{\widehat{f}}_n$ est inférieure à celle de $\widehat{f}_n$.

$$MISE(\widehat{f}_n) \geq MISE(\widehat{\widehat{f}}_n).$$

c. Cas d'un noyau associé gaussien inverse IG

Soit $g(t)$ la densité de loi gaussienne inverse telle que

$$g_{IG(a,b)}(t) = \frac{\sqrt{b}}{\sqrt{2\pi t^3}} \exp\left\{\frac{-b}{2a}\left(\frac{t}{a} - 2 + \frac{a}{t}\right)\right\},$$

où $t > 0$ et (a,b) est un couple de deux réels strictement positifs. La figure 3.10 donne l'allure générale de la densité gaussienne inverse. Si X est une variable aléatoire qui suit la loi gaussienne inverse alors

$$\mathbb{E}(X) = a \text{ et } Var(X) = a^3/b.$$

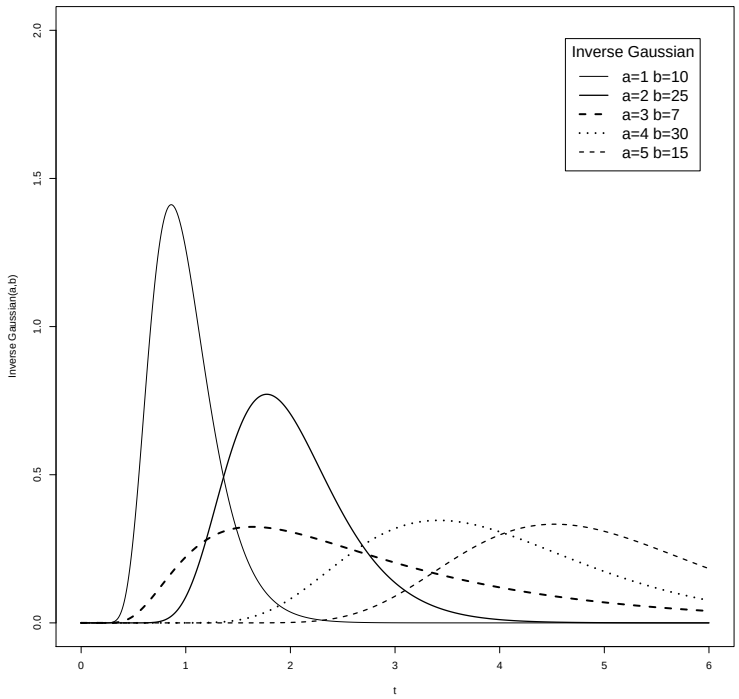
Soit $K_{IG(x;1/h)}$ le noyau gaussien inverse associé à la variable aléatoire $\mathcal{K}_{IG(x;1/h)}$ défini sur $\aleph_{x,h} = [0, +\infty[$, de paramètres x et $1/h$. Ce noyau associé $K_{IG(x;1/h)}$ se définit comme suit :

$$K_{IG(x;1/h)}(t) = \frac{1}{\sqrt{2\pi h t^3}} \exp\left\{\frac{-1}{2hx}\left(\frac{t}{x} - 2 + \frac{x}{t}\right)\right\}.$$

Nous vérifions chacune des hypothèses du noyau associé :

i. $\mathbb{R}_+ \cap \mathbb{R}_+ \neq \varnothing$.
ii. $\cup_x \mathbb{R}_+ = \mathbb{R}_+$.
iii. $\mathbb{E}(\mathcal{K}_{IG(x;1/h)}) = x$.
iv. $Var(\mathcal{K}_{IG(x;1/h)}) = x^3 h < \infty$.
v. $h \to 0 \Rightarrow Var(\mathcal{K}_{IG(x;1/h)}) = 0$.

FIG. 3.10 – *Allure générale de la densité gaussienne inverse*



Ainsi le noyau $K_{IG(x;1/h)}$ est un noyau associé. Les graphiques 3.11 et 3.12 présentent l'allure d'une densité gaussienne inverse quand nous varions x et h.
Pour un échantillon de variables aléatoires i.i.d., X_1,X_2,...,X_n, nous considérons la densité de probabilité f inconnue définie continue sur $\mathbb{R}_+$. Soit l'estimateur $\widehat{f_n}$ de f à noyau inverse gaussien défini sur $[0,+\infty[$ tel que :

$$\begin{aligned}\widehat{f_n}(x) &= \frac{1}{n}\sum_{i=1}^{n} K_{IG(x;1/h)}(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi h X_i^3}} \exp\left\{\frac{-1}{2hx}\left(\frac{X_i}{x} - 2 + \frac{x}{X_i}\right)\right\},\end{aligned}$$

où le paramètre h est strictement positif et x est dans $\mathbb{R}_+$.
En tenant compte de ce qui était cité précédemment, le biais est

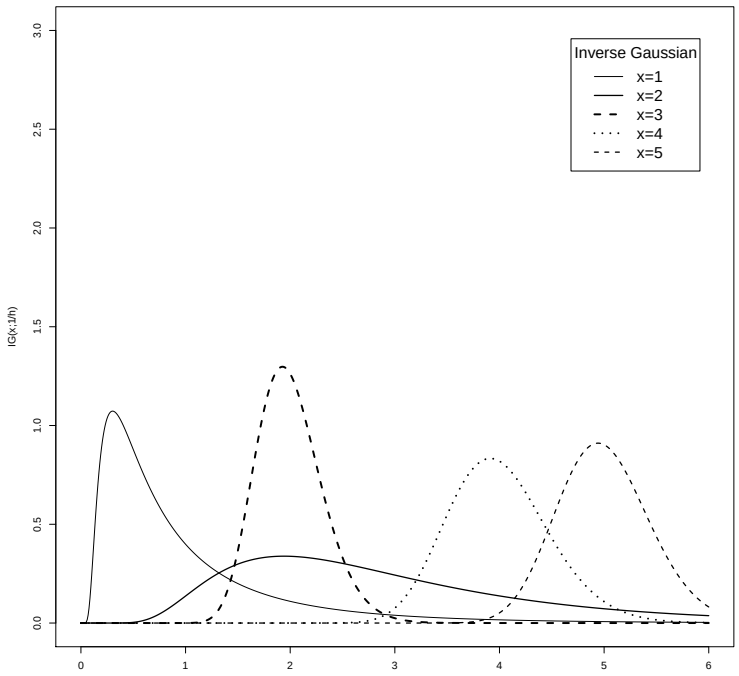
$$Biais\left\{\widehat{f_n}(x)\right\} = \frac{1}{2}x^3 f''(x)h + o(h),$$

donc

$$\int_{\mathbb{R}_+} Biais^2\left\{\widehat{f_n}(x)\right\} dx = \frac{1}{4}h^2 \int_{\mathbb{R}_+} \left\{x^3 f''(x)\right\}^2 dx + o(h^2).$$

Comme $\int_{\mathbb{R}_+} \left\{x^3 f''(x)\right\}^2 dx$ est finie alors, pour tout x qui tend vers $+\infty$, $x^3 f''(x)$ converge vers 0. D'où le biais diminue quand x augmente.

FIG. 3.11 – *Allure du noyau associé gaussien inverse pour $h = 0.1$ et x varié*



Nous calculons la variance sur la base des calculs effectués au préalable.

$$Var\left\{\widehat{f}_n(x)\right\} = \frac{1}{n}\mathbb{E}\left[\left\{K_{IG(x;1/h)}(X_1)\right\}^2\right] + O(n^{-1}).$$

$$\left\{K_{IG(x;1/h)}(X_1)\right\}^2 = \frac{1}{2\pi h}X_1^{-3}\exp\left\{\frac{-1}{xh}\left(\frac{X_1}{x} - 2 + \frac{x}{X_1}\right)\right\}.$$

Soit $K_{IG(x;2/h)}(X_1)$ le noyau gaussien inverse de paramètre x et $2/h$ associé à $\mathcal{K}_{IG(x;2/h)}$ et définie sur $[0, +\infty[$. Nous vérifions simplement les différentes hypotèses liées à cette variable aléatoire :

i. $\mathbb{R}_+ \cap \mathbb{R}_+ \neq \varnothing$
ii. $\cup_x \mathbb{R}_+ = \mathbb{R}_+$.
iii. $\mathbb{E}(\mathcal{K}_{IG(x;2/h)}) = x$.
iv. $Var(\mathcal{K}_{IG(x;2/h)}) = x^3\frac{h}{2} < \infty$.
v. $h \to 0 \Rightarrow Var(\mathcal{K}_{IG(x;2/h)}) = 0$.

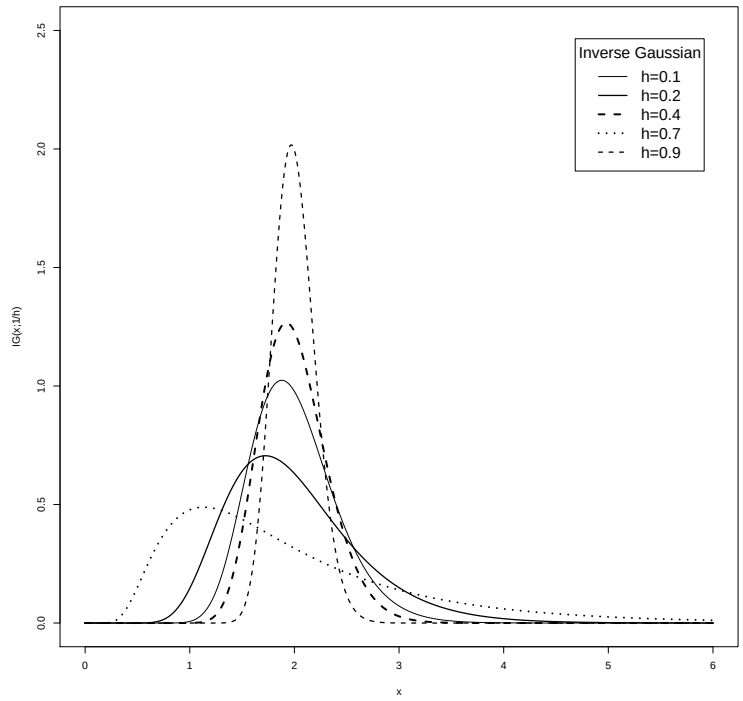
En conclusion, il s'agit d'un noyau associé. Tout bien considéré :

$$K_{IG(x;2/h)}(X_1) = \frac{\sqrt{2}}{\sqrt{2\pi h X_1^3}}\exp\left\{\frac{-1}{xh}\left(\frac{X_1}{x} - 2 + \frac{x}{X_1}\right)\right\}.$$

Ce qui implique

$$\exp\left\{\frac{-1}{xh}\left(\frac{X_1}{x} - 2 + \frac{x}{X_1}\right)\right\} = \sqrt{\pi h X_1^3}K_{IG(x;2/h)}(X_1),$$

FIG. 3.12 – *Allure du noyau associé gaussien inverse pour $x = 2$ et h varié*



et par la suite, nous avons

$$\mathbb{E}\left[\left\{K_{IG(x;1/h)}(X_1)\right\}^2\right] = \frac{1}{2\sqrt{\pi h}}\mathbb{E}\left\{X_1^{-3/2}K_{IG(x;2/h)}(X_1)\right\}.$$

A partir de l'approximation de Taylor-Lagrange, nous obtenons

$$\begin{aligned}\mathbb{E}\left\{X_1^{-3/2}K_{IG(x;2/h)}(X_1)\right\} &= \mathbb{E}\left\{\mathcal{K}_{IG(x;1/h)}^{-3/2}f(\mathcal{K}_{IG(x;1/h)})\right\} \\ &= x^{-3/2}f(x) + O(h).\end{aligned}$$

En conclusion, quand $x > 0$ se situe à l'intérieur du support, la variance est

$$Var\left\{\widehat{f}_n(x)\right\} = \frac{1}{2\sqrt{h\pi}}\frac{x^{-3/2}}{n}f(x) + o(n^{-1}h^{-1}).$$

La variance au bord, quand $x/h \to k$, présente quelques différences. Elle est égale à

$$Var\left\{\widehat{f}_n(x)\right\} = \frac{1}{2\sqrt{h\pi}}\frac{k^{-3/2}}{n}f(x) + o(n^{-1}h^{-2}),$$

k étant une constante positive.
L'erreur globale de cet estimateur est

$$MISE \doteq \frac{1}{4}h^2\int_{\mathbb{R}_+}\left\{x^3f''(x)\right\}^2dx + \frac{1}{2\sqrt{h\pi}}\frac{1}{n}\int_{\mathbb{R}_+}x^{-3/2}f(x)dx.$$

Nous cherchons à déterminer le h optimal. Pour cela, nous minimisons le MISE par rapport à h, nous trouvons

$$\frac{1}{2}h\int_{\mathbb{R}_+}\left\{x^3f''(x)\right\}^2dx-\frac{1}{2h^2\sqrt{\pi}}\frac{1}{2n}\int_{\mathbb{R}_+}x^{-3/2}f(x)dx=0,$$

c'est à dire

$$h^{5/2}\int_{\mathbb{R}_+}\left\{x^3f''(x)\right\}^2dx=\frac{1}{2\sqrt{\pi}}\frac{1}{n}\int_{\mathbb{R}_+}x^{-3/2}f(x)dx.$$

Enfin, la fenêtre optimale est

$$h_{opt}=\frac{\left\{\frac{1}{2\sqrt{\pi}}\int_{\mathbb{R}_+}x^{-3/2}f(x)dx\right\}^{2/5}}{\left[\int_{\mathbb{R}_+}\left\{x^3f''(x)\right\}^2dx\right]^{2/5}}n^{-2/5}.$$

En l'exploitant dans la formule du MISE, nous trouvons

$$MISE(h_{opt})=\frac{5}{4}\left\{\frac{1}{2\sqrt{\pi}}\int_{\mathbb{R}_+}x^{-3/2}f(x)dx\right\}^{4/5}\left\{\int_{\mathbb{R}_+}x^3f''(x)dx\right\}^{2/5}n^{-4/5}.$$

d. Cas d'un noyau associé gaussien inverse réciproque RIG

Nous considérons $g(t)$ la densité de loi gaussienne inverse réciproque :

$$g_{RIG(a,b)}(t)=\frac{\sqrt{b}}{\sqrt{2\pi t}}\exp\left\{\frac{-b}{2a}\left(at-2+\frac{1}{at}\right)\right\},$$

où $t>0$, $a>0$ et $b>0$. La figure 3.13 donne l'allure générale d'une densité gaussienne inverse réciproque. Si X est une variable aléatoire qui suit la loi gaussienne inverse réciproque alors

$$\mathbb{E}(X)=\frac{1}{a}+\frac{1}{b}\text{ et }Var(X)=\frac{1}{ab}+\frac{2}{b^2}.$$

Soit $K_{RIG(1/(x-h);1/h)}$ le noyau gaussien inverse réciproque associé à la variable aléatoire $\mathcal{K}_{RIG(1/(x-h);1/h)}$ défini sur $\aleph_{x,h}=[0,+\infty[$, de paramètres $1/(x-h)$ et $1/h$. Ce noyau se présente comme suit :

$$K_{RIG(1/(x-h);1/h)}(t)=\frac{1}{\sqrt{2\pi ht}}\exp\left\{-\frac{x-h}{2h}\left(\frac{t}{x-h}-2+\frac{x-h}{t}\right)\right\}.$$

Nous commençons par vérifier chacune des hypothèses du noyau associé :

i. $\{\aleph=[0,+\infty[\}\cap\{\aleph_{x,h}=[0,+\infty[\}=[0,+\infty[\neq\varnothing$.

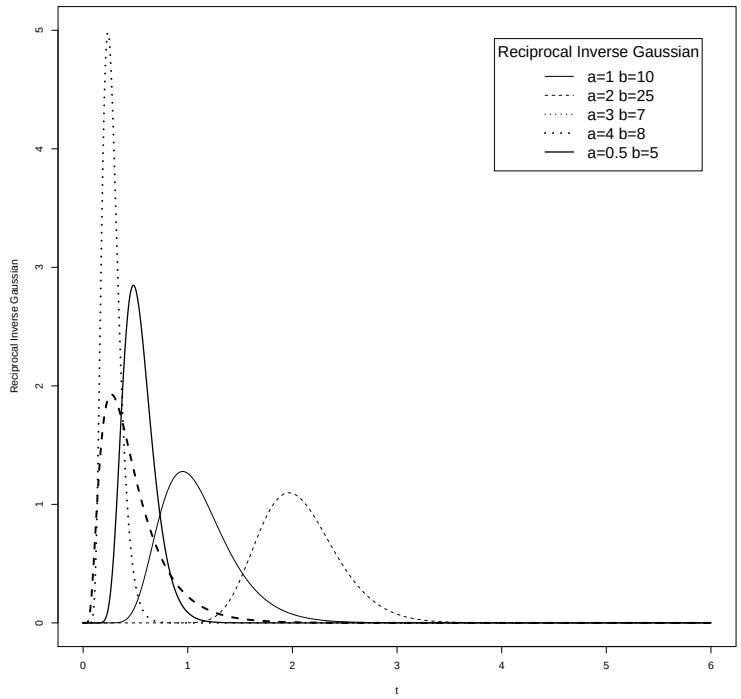
ii. $\cup_x[0,+\infty[=[0,+\infty[$.

iii. $\mathbb{E}(\mathcal{K}_{RIG(1/(x-h);1/h)})=x-h+h=x$.

iv. $Var(\mathcal{K}_{RIG(1/(x-h);1/h)})=(x-h)h+2h^2=xh+h^2<\infty$.

v. $h\to 0\Rightarrow Var(\mathcal{K}_{RIG(1/(x-h);1/h)})=0$.

FIG. 3.13 – *Allure générale de la densité gaussienne inverse réciproque*



Ainsi toutes les conditions du noyau associé sont satisfaites.

Soit X_1,X_2,...,X_n l'échantillon de variables aléatoires i.i.d. de densité de probabilité f inconnue définie continue sur $\aleph = \mathbb{R}_+$. Nous considérons l'estimateur $\widehat{f}_n$ de f à noyau gaussien inverse réciproque défini sur $[0, +\infty[$ tel que

$$\begin{aligned}\widehat{f}_n(x) &= \frac{1}{n}\sum_{i=1}^{n} K_{RIG(1/(x-h);1/h)}(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi h X_i}} \exp\left\{-\frac{x-h}{2h}\left(\frac{X_i}{x-h} - 2 + \frac{x-h}{X_i}\right)\right\},\end{aligned}$$

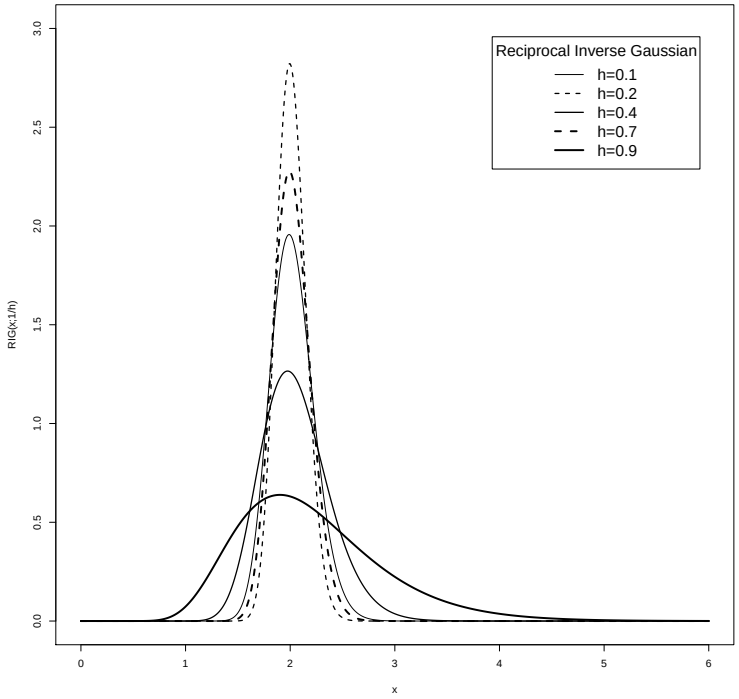
avec $h > 0$ et $x \in \mathbb{R}_+$.
En tenant compte des résultats obtenus précédemment :

$$Biais\left\{\widehat{f}_n(x)\right\} = \frac{1}{2}xf''(x)h + o(h),$$

et donc

$$\int_{\mathbb{R}_+} Biais^2\left\{\widehat{f}_n(x)dx\right\} = \frac{1}{4}h^2 \int_{\mathbb{R}_+} \left\{xf''(x)\right\}^2 dx + o(h^2).$$

FIG. 3.14 – *Allure du noyau associé gaussien inverse réciproque pour $x = 2$ et h varié*



En refaisant les calculs de la variance de la même façon, nous trouvons

$$Var\left\{\widehat{f}_n(x)\right\} \sim \begin{cases} \frac{1}{2n\sqrt{h\pi}} x^{-1/2} f(x) + O(n^{-1}) & \text{si } x/h \to \infty \\ \frac{1}{2nh\sqrt{\pi}} (k^{-1/2} + \frac{7}{16} k^{-3/2}) + O(n^{-1}) & \text{si } x/h \to k. \end{cases}$$

La fenêtre optimale est égale à

$$h_{opt} = \left(\frac{1}{2\sqrt{\pi}}\right)^{2/5} \frac{\left\{\int_{\mathbb{R}_+} x^{-1/2} f(x) dx\right\}^{2/5}}{\left[\int_{\mathbb{R}_+} \{x f''(x)\}^2 dx\right]^{2/5}} n^{-2/5}.$$

En conclusion, nous évaluons le MISE en fonction de cette valeur h_{opt} :

$$MISE(h_{opt}) = \left(\frac{1}{2\sqrt{\pi}}\right)^{2/5} \frac{\left\{\int_{\mathbb{R}_+} x^{-1/2} f(x) dx\right\}^{2/5}}{\left[\int_{\mathbb{R}_+} \{x f''(x)\}^2 dx\right]^{2/5}} n^{-2/5}.$$

e. Remarques :

i. *Quand le support du noyau associé $\aleph_{x,h}$ ne dépend pas ni de x ni de h alors nécessairement il coincide avec le support des observations $\aleph$. Ceci est vrai pour tout les*

exemples que nous avons traité.

ii. *Dans l'expression de la fenêtre optimale, la densité inconnue f et sa dérivée seconde figurent dans le numérateur et le dénumérateur contrairement au cas symétrique où f se trouve seulement au dénumérateur. Si nous voulons faire une analogie entre le cas symétrique et asymétrique, alors pour admettre la méthode "Plug-in" comme méthode d'estimation du paramètre h, f doit systématiquement suivre une loi continue asymétrique de support $\aleph$. Par exemple ; dans le cas des noyaux associés gamma, gaussien inverse et gaussien inverse réciproque, nous supposons que f suit une loi asymétrique défini sur $\mathbb{R}_+$. Dans le cas des noyaux associés bêta, la densité suit une loi de support* [0,1].

3.2 Cas multivarié

Dans cette section, nous généralisons l'estimateur à noyau associé continu asymétrique au cas multidimensionnel. Pour cela, nous considérons un échantillon de variables aléatoires $X_1,\ldots,X_n$ i.i.d., de densité de probabilité f continue asymétrique inconnue defini sur $\aleph$ de dimension d. L'estimateur $\widehat{f}_n$ de f à noyau associé continu asymétrique est

$$\widehat{f}_n(\underline{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{\underline{x},H}(X_i), \tag{3.17}$$

où la cible $\underline{x} =^t (x_1,\ldots,x_d)$, H est la matrice pleine de variance-covariance des fenêtres h de dimension $d \times d$, et $X_i =^t (X_{i1},\ldots,X_{id})$. La fonction $K_{\underline{x},H}$ est le noyau associé asymétrique sur $\aleph_{x,h}$.
Pareillement, nous donnons un estimateur qui se base sur le produit des noyaux univariés asymétriques. Cet estimateur a une forme plus vulgarisée que celle de (3.17). En effet, nous avons

$$\widehat{f}_n(\underline{x}) = \frac{1}{n}\sum_{i=1}^{n}\left\{\prod_{j=1}^{d} K^j_{x_j,h_j}(X_{ij})\right\}, \tag{3.18}$$

où x_j est la jème composante du vecteur $\underline{x}$, h_j est la jème fenêtre et X_{ij} est la ième observation de la jème composante.

Chapitre 4

Noyau associé discret

Le nombre de travaux abordant les estimateurs à noyau pour des données discrètes reste limité. Dans ce chapitre, nous présentons deux types de noyaux associés discrets. La première section porte sur les noyaux associés discrets pour des données catégorielles où les données sont qualitatives ordonnées et définis sur un ensemble fini inclu dans $\mathbb{N}$ que nous désignons $\aleph_{x,h}$. Ensuite, dans une deuxième partie, nous introduisons le noyau associé discret pour des données de comptages. Une première tentative dans ce cadre, uniquement de manière expérimentale, a été proposé par Marsh & Mukhopadhyay (1999). Nous étudions les propriétes ponctuelles et globales de chacun des deux estimateurs à noyau associé discret. Différentes techniques de selection de la fenêtre du lissage sont proposées. Enfin, nous généralisons l'estimateur à noyau associé au cas multivarié.

Définition 1 : *Soit x fixé dans $\aleph$ et $h > 0$. Nous appelons "noyau associé discret" $K_{x,h}$, toute fonction de masse de probabilité discrète d'une variable aléatoire $\mathcal{K}_{x,h}$ sur le support $\aleph_{x,h}$, tels que :*

$$\aleph_{x,h} \cap \aleph \neq \varnothing \tag{4.1}$$

$$\cup_x \aleph_{x,h} \supseteq \aleph \tag{4.2}$$

$$\mathbb{E}(\mathcal{K}_{x,h}) \sim x \text{ quand } h \to 0 \tag{4.3}$$

$$Var(\mathcal{K}_{x,h}) < \infty \tag{4.4}$$

$$Var(\mathcal{K}_{x,h}) \to 0 \text{ quand } h \to 0. \tag{4.5}$$

Commentaire : Nous vérifions dans ce qui suit, que dans le cas du noyau associé discret pour des données catégorielles, le support $\aleph_{x,h}$ coincide avec $\aleph$. Nous verrons que dans certaine situation, ce n'est pas toujours vérifié comme dans le cas des données de comptage ; $\aleph_{x,h}$ dépend de x et ne se colle pas avec le support $\aleph$.

Définition 2 : *Soit $X_1, \ldots, X_n$ un échantillon de variables aléatoires i.i.d. de fonction de masse de probabilité f discrète inconnue sur $\aleph$. L'estimateur à noyau associé discret $\widehat{f}_n \equiv \widehat{f}_{n,h,K}$ de f est défini par*

$$\widehat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} K_{x,h}(X_i), \tag{4.6}$$

avec $x \in \aleph$ et $h > 0$.

Propriété 1 : Soit x fixé dans $\aleph$. Nous avons

$$\mathbb{E}\left\{\widehat{f}_n(x)\right\} = \mathbb{E}\left\{f(\mathcal{K}_{x,h})\right\}. \tag{4.7}$$

Démonstration : En effet, nous trouvons successivement

$$\begin{aligned}
\mathbb{E}\left\{\widehat{f}_n(x)\right\} &= \mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i)\right\} \\
&= \mathbb{E}\left\{K_{x,h}(X_1)\right\} \\
&= \sum_{y\in\aleph x,h} K_{x,h}(y) f(y) \\
&= \sum_{y\in\aleph x,h} f(y)\Pr(\mathcal{K}_{x,h} = y) \\
&= \mathbb{E}\left\{f(\mathcal{K}_{x,h})\right\}.\blacksquare
\end{aligned}$$

Propriété 2 : Soit f une fonction discrète de support $\aleph$. Soit $\widehat{f}_n$ l'estimateur de f à noyau associé discret $K_{x,h}$ sur $\aleph_{x,h}$. Nous supposons que $\forall x \in \aleph$, $\aleph_{x,h} \subseteq \aleph$. Alors, nous avons

$$\mathbb{E}\left\{\widehat{f}_n(x)\right\} = \sum_{t\in\aleph\cap\aleph_{x,h}} f(t)K_{x,h}(t) \to f(x) \text{ quand } n \to +\infty.$$

Démonstration : Nous partons de l'espérance de $\widehat{f}_n(x)$ qui est égale à $\sum_{t\in\aleph\cap\aleph_{x,h}} f(t)K_{x,h}(t)$. Nous calculons sa différence avec $f(x)$. Pour cela, $\exists\ \delta > 0$ tel que

$$\begin{aligned}
\left|\mathbb{E}\left\{\widehat{f}_n(x)\right\} - f(x)\right| &= \left|\sum_{t\in\aleph\cap\aleph_{x,h}} \left\{f(t) - f(x)\right\} K_{x,h}(t)\right| \\
&\leq \sum_{|t-x|<\delta} |f(t) - f(x)|\, K_{x,h}(t) + \sum_{|t-x|>\delta} |f(t) - f(x)|\, K_{x,h}(t).
\end{aligned}$$

Pour calculer le premier terme, nous avons recours à la définition de la continuité dans le cas discret (cette notion de continuité est différente par rapport à celle du cas continu): f est continue en $x \Leftrightarrow \forall\ \epsilon > 0$, $\exists\ \delta > 0$ tel que $\forall\ t \in]x-\delta, x+\delta[\cap \aleph_{x,h} \Rightarrow |f(t)-f(x)| < \epsilon$. Ce qui implique

$$\sum_{|t-x|<\delta} |f(t) - f(x)| K_{x,h}(t) \leq \epsilon.$$

La fonction f est discrète donc elle est bornée par 1 et nous obtenons successivement

$$\begin{aligned}
\sum_{|t-x|>\delta} |f(t) - f(x)| K_{x,h}(t) &\leq \frac{2}{\delta^2}\Pr\left(|\mathcal{K}_{x,h} - x| > \delta\right) \\
&= \frac{2}{\delta^2} Var\left(\mathcal{K}_{x,h}\right) + \frac{2}{\delta^2}\left\{\mathbb{E}\left(\mathcal{K}_{x,h}\right) - x\right\}^2.
\end{aligned}$$

Finalement, sous les deux conditions (4.3) et (4.5), toute cette quantité converge vers 0.$\blacksquare$

Propriété 3 : Soit x fixé dans $\aleph$. Le biais ponctuel de l'estimateur $\widehat{f}_n$ de f à noyau associé discret est

$$\begin{aligned} Biais\left\{\widehat{f}_n(x)\right\} &= \mathbb{E}\left\{f(\mathcal{K}_{x,h})\right\} - f(x) \\ &= f\left\{\mathbb{E}(\mathcal{K}_{x,h})\right\} - f(x) + \frac{1}{2}Var(\mathcal{K}_{x,h})f^{(2)}(x) + o(h). \end{aligned} \tag{4.8}$$

Démonstration : Par définition, le biais est la différence entre l'espérance de l'estimateur $\widehat{f}_n$ et la densité inconnue f. En effet, d'après le résultat (4.7) nous avons

$$\mathbb{E}\left\{\widehat{f}_n(x)\right\} = \mathbb{E}\left\{f(\mathcal{K}_{x,h})\right\}.$$

Or, en utilisant un develeppoment limité au point moyen $m_{x,h} = \mathbb{E}(\mathcal{K}_{x,h})$, nous obtenons

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

Et en prenant l'espérance mathématique, nous avons finalement

$$\mathbb{E}\left\{f(\mathcal{K}_{x,h})\right\} = f\left\{\mathbb{E}(\mathcal{K}_{x,h})\right\} + \frac{1}{2}Var(\mathcal{K}_{x,h})f^{(2)}(x) + o(h).\blacksquare$$

Remarque : Nous mentionnons que les $f^k(x)$ d'ordre $k \geq 1$ représentent les différences finies qui viennent remplacer les dérivées dans le cas continu et qui vérifient

$$f^{(k)}(x) = \left\{f^{(k-1)}(x)\right\}' \text{ et } f'(x) = \begin{cases} \left\{f(x+1) - f(x-1)\right\}/2 & \text{si } x \in \mathbb{N}^* \\ f(1) - f(0) & \text{si } x = 0. \end{cases}$$

Propriété 4 : Soit x fixé dans $\aleph$. La variance ponctuelle de l'estimateur $\widehat{f}_n \equiv \widehat{f}_{n,h,K}$ de f à noyau associé discret est

$$Var\left\{\widehat{f}_n(x)\right\} \doteq \frac{1}{n}f(x)\Pr(\mathcal{K}_{x,h} = x). \tag{4.9}$$

Démonstration : La variance est donnée de manière successive par

$$\begin{aligned} Var\left\{\widehat{f}_n(x)\right\} &= Var\left\{\frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i)\right\} \\ &= \frac{1}{n}Var\left\{K_{x,h}(X_1)\right\} \\ &= \frac{1}{n}\mathbb{E}\left\{K_{x,h}(X_1)\right\}^2 - \frac{1}{n}\left[\mathbb{E}\left\{K_{x,h}(X_1)\right\}\right]^2 \\ &= \frac{1}{n}\left[\sum_{y\in\aleph x,h} f(y)\left\{\Pr(\mathcal{K}_{x,h} = y)\right\}^2\right] - \frac{1}{n}\left\{\sum_{y\in\aleph x,h} f(y)\Pr(\mathcal{K}_{x,h} = y)\right\}^2 \\ &= \frac{1}{n}\left\{f(x)\sum(\mathcal{K}_{x,h}^2) - f^2(x)\right\} + O(\frac{h}{n}) \\ &= \frac{1}{n}f(x)\Pr(\mathcal{K}_{x,h} = x).\blacksquare \end{aligned}$$

Nous précisons que le terme $\sum(\mathcal{K}_{x,h}^2) := \sum_{y\in\aleph x,h}\{\Pr(\mathcal{K}_{x,h}=y)\}^2$ est majoré par 1. Le résultat final se base sur la condition (4.3) à travers la probabilité modale $\Pr(\mathcal{K}_{x,h}=x)$.

Propriété 5 : L'erreur quadratique moyenne intégrée que nous appelons MISE est

$$\begin{aligned} MISE &= \sum_{x\in\aleph}\mathbb{E}\left\{\widehat{f}_n(x)-f(x)\right\}^2 \\ &= \sum_{x\in\aleph} Biais^2\left\{\widehat{f}_n(x)\right\}+\sum_{x\in\aleph} Var\left\{\widehat{f}_n(x)\right\} \\ &= \sum_{x\in\aleph}\left\{\mathbb{E}(\mathcal{K}_{x,h})-f(x)+\frac{1}{2}Var(\mathcal{K}_{x,h})f^{(2)}(x)+o(h)\right\}^2 \\ &\quad +\sum_{x\in\aleph}\frac{1}{n}f(x)\Pr(\mathcal{K}_{x,h}=x). \end{aligned} \tag{4.10}$$

4.1 Noyau associé discret pour des données catégorielles

Dans cette partie, nous nous focalisons sur les données discrètes catégorielles (i.e. données qualitatives). Nous travaillons essentiellement sur un ensemble discret fini $\aleph \subset \mathbb{R}$. Nous signalons que durant les dernières années, il y avait une croissance considérable dans le domaine des noyaux discrets pour des données catégorielles, les premiers travaux sont dûs aux innovateurs Aitchison & Aitken (1976) puis Simonoff & Tutz (2000) et enfin, Racine & Li (2007). (voir bibliographie pour plus de détails.)

Définition 3 : *Soit X la variable aléatoire de loi d'Aitchison & Aitken que nous notons $\mathcal{D}(c;c_0,\lambda)$, où $c\in\mathbb{N}\setminus\{0,1\}$ est le cardinal du support, $c_0\in\{0,1,\ldots,c-1\}$ est le point de référence et $\lambda\in]0,1]$, de densité de probabilité sur le support $\aleph=\{0,1,\ldots,c-1\}$ définie par*

$$\Pr(X=x)=(1-\lambda)1_{x=c_0}+\frac{\lambda}{c-1}1_{x\neq c_0}.$$

Propriété 6 : L'espérance de la variable aléatoire X de loi d'Aitchison & Aitken est

$$\mathbb{E}(X)=c_0\left(1-\lambda-\frac{\lambda}{c-1}\right)+\frac{\lambda c}{2}. \tag{4.11}$$

Démonstration : L'espérance de cette variable aléatoire est donnée de manière successive par :

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{x\in\{0,1,\ldots,c-1\}} x\Pr(X=x) \\
&= \sum_{x\in\{0,1,\ldots,c-1\}} \left\{ x(1-\lambda)1_{x=c_0} + \frac{x\lambda}{c-1}1_{x\neq c_0} \right\} \\
&= \left\{ c_0(1-\lambda) + \frac{\lambda}{c-1}(0+1+\ldots+(c_0-1)+(c_0+1)+\ldots+c-1) \right\} \\
&= c_0(1-\lambda) + \frac{\lambda}{c-1}\left\{ \left(\sum_{i=0}^{c-1} i \right) - c_0 \right\} \\
&= c_0(1-\lambda) + \frac{\lambda}{c-1}\left\{ \frac{c(c-1)}{2} - c_0 \right\} \\
&= c_0\left(1-\lambda-\frac{\lambda}{c-1}\right) + \frac{\lambda c}{2}.\blacksquare
\end{aligned}$$

Propriété 7 : La variance de la variable aléatoire X de loi d'Aitchison & Aitken est

$$Var(X) = c_0^2\frac{c^2\lambda\,(1-\lambda)-\lambda c}{(c-1)^2} - c_0\frac{c^2\lambda(1-\lambda)-\lambda c}{c-1} + \frac{\lambda c}{2}\left(\frac{2c-1}{3}-\frac{\lambda c}{2}\right). \qquad (4.12)$$

Démonstration : La variance est obtenue de manière successive par :

$$\begin{aligned}
Var(X) &= \mathbb{E}(X^2) - \{\mathbb{E}(X)\}^2 \\
&= c_0^2(1-\lambda) + \frac{\lambda}{c-1}\left\{ \left(\sum_{i=0}^{c-1} i^2 \right) - c_0^2 \right\} - \left\{ c_0\left(1-\lambda-\frac{\lambda}{c-1}\right) + \frac{\lambda c}{2} \right\}^2 \\
&= c_0^2(1-\lambda) + \frac{\lambda c(2c-1)}{6} - \frac{\lambda c_0^2}{c-1} - \left\{ c_0\left(1-\lambda-\frac{\lambda}{c-1}\right) + \frac{\lambda c}{2} \right\}^2 \\
&= c_0^2\left(1-\lambda-\frac{\lambda}{c-1}\right) + \frac{\lambda c(2c-1)}{6} - c_0^2\left(1-\lambda-\frac{\lambda}{c-1}\right)^2 - \frac{\lambda^2c^2}{4} \\
&\quad -c_0\lambda c\left(1-\lambda-\frac{\lambda}{c-1}\right) \\
&= c_0^2\frac{c^2\lambda(1-\lambda)-\lambda c}{(c-1)^2} - c_0\frac{c^2\lambda(1-\lambda)-\lambda c}{c-1} + \frac{\lambda c}{2}\left(\frac{2c-1}{3}-\frac{\lambda c}{2}\right).\blacksquare
\end{aligned}$$

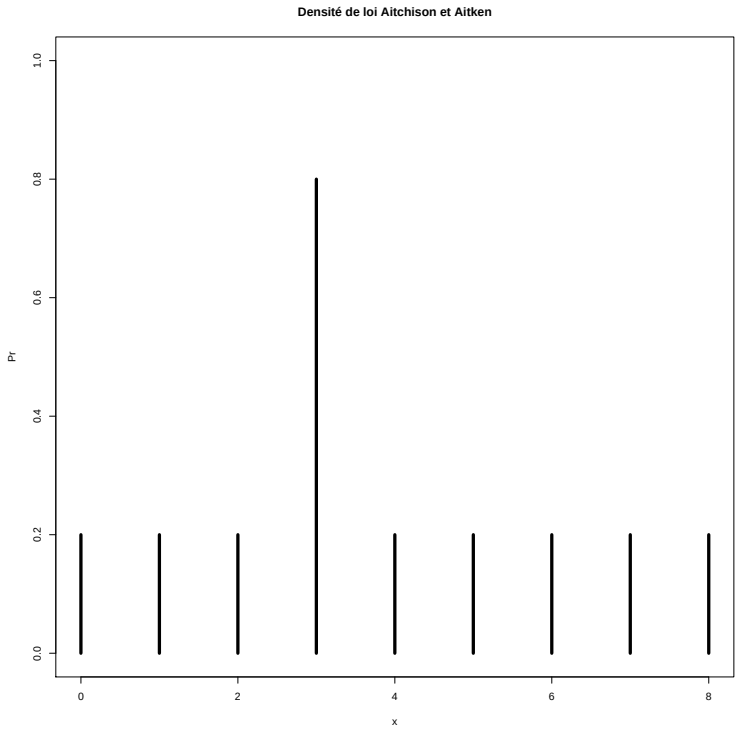
Commentaires :

a. Lorsque $c = 2$, nous nous retrouvons dans le cas d'une loi Bernoulli de paramètre λ ou $1-\lambda$. Le type de la loi Bernoulli change selon que le point de référence se trouve en 0 ou en 1. Nous verrons dans le cas de l'estimateur à noyau associé discret que le choix du point de référence sera la cible.

b. Lorsque $c \mapsto +\infty$, le support $\aleph = \mathbb{N}$.

c. Si $\lambda = 0$, ceci revient à dire que notre loi est la loi de dirac qui ne dépend plus

FIG. 4.1 – *Illustration de la loi d'Aitchison et Aitken*



de c et que nous la notons δ_x. Si maintenant, λ prend la deuxième valeur limite qui est égale à 1 alors $\Pr(X = x) = \frac{1}{c-1} 1_{x \neq c_0}$.

Nous sommes en mesure de donner une définition précise d'un estimateur à noyau associé discret pour une densité de probabilité f sur un ensemble discret $\aleph$ et de présenter les propriétés fondamentales relatives.

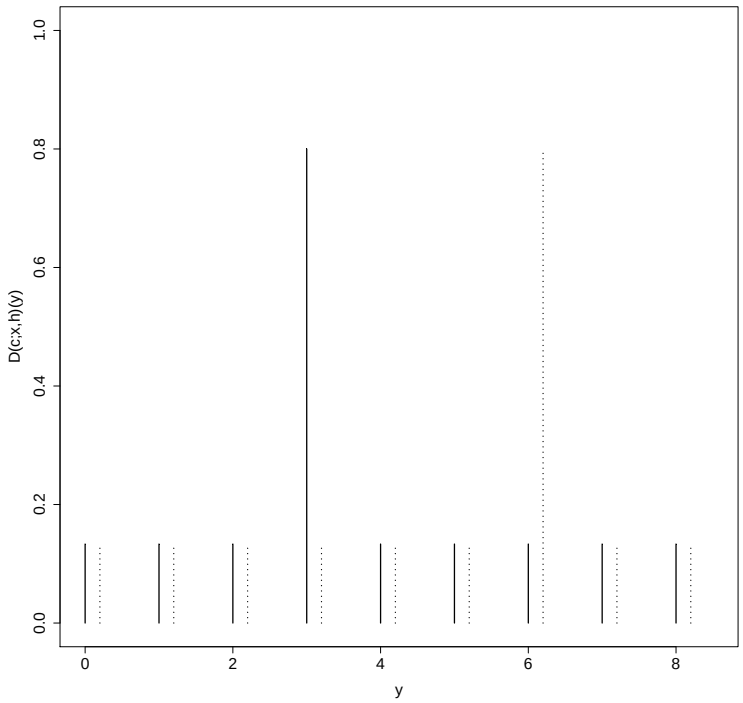
Définition 4 : *Soit $X_1, X_2, \ldots, X_n$ un échantillon de variables aléatoires i.i.d. de fonction de masse de probabilité discrète catégorielle ordonnée inconnue f sur $\aleph = \{0,1,\ldots,c-1\}$, où c est connu et fixé dans $\mathbb{N} \setminus \{0,1\}$. Un estimateur $\widehat{f}_n(x) \equiv \widehat{f}_{n,h,K}(x)$ de $f(x)$ à noyau associé discret $K_{D(c;x,h)}$ qui suit la loi d'Aitchison & Aitken est défini par*

$$\begin{aligned} \widehat{f}_n(x) &= \frac{1}{n} \sum_{i=1}^{n} K_{D(c;x,h)}(X_i) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left\{ (1-h) 1_{X_i = x} + \frac{h}{c-1} 1_{X_i \neq x} \right\}. \end{aligned} \tag{4.13}$$

avec x est dans $\aleph$ et $h \in]0,1]$ est le paramètre de lissage discret (ou encore la fenêtre).
Nous examinons les différents points que doit vérifier le noyau associé $K_{D(c;x,h)}$:

i. $\aleph_{c;x,h} = \{0,1,\ldots,c-1\} = \aleph$.

FIG. 4.2 – *Illustration du noyau associé d'Aitchison et Aitken pour* $h = 0.2$ *et* x *varié*



ii. $\cup_x \aleph_{c;x,h} = \{0,1,\ldots,c-1\} = \aleph$.

iii. $\mathbb{E}\left(\mathcal{K}_{D(c;x,h)}\right) = x\left(1-h-\frac{h}{c-1}\right) + \frac{hc}{2} \sim x$ quand $h \to 0$.

iv. $Var\left(\mathcal{K}_{D(c;x,h)}\right) = x^2 \frac{hc^2(1-h)-hc}{(c-1)^2} - x\frac{hc^2(1-h)-hc}{c-1} + \frac{hc}{2}\left(\frac{2c-1}{3} - \frac{hc}{2}\right) < \infty$.

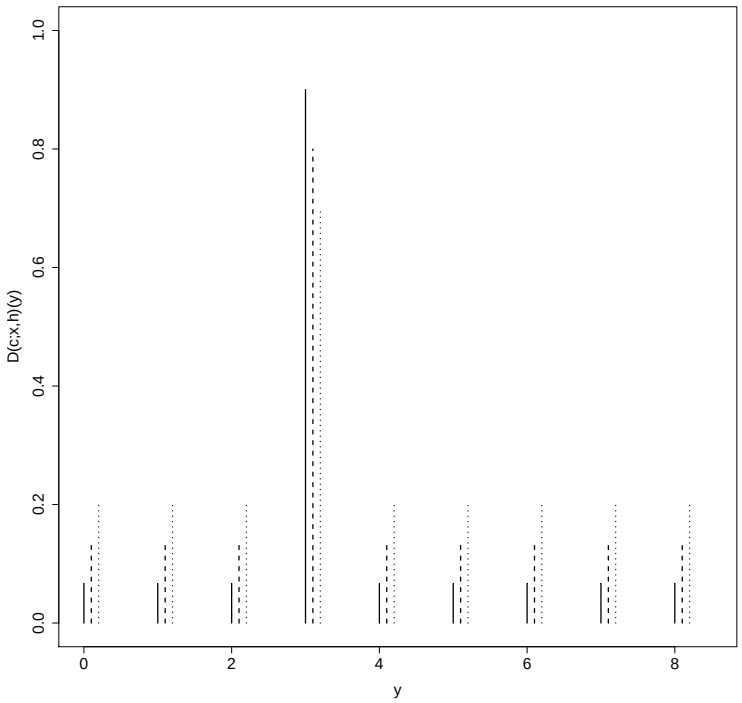
v. $h \to 0 \Rightarrow Var\left(\mathcal{K}_{D(c;x,h)}\right) = 0$.

Propriété 8 : A travers la formule (4.8), la fonction $x \mapsto \widehat{f}_n(x)$ est une fonction de masse de probabilité.

Démonstration : Comme les X_i sont i.i.d., nous avons successivement :

$$\begin{aligned}
\sum_{x=0}^{c-1} \widehat{f}_n(x) &= \sum_{x=0}^{c-1} \left\{ \frac{1}{n} \sum_{i=1}^{n} K_{D(c;x,h)}(X_i) \right\} \\
&= \sum_{x=0}^{c-1} \left\{ (1-h)1_{X_1=x} + \frac{h}{c-1} 1_{X_1 \neq x} \right\} \\
&= (1-h) + \frac{h}{c-1}(1+1+\ldots+1) \\
&= (1-h) + \frac{h}{c-1}(c-1) \\
&= 1.\blacksquare
\end{aligned}$$

FIG. 4.3 – *Illustration du noyau associé d'Aitchison et Aitken pour $x = y = 2$ et h varié*



Propriété 9 : D'après la relation (4.8), le biais de l'estimateur $\widehat{f}_n$ de f à noyau associé discret $K_{D(c;x,h)}$ de la loi d'Aitchison & Aitken est

$$Biais\left\{\widehat{f}_n(x)\right\} = \frac{hc}{2}f^{(1)}(x) + \frac{1}{2}\left\{\frac{x^2hc}{c-1} - xhc + \frac{hc}{2}\left(\frac{2c-1}{3} + \frac{hc}{2}\right)\right\}f^{(2)}(x) + o(h^2). \tag{4.14}$$

Remarque : *Nous remarquons d'après (4.14) que le biais est très important. Il dépend à la fois de c, h et des dérivées première et seconde. Chaque fois que le cardinal du support c augmente le biais s'accroit. Nous devons ainsi penser à réduire le biais en modifiant les paramètres comme dans le cas des noyaux associés asymétriques (plus précisémment les noyaux gamma et bêta de Chen, gaussien inverse et gaussien inverse réciproque de Scaillet). Nous montrons que ce n'est pas évident de déterminer ces paramètres. Cependant, une façon possible pour le réduire consiste à prendre une loi centrée en c_0 comme a fait T. Senga Kiessé pour les noyaux triangulaires (voir bibliographie pour plus de détails).*

Propriété 10 : D'après la relation (4.9), la variance de l'estimateur $\widehat{f}_n$ de f à noyau associé discret $K_{D(c;x,h)}$ de loi d'Aitchison & Aitken est

$$Var\left\{\widehat{f}_n(x)\right\} = \frac{1}{n}\left[f(x)(1-h)^2 + \left(\frac{h}{c-1}\right)^2\left\{\sum_{i=0}^{c-1} f(i) - f(x)\right\}\right]. \tag{4.15}$$

La variance de cet estimateur dépend à son tour de c, h et de la fonction inconnue f.

4.2 Noyau associé discret pour des données de comptage

Pareillement à la section précédente, nous donnons dans cette partie l'estimateur à noyau associé discret pour des données de dénombrement. Nous travaillons essentiellement sur un ensemble fini $\aleph$ (ou encore n'importe quel ensemble dénombrable notamment $\mathbb{Z}$, $\mathbb{N}+q\mathbb{N}$, etc). Nous calculons les propriétés fondamentales pour cet estimateur en utilisant les différences finies à la place des dérivées. Nous présentons dans la suite 4 exemples de noyaux associés discrets symétriques et standards asymétriques.

4.2.1 Noyau associé poissonien

Nous rappelons qu'une loi de Poisson $Po(\lambda)$ de paramètre λ est une loi discrète définie sur $\mathbb{N}$ de fonction de masse de probabilité $\Pr(X=x)$ telle que pour tout x dans $\mathbb{N}$, nous avons

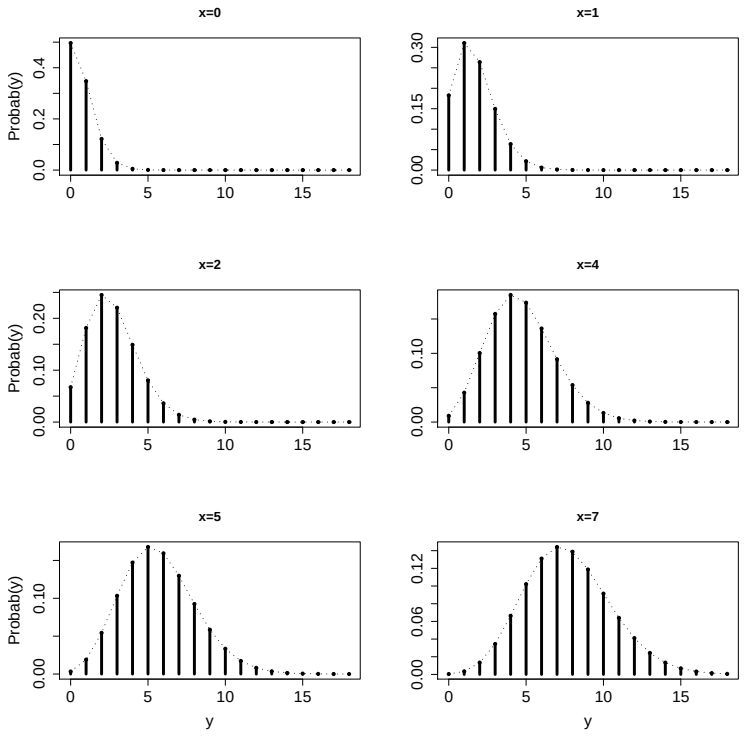
$$\Pr(X=x)=e^{-\lambda}\frac{\lambda^x}{x!}.$$

Si X est une variable aléatoire qui suit la loi Poisson, alors l'espérance et la variance sont respectivement égales à

$$\mathbb{E}(X)=\lambda \text{ et } Var(X)=\lambda.$$

La figure 4.4 illustre la variation de la fonction de masse poissonienne pour $h=0.1$.
Soit $K_{Po(x+h)}$ le noyau poissonien associé à la variable aléatoire $\mathcal{K}_{Po(x+h)}$ sur $\aleph_{x,h}=\mathbb{N}$

FIG. 4.4 – *Illustration du noyau associé poissonnien pour $h=0.1$ et x variée*



tel que :

$$K_{Po(x+h)}(y) = e^{-(x+h)}\frac{(x+h)^y}{y!},$$

avec $x \in \mathbb{N}$, $y \in \mathbb{N}$ et $h > 0$ est le paramètre de lissage discret. Ce noyau poissonien vérifie-t-il la définition d'un noyau associé ? En effet, nous avons

i. $\aleph \cap \aleph_{x,h} = \mathbb{N} \cap \mathbb{N} = \mathbb{N} \neq \varnothing$.
ii. $\cup_x \mathbb{N} = \mathbb{N}$.
iii. $\mathbb{E}\left(\mathcal{K}_{Po(x+h)}\right) = x + h \sim x$ quand $h \to 0$.
iv. $Var\left(\mathcal{K}_{Po(x+h)}\right) = x + h < \infty$.
v. $h \to 0 \Rightarrow Var\left(\mathcal{K}_{Po(x+h)}\right) = x$.

Soit X_1,X_2,...,X_n un échantillon de variables aléatoires i.i.d. de fonction de masse de probabilité discrète inconnue f définie sur un ensemble discret $\aleph = \mathbb{N}$. L'estimateur $\widehat{f}_n$ de f à noyau associé poissonien est défini par

$$\begin{aligned}\widehat{f}_n(x) &= \frac{1}{n}\sum_{i=1}^{n} K_{Po(x+h)}(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} e^{-(x+h)}\frac{(x+h)^{X_i}}{X_i!},\end{aligned}$$

avec $x \in \mathbb{N}$ et $h > 0$.
Cet estimateur est-il une fonction de masse de probabilité ? Non, en effet

$$\begin{aligned}\sum_{x=0}^{\infty}\widehat{f}_n(x) &= \sum_{x=0}^{\infty}\left\{\frac{1}{n}\sum_{i=1}^{n} K_{Po(x+h)}(X_i)\right\} \\ &= \sum_{x=0}^{\infty}\left\{K_{Po(x+h)}(X_1)\right\} \\ &= \sum_{x=0}^{\infty}\left\{e^{-(x+h)}\frac{(x+h)^{X_1}}{X_1!}\right\}.\end{aligned}$$

Nous avons calculé cette quantité numériquement sous R, pour plusieurs valeurs de h et de X_1, nous avons abouti à des valeurs très faibles par rapport à 1. Enfin, l'estimateur $\widehat{f}_n(x)$ n'est pas une fonction de masse de probabilité.
Nous évaluons ainsi le biais et la variance de l'estimateur à noyau associé poissonien.
En se basant sur la relation (4.8), le biais ponctuel de $\widehat{f}_n(x)$ en un point x fixé est

$$Biais\left\{\widehat{f}_n(x)\right\} = hf^{(1)}(x) + \frac{1}{2}(x+h)f^{(2)}(x) + o(h).$$

De même, d'après la relation (4.9), la variance de $\widehat{f}_n(x)$ en un point x fixé est

$$Var\left\{\widehat{f}_n(x)\right\} = \frac{1}{n}f(x)\frac{(x+h)^x}{x!}e^{-(x+h)}.$$

Enfin, la valeur du MISE est

$$MISE(n,h,f) = \frac{1}{n}\sum_{x\in\mathbb{N}} f(x)\frac{(x+h)^x}{x!}e^{-(x+h)} + \sum_{x\in\mathbb{N}}\left\{hf^{(1)}(x) + \frac{1}{2}(x+h)f^{(2)}(x) + o(h)\right\}^2.$$

Le MISE dépend de la densité inconnue et des ses dérivées première et seconde.

4.2.2 Noyau associé binomial

Nous rappelons qu'une loi binomiale de paramètres N et p, $B(N,p)$ est une loi discrète définie sur l'ensemble $\{0,\ldots,N\}$, avec N est un entier fixé dans $\mathbb{N}$, de fonction de masse de probabilité $g_{B(N,p)}$ telle que

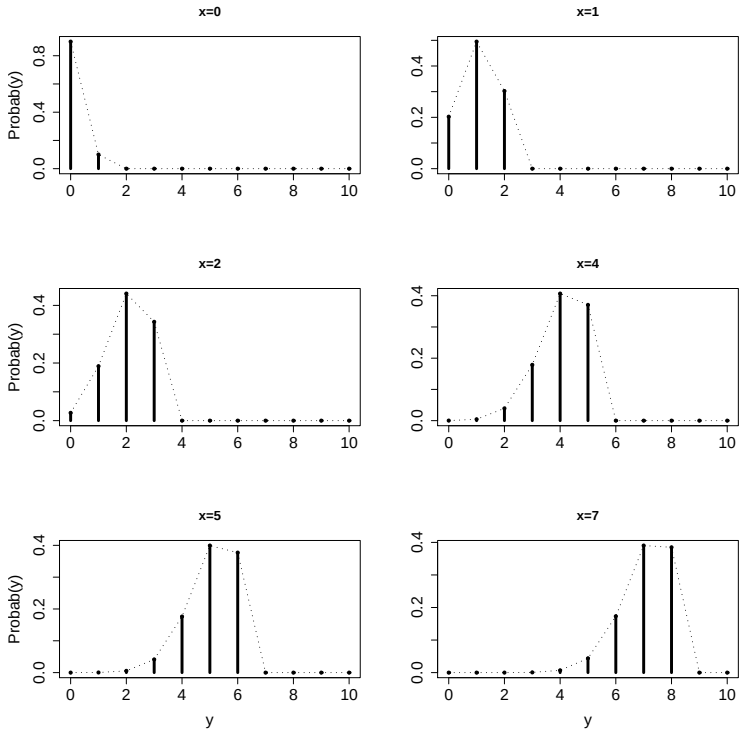
$$\Pr(X = x) = \frac{N!}{x!(N-x)!} p^x (1-p)^{N-x}.$$

Si X est une variable aléatoire qui suit la loi binomiale, alors l'espérance et la variance sont respectivement

$$\mathbb{E}(X) = Np \text{ et } Var(x) = Np(1-p).$$

La figure 4.5 présente l'allure de la densité d'une loi binomiale quand nous fixons la fenêtre h et nous faisons varier x. Par ailleurs, la figure 4.6 donne la variation de cette densité quand nous varions h et nous gardons x fixé en 7.

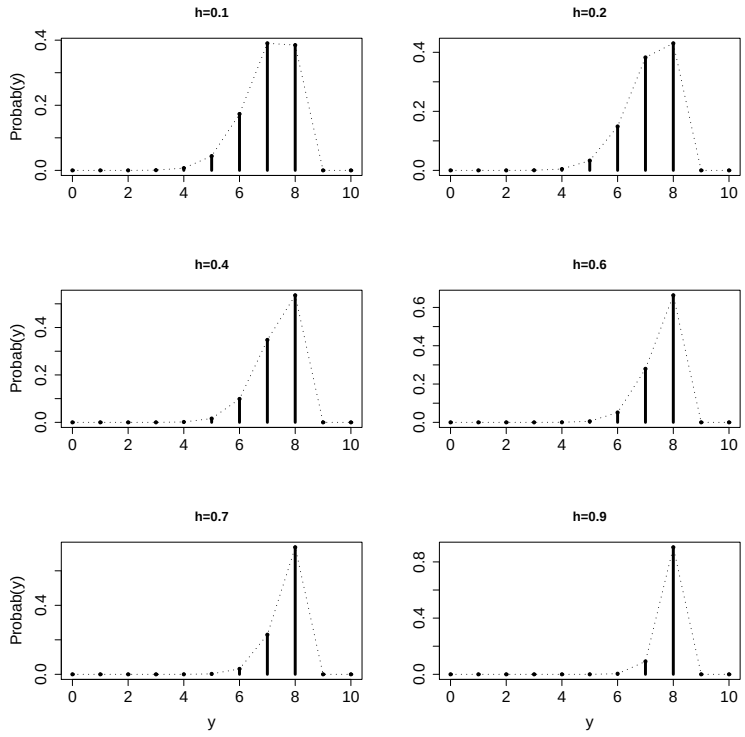
FIG. 4.5 – *Illustration du noyau associé binomial pour $h = 0.1$ et x varié.*



Le noyau $K_{B(x+1,(x+h)/(x+1))}$ est le noyau discret associé à la variable aléatoire $\mathcal{K}_{B(x+1,(x+h)/(x+1))}$ de loi binomiale défini sur le support $\aleph_{x,h} = \{0,1,\ldots,x+1\}$ tel que

$$K_{B(x+1,(x+h)/(x+1))}(y) = \frac{(x+1)!}{y!(x+1-y)!} \left(\frac{x+h}{x+1}\right)^y \left(\frac{1-h}{x+1}\right)^{x+1-y},$$

FIG. 4.6 – *Illustration du noyau associé binomial pour $x == y7$ et h varié.*



où x est dans $\mathbb{N}$ et h est dans [0,1]. Nous vérifions à ce niveau que $K_{B(x+1,(x+h)/(x+1))}$ est un noyau associé. En effet, nous avons

i. $\aleph_{x,h} \cap \aleph = \{0,1,\ldots,x+1\} \cap \mathbb{N} = \{0,1,\ldots,x+1\} \neq \varnothing$.

ii. $\cup_{x\in\mathbb{N}} \{0,1,\ldots,x+1\} = \mathbb{N}$.

iii. $\mathbb{E}(\mathcal{K}_{B(x+1,(x+h)/(x+1))}) = (x+1)(x+h)/(x+1) = x+h \sim x$ quand $h \to 0$.

iv. $Var(\mathcal{K}_{B(x+1,(x+h)/(x+1))}) = (x+h)\left(\frac{1-h}{x+1}\right) < \infty$.

v. $h \to 0 \Rightarrow Var(\mathcal{K}_{B(x+1,(x+h)/(x+1))}) = \frac{x}{x+1} < 1$.

Pour le même échantillon de variables aléatoires considéré, nous donnons l'estimateur $\widehat{f_n}$ de f à noyau associé binomial défini sur $\aleph_{x,h} = \{0,1,\ldots,x+1\}$ comme étant

$$\begin{aligned}\widehat{f_n}(x) &= \frac{1}{n}\sum_{i=1}^{n} K_{B(x+1,(x+h)/(x+1))}(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{(x+1)!}{X_i!(x+1-X_i)!}\left(\frac{x+h}{x+1}\right)^{X_i}\left(\frac{1-h}{x+1}\right)^{x+1-X_i},\end{aligned}$$

avec $x \in \mathbb{N}$ et $h \in]0,1]$.

D'après (4.8), le biais est

$$Biais\left\{\widehat{f_n}(x)\right\} = hf^{(1)}(x) + \frac{1}{2}(x+h)\left(\frac{1-h}{x+1}\right)f^{(2)}(x) + o(h).$$

Pareillement, d'après (4.9), la variance est

$$Var\left\{\widehat{f}_n(x)\right\} = \frac{1-h}{n}\left(\frac{x+h}{x+1}\right)^x f(x).$$

Enfin, le MISE est obtenue en sommant les deux quantités calculées précédemment. Nous trouvons

$$\begin{aligned} MISE(n,h,f) &= \frac{1-h}{n}\sum_{x\in\mathbb{N}}\left(\frac{x+h}{x+1}\right)^x f(x) \\ &\quad + \sum_{x\in\mathbb{N}}\left\{hf^{(1)}(x) + \frac{1}{2}(x+h)\left(\frac{1-h}{x+1}\right)f^{(2)}(x) + o(h)\right\}^2. \end{aligned}$$

4.2.3 Noyau associé binomial négatif

Nous rappelons qu'une loi binomiale négative de paramètres s et p, $BN(s,p)$ est une loi discréte définie sur le support $\mathbb{N}$ de fonction de masse de probabilité $g_{BN(s,p)}$ telle que

$$g_{BN(s,p)}(x) = \frac{(x+s)!}{x!s!}p^s(1-p)^x.$$

Si X est une variable aléatoire qui suit la loi binomiale négative, alors l'espérance et la variance sont respectivement

$$\mathbb{E}(X) = s(1-p)/p \text{ et } Var(x) = s(1-p)/p^2.$$

Soit $K_{BN(x+1,(x+1)/(2x+1+h))}$ le noyau associé à la variable aléatoire $\mathcal{K}_{BN(x+1,(x+1)/(2x+1+h))}$ de loi binomiale négative défini sur le support $\aleph_{x,h} = \mathbb{N}$ tel que

$$K_{BN(x+1,(x+1)/(2x+1+h))}(y) = \frac{(x+y)!}{y!x!}\left(\frac{x+h}{2x+1+h}\right)^y\left(\frac{x+1}{2x+1+h}\right)^{x+1},$$

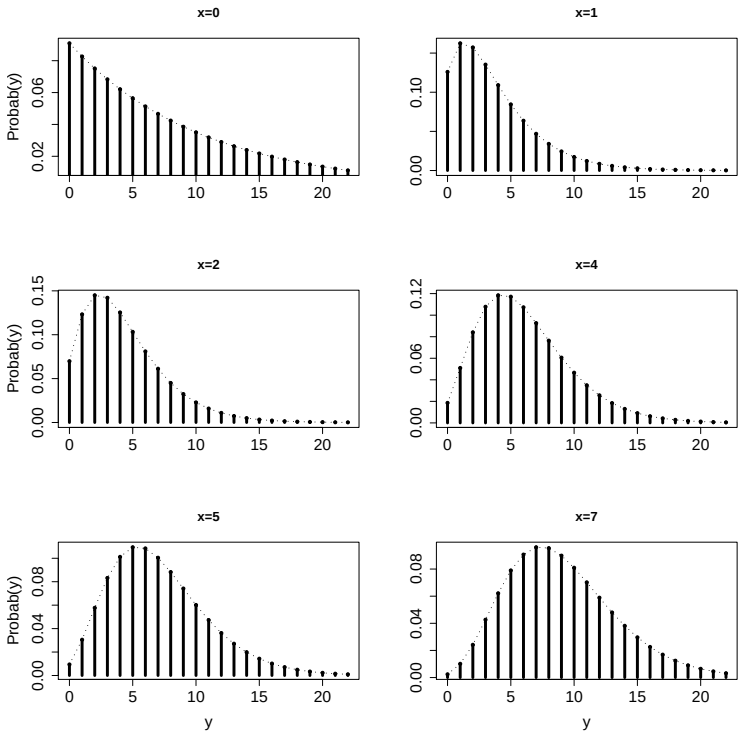
où x et y appartiennent à $\mathbb{N}$ et h est strictement positif. Nous vérifions qu'il s'agit d'un noyau associé :

i. $\mathbb{N}\cap\mathbb{N} = \mathbb{N} \neq \varnothing$.
ii. $\cup_x \aleph_{x,h} = \cup_x \mathbb{N} = \mathbb{N}$.
iii. $\mathbb{E}(\mathcal{K}_{BN(x+1,(x+h)/(2x+1+h))}) = x+h \sim x$ quand $h \to 0$.
iv. $Var(\mathcal{K}_{B(x+1,(x+h)/(x+1))}) = (x+h)\left(\frac{2x+1+h}{x+1}\right) < \infty$.
v. $h \to 0 \Rightarrow Var(\mathcal{K}_{B(x+1,(x+h)/(x+1))}) = x\left(\frac{2x+1}{x+1}\right)$.

Pour notre même échantillon de variables aléatoires, nous donnons l'estimateur $\widehat{f}_n$ de f à noyau associé binomial négatif défini sur $\aleph_{x,h} = \mathbb{N}$ comme étant

$$\begin{aligned} \widehat{f}_n(x) &= \frac{1}{n}\sum_{i=1}^n K_{BN(x+1,(x+1)/(2x+1+h))}(X_i) \\ &= \frac{1}{n}\sum_{i=1}^n \frac{(x+X_i)!}{X_i!x!}\left(\frac{x+h}{2x+1+h}\right)^{X_i}\left(\frac{x+1}{2x+1+h}\right)^{x+1}, \end{aligned}$$

FIG. 4.7 – *Illustration du noyau associé binomial négative pour $h = 0.1$ et x varié*



avec $x \in \mathbb{N}$ et $h \in \mathbb{R}_+^*$.
Le biais de cet estimateur est

$$Biais\left\{\widehat{f}_n(x)\right\} = hf^{(1)}(x) + \frac{1}{2}(x+h)\left(\frac{2x+1+h}{x+1}\right)f^{(2)}(x) + o(h).$$

D'après (4.9), la variance est

$$Var\left\{\widehat{f}_n(x)\right\} = \frac{1}{n}\frac{2}{x!}\left(\frac{x+h}{2x+1+h}\right)^x\left(\frac{x+1}{2x+1+h}\right)^{x+1} f(x).$$

En final, le MISE est la somme des deux derniers resultats. Il est égal à

$$\begin{aligned} MISE(n,h,f) &= \frac{1}{n}\sum_{x\in\mathbb{N}}\frac{2}{x!}\left(\frac{x+h}{2x+1+h}\right)^x\left(\frac{x+1}{2x+1+h}\right)^{x+1} f(x) \\ &\quad + \sum_{x\in\mathbb{N}}\left\{hf^{(1)}(x) + \frac{1}{2}(x+h)\left(\frac{2x+1+h}{x+1}\right)f^{(2)}(x) + o(h)\right\}^2. \end{aligned}$$

4.2.4 Noyau associé triangulaire

En se référant aux travaux de Kokonendji et Senga Kiessé (2007) sur les distributions triangulaires discrètes, nous rappelons qu'une loi triangulaire $T_{a,h,c}$ de paramètres a

et c dans $\mathbb{N}$ et h dans $\mathbb{R}_+$ est une loi discrète centrée en c et de bras a défini sur $\aleph_{a,c} = \{c,c \pm 1, \ldots, c \pm a\}$ de fonction de masse de probabilité :

$$\Pr\left(T_{a,h,c} = y\right) = \frac{(a+1)^h - |y-c|^h}{P(a,h)},$$

où $P(a,h)$ est la constante de normalisation telle que

$$P(a,h) = (2a+1)(a+1)^h - 2\sum_{i=0}^{a} i^h.$$

Nous remarquons que le cas $h = 1$ correspond à la variable aléatoire triangulaire symétrique. Le cas $h \leq 0$ n'est pas défini en c et en particulier, si $h = 0$ nous nous retrouvons la loi de Dirac d'espérance c. Si h tend vers l'infini, nous trouvons la loi uniforme. Pour des entiers non nuls $h \in \mathbb{R}^*$, la constante de normalisation peut s'écrire :

$$P(a,h) = (2a+1)(a+1)^h - 2\sum_{i=0}^{a} \frac{(-1)^{h-i+1)} h! B_{h-i+1}}{i!(h-i+1)!} a^i,$$

où B_{h-i+1} est le nombre de Bernoulli. La figure 4.8 présente l'allure de la densité triangulaire par rapport aux autres noyaux discrète que nous avons étudié.
Si X est une variable aléatoire qui suit la loi triangulaire alors l'espérance et la variance sont respectivement :

$$E(X) = c \text{ et } Var(X) = \frac{1}{P(a,h)} \left(\frac{a(a+1)^{h+1}(2a+1)}{3} - 2\sum_{i=0}^{a} i^{h+2} \right).$$

La loi $T_{a,h,c}$ est symétrique autour de sa moyenne. De plus, la variance ne dépend pas de c.

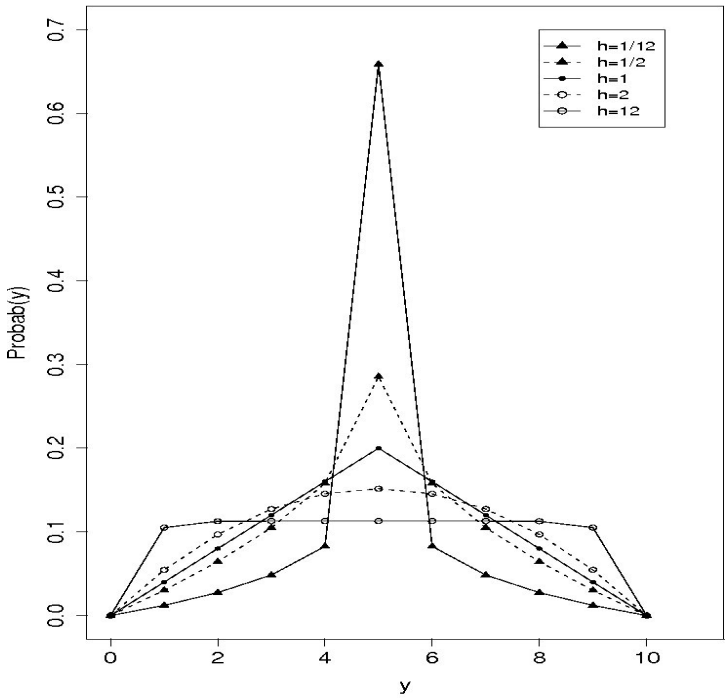
Soit $K_{T(a,h,x)}$ le noyau triangulaire associé à la variable aléatoire $\mathcal{K}_{T(a,h,x)}$, défini sur $\{x,x \pm 1, \ldots, x \pm a\}$ et donné par

$$K_{T(a,h,x)}(y) = \frac{(a+1)^h - |y-x|^h}{(2a+1)(a+1)^h - 2\sum_{j=0}^{a} j^h},$$

avec $x \in \mathbb{N}$, $h > 0$ et $a \in \mathbb{N}$.
Nous nous assurons des différents points de la définition 1.

i. $\{x,x \pm 1, \ldots, x \pm a\} \cap \mathbb{N} = \{x,x \pm 1, \ldots, x \pm a\} \neq \varnothing$.
ii. $\cup_{x \in \mathbb{N}} \{x,x \pm 1, \ldots, x \pm a\} = \mathbb{N}$.
iii. $\mathbb{E}\left(\mathcal{K}_{T(a,h,x)}\right) = x$.
iv. $Var\left(\mathcal{K}_{T(a,h,x)}\right) = \frac{1}{P(a,h)} \left(\frac{a(a+1)^{h+1}(2a+1)}{3} - 2\sum_{j=0}^{a} j^{h+2} \right) < \infty$.
v. Lorsque $h \to 0$, la variance de $\mathcal{K}_{T(a,h,x)}$ tend aussi vers 0. En effet, ce résultat a été obtenu dans la proposition (2.4) des travaux de Kokonendji, Senga Kiessé et Zocchi (2007). Dans cette proposition, nous montrons que la variance de la variable aléatoire converge vers une loi de Dirac, ce qui implique une variance nulle (voir aussi la remarque 2.3(ii)).

FIG. 4.8 – *Illustration du noyau associé triangulaire pour différentes valeurs de h.*



Soit $X_1,\ldots,X_n$ l'échantillon de variables aléatoires i.i.d. de densité f inconnue définie sur $\mathbb{N}$. Nous donnons l'estimateur $\widehat{f}_n$ de f à noyau associé triangulaire défini sur $\aleph_{x,h} = \{x,x\pm 1,\ldots,x\pm a\}$ comme étant :

$$\begin{aligned}\widehat{f}_n(x) &= \frac{1}{n}\sum_{i=1}^{n} K_{T(a,h,x)}(X_i)\\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{(a+1)^h - |X_i - x|^h}{(2a+1)(a+1)^h - 2\sum_{j=0}^{a} j^h},\end{aligned}$$

avec $x\in\mathbb{N}$, $h>0$ et $a\in\mathbb{N}$. Le noyau $K_{T(a,h,x)}$ est le noyau associé défini sur $\aleph_{a,x,h} = \{x,x\pm 1,\ldots,x\pm a\}$. Nous remarquons que le support du noyau associé ne dépend pas de h. Si $a=0$, alors $\aleph_{0,x}=\{x\}$ et $\cup_x \aleph_{0,x} = \mathbb{N}$. Par contre, si $a\neq 0$ nous avons

$$\cup_{x\in\mathbb{N}}\aleph_{a,x} = \{-a,\ldots,-1\}\cup\mathbb{N}. \tag{4.16}$$

Le fait que le support du noyau discret triangulaire (4.16) à $a\neq 0$ fixé contienne strictement $\mathbb{N}$ induit un biais de bordure à gauche du support de f. Nous y remédions en modifiant le bras a par a_0 de sorte que, $\forall a_0$ nous avons

$$\cup_{x\in\mathbb{N}}\aleph_{a_0,x} = \mathbb{N}.$$

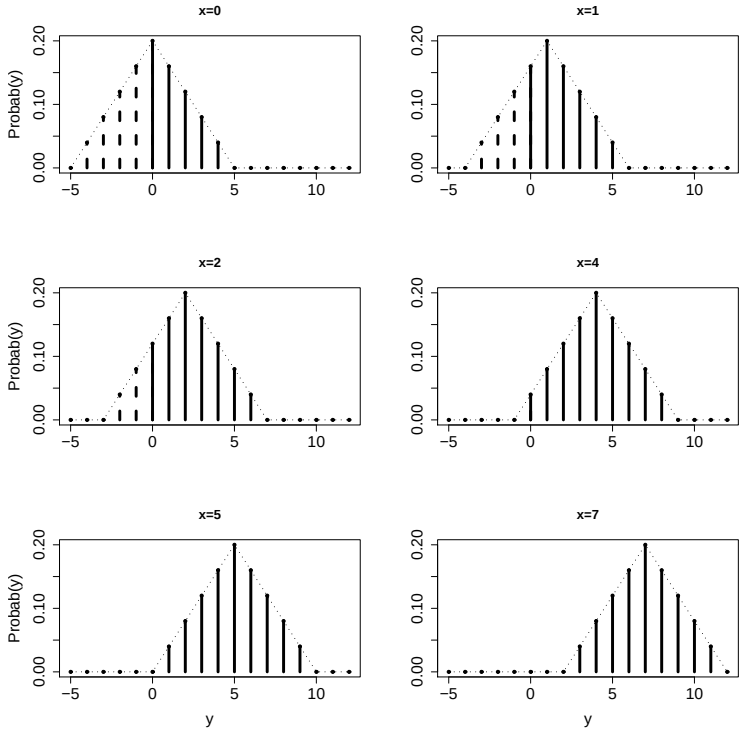
Nous proposons ainsi une solution à ce problème de biais de bordure. En cas d'observations importantes au bord $\{0,1,\ldots,k\}$ du support $\aleph=\mathbb{N}$ de f (k très petit, de l'ordre

de 0, 1 ou 2), nous considérons le bras modifié a_0 de a tel que $\forall k \in \mathbb{N} \setminus \{0\}$ donné et $x \in \mathbb{N}$, nous avons

$$a_0 = k \Leftrightarrow a = \begin{cases} \text{j} & \text{si } x = j \in \{0,1,\ldots,k-1\} \\ k & \text{si } x \in \{k,k+1,\ldots\} \ . \end{cases}$$

Nous illustrons ce problème du biais de bordure dans les figures 4.9 et 4.10. Nous avons fixé $h = 1$, $a = 4$ et $a_0 = 4$.

FIG. 4.9 – *Illustration du noyau associé triangulaire sans modification du bras*



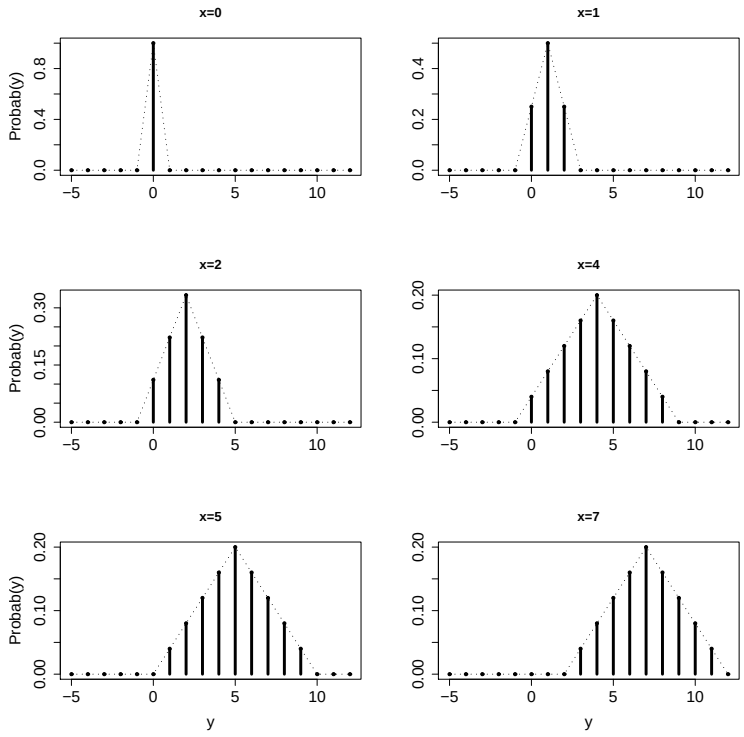
Le biais de cet estimateur est

$$Biais\left\{\widehat{f}_n(x)\right\} = \frac{1}{2}\frac{1}{P(a,h)}\left(\frac{a(a+1)^{h+1}(2a+1)}{3} - 2\sum_{i=0}^{a} i^{h+2}\right) f^{(2)}(x) + o(h).$$

D'après (4.9), la variance est

$$Var\left\{\widehat{f}_n(x)\right\} \doteq \frac{(a+1)^h}{nP(a,h)} f(x),$$

FIG. 4.10 – *Illustration du noyau associé triangulaire avec modification du bras*



où $P(a,h) = (2a+1)(a+1)^h - 2\sum_{j=0}^{a} j^h$.
En final, le MISE est la somme des deux derniers resultats. Il est égal à :

$$MISE(n,h,f) = \frac{(a+1)^h}{nP(a,h)} \sum_{x\in\mathbb{N}} f(x) + \sum_{x\in\mathbb{N}} \left\{ \frac{1}{2}\frac{1}{P(a,h)} \left(\frac{a(a+1)^{h+1}(2a+1)}{3} - 2\sum_{i=0}^{a} i^{h+2} \right) f^{(2)}(x) + o(h) \right\}^2 .$$

f. Remarques :

a. *Nous remarquons qu'il existe des lois discrètes qui ne peuvent pas être associées à aucun noyau discret notamment la loi uniforme discrète. En effet, si nous considérons la loi uniforme discrète centrée en x et de largeur $2a$, le noyau associé $U_{x,a}$ défini sur $\aleph_{x,a} = \{x, x\pm 1, \ldots, x\pm a\}$ s'écrit comme suit :*

$$U_{x,a}(y) = \frac{1}{2a+1} 1_{x,x\pm 1,\ldots,x\pm a}(y),$$

où y est dans $\mathbb{N}$. Nous apercevons que les paramètres propres de cette loi se trouvent sur l'ensemble des valeurs entières $\mathbb{N}$. Or le paramètre de lissage h est dans $\mathbb{R}_+^*$ ce qui fait que nous ne pouvons pas créer une substitution au niveau de ces paramètres.

b. *Le rôle du paramètre de lissage discret $h > 0$ reste semblable au cas continu, car il permet de tenir compte des observations X_i qui sont proches de la cible $x \in \mathbb{N}$ lorsque $h = h(n) \to 0$. Cependant la dispersion locale en tout point d'estimation x se traduit par l'importance du noyau associé discret $K_{x,h}$ choisi. Ainsi, le choix d'un type de noyau discret s'oriente vers des distributions de $K_{x,h}$ qui soient moins dispersées autour de $x \in \mathbb{N}$ et $h > 0$ fixées.*

c. *Pour une loi triangulaire $T_{a,h,x}$, si $a = 0$ alors la loi discrète $T_{0,h,x}$ correspond à une loi de Dirac $D(x)$ en x. Nous donnons le noyau associé discret de loi de Dirac (noyau "naïf") $D_{x,0}$. Pour tout $x \in \mathbb{N}$ et $h > 0$,*

$$D_{x,0}(y) = \delta_x(y),\ y \in \mathbb{N}.$$

d. *Nous remarquons que la cinquième condition de la définition d'un noyau associé n'est pas vérifiée dans le cas d'un noyau discret standard tel que le noyau poissonnien, binomial et binomial négatif. Faut il donc avoir une deuxième définition pour ces types de noyau?*

4.2.5 Choix de fenêtres

Nous présentons à ce niveau trois méthodes de choix de fenêtres pour approcher la valeur idéale de la fenêtre h définie par

$$h_{id} = \arg\min_{h>0} MISE(n,h,K,f) = h_{id}(n,K,f). \tag{4.17}$$

a. Minimisation des erreurs quadratiques
Du point de vue purement pratique où $\underline{X} = (X_1, \ldots, X_n)$ est un échantillon de variables aléatoires de fonction de masse de probabilité f, associé aussi à la distribution empirique f_0 de f, nous proposons maintenant quelques types de fenêtres liées aux erreurs d'estimations. La première est déduite de *l'erreur quadratique intégrée* (en anglais " Integrated Squared Error") définie par

$$ISE := \sum_{x\in\mathbb{N}} \left\{\widehat{f}_n(x) - f(x)\right\}^2 = ISE(\underline{X}; h,K,f), \tag{4.18}$$

laquelle mesure sur un seul échantillon $\underline{X}$ l'écart (au sens quadratique) entre $\widehat{f}$ et f. Par conséquent, la minimisation en h de l'ISE (4.18) conduit à choisir une fenêtre adéquate

$$h^{**} = \arg\min_{h>0} ISE(\underline{X}; h,K,f) = h^{**}(n,K,f). \tag{4.19}$$

En remplaçant f par f_0 dans (4.19), nous utilisons $h_0^{**} = h^{**}(n,K,f_0)$ pour le lissage discret d'un f_0 de f. Autrement dit, nous avons

$$h_0^{**} = \arg\min_{h>0} ISE(\underline{X}; h,K,f) = h^{**}(n,K,f_0). \tag{4.20}$$

Basé sur la convergence de f_0 vers f quand $n \to +\infty$, nous avons immédiatement

$$\lim_{n\to+\infty} h_0^{**}(n,K,f_0) = \lim_{n\to+\infty} h^{**}(n,K,f), \tag{4.21}$$

pour un type de noyau associé K donné. L'importance de la fenêtre adéquate h^{**} (4.19) de h est due, en partie, aux relations suivantes :

$$MISE = \mathbb{E}\left(ISE\right) = \sum_{x \in \aleph} MSE\left(x\right). \tag{4.22}$$

b. Validation croisée

Tout comme le cas continu, la méthode classique de validation croisée (en anglais "Cross Validation") ne fait 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 h_{id} (4.17) de h.

Le principe de cette méthode est de minimiser par rapport à h un estimateur de MISE pour trouver le paramètre optimal. Pour cela, la forme du MISE peut être développé comme suit :

$$MISE = \mathbb{E}\left\{\sum_{x \in \mathbb{N}} \widehat{f}_n^2(x)\right\} - 2\mathbb{E}\left\{\sum_{x \in \mathbb{N}} \widehat{f}_n(x) f(x)\right\} + \sum_{x \in \mathbb{N}} f(x)^2.$$

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

$$MISE_{cv} = \mathbb{E}\left\{\sum_{x \in \mathbb{N}} \widehat{f}_n^2(x)\right\} - 2\mathbb{E}\left\{\sum_{x \in \mathbb{N}} \widehat{f}_n(x) f(x)\right\} = MISE_{cv}(h),$$

le terme $MISE$ qui dépend de h. Dans la suite, nous déterminons un estimateur $CV(h)$ de $MISE_{cv}$. D'abord, nous avons évidemment $\sum_{x \in \mathbb{N}} \widehat{f}_n^2(x)$ qui est un estimateur sans biais de $\mathbb{E}\left\{\sum_{x \in \mathbb{N}} \widehat{f}_n^2(x)\right\}$.

Ensuite, soit

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

Par construction,

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

est un estimateur de $\mathbb{E}\left\{\sum_{x \in \mathbb{N}} \widehat{f}_n(x) f(x)\right\}$ et on vérifie de plus qu'il est sans biais. En effet, d'une part, comme les X_i sont i.i.d., nous avons

$$\begin{aligned} \widehat{G} &= \mathbb{E}\left\{\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq 1} K_{X_1,h}\left(X_j\right)\right\} \\ &= \mathbb{E}\left\{\frac{1}{n-1} \sum_{j \neq 1} K_{X_1,h}\left(X_j\right)\right\} \\ &= \mathbb{E}\left\{K_{X_1,h}\left(X_2\right)\right\}. \end{aligned}$$

Finalement, nous venons de montrer que

$$\begin{aligned} CV(h) &= \sum_{x\in\mathbb{N}} \widehat{f}_n^2(x) - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_{n,-i}(X_i) \\ &= \sum_{x\in\mathbb{N}} \left\{ \frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i) \right\}^2 - \frac{2}{n(n-1)} \sum_{i=1}^{n}\sum_{j\neq 1} K_{X_i,h}(X_j). \end{aligned} \tag{4.23}$$

est un estimateur sans biais de $MISE_{cv}$. Par conséquent, la fenêtre optimale par la méthode de la validation croisée s'obtient par

$$h_{cv} = \arg\min_{h>0} CV(h) \tag{4.24}$$

où CV(h) est donné en (4.23). Pour quelques détails, nous pouvons nous référer à de nombreux auteurs tels Bowman (1984), Marron (1984), Rudemo (1982), Stone (1984) et leurs références.

c. Excès de zéros
Pour cette section, le choix de la fenêtre repose sur une particularité des données de comptage avec $\aleph = \mathbb{N}$ qui n'est autre que l'excès des zéros dans l'échantillon $\underline{X} = (X_1,\ldots,X_n)$. Pour ce phénomène bien connu (voir, par exemple, Kokonendji *et al.*, 2007, et leurs références) et étant donné un noyau discret associé $K_{x,h}$, nous pouvons choisir une fenêtre adaptée $h_0 = h_0(\underline{X}; K)$ de h satisfaisant

$$\sum_{i=1}^{n} \Pr(\mathcal{K}_{X_i,h_0} = 0) = n_0, \tag{4.25}$$

où n_0 désigne le nombre des zéros dans $\underline{X}$; voir Marsh & Mukhopadhyay (1999) pour leur noyau du type poissonnien. Cette fenêtre h_0 ajuste le nombre de zéros théorique au nombre de zéros observé.
L'équation (4.25) s'obtient à partir de l'expression

$$\mathbb{E}\left\{\widehat{f}_n(x)\right\} = \sum_{y\in\mathbb{N}} \Pr(\mathcal{K}_{x,h})\, f(y),$$

dans laquelle nous prenons $y = 0$ et $f(0) = 1$ afin d'identifier le nombre de zéros théoriques au nombres de zéros empiriques n_0.
Dans le cas du noyau associé poissonien, la fenêtre adaptée h_0 est connue explicitement. Tandis que dans le cas des noyaux associés binomial et binomial négatif, la fenêtre h_0 est obtenue par la résolution numérique d'une équation non-linéaire (voir Table 4.1).

4.3 Noyau associé discret multiple

Nous généralisons l'estimateur à noyau associé discret au cas multidimensionnel. Pour cela, nous considérons un échantillon de variables aléatoires $X_1,\ldots,X_n$ i.i.d., de

TAB. 4.1 – *Solutions h_0 pour les noyaux associés discrets standards*

Type de noyau	h_0
Poisson	$h_0 = \log\left(\frac{1}{n_0}\sum_{i=1}^{n} e^{X_i}\right)$
Binomial	$\sum_{i=1}^{n}\left(\frac{1-h_0}{X_i+1}\right)^{X_i+1} = n_0$
Binomial négatif	$\sum_{i=1}^{n}\left(\frac{X_i+1}{2X_i+1+h_0}\right)^{X_i+1} = n_0$

fonction de masse de probabilité f et inconnue defini sur $\aleph = \mathbb{N}$ de dimension d. L'estimateur $\widehat{f}_n$ de f à noyau asocié discret est

$$\widehat{f}_n(\underline{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{\underline{x},H}(X_i), \tag{4.26}$$

où la cible $\underline{x} =^t (x_1,\ldots,x_d)$, H est la matrice pleine inversible de variance-covariance des fenêtres h de dimension $d\times d$ (présentée dans la section 2.2), et $X_i =^t (X_{i1},\ldots,X_{id})$. La fonction $K_{\underline{x},H}$ est le noyau associé asymétrique sur $\aleph_{x,h}$.
Dans le but d'avoir une forme plus sympathique et qui ne dépend pas des coefficients de corrélation, nous présentons l'estimateur (4.26) qui utilise le produit des noyaux associés univariés. En effet, nous avons

$$\widehat{f}_n(\underline{x}) = \frac{1}{n}\sum_{i=1}^{n}\left\{\prod_{j=1}^{d} K^j_{x_j,h_j}(X_{ij})\right\}, \tag{4.27}$$

où x_j est la jème composante du vecteur $\underline{x}$, h_j est la jème fenêtre et X_{ij} est la ième observation de la jème composante. Le noyau associé K^j est la fonction noyau associé univarié décrite tout au long de cette partie.

Chapitre 5

Régression multiple à noyaux associés mixtes

Nous rappelons que si nous avons un couple de variables aléatoires réelles telles que Y soit intégrable ($\mathbb{E}(Y) < \infty$) alors la fonction

$$r(x) = \mathbb{E}(Y|X = x)$$

est appelée fonction de régression de Y sur X où nous n'avons aucune spécification sur $r(x)$, avec $x \in \aleph$. Supposons que nous disposons de n-échantillon $(X_1,Y_1), \ldots, (X_n,Y_n)$ de variables aléatoires de même loi que (X,Y), de densité (fonction de masse) de probabilité inconne. Nous nous proposons ainsi de construire un estimateur $\widehat{r}_n$ de la fonction densité (de masse) inconnue. En effet, dans l'étude de la régression non-paramétrique, nous distinguons deux modèles principaux ; la régression non-paramétrique à effets aléatoires et la régression non-paramétrique à effets fixes. Dans le premier cas, les observations X_i sont aléatoires, alors que dans le cas d'effets fixes les X_i sont i.i.d., fixé dans $\mathbb{R}$ $(X_i = i/n)$ et déterministes.
Soit ainsi le modèle général

$$Y_i = r(X_i) + \epsilon_i \text{ pour } i = 1, \ldots, n, \tag{5.1}$$

où les ϵ_i sont i.i.d., non corrélés avec X_i, de moyenne nulle et de variance σ^2.

5.1 Estimateur de Nadaraya-Watson

Il existe plusieurs types d'estimateurs à noyau pour la régression dont le plus fameux est celui de Nadaraya-Watson. Dans le cas univarié, l'estimateur de Nadaray-Watson de la fonction régression r est défini par

$$\widehat{r}_n(x) = \frac{\sum_{i=1}^{n} Y_i K_{x,h}(X_i)}{\sum_{i=1}^{n} K_{x,h}(X_i)}, \tag{5.2}$$

si la quantité $\sum_{i=1}^{n} K_{x,h}(X_i)$ est strictement différente de zéro et où $K_{x,h}$ est le noyau associé univarié (continu ou discret), $h > 0$ est le paramètre de lissage et $x \in \aleph$. A

contrario, l'estimateur $\widehat{r}(x)$ est nul. Nous pouvons représenter l'estimateur de Nadaraya-Watson comme une somme pondérée des Y_i :

$$\widehat{r}_n(x) = \sum_{i=1}^{n} w_{x,h}(X_i)Y_i \text{ pour } x \in \aleph, \tag{5.3}$$

où

$$w_{x,h}(X_i) = \frac{K_{x,h}(X_i)}{\sum_{i=1}^{n} K_{x,h}(X_i)} \tag{5.4}$$

est la fonction poid telle que $\sum_{i=1}^{n} w_{x,h}(X_i) = 1$, par convention $0/0{=}0$. La fonction $K_{x,h}$ est la fonction noyau associé présentée dans les chapitres précédents, défini sur $\aleph_{x,h}$. Nous pouvons mélanger plusieurs types de noyau associé à savoir les noyaux associés continus symétriques ou asymétriques avec les noyaux discrets standards. La fenêtre $h = h(n,K)$ détermine le niveau de lissage de l'estimation.
En se référant au quatrième chapitre de la thèse (en préparation) de Senga Kiessé (2008), il est convenable de donner l'estimateur de Nadaraya-Watson sous une forme plus souple. Pour cela, soit

$$\widehat{r}_n(x) = \frac{N_n(x;h)}{D_n(x;h)} \tag{5.5}$$

avec

$$N_n(x;h) = \frac{1}{n}\sum_{i=1}^{n} Y_i K_{x,h}(X_i),$$

et

$$D_n(x;h) = \frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i) = \widehat{f}_n(x).$$

Nous généralisons la définition (5.2) de cet estimateur au cas multidimensionnel. En effet, en utilisant (3.18) et (4.27), l'estimateur de Ndaraya-Watson devient

$$\widehat{r}_n(\underline{x}) = \frac{\sum_{i=1}^{n} Y_i \left\{\prod_{j=1}^{p} K^j_{x_j,h_j}(X_{ij})\right\}}{\sum_{i=1}^{n}\left(\prod_{j=1}^{p} K^j_{x_j,h_j}\right)} \tag{5.6}$$

où $\underline{x} =^t (x_{1i},\ldots,x_{pi}) \in \aleph$, $K^j_{x_j,h_j}$ est le jème noyau associé et X_{ij} est la ième observation de la jème composante.

Chapitre 6

Données de Panel à l'étude

6.1 Notions élémentaires :

Données de Panel :

Un panel est un échantillon stable de consommateurs ou de distributeurs interrogé régulièrement ou périodiquement et dont la composition ne se renouvelle que lentement. Son étude permet une analyse dynamique de la population considérée et la prise en compte du contexte concurrentiel. Le panel de distributeurs permet la collecte d'informations commerciales.
Il est ainsi possible de mesurer plus précisément la nature du référencement d'une marque ou d'un produit en fonction du type de point de vente ou encore de la zone géographique. Le panel de consommateurs procure, quant à lui, des informations marketing et revêt un intérêt particulier pour l'analyse de l'évolution du comportement d'achat des consommateurs.
Un panéliste est donc l'individu ou le ménage membre d'un panel dont nous observons le comportement et/ou les attitudes. Selon la nature du panel, la transmission des données par le panéliste peut se faire automatiquement et passivement vers le système d'information de la société d'étude ayant crée le panel.

Du Marketing Transactionnel au Marketing Relationnel :

Le marketing est traditionnellement orienté vers l'acquisition de clients et la réalisation de transactions. Dans les années 90, de nombreux facteurs vont inciter les entreprises à utiliser les nouvelles technologies, avec notamment les bases de données et les nouveaux canaux de communication personnalisables et interactifs, pour développer des programmes de fidélisation. Le marketing n'est plus simplement transactionnel, il devient aussi relationnel.
Par conséquent le Marketing relationnel, dont la vision à plus long terme devrait permettre la fidélisation du consommateur, souhaite obtenir et renforcer la fidélité du client, grâce à son consentement volontaire, à une communication personnalisée et des offres sur-mesure. La fidélisation du client et les revenus futurs qu'il peut ainsi générer sont mis en perspective dans une optique financière et comptable. Se développe dès lors la notion de " valeur à vie " (lifetime value) qui permet de définir la valeur à terme d'un client tout le temps qu'est maintenue sa relation avec l'entreprise.

TAB. 6.1 – *Tableau comparatif du marketing transactionnel et relationnel*

Le marketing transactionnel favorise	Le marketing relationnel favorise
le produit	la relation avec le client
l'acte d'achat	la durée de la relation
le moment de la transaction	l'individualisation
le montant de la transaction	la fidélisation

Valeur à vie (Lifetime Value) :

Cette notion de valeur du client ou d'une clientèle a été développée initialement par les spécialistes de la vente à distance. C'est en effet dans ce secteur que sont apparues les premières bases de données clientèle permettant de tels calculs.
En marketing direct, la valeur à vie ou " Life time Value " se définit comme étant la somme des profits actualisés attendus sur la durée de vie d'un client. Elaborée à partir de la durée de vie moyenne d'un client et de l'évolution théorique de sa consommation, la life time value doit permettre de déterminer la limite haute du coût d'acquisition client. Elle peut être surestimée par des hypothèses trop optimistes en termes de fidélité. Par ailleurs, les différentes techniques et canaux de recrutement utilisés influencent la valeur vie client. La question qui se pose :

Par quel moyen nous allons arriver à impacter la valeur d'un client?

L'objectif dans la partie qui suit est de donner, en premier lieu, un résumé statistique sur les variables d'étude ; nous étudions la contribution et la corrélation des variables principales. En second lieu, nous faisons appel aux estimateurs à noyaux associés discrets pour représenter et prédire les actes d'achats effectué par chaque panéliste. En guise d'avoir un résultat lisible et explicatif, nous nous sommes arrêtés aux 100 premières observations.

6.2 Traitements préliminaires

Par respect à l'article 15 du code de la déontologie statistique « Le statisticien dépositaire de données [...] doit respecter les obligations de secrets particulières à la source qu'il exploite.» Nous garantissons ainsi la confidentialité des données que nous possédons et nous assurons l'anonymat de nos diffuseurs.

L'enquête sujet d'étude s'est déroulée d'une manière régulière dans sept supermarchés différents que nous désignons de manière anonyme : magasin 1, magasin 2, magasin 3, magasin 4, magasin 5, magasin 6 et magasin 7. Notre étude s'est limité au magasin 1 parce qu'il présente le plus grand nombre de foyers clients. Afin d'avoir une idée claire des comportements et des attitudes de consommations, nous avons traité des variables quantitatives scalées qui permettent, par leur nature, les calculs scientifiques les plus souvent utilisées dans l'analyse multivariée. Il s'agit donc d'une étude quantitative qui vise à comparer ou à mettre en relief un certains nombre de comportements. Les bases dont nous disposons sont les données brutes stockées par le système de l'entreprise responsable de la collecte de ces données. Nous avons dû effectuer des agrégations et des fusions pour aboutir enfin à une base exhaustive. En effet, cette partie est une étape essentielle à toute étude exploratoire : ces données sont le résultat de l'agrégation effectuée sur la base "Achats" et "Foyers", qui sont stockées dans une base de données sous SPSS. Cette base contient 46 variables quantitatives et 4922 actes d'achats. Les variables présentées sont très pertinentes et définissent les caractéristiques personnelles de chacun des panélistes pour cet échantillon supposé être représentatif d'une commune française de taille moyenne. Enfin, nous nous restreignons sur un échantillon de 100 actes d'achats choisi aléatoirement. Les variables principale sont :

Foyer : L'identificateur du panéliste.
ReAchats : Le nombre des actes d'achats (passage en caisse) répétés pendant la période des 26 premières semaines.
N26etplusAchats : Le nombre des actes d'achat effectués pendant la deuxième période, donnée par 26 semaines.
Cohorte : Prend 1 si l'acte d'achat est fait pendant les premières 26 semaines (période d'estimation) et prend 2 pour les 26 semaines restantes (période de validation).
HHsize : La taille du ménage.
CSPchef : La catégorie socio-professionnelle du chef de ménage.
Nrevenu : Le revenu net du chef de ménage.
Quartierhab : Le quartier habité par le panéliste.
Dureeobservation : La durée d'observation.

Nous effectuons en premier lieu, une étude descriptive unidimensionnelle qui nous précise les caractéristiques principales de la distribution sujet d'étude, elle nous fournit des renseignements sur la forme de nos observations, et ce, numériquement au biais de la comparaison des paramètres de la distribution. Nous commençons par présenter le tableau (6.2) suivant qui résume quelques aspects des achats effectués par les panélistes.

Remarques : *A partir du Tab. 6.2, nous constatons que la moyenne de la variable " ReAchats " ainsi de " N26etplusAchats " (49.19 resp. 51.02) est très éloignée du maximum (Max=317 resp. 336) et relativement distante de la médiane (18.5 resp. 10). Ceci nous mène à conclure que notre distribution est assez dispersées autour de la moyenne (la variance est respectivement égal à 3727.29 et 4522.95). Le mode est inférieur à la médiane qui est elle même inférieure à la moyenne, ainsi, la distribution est légèrement étalée vers la droite.*

Nous vérifions que ces deux variables sont fortement corrélées (avec un coefficient 0.971) :

TAB. 6.2 – *Statistique descriptives fondamentales*

	ReAchats	N26etplusAchats
N	100	100
Minimun	0	0
Maximum	317	336
Moyenne	49.19	51.02
Médiane	18.5	10
Mode	0	0
Asymétrie	1.83	1.77
Variance	3727.29	4522.95

ceci paraît tout à fait logique. Un consommateur fréquent pendant la première période reste évidemment fidéle pendant la seconde période. A un moment, le comportement de consommation passé explique le comportement de consommation à venir.

6.2.1 Répartition des panélistes selon les variables caractéristiques

a. Clients et lieu d'habitation

Nous essayons, à partir de la figure (6.1), de voir la proportion des clients pour chacun des 13 quartiers. Nous donnons un deuxième graphique (6.2) qui traduit les achats effectués par les consommateurs selon l'emplacement de leurs lieux d'habitations par rapport à la position du magasin 1.

Commentaires : *Nous pouvons conclure, à partir de ces deux figures, que cette représentation reflète la fidélisation des clients du magasin 1 ; ceux qui le fréquentaient pendant les 26 premières semaines continuent de manière habituelle à faire leurs achats auprès de ce super marché durant la deuxième période de 26 semaines (nous voyons qu'il y a une petite hausse du panier moyen). Ces clients sont majoritairement les habitants du quartier 6, 8 et 12. Ceci est tout à fait cohérent puisqu'il résident dans les quartiers qui environnent le magasin 1. Par contre, ceux qui résident aux quartiers 1, 5, 11 et 13 baissent leurs actes achats pendant la deuxième période.*

b. Clients et catégories socio-profesionnelles

Concernant la catégorie socio-professionnelle des pénalistes, nous avons utilisé la no-

FIG. 6.1 – *Dispersion des clients selon le lieu d'habitation*

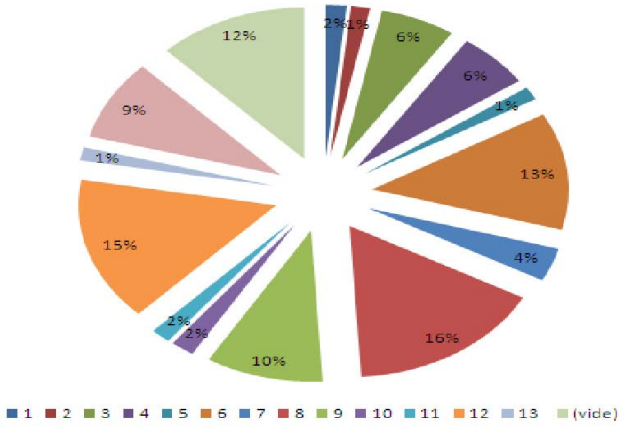
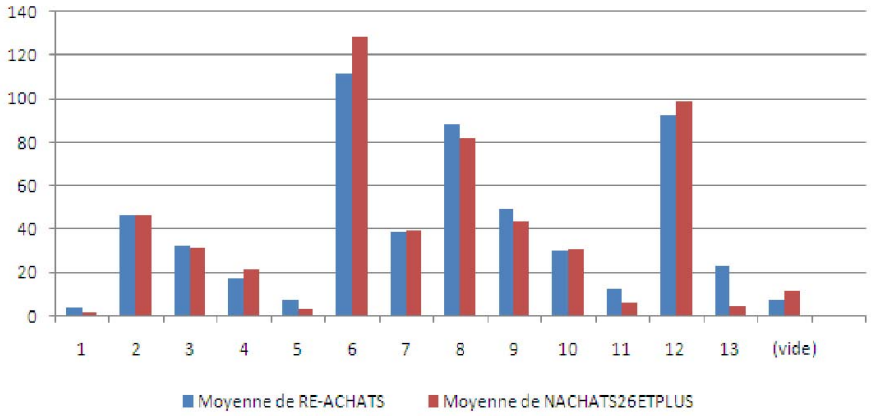


FIG. 6.2 – *Localisation des panélistes du magasin 1*

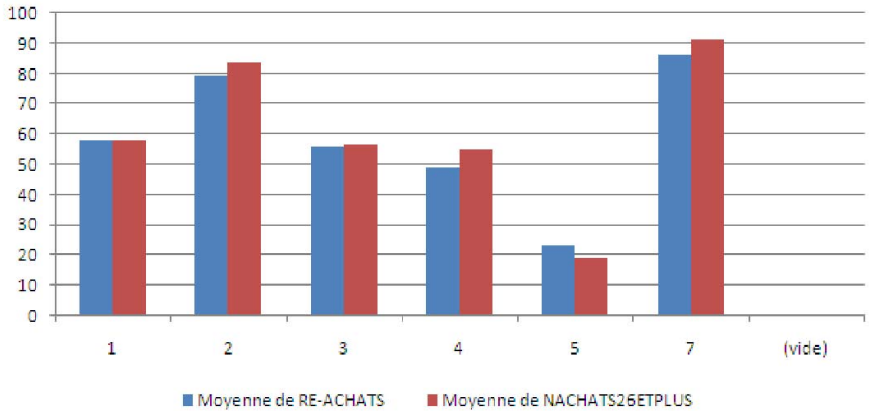


menclature suivante pour coder la variable "CSPchef" :
1 = un agriculteur ; 2 = un cadre ; 3 = un technicien ; 4 = un employé ; 5 = un ouvrier ; 7 = un chômeur.

Commentaires : *Nous présentons dans un troisième graphique la répartition d'actes d'achats selon la catégorie socio-professionnelle (voir 6.3). L'effectif le plus grand est accordé à la classe des non travailleurs et des cadres. Donc, ce qui est vraiment intéressant à comprendre est que le nombre de passage en caisse le plus élevé est réalisé par ceux qui ne travaillent pas ; donc le plus grand nombre des fidèles sont des non-travailleurs. Ceci nous amène à poser différentes questions : comment s'explique ce phénomène ? Sont-ils des chômeurs qui profitent des aides sociales et dépensent leurs revenus dans des produits alimentaires ? La proportion augmente pendant la seconde période, ceci se justifie par quoi ? Nous pouvons penser à une explication raisonnable : il s'agit d'une question de disponibilité (facteur temps). Les autres valeurs sont pratiquement proches.*

Les clients les moins fréquents sont les ouvriers.

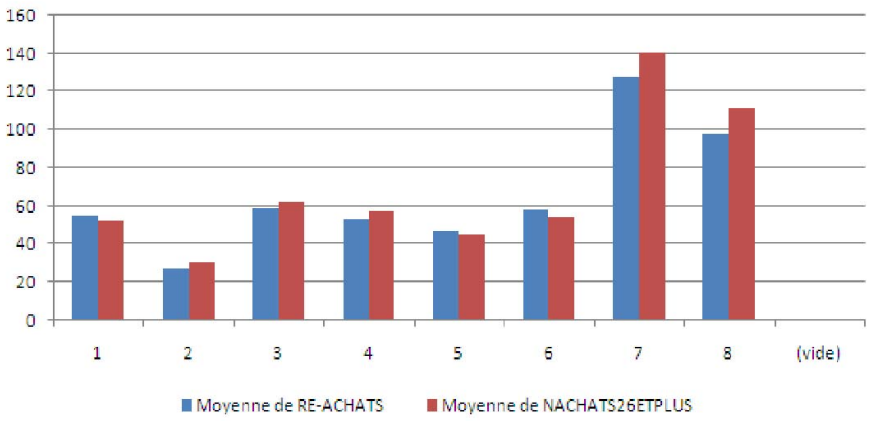
FIG. 6.3 – *Catégorie socio-professionnelle des panélistes et actes d'achats*



c. Clients et revenu net

Nous avons utilisé la nomenclature suivante pour coder la variable "AGEchef" : 1 = moins de 6500 FR ; 2 = 6500 FR - 8500 FR ; 3 = 8500 FR - 12000 FR ; 4 = 12000 FR - 15000 FR ; 5 = 15000 FR - 18000 FR ; 6 = 18000 FR - 22000 FR ; 7 = 22000 FR - 25000 FR ; 8 = 25000 FR - 30000 FR ; 9 = plus que 30000 FR ; 10 = pas de réponse. L'étude de l'évolution du nombre d'articles achetés en fonction du revenu des consommateurs est un élément déterminant et fondamental, ça nous reflète la qualité des clients qui font leurs achats dans le magasin 1, ensuite les dépenses qui mettent par rapport à ce revenu. Dans la figure 6.4, nous voyons la répartition de notre échantillon.

FIG. 6.4 – *Revenu net des panélistes du magasin 1*



Commentaire : *La consommation continue à croitre légèrement pour les catégories 2, 3, 4, 7 et 8 des panélistes et dimiue pour les autres.*

d. Clients et taille du foyer

Nous avons utilisé la nomenclature suivante pour coder la variable "HHsize" :
1 = une personne (femme) ; 2 = 2 personnes ; 3 = 3 personnes ; 4 = 4 personnes ; 5 = 5 personnes ; 6 = 6 personnes ; 7 = 7 personnes ; 8 = 8 personnes ; 9 = 9 personnes ;
La taille de la famille contribue à son tour dans l'accroissement du nombre d'achats. Le graphique 6.6 met en évidence cette nomenclature.

FIG. 6.5 – *Répartition des clients du magasin 1 selon la taille du foyer*

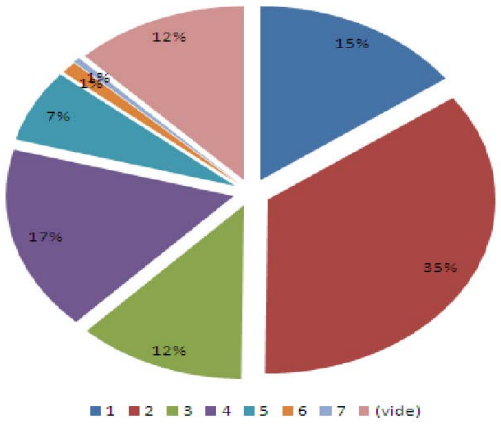
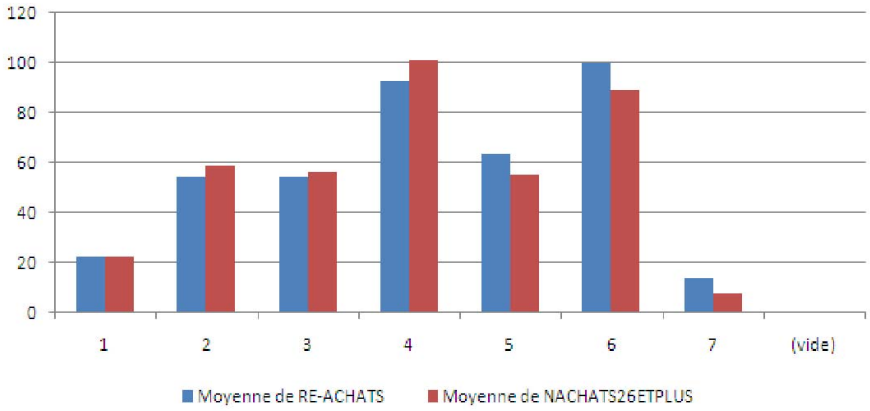


FIG. 6.6 – *Taille de famille des panélistes du magasin 1*



Commentaires : *Nous constatons que l'effectif des achats effectués est considérablement plus haut pour les familles à quatre et six personnes et moins peu pour les ménages de deux, trois et cinq personnes (voir 6.5).*

g. Conclusion

En conclusion, le comportement de consommation dépend de deux grands élements :
-En premier lieu, l'historique des actes d'achats effectués (s'il s'agit d'un grand ou un petit nombre) et son influence sur la consommation à venir.
-En second lieu, la fréquence de consommation dépend d'un certain nombre de caractéristiques personnelles.

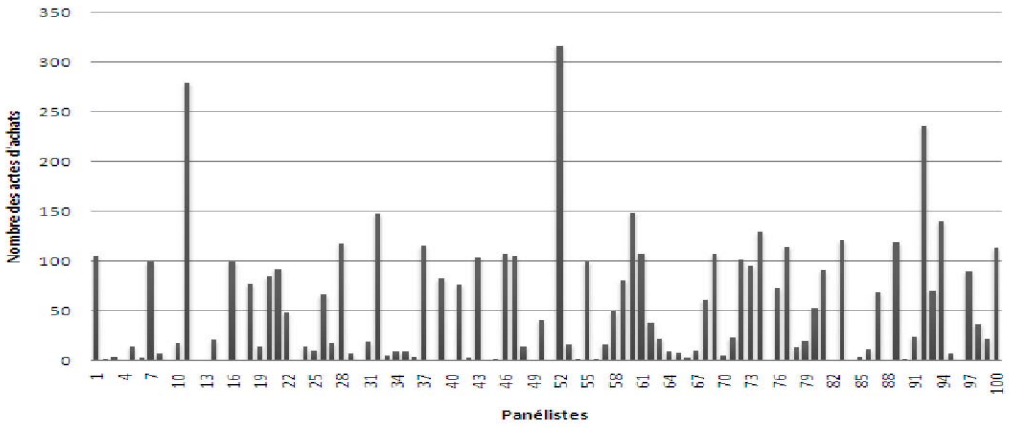
6.3 Application

Dans cette partie, nous approprions l'estimateur à noyau associé pour l'estimation des actes d'achats effectués par l'ensemble des panélistes. En effet, comme nous avons des données de dénombrement, le cas du noyau associé continu est automatiquement éliminé. Nous optons plutôt pour le cas discret. Or, nous avons des variables surdispersés (la variance est beaucoup plus supérieure que la moyenne). Nous utilisons alors l'approche non-paramétrique pour estimer les actes d'achats en faisant appel aux estimateurs à noyaux associés triangulaire et binomial.

6.3.1 Dans le cas d'un estimateur à noyau associé triangulaire

Nous donnons d'abord les graphiques 6.7 et 6.8 qui illustrent le comportement de consommation des foyers pour la cohorte 1 et 2. En abscisses nous avons les panélistes, et en ordonnées nous observons l'effectif de leurs actes d'achats. Nous remarquons que cet effectif est différent d'un panéliste à un autre. Il n'y a pas un comportement de consommation homogène. Nous donnons ainsi une estimation des actes d'achats pour la première et la deuxième période (chaque période est donnée par 26 semaines). Les figures 6.9 et 6.10 mettent en évidence ce comportement.

FIG. 6.7 – *Comportement des achats individuels pendant la première tranche*



Commentaire : *D'après 6.9 et 6.10, nous remarquons que nous avons plusieurs observations en zéro ; c'est ce que nous appelons "excès de zéro". L'ajustement par le noyau associé triangulaire (marqué en rouge dans chaque graphique) réagit mieux avec*

FIG. 6.8 – *Comportement des achats pendant la deuxième tranche*

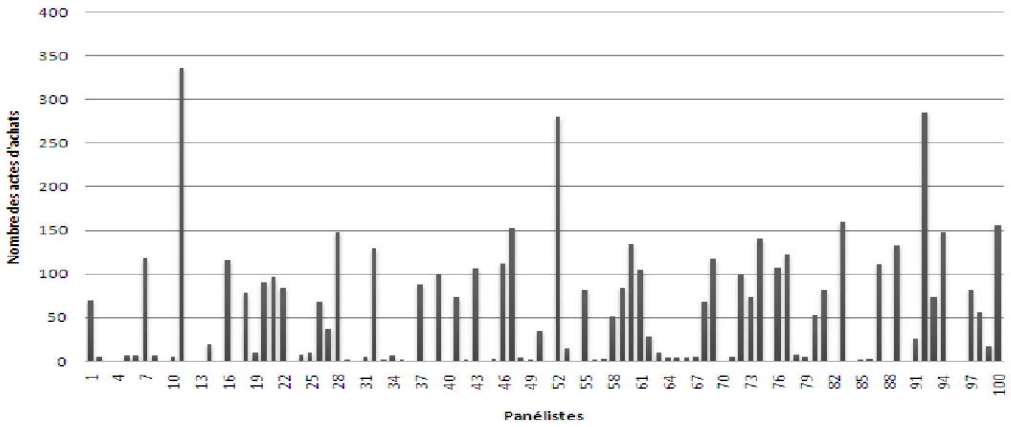
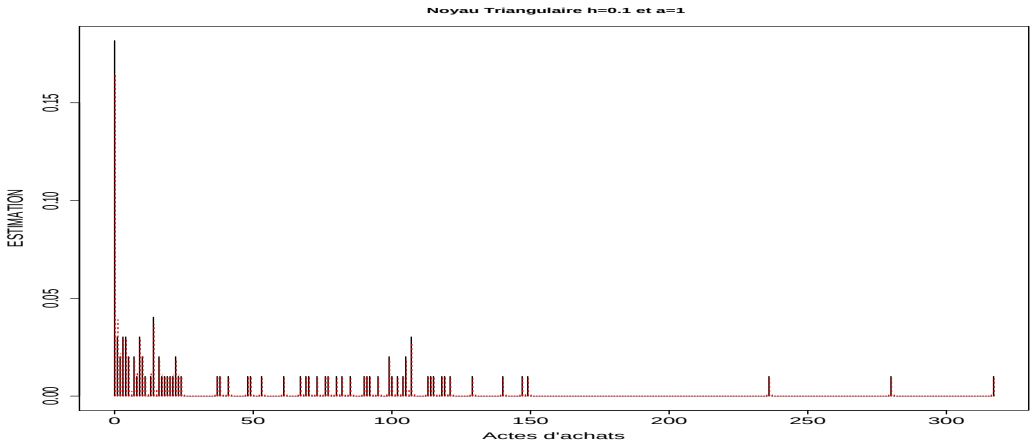


FIG. 6.9 – *Estimation des actes d'achats pour la première période*



cette distribution. Nous constatons que ce noyau associé tient compte de chaque observation et détecte les plateaux et la dispersion des données. Il estime correctement la distribution empirique. En plus, comme les observations sont assez dispérsées à droite, nous pouvons conclure que ces données de panel sont des données parsemées (en anglais "Sparse Data") et pour les estimer rigoureusement, il est préfèrable de prendre les modèles non-paramétriques. Dans les figures 6.11 et 6.12 nous avons agrandi l'échelle pour mieux voir l'ajustement.

6.3.2 Dans le cas d'un estimateur à noyau associé binomial

Nous estimons à ce niveau les mêmes actes d'achats parsemés en utilisant un estimateur à noyau associé binomial. Pour cela, nous donnons d'abord les graphiques 6.13 et 6.13.

FIG. 6.10 – *Estimation des actes d'achats pour la deuxième période*

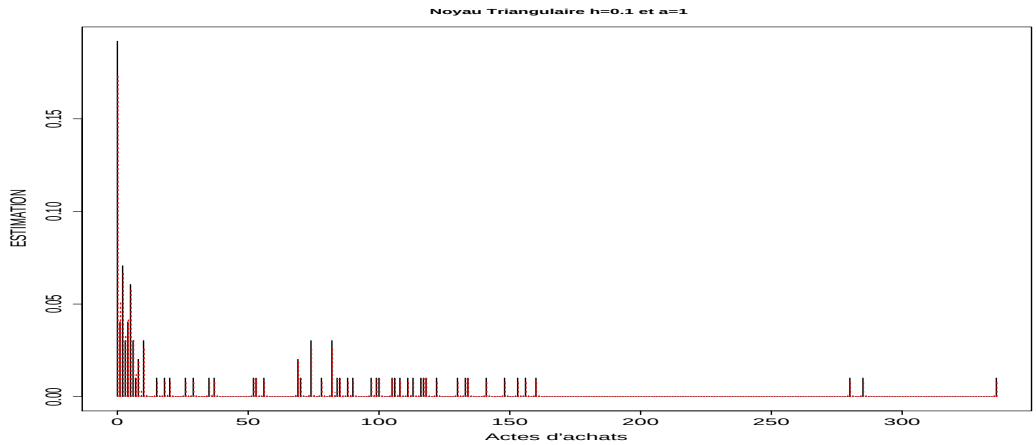
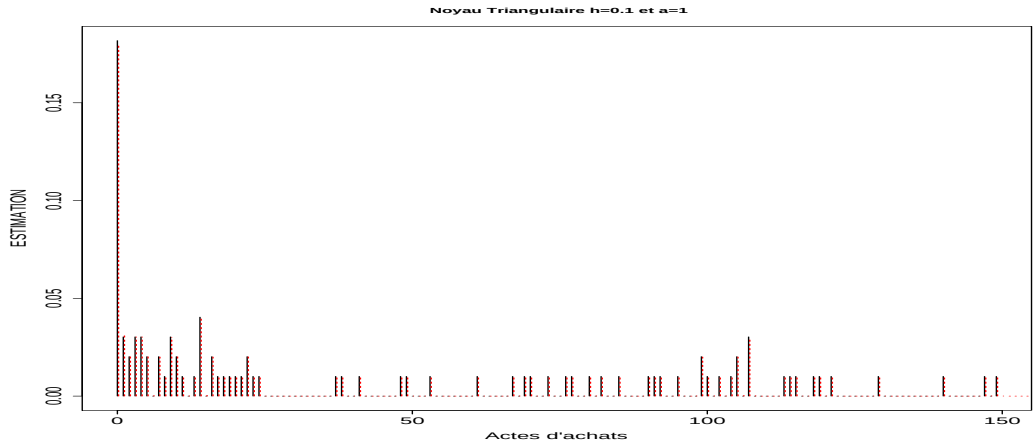


FIG. 6.11 – *Estimation de la première période agrandie*



Commentaire : *Concernant ce type de noyau associé, nous remarquons qu'il y a des masses qui apparaissent à gauche et à droite des observations étalées. Ce noyau associé binomial s'attarde à prendre en compte fortement les différentes observations. Même si ce noyau associé binomial est reconnu pour sa flexibilité, il ne réussit pas à estimer correctement des données parsemées.*

FIG. 6.12 – *Estimation de la première période plus agrandie*

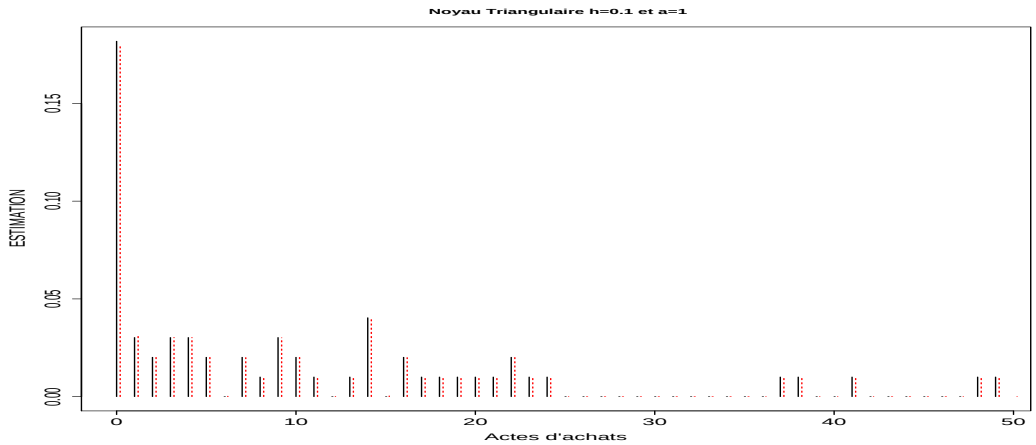


FIG. 6.13 – *Estimation des actes d'achats pour la première période*

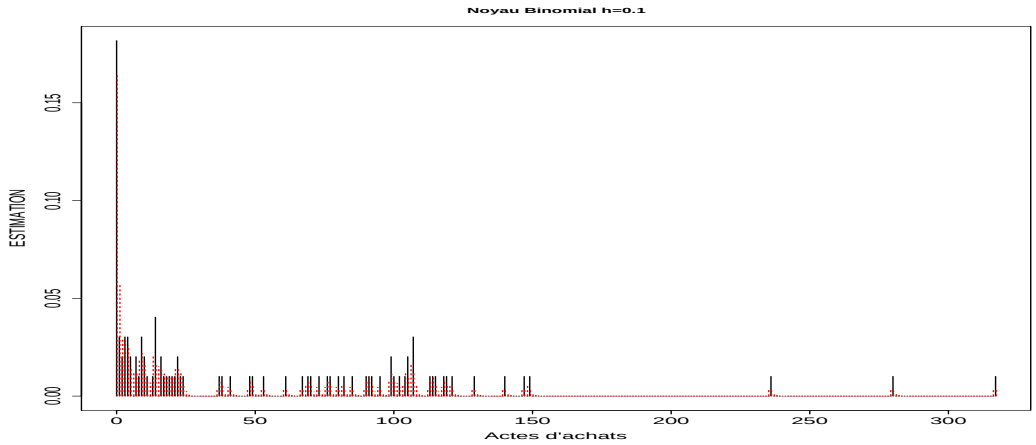


FIG. 6.14 – *Estimation des actes d'achats pour la deuxième période*

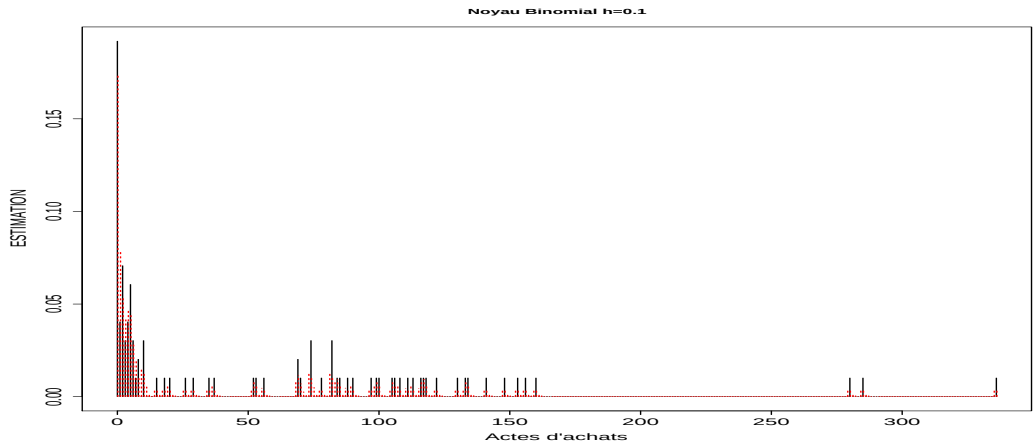


FIG. 6.15 – *Estimation des actes d'achats de la première période agrandie (150 observations)*

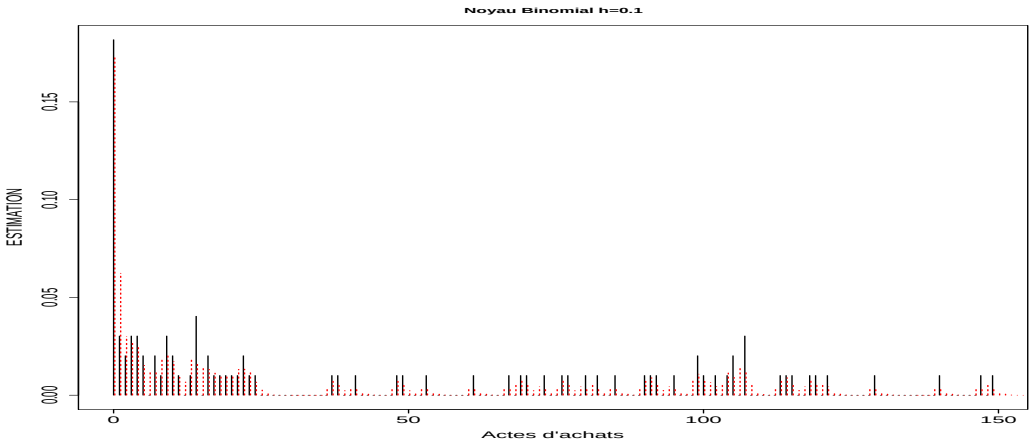
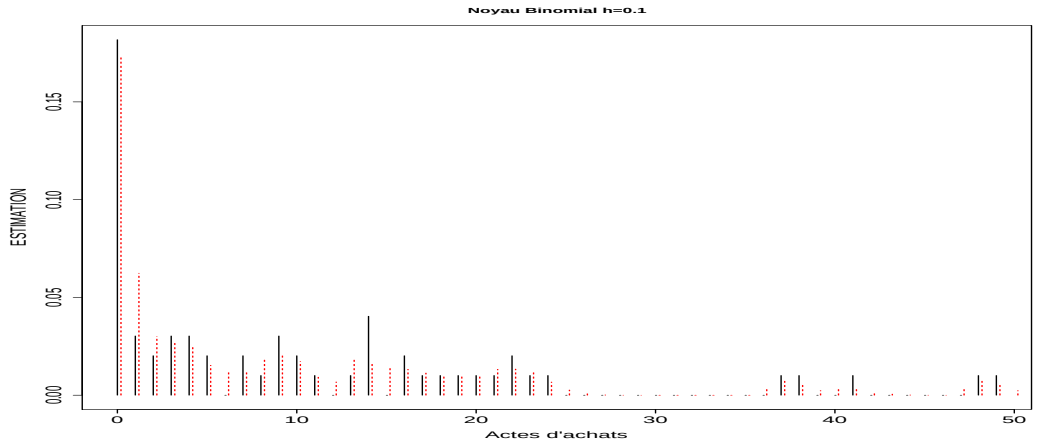


FIG. 6.16 – *Estimation des actes d'achats de la première période plus agrandie (50 observations)*



Chapitre 7

Conclusions et perspectives

7.1 Conclusions

Ce rapport a permis de couvrir une étendue assez large du domaine de l'estimation non-paramétrique d'une densité (fonction de masse) de probabilité inconnue f basée sur la technique des noyaux associés. Nous avons vulgarisé les travaux des pionniers de ce domaine, et aussi unifié la définition d'un noyau associé dans chacun des cas continu et discret. Nous avons pu ainsi donner l'estimateur et calculer ses propriétés. Les illustrations faites simplifient la compréhension de cette méthode. L'application de cette approche sur les données parsemées met en évidence que le noyau associé triangulaire est performant. Enfin, la méthode d'estimation non-paramétrique par noyaux associés permet d'avoir de bons résultats si nous choisissons adéquatement les paramètres mis en jeu.

7.2 Perspectives

Les travaux présentés dans ce document offrent de nombreuses perspectives.

Sur un plan théorique, nous aurions aimé nous attarder sur l'estimateur à noyau associé multiple et nous intéresser à ses propriétés fondamentales. Nous aurions aussi aimé appliquer ces noyaux associés sur des données de panel dans un cadre de régression. C'est à dire que sur ces données parsemées, nous attribuons un mélange de noyaux discrets et continus afin d'améliorer la qualité d'estimation.

Il sera également intéressant de penser à une combinaison entre les noyaux associés continus et les noyaux associés discrets. Quelques travaux dans cette direction vont d'ailleurs être entrepris.

Chapitre 8

Annexe 1 : commandes sous le logiciel R

Progamme des simulations de l'estimateur à noyau continu symétrique

Nous avons utilisé le code suivant pour la méthode de Plug-in :

density.default(x=x,bw="nrd0",kernel="epanechnikov",n=100),

où la commande " bw= 'nrd0' " permet de choisir la fenêtre de lissage.
Nous avons utilisé le code suivant pour la méthode de Validation croisée :

density.default(x=x,bw="ucv",kernel="epanechnikov",n=100),

où la commande " bw= 'ucv' " permet de choisir la fenêtre de lissage.

Nous avons crée nos propres codes pour présenter les graphiques des différents types de noyaux .
En particulier, nous avons eu recours aux fonctions "dgamma" et "dbeta" qui existent déjà sous R. Pour La loi inverse gaussienne (IG) et réciproque inverse gaussienne (RIG), nous les avons programmmé puisque le code n'existe pas.

```
dinvgauss <- function(x, mu = stop("no shape arg"), lambda = 1)
{
if(any(mu<=0)) stop("mu must be positive")
if(any(lambda<=0)) stop("lambda must be positive")
d <- ifelse(x>0,sqrt(lambda/(2*pi*x^3))*exp(-lambda*(x - mu)^2/(2*mu^2*x)),0)
if(!is.null(Names <- names(x)))
names(d) <- rep(Names, length = length(d))
d
}
pinvgauss <- function(q, mu = stop("no shape arg"), lambda = 1)
{
if(any(mu<=0)) stop("mu must be positive")
if(any(lambda<=0)) stop("lambda must be positive")
n <- length(q)
if(length(mu)>1 && length(mu)!=n) mu <- rep(mu,length=n)
if(length(lambda)>1 && length(lambda)!=n) lambda <- rep(lambda,length=n)
lq <- sqrt(lambda/q)
```

```
qm <- q/mu
p <- ifelse(q>0,pnorm(lq*(qm-1))+exp(2*lambda/mu)*pnorm(-lq*(qm+1)),0)
if(!is.null(Names <- names(q)))
names(p) <- rep(Names, length = length(p))
p
}
rinvgauss <- function(n, mu = stop("no shape arg"), lambda = 1)
{
if(any(mu<=0)) stop("mu must be positive")
if(any(lambda<=0)) stop("lambda must be positive")
if(length(n)>1) n <- length(n)
if(length(mu)>1 && length(mu)!=n) mu <- rep(mu,length=n)
if(length(lambda)>1 && length(lambda)!=n) lambda <- rep(lambda,length=n)
y2 <- rchisq(n,1)
u <- runif(n)
r1 <- mu/(2*lambda) * (2*lambda + mu*y2 - sqrt(4*lambda*mu*y2 + mu^2*y2^2))
r2 <- mu^2/r1
ifelse(u < mu/(mu+r1), r1, r2)
}
```

Nous avons crée nos propres codes pour appliquer les estimateurs aux données de panel.

Progamme de l'estimateur à noyau associé discret triangulaire

Description : Lissage d'une distribution de probabilité discrète par un estimateur à noyau associé discret triangulaire.

Arguments:
x : vecteur des points
h : paramètre de lissage
a : bras (paramètre)
V : vecteur des observations de l'échantillon
N : effectifs des observations
n=sum(N) : nombre total d'observations = taille de l'échantillon

Usage :

```
trng=function(x,h,V,N,n,a)
trng=edit(trng,editor="nedit")
Y=trng(x,h,V,N,n,a)
```

Détails : La loi de probabilité discrète triangulaire d'ordre h, de bras a et de centre x se définit par

$$\Pr(z) = \big((a+1)^h - \big(|z-x|\big)^h\big)/A,$$

avec $z = x \pm 1, x \pm 2, \ldots, x \pm a$, et où $A = (2a+1)(a+1)^h - 2\sum(k^h)$, $k = 1, 2, \ldots$, a est la constante de normalisation.

Code de l'estimateur à noyau associé discret triangulaire :

```
function(x,a,V,N,n,h)
{
y=0
s=rep(0,length(x))
n=sum(N)          # Taille de l'échantillon
f0=c(N/n,rep(0,length(x)-length(N)))          # Estimateur fréquence
u=0;
m=0;
for (k in 1:a)
{ m=k^h
u=u+m
}
A=(2*a+1)*(a+1)^h-2*u          # Constante de normalisation P(a,h)
for (i in 1:length(x))
{for (j in 1:length(N))
{if (V[j]≥(x[i]-a) & V[j]≤(x[i]+a))          # Support {x ± 1,...,x ± a}
{K=((a+1)^h - (abs(V[j]-x[i]))^h)/A          # Noyau associé
          y=(N[j]/n)*K          # Estimation à noyau associé discret triangulaire
}
else{
y=0
}
s[i]=s[i]+y
}
}
fn=s/sum(s)          # Estimations f̂n
E=sum(s)          # Constante de normalisation C
E[2]=sum((f0-fn) ^2)          # ISE0
return(E)
}
```

Progamme de l'estimateur à noyau associé discret binomial

Description: Lissage d'une distribution de probabilité discrète par un estimateur à noyau associé discret binomial.
Arguments :
x : vecteur des points
h : paramètre de lissage
V : vecteur des observations de l'échantillon
N : effectifs des observations
n=sum(N) : nombre total d'observations = taille de l'échantillon

Usage :
binom=function(x,h,V,N,n)
binom=edit(binom,editor="nedit")
Yb=binom(x,h,V,N,n)
Détails : La loi de probabilité binomiale de paramètres p et n se définit par

Pr(z)= choose(n,z)*(p)^z*(1-p)^(n-z),

z = 0,1,..., n. Le noyau associé discret se construit avec p=(x+h)/(x+1) et n=x+1.
Code de l'estimateur à noyau associé discret binomial :

```
function(x,V,N,n,h)
{ y=0    s=rep(0,length(x))
n=sum(N)          # Taille de l'échantillon
f0=c(N/n,rep(0,length(x)-length(N)))          # Estimateur fréquence
for (i in 1:length(x))
{for (j in 1:length(N))
{if(V[j]<=x[i]+1)          # Support {0,1,...,x + 1}
{ K= choose(x[i]+1,V[j])*((x[i]+h)/(x[i]+1))^(V[j])
*((1-h)/(x[i]+1))^(x[i]+1-V[j])) # noyau associé
y=(N[j]/n)*K          # Estimation à noyau associé discret binomial
          }
s[i]=s[i]+y
}       }
fn=s/sum(s)          # Estimations f̂_n
E=sum(s)          # Constante de normalisation C
E[2]=sum((f0-fn)^2)          # ISE0
return(E)
}
```

Bibliographie

[1] Aitchison, J. & Aitken, C.G.G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika* **63**, 413-420.

[2] Chaubey, Y.P., Sen A. & Sen P.K. (2007). A New Smooth Density Estimator For Non-Negative Random Variables. *Technical Report No 01/07.* Concordia University. Montréal.

[3] Chen, S.X. (1999). Beta Kernels estimators for density functions. *Computational Statistics and Data Analysis* **31**, 131-145.

[4] Chen, S.X. (2000). Gamma Kernels estimators for density functions. *Annals of the Institute of Statistical Mathematics* **52**, 471-480.

[5] Duong, T. (2004). *Bandwidth selectors for multivariate kernel density estimation*, thesis for the degree of Doctor of philosophy at the University of Western Australia. School of Mathematics and Statistics.

[6] Feller, W. (1966). *An Introduction to Probability and Its Applications.* John Wiley and Sons, New York.

[7] Hall, P. (1981). On nonparametric multivariate binary discrimination. *Biometrika* **68**, 287-294.

[8] Hall, P., Racine, J.S. & Li, Q. (2004). Cross validation and the estimation of conditional probability densities. *Journal of the American Statistical Association* **99**, 1015-1026.

[9] Hille, E. (1948). *Functional Analysis and Semigroups.* American Mathematical Society Colloquium, New York.

[10] Senga Kiessé, T. (2008). *Approche non-paramétrique des données de dénombrement*, thèse en préparation pour obtenir le grade d'un Docteur d'Université de Pau et des Pays de l'Adour.

[11] Kokonendji, C.C., Senga Kiessé, T. & Zocchi, S.S. (2007). Discrete triangular distributions and non-parametric estimation for probability mass function. *Journal of Nonparametric Statistics* **19**, 241-254.

[12] Li, Q. & Racine, J.S. (2007). *Nonparametric Econometrics: Theory and Practice.* Princeton University Press, New York.

[13] Michels, P. (1992). Assymetric Kernels Functions in Non-Parametric Regression Ananlysis and Prediction. *The Statistician* **41**, 439-454.

[14] Scaillet, O. (2004). Density estimation using inverse and reciprocal inverse Gaussian kernels. *Journal of Nonparametric Statistics* **16**, 217-226.

[15] Seshadri, V. (1993). *The Inverse Gaussian Distribution: A Case Study in Exponential Families.* Oxford University Press, New York.

[16] Simonoff, J.S. (1996). *Smoothing Methods in Statistics.* Springer, New York.

[17] Tsybakov, A.B. (2004). *Introduction à l'Estimation Non Paramétrique.* Springer, Paris.

[18] Wand, M.P. & Jones, M.C. (1995). *Kernel Smoothing.* Chapman and Hall, London.