RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
UNIVERSITÉ A/MIRA DE BÉJAÏA
Faculté des Sciences Exactes
Département de Recherche Øpérationnelle

MÉMOIRE DE MAGISTER

En

Mathématiques Appliquées

Option

Modélisation Mathématique et Techniques de Décision

Thème

Bootstrap dans l'estimation non paramétrique de la densité de probabilité et la courbe de régression de la moyenne
Bootstrap in the nonparametric estimation of a probability density and the curve regression of the mean.

Présenté par :
Mr CHERFAOUI Mouloud

Soutenu devant le jury composé de :

Président	Djamil AISSANI	Professeur	Université de Béjaïa
Rapporteur	Smaïl ADJABI	Maître de Conférences	Université de Béjaïa
Examinateur	Mohand Ouamer BIBI	Professeur	Université de Béjaïa
Examinateur	Zaher MOHDEB	Professeur	Université de Constantine
Invitée	Karima CHERFI-LAGHA	Docteur	Université de Béjaïa

Université de Béjaïa, 2009

Remerciements

Je tiens à remercier en premier lieu Monsieur Smaïl ADJABI, Maître de Conférences à l'Université de Béjaïa pour m'avoir proposé ce travail, pour l'avoir dirigé, pour la confiance qu'il m'a accordée et pour l'aide qu'il m'a apportée pendant toute la durée de ce mémoire.

J'adresse aussi mes sincères remerciements à Monsieur Djamil AISSANI, Professeur à l'Université de Béjaïa qui a accepté de présider le jury de ce mémoire.

Je remercie profondèment Monsieur Mohand Ouamer BIBI, Professeur à l'Université de Béjaïa, Monsieur Zaher MOHDEB, professeur à l'université de Constantine, d'avoir accepté d'examiner ce travail ainsi que Madame Karima CHERFI-LAGHA, pour avoir accepté l'invitation.

Qu'il me soit permis enfin d'avoir une pensée bien spéciale pour ma famille et mes amis (es). Je suis très reconnaissant envers eux pour leurs encouragements et sacrifices dans tout ce que j'ai décidé d'entreprendre.

Dédicace

Je rends ce qui est à César à César,
et
ce qui est à Dieu à Dieu.

Table des matières

Introduction générale

Parmi les problèmes d'actualité dans le domaine des statistiques, on trouve le problème de la statistique fonctionnelle. L'essor de ce domaine de la statistique tient aussi bien de l'augmentation des problèmes concrets pour lesquels les jeux de données sont à caractère fonctionnel qu'à l'accumulation des connaissances théoriques acquises depuis quelques années dans plusieurs domaines fonctionnels de la statistique. Parmi les problèmes d'estimation fonctionnelle on cite : l'estimation de la densité de probabilité et l'estimation de la courbe de régression de la moyenne.

L'estimation de la densité de probabilité est l'un des plus vieux problèmes de l'estimation non paramétrique. En effet, l'une des premières analyses consacrées au sujet de l'estimation de la densité de probabilité est due au biométricien Karl Pearson, il y a une centaine d'années. Dans son article de 1902 [78], une première approche (dite paramétrique) consiste à supposer que la densité de probabilité f appartient à une famille de densités qui peuvent être décrites par un petit nombre (connu) de paramètres réels. Le statisticien qui opte pour une telle approche possède une bonne connaissance a priori du phénomène aléatoire. Il sait, par intuition ou par expérience, que la variable aléatoire X suit une loi f, tout en ignorant la valeur de son espérance ou de sa variance. Dans un tel contexte, l'estimation de la densité se réduit alors à un problème d'estimation de paramètres.

Ainsi, à moins d'avoir sur le phénomène aléatoire étudié des informations à priori très précises et indiscutables, le champ d'application d'un modèle paramétrique n'est satisfaisant que lorsque l'inflation du nombre de paramètres est telle que les méthodes d'estimation du modèle deviennent tout à fait inefficaces.

Pour pallier les insuffisances et les défauts des familles paramétriques, une seconde approche dite non paramétrique propose de "laisser parler les données", sans spécifier au préalable de forme sur f. Actuellement, il existe plusieurs méthodes non paramétriques pour l'estimation la densité de probabilité, on peut citer : la méthode de l'histogramme, méthode d'estimation par les séries orthogonales et la méthode du noyau. Cette dernière méthode qui fait l'objet de notre travail sera présentée en détail. Nous avons opté pour cet estimateur vu sa souplesse d'utilisation et ces bonnes propriétés de convergence.

C'est Rosenblatt [80] en 1956, suivi de Parzen [77] en 1962, qui ont proposé une classe d'estimateurs à noyau d'une densité univariée. Les estimateurs à noyau sont fonction de deux paramètres K, appelé noyau, et h dit paramètre de lissage (largeur de fenêtre). Rosenblatt reprenait l'idée de Fix et Hodges [40] en 1951, qui consistait à estimer la densité en un point, en comptant le nombre d'observations situées dans l'intervalle de longueur $2h$ et centré en ce point.

Les propriétés de convergence de l'estimateur à noyau ont été établies par Parzen [77], Silverman [90] et Nadaraya [72]. Devroye [25] en 1985 a fait une étude complète sur la convergence L_1. Les théorèmes relatifs à l'erreur quadratique asymptotique et l'erreur quadratique intégrée asymptotique ont été obtenus sous forme élémentaire par Parzen [77]. Enfin, c'est Epanechnikov

en 1969 [33] qui s'est rendu compte de l'existence d'un noyau asymptotiquement optimal K_E. Mais l'erreur quadratique moyenne asymptotiquement intégrée varie peu en fonction du choix de K.

Si le choix du noyau n'est pas un problème dans l'estimation de la densité, il n'en est pas de même pour le choix de la largeur de fenêtre qui ne dépend que de la taille n de l'échantillon. Plusieurs travaux ont montré que l'estimateur peut changer dramatiquement pour de petites variations du paramètre de lissage. Actuellement, il n'existe pas de choix optimal pour ce paramètre de lissage. Le choix optimal qui minimise l'erreur relative globale ($MISE$) dépend de la dérivée seconde de la densité inconnue. Les auteurs se sont alors attachés à introduire des procédures de sélection automatiques et donc moins subjectives que le simple choix à l'œil.

On peut regrouper ces procédures en deux familles : famille des méthodes plug-in (re-injection) et famille des méthodes cross-validation (validation croisée). Mais aucune méthode n'est meilleure que les autres. Ces méthodes de sélection du paramètre de lissage ne fournissent pas une estimation graphiquement satisfaisante, laissant subsister des variations locales. L'estimateur a tendance à dévier systématiquement de la vraie valeur de la densité au voisinage de certains points critiques. Des travaux [109], ont montré que le choix de la méthode de sélection dépend de la forme de la densité que l'on cherche à estimer.

L'un des objectifs de ce travail est de contribuer au choix du paramètre de lissage en appliquant une ancienne technique appelé Boostsrap qui est un ensemble de méthodes qui consistent à faire de l'estimation sur de nouveaux échantillons tirés à partir d'un échantillon initial. Cette technique de Boostrap, s'est rapidement développée ces dernières années, grâce au développement de l'outil informatique.

Souvent, on se trouve face à un problème de recherche d'explication d'une variable Y (dite variable expliquée) en fonction d'une ou de plusieurs autres variables X (dite variables explicatives). Ce type de relation s'appelle courbe de régression de la moyenne. Elle permet de développer des outils théoriques performants et elle offre un énorme potentiel en termes d'applications (imagerie, agro-alimentaire, reconnaissance de forme, géographique, économétrie. . . etc.).

Plusieurs méthodes paramétriques ont été élaborées pour l'estimation des paramètres d'un modèle proposé. On peut citer : la méthode des moindres carrées, régression linéaire, etc. Mais le choix du modèle de régression reste un problème colossal, car on peut être face à des situations où la validation des modèles proposés se résume toujours au rejet du modèle .

Pour remédier à ce problème, les méthodes paramétriques cèdent leurs places aux méthodes non paramétriques telles que la méthode des splines, des polynômes locaux, des ondelettes et la méthode du noyau qui fait l'objet de notre travail.

Les estimateurs de type noyau de la courbe de régression de la moyenne, sont introduits indépendamment par Nadaraya 1964 [71] et Watson (1964) [104]. Comme l'estimateur à noyau de la densité de probabilité, l'estimateur à noyau de la courbe de régression de la moyenne dépend essentiellement de deux paramètres : le noyau K et le paramètre de lissage h.

Il existe une littérature abondante sur l'estimation non parmétrique de la courbe de régression de la moyenne, à laquelle Gérard Collomb apporta une contribution déterminante comme en témoigne la compilation de ses travaux (Collomb 1983 [20])(voir aussi Révész 1977 [82] pour d'autres résultats précurseurs en ce domaine). Sarda et Vieu [83] en 2000, ont étudié la convergence de cet estimateur en norme L_2 et L_∞.

La méthode utilisée pour le choix du paramètre de lissage est la méthode validation croisée, développée par : Marron (1988)[68], Jones et al. (1996) [56], Vieu (1993) [100] et Herrmann (2000) [52]. L'alternative à cette méthode est celle basée sur les techniques de rééchantillonnage Boostrap (Cao-Abad ,1991 [12]) [51]), que nous présenterons et appliquerons dans ce travail.

Concernant le choix du noyau, nous renvoyons à l'article initial de Gasser et Müller (1979)[41] qui permet de situer les problèmes, à celui de Berlinet (1993)[3] qui propose une méthode automatique de choix de noyau ainsi qu'à celui de Vieu (1999) [101] pour des résultats récents.

Pour les techniques de rééchantillonnage, signalons les travaux de Léger et Altman (1990)[63] et Mammen (1992)[66] pour une présentation de ces méthodes dans un contexte dépassant largement celui du simple choix de paramètre de lissage.

L'objectif de ce travail présenté dans ce mémoire se résume principalement dans :

- L'étude par simulation, de l'influence de la technique bootstrap sur l'estimation du paramètre de lissage dans l'estimation de la densité de probabilité par la méthode du noyau de Parzen-Rosenblatt ;
- L'étude par simulation de l'influence du choix du noyau sur les performances de l'estimateur à noyau de la densité de probabilité de Parzen-Rosenblatt ;
- L'étude par simulation de l'influence de la technique bootstrap, sur l'estimateur à noyau de la courbe de régression de la moyenne.

Pour répondre à ces objectifs, nous avons organisé notre travail comme suit :

- Le premier chapitre est consacré à la description de la technique de bootstrap, en particulier, nous nous intéresserons à l'estimation d'un paramètre (moyenne, médiane, variance), la bootstrap des résidus et enfin aux différentes méthodes de construction d'un intervalle de confiance par la technique bootstrap.

- Dans le second chapitre, nous nous intéresserons à l'estimation de la densité de probabilité par la méthode du noyau en donnant les propriétés statistiques de cet estimateur, les différentes méthodes de sélection du paramètre de lissage h, les différents noyaux usuels, ainsi que de nouveaux noyaux pour remédier au problème du biais aux bornes pour des densités définies sur des supports fermés ou semi-fermés. Enfin, dans ce second chapitre, nous présenterons les résultats des simulations conduites à partir de plusieurs densités cibles connues : loi normale, loi exponentielle et loi de khi-deux, afin :
  1. D'étudier l'influence du nombre de réplications de bootstrap sur les différentes méthodes de sélection et différentes densités cibles ;
  2. D'étudier les performances de la technique bootstrap ;
  3. D'étudier l'influence du choix du noyau sur les performances de l'estimateur à noyau de la densité de probabilité.
- Le chapitre trois est consacré à une présentation de l'estimation non paramétrique de la courbe de régression de la moyenne par la méthode du noyau, ainsi que quelques méthodes de sélection du paramètre de lissage. Enfin, dans ce chapitre, nous présenterons les résultats des simulations conduites pour :
  1. Étudier l'influence de la taille de l'échantillon sur les performances de l'estimateur de la courbe de régression de la moyenne.
  2. Étudier l' influence du nombre de réplications de bootstrap sur les performances de l'estimateur à noyau de la courbe de régression de la moyenne.
- Ce mémoire se termine par une conclusion générale et quelques perspectives de recherche sur l'application de la technique de bootstrap dans l'estimation de la densité de probabilité et l'estimation de la courbe de régression de la moyenne par la méthode du noyau.

Chapitre 1

La Technique de *bootstrap*

1.1 Introduction :

Le terme de rééchantillonnage, ou, en anglais, "*bootstrap*", qui évoque l'action de "se hisser en tirant sur ses propres lacets", désigne un ensemble de méthodes qui consistent à faire de l'inférence statistique sur de "nouveaux" échantillons tirés à partir d'un échantillon initial. Disposant d'un échantillon destiné à donner une certaine information sur une population, on tire au sort, parmi la sous-population réduite à cet échantillon, un nouvel échantillon de même taille n. On répète cette opération B fois, avec B grand. On analyse ensuite les nouvelles observations ainsi obtenues pour affiner l'inférence faite sur les observations initiales. A priori, on peut avoir des doutes sur l'efficacité d'une telle méthode et penser qu'il n'y a aucune amélioration à espérer en rééchantillonnant à partir du même échantillon. En effet, aucune information supplémentaire ne peut être espérée, toute l'information étant contenue dans l'échantillon initial. Cependant, comme on va le voir, ce rééchantillonnage, s'il ne rajoute aucune information, permet, dans certains cas, d'extraire de l'échantillon de base l'information souhaitée.

L'objectif de ce chapitre est de décrire l'application de la technique de *bootstrap* pour la résolution des problèmes d'inférence statistique en relation avec l'estimation des paramètres. Nous présentons d'abord les techniques de rééchantillonnage de *bootstrap*, ensuite, nous examinons l'estimation de l'erreur-standard et du biais d'un estimateur et enfin nous exposons quelques méthodes de détermination de l'intervalle de confiance d'un paramètre avant de conclure.

1.2 Exemple canonique (Evaluation de la précision d'une estimation)

Un exemple proposé par Efron [29] : A l'origine, le *bootstrap* a été employé pour évaluer la précision d'un estimateur. Par exemple, lors d'une petite expérimentation sur des souris, on a tiré au sort parmi 16 souris, 7 qui reçoivent le nouveau traitement alors que les 9 autres sont des contrôles qui reçoivent un placebo. Leurs durées de vie sont mesurées, en jours, et donnent les résultats suivants :

											Moyenne	écart-type
Traitées	X	94	197	16	38	99	141	23	$\times$	$\times$	86,86	25,24
Controlées	Y	52	104	146	10	51	30	40	27	46	56,22	14,14

La différence des moyennes est égale à : 30.63

Soit la statistique suivante qui représente l'erreur standard associée à la différence :

$$T = \frac{\bar{X} - \bar{Y}}{Se} \tag{1.1}$$

avec

- $\bar{X}$ la moyenne du premier échantillon.
- $\bar{Y}$ la moyenne du deuxième échantillon.
- Se_1 l'écart-type du premier échantillon.
- Se_2 l'écart-type du deuxième échantillon.
-

$$Se = \sqrt{Se_1 + Se_2} = \sqrt{14.14^2 + 25.24^2} = 28.93. \tag{1.2}$$

Si on compare directement les deux moyenne on aura l'impression que le traitement assure une meilleure survie que le placebo, car les durées moyennes observées sont respectivement : 86,86 et 56,22. Mais les deux échantillons sont petits et la précision de ces deux estimateurs est certainement très mauvaise.

Et même, si on calcule la statistique T donnée par l'équation (1.1) qui vaut 1.05, on ne peut pas juger si elle est significative ou non, car on ne connaît pas la loi des deux échantillons pour déterminer la valeur critique qui nous permet de prendre la décision sur le rejet de l'hypothèse d'égalité de ces deux moyennes.

Comment, donc mesurer cette précision ? Si l'on disposait d'une taille d'échantillon suffisamment grande pour pouvoir appliquer l'approximation normale (Théorème Centrale Limite), on utiliserait le fait que :

$$\mathcal{L}(\overline{X}|F) \approx \mathcal{N}(\mu, \frac{s^2}{n}), \tag{1.3}$$

avec

- F est la loi empirique de l'échantillon X .
- μ la moyenne empirique de l'échantillon X.
- s^2 la variance empirique de l'échantillon X.

Mais les tailles des deux échantillons sont trop faibles. De plus, si au lieu de comparer les moyennes, on veut comparer les médianes, qui sont ici respectivement 94 et 46, que faire pour estimer la précision et savoir à quel point elles sont effectivement différentes ?

Pour résoudre ce problème on fait appel à la technique de *bootstrap*.

1.3 Méthodes de rééchantillonnage de *bootstrap*

1.3.1 Définition

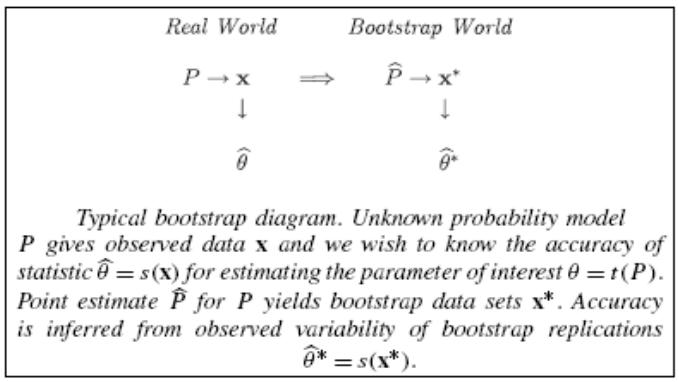
Le mot *bootstrap* provient de l'expression anglaise "to pull oneself up by one's *bootstrap*" (Efron, Tibshirani, 1993)[32], qui signifie littéralement "se soulever en tirant sur les languettes de ses bottes".

Le mot *bootstrap* fait penser à des traductions telles que "à la force du poignet" ou "par soi-même" ou "passe partout" (Dagnelie, 1998 [22]), mais en fait il n'est jamais traduit dans la littérature scientifique d'expression française.

1.3.2 Principe de *bootstrap*

La technique de *bootstrap* est conçue pour être utilisée dans le contexte du travail empirique ; comme le nom suggère, l'idée du principe original de la méthode est d'utiliser le seul

ensemble de données disponible pour approximer la distribution des aléas ou d'autres quantités du modèle (Voir le schéma (Fig.1.1) qui est donné par Bradly Effron pour expliquer le principe de *bootstrap* [30]), et cela en construisant toutes les combinaisons possibles (toutes les fonctions de répartitions empiriques possible) de ces données.



Typical bootstrap diagram. Unknown probability model P *gives observed data* $\mathbf{x}$ *and we wish to know the accuracy of statistic* $\widehat{\theta} = s(\mathbf{x})$ *for estimating the parameter of interest* $\theta = t(P)$. *Point estimate* $\widehat{P}$ *for* P *yields bootstrap data sets* $\mathbf{x}^*$. *Accuracy is inferred from observed variability of bootstrap replications* $\widehat{\theta}^* = s(\mathbf{x}^*)$.

FIG. 1.1: *Principe de Bootstrap*

A titre d'illustration prenons un exemple concret : supposons qu'on n'a aucun renseignement sur F la loi de la population étudiée et on veut estimer la statistique $\theta(F)$ qui est définie par $(EX)^2$ et sachant que nous ne disposons que de 3 observations : $x_1 = 1.1$; $x_2 = 1.2$ et $x_3 = 2.5$ qui sont issues de cette population ;

Donc $\theta(F)$ est définie dans ce cas comme suit :

$$\theta(F) = \left[\int x dF(x)\right]^2.$$

On est dans le cas non paramétrique. F_0 est donc la fonction de répartition empirique de cet échantillon qui donne à chacune des trois observations ci-dessus la probabilité $\frac{1}{3}$ alors :

$$\theta(F_0) = [(1.1 + 1.2 + 2.5)/3]^2 = 2.56.$$

Si on applique la *bootstrap* sur cet échantillon, dans ce cas la loi empirique F_b des échantillons de bootstrap peut prendre 10 valeurs différentes, en effet, quand on tire un échantillon de taille $n = 3$ de la loi F_0, tous ce passe comme si on tirait trois fois avec remise dans une urne à trois boules numérotés 1.1, 1.2 et 2.5 chaque tirage peut être résumé par le nombre de fois d'apparitions de la même boule (voir TAB.1.1) :

Fréquences			probabilité	θ
1.1	1.2	2.5		
3	0	0	$(1/3)^3$	1.210
2	1	0	$(1/3)^2$	1.284
2	0	1	$(1/3)^2$	2.454
1	1	1	2/9	2.560
1	2	0	$(1/3)^2$	1.361
1	0	2	$(1/3)^2$	4.134
0	2	1	$(1/3)^2$	2.667
0	1	2	$(1/3)^2$	4.271
0	3	0	$(1/3)^3$	1.440
0	0	3	$(1/3)^3$	6.250

TAB. 1.1: *Les 10 combinaisons possible de* F_b.

Donc l'estimation de *bootstrap* pour $(EX)^2$ est :

$$E\left(\theta(F_b)/F_0\right) = 1.210/27 + 1.284/9 + ... + 6.25/27 = 2.6955$$

On voit clairement sur cet exemple comment effectuer le calcul explicite de *bootstrap*. Mais on constate aussi que, même pour un cas aussi simple où la taille n de l'échantillon est égale à 3, il y'a un grand nombre de tirages possibles pour F_b (10 possibilités). Le tableau (TAB.1.2) résume l'évolution du nombre de tirages possibles pour F_b en fonction de n. On constat que ce nombre est exponentiellement proportionnelle avec n (voir Fig.1.2).

On constate ainsi que le principe est impossible à appliquer en pratique lorsque la taille de l'échantillon est grande. A cet effet, une autre version de la technique de bootstrap est née.

Tout en respectant le principe de rééchantillonnage aléatoire simple à partir de l'échantillon initial, mais au lieu d'énumérer toutes les combinaisons possibles en se contente d'un nombre fini B de rééchantillonages fixés préalablement par l'utilisateur du *bootstrap* ; autrement dit on se contente d'un sous ensemble de fonctions de répartitions empiriques issues de l'échantillon initial.

Valeur de n	Nombre de possibilité
2	3
3	10
4	35
5	126
10	92378
20	$> 6.8 \times 10^{10}$
25	$> 6 \times 10^{13}$
50	$> 5 \times 10^{28}$
100	$> 4 \times 10^{58}$
1000	∞

TAB. 1.2: *Evolution du nombre de possibilités de F_b en fonction de la taille de l'échantillon n.*

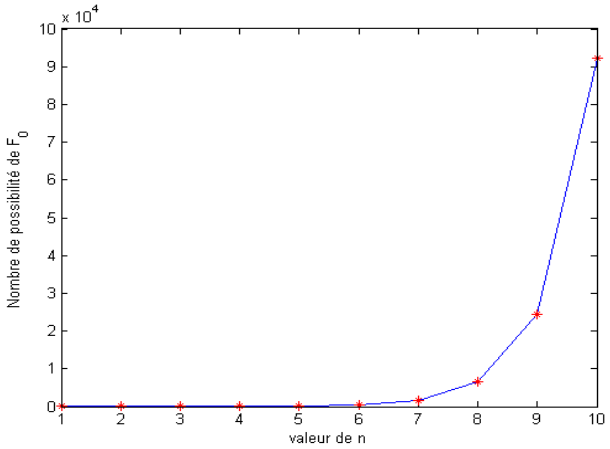


FIG. 1.2: *Nombre de cas possible de F_0 en fonction de la taille de l'échantillon n*

1.3.3 *Bootstrap* des individus

On considère un échantillon de n observations : x_1, x_2, . . ., x_i ,. . ., x_n, prélevé de manière aléatoire et simple dans une population. Ces observations peuvent concerner une seule variable, ou être relatives à plusieurs variables. Dans ce cas, les x_i représentent des vecteurs de dimension p, p étant le nombre de variables. Afin de ne pas alourdir les notations, nous ne distinguerons pas ces deux situations et, de manière plus générale, nous désignerons l'échantillon initial par x, qu'il s'agisse d'un vecteur ou d'une matrice.

Le principe de la méthode du *bootstrap* est de prélever une série d'échantillons aléatoires et simples avec remise de n observations dans l'échantillon initial, considéré comme une population. Ces échantillons successifs seront notés :

$$x_1^*, x_2^*, ..., x_b^*, ..., x_B^*,$$

B étant le nombre de rééchantillonnages effectués.

À titre d'illustration, nous présentons l'exemple exposé par R.Palm [75] dans lequel, on retrouve le problème de l'estimation de diverses caractéristiques de la population des tailles des exploitations agricoles de la région Wallonne, à partir d'un échantillon aléatoire et simple de 100 observations.

La population a été simulée à partir de la distribution groupée résultant du recensement agricole et horticole au 15 mai 1995 (INS, 1996). Pour une classe donnée, d'effectif n_i et de limites de classe $x_{i\,\mathrm{inf}}$ et $x_{i\,\mathrm{sup}}$, on a généré n_i nombres aléatoires appartenant à une distribution uniforme dans le domaine $(x_{i\,\mathrm{inf}}, x_{i\,\mathrm{sup}})$. On a ainsi obtenu les tailles simulées des 24 719 exploitations. La figure (Fig1.3) reprend l'histogramme et les principaux paramètres de cette population dont la caractéristique la plus marquante est la très forte dissymétrie gauche.

L'exemple présente donc un caractère artificiel, dans la mesure où on connaît exactement les caractéristiques de la population, celle-ci étant simulée. Dans la population théorique en question, on a sélectionné, de manière aléatoire et simple, un échantillon de 100 observations. La deuxième colonne du tableau (TAB.1.3), notée x, donne les premières et les dernières observations de l'échantillon, après classement des données par ordre croissant. Les trois colonnes suivantes donnent les premières et les dernières observations de trois échantillons de 100 observations prélevés dans l'échantillon initial et notés x_1^*, x_2^* et x_3^*, ceux-ci ayant également été classés par ordre croissant. On constate, par exemple, que pour l'échantillon x_1^*, la première

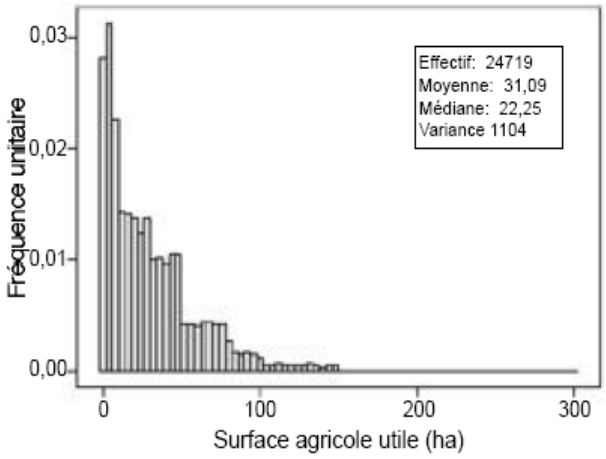


FIG. 1.3: *Population simulée des tailles des exploitations de la région Wallonne en 1995*

observation de l'échantillon initial a été sélectionnée deux fois et que la deuxième observation de l'échantillon initial n'a, par contre, pas été sélectionnée. Le rééchantillonnage se faisant avec remise, il est tout à fait normal que certaines observations de l'échantillon initial soient absentes, ou au contraire apparaissent plus d'une fois.

Numéro d'ordre	x	x_1^*	x_2^*	x_3^*
1	0,00	0,00	0,00	0,00
2	0,18	0,00	0,00	0,00
3	0,36	0,36	0,18	0,36
4	0,46	1,04	0,18	0,36
...				
97	96,81	85,55	96,81	91,61
98	98,60	85,55	98,60	96,81
99	133,60	91,61	133,60	98,60
100	145,21	96,81	133,60	98,60

TAB. 1.3: *Échantillon initial x et résultats de trois rééchantillonnages, x_1^*, x_2^* et x_3^* (données partielles).*

Pour l'ensemble des B échantillons obtenus par *bootstrap* (*bootstrap* sample), les observations x_i n'apparaissent pas en nombre égal et on peut définir les proportions d'apparition P_i^* de chacune des observations, P_i^* étant égal au nombre de fois que l'observation x_i a été prélevée pour l'ensemble des B échantillons, divisé par le nombre total de prélèvements, qui est égal à nB. Ces proportions P_i^* interviennent dans certaines estimations. Des méthodes de rééchantillonnage assurant l'égalité de ces proportions sont également proposées. Cette approche porte le nom de rééchantillonnage balancé (*bootstrap* bayésien) (voir [18, 30, 4]).

1.4 *Bootstrap* des résidus

La technique de rééchantillonnage présentée ci-dessus est la plus simple et la plus courante. Des méthodes un peu différentes sont utilisées pour des applications particulières. Ainsi, dans les problèmes de régression lorsque les valeurs des variables explicatives sont fixées a priori par l'utilisateur, le rééchantillonnage d'individus peut difficilement se justifier. Dans une telle situation, on peut remplacer le *bootstrap* des individus par le *bootstrap* des résidus.

Soit y le vecteur de la variable à expliquer et Z la matrice des variables explicatives. L'échantillon initial est donc décrit par la juxtaposition du vecteur y et de la matrice Z.

Soit $\hat{\beta}$ le vecteur des coefficients estimés par une méthode donnée d'ajustement, qui n'est pas nécessairement la méthode des moindres carrés. On peut calculer le vecteur des valeurs estimées de la variable à expliquer et en déduire le vecteur des résidus :

$$e = y - \hat{y}.$$

Dans le cas du modèle linéaire, on a :

$$\hat{y} = Z\hat{\beta}.$$

Mais, de manière plus générale, y_i peut être une fonction quelconque des valeurs observées des variables explicatives et des paramètres estimés :

$$\hat{y} = f(Z, \hat{\beta}).$$

Soit e_k^*, un échantillon aléatoire et simple prélevé avec remise dans le vecteur e. Si le vecteur e n'est pas de moyenne nulle, il est nécessaire de soustraire cette moyenne de chacun des résidus, avant de procéder au rééchantillonnage (Léger et al., 1992 [64]).

En additionnant e_k^* à la partie déterministe du modèle $f(Z, \hat{\beta})$, on obtient le vecteur y_k^* :

$$y_k^* = f\left(Z, \hat{\beta}_k\right).$$

La juxtaposition de y_k^* et de Z constitue le résultat du rééchantillonnage x_k^*. Comme précédemment, la procédure décrite est répétée B fois.

Il faut noter que l'application du *bootstrap* à la régression n'implique pas nécessairement le *bootstrap* des résidus. Le choix de l'une ou de l'autre méthode dépend du caractère fixe ou aléatoire de la matrice Z, mais dépend aussi des hypothèses relatives au modèle. Le rééchantillonnage des résidus suppose en effet que les résidus ne sont pas fonction des variables explicatives, ce qui n'est, par exemple, pas le cas en présence d'inégalité des variances conditionnelles ou d'inadéquation de la relation.

Soit l'exemple suivant, afin d'illustrer le principe de cette méthode, qui représente la variation du taux de cholestérol (y) en fonction du pourcentage de la dose prescrite effectivement absorbée (x).

$x_i(\%)$	0	2	7	8	16	33	43	...	100
y_i	11.5	5.75	-10.5	36.25	29.75	27.75	33.25		86.75

Supposons que le modèle d'ajustement adéquat qui lie ces deux variables (x et y) est sous la forme suivante :

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i.$$

Pour l'estimation des paramètres β_1, β_2 et β_3 on utilise, par exemple la méthode des moindres carrés ; à cet effet on aura : $\hat{\beta}_1, \hat{\beta}_2$ et $\hat{\beta}_3$. Le modèle des prédites est alors sous la forme :

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2.$$

L'application de *bootstrap* pour ce cas peut se faire de deux manières différentes.

1.4.1 1^{ere} approche(*bootstrap* des paires)

La premiere approche consiste a fair un tirage aléatoire simple avec remise (principe de *bootstrap*) d'un couple de variable (y_i, x_i) et la statistiques d'intérêt dans ce cas est les valeurs prédites $\hat{y}_i$. A cet effet, on estime les paramètres $\hat{\beta}_0, \hat{\beta}_1$ et $\hat{\beta}_2$ pour chaque échantillon toute en calculant les valeurs des prédites $\hat{y}_i$ correspondantes :

$$\begin{cases} x_1^* & 0 & 54 & 43 & 2 & ... & 16 \\ y_1^* & 11.5 & 47.25 & 33.25 & 5.75 & ... & 29.75 \end{cases} \Rightarrow \text{estimation}(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) \Rightarrow \text{calcule de } y_i^{*1}$$

$$\begin{cases} x_b^* & 33 & 95 & 7 & 43 & ... & 72 \\ y_b^* & 27.75 & 77.00 & -10.5 & 33.25 & ... & 63.00 \end{cases} \Rightarrow \text{estimation}(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) \Rightarrow \text{calcule de } y_i^{*b}$$

....

$$\begin{cases} x_B^* & 100 & 72 & 43 & 28 & ... & 7 \\ y_B^* & 86.75 & 63.00 & 33.25 & 23.5 & ... & -10.5 \end{cases} \Rightarrow \text{estimation}(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) \Rightarrow \text{calcule de } y_i^{*B}$$

dans ce cas :

$$\hat{y}_i^* = \frac{1}{B} \sum_{b=1}^{B} \hat{y}_i^{*b}$$

et l'écart-standard

$$\sigma_{\hat{y}_i^*} = \sqrt{\frac{\sum_{b=1}^{B} (\hat{y}_i^{*b} - \hat{y}_i^*)^2}{B-1}}.$$

1.4.2 2^{eme} approche (*bootstrap* des résidus)

La deuxième approche consiste a faire un tirage aléatoire sur les résidus après l'ajustement du modèle : $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i$ c'est-à-dire une fois les paramètres $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ sont estimer par une méthode bien déterminée (moindres carrés par exemple), on construit l'échantillon des résidus

$$\hat{e}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2)$$

$$\hat{e} = 1.2 \;\; 2.4 \;\; -1.3 \;\; ... \;\; -0.8.$$

On applique ensuite la méthode de rééchantillonnage de *bootstrap* classique sur ce dernier échantillon (l'échantillon e)

e_i^{*1} 2.4 -1.3 0.7 ... 0.6 $\Rightarrow$ modèle $y_i^{*1} = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 + e_i^{*1} \Rightarrow$ calcule de y_i^{*1}

e_i^{*b} -1.3 -0.8 1.6 ... 1.2 $\Rightarrow$ modèle $y_i^{*b} = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 + e_i^{*b} \Rightarrow$ calcule de y_i^{*b}

....

e_i^{*B} 2.4 1.2 0.5 ... -0.1 $\Rightarrow$ modèle $y_i^{*B} = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 + e_i^{*B} \Rightarrow$ calcule de y_i^{*B}.

Enfin le calcul de $\hat{y}_i^*$ et $\sigma_{\hat{y}_i^*}$ se fait de la même manière que la première approche (section 1.4.1).

Remarque 1.1

1. *Si le modèle est incertain, on opte pour le bootstrap des paires, car cette dernière suppose que les paires sont des réalisations aléatoires de la population.*
2. *la bootstrap des paires est aussi recommandé pour le cas ou le modèle nécessaire n'est pas clairement défini.*
3. *l'inconvénient de la bootstrap des résidus est qu'elle est sensible au modèle.*

1.5 Erreur-standard et biais d'un paramètre

1.5.1 Estimation de l'erreur-standard

Soit un paramètre θ d'une population donnée et soit :

$$\hat{\theta} = f(x_1, x_2, ..., x_n) = f(x),$$

une estimation de ce paramètre, obtenue à partir des données de l'échantillon initial X.

Chaque échantillon obtenu par rééchantillonnage permet de calculer une répétition du *bootstrap* (*bootstrap* replication) de l'estimation $\hat{\theta}$:

$$\hat{\theta}_k^* = f(x_k^*), \quad (k = 1, ..., B),$$

la fonction f étant la même que celle utilisée pour la définition de θ.

Supposons qu'on s'intéresse à la moyenne, à la médiane et à la variance de la distribution des tailles des exploitations agricoles et qu'on se propose d'estimer ces trois paramètres à partir de l'échantillon X. Si on utilise les estimateurs classiques, le paramètre $\hat{\theta}$ s'écrit, successivement pour les trois paramètres considérés :

$$\bar{x} = \frac{1}{100}\sum_{i=1}^{100} x_i \ , \quad \tilde{x} = \frac{1}{2}\left(x_{[50]} + x_{[51]}\right) \quad \text{et} \quad \hat{\sigma}^2 = \frac{1}{99}\sum_{i=1}^{100}(x_i - \bar{x})^2,$$

$x_{[50]}$ et $x_{[51]}$ étant les observations de rangs 50 et 51 de l'échantillon initial. Les valeurs numériques pour ces trois estimations sont données dans la première partie du tableau (TAB.1.4), sur la ligne intitulée $\hat{\theta}$.

Paramètre	Moyenne	Médiane	Variance
θ	28,13	21,56	854,63
$\hat{\theta}^*_1$	27,84	18,91	667,19
$\hat{\theta}^*_2$	26,32	19,95	796,93
$\hat{\theta}^*_3$	31,22	23,15	708,62
.			
.			
.			
$\hat{\theta}^*$	27,99	20,44	843,58
$\hat{\sigma}_{\hat{\theta}^*}$	2,89	2,53	184,25

TAB. 1.4: *Paramètres estimés pour l'échantillon initial($\hat{\theta}$) et pour les trois premiers échantillons obtenus par rééchantillonnage ($\hat{\theta}^*_1, \hat{\theta}^*_2, \hat{\theta}^*_3$) ; moyennes ($\hat{\theta}^*$) et écarts-types ($\sigma_{\hat{\theta}^*}$) des paramètres estimés pour 1 000 rééchantillonnages.*

Les calculs des trois paramètres peuvent être répétés pour les échantillons x^*_1, x^*_2 et x^*_3. Les résultats obtenus sont repris dans la seconde partie du tableau (Tab.1.4), sur les lignes intitulées θ^*_1, θ^*_2 et θ^*_3. Cette seconde partie du tableau peut évidemment être complétée au fur et à mesure des rééchantillonnages fournissant x^*_4, x^*_5 , . . ., x^*_B.

Disposant des B répétitions, on peut déterminer la moyenne :

$$\hat{\theta}^* = \frac{1}{B}\sum_{k=1}^{B}\hat{\theta}^*_k$$

et l'écart-type des $\hat{\theta}^*_k$:

$$\hat{\sigma}_{\hat{\theta}^*} = \sqrt{\frac{1}{(B-1)}\sum_{k=1}^{B}(\hat{\theta}^*_k - \hat{\theta}^*)} \quad .$$

On peut résumer le calcul de l'erreur standard de la méthode de *bootstrap* par l'algorithme suivant :

Algorithme d'estimation des erreurs-standards

1- Tirer B échantillons *bootstrap* $X^*_1,\ X^*_2,\ \ldots, X^*_B$ à partir de X

2- Calculer la copie *bootstrap* $\hat{\theta^*}_k = s(X^*_k)$

3- Calculer l'erreurs-standard pour les B copies

$$\hat{\sigma}_{\hat{\theta}^*} = \left\{\sum_{b=1}^{B}(\hat{\theta^*}_b - \hat{\theta}^*)^2/(B-1)\right\}^{1/2} \quad \text{avec} \quad \hat{\theta}^* = \sum_{b=1}^{B}\hat{\theta^*}_b/B.$$

Pour l'exemple ci-dessus, 1000 rééchantillonnages ont été réalisés. Les figures (Fig.1.4) donnent la distribution des valeurs obtenues pour les trois paramètres considérés. Pour la moyenne et pour la variance, on constate que la distribution est en cloche et relativement symétrique. Par contre, la distribution des médianes est assez différente puisqu'elle présente un caractère bimodal assez prononcé. La moyenne et l'écart-type des 1000 moyennes, des 1000

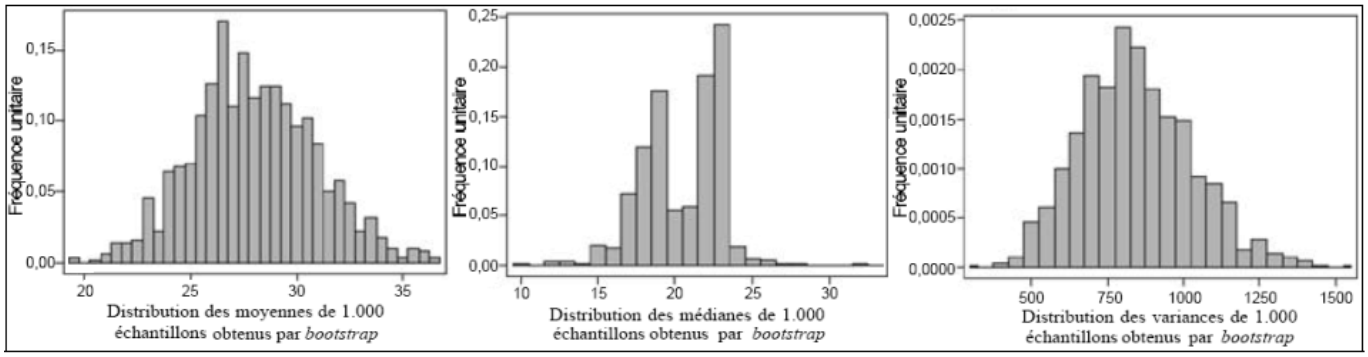


FIG. 1.4: *Distribution des valeurs de la moyenne, médiane et variance des échantillons bootstrapé.*

médianes et des 1000 variances sont donnés dans la troisième partie du tableau (TAB.1.4), dans les lignes intitulées $\hat{\theta}^*$ et $\hat{\sigma}_{\hat{\theta}^*}$.

L'écart-type $\hat{\sigma}_{\hat{\theta}^*}$ est une estimation de l'erreur standard de l'estimateur du paramètre θ. Pour les situations où on dispose d'un estimateur de cette erreur-standard, et pour autant que les conditions d'application soient remplies, on peut montrer que l'écart-type des $\hat{\sigma}_{\hat{\theta}^*_k}$ tend vers le résultat analytique, lorsque B tend vers l'infini.

Ainsi, pour la moyenne d'un échantillon aléatoire et simple, on sait que l'erreur-standard de la moyenne est égale à $\hat{\sigma}/\sqrt{n}$. Si B tend vers l'infini, l'écart-type $\hat{\sigma}_{\hat{\theta}^*}$ tend vers $\hat{\sigma}$, avec :

$$\hat{\sigma}_{plug} = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2/n} \quad .$$

L'estimation $\hat{\sigma}_{plug}$ de l'écart-type de la population donnée ci-dessus est appelée estimation par insertion (plug-in estimator). Pour un estimateur par insertion, la formule conduisant à l'estimation est la même que celle utilisée pour la définition du paramètre de la population. En effet, pour une population finie de taille N et de moyenne m_X, on a :

$$\sigma = \sqrt{\sum_{i=1}^{N}(x_i - m_X)^2/N} \quad .$$

En d'autres mots, pour un estimateur par insertion, on considère l'échantillon comme une population particulière et on utilise la formule relative au paramètre de la population.

Pour l'exemple ci-dessus, l'écart-type des moyennes obtenues pour les différents échantillons x_k^* est égal à 2,89 (voir Tab.1.4). Si on augmentait indéfiniment le nombre de répétitions B, cet écart-type se rapprocherait de 2,91 puisque :

$$\hat{\sigma}^2_{plug} = (n-1)\hat{\sigma}^2/n = 99(854,63)/100 = 846,08,$$

et

$$\sqrt{\hat{\sigma}^2_{plug}/n} = \sqrt{846,08/100} = 2,91 \ .$$

D'une manière générale, lorsque B tend vers l'infini, la valeur $\hat{\sigma}_{\hat{\theta}^*}$ tend vers une valeur fixée qui correspond à l'estimation de l'erreur-standard du *bootstrap* idéal.

Efron et Tibshirani (1993) [32] proposent les règles empiriques suivantes pour le choix de B :

– Un nombre réduit de répétitions permet d'obtenir une première information et $B = 50$ est généralement suffisant pour avoir une bonne estimation de l'erreur-standard ;
– Il est très rare que plus de 200 répétitions soient nécessaires pour estimer une erreur-standard.

On peut noter que le choix de B n'est pas en fonction de la taille n de l'échantillon.

1.5.2 Estimation du biais

Le biais d'un paramètre peut être estimé par la méthode du *bootstrap* de la manière suivante :

$$biais_B(\hat{\theta}) = \hat{\theta}^* - \hat{\theta}_{plug} \quad , \tag{1.4}$$

$\hat{\theta}_{plug}$ étant l'estimation par plug-in du paramètre θ. la formule (1.4) est valable même si $\hat{\theta}$ n'est pas une estimation par plug-in.

Ainsi, pour la moyenne et pour la médiane $\hat{\theta}$ est égale à $\hat{\theta}_{plug}$ et l'estimation du biais pour ces deux paramètres est égale à :

$$27,99 - 28,13 = -0,14 \quad \text{et} \quad 20,44 - 21,56 = -1,12.$$

Pour la variance, l'estimation par insertion est égale à 846,08 et l'estimation du biais est égale à :

$$843,58 - 846,08 = -2,50.$$

Disposant d'une estimation du biais, on peut éventuellement corriger l'estimation initiale. On obtient alors :

$$\hat{\theta}_C = \hat{\theta} - biais_B(\hat{\theta}).$$

On notera cependant que la correction systématique du biais, par la relation ci-dessus, peut s'avérer dangereuse dans la mesure où cette correction peut augmenter l'erreur-standard de manière importante.

En pratique, lorsque le rapport du biais à l'erreur-standard est inférieur à 0,25, il est souvent préférable de ne pas corriger l'estimateur pour le biais. D'autre part, un rapport supérieur à 0,25 peut être une indication que la statistique $\hat{\theta} = f(x_1, ..., x_n)$ est inappropriée pour estimer θ.

Pour l'exemple ci-dessus, il n'y a certainement pas intérêt à corriger les estimations pour le biais, puisqu'on peut montrer que les trois estimations sont non biaisées (Dagnelie, 1998 [22]).

Lorsque $\hat{\theta}$ est une estimation par insertion, on montre qu'une meilleure estimation du biais est obtenue en tenant compte des proportions d'apparition P_i^*, définies dans le dernier paragraphe de la section (1.3). Il suffit de remplacer, dans la formule ci-dessus, $\hat{\theta}_{plug}$ par $\hat{\theta}'_{plug}$, $\hat{\theta}'_{plug}$ étant une fonction des observations de l'échantillon x qui attribue aux observations x_i un poids égal à P_i^*, au lieu de leur attribuer des poids égaux à $P_0 = 1/n$.

1.6 Intervalles de confiance

1.6.1 Méthode de l'erreur-standard

Une première solution consiste à définir l'intervalle de confiance par la méthode de l'erreur-standard (standard *bootstrap* confidence interval) (voir [30]) :

$$\hat{\theta} \pm u_{(1-\alpha/2)}\hat{\sigma}_{\hat{\theta}^*},$$

$u_{1-\alpha/2}$ étant le pourcentile $1-\alpha/2$ de la distribution normale réduite et $1-\alpha$ étant le degré de confiance retenu.

Pour que cette approche soit satisfaisante, il faut que la distribution d'échantillonnage du paramètre étudié soit approximativement normale, que l'estimateur soit non biaisé, et que $\hat{\sigma}^*_{\hat{\theta}^*}$ soit une bonne estimation de l'erreur-standard de la distribution du paramètre.

Le fait que ces conditions soient remplies ou non dépend des circonstances. La condition de normalité peut être vérifiée à partir de la distribution des $\hat{\theta}^*_k$ et il peut être utile éventuellement d'effectuer une transformation de manière à rendre la distribution plus proche de la normale.

Le biais de l'estimateur peut être estimé, comme nous l'avons vu au paragraphe (1.5.2), mais sa prise en compte risque d'augmenter la variance de l'estimateur.

Enfin, la qualité de l'estimation de l'erreur-standard est liée au nombre de répétitions B considéré et nous avons signalé précédemment que 50 répétitions sont généralement suffisantes.

Pour l'exemple considéré et pour un degré de confiance de 95%, on obtient les limites de confiance suivantes, respectivement pour la moyenne, pour la médiane et pour la variance :

$$28,13 \pm (1,96)(2,89), \text{ soit } [22,47\ ,\ 33,79],$$
$$21,56 \pm (1,96)(2,53), \text{ soit } [16,60\ ,\ 26,52]$$
$$\text{et } 854,63 \pm (1,96)(184,25), \text{ soit } [493,50\ ,\ 1215,76].$$

1.6.2 Méthode des pourcentiles simples

Dans la méthode des pourcentiles simples (simple percentile confidence interval) (voir [30]), les limites de confiance sont données par les pourcentiles $\alpha/2$ et $1-\alpha/2$ de la distribution d'échantillonnage empirique, c'est-à-dire de la distribution des $\hat{\theta}^*$. Nous les notons $\hat{\theta}^*_{[\alpha/2]}$ et $\hat{\theta}^*_{[1-\alpha/2]}$.

Contrairement à la méthode de l'erreur-standard, la distribution d'échantillonnage du paramètre étudié ne doit pas être normale pour que la méthode des pourcentiles soit satisfaisante. Par contre, le nombre de rééchantillonnages B doit être plus élevé que dans le cas de la méthode de l'erreur-standard, car il faut un plus grand nombre d'observations pour estimer, avec une précision suffisante, un pourcentile que pour estimer un écart-type. B sera par exemple de l'ordre de 1000.

Pour 1000 rééchantillonnages et pour un degré de confiance de 95%, les pourcentiles 0,025 et 0,975 correspondent approximativement à l'observation de rang 25 et à l'observation de rang 975, la valeur exacte pouvant dépendre de l'algorithme utilisé pour le calcul de ces pourcentiles. Les résultats obtenus pour les trois paramètres considérés dans l'exemple sont les suivants : 22,48 et 33,89 pour la moyenne, 14,82 et 24,01 pour la médiane, et 512,92 et 1248,80 pour la variance.

Il faut noter aussi qu'une procédure de calcul un peu différente a été proposée par Hall (1992)[46] et est décrite par Manly (1997) [67]. La méthode consiste à calculer les écarts :

$$\hat{e}^*_k = \hat{\theta}^*_k - \hat{\theta}\ \ ,$$

et à déterminer les pourcentiles $\alpha/2$ et $1-\alpha/2$, notés $\hat{e}^*_{[\alpha/2]}$ et $\hat{e}^*_{[1-\alpha/2]}$ de cette distribution. Les limites de confiance sont alors données par les relations :

$$\hat{\theta} - \hat{e}^*_{[1-\alpha/2]} \ \text{ et } \ \hat{\theta} - \hat{e}^*_{[\alpha/2]}\ .$$

Par cette méthode on obtient, dans le cas de l'exemple 22,37 et 33,78 pour la moyenne, 19,11 et 28,31 pour la médiane, et 460,46 et 1196,34 pour la variance.

1.6.3 Méthode des pourcentiles corrigés pour le biais

On détermine d'abord la proportion p de valeurs $\hat{\theta}^*_k$ inférieures à $\hat{\theta}$ et on calcule le pourcentile u_p relatif à la distribution normale réduite.

Soit α_1 et α_2 les valeurs de la fonction de répartition de la normale réduite aux points u_1 et u_2 :

$$\alpha_1 = \Phi(u_1) \quad \text{et} \quad \alpha_2 = \Phi(u_2),$$

avec $u_1 = 2u_p + u_{\alpha/2}$ et $u_2 = 2u_p + u_{1-\alpha/2}$.

Les limites de confiance déterminées par la méthode des pourcentiles corrigés pour le biais (bias corrected percentile confidence interval) sont alors les pourcentiles $\hat{\theta}^*_{\alpha_1}$ et $\hat{\theta}^*_{\alpha_2}$ de la distribution des $\hat{\theta}^*_k$.

Des informations concernant l'origine de cette correction sont données dans Efron et Tibshirani(1993) [32] et dans Chernick(1999)[17].

On remarque que si $p = 0,5$, c'est-à-dire si $\hat{\theta}$ est la médiane de la distribution des $\hat{\theta}^*_k$, il n'y a pas de correction pour le biais, puisque $u_p = 0$, et on retrouve la méthode précédente. Si p est inférieur à $0,5$, les limites de confiance correspondent à des pourcentiles inférieurs respectivement à $\alpha/2$ et $1-\alpha/2$. Au contraire, si p est supérieur à 0,5, les limites correspondent à des pourcentiles supérieurs à $\alpha/2$ et $1-\alpha/2$. Par exemple, si p = 0,4, les limites de confiance, pour un degré de confiance de 0,95, sont les pourcentiles $\hat{\theta}^*_{[0,0068]}$ et $\hat{\theta}^*_{[0,9269]}$. Pour $p = 0,6$ les limites correspondent à $\hat{\theta}^*_{[0,0731]}$ et $\hat{\theta}^*_{[0,9932]}$.

Pour les exploitations agricoles, les valeurs de p sont égales à 0,542, 0,533 et 0,555 respectivement pour la moyenne, pour la médiane et pour la variance. Les limites de confiance seront par conséquent toutes légèrement plus grandes que dans le cas de la méthode des pourcentiles simples. On obtient en effet 23,03 et 34,62 pour la moyenne,15,78 et 24,17 pour la médiane et 554,15 et 1303,12 pour la variance.

1.6.4 Méthode des pourcentiles avec correction pour le biais et accélération

La méthode précédente, qui prend en compte le biais, peut être étendue de manière à tenir compte d'un éventuel changement de l'erreur-standard de $\hat{\theta}$ lorsque θ varie. Elle porte alors le nom de méthode des pourcentiles avec correction pour le biais et accélération (bias corrected and accelerated confidence interval). Une justification de cette méthode est donnée par Efron et Tibshirani (1993)[32].

Les limites de confiance sont les pourcentiles $\hat{\theta}^*_{[\alpha_1]}$ et $\hat{\theta}^*_{[\alpha_2]}$ de la distribution des $\hat{\theta}^*_k$, α_1 et α_2 étant les valeurs de la fonction de répartition de la variable normale réduite aux points u_1 et u_2 définis de la manière suivante :

$$u_1 = u_p + (u_p + u_{\alpha/2})/[1 - a(u_p + u_{\alpha/2})]$$
$$u_2 = u_p + (u_p + u_{1-\alpha/2})/[1 - a(u_p + u_{1-\alpha/2})].$$

Dans ces relations, u_p est défini comme précédemment et la constante a est appelée accélération, car elle est liée au taux de variation de l'erreur-standard de $\hat{\theta}$ lorsque le paramètre θ varie. Cette constante peut être estimée de différentes manières. Une solution consiste à utiliser la technique du jackknife. On obtient alors le paramètre a par la relation suivante :

$$a = \frac{\sum_{i=1}^{n}(\hat{\theta}_J - \hat{\theta}_{-i})^3}{6\left[\sum_{i=1}^{n}(\hat{\theta}_J - \hat{\theta}_{-i})^2\right]^{3/2}}.$$

Dans cette relation, $\hat{\theta}_{(-i)}$ est l'estimation du paramètre θ obtenue à partir de l'échantillon initial, dont on a enlevé la i^{eme} observation et $\hat{\theta}_J$ est la moyenne des n valeurs $\hat{\theta}_{(-i)}$.

On peut constater que si $a = 0$, on retrouve la méthode des pourcentiles corrigés pour le biais. La prise en compte de l'accélération constitue donc bien une extension de la méthode précédente.

Les limites obtenues par cette méthode sont, pour l'exemple examiné, 23,21 et 35,33 pour la moyenne, 15,78 et 24,17 pour la médiane et 589,75 et 1405,85 pour la variance.

On constate que pour la médiane les résultats sont identiques à ceux obtenus par la méthode précédente, le paramètre a étant nul, du fait de la parfaite symétrie de la distribution des $\tilde{x}_{(-i)}$.

1.6.5 Méthode du *bootstrap* $- t$

L'idée mise en application dans le *bootstrap* $- t$ est de définir une statistique dont la distribution ne soit pas en fonction de la valeur réelle et inconnue du paramètre θ. La statistique T :

$$T = \frac{\hat{\theta} - \theta}{\sigma_\theta},$$

peut remplir ce rôle.

Il s'agit alors d'approcher la distribution théorique de T par rééchantillonnage. Dans ce but, on calcule :

$$t_k^* = \frac{\hat{\theta}_k^* - \hat{\theta}}{\hat{\sigma}\left(\hat{\theta}_k^*\right)},$$

$\hat{\sigma}\left(\hat{\theta}_k^*\right)$ étant l'erreur-standard de $\hat{\theta}_k^*$, qui dépend donc de l'échantillon k. Elle peut être calculée par une formule théorique, lorsqu'une telle formule est disponible, ou à partir du rééchantillonnage de l'échantillon x_k^* utilisé pour calculer $\hat{\theta}_k^*$. Il s'agit alors d'un *bootstrap* à deux niveaux, puisque chaque échantillon x_k^* fait lui-même l'objet d'un rééchantillonnage, permettant de calculer l'erreur standard.

Disposant de la distribution des t_k^*, on en détermine les pourcentiles $\alpha/2$ et $1 - \alpha/2$ notés $t_{[\alpha/2]}^*$ et $t_{[1-\alpha/2]}^*$, et on obtient les limites de confiance du paramètre θ par les relations :

$$\hat{\theta}_1 = \hat{\theta} - t_{[1-\alpha/2]}^* \hat{\sigma}_{\hat{\theta}^*} \qquad \text{et} \qquad \hat{\theta}_2 = \hat{\theta} - t_{[\alpha/2]}^* \hat{\sigma}_{\hat{\theta}^*}.$$

Le *bootstrap* à deux niveaux a été appliqué à l'échantillon des 100 tailles d'exploitations agricoles, le nombre de répétitions au deuxième niveau étant égal à 50. Pour la moyenne, les pourcentiles 0,025 et 0,975 de la distribution observée des t_k^* sont les suivants :

$$t_{[0,025]}^* = -2,3364 \qquad \text{et} \qquad t_{[0,975]}^* = 1,7574.$$

Les limites de confiance de la moyenne sont par conséquent égales à :

$$28,13 - (1,7574)(2,89) = 23,05 \quad \text{et} \quad 28,13 - (-2,3364)(2,89) = 34,88.$$

Des calculs identiques ont été réalisés pour la médiane et pour la variance ; les limites de confiance obtenues sont respectivement 18,45 et 27,66 pour la médiane et 561,19 et 1523,14 pour la variance.

Remarque 1.2 *Des études de comparaison des différentes méthodes proposées montrent que la méthode des pourcentiles corrigés pour le biais et accélération offre, dans l'ensemble, le plus d'avantages et est, de ce fait, préconisée par Efron et Tibshirani [1993].*

1.7 Conclusion

"Le *bootstrap* est une méthode d'inférence statistique basée sur l'utilisation de l'ordinateur qui peut répondre sans formules à beaucoup de questions statistiques réelles". C'est en ces termes qu'Efron et Tibshirani (1993)[32] présentent le *bootstrap* dans la préface de leur ouvrage.

Il est incontestable que l'utilisation des techniques de rééchantillonnage a été rendue possible grâce à la généralisation des moyens de calcul performants. Ces techniques reposent, au départ, sur des idées simples. Toutefois, il faut bien admettre que les développements apportés aux méthodes de base leur ont fait perdre une partie de cette simplicité.

Dans ce chapitre, nous nous sommes limités aux problèmes de l'estimation du biais et de l'erreur-standard d'un paramètre, et à la détermination des limites de confiance d'un paramètre. Il ne s'agit cependant pas là des seules applications des méthodes de rééchantillonnage. Celles-ci peuvent, en effet, aussi être utilisées pour la réalisation de différents tests d'hypothèses, pour le choix des variables et l'estimation de l'erreur de prédiction en régression, pour l'estimation du taux d'erreur en analyse discriminante, notamment. Bien qu'elles puissent être utilisées dans des situations très variées, leur mise en oeuvre ne présente guère d'intérêt lorsque l'inférence statistique peut être réalisée par des méthodes analytiques classiques, pour lesquelles les conditions d'application sont remplies.

Elles ne sont donc pas destinées à remplacer les méthodes d'inférence statistique classiques lorsque celles-ci sont applicables mais plutôt à fournir des réponses à des questions pour lesquels les méthodes classiques sont inapplicables ou non disponibles. Ainsi, pour l'exemple traité (section 1.5.1), le recours au *bootstrap* ne se justifie certainement pas dans le cas de la moyenne et de la médiane. Pour la moyenne, on peut en effet utiliser la méthode de l'erreur-standard, compte tenu du caractère approximativement normal de la distribution d'échantillonnage de la moyenne, vu la taille de l'échantillon. Pour la médiane, on dispose d'une approche non paramétrique. Par contre, pour la variance, l'utilisation du *bootstrap* constitue une alternative valable.

Le caractère relativement général des problèmes qui peuvent être résolus par *bootstrap* ne doit pas faire perdre de vue que la qualité de l'inférence dépend de la nature de la question posée et de la disponibilité des données. Comme le suggère Manly (1997)[67], le *bootstrap* doit être utilisé avec prudence dans les situations où il n'a pas encore été testé de manière approfondie. Une discussion assez générale sur l'intérêt et les limites du *bootstrap* est donnée dans la synthèse proposée par Young (1994)[106].

Dans l'estimation de la densité de probabilité et la courbe de régression de la moyenne par la méthode du noyau de Parzen-Rosenblatt, se pose le problème du choix du paramètre de lissage. Plusieurs méthodes existent (plug-in et cross validation) mais aucune n'est meilleure que les autres. C'est pourquoi nous allons appliquer cette technique de bootstrap à l'estimation du paramètre de lissage qui intervient dans l'estimation de la densité de probabilité.

Chapitre 2

Estimation de la densité de probabilité par la méthode du noyau

2.1 Introduction

Soit $x_1, x_2, \ldots, x_n$ n observations équipondérées issues d'une variable aléatoire réelle X de densité de probabilité réelle $f(x)$ inconnue. Comment obtenir une estimation de $f(x)$ à partir de la seule information contenue dans l'échantillon ?

Ce problème, que l'on désigne généralement par *estimation non paramétrique de la densité de probabilité* a fait l'objet de multiples travaux par des méthodes diverses, citons :

- L'estimateur par histogramme
- L'estimateur par les séries orthogonales
- Les estimateurs par histogrammes modifiés
- Les méthodes à base de splines
- L'estimateur par la méthode du noyau.

Dans ce chapitre, nous allons présenter une étude détaillée de l'estimateur par la méthode du noyau ainsi que ses propriétés statistiques.

2.2 Critères d'erreur et définitions

Avant de présenter l'estimateur à noyau de $f(x)$ ainsi que ses propriétés, il est intéressant de citer quelques normes de mesure d'erreur qui sont un critère de performance de cet estimateur.

2.2.1 Les différents critères d'erreur

Soit f_h un estimateur de la densité de probabilité f.

• **Les distances L_p :**

La distance L_p entre f et f_h est définie par :

$$L_p(f, f_h) = \begin{cases} \left(\int \left| f(x) - f_h(x) \right|^p dx \right)^{1/p}, & (0 < p < \infty) ; \\ \sup_x \left| f(x) - f_h(x) \right|, & (p = \infty). \end{cases}$$

- **Les distances de Hellinger H_p :**

$$H_p(f, f_h) = \left(\int \left[f^{1/p}(x) - f_h^{1/p}(x) \right]^p dx \right)^{1/p} \qquad (p < \infty).$$

- **L'information de Kullback-Leibler D :**

L'information de **Kulback-Leibler** est définie par :

$$D(f, f_h) = \begin{cases} \int f(x) ln(\frac{f(x)}{f_h(x)}) & f \ll f_h, \\ \infty & sinon. \end{cases}$$

Par convention ln(0/0) est pris égal à 0.

- **La distance de χ^2 :**

$$\chi^2(f, f_h) = \begin{cases} \int \frac{(f(x)-f_h(x))^2}{f_h(x)} & f \ll f_h, \\ \infty & sinon. \end{cases}$$

Par convention (0/0) est pris égal à 0.

- **L'erreur quadratique intégrée : ISE**

$$\begin{aligned} ISE(f, f_h) &= \int [f(x) - f_h(x)]^2 dx \\ &= \int [f(x)^2 - 2f(x)f_h(x) + f_h^2(x)] dx. \end{aligned}$$

- **L'erreur quadratique moyenne : MSE**

$$\begin{aligned} MSE(f(x), f_h(x)) &= \mathbb{E}(f_h(x) - f(x))^2 \\ &= [f(x) - \mathbb{E}f_h(x)]^2 + \mathbb{E}(f_h^2(x)) - [\mathbb{E}(f_h(x))]^2 \end{aligned}$$

$$MSE(f(x), f_h(x)) = \mathbb{V}(f_h(x)) + Biais^2 f_h(x). \tag{2.1}$$

- **L'erreur quadratique moyenne intégrée : MISE**

$$\begin{aligned} MISE(f, f_h) &= \int MSE(f(x), f_h(x)) dx = \int \mathbb{E}(f(x) - f_h(x))^2 dx \\ &= \int [Biais^2 f_h(x) + \mathbb{V}(f_h(x))] dx. \end{aligned}$$

2.2.2 Quelques définitions

Définition 2.1 *On dit qu'un estimateur f_h de f est sans biais si :* $\mathbb{E}(f_h) = f$.

Définition 2.2 *On dit qu'un estimateur f_h de f est asymptotiquement sans biais si :*

$$\lim_{n \to \infty} \mathbb{E}(f_h) = f.$$

Définition 2.3 *Un estimateur f_h de f est dit asymptotiquement uniformément sans biais si :*

$$\lim_{n\to\infty} \sup_x |\mathbb{E}[f_h(x) - f(x)]| = 0.$$

Définition 2.4 *On dit qu'un estimateur f_h de f est ponctuellement consistant en moyenne quadratique si :*

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

Définition 2.5 *On dit qu'un estimateur f_h de f est uniformément consistant en moyenne quadratique intégrée si :*

$$\lim_{n\to\infty} MISE(f, f_h) = 0.$$

Définition 2.6 *On dit qu'un estimateur f_h de f est asymptotiquement normal si :*

$$f_h \xrightarrow{cv.loi} \mathcal{N}(\mathbb{E}(f_h), \mathbb{V}(f_h)).$$

2.3 L'estimateur de Rosenblatt

En 1962, Parzen [77] a étudié les propriétés fondamentales de l'estimateur à noyau de la densité, juste après son introduction par Rosenblatt [80]. A partir de ce moment, cet estimateur à noyau de la densité est devenu un objet classique étudié par les statisticiens. Pour les statisticiens, il est déjà devenu un exemple canonique d'estimateur non paramétrique de courbe, qui utilise des résultats de la théorie d'approximation et l'analyse harmonique.

L'estimateur de la densité de probabilité par la méthode du noyau est le plus répandu aujourd'hui, car il répond au problème du choix des différents paramètres dans l'estimation à histogramme et possède de bonnes propriétés. L'idée consiste à évaluer la densité f au point x en comptant le nombre d'observations tombées dans un certain voisinage de x sur $\mathbb{R}$.

Définition 2.7 *Soit $x_1, ..., x_n$ un échantillon de loi $f(x)$ sur $\mathbb{R}$, de fonction de répartition $F(x) = \int_{-\infty}^{x} f(t)dt$. On appelle fonction de répartition empirique associé à $x_1, ..., x_n$, la fonction aléatoire $F_n : \mathbb{R} \to [0, 1]$ définie pour tout $x \in \mathbb{R}$ par $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{x_i<x\}}$. On peut également écrire de manière équivalente*

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1(x_i)_{]-\infty,x[}. \tag{2.2}$$

La fonction de répartition empirique F_n est un estimateur simple de F. Il s'avère que cette fonction est un très bon estimateur de F.

$$nF_n(x) = \sum_{i=1}^{n} 1_{\{x_i<x\}} \xrightarrow{loi} \mathcal{B}(n, F(x)).$$

où $\mathcal{B}$ est la loi binomiale, dont l'espérance et la variance de $F_n(x)$ sont données respectivement par :

$$\mathbb{E}(F_n(x)) = F(x) \qquad \text{et} \qquad \mathbb{V}(F_n(x) = \frac{1}{n}[1 - F(x)]F(x).$$

A partir de la définition d'une densité de probabilité (basée sur la dérivée de la fonction de répartition) et en utilisant l'équation (2.2), la densité f peut s'écrire en ses points de continuité :

$$f_h(x) = \lim_{h \longrightarrow 0} \frac{F_n(x+h) - F_n(x-h)}{2h}. \tag{2.3}$$

$$\begin{aligned} f_h(x) &= \frac{1}{2nh} \sum_{i=1}^{n} \mathbf{1}(x_i)_{]x-h, x+h[} \\ &= \frac{1}{2nh} \sum_{i=1}^{n} \mathbf{1}_{\{x-h<x_i<x+h\}}. \end{aligned}$$

En posant

$$\omega(u) = \begin{cases} 1/2, & -1 < u \leq 1, \\ 0, & \text{sinon.} \end{cases}$$

On peut réécrire (2.3) sous la forme

$$f_h(x) = \frac{1}{nh} \sum_{i=1}^{n} \omega\big(\frac{x - x_i}{h}\big). \tag{2.4}$$

Nous venons de définir l'estimateur à noyau dit de **Rosenblatt** (uniforme).

Parzen[77] a étudié une classe générale d'estimateurs. En remplaçant la fonction w par une fonction noyau K(Kernel) satisfaisant la condition

$$\int_{-\infty}^{\infty} K(u)du = 1. \tag{2.5}$$

Généralement, K est une densité de probabilité. Par analogie avec la définition de l'estimateur de Rosenblatt l'estimateur à noyau (de Parzen) est :

$$f_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - x_i}{h}), \tag{2.6}$$

où $h = h(n)$ est un paramètre qui est fonction de n, appelé paramètre de lissage.
K est une fonction définie sur $\mathbb{R}$ appelée noyau.

Le noyau K vérifie les conditions suivantes :

$$\int_{\mathbb{R}} K(y)dy = 1, \quad \int_{\mathbb{R}} yK(y)dy = 0, \quad \text{et} \quad \int_{\mathbb{R}} y^2K(y)dy = \sigma_K^2 < \infty.$$

On peut vérifier que $f_h(x)$ est une densité de probabilité. Car $f_h(x) \geq 0 \ \forall x$, et

$$\int_{\mathbb{R}} f_h(x)dx = \int_{\mathbb{R}} \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - x_i}{h})dx = \frac{1}{nh} \sum_{i=1}^{n} \int_{\mathbb{R}} K\big(\frac{x - x_i}{h}\big)dx = \frac{1}{n} \sum_{i=1}^{n} \int_{\mathbb{R}} K(u)du = 1.$$

Pour assurer la convergence de l'estimateur $f_h(x)$, les seules conditions imposées sont :

$$h(n) \longrightarrow 0 \quad \text{et} \quad nh(n) \longrightarrow \infty \quad \text{quand} \quad n \longrightarrow \infty.$$

2.4 Propriétés de l'estimateur à Noyau

Cette section est consacrée à quelques résultats théoriques sur les propriétés de l'estimateur à noyau, à savoir :

- Le comportement asymptotique du biais et de la variance.
- La convergence en moyenne quadratique et en moyenne quadratique intégrée.
- La convergence uniforme (en probabilité,presque complète).
- La convergence en norme L_1.

Bochner a donné le premier résultat de convergence sous forme d'un lemme sur lequel les principaux théorèmes de convergences sont basés.

Lemme 2.4.1 *(de Bochner) :*

i) Soit K un noyau de Parzen-Rosenblatt et g une fonction de $\mathbb{L}^1$.
Alors, en tout point x, où g est continue :

$$\lim_{h\longrightarrow 0}(K_h * g)(x) = g(x).$$

ii) Soit K un noyau quelconque et g une fonction de $\mathbb{L}^1$ uniformément continue, alors

$$\lim_{h\longrightarrow 0}\sup_x |(K_h * g)(x) - g(x)| = 0.$$

L'interprétation de ce lemme est que lorsque la fenêtre h est petite, la convolution d'une fonction g de $\mathbb{L}^1$ avec K_h perturbe peu cette fonction.

2.4.1 Espérance, Biais et Variance de l'estimateur

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

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

En posant $y = \frac{x-u}{h} \Rightarrow dy = -\frac{du}{h}$

$$\mathbb{E}f_h(x) = \int_{-\infty}^{\infty} K(y)f(x-hy)dy \tag{2.7}$$

• Le biais de $f_h(x)$ est :

$$Biais\ f_h(x) = \mathbb{E}f_h(x) - f(x) = \int_{-\infty}^{\infty} K(y)f(x-hy)dy - f(x). \tag{2.8}$$

• La variance de $f_h(x)$ est :

$$\begin{aligned}\mathbb{V}f_h(x) &= \mathbb{V}\sum_{i=1}^{n}\frac{1}{nh}K\big(\frac{x-x_i}{h}\big)\\ &= \frac{1}{n^2h^2}\sum_{i=1}^{n}\mathbb{V}K\big(\frac{x-x_i}{h}\big)\\ &= \frac{1}{n^2h^2}\sum_{i=1}^{n}\left[\mathbb{E}\left(K\big(\frac{x-x_i}{h}\big)\right)^2\right] - \frac{1}{n^2h^2}\sum_{i=1}^{n}\left[\mathbb{E}K\big(\frac{x-x_i}{h}\big)\right]^2\\ &= \frac{1}{nh^2}\int_{-\infty}^{\infty}\big[K\big(\frac{x-y}{h}\big)\big]^2 f(y)dy - \frac{1}{n}\left(\frac{1}{h}\int_{-\infty}^{\infty}K\big(\frac{x-y}{h}\big)f(y)dy\right)^2.\end{aligned}$$

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

$$\mathbb{V}f_h(x) = \frac{1}{nh}\int_{-\infty}^{\infty}\big(K(y)\big)^2 f(x-hy)dy - \frac{1}{n}\bigg(\int_{-\infty}^{\infty}K(y)f(x-hy)dy\bigg)^2. \tag{2.9}$$

En faisant le développement de *Taylor* à l'ordre 2 au point $y = 0$ de $f(x-hy)$.
On obtient :

$$f(x-hy) = f(x) - \frac{hy}{1}f'(x) + \frac{h^2y^2}{2!}f''(x) + o(h^2).$$

$$\begin{aligned}\mathbb{E}f_h(x) &= \int_{-\infty}^{\infty}K(y)[f(x) - huf'(x) + \frac{h^2y^2}{2!}f''(x)]dy + o(h^2)\\ &= f(x)\int_{-\infty}^{\infty}K(y)dy - hf'(x)\int_{-\infty}^{\infty}yK(y)dy + \frac{h^2}{2}f''(x)\int_{-\infty}^{\infty}y^2K(y)dy + o(h^2).\end{aligned}$$

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

$$\int_{-\infty}^{\infty}yK(y)dy = 0 \quad et \quad \int_{-\infty}^{\infty}y^2K(y)dy < \infty$$

Alors les expressions finales sont données par :

$$\mathbb{E}f_h(x) = f(x) + \frac{h^2}{2}f''(x)\mu_2(K) + o(h^2), \quad \mu_2(K) = \int_{-\infty}^{\infty}y^2K(y)dy. \tag{2.10}$$

$$biais\ f_h(x) = \mathbb{E}f_h(x) - f(x) = \frac{h^2}{2}f''(x)\mu_2(K), \quad \mu_2(K) = \int_{-\infty}^{\infty}y^2K(y)dy. \tag{2.11}$$

$$\mathbb{V}f_h(x) = \frac{f(x)}{nh}\int_{-\infty}^{\infty}K^2(y)dy - \frac{f'(x)}{n}\int_{-\infty}^{\infty}yK^2(y)dy - \frac{1}{n}\bigg(f(x) + biais f_h(x)\bigg)^2. \tag{2.12}$$

2.4.2 Comportement asymptotique du biais et de la variance

2.4.2.1 Comportement asymptotique du biais

Théorème 2.4.1 *(Parzen [77])*
Si on a :

1. $\lim\limits_{n\to+\infty} h(n) = 0 \quad et \quad \lim\limits_{y\to+\infty}|yK(y)| = 0$
2. $\sup\limits_{y}|K(y)| < \infty \quad et \quad \int_{-\infty}^{\infty}|K(y)|dy < \infty$
3. $\int_{-\infty}^{\infty}K(y)dy = 1$

Alors, l'estimateur $f_h(x)$ est asymptotiquement sans biais c'est -à-dire :

$$\lim_{n\longrightarrow\infty}\mathbb{E}f_h(x) = f(x).$$

pour tout point x pour lequel la densité f est continue.

2.4.2.2 Comportement asymptotique de la variance

Théorème 2.4.2 *(Parzen [77])*
Si on a

1. $\lim_{n\to+\infty} h(n) = 0 \quad et \quad \lim_{y\to+\infty} |yK(y)| = 0.$
2. $\sup_y |K(y)| < \infty \quad et \quad \int_{-\infty}^{\infty} |K(y)|dy < \infty.$
3. $\int_{-\infty}^{\infty} K(y)dy = 1.$

Alors

$$\lim_{n\longrightarrow\infty} nh\mathbb{V}f_h(x) = f(x)\int_{-\infty}^{\infty} K^2(y)dy.$$

pour tout point x pour lequel la densité f est continue.

2.4.3 Convergence en moyenne quadratique

En remplaçant les expressions finales des deux termes, le biais et la variance dans l'équation (2.1) on obtient :

$$MSE(f(x), f_h(x)) = \frac{f(x)}{nh}\int_{-\infty}^{\infty} K^2(y)dy + \frac{1}{4}h^4(f''(x))^2\left(\int_{-\infty}^{\infty} y^2K(y)dy\right)^2 + o(\frac{1}{nh} + h^2). \quad (2.13)$$

Théorème 2.4.3 *(Parzen [77])*
Si, $\lim_{n\to\infty} h(n) = 0 \quad et \quad \lim_{n\to\infty} nh(n) = \infty,$
et K satisfait aux conditions suivantes :

- $\sup_y |K(y)| < \infty \quad et \quad \lim_{y\longrightarrow\infty} |yK(y)| = 0,$
- $\int_{-\infty}^{\infty} |K(y)|dy < \infty \quad et \quad \int_{-\infty}^{\infty} K(y)dy = 1,$

Alors l'estimateur $f_h(x)$ est consistant en moyenne quadratique c'est-à-dire :

$$\lim_{n\longrightarrow\infty} MSE(f_h(x), f(x)) = 0,$$

pour tout point x pour lequel la densité f est continue.

2.4.4 Convergence en moyenne quadratique intégrée

Théorème 2.4.4 *(Parzen [77])*
Si K est un noyau de Parzen-Rosenblatt, c'est-à-dire K verifie :

- $\int_{\mathbb{R}} K(x)dx = 1.$
- $\int_{\mathbb{R}} \mid K(x) \mid dx < \infty.$
- $\sup_x \parallel K(x) \parallel dx < \infty.$
- $\lim_{|x|} K(x) = 0.$

$$\lim_{n\longrightarrow\infty} h(n) = 0, \quad \lim_{n\longrightarrow\infty} nh(n) = \infty \Longleftrightarrow (\forall f \in \mathbb{L}^p), \lim_{n\longrightarrow\infty} MISE(f_h, f) = 0.$$

On note par $\mathbb{L}^p$: l'ensemble des fonctions f définies sur $\mathbb{R}$, telles que $\int |f(x)|^p dx < \infty$.

2.4.5 Convergence uniforme

2.4.5.1 Convergence uniforme en probabilité

Théorème 2.4.5 *(Parzen [77])*

$$Si \lim_{n\longrightarrow\infty} nh(n)^2 = \infty,$$

si la fonction K satisfait aux conditions suivantes :

1. $\sup_y |K(y)| < \infty$ *et* $\lim_{y\longrightarrow+\infty} |yK(y)| = 0$,
2. $\int_{-\infty}^{\infty} |K(y)|dy < \infty$ *et* $\int_{-\infty}^{\infty} K(y)dy = 1$,

et si la transformée de Fourier $\tilde{K}(z) = \int_{-\infty}^{\infty} \exp(-izy)K(y)dy$ *est absolument intégrable, alors $f_h(x)$ est un estimateur uniformément consistant en probabilité c'est-à-dire :*

$$\forall\epsilon > 0, \quad P\left(\sup_{x\in\mathbb{R}} |f_h(x) - f(x)| < \epsilon\right) = 1.$$

2.4.5.2 Convergence uniforme presque complète

Théorème 2.4.6 *(Nadaraya [72])*
Si K est un noyau positif à variation bornée et f est uniformément continue,

$$si \lim_{n\longrightarrow\infty} h(n) = 0 \quad et \quad \sum_{n=1}^{\infty} \exp(-\gamma nh(n)^2) < \infty, \quad \forall\gamma > 0,$$

alors :

$$\sup_x |f_h(x) - f(x)| \longrightarrow 0 \quad avec\ une\ probabilité\ 1.$$

Silverman [90] a donné le même théorème sur la convergence presque complète en remplaçant la condition $\sum_{n=1}^{\infty} \exp(-\gamma nh^2) < \infty$, par les deux conditions suivantes :

$$\lim_{n\longrightarrow\infty} h(n) = 0 \ \text{ et } \ \lim_{n\longrightarrow\infty} \frac{\log n}{nh(n)} = 0.$$

Théorème 2.4.7 *(Silverman[90])*
Si on a :

$$\lim_{n\longrightarrow\infty} h(n) = 0 \quad et \quad \lim_{n\longrightarrow\infty} \frac{\log n}{nh(n)} = 0,$$

et K satisfait aux conditions suivantes :

- *K est uniformement continue et à variation bornée sur $\mathbb{R}$,*
- *Supposons aussi que f est uniformément continue,*
- $\int_{-\infty}^{\infty} |K(y)|dy < \infty$, $\int_{-\infty}^{\infty} \sqrt{|y\log|y||}\ |dK(y)| < \infty$,
- $\int_{-\infty}^{\infty} K(y)dy = 1$,

alors :

$$\lim_{n\longrightarrow\infty} \sup_x |f_h(x) - f(x)| = 0 \quad Presque\ Sûrement.$$

2.4.6 Convergence L_1 presque complète

Théorème 2.4.8 *(Devroye[25])*
Si :

$$\lim_{n\longrightarrow\infty} h(n) = 0, \quad \lim_{n\longrightarrow\infty} nh(n) = \infty,$$

alors

$$\forall (f \in \mathcal{F}), \lim_{n\longrightarrow\infty} \int |f_h(x) - f(x)|dx = 0, \ \textit{Presque Complètement},$$

où $\mathcal{F}$: Ensemble des densités de probabilité.

2.4.7 Comportement asymptotique

Théorème 2.4.9 *(Parzen[77])*
Si h satisfait : $\lim_{n\longrightarrow\infty} nh(n) = \infty$,
et si le noyau K satisfait aux conditions suivantes :

- $\int_{-\infty}^{\infty} |K(y)|dy < \infty$ *et* $\sup_{y\in\mathbb{R}} |K(y)| < \infty$,
- $\int_{-\infty}^{\infty} K(y)dy = 1$ *et* $\lim_{y\longrightarrow\infty} |yK(y)| = 0$,

alors $f_h(x)$ est un estimateur asymptotiquement normal c'est-à-dire :

$$f_h(x) \xrightarrow{cv.loi} \mathcal{N}(\mathbb{E}f_h(x), \mathbb{V}f_h(x)).$$

2.4.8 Vitesse de convergence

Wahba [102] a montré qu'on ne peut améliorer indéfiniment la convergence d'un estimateur f_h vers f, même pour la fonction la plus régulière possible (indéfiniment dérivable, bornée), mais bien sûr inconnue, c'est-à-dire : $MSE(f_h(x), f(x))$ ne peut tendre vers 0 qued'un ordre $\frac{e}{n}$, où e est une constante.
Précisément :

$$\sup_{f\in W(r,m,M)} \left(\left| f_h(x) - f(x) \right|^2 \right) \text{ ne peut tendre vers 0 que } \frac{1}{n^{G(m,r)}},$$

$$\text{où} \quad \frac{1}{n^{G(m,r)}} = \frac{2m - 2/r}{2m + 1 - 2/r} \text{ est une fonction croissante par rapport à } m \text{ et } r$$

$$\text{et } \lim_{m\longrightarrow\infty} G(m, r) = 1.$$

$f \in W(r, m, M)$ si :

a) les $(m-1)$ dérivées $f^{(i)}$ sont absolument continues ;
b) $\int \left| f^{(m)}(x) \right|^r dx < \infty$,
c) $\sup_x \left| f^{(m)}(x) \right| \leq M < \infty$,

2.4.8.1 Théorème sur la vitesse de convergence

Théorème 2.4.10 *(Wahba[102])*
Soit $m \geq 1$ un entier, soit $r \geq 1$ un nombre réel et $f \in \mathcal{W}(r, m, M)$.
Soit

$$f_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h} \right).$$

$$où \lim_{n\longrightarrow\infty} h(n) = 0, \quad \lim_{n\longrightarrow\infty} nh(n) = \infty.$$

et le noyau K satisfait aux conditions suivantes :

1. $\sup_y \big|yK(y)\big| < \infty$ *et* $\int_{-\infty}^{\infty} \big|K(y)\big| dy < \infty$,
2. $\int_{-\infty}^{\infty} K(y)dy = 1$ *et* $\lim_{y\longrightarrow\infty} \big|yK(y)\big| = 0$,
3. $\int_{-\infty}^{\infty} y^i K(y)dy = 0, i = 1, 2, ...m-1,$ *et* $\int_{-\infty}^{\infty} |y|^{\frac{m-1}{r}} \big|K(y)\big| < \infty$,

$$h = \frac{c_h}{n} \quad avec \quad c_h = \big(\frac{1}{2m-2/r}\frac{1}{M^2 A_2}\big)^{\frac{1}{2m+1-2/r}} n^{\frac{2m-2/r}{2m+1-2/r}}.$$

Alors :

$$MSE(f_h(x), f(x)) = \mathbb{E}\big(f_h(x) - f(x)\big)^2 \leq D_2 n^{-\frac{2m-2/r}{2m+1-2/r}}(1 + o(1)).$$

avec

$$D_2 = \theta\big(M^2 A_2 B_2^{2m-2/r}\big)^{\frac{1}{2m+1-2/r}}, \quad \theta = \frac{2m+1-2/r}{(2m-2/r)^{2m-2/r}}.$$

$$A_2 B_2^{(2m-2/r)} = \frac{1}{\Big((m-1)!(\frac{m-1}{1-1/r}+1)^{1-1/r}\Big)^2}\left(\int_{-\infty}^{\infty} \big|K(z)\big||z|^{m-1/r} dz\right)^2 \left(\omega \int_{-\infty}^{\infty} K^2(y)dy\right)^{2m-2/r}.$$

où

$$B_2 = \omega \int_{-\infty}^{\infty} K^2(y)dy, \quad f(u) \leq \omega \quad \forall u \in \mathbb{R}.$$

et

$$A_2 = \frac{1}{\big((m-1)!\big)^2\big((m-1)r+1\big)^{2/r}}\left(\int_{-\infty}^{\infty} \big|K(z)\big||z|^{m-1/r} dz\right)^2.$$

2.5 Choix du noyau

2.5.1 Noyau Uniforme (Rosenblatt)

Ce noyau a été proposé par Rosenblatt en 1956[80], l'avantage de ce noyau est la simplicité de sa forme. Il s'écrit sous la forme :

$$K(u) = \begin{cases} \frac{1}{2}, & \text{Si } |u| \leq 1\,; \\ \\ 0, & \text{Sinon.} \end{cases} \tag{2.14}$$

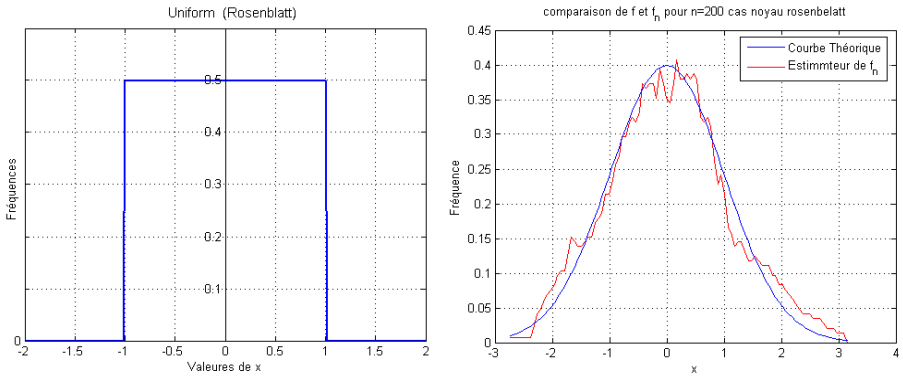


FIG. 2.1: *Noyau Uniform (Rosenblatt)*

2.5.2 Noyau Box(boite)

$$K(u) = \begin{cases} \frac{1}{2\sqrt{3}}, & \text{si } -\sqrt{3} \leq u \leq \sqrt{3}\,; \\ 0, & \text{Sinon.} \end{cases} \tag{2.15}$$

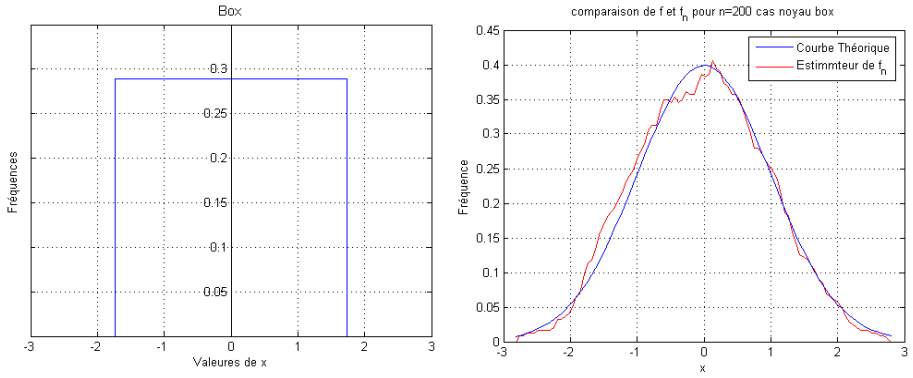


FIG. 2.2: *Noyau Box*

2.5.3 Noyau Triangulaire

Ce noyau a un avantage par rapport au noyau de Rosenblatt, il est continu partout, ce qui conduit à une estimation de f_h continue. Ce noyau s'écrit sous la forme :

$$K(u) = \begin{cases} (1 - |u|), & \text{Si } -1 \leq u \leq 1\,; \\ 0, & \text{Sinon.} \end{cases} \tag{2.16}$$

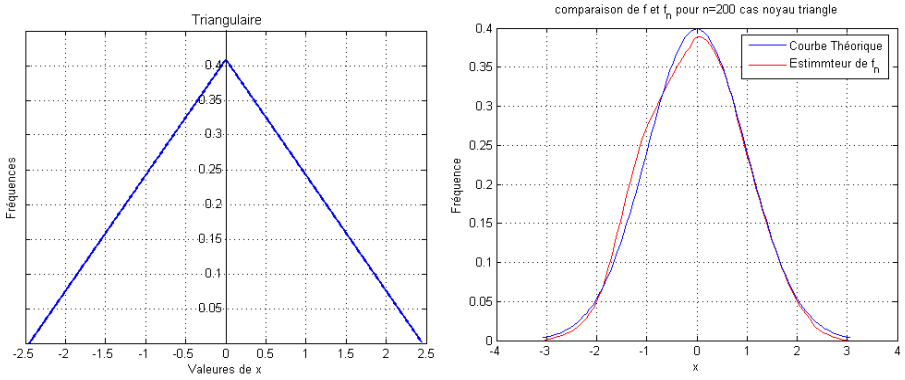


FIG. 2.3: *Noyau Triangulaire*

2.5.4 Noyau Cosine

$$K(u) = \begin{cases} \frac{\pi}{4}\cos(\frac{\pi u}{2}), & \text{Si} -1 \leq u \leq 1\,; \\ 0, & \text{Sinon.} \end{cases} \tag{2.17}$$

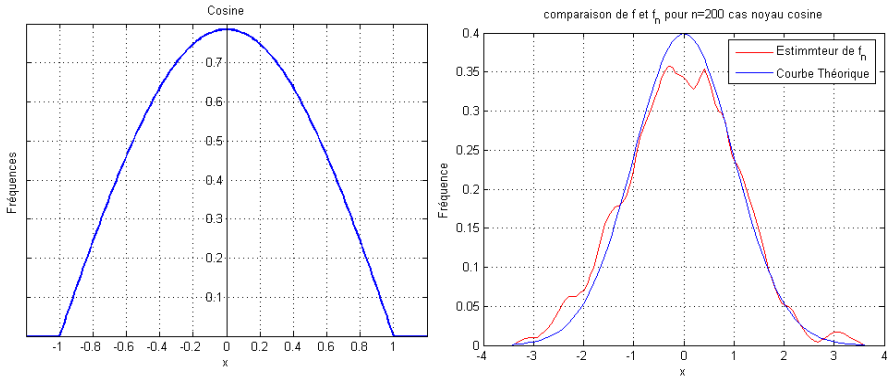


FIG. 2.4: *Noyau Cosine*

2.5.5 Noyau Gaussien

L'avantage du noyau gaussien est que plus la valeur de h est élevée plus on élargit la fenêtre, ce qui a un effet de lissage globale important ; mais le coût de calcul dans le cas de ce noyau est très élevé du fait de son support infini. Ce noyau s'écrit sous la forme :

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}; \qquad \forall u \in \mathbb{R}. \tag{2.18}$$

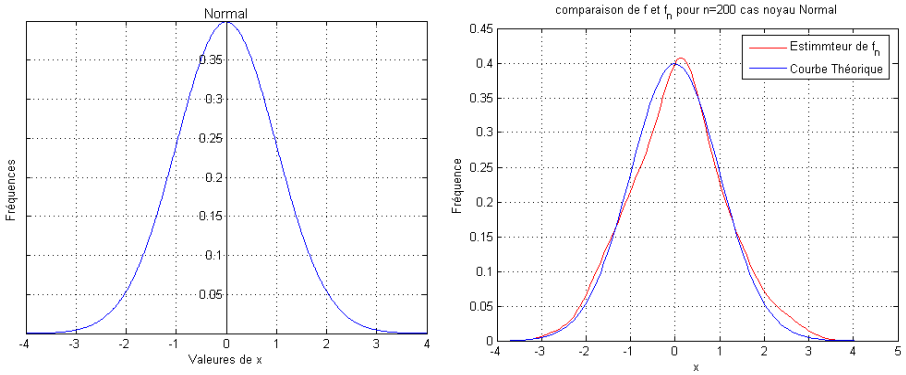


FIG. 2.5: *Noyau Normal (Gaussien)*

2.5.6 Noyau Biweight (Tukey)

Le noyau de *Tukey* ou *biweight* est, à notre sens le plus intéressant car donnant une estimation dérivable partout tout en étant simple à mettre en œuvre. En fait, il s'agit du noyau le plus simple parmi les noyaux de forme polynômial dérivable partout. Ainsi, il assure le lissage locale de la fonction f_h. Ce noyau est d'une forme très proche du noyau gaussien, il est donc préférable. il s'écrit sous la forme :

$$K(u) = \begin{cases} \frac{15}{16}(1-u^2)^2, & \text{Si } -1 \leq u \leq 1\,; \\ \\ 0 & \text{Sinon.} \end{cases} \tag{2.19}$$

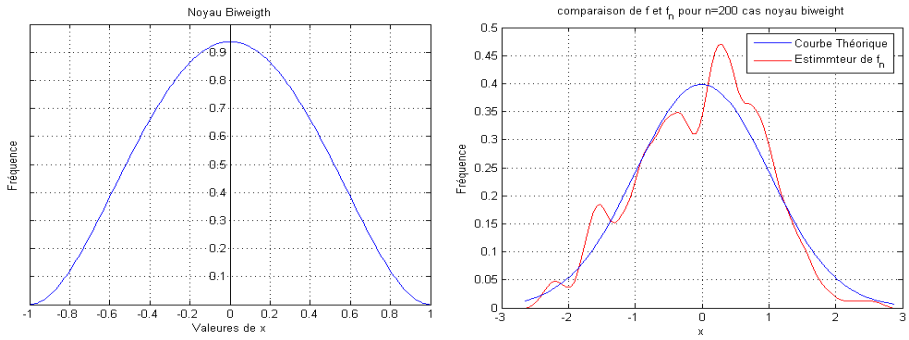


FIG. 2.6: *Noyau Biweight(Tukey)*

2.5.7 Noyau Triweight

Le noyau triweight s'écrit sous la forme :

$$K(u) = \begin{cases} \frac{35}{32}(1-u^2)^3, & \text{Si } -1 \leq u \leq 1\,; \\ \\ 0 & \text{Sinon.} \end{cases} \tag{2.20}$$

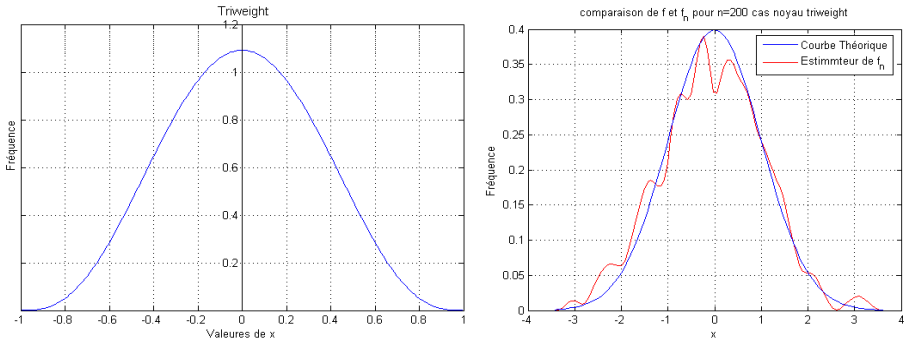


FIG. 2.7: *Noyau Triweight*

2.5.8 Noyau Epanechnikov

En 1969, Epanechnikov [33], a donné la forme du noyau K_E défini par :

$$K_E = \begin{cases} \frac{3}{4\sqrt{5}}\left(1 - \frac{x^2}{5}\right), & \text{Si } \; x \in \left[-\sqrt{5}, \sqrt{5}\right] ; \\ \\ 0 & \text{Sinon.} \end{cases} \tag{2.21}$$

qui minimise le MISE asymptotique.

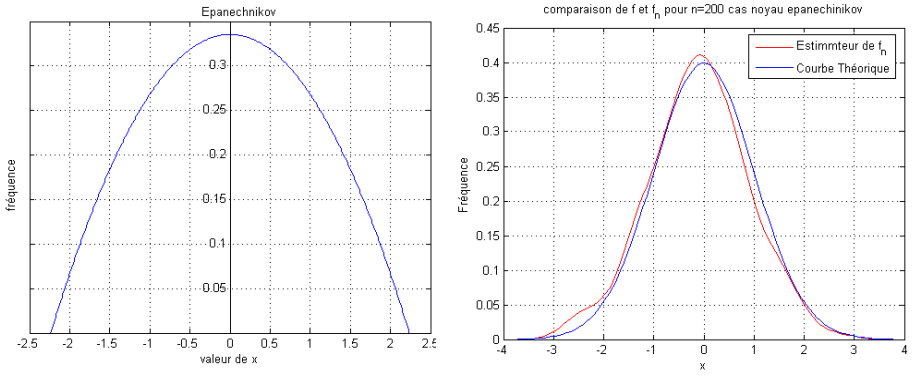


FIG. 2.8: *Noyau Epanechnikov*

On utilise le noyau gaussien en raison de sa simplicité et ses propriétés asymptotiques qui sont établies, pour les variables aléatoire $i.i.d$ ou des séries chronologiques, par plusieurs auteurs (Silverman (1986) [91], Pagan et Ullah (1999)[74] et Fan et Yao (2003) [34]. De plus, il est utilisé dans tous les packages dans la plupart des applications (logiciels). Cependant le crucial inconvénient de ce noyau est qu'il attribue des poids positif pour des coordonnées (valeurs de x) qui sont à l'extérieur du support lorsque on cherche l'estimateur de la densité de probabilité d'une variable aléatoire bornée ou semi bornée (cas de la loi exponentielle). Cela cause le problème du biais aux bornes et donne un estimateur non consistant.

Ce problème, connu sous le problème de biais aux bornes à donné un essor pour des nouvelles méthodes et de nouveaux noyaux pour l'estimation d'une densité de probabilité par la méthode du noyau pour le cas des données $i.i.d$. Parmi ces méthodes on peut citer :

Les méthodes de réflexion (noyau miroir) de Schuster(1985)[84] et la méthode de rénormalisation local de Diggle 1985[53] et Härdle(1990)[27]. D'autres auteurs ont proposé d'utiliser les noyaux adaptés aux bornes et les noyaux standard (classique) à l'intérieur du support.

2.5.9 Noyau Miroir (Schuster)

L'idée de cette méthode, développée par Deheuvels et Hominal (1979)[24] et Schuster (1985)[84], est d' ajouter une "masse manquante" par réflexion de l'échantillon et qui concerne les données aux frontières. Elles se focalisent sur le cas où les variables sont positives, c'est-à-dire, dont le support est $[0, \infty[$. Formellement et sous sa forme plus simple, il consiste à remplacer $K\left(\frac{x-X_i}{h}\right)$ par $K\left(\frac{x-X_i}{h}\right) + K\left(\frac{x+X_i}{h}\right)$. L'estimateur de la densité est alors de la forme :

$$f_h = \frac{1}{nh}\sum_{i=1}^{n}\left[K\left(\frac{x-X_i}{h}\right) + K\left(\frac{x+X_i}{h}\right)\right]. \tag{2.22}$$

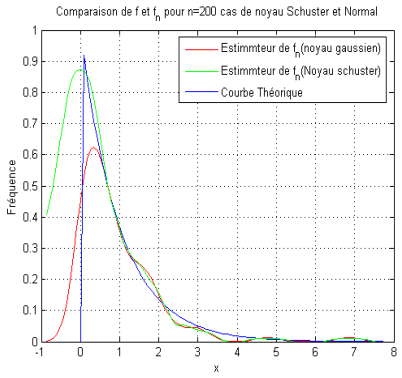


FIG. 2.9: *Noyau Miroir (Schuster)*

Dans le cas des densités dont le support est $[0, 1]$, la non consistance peut être corrigé aux bornes. Mais le taux de convergence du biais reste très faible, il est d'ordre $0(h)$ aux bornes qui est grand au taux usuel qui est d'ordre $0(h^2)$.

2.5.10 Les noyaux Gamma et Beta

Quoique les méthodes précédentes diminuent le biais, aux bornes, elles restent peu efficaces car le biais reste considérable si on le compare aux biais de l'intérieur du support. Pour obtenir un biais aux bornes de même ordre que celui de l'intérieur, Devroy et Gyorfi (1985)[26] et Marron et Ruppert (1994)[69], ont proposé d'appliquer une transformation sur les données originales de telle façon que la dérivée d'ordre 1 de la densité des variables transformées soit égale à zéro et ensuite utiliser la méthode de réflexion pour estimer la densité des données transformées. L'objectif étant de trouver cette fois un biais du même ordre mais sans transformation des données. Plusieurs autres auteurs ont proposé d'utiliser les noyaux adaptés dans la région des bornes et le noyau standard à l'intérieur du support (voir Jones (1993) [57]). Pour l'estimation à noyau aux bornes, Müller (1991) [70] pour l'estimateur à noyau optimal aux bornes et Lejeune et Sarda (1992) [62] pour l'estimation linéaire local.

L'inconvénient de ces estimateurs est qu'ils attribuent des poids négatifs aux valeurs du voisinage des bornes.

La solution la plus récente est d'utiliser des noyaux asymétriques et adaptés qui n'assignent aucun poids à l'extérieur du support. Chen(1999)[13] et Chen(2000)[15] propose respectivement le noyau *Beta* pour les densités à support compact et le noyau gamma pour les densités à variables à support positif (c'est-à-dire sur $[0, +\infty[$).

2.5.10.1 Noyau Beta

Dans cette section on présente le noyau bêta proposé par Brown et Chen (1999)[10], et Chen (1999,2000)[13, 14] pour l'estimation non paramétrique de la courbe de régression et des densités unidimensionnelles defini sur un support compact.

L'idée de Harrell et Davis (1982)[50], Chen (1999,2000)[13, 14] est d'utiliser le noyau bêta pour estimer la densité de probabilité à support compact [0, 1] et ainsi de régler le problème du biais aux bornes. L'estimateur de la densité sera alors de la forme :

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K(X_i, \frac{x}{h}+1, \frac{(1-x)}{h}+1), \tag{2.23}$$

où $K(., \alpha, \beta)$ représente la densité de la distribution *Beta* de paramètres α et β,

$$K(x, \alpha, \beta) = \frac{x^{\alpha}(1-x)^{\beta}}{\mathcal{B}(\alpha, \beta)}, \quad x \in [0, 1],$$

avec,

$$\mathcal{B}(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}.$$

Le noyau bêta a deux avantages, premièrement il peut parfaitement estimer les densités à support compact et deuxièmement il possède une forme flexible qui change le lissage dans le sens naturel quand on s'éloigne des bornes. Par conséquent, le noyau bêta élimine le biais des bornes et fourni une réduction de la variance (voir figure (FIG.2.10)). Charpentier, Fermanian et Scaillet [37] ont montré par simulation que l'estimateur à noyau bêta est plus performant quand on le compare à d'autre estimateurs avec des noyaux standards.

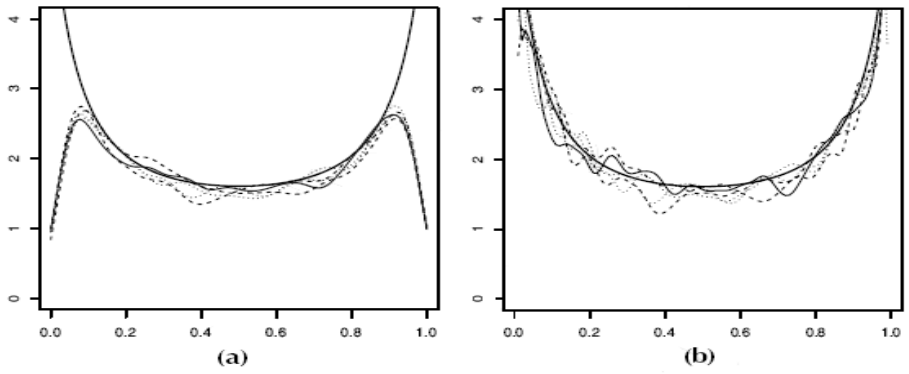


FIG. 2.10: *Estimation d'une densité de probabilité à support compact* $[0, 1]$ *pour* $n = 10000$ *:(a) quand on utilise le noyau standard (gaussien) , (b) quand on utilise le noyau bêta.*

2.5.10.2 Noyau Gamma

On observe $X_1, X_2, ..., X_n$ à partir d'une densité f inconnue. L'objectif est d'estimer la fonction $f(x)$ (par la méthode de noyau) pour $x \in [0, +\infty[$. L'estimateur à noyau gamma est défini comme suit :

$$f_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_{(\frac{x}{h}+1,h)}(X_i).$$

La première forme du noyau gamma est définie comme suit (voir Bouezmarni et Scaillet [7]) :

$$K_{(\frac{x}{h}+1,h)}(t) = \frac{t^{(x/h)}e^{-(t/h)}}{h^{(x/h)+1}\Gamma((x/h)+1)}. \tag{2.24}$$

Ce dernier noyau reste inefficace au voisinage de zéro (voir figure 2.12 à gauche), pour cela une autre version du noyau gamma avait été proposer [7]. La forme de ce nouveau noyau est définie comme suit :

$$K_{(\rho_h(x),h)}(t) = \frac{t^{\rho_h(x)-1}e^{-t/h}}{h^{\rho_h(x)}\Gamma(\rho_h(x))}; \tag{2.25}$$

avec

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

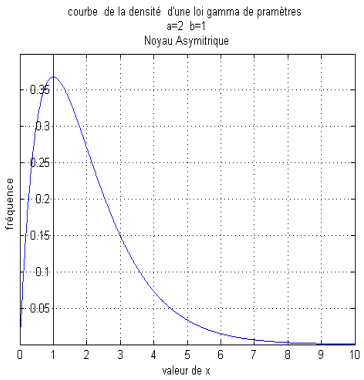


FIG. 2.11: *Courbe de la densité gamma*

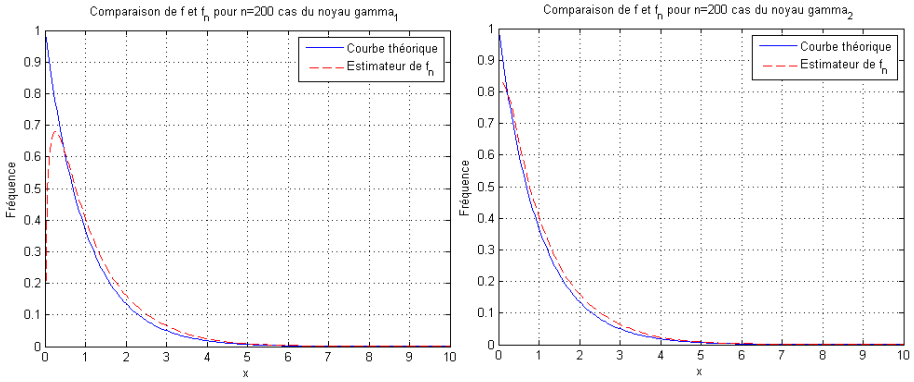


FIG. 2.12: *Noyau Gamma*

On constate que la forme du noyau gamma defini dans (2.25) et la qualité du lissage changent selon la position où la densité est estimée (voir figure 2.12 à droite), ce qui implique que l'estimateur du noyau gamma est un estimateur adaptatif de la densité. C'est la différence avec le noyau gaussien, ou tout autre noyau symétrique. Le support du noyau gamma est égal au support de la densité à estimer, ainsi on ne perd pas de poids lorsque on estime la densité aux voisinage des bornes. Le noyau gamma est facile à implémenter, il élimine le biais des bornes et souvent non négative. Il atteint le taux de convergence optimale pour les variables $i.i.d$ au sens de MISE dans la classe des estimateurs à noyaux non négatifs. De plus, il permet une réduction de la variance lors du lissage en s'éloignant des bornes. Particulièrement lorsque on utilise la normalité asymptotique de l'estimateur à noyau gamma.

Bouezmarni et Scaillet (2003)[7] ont donné les conditions de convergence faible de l'estimateur à noyau gamma sur un compact $[0, +\infty[$ lorsque f est continue sur ce support et la

convergence faible au sens MIAE (Mean Integer Absolute Error). Pour les densités non bornées à l'origine c'est-à-dire au voisinage de zéro, ils ont examiné les performances de cet estimateur par simulation et ils ont prouvé la convergence en probabilité vers l'infini à l'origine.

Fermandez et Monteira (2005)[38] ont établi le théorème centrale limite pour l'estimateur fonctionnelle pour le noyau gamma. Bouezmarni et Ronbouts (2006)[6] ont démontré la convergence presque sûre au sens du MISE et la normalité asymptotique de cet estimateur.

2.6 Choix du paramètre de lissage

D'après la formule (2.6) on constate que l'estimateur $f_h(x)$ de $f(x)$ ne dépend pas seulement du noyau K mais aussi du paramètre h, appelé paramètre de lissage ou fenêtre (bandwidth or window). Une petite perturbation de ce dernier est suffisante pour que $f_h(x)$ change complètement ses caractéristiques (performances numériques ou graphiques), ce qui signifie $f_h(x)$ est fortement lié à ce paramètre. C'est pour cette raison que plusieurs travaux ont été consacrés au choix de ce paramètre.
Il existe plusieurs méthodes de sélection de ce paramètre que l'on peut regrouper en deux familles :

- Méthodes de plug-in (re-injection)
- Méthodes de Cross-Validation (Validation-croisée).

La multitude de ces méthodes et leurs diversités du point de vue principe, sont dues au fait que ces méthodes restent incomplètes ou autrement dit, ces méthodes ont toujours des inconvénients [109], soit au sens de la qualité de l'estimateur f_h par rapport à une norme d'erreur bien déterminée, soit par l'allure graphique de la courbe (lissée ou non).

Dans cette section on va présenter quelques méthodes de sélection.

2.6.1 Méthodes plug-in(re-injection)

2.6.1.1 Choix optimal

La décision d'un choix optimal pour le paramètre de lissage suppose la spécification d'un critère d'erreur qui puisse être optimisé. L'optimalité n'est pas un concept absolu : elle est intimement liée aux choix du critère, qui peut faire intervenir à la fois la densité inconnue f et l'estimateur f_h(donc h et le noyau K).
Dans ce cas, on cherche à minimiser l'Erreur Quadratique Intégrée Moyenne ($MISE$).

$$MISE(f, f_h) = \int \mathbb{E}\big[f_h(x) - f(x)\big]^2 dx.$$

Théorème 2.6.1 *(Scott[85])*
Si f a une dérivée seconde absolument continue, si $f^{(3)} \in \mathbb{L}^2$ et si le noyau $K \in \mathbb{L}^2$ est une densité de probabilité continue, symétrique de variance $\sigma_K^2 > 0$, alors, sous les conditions $h(n) \longrightarrow 0$ et $nh(n) \longrightarrow \infty$, on a le développement asymptotique :

$$MISE = \frac{h^4}{4}\sigma_K^4 \int (f'')^2 \ + \ \frac{\int K^2}{nh} \ + \ O(h^5 + \frac{1}{n}). \tag{2.26}$$

où $\mathbb{L}^2$: l'ensemble des fonctions f définies sur $\mathbb{R}$, telles que $\int |f(x)|^2 dx < \infty$.

L'Erreur Quadratique Intégrée Moyenne Asymptotique est alors de la forme :

$$AMISE = \frac{h^4}{4}\sigma_K^4 R(f'') + \frac{R(K)}{nh}. \tag{2.27}$$

où

$$R(g) = \int g^2(x)dx, \quad \text{Pour toute fonction } g.$$

On remarque que le premier terme du membre à droite du développement (2.27) est un terme de biais, alors que le second est un terme de variance. On constate que dans l'$AMISE$, le terme de biais est fonction croissante en h alors que le terme de la variance est une fonction décroissante en h c'est à dire les deux termes varient en sens inverse par rapport à h : une largeur de fenêtre h trop importante entraînera une augmentation du biais et une diminution de la variance (phénomène de surlissage), alors qu'une largeur de fenêtre trop petite provoquera une augmentation de la variance et une diminution du biais (phénomène de sous-lissage)(voir figure 2.13).

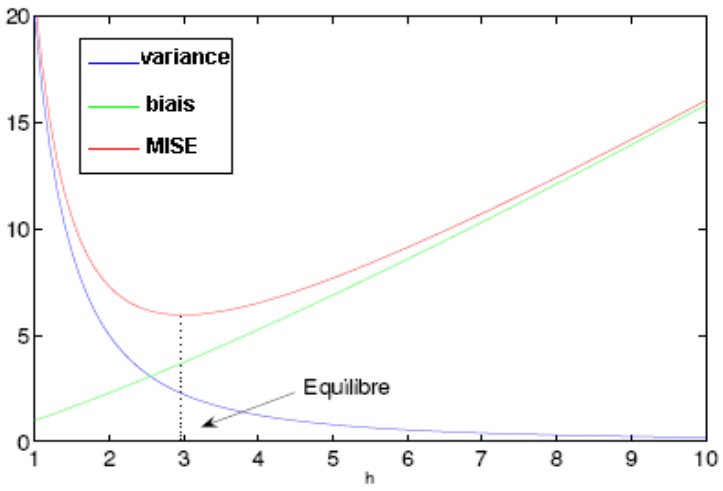


FIG. 2.13: *Illustration du principe de l'équilibre biais-variance à l'aide d'une simulation.*

Pour obtenir le paramètre de lissage h^* qui minimise l'Erreur Quadratique Intégrée Moyenne Asymptotique : il suffit de résoudre le système suivant :

$$\begin{cases} \frac{dAMISE}{dh} = 0 \\ \text{et} \\ \frac{d^2AMISE}{dh^2} > 0 \end{cases} \tag{2.28}$$

A partir de l'expression (2.27) on a :

$$\frac{dAMISE}{dh} = h^3\sigma_K^4 R(f'') - \frac{1}{nh^2}R(K) = 0$$

$$nh^5\sigma_K^4 R(f'') - R(K) = 0 \Rightarrow h^5 = \frac{R(K)}{n\sigma_K^4 R(f'')}$$

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

$$\frac{d^2AMISE}{dh^2} = 3h^2\sigma_K^4 R(f'') + \frac{1}{nh^3}R(K) > 0 \Rightarrow h^* \text{ minimise } AMISE$$

h^* peut aussi s'écrire :

$$h^* = \psi(K)\varphi(f)n^{-1/5}, \tag{2.29}$$

où

$$\psi(K) = \left[\frac{R(K)}{\sigma_K^4}\right]^{1/5} \text{ et } \varphi(f) = \left[\frac{1}{R(f'')}\right]^{1/5} \text{ avec } R(f'') \neq 0.$$

Notons que h^* est une quantité déterministe qui dépend du nombre d'observations n.

La valeur du $AMISE$ optimale $AMISE^* = AMISE(h^*)$ est donnée par :

$$AMISE^* = \frac{5}{4}\left[\sigma_K R^4(K)R(f'')\right]^{1/5} n^{-4/5}. \tag{2.30}$$

Le paramètre de lissage h^* optimal au sens du critère de l'Erreur Quadratique Intégrée Moyenne Asymptotique, devra réaliser un compromis entre les valeurs de la variance et celle du biais. Outre sa nature asymptotique, la largeur de fenêtre optimale h^* dépend de la densité inconnue f au travers du paramètre $R(f'')$. Cette largeur de fenêtre "idéale" (relativement au critère d'erreur retenu) n'est donc pas directement calculable. Une façon classique de remédier à ce dernier problème consiste à remplacer la quantité $R(f'')$ par un estimateur approprié.

2.6.1.2 Estimateur Rule Of Thumb

Si on choisit f comme étant la distribution de loi normale de moyenne 0 et de variance σ^2 on aura alors :

$$R(f'') = \int (f''(x))^2 dx = \frac{3}{8}\sqrt{\pi}\sigma^{-5}. \tag{2.31}$$

De plus, si K est un noyau gaussien, alors la valeur pour le h^* notée dans ce cas par h_{rot} est obtenue en substituant ce noyau et la valeur $R(f'')$ obtenue dans (2.31) dans la formule (2.29)

$$\begin{aligned} h_{rot} &= (4\pi)^{-1/10}\big[\frac{3}{8}\pi^{-1/2}\sigma\big]n^{-1/5} && (2.32)\\ &= (\frac{4}{3})^{1/5}\sigma n^{-1/5} && (2.33)\\ &= 1.06\sigma n^{-1/5}. && (2.34) \end{aligned}$$

Il suffit donc d'estimer σ à partir des données et de substituer cet estimateur dans la formule ci-dessus. D'après Silverman [91] en (1986), cette formule donnera de bons résultats si la population est réellement normalement distribuée. Mais celle ci peut donner une distribution trop lissée si la population est plutôt multimodale. Dans ce cas, de meilleurs résultats peuvent être obtenus si l'on utilise une mesure plus robuste de l'étendue. Au lieu d'utiliser σ on utilise plutôt l'interquartile IQ où

$$IQ = \frac{X_{(3n/4)} - X_{(n/4)}}{1.34}.$$

où $X_{(n/4)}$ et $X_{(3n/4)}$ sont respectivement le premier et le troisième quartile.
Dans le cas où X suit une loi normale l'écart interquartile est $IQ = 1.394$.
h_{rot} de l'équation (2.32) devient alors

$$h_{rot} = 1.06 IQ n^{-1/5}.$$

Cette formule peut aussi donner une distribution trop lissée si la vraie densité est multimodale et parfois cette dernière donne des résultats moins bons que si l'on avait utilisé l'écart type, d'ou le meilleur des deux méthodes peut être obtenu en utilisant un estimateur adaptatif de l'étendue. C'est à dire, en utilisant A au lieu de σ dans la formule (2.32) où A est défini par $A = \min(\sigma, IQ)$, donc la formule pour h_{rot} devient alors :

$$h_{rot} = 1.06 A n^{-1/5}. \tag{2.35}$$

Cette correction est insuffisante dans de nombreux cas. Par exemple si la vraie densité est multimodale.

2.6.1.3 Estimateur de Sheather et Jones

En 1991, Sheather et Jones [89] recommandent l'utilisation de l'estimateur naturel $\hat{R}_a(f'')$, en faisant observer que le terme de biais $\frac{R(K'')}{na^5}$ est positif et peut donc servir à annuler le terme de biais (négatif) de l'erreur quadratique moyenne entre $\hat{R}_a(f'')$ et $R(f'')$. Afin de faire disparaître quelques effets indésirables du terme du biais. Les deux auteurs sont contraints de mettre en place une procédure de type plug-in en trois étapes.

Sheather et Jones choisissent d'estimer $R(f'') = \int_{-\infty}^{+\infty} (f''(x))^2 dx$ par :

$$S(a) = \frac{1}{n(n-1)a^5} \sum_{i=1}^{n} \sum_{j=1}^{n} L^{(4)}\left(\frac{x_i - x_j}{a}\right), \tag{2.36}$$

où $L^{(4)}$ désigne la dérivée quatrième du noyau suffisamment lisse L et où a est un nouveau paramètre de lissage appelé paramètre pilote. Les deux auteurs choisissent le noyau gaussien :

$$K(u) = L(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{u^2}{2}).$$

L'estimateur $S(a)$ est obtenu en écrivant $\hat{R}_a(f'') = R(f_a'')$ et en remarquant que, sous des conditions de régularité suffisantes suivantes :

- **H1** Le noyau K est quatre fois différentiable et ses dérivées sont lipschitziennes.
- **H2** f admet une dérivée seconde absolument continue.
- **H3** f admet une dérivée quatrième lipschitzienne.
- **H4** f admet un degré de lissage d'ordre (l, η).
- **H5** Il existe deux constantes $0 < \underline{B} < \overline{B}$ telles que :

$$h \in [\underline{B} n^{-1/5}, \overline{B} n^{-1/5}].$$

$$\int_{\mathbb{R}} f''^2(x) dx = \int_{\mathbb{R}} f^{(4)}(x) f(x) dx = \mathbb{E}\big(f^{(4)}(x)\big).$$

$R(f'')$ peut donc être estimer par :

$$S(a) = R(f_a'') = \frac{1}{n} \sum_{i=1}^{n} f_a^{(4)}(x_i).$$

En utilisant la dérivée quatrième du noyau L. $S(a)$ peut s'écrire comme suit :

$$S(a) = \frac{1}{n(n-1)a^5} \sum_{i=1}^{n} \sum_{j=1}^{n} L^{(4)}\left(\frac{x_i - x_j}{a}\right).$$

Ainsi le paramètre de lissage optimal est estimé par :

$$\hat{h} = \left(\frac{R(K)}{\sigma_K^4 S(a)}\right)^{1/5} n^{-1/5}, \tag{2.37}$$

avec

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

Sheather et Jones[89] affirment alors, que le paramètre de lissage $\hat{a}$ qui minimise L'erreur Quadratique Moyenne $\mathbb{E}[S(a) - R(f'')]^2$ admet la représentation asymptotique suivante :

$$\hat{a} = \left[\frac{-2K^{(4)}(0)}{\sigma_K^2 n \int_{\mathbb{R}} f^{(6)}(x) f(x) dx}\right]^{1/7}. \tag{2.38}$$

En combinant l'équation (2.38) avec l'équation (2.37), on peut fournir une expression reliant les deux largeurs de fenêtre $\hat{a}$ et $\hat{h}$.
Soit $\hat{\lambda}$ un estimateur de λ qui représente une mesure d'échelle de f (par exemple son écart interquartile). Il vient ainsi :

$$\hat{\alpha}(h) = 1.357\left(\frac{S(a)}{T(b)}\right)^{1/7} h^{5/7}; \tag{2.39}$$

avec

$$T(b) = \frac{1}{n(n-1)b^7} \sum_{i=1}^{n} \sum_{j=1}^{n} L^{(6)}\left(\frac{x_i - x_j}{b}\right),$$

est l'estimateur de $\int_{\mathbb{R}} f^{(6)}(x) f(x) dx$.

et

$$a = 0.920\hat{\lambda} n^{-1/7} \quad , \quad b = 0.912\hat{\lambda} n^{-1/9}.$$

Le paramètre de lissage final est solution de l'équation :

$$\hat{h} = \left(\frac{R(K)}{\sigma_K^4 S(\hat{\alpha}(\hat{h}))}\right)^{1/5} n^{-1/5}. \tag{2.40}$$

Le critère $SJ(h)$ est définie par cette formule :

$$SJ(\hat{h}) = \left(\frac{1}{2\sqrt{\pi}}\right)^{1/5} S(\hat{\alpha}(\hat{h}))\left(f''\right)^{-1/5} n^{-1/5} - \hat{h}. \tag{2.41}$$

Définition 2.8 *On dit que f a un degré de lissage d'ordre (l, η)pour un nombre entier l et un $\eta \in\ ]0, 1]$ si f est au moins deux fois dérivable et si il existe une constante réelle $Q > 0$ telle que*

$$|f^{(2+l)}(x) - f^{(2+l)}(y)| \leq Q|x - y|^{\eta}, \quad \forall \quad x, y.$$

2.6.2 Méthodes Cross-Validation (Validation Croisée)

2.6.2.1 Validation croisée non biaisée

Cette méthode appelée Validation Croisée non Biaisée a été proposée par Rudemo [81] en 1982 et Bowman [9] en 1984. Le critère consiste à choisir le paramètre de lissage qui minimise un estimateur convenable de :

$$UCV(h) = \int_{\mathbb{R}} \big[f_h(x) - f(x)\big]^2 dx - \int_{\mathbb{R}} f^2(x)dx = \int_{\mathbb{R}} f_h^2(x)dx - 2\int_{\mathbb{R}} f_h(x)f(x)dx.$$

Puisque $\int_{\mathbb{R}} f^2(x)dx$ ne dépend pas du paramètre de lissage h. On peut choisir le paramètre de lissage de façon à ce qu'il minimise un estimateur de :

$$\int_{\mathbb{R}} f_h^2(x)dx - 2\int_{\mathbb{R}} f_h(x)f(x)dx.$$

On veut premièrement trouver un estimateur de $\int_{\mathbb{R}} f_h(x)f(x)dx$. Remarquons que

$$\int_{\mathbb{R}} f_h(x)f(x)dx = \mathbb{E}(f_h(x)).$$

L'estimateur empirique de $\int_{\mathbb{R}} f_h(x)f(x)dx$, est alors $\frac{1}{n}\sum_{i=1}^{n} f_{h,i}(x_i)$.
Le critère à optimiser est alors :

$$UCV(h) = \int_{\mathbb{R}} f_h^2(x)dx - \frac{2}{n}\sum_{i=1}^{n} f_{h,i}(x_i). \tag{2.42}$$

où $f_{h,i}(x_i) = \frac{1}{(n-1)h}\sum_{j=1,j\neq i}^{n} K\big(\frac{x_i - x_j}{h}\big)$ est l'estimateur de la densité construit à partir de l'ensemble de points sauf le point x_i.

Montrons maintenant que $UCV(h)$ est un estimateur *sans biais* de $MISE(h) - R(f)$.
On a :

$$MISE(h) - R(f) = \mathbb{E}\int \big(f(x) - f_h(x)\big)^2 dx - R(f) = \mathbb{E}\bigg[\int f_h^2(x)dx - 2\int f_h(x)f(x)dx\bigg].$$

Il suffit de montrer que $\int f_h^2 dx$ et $\frac{1}{n}\sum_{i=1}^{n} f_{h,i}(x_i)$ sont des estimateurs *sans biais* de $\mathbb{E}\big[\int f_h^2(x)dx\big]$ et $\mathbb{E}\big[\int f_h(x)f(x)dx\big]$ respectivement. Or $\mathbb{E}\big[\int f_h^2(x)dx\big]$ admet l'estimateur sans biais trivial $\int f_h^2(x)dx$. Il reste donc à montrer que $\frac{1}{n}\sum_{i=1}^{n} f_{h,i}(x_i)$ est un estimateur *sans biais* de $\mathbb{E}\big[\int f_h(x)f(x)dx\big]$.

On a d'une part,

$$\begin{aligned}
\mathbb{E}\big[\frac{1}{n}\sum_{i=1}^{n} f_{h,i}(x_i)\big] &= \mathbb{E}\big[\frac{1}{n(n-1)h}\sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n} K\big(\frac{x_i - x_j}{h}\big)\big] \\
&= \mathbb{E}\big[\frac{1}{(n-1)h}\sum_{j\neq 1}^{n}\int K\big(\frac{x - x_j}{h}\big)f(x)dx\big] \\
&= \frac{1}{h}\int f(z)\int K\big(\frac{x-z}{h}\big)f(z)dxdz.
\end{aligned}$$

D'autre part,

$$\begin{aligned}\mathbb{E}\big[\int f_h(x)f(x)\big] &= \mathbb{E}\big[\frac{1}{nh}\sum_{i=1}^{n}\int K\big(\frac{x-x_i}{h}\big)f(x)dx\big] \\ &= \frac{1}{h}\int f(z)\int K\big(\frac{x-z}{h}\big)f(x)dxdz.\end{aligned}$$

Ce qui implique que

$$\mathbb{E}\big[\int f_h(x)f(x)dx\big] = \mathbb{E}\big[\frac{1}{n}\sum_{i=1}^{n} f_{h,i}(x_i)\big].$$

Finalement, un estimateur *sans biais* de $MISE(h) - R(f)$ est donnée donc par $UCV(h)$.

En utilisant l'équation (2.42), le critère $UCV(h)$ devient :

$$UCV(h) = \frac{R(K)}{nh} + \sum_{i=1}^{n}\sum_{i\neq j,j=1}^{n}\left[\int \frac{1}{n^2h^2}K\big(\frac{x-x_i}{h}\big)K\big(\frac{x-x_j}{h}\big)dx - \frac{2}{n(n-1)h}K\big(\frac{x_i-x_j}{h}\big)\right]. \tag{2.43}$$

Nous noterons h_{ucv} l'estimateur de h qui minimise $UCV(h)$

La popularité de cette méthode est due à la motivation intuitive et au fait que cet estimateur est asymptotiquement optimal sous de faibles conditions. L'optimalité asymptotique de la validation croisée non biaisée a été obtenue par Stone [96].

Proposition 2.6.1 *Soit $X_1, X_2, ..., X_n$ un n-échantillon i.i.d issu d'une variable aléatoire X de fonction de densité f. Utilisant le noyau gaussien on obtient :*

$$\begin{aligned}UCV(h) &= \frac{1}{2n^2h\sqrt{\pi}}\left(n + 2\sum_{i=1}^{n}\sum_{i\neq j,j=1}^{n}\exp\left(-\big(\frac{x_i-x_j}{2h}\big)^2\right)\right) \\ &- \frac{2}{\sqrt{2}\pi n(n-1)h}\sum_{i=1}^{n}\sum_{i\neq j,j=1}^{n}\exp\left(-\frac{(x_i-x_j)^2}{2h^2}\right).\end{aligned}$$

Démonstration

$$\int_{\mathbb{R}} f_h^2(x)dx = \frac{1}{n^2h^2}\left(\sum_{i=1}^{n}\int K^2\big(\frac{x-x_i}{h}\big)dx + 2\sum_{i\neq j}^{n}\int K\big(\frac{x-x_i}{h}\big)K\big(\frac{x-x_j}{h}\big)\right).$$

Utilisant le noyau Gaussien on obtient :

$$\int_{\mathbb{R}} f_h^2(x)dx = \frac{1}{2\pi n^2h^2}\left(\sum_{i=1}^{n}\int \exp\left(-2\big(\frac{x-x_i}{\sqrt{2}h}\big)^2\right)dx + 2\sum_{i\neq j}^{n}\int \exp\left(-\big(\frac{x-x_i}{\sqrt{2}h}\big)^2\right)\exp\left(-\big(\frac{x-x_j}{\sqrt{2}h}\big)^2\right)\right)$$

$$I_1 = \int_{-\infty}^{+\infty}\exp\left(-2\big(\frac{x-x_i}{\sqrt{2}h}\big)^2\right)dx = \sqrt{\pi}h;$$

et

$$I_2 = \int_{-\infty}^{+\infty}\exp\left(-\big(\frac{x-x_i}{\sqrt{2}h}\big)^2\right)\exp\left(-\big(\frac{x-x_j}{\sqrt{2}h}\big)^2\right) = \sqrt{\pi}h, \quad \forall i \neq j;$$

donc

$$\int_{\mathbb{R}} f_h^2(x)dx = \frac{1}{2n^2h\sqrt{\pi}}\Big(n + 2\sum_{i\neq j}^{n}\exp\Big(-\big(\frac{x_i - x_j}{2h}\big)^2\Big)\Big).$$

Finalement

$$\begin{aligned} UCV(h) &= \frac{1}{2n^2h\sqrt{\pi}}\Big(n + 2\sum_{i=1}^{n}\sum_{i\neq j,j=1}^{n}\exp\Big(-\big(\frac{x_i - x_j}{2h}\big)^2\Big)\Big) \\ &- \frac{2}{\sqrt{2}\pi n(n-1)h}\sum_{i=1}^{n}\sum_{i\neq j,j=1}^{n}\exp\Big(-\frac{(x_i - x_j)^2}{2h^2}\Big). \end{aligned}$$

Cette méthode présente deux problèmes majeurs (ou points faibles) : d'une part son manque de robustesse par rapport aux changements de taille de l'échantillon c'est-à-dire le résultat de simulation peut se révéler extrêmement variable d'un échantillon à l'autre, d'autre part, la fonctionnelle à minimiser a souvent tendance à présenter plusieurs minimums locaux [49]. Pour d'autres études, voir Hall [43], Burman [11], Scott et Terrell[87].

Algorithme de la méthode

Pour trouver le paramètre de lissage optimal, noté h_{ucv} par validation croisée non biaisée, on minimise numériquement $UCV(h)$. Par exemple, on peut faire une recherche exhaustive du minimum $UCV(h)$ sur un grand nombre de valeurs possibles du paramètre de lissage, ou il suffit de donner un algorithme afin de calculer le paramètre de lissage optimal. En utilisant le noyau gaussien les étapes de l'algorithme sont :

Algorithme 1 (validation croisée non biaisée UCV)

Début (Génération d'un échantillon $x_{1\leq i\leq n}$)
$Somme1 = 0$, $Somme2 = 0$;
Pour $i = 1$ à n faire
Pour $j = 1$ à n faire
Si $i \neq j$
$Som1 = \frac{1}{\sqrt{2\pi}}\exp\big(-(\frac{x_i - x_j}{\sqrt{2}h})^2\big)$,
$Somme1 = Somme1 + Som1$,
$Som2 = \frac{1}{\sqrt{2\pi}}\exp\big(-(\frac{x_i - x_j}{\sqrt{2}h})^2\big)$,
$Somme2 = Somme2 + Som2$,
Fin pour
$UCV(h) = \frac{2}{\sqrt{2\pi}(n-1)h} + \frac{1}{n^2h}Somme1 - \frac{2}{n(n-1)h}Somme2$,
$h_{lcv} = \min_h UCV(h)$.

2.6.2.2 Validation croisée biaisée

Un critère de validation croisée biaisée, a été introduit par Scott et Terrell [87] en 1987 pour remédier aux problèmes de validation croisée non biaisée. Il s'agit d'introduire un biais dans le UCV afin de réduire sa variance.

L'Erreur Quadratique Intégrée Moyenne Asymptotique s'écrit sous la forme (2.27).

$$AMISE = \frac{h^4}{4}\sigma_K^4 R(f'') + \frac{R(K)}{nh}.$$

Le paramètre de lissage basé sur la méthode de validation croisée biaisée est la valeur h qui minimise un estimateur du $AMISE$. On peut estimer le $AMISE$ si l'on estime $R(f'')$. Un estimateur naturel de ce terme est donné par $R(f_h'')$ où f_h est l'estimateur de la densité qui utilise la méthode du noyau.

Lemme 2.6.1 *(Scott et Terrell [87])*
Supposant que le noyau K satisfait aux conditions suivantes :

$$\int K''(u)du = 0, \quad \mu_1(K'') = \int uK''(u) = 0, \quad \mu_2(K'') = \int u^2K''(u) = 2.$$

On obtient le développement asymptotique :

$$\mathbb{E}\big[R(f_h'')\big] = R(f'') + \frac{R(K'')}{nh^5} + O(h^2).$$

Proposition 2.6.2 *(Scott et Trell[87])*
Soit $X_1, X_2, ..., X_n$ un n-échantillon i.i.d issu d'une variable aléatoire X de fonction de densité f. Pour un noyau K on obtient :

$$BCV(h) = \frac{R(K)}{nh} + h^4\frac{\mu_2^2(K)}{4n^2}\sum_i\sum_{j,j\neq i} K_h^{(2)}K_h^{(2)}(X_i - X_j). \tag{2.44}$$

Proposition 2.6.3
Soit $X_1, X_2, ..., X_n$ un n-échantillon i.i.d issu d'une variable aléatoire X de fonction de densité f. en choisissant le noyau gaussien on obtient :

$$BCV(h) = \frac{1}{2nh\sqrt{\pi}} + \frac{1}{64n^2h\sqrt{\pi}}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\left[\left(\frac{x_i - x_j}{h}\right)^4 - 12\left(\frac{x_i - x_j}{h}\right)^2 + 12\right]\exp\left[-\frac{(x_i - x_j)^2}{4h^2}\right]. \tag{2.45}$$

Preuve :
On a $K(x) = \frac{1}{\sqrt{2\pi}}\exp(-x^2/2)$ et $\int x^2K(x) = 1$.
Calculons la valeur des deux expressions suivantes $\int\big[K(x)\big]^2dx$ et $\int\big[K''(x)\big]^2dx$.

$$\int\big[K(x)\big]^2dx = \frac{1}{2\pi}\int\exp(-x^2)dx = \frac{1}{2\sqrt{\pi}}. \tag{2.46}$$

De plus, si on dérive $K(x)$ deux fois par rapport à x, on obtient $K''(x) = \frac{-1}{\sqrt{2\pi}}\exp(-x^2/2) + \frac{-1}{\sqrt{2\pi}}x^2\exp(-x^2/2)$. On a donc :

$$\int\big[K''(x)\big]^2 = \int\left[\frac{-1}{\sqrt{2\pi}}\exp(-x^2/2) + \frac{-1}{\sqrt{2\pi}}x^2\exp(-x^2/2)\right]^2 \tag{2.47}$$

$$= \frac{3}{8\sqrt{\pi}}. \tag{2.48}$$

Reste maintenant à évaluer $\int \big[f_h(x)\big]^2 dx$. Puisque le noyau K est gaussien on a :

$$\begin{aligned} f_h(x) &= \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x-x_i}{h}\Big) \\ &= \frac{1}{nh\sqrt{2\pi}}\sum_{i=1}^{n} \exp\big(-(\frac{x-x_i}{\sqrt{2}h})^2\big). \end{aligned}$$

Il est facile de montrer que si l'on dérive $f_h(x)$ deux fois par rapport à x, on obtient :

$$f_h''(x) = \frac{1}{nh^3\sqrt{2\pi}}\sum_{i=1}^{n}\left[(\frac{x-x_i}{h})^2 \exp\left(-(\frac{x-x_i}{\sqrt{2}h})^2\right) - \exp\left(-(\frac{x-x_i}{\sqrt{2}h})^2\right)\right].$$

D'où

$$\begin{aligned} \int \big[f_h''(x)\big]^2 dx &= \int \frac{1}{nh^3\sqrt{2\pi}}\sum_{i=1}^{n}\left[(\frac{x-x_i}{h})^2 \exp\left(-(\frac{x-x_i}{\sqrt{2}h})^2\right) - \exp\left(-(\frac{x-x_i}{\sqrt{2}h})^2\right)\right] \times \\ &\quad \frac{1}{nh^3\sqrt{2\pi}}\sum_{j=1}^{n}\left[(\frac{x-x_j}{h})^2 \exp\left(-(\frac{x-x_j}{\sqrt{2}h})^2\right) - \exp\left(-(\frac{x-x_j}{\sqrt{2}h})^2\right)\right] dx \\ &= \frac{1}{2\pi n^2h^6}\sum_{i=1}^{n}\sum_{j=1}^{n}\int (\frac{x-x_i}{h})^2(\frac{x-x_j}{h})^2 \exp\left[-\left((\frac{x-x_i}{\sqrt{2}h})^2 + (\frac{x-x_j}{\sqrt{2}h})^2\right)\right] \\ &- (\frac{x-x_i}{h})^2 \exp\left[-\left((\frac{x-x_i}{\sqrt{2}h})^2 + (\frac{x-x_j}{\sqrt{2}h})^2\right)\right] - (\frac{x-x_j}{h})^2 \times \\ &\quad \exp\left[-\left((\frac{x-x_i}{\sqrt{2}h})^2 + (\frac{x-x_j}{\sqrt{2}h})^2\right)\right] + \exp\left[-\left((\frac{x-x_i}{\sqrt{2}h})^2 + (\frac{x-x_j}{\sqrt{2}h})^2\right)\right] dx. \end{aligned}$$

Afin d'obtenir l'expression pour l'intégrale ci-dessus, il suffit de calculer les trois intégrales suivantes :

- $\int_{\mathbb{R}} \big(\frac{x-x_i}{h}\big)^2\big(\frac{x-x_j}{h}\big)^2 \exp\left[-\left((\frac{x-x_i}{\sqrt{2}h})^2 + (\frac{x-x_j}{\sqrt{2}h})^2\right)\right] dx,$
- $\int_{\mathbb{R}} \big(\frac{x-x_i}{h}\big)^2 \exp\left[-\left((\frac{x-x_i}{\sqrt{2}h})^2 + (\frac{x-x_j}{\sqrt{2}h})^2\right)\right] dx,$
- $\int_{\mathbb{R}} \exp\left[-\left((\frac{x-x_i}{\sqrt{2}h})^2 + (\frac{x-x_j}{\sqrt{2}h})^2\right)\right] dx.$

Les expressions obtenues pour chacune de ces intégrales sont respectivement :

- $\exp\left[-(\frac{x_i-x_j}{2h})^2\right]\left[\frac{3h\sqrt{\pi}}{4} - \frac{\sqrt{\pi}}{4h}(x_i-x_j)^2 + \frac{\sqrt{\pi}}{16h^3}(x_i-x_j)^4\right],$
- $\exp\left[-(\frac{x_i-x_j}{2h})^2\right]\left[\frac{\sqrt{\pi}h}{2} + \frac{\sqrt{\pi}}{4h}(x_i-x_j)^2\right],$
- $\sqrt{\pi}h \exp\left[-(\frac{x_i-x_j}{2h})^2\right].$

On obtient alors après calcul :

$$\int \big[f_h''(x)\big]^2 dx = \frac{1}{n^2h^6\sqrt{\pi}}\sum_{i=1}^{n}\sum_{i\neq j,j=1}^{n} \exp\left[-(\frac{x_i-x_j}{2h})^2\right]\left[\frac{3h}{4} - \frac{3}{4h}(x_i-x_j)^2 + \frac{1}{16h^3}(x_i-x_j)^4\right] + \frac{3}{8\sqrt{\pi}nh^5}. \tag{2.49}$$

Si l'on substitue les équations (2.46), (2.47) et (2.49) dans l'équation (2.44). On obtient l'expression pour le $BCV(h)$ donnée par :

$$\frac{1}{2nh\sqrt{\pi}} + \frac{1}{64n^2h\sqrt{\pi}} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \left[\left(\frac{x_i - x_j}{h} \right)^4 - 12 \left(\frac{x_i - x_j}{h} \right)^2 + 12 \right] \exp\left[-\frac{(x_i - x_j)^2}{4h^2} \right].$$

Des résultats de simulations ont été obtenus pour la méthode de validation croisée biaisée dans le travail de Park et Marron [76]. Les auteurs ont constaté que la méthode validation croisée biaisée présente le même point faible que celui de la méthode validation croisée non biaisée. Cette méthode nous donne plusieurs minimums locaux pour la fonctionnelle cible a minimiser. Cependant, d'après plusieurs simulations, les auteurs proposent de choisir la valeur inférieure parmis les minimums locaux.

Algorithme de la méthode validation croisée biaisée

Afin de calculer le paramètre de lissage optimal noté h_{bcv} qui minimise $BCV(h)$. En utilisant le noyau gaussien, les principales étapes de l'algorithme sont :

Algorithme 2 (validation croisée biaisée BCV)

Début (Génération d'un échantillon $x_{1\leq i\leq n}$)
$BCV(h) = 0$;
Pour $i = 1$ à n faire
Pour $j = 1$ à n faire
Si $i \neq j$, $x = \left(\frac{x_i - x_j}{h} \right)$;
$BCV(h) = BCV(h) + \exp(-\frac{x^2}{4})\big(3 - 3x^2 + \frac{1}{4}x^4\big)$,
Fin pour
$BCV(h) = \frac{1}{2nh\sqrt{\pi}} + \frac{1}{16n^2h\sqrt{\pi}} BCV(h)$,
$h_{bcv} = \min_h BCV(h)$.

2.7 Bootstrap dans l'estimation de la densité de probabilité

2.7.1 Bootstrap dans l'estimation locale de la densité de probabilité

Soit f une densité de probabilité réelle inconnue et $X_1, X_2, ..., X_n$ un n-échantillon $i.i.d$ issu de f. Pour estimer f à un point x nous utilisons l'estimateur de Parzen-Rosenblatt :

$$f_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{2.50}$$

où K est un noyau tel que $\int_R K(x)dx = 1$ et $h > 0$ le paramètre de lissage.
Le MSE (Mean Squared Error) qui est donné par :

$$MSE(f_h(x), f(x)) = E(f_h(x) - f(x))^2 \tag{2.51}$$

est la mesure usuelle la plus appropriée pour l'évaluation de la performance d'un estimateur local de f en un point x. Il est bien connu que sous quelques conditions de régularité sur f et K (Silverman, 1986 [91]), le h optimal au sens de minimisation de $MSE(x)$ par rapport au paramètre h, est asymptotiquement :

$$h^* \sim n^{-1/5} \, . \, S^* \, , \tag{2.52}$$

où :

$$S^* \equiv S^*(x) = \left(\frac{f(x)R(K)}{[f''(x)\sigma_k^2]^2}\right)^{1/5} . \tag{2.53}$$

Cela n'est vérifie que si et seulement si $f(x) > 0$ et $f''(x) \neq 0$.
Or, S_{opt} dépend de deux termes inconnus $f(x)$ et $f''(x)$. L'approche la plus naturelle pour le choix du paramètre de lissage réalisable optimal est la méthode de plug-in qui consiste à remplacer l'estimateur de $f(x)$ et $f''(x)$ dans la formule de S_{opt} ; qui résulte, le paramètre h s'écrit sous la forme :

$$h = n^{-1/5} S, \text{ avec } S \equiv S(x). \tag{2.54}$$

Il est bien connu que si on remplace h^* par h on aura pas d'impact sur l'optimalité au sens de minimisation de MSE asymptotique (Woodroofe(1970) [105]). Hall (1993) [47] a démontré que l'ordre de MSE peut être améliorer par la procédure de plug-in , on choisissant un noyau pilote pour f et f''.
Une procédure de sélection, plus récente, de largeur de la fenêtre qui suscite beaucoup d'attention consiste à remplacer le MSE par la version bootstrapé MSE^* qui peut être réduit au minimum directement puisqu'il est entièrement basé sur une distribution connue. L'idée est de construire B échantillons par la technique de bootstrap à partir de l'échantillon $X_1,\ X_2\ , ...,\ X_n$ et d'estimer $f(x)$ pour ces différents échantillons. C'est-à-dire :

$$f_h^j(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i^j}{h}\right), \text{ pour } j = 1, 2, ..., B. \tag{2.55}$$

Nous constatons que à partir (2.52), le choix du paramètre est réduit à un problème de sélection d'un scalaire s dans $h = n^{-1/5}s$, on remplace dans (2.55) on aura :

$$f_s^j(x) = \frac{1}{n^{4/5}s} \sum_{i=1}^{n} K\left(\frac{x - X_i^j}{n^{-1/5}s}\right), \text{ pour } j = 1, 2, ..., B. \tag{2.56}$$

et le MSE^* qui correspond au MSE bootstrapé sera défini comme suite :

$$MSE^*(f_s^*(x), f_h(x)) = \mathbb{E}\left((f_s^*(x) - f_h(x))^2\right) \tag{2.57}$$

avec

$$f_s^*(x) = \frac{1}{B}\sum_{j=1}^{B} f_s^j(x),$$

et $f_h(x)$ l'estimateur de $f(x)$ obtenu a partir de l'échantillon initial et h est calculé par la formule (2.54).

Enfin, le paramètre de lissage est alors sélectionné par :

$$h^* = n^{-1/5}.\arg\min_s MSE^*(f_s^*(x), f_h(x)),$$

dans le cas d'existence du minimum.

2.7.2 Bootstrap dans l'estimation globale de la densité de probabilité

La technique de bootstrap a été introduite aussi pour la sélection du paramètre de lissage h qui minimise le MISE (Mean Integer Square Error).

Soit $X_1, X_2, ..., X_n$ des observations indépendantes issues d'une densité de probabilité f inconnue. Considérons l'estimateur à noyau de Parzen-Rosenblatt de f de la forme :

$$f_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{2.58}$$

où h est le paramètre de lissage est $K(.)$ un noyau. Le h approprié est celui qui permet d'obtenir un bon estimateur de $f(x)$. En générale le critère utiliser pour l'évaluation de l'efficacité de cette technique est la minimisation de $MISE$ par rapport à h qui est définie comme suite :

$$MISE(f_h, f) = \int E\left(f_h(x) - f(x)\right)^2 dx \tag{2.59}$$

Soit h_0 le paramètre qui minimise ce critère.

Pour calculer la valeur de la fenêtre par la technique de bootstrap on doit rééchantillonner par cette technique à partir de l'échantillon initial X et construisant ensuite l'estimateur de bootstrap sous la forme :

$$f_h^j(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i^j}{h}\right), \quad \text{pour } j = 1, 2, ..., B; \tag{2.60}$$

où B est le nombre de réplications de bootstrap.

L'estimateur de la variance de $f_h^j(x)$ dans ce cas est donné sous la forme :

$$\frac{1}{B}\sum_{j=1}^{B}\int \left(f_h^j(x) - \bar{f}_h^*(x)\right)^2,$$

avec,

$$\bar{f}_h^*(x) = \frac{1}{B}\sum_{j=1}^{B} f_h^j(x).$$

Nous construisons un estimateur initial de la densité f_{h_0} ensuite rééchantillonné par la technique de bootstrap à partir de l'échantillon initial pour construire les estimateurs $f_h^j(x)(j = 1, ..., B.)$.

Enfin, nous obtenons la fenêtre bootstrapée h_{boot} par la minimisation de $BMISE(h, h_0)$ sous h qui s'écrit sous la forme :

$$BMISE(h, h_0) = \frac{1}{B}\sum_{j=1}^{B}\int \left(f_h^j(x) - f_{h_0}(x)\right)^2 dx. \tag{2.61}$$

Remarque 2.1 *On peut définir un autre h_{boot} par :*

$$h_{boot} = \arg\min_h \left(\int \left(\frac{1}{B}\sum_{j=1}^{B} f_h^j(x) - f_{h_0}(x)\right)^2 dx\right) \tag{2.62}$$

si on suppose qu'on s'intéresse à la statistique $f_h(x)$.

2.8 Simulations et résultats

2.8.1 Introduction

La technique de bootstrap a apportée des solutions à divers problèmes. Peut-elle améliorer l'estimateur de la densité de probabilité par la méthode du noyau ? C'est l'objet de la section suivante dans laquelle, nous allons par simulation étudier l'influence du bootstrap sur la qualité de l'estimateur de la densité.

Afin d'illustrer ou de vérifier l'influence de l'application de la technique de bootstrap sur la qualité (performance) de l'estimateur non paramétrique d'une densité de probabilité au sens de minimisation du $AMISE$, on effectue quelques simulations. Ces simulations sont organisées selon les étapes suivantes :

Etape1 : Générer un n-échantillon $i.i.d$ d'une densité de probabilité cible.

Etape2 : Construire B échantillons de *Bootstrap*.

Etape3 : Estimer h^* et h_{boot} par une méthode de sélection .

Etape4 : Estimer $f_{h^*}(x)$ et $f_{h_{boot}}(x)$.

Etape5 : Calculer l'$AMISE(f_{h^*}, f)$ qui sera notée $AMISE^*$ et $AMISE(f_{h_{boot}}, f)$ qui sera notée $AMISE_{boot}$

Après l'implémentation de ces étapes sous le logiciel R, des exécutions ont été réalisées pour les différentes variantes suivantes :

- **Densités cibles :**
  1. Normale centrée réduite,
  2. Exponentielle de paramètre $\lambda = 1$,
  3. Khi-deux de degré de liberté 4.
- **Taille d'échantillon d'étape 1 :**
  1. $n = 1000$
  2. $n = 5000$
  3. $n = 10000$
- **Nombre de réplications de *Bootstrap* :**

1. $B = 10$
2. $B = 50$
3. $B = 100$
4. $B = 250$
5. $B = 500$

– **Méthodes de sélection :**

1. NRD implémente la méthode "rule-of-thumb" de Silverman, l'estimateur sera noté h_1.
2. UCV utilise la méthode de validation croisée non biaisée, l'estimateur sera noté h_2.
3. BCV utilise la méthode de validation croisée biaisée, l'estimateur sera noté h_3.
4. SJ implémente la méthode Sheather&Jones, l'estimateur sera noté h_4.

– **Les noyaux**

1. Noyau Gaussien
2. Noyau Biweight (Tukey)
3. Noyau Epanechnikov
4. Noyau Miroir (Schuster)
5. Noyau Gamma(2^{eme} version)

Pour le calcul de h_{boot} on a utilisé l'algorithme suivant :

Etape1 : Générer un n-échantillon *i.i.d* d'une densité de probabilité cible.

Etape2 : Construire B échantillons de *Bootstrap*.

Etape3 : Estimer h_b (b=1,...,B) pour chaque échantillon de bootstrap par une méthode de sélection classique.

Etape4 : Calculer h_{boot} par :

$$h_{boot} = \frac{1}{B}\sum_{b=1}^{B} h_b \quad .$$

Remarque 2.2 *Dans la simulation, on a utilisé l'approche "simulation de Monté Carlo" et on a fixé le nombre de réplications à 100.*

2.8.2 Résultats

2.8.2.1 Estimation du paramètre de lissage

Dans cette section, nous allons présenter les résultats des simulations obtenus pour l'estimation du paramètre de lissage d'une densité de probabilité dans le cas : d'une loi normale, exponentielle et khi-deux par les méthodes de sélection : NRD, UCV, BCV et SJ en appliquant la technique de bootstrap ainsi que leur AMISE associé.

2.8.2.1.1 Loi normale

Paramètre	n	B	h^*	h_{boot}	$AMISE^*$	$AMISE_{boot}$
h_1	1000	10	0.2233067	0.2221746	0.0141290	0.0138793
		50	0.2229254	0.2217659	0.0143423	0.0140725
		100	0.2248300	0.2237004	0.0149243	0.0143634
		250	0.2241853	0.2228865	0.0147292	0.0145423
		500	0.2240398	0.2229024	0.0144903	0.0140256
	5000	10	0.1613476	0.1601308	0.0043889	0.0039225
		50	0.1617033	0.1601936	0.0039531	0.0037374
		100	0.1612495	0.1608970	0.0038098	0.0035195
		250	0.1612749	0.1607906	0.0039140	0.0036191
		500	0.1613403	0.1601697	0.0040798	0.0035145
	10000	10	0.1505124	0.1492378	0.0022290	0.0021525
		50	0.1497519	0.1492374	0.0024073	0.0021074
		100	0.1502305	0.1491917	0.0021739	0.0020193
		250	0.1501873	0.1497183	0.0022058	0.0021111
		500	0.1492824	0.1498973	0.0023249	0.0023009
h_2	1000	10	0.2637520	0.2626835	0.0131875	0.0127912
		50	0.2713318	0.2643178	0.0133676	0.0133859
		100	0.2690290	0.2624879	0.0138132	0.0134523
		250	0.2741768	0.2644120	0.0135987	0.0130460
		500	0.2782387	0.2638493	0.0134459	0.0129456
	5000	10	0.2029686	0.1963668	0.0040273	0.0038783
		50	0.1987253	0.1959832	0.0038926	0.0038732
		100	0.1996528	0.1979249	0.0039821	0.0038867
		250	0.2017549	0.1963248	0.0041217	0.0038927
		500	0.2004525	0.1942246	0.0039784	0.0038623
	10000	10	0.1665237	0.1663452	0.0021732	0.0020348
		50	0.1663478	0.1660485	0.0023756	0.0020235
		100	0.1669836	0.1663073	0.0020124	0.0020099
		250	0.1669893	0.1665347	0.0022384	0.0020222
		500	0.1666928	0.1663834	0.0021386	0.0020237
h_3	1000	10	0.2803133	0.2906413	0.0128093	0.0130671
		50	0.2816399	0.2923112	0.0129138	0.0134826
		100	0.2796954	0.2902684	0.0134912	0.0137561
		250	0.2808413	0.2920106	0.0125973	0.0133967
		500	0.2811230	0.2924904	0.0133174	0.0133478
	5000	10	0.2038746	0.2090397	0.0039286	0.0039874
		50	0.1972895	0.2032374	0.0036334	0.0039947
		100	0.1996865	0.2043401	0.0035789	0.0039176
		250	0.2023900	0.2058912	0.0037923	0.0039009
		500	0.1989920	0.2013265	0.0038469	0.0039384
	10000	10	0.1619373	0.1650896	0.0023391	0.0023739
		50	0.1618341	0.1682934	0.0023598	0.0023912
		100	0.1611349	0.1690093	0.0023237	0.0024374
		250	0.1620075	0.1644465	0.0023883	0.0023982
		500	0.1617254	0.1624787	0.0023113	0.0024120
h_4	1000	10	0.2603793	0.2509944	0.0129250	0.0125867
		50	0.2609745	0.2432584	0.0131968	0.0129331
		100	0.2618081	0.2434982	0.0136932	0.0126986
		250	0.2605905	0.2428563	0.0130537	0.0122063
		500	0.2635642	0.2452238	0.0131986	0.0127583
	5000	10	0.1887019	0.1887826	0.0042946	0.0039931
		50	0.1886753	0.1889731	0.0037947	0.0032656
		100	0.1883298	0.1886132	0.0037365	0.0035413
		250	0.1882765	0.1883589	0.0039764	0.0036914
		500	0.1885491	0.1884050	0.0038349	0.0037190
	10000	10	0.1589912	0.1589214	0.0020907	0.0020589
		50	0.1589735	0.1589373	0.0022064	0.0021975
		100	0.1589273	0.1589910	0.0021834	0.0021758
		250	0.1589974	0.1589327	0.0021681	0.0021497
		500	0.1589017	0.1589130	0.0020899	0.0020512

TAB. 2.1: *Résultats des simulations effectués sur la loi Normale(0,1) pour déterminer les largeurs de fenêtres optimales h^* et h_{boot}.*

2.8.2.1.2 Loi exponentielle

Paramétre	n	B	h^*	h_{boot}	$AMISE^*$	$AMISE_{boot}$
h_1	1000	10	0.1852068	0.1856418	0.5360971	0.5356823
		50	0.1836891	0.1839001	0.5409105	0.5406917
		100	0.1857607	0.1858214	0.5540642	0.5540037
		250	0.1841570	0.1842001	0.5355922	0.5355492
		250	0.1863983	0.1864592	0.5405894	0.5405584
		500	0.1854954	0.1853820	0.5511444	0.5512484
	5000	10	0.1339753	0.1340086	0.3225942	0.3225909
		50	0.1323279	0.1319427	0.3207328	0.3207317
		100	0.1334372	0.1340642	0.3207753	0.3207794
		250	0.1328347	0.1327992	0.3213329	0.3213346
		500	0.1315546	0.1314851	0.3209876	0.3209912
	10000	10	0.1170578	0.1175697	0.2650570	0.2650566
		50	0.1167590	0.1164561	0.2643568	0.2643561
		100	0.1186543	0.1181330	0.2586017	0.2586029
		250	0.1170987	0.1175809	0.2644136	0.2644145
		500	0.1166543	0.1160719	0.2663562	0.2663587
h_2	1000	10	0.0602429	0.0593424	0.2800012	0.2859504
		50	0.0585438	0.0597847	0.2791145	0.2865065
		100	0.0595894	0.0601354	0.2793140	0.2803535
		250	0.0606347	0.0596781	0.2829039	0.2859024
		500	0.0614829	0.6251291	0.2783237	0.2812265
	5000	10	0.0217138	0.0214390	0.0575054	0.0604128
		50	0.0220979	0.0218324	0.0587671	0.0603145
		100	0.0215915	0.0216328	0.0531080	0.0563274
		250	0.0209312	0.0218050	0.0543056	0.0582154
		500	0.0221009	0.0218694	0.0578763	0.0601623
	10000	10	0.0189512	0.0187808	0.0172925	0.0175139
		50	0.0187356	0.0184546	0.0159463	0.0175871
		100	0.0190909	0.0185561	0.0155283	0.0163543
		250	0.0188876	0.0188732	0.0164934	0.0179058
		500	0.0189219	0.0188053	0.0165632	0.0175967
h_3	1000	10	0.0957601	0.0923327	0.3177505	0.3197029
		50	0.0937565	0.0965328	0.3209693	0.3211784
		100	0.0967867	0.0954992	0.3166561	0.3104783
		250	0.0973491	0.0973780	0.3323125	0.3324214
		500	0.0948340	0.0934896	0.3259604	0.3274038
	5000	10	0.0574679	0.0567974	0.0931346	0.0957324
		50	0.0569907	0.0565325	0.0887760	0.0965220
		100	0.0573248	0.0559323	0.0917338	0.0928959
		250	0.0570746	0.0569642	0.0925603	0.0931216
		500	0.0567344	0.0563477	0.0909375	0.0925467
	10000	10	0.0199352	0.0197987	0.0161917	0.0163978
		50	0.0209462	0.0197764	0.0167019	0.0170754
		100	0.0198665	0.0199819	0.0167005	0.0172789
		250	0.0197796	0.0195088	0.0161098	0.0167899
		500	0.0196036	0.0197593	0.0164572	0.0168256
h_4	1000	10	0.0992531	0.0981050	0.3217277	0.338063
		50	0.0997447	0.0989786	0.3260613	0.337089
		100	0.0997769	0.0989617	0.2747429	0.308961
		250	0.0996543	0.0987863	0.2823170	0.316849
		500	0.0994548	0.0981204	0.3205451	0.375097
	5000	10	0.0413247	0.0405765	0.1682984	0.1701790
		50	0.0410946	0.0409486	0.1844357	0.1910789
		100	0.0394319	0.0396464	0.1770978	0.1916037
		250	0.0414704	0.0390642	0.1682692	0.1808675
		500	0.0408061	0.0404292	0.1787701	0.1847453
	10000	10	0.0342617	0.0348648	0.0886295	0.0900169
		50	0.0346901	0.0346344	0.0783287	0.0887651
		100	0.0341375	0.0346091	0.0796114	0.0878765
		250	0.0345903	0.0340978	0.0823089	0.0840856
		500	0.0348746	0.0347605	0.0882560	0.0906524

TAB. 2.2: *Résultats des simulations effectués sur la loi exponentielle* ($\lambda = 1$) *pour déterminer les largeurs de fenêtres optimales* h^* *et* h_{boot}.

2.8.2.1.3 Loi de khi-deux

Paramétre	n	B	h^*	h_{boot}	$AMISE^*$	$AMISE_{boot}$
h_1	1000	10	0.5836305	0.5893339	0.0127878	0.0124030
		50	0.5803934	0.5836417	0.0124754	0.0121900
		100	0.5824593	0.5867408	0.0121792	0.0125347
		250	0.5818564	0.5850120	0.0126345	0.0129751
		500	0.5825956	0.5843754	0.0122367	0.0123489
	5000	10	0.4231369	0.4231312	0.0048128	0.0048005
		50	0.4230813	0.4231483	0.0048619	0.0048403
		100	0.4237991	0.4232442	0.0048098	0.0048575
		250	0.4237457	0.4232293	0.0048704	0.0048934
		500	0.4236310	0.4238404	0.0048975	0.0047057
	10000	10	0.3693451	0.3695123	0.0034068	0.0034051
		50	0.3690158	0.3690975	0.0033450	0.0033934
		100	0.3699316	0.3698498	0.0033910	0.0034123
		250	0.3695733	0.3694575	0.0033929	0.0034050
		500	0.3697895	0.3694193	0.0033717	0.0033813
h_2	1000	10	0.3715180	0.3799930	0.0093386	0.0092105
		50	0.3774936	0.3755439	0.0093535	0.0092893
		100	0.3737951	0.3756513	0.0090049	0.0094754
		250	0.3703175	0.3718493	0.0097327	0.0105745
		500	0.3792440	0.3886014	0.0093183	0.0093410
	5000	10	0.2420217	0.2419986	0.0029242	0.0029076
		50	0.2427154	0.2411235	0.0029134	0.0029109
		100	0.2429067	0.2399971	0.0028575	0.0029451
		250	0.2431584	0.2421799	0.0027802	0.0029545
		500	0.2427680	0.2409193	0.0028379	0.0029098
	10000	10	0.2013090	0.2010148	0.0018916	0.0018101
		50	0.2018974	0.2018163	0.0019573	0.0019043
		100	0.2017896	0.2011041	0.0018454	0.0018250
		250	0.2011093	0.2013891	0.0018659	0.0018797
		500	0.2014169	0.2016413	0.0018874	0.0018924
h_3	1000	10	0.4732024	0.4741229	0.0102491	0.0103973
		50	0.4700045	0.4712519	0.0100890	0.0108957
		100	0.4695827	0.4723144	0.0098124	0.0108345
		250	0.4668769	0.4695810	0.0091242	0.0108565
		500	0.4652894	0.4688928	0.0097429	0.0105311
	5000	10	0.2910544	0.2957454	0.0031230	0.00316345
		50	0.2924013	0.2943578	0.0030944	0.00318912
		100	0.2917951	0.2939004	0.0030679	0.00315081
		250	0.2914935	0.2929394	0.0030947	0.00315390
		500	0.2921354	0.2935894	0.0030394	0.00314301
	10000	10	0.1759574	0.1761958	0.0019034	0.00191983
		50	0.1758596	0.1760948	0.0019113	0.00195135
		100	0.1759544	0.1761748	0.0018923	0.00193659
		250	0.1757940	0.1761814	0.0018710	0.00190834
		500	0.1758103	0.1762058	0.0018812	0.00195643
h_4	1000	10	0.4570778	0.4591045	0.0099834	0.0098031
		50	0.4564454	0.4581978	0.0096973	0.0104813
		100	0.4555458	0.4569107	0.0097293	0.0101051
		250	0.4575913	0.4542377	0.0093432	0.0103572
		500	0.4565134	0.4568251	0.0095947	0.0104359
	5000	10	0.3007191	0.3005974	0.0031480	0.0030732
		50	0.3003459	0.3002591	0.0031536	0.0030908
		100	0.3000134	0.3008515	0.0030832	0.0030824
		250	0.3007581	0.3009493	0.0031324	0.0030249
		500	0.3000948	0.3005417	0.0030595	0.0030748
	10000	10	0.2519048	0.2518586	0.0020051	0.0020007
		50	0.2520102	0.2520024	0.0020814	0.0020334
		100	0.2520473	0.2520128	0.0020492	0.0020983
		250	0.2519585	0.2520394	0.0020093	0.0020497
		500	0.2518949	0.2520813	0.0020910	0.0020908

TAB. 2.3: *Résultats des simulations effectués sur la loi χ^2_4 pour déterminer les largeurs de fenêtres optimales h^* et h_{boot}.*

Performances des estimateurs

1. **Loi normale :**

   A partir des résultats des simulations obtenus pour la loi normale (tableau 2.1), on constate que :

   – La méthode BCV (biased cross validation) est meilleure au sens du AMISE (elle admet le AMISE le plus petit).

   – L'augmentation de n entraîne la décroissance de h^*, h_{boot}, $AMISE^*$ et de $AMISE_{boot}$.

   – La bootstrap améliore (diminue) l'$AMISE$ pour toutes les méthodes considérées sauf pour la BCV qui est dû au biais important de cette méthode.

– On constate aussi que l'amélioration apportée par la bootstrap est indépendant de nombre de réplication B de bootstrap.

– - La technique de bootstrap nous donne des estimateurs du paramètre de lissage h plus stables que ceux obtenus par les méthodes classiques (de même pour les AMISE). Cette remarque est valable même dans le cas d'échec de la technique bootstrap (méthode BCV).

2. **Loi exponentielle :**

Pour le cas de la loi exponentielle, qui appartient à la famille des densités à queue lourde, les résultats obtenus (tableau 2.2) nous permettent de conclure que :

– La méthode de UCV est la meilleure au sens du critère du AMISE.

– Les échantillons suivant une telle loi ne permet pas une estimation correcte de cette densité. En effet, pour cette loi, on constate que les méthodes classiques nous fournissent un AMISE plus petit que la technique de bootstrap (voir les valeurs d'AMISE du tableau 2.2). Dans ce cas, les valeurs du paramètre de lissage calculées sont soit $h_{boot} > h^*$ (phénomène de sur-lissage) soit $h_{boot} < h^*$ (phénomène de sous-lissage).

– Contrairement au cas de la densité de la loi normale, l'estimation de la densité de la loi exponentielle dépend du nombre de réplications B de bootstrap. En effet, on constate que pour $B \in \{10, 50\}$ une amélioration du l'AMISE aura lieu (sauf pour la BCV) ; pour d'autre valeurs de B soit la technique de bootstrap échoue ou le résultat dépend des échantillons bootstrapés. Pour certains échantillons l'application de la bootstrap engendre un gain dans l'AMISE et pour d'autres la technique de bootstrap échoue.

– L'échec de la bootstrap, peut être expliqué par la présence du biais au voisinage de zéro de l'estimateur de la densité exponentielle, car la bootstrap classique échoue dans l'estimation d'une statistique biaisée.

3. **Loi de khi-deux :**

Les résultats des simulations obtenus pour la loi de khi-deux de degré de liberté 4 (tableau 2.3), nous permettent de tirer les mêmes conclusions que dans le cas de la densité de probabilité exponentielle, soit pour les méthodes classiques ou pour la technique de bootstrap :

– La méthode UCV est meilleure au sens du AMISE par rapport aux autres méthodes.

– Pour une augmentation de la taille de l'échantillon de 10 fois on obtient une diminution de quatre fois de l'AMISE (de n=1000 à n=10 000) pour toutes les méthodes de sélection. Cette très faible diminution de l'$AMISE$ peut s'expliquer par la convergence lente de l'estimateur et la non convergence au voisinage de zéro (la densité de la loi de khi-deux appartient à la famille des lois à convergence lente).

2.8.2.2 Analyse sensitive

D'après les résultats précédents la technique de bootstrap peut dans certains cas améliorer la qualité de l'estimateur au sens du AMISE. Mais, on constate que la durée d'exécution des simulateurs est très grande lorsque on applique la technique bootstrap. A cet effet, une étude sensitive temporelle a été effectuée pour les différentes méthodes de sélection du paramètre de

lissage h et pour différentes valeurs de B. Les résultats obtenus sont résumés dans les deux tableaux 2.4 et 2.5.

Le tableau 2.4 représente les durées moyennes nécessaires pour le calcul du paramètre de lissage par les méthodes classiques pour différentes tailles des échantillons issus d'une loi normale, exponentielle et de khi-deux.

La loi	n	h_1	h_2	h_3	h_4
Loi normale	1000	0.0024	0.0676	0.0636	0.0648
		0.0016	0.0680	0.0656	0.0680
	2500	0.0040	0.3664	0.3572	0.3608
		0.0040	0.3736	0.3780	0.3744
	5000	0.0048	1.4704	1.4304	1.4348
		0.0032	1.3756	1.3572	1.3676
	10000	0.0064	5.4288	5.3912	5.4252
		0.0056	5.4352	5.3952	5.4332
Loi exponentielle	1000	0.0020	0.0564	0.0544	0.0608
		0.0020	0.0632	0.0596	0.0672
	2500	0.0020	0.3348	0.3328	0.3408
		0.0032	0.3348	0.3296	0.3404
	5000	0.0036	1.3196	1.3152	1.3304
		0.0028	1.3196	1.3168	1.3296
	10000	0.0056	5.2580	5.2540	5.2828
		0.0060	5.2584	5.2548	5.2780
Loi de khi-deux	1000	0.0032	0.0616	0.0568	0.0632
		0.0024	0.0592	0.0580	0.0644
	2500	0.0036	0.3384	0.3352	0.3452
		0.0036	0.3344	0.3336	0.3444
	5000	0.0032	1.3232	1.3160	1.3360
		0.0048	1.3248	1.3148	1.3376
	10000	0.0060	5.2728	5.2676	5.2948
		0.0060	5.2640	5.2468	5.2924

TAB. 2.4: *Temps moyen nécessaire en secondes pour l'estimation du paramètre de lissage par les différentes méthodes de sélection. Cas : Loi normale, exponentielle et de khi-deux.*

Le tableau 2.5 représente les durées moyennes nécessaires pour le calcul du paramètre de lissage par les méthodes classiques et la technique de bootstrap pour différentes tailles des échantillons issus d'une loi normale, ainsi que les durées moyennes nécessaires pour le calcul du paramètre de lissage, en appliquant la technique de bootstrap. Le temps moyen nécessaire pour le calcul du paramètre de lissage par la technique de bootstrap est donné sous forme d'un intervalle (colonne 3 du tableau 2.5). La première borne correspond à la durée moyenne pour B=10 et la deuxième pour B=500.

Méthode	n	classique	Bootstrap
NRD	1000	0.0016	[0.2608, 9.2796]
	5000	0.0048	[0.1316, 15.1254]
	10000	0.0056	[0.2640, 19.2860]
UCV	1000	0.0680	[3.1500, 38.0160]
	2500	0.3664	[3.3564, 181.5647]
	5000	1.4704	[13.2112, 723.3461]
	10000	5.4352	[52.0948, 2804.7856]
BCV	1000	0.0656	[3.1620, 38.2492]
	2500	0.3780	[3.3616, 181.3129]
	5000	1.4304	[13.2164, 727.5642]
	10000	5.3952	[52.0704, 2897.1543]
SJ	1000	0.0680	[3.0408, 36.7000]
	2500	0.3744	[3.3626, 182.7866]
	5000	1.4348	[13.2796, 761.6412]
	10000	5.4332	[52.3260, 2970.5476]

TAB. 2.5: *Temps moyen nécessaire (en secondes) pour le calcul du paramètre de lissage par les méthodes classiques et la technique de bootstrap pour $B = 10$ et $B = 500$. Cas : Loi normale.*

2.8.2.3 Interprétation

- En comparant les résultats du tableau 2.4, on constate que la durée moyenne de calcul du paramètre de lissage est plus importante dans le cas de la loi normale. Ceci est dû à la complexité de la forme mathématique de cette loi.
- Les méthodes BCV, UCV et SJ nécessitent un temps de calcul très important pour la détermination du h optimal. Ceci est dû à la forme multiplicatif et exponentielle de h_2, h_3 et h_4.
- Si on compare les résultats du tableau 2.5, on constate que les temps moyens de calcul de h par la technique de bootstrap sont très élevés et sont trop sensibles (fortement dépendant) du nombre de réplications B. Pour cette raison, une question reste posée : est-il préférable d'utiliser la technique bootstrap qui nécessite un temps d'exécution très important pour un gain de précision d'ordre 10^{-2} au maximum de la valeur du l'$AMISE$, ou se contenter des méthodes classiques qui nécessitent un temps d'exécution raisonnable ?

2.8.2.4 Calcul du AMISE pour différents noyaux

Dans cette section, on veut calculer l'erreur commise lors de l'estimation d'une densité de probabilité $f(x)$ par la méthode du noyau, en s'intéressant à l'influence de la technique de bootstrap sur le noyau utilisé pour l'estimation de $f(x)$. Les résultats obtenus dans le cas d'une loi normale centrée et réduite sont classés dans le tableau 2.6. D'autres résultats de simulation pour l'estimation d'une densité de probabilité exponentielle de paramètre $\lambda = 1$ et d'une densité de probabilité de khi-deux de degré de liberté 4 sont également donnés. Ils sont classés respectivement dans le tableau 2.7 et le tableau 2.8, où :

- $amise_1$ correspond à l'AMISE obtenu pour le noyau Epanechnikov et $amise_{b1}$ correspond à l'AMISE bootstrapée pour le même noyau.
- $amise_2$ correspond à l'AMISE obtenu pour le noyau Biweight et $amise_{b2}$ correspond à l'AMISE bootstrapée pour le même noyau.
- $amise_3$ correspond à l'AMISE obtenu pour le noyau Schuster et $amise_{b3}$ correspond à l'AMISE bootstrapée pour le même noyau.
- $amise_4$ correspond à l'AMISE obtenu pour le noyau gamma 2 et $amise_{b4}$ correspond à l'AMISE bootstrapée pour le même noyau.

n	h	Gauss.	Epane.	Biwei.
1000	h_1	0.0141	0.0131	0.0436
		0.0143	0.0144	0.0439
		0.0149	0.0140	0.0430
	h_2	0.0131	0.0124	0.0385
		0.0133	0.0135	0.0388
		0.0138	0.0129	0.0380
	h_3	0.0128	0.0122	0.0343
		0.0129	0.0131	0.0346
		0.0134	0.0124	0.0338
	h_4	0.0129	0.0123	0.0369
		0.0131	0.0133	0.0373
		0.0136	0.0126	0.0364
5000	h_1	0.0043	0.0037	0.0107
		0.0039	0.0036	0.0109
		0.0038	0.0038	0.0106
	h_2	0.0040	0.0035	0.0092
		0.0036	0.0034	0.0094
		0.0035	0.0036	0.0091
	h_3	0.0039	0.0035	0.0087
		0.0036	0.0033	0.0088
		0.0035	0.0035	0.0086
	h_4	0.0040	0.0035	0.0090
		0.0036	0.0034	0.0092
		0.0035	0.0035	0.0089
10000	h_1	0.0022	0.0022	0.0058
		0.0024	0.0022	0.0062
		0.0021	0.0021	0.0059
	h_2	0.0021	0.0021	0.0052
		0.0023	0.0021	0.0056
		0.0020	0.0020	0.0053
	h_3	0.0023	0.0023	0.0062
		0.0025	0.0023	0.0065
		0.0022	0.0022	0.0063
	h_4	0.0020	0.0021	0.0049
		0.0022	0.0021	0.0053
		0.0020	0.0020	0.0050

TAB. 2.6: *Résultats des simulations effectués sur la loi Normale(0,1) pour déterminer les AMISE pour différents noyaux.*

h	n	B	$amise_1$	$amise_{b1}$	$amise_2$	$amise_{b2}$	$amise_3$	$amise_{b3}$	$amise_4$	$amise_{b4}$
h_2	1000	10	0.8158900	0.8172653	0.8773973	0.8785201	0.0628054	0.0635382	0.0161366	0.0161424
		50	0.9144138	0.9119522	0.8007408	0.7992636	0.0459392	0.0447051	0.0181240	0.0183341
		100	0.8194956	0.8185549	0.8556567	0.8549958	0.0760343	0.0756930	0.0245195	0.0245780
		250	0.8900072	0.8917226	0.8411298	0.8424278	0.0563084	0.0575163	0.0160835	0.0165922
		500	0.8159415	0.8158520	0.8421741	0.8411209	0.0625738	0.0615720	0.0165029	0.0166452
	2500	10	0.0784133	0.0784048	0.0821921	0.0852894	0.0233244	0.0232641	0.0077749	0.0077787
		50	0.0804945	0.0803803	0.0864820	0.0836296	0.0208841	0.0204025	0.0095261	0.0094218
		100	0.0893738	0.0893033	0.0849527	0.0848954	0.0142715	0.0140490	0.0075513	0.0075001
		250	0.0816903	0.0816685	0.0721977	0.0721732	0.0156081	0.0154695	0.0094949	0.0094628
		500	0.0933548	0.0933838	0.0851609	0.0851780	0.0154833	0.0155898	0.0041001	0.0040925
	5000	10	0.0083725	0.0083767	0.0091097	0.0091140	0.0023665	0.0023992	0.0003582	0.0003638
		50	0.0083762	0.0083800	0.0087144	0.0087180	0.0026190	0.0026552	0.0005486	0.0005489
		100	0.0073624	0.0073624	0.0084080	0.0084075	0.0022619	0.0022572	0.0003509	0.0003508
		250	0.0082011	0.0081918	0.0087243	0.0087127	0.0030365	0.0029019	0.0006140	0.0006094
		500	0.0079490	0.0079508	0.0073386	0.0073403	0.0020054	0.0020248	0.0003270	0.0003275
h_4	1000	10	0.9274914	0.9287839	0.7721525	0.7730992	0.0256172	0.0257814	0.0966429	0.0942967
		50	0.8490935	0.8476073	0.7999490	0.7986728	0.0499935	0.0501605	0.0751576	0.0750330
		100	0.9726110	0.9715667	0.8303785	0.8296706	0.0198851	0.0200197	0.1237604	0.1234712
		250	0.9325880	0.9337777	0.8133108	0.8144709	0.0107349	0.0109441	0.1048228	0.1038240
		500	0.8515873	0.8493262	0.8121030	0.8108708	0.0344300	0.0350962	0.0770548	0.0769996
	2500	10	0.0846891	0.0846696	0.0856098	0.0855945	0.0116371	0.0117638	0.0089344	0.0089224
		50	0.0898376	0.0896810	0.0928338	0.0926633	0.0164453	0.0165567	0.0058991	0.0058833
		100	0.0921760	0.0921180	0.0820543	0.0819981	0.0176702	0.0178116	0.0048394	0.0048152
		250	0.0765944	0.0765527	0.0833211	0.0832884	0.0139549	0.0140172	0.0070466	0.0070308
		500	0.0832602	0.0832713	0.0834229	0.0834372	0.0162073	0.0163378	0.0077377	0.0076767
	5000	10	0.0085693	0.0085716	0.0085685	0.0085717	0.0012422	0.0012667	0.0004133	0.0004130
		50	0.0078618	0.0078635	0.0088440	0.0088477	0.0017420	0.0017475	0.0005012	0.0005009
		100	0.0086987	0.0086979	0.0079939	0.0079935	0.0013112	0.0013028	0.0003543	0.0003533
		250	0.0087981	0.0087871	0.0087314	0.0087200	0.0010652	0.0011266	0.0003701	0.0003792
		500	0.0077783	0.0077824	0.0088969	0.0088999	0.0020464	0.0020625	0.0003740	0.0003743

TAB. 2.7: *Résultats des simulations effectués sur la loi Exponentielle de paramètre $\lambda = 1$ pour déterminer les AMISE pour différents noyaux.*

h	n	B	$amise_1$	$amise_{b1}$	$amise_2$	$amise_{b2}$	$amise_3$	$amise_b$	$amise_4$	$amise_b$
h_2	1000	10	0.9363606	0.8778292	0.1808852	0.1696179	0.4519388	0.4638809	0.0695026	0.0722178
		50	0.8566707	0.8460771	0.1071372	0.1052078	0.4099049	0.4103673	0.1079991	0.1072986
		100	0.8617145	0.8557799	0.1252302	0.1268012	0.5655759	0.5675969	0.0550508	0.0555425
		250	0.9539037	0.9399877	0.0564399	0.0554247	0.7275496	0.7275970	0.0686774	0.0691318
		500	0.8831730	0.8397480	0.1114643	0.1096484	0.8162375	0.8213873	0.0766108	0.0791258
	2500	10	0.2625055	0.2573464	0.0193675	0.0183431	0.2081980	0.2133934	0.0339305	0.0352997
		50	0.2761927	0.2805883	0.0225200	0.0232353	0.1640807	0.1634730	0.0248736	0.0241340
		100	0.2408610	0.2437339	0.0372865	0.0411860	0.2997061	0.2979677	0.0297090	0.0285677
		250	0.2796568	0.2720575	0.0850740	0.0504704	0.2374434	0.2392464	0.0269804	0.0296232
		500	0.2872840	0.2862704	0.0569108	0.0565438	0.2812672	0.2814552	0.0469963	0.0473646
	5000	10	0.0747327	0.0747445	0.0092726	0.0092734	0.0214448	0.0214435	0.0024728	0.0024724
		50	0.0291385	0.0297314	0.0020917	0.0021273	0.0224983	0.0225066	0.0045487	0.0045071
		100	0.0959966	0.0975933	0.0033308	0.0034034	0.0204585	0.0203338	0.0021140	0.0020565
		250	0.0814317	0.0819678	0.0067716	0.0068033	0.0146135	0.0145578	0.0029938	0.0029711
		500	0.0726518	0.0736658	0.0024407	0.0024836	0.0204167	0.0203821	0.0034035	0.0033586
h_4	1000	10	0.0747541	0.0765356	0.2377178	0.2411503	0.7761826	0.7777335	0.0979241	0.0986127
		50	0.0559749	0.0569958	0.2593883	0.2634876	0.6508544	0.6529907	0.1065837	0.1072049
		100	0.0973497	0.0996912	0.2825637	0.2826644	0.4876720	0.4898676	0.0955930	0.0960410
		250	0.0959126	0.0906911	0.2755284	0.2727520	0.6457097	0.6419924	0.1159084	0.1147037
		500	0.0982477	0.0987634	0.2547683	0.2553373	0.5327143	0.5331179	0.1017893	0.1019001
	2500	10	0.0171221	0.0171889	0.0123748	0.0124052	0.2844248	0.2833624	0.0615213	0.0612067
		50	0.0286755	0.0286039	0.0174492	0.0174337	0.3179216	0.3184687	0.0727576	0.0729482
		100	0.0243530	0.0238955	0.0159918	0.0157434	0.2511202	0.2516391	0.0647242	0.0653008
		250	0.0137654	0.0137832	0.0135159	0.0135384	0.2976292	0.2989120	0.0619958	0.0618401
		500	0.0860047	0.0871384	0.0182324	0.0180625	0.4223356	0.4218068	0.0668304	0.0675370
	5000	10	0.0019225	0.0019237	0.0068358	0.0068396	0.0201431	0.0202496	0.0343593	0.0343287
		50	0.005148	0.005153	0.0061187	0.0061210	0.0150969	0.0151360	0.0294774	0.0294518
		100	0.0010282	0.0010219	0.0067590	0.0067494	0.0178559	0.0178502	0.0245536	0.0247956
		250	0.003335	0.003333	0.0059116	0.0059077	0.0147535	0.0147669	0.0293638	0.0294083
		500	0.003604	0.003599	0.0061806	0.0061666	0.0175102	0.0175173	0.0369421	0.0370644

TAB. 2.8: *Résultats des simulations effectués sur la loi de χ^2 de ddl $n = 4$ pour déterminer les AMISE pour différents noyaux.*

Performances des estimateurs

1. Loi normale

Les résultats des simulations obtenus pour les différents noyaux (Gaussien, Epanechnikov et Biweight) (tableau 2.6) nous permettent de conclure que :

- Le choix du noyau n'est pas important au sens d'$AMISE$, car les résultats sont proches.
- L'$AMISE$ maximale est donnée dans le cas du noyau biweight car ce noyau nous donne un estimateur sous lissé. Ceci peut être confirmé sur la figure 2.6.

– La convergence de l'$AMISE$ vers zéro est indépendante du noyau mais dépend essentiellement de la taille n de l'échantillon étudié.
– Enfin, les résultats confirment l'optimalité du noyau Epanechnikov. Car l'$AMISE$ minimale est obtenu pour ce noyau.

2. **Loi exponentielle**

Les résultats obtenus (tableau 2.7) nous permettent de conclure que :
– Le noyau gamma 2 améliore l'estimateur de la densité au sens de l'$AMISE$. En effet, la valeur de l'AMISE diminue de plus de 90% par rapport aux noyaux biweight et Epanechnikov, malgré sa lente convergence. Cette constatation est confirmée graphiquement (graphes fig.1 et fig.2 de l'annexe A).
– L'amélioration (la diminution) de l'$AMISE$ engendrée par la technique bootstrap dépend de la taille n de l'échantillon et du nombre de réplications B. En effet, par exemple, pour n=1000 :
  • Si $B = 10$: la bootstrap échoue pour les différents noyaux.
  • Si $B = 50$: la bootstrap améliore (diminue la valeur) l'$AMISE$ pour tous les noyaux sauf pour le noyau gamma 2.
  • Si on augmente le nombre de réplications à 100 ($B = 100$), la technique bootstrap échoue de nouveau pour tous les noyaux.
  • Si B=500, un gain d'$AMISE$ est engendré par la bootstrap.

  Ceci peut s'expliquer essentiellement par :
  • La dépendance du paramètre de lissage bootstrapé de trois paramètres : la taille n de l'échantillon étudié, le noyau K et le nombre B de réplications de bootstrap.
  • L'existence de plusieurs minimums locaux.
– Enfin, on peut conclure que pour l'estimation d'une densité de probabilité de la loi exponentielle par la méthode du noyau, il est préférable d'utiliser le noyau gamma 2 et la BCV comme procédure de sélection du paramètre de lissage.

3. **Loi de khi-deux**

Les résultats obtenus (tableau 2.8) nous permettent de conclure que :

– Le choix du noyau est important. Mais ce choix dépend aussi de la méthode de sélection utilisée pour le calcul du paramètre de lissage. Par exemple, si nous choisissons la méthode BCV, il est préférable d'utiliser le noyau gamma, si nous choisissons la méthode de SJ dans ce cas, le noyau qui minimise l'$AMISE$ est le noyau Epanechnikov. Enfin, on peut conclure que pour estimer la densité de probabilité de la loi de khi-deux il est préférable d'utiliser le noyau Epanechnikov et SJ comme méthode de sélection du paramètre de lissage.

– Concernant la technique de bootstrap, les mêmes conclusions que pour la densité exponentielle peuvent être faits.

2.9 Conclusion

Les résultats des simulations mettent en relief le problème de l'application de la technique de bootstrap dans le choix du paramètre de lissage dans l'estimation de la densité de probabilité par la méthode du noyau.

En effet, la performance de la technique de bootstrap varie en fonction de la densité à estimer, de la taille de l'échantillon étudié, du nombre de réplications de bootstrap B et du noyau K. Cependant, plusieurs points importants peuvent être soulignés :

- La technique bootstrap échoue lorsque l'estimateur contient un biais, comme dans le cas des densités à queues lourdes (loi exponentielle) ou encore des densités à convergence lente(loi de khi-deux qui est caractérisée par une convergence lente est la non convergence au voisinage de zéro).
- Le choix du noyau est important pour les lois définies sur des supports fermés ou semi-fermés. Dans ce cas, on doit utiliser des noyaux de correction du biais aux bornes. Pour le cas de la densité exponentielle, le meilleur noyau est le noyau gamma car ce dernier nous permet de réduire la valeur du AMISE de 90% par rapport au noyau gaussien ou biweight. L'utilisation de la technique bootstrap dépend du nombre de réplications B et de la procédure de sélection utilisée pour le choix du paramètre de lissage.
- L'inconvénient majeur de la technique de bootstrap est qu'elle nécessite un temps de calcul important.

Chapitre 3

Régression non-paramétrique réelle

3.1 Le modèle non-paramétrique

Nous nous plaçons dans ce chapitre, dans le cadre de l'estimation de la fonction (de la courbe) de régression de la moyenne

$$r(x) = E(Y|X = x),$$

d'une variable réelle Y sur une variable réelle X, et nous supposons que nous disposons pour cette estimation d'un échantillon $\{(X_i, Y_i),\ i = 1,\ \ldots,\ n\}$ de couples indépendants et ayant chacun la même loi que (X, Y). Le modèle qui va nous intéresser et un modèle non-paramétrique, en ce sens que la seule condition que nous ferons sur la fonction r est une condition de régularité

$$r \text{ est } k \text{ fois continûment dérivable,} \tag{3.1}$$

k étant un entier positif ou nul (le cas $k = 0$ correspondent évidemment à l'hypothèse simple de continuité de r). Ce problème a été abondamment étudié dans la littérature, et notre objectif ici n'est de présenter une discussion exhaustive des différentes méthodes d'estimation existantes. Nous souhaitons d'avantage nous attarder sur la nature même du modèle défini par (3.1). Bien sûr, comprendre ce modèle statistique nécessite de connaître un estimateur qui, sous ce modèle, possède de bonnes propriétés mathématiques. C'est la raison pour laquelle nous allons nous limiter à l'étude des estimateurs de type noyau de convolution, puisque la simplicité de leur construction et leur facilité d'utilisation vont de pair avec leur bonnes propriétés asymptotiques. En particulier, ces estimateurs nous permettront de mettre en évidence le problème essentiel lié à ces modèles : la question du choix du paramètre de lissage.

3.2 La méthode du noyau

Les estimateurs de type noyau, introduits indépendamment par Nadaraya (1964)[71] et Watson (1964)[104], sont une des techniques les plus populaires d'estimation sous des modèles de régression de type (3.1). Pour comprendre les idées qui ont amené à l'introduction de ces estimateurs, peut-être faut-il remonter au régressogramme de Tukey (1961)[98] défini de la manière suivante :

$$\hat{r}_{reg}(x) = \frac{\sum_{i=1}^{n} Y_i \mathbf{1}(X_i \in B_j)}{\sum_{i=1}^{n} \mathbf{1}(X_i \in B_j)}, \quad \forall\, x \in B_j, \tag{3.2}$$

où $B_j,\ j = 1, \ldots, J$ est une partition du support de X fixée à priori. Comme l'histogramme et l'estimation de densité, cet estimateur primitif présente comme inconvénient d'avoir à choisir à la fois la finesse de la déscrétisation (i.e. le nombre de J de découpages) et la position exacte des bornes des intervalles B_j.

Afin de résoudre ce second problème, un nouvel estimateur peut être construit en remplaçant la discrétisation à priori en intervalles B_j par un seul intervalle mais qui varie de manière continue. Concrètement, cela donne l'estimateur de la fenêtre mobile défini de la manière suivante :

$$\hat{r}_{FM}(x) = \frac{\sum_{i=1}^{n} Y_i \mathbf{1}(X_i \in [x-h; x+h])}{\sum_{i=1}^{n} \mathbf{1}(X_i \in [x-h; x+h])}, \quad \forall\, x, \tag{3.3}$$

où h est un paramètre réel strictement positif.

L'estimateur précédent présente encore le désavantage d'être discontinu par nature. Ainsi sa généralisation naturelle est l'estimateur à noyau, appelé aussi estimateur de Nadaraya- Watson, défini de la manière suivante :

$$\hat{r}_{NW}(x) = \begin{cases} \dfrac{\sum_{i=1}^{n} Y_i K(\frac{x-X_i}{h})}{\sum_{i=1}^{n} K(\frac{x-X_i}{h})}, & \text{si } \sum_{i=1}^{n} K(\frac{x-X_i}{h}) \neq 0\,; \\ & \qquad \forall x \in \mathbb{R} \\ \frac{1}{n}\sum_{i=1}^{n} Y_i, & \text{sinon.} \end{cases} \tag{3.4}$$

Dans cette définition K est une fonction de $\mathbb{R}$ dans $\mathbb{R}$ et h est un paramètre réel strictement positif (dont nous verrons qu'il sera intéressant de le faire dépendre de n). Il faut noter que tous les résultats obtenus pour l'estimateur $\hat{r}_{NW}$ resteront valables pour l'estimateur de la fenêtre mobile $\hat{r}_{FM}$ puisque ce dernier est un estimateur à noyau particulier correspondant au cas où K est le noyau uniforme

$$K(t) = \mathbf{1}(t \in [-1; +1]).$$

Notons que dans toutes ces définitions nous adoptons implicitement la convention 0/0=0.

3.3 Convergence presque complète

3.3.1 Résultats sous hypothèse de dérivabilité

Nous allons commencer par donner un résultat de convergence presque complète sous le modèle non-paramétrique (3.1). Cette notion de convergence presque complète entraîne à la fois la convergence presque sûr et la convergence en probabilité. Dans un premier temps nous nous plaçons en un point fixé x. Le modèle (3.1) est renforcé par la condition

$$r \text{ et } f \text{ sont } k \text{ fois continûment dérivables autour de } x, \tag{3.5}$$

et nous supposons en outre que

$$f(x) > 0, \tag{3.6}$$

f désignant la densité de X par rapport à la mesure de Lebesgue sur $\mathbb{R}$ (densité supposée exister). Nous verrons que nous aurons besoin des conditions suivantes sur le paramètre de lissage $h = h(n)$

$$\lim_{n\to\infty} h(n) = 0 \quad \text{et} \quad \lim_{n\to\infty} \frac{nh(n)}{\log n} = \infty. \tag{3.7}$$

Concernant la pondération K, nous supposerons que

$$K \text{ est borné, intégrable et à support compact.} \tag{3.8}$$

Nous verrons que lorsque $k > 0$ (voir la formule 3.1) il sera intéressant de rajouter l'hypothèse que K est un noyau d'ordre k au sens de Gasser et Müller (1979)[41], c'est à dire qu'il vérifie :

$$\int t^j K(t)dt = 0, \ \forall\ j = 1, \ \ldots, \ k-1 \ \text{ et } \ 0 < |\int t^k K(t)dt| < \infty. \tag{3.9}$$

Notons que cette condition d'ordre est, dès que $k > 2$, incompatible avec une hypothèse de positivité du noyau. Soit l'hypothèse suivante

$$|Y| < M < \infty, \tag{3.10}$$

hypothèse qui pourrait être allégée (au moyen par exemple d'une technique de troncation du type de celle introduite par Mack et Silverman (1982)[65] et reprise entre autres dans Györfi et al., 1989 [42], ou Bosq, 1996 [5]), mais au prix d'un accroissement sensible de la lourdeur des démonstrations. Dans la forme simplifiée sous laquelle nous le présentons ci-dessous, ce résultat est issu de Sarda et Vieu (2000)[83], mais les idées qui servent de base à cette démonstration reviennent à Collomb (1976)[19] et (1984) [21].

Théorème 3.3.1 Vitesse de convergence presque complète ponctuelle sous condition de dérivabilité. *(Sarda et Vieu)[83]*
Considérons le modèle (3.5) avec $k > 0$ et supposons que les conditions (3.6)-(3.10) soient réalisées. On a

$$\hat{r}_{NW}(x) - r(x) = O(h^k) + O(\sqrt{\frac{\log n}{nh}}), \ \textit{Presque Complète.} \tag{3.11}$$

Nous allons maintenant établir une version uniforme du résultat précédent. La démonstration de ces théorèmes s'inspire de Collomb (1976)[19] et (1984) [21]. Nous avons bien évidement besoin d'un version uniforme des hypothèses (3.6) et (3.5), et pour cela nous nous plaçons sur un compact S de $\mathbb{R}$, tel qu'il existe $\theta > 0$ pour lequel on a

$$r \text{ et } f \text{ sont } k \text{ fois continûment dérivables autour de } S, \tag{3.12}$$

et

$$\inf_{x\in S} f(x) > \theta. \tag{3.13}$$

Pour le reste, les conditions dont nous aurons besoin sont les mêmes que celles nécessitées pour l'obtention des résultats ponctuels ci-dessus, auxquelles s'ajoute la restriction suivante de type Lipschitz sur le noyau

$$\exists\ \beta > 0, \ \ \exists\ C < \infty, \ \ \forall\ x \in S, \ \ \forall\ y \in S, \ \ |K(x) - K(y)| \leq C|x - y|^\beta. \tag{3.14}$$

Théorème 3.3.2 Vitesse de convergence presque complète uniforme sous condition de dérivabilité. *(Sarda et Vieu)[83]*
Considérons le modèle (2.28) avec $k > 0$ et supposons que les conditions (2.7), (2.8), (2.9), (3.10), (2.29) et (2.30) soient réalisées. On a

$$\sup_{x\in S}|\hat{r}_{NW}(x) - r(x)| = O(h^k) + O(\sqrt{\frac{\log n}{nh}}),\ \textit{Presque Complète.} \tag{3.15}$$

3.3.2 Résultats sous hypothèse de continuité

Le prochain résultat est établi sous un modèle de régression plus général que celui du Théorème 3.3.1, puisque l'hypothèse d'existence de dérivées continues pour les fonctions que l'on estime (i.e. pour f et r) est remplacée par la condition de continuité. Le gain obtenu ainsi en terme de généralité du modèle statistique est à mettre en balance avec la perte des vitesses de convergence.

Théorème 3.3.3 Convergence presque complète ponctuelle sous condition de continuité. *(Ferraty et Vieu [39])*
Considérons le modèle (3.5) avec $k = 0$ et supposons que les conditions (3.6) , (3.7), (3.8) et (3.10) soient réalisées. On a

$$\hat{r}_{NW}(x) \to r(x),\ \textit{Presque Complète.} \tag{3.16}$$

Le prochain résultat est une version uniforme du théorème précédent. Il est établi sous un modèle de régression plus général que celui du Théorème 3.3.2, puisque l'hypothèse d'existence de dérivées continues pour les fonctions que l'on estiment (c'est-à-dire pour f et r) est remplacé par la condition de continuité. Le gain obtenu ainsi en terme de généralité du modèle statistique s'accompagne de la perte des vitesses de convergence.

Théorème 3.3.4 Convergence uniforme presque complète ponctuelle sous condition de continuité. *(Ferraty et Vieu [39])*
Considérons le modèle (3.12) avec $k = 0$ et supposons que les conditions (3.7), (3.8),(3.10), (3.13) et (3.14) soient réalisées. On a

$$\sup_{x\in S}|\hat{r}_{NW}(x) - r(x)| \to 0,\ \textit{Presque Complète.} \tag{3.17}$$

3.3.3 Résultats sous hypothèse de type Lipschitz

Mentionnons, pour terminer cet exposé des résultats de convergence presque complète, que d'autres modèles non-paramétriques, caractérisables par d'autres conditions de régularité sur r et f, peuvent être étudiés par des techniques similaires. C'est le cas en particulier lorsque les conditions (3.12) ou (3.5) sont remplacées par des condition de type Lipschitz. Nous donnons l'énoncé de ces résultats. Les conditions de Lipschitz sont soit des conditions ponctuelles en x fixé du type

$$\exists\ \beta > 0,\ \exists\ C < \infty,\ \exists\ \epsilon > 0,\ \forall\ y \in]x-\epsilon, x+\epsilon[,\ |\phi(x) - \phi(y)| \leq C|x-y|^\beta, \tag{3.18}$$

soit des conditions uniformes sur un compact S du type

$$\exists\ \beta > 0,\ \exists\ C < \infty,\ \forall\ x \in S,\ \forall\ y \in S,\ |\phi(x) - \phi(y)| \leq C|x-y|^\beta, \tag{3.19}$$

où ϕ désigne indifféremment f ou r.

Théorème 3.3.5 Vitesse de convergence presque complète ponctuelle sous condition de Lipschitz. *(Ferraty et Vieu [39])*
Considérons le modèle (3.18) et supposons que les conditions (3.6), (3.7),(3.8) et (3.10) soient réalisées. On a

$$\hat{r}_{NW}(x) - r(x) = O(h^{\beta}) + O(\sqrt{\frac{\log n}{nh}}), \ \textit{Presque Complète.} \tag{3.20}$$

Théorème 3.3.6 Vitesse de convergence presque complète uniforme sous condition de Lipschitz. *(Ferraty et Vieu [39])*
Considérons le modèle (3.19) et supposons que les conditions (3.7), (3.8),(3.9),(3.10),(3.13) et (3.14) soient réalisées. On a

$$\sup_{x \in S} |\hat{r}_{NW}(x) - r(x)| = O(h^{\beta}) + O(\sqrt{\frac{\log n}{nh}}), \ \textit{Presque Complète.} \tag{3.21}$$

3.4 Convergence en moyenne quadratique

3.4.1 Erreur quadratique en moyenne ponctuelle

Nous allons nous intéresser à des résultats asymptotiques en terme de convergence quadratique. La structure de ce paragraphe est similaire à celle du précédent. Les hypothèses dont nous aurons besoin sont sensiblement les mêmes que pour les résultats de convergence presque complète, à quelques exceptions prés décrites ci-dessous. Tout d'abord, la condition sur le paramètre de lissage peut être rendue légèrement moins restrictive en ce sens on peut remplacer (3.7) par

$$\lim_{n \to \infty} h(n) = 0 \text{ et } \lim_{n \to \infty} nh(n) = \infty. \tag{3.22}$$

Par ailleurs, afin de pouvoir spécifier les constantes intervenant dans nos développements asymptotiques nous aurons besoin parfois de rajouter la condition suivante sur la loi de (X, Y) :

$$\phi(u) = E(Y^2 \mid X = u) \text{ est continue au point } x \tag{3.23}$$

Pour les raisons techniques qui serons explicitées ultérieurement, nous aurons besoin de nous restreindre à des noyaux positifs. Or, cette condition est incompatible avec un noyau d'ordre supérieur à 2. Par conséquent, nous on resterons à un modèle de type (3.5) avec $k = 2$, et les conditions sur le noyau K deviennent alors :

$$K \text{est borné, intégrable, positif, symétrique et à support compact.} \tag{3.24}$$

Le premier résultat établit la vitesse de convergence en moyenne quadratique pour l'estimateur à noyau de la régression sous hypothèses de dérivabilité, en un point x fixé. La démonstration de ce résultat s'inspire de celle de Collomb (1976)[19].

Théorème 3.4.1 Convergence convergence en moyenne quadratique ponctuelle sous condition de dérivabilité. *(Ferraty et Vieu [39])*
Considérons le modèle (3.5) avec $k = 2$ et supposons que les conditions (3.6), (3.10),(3.22), (3.23) et (3.24) soient réalisées. On a

$$E[\hat{r}_{NW}(x) - r(x)]^2 = B^2(x)h^4 + V(x)\frac{1}{nh} + o(h^4 + \frac{1}{nh}), \tag{3.25}$$

où

$$B(x) = \frac{\int t^2 K(t)dt}{2} \frac{(g^{(2)}(x) - r(x)f^{(2)}(x))}{f(x)}, \tag{3.26}$$

et

$$V(x) = \int K^2(t)dt \frac{(\phi(x) - r^2(x))}{f(x)}. \tag{3.27}$$

Le prochain résultat est établi sous un modèle de régression plus général que celui du Théorème 3.4.1, puisque l'hypothèse d'existence de dérivées continues pour les fonctions que l'on estime (c'est-à-dire pour f et r) est remplacée par la simple condition de continuité. Comme toujours en pareil cas, le gain obtenu ainsi en terme de généralité du modèle statistique est à mettre en balance avec la perte des vitesses de convergence.

Théorème 3.4.2 Convergence en moyenne en moyenne quadratique ponctuelle sous condition de continuité. *(Ferraty et Vieu [39])*
Considérons le modèle (3.5) avec $k = 0$ et supposons que les conditions (3.6), (3.10),(3.22), (3.24) et (3.14) soient réalisées. On a

$$E[\hat{r}_{NW}(x) - r(x)]^2 \to 0. \tag{3.28}$$

3.4.2 Erreur quadratique moyenne intégrée

Nous allons maintenant des versions uniformes sur un compact des deux théorèmes précédents. Pour cela, nous allons nous intéresser aux erreurs quadratiques moyennes intégrées, notées $MISE$ et définies par :

$$MISE(\hat{r}_{NW}) = E\left(\int [\hat{r}_{NW}(x) - r(x)]^2 w(x)dx\right). \tag{3.29}$$

La fonction w est une fonction de poids vérifiant :

$$w \text{ est positive, bornée et à support compact } S. \tag{3.30}$$

Cette fonction est fixée à priori, et sera souvent dans la pratique prise égale à une indicatrice sur un intervalle borné de $\mathbb{R}$, ou bien égale à un produit d'une telle indicatrice par la densité f de X.Nous aurons aussi besoin d'une version uniforme de la condition (3.23), à savoir que

$$\phi(u) = E(Y^2|X = u) \text{ est continue autour de } S. \tag{3.31}$$

Théorème 3.4.3 Erreur quadratique moyenne intégrée sous condition de dérivabilité. *(Ferraty et Vieu [39])*
Considérons le modèle (3.12) avec $k = 2$ et supposons que les conditions (3.13), (3.22),(3.24), (3.30) et (3.31) sont réalisées. On a

$$MISE(\hat{r}_{NW}) = B^2h^4 + \frac{V}{nh} + o(h^4 + \frac{1}{nh}), \tag{3.32}$$

où

$$B^2 = \int B^2(x)w(x)dx \quad et \quad V = \int V(x)w(x)dx, \tag{3.33}$$

$B(x)$ et $V(x)$ étant définis par (3.26) et (3.27).

Pour terminer, nous allons donner un résultat analogue sous le modèle plus général $k = 0$.

Théorème 3.4.4 Erreur quadratique moyenne intégrée sous condition de continuité. *(Ferraty et Vieu [39])*
Considérons le modèle (3.12) avec $k = 0$ et supposons que les conditions (3.13), (3.22),(3.24) et (3.30) soient réalisées. On a

$$MISE(\hat{r}_{NW}) \to 0. \tag{3.34}$$

3.5 Choix du paramètre de lissage

Par construction, les estimateurs à noyau $\hat{r}_{NW}$ dépendent de deux paramètres : le noyau K et le fenêtre h. Dans la pratique, on aura besoin de décider quels choix effectuer pour ces deux paramètres. Au vu des résultats asymptotiques des deux paragraphes précédents, il est clair que le paramètre h, en contrôlant la régularité de la fonction estimée, jouera un rôle essentiel. Nous nous en tiendrons ici à quelques idées générales concernant l'influence de ce paramètre de lissage h et à la présentation de quelques méthodes permettant de le choisir de manière automatique.

3.5.1 Optimisation des vitesses de convergence

Tous les résultats de convergence pour lesquels nous avons été en mesure de spécifier les vitesses de convergence, c'est à dire les Théorèmes 3.3.1 ,3.3.2, 3.4.1 et 3.4.3, mettent en évidence le rôle du paramètre de lissage h. En regardant par exemple le résultat du Théorème 3.4.3, on constate que le terme de biais est le terme en h^4, tandis que le terme de variance est en $1/(nh)$. L'un est proportionnel à h tandis que l'autre est inversement proportionnel à h. Notons ici que l'existence d'un biais n'est pas liée à la méthode d'estimation par noyau, mais que c'est une chose inhérente au modèle non-prarmétrique lui-même (voir Collomb (1976) [19] ou Sarda et Vieu (2000)[83] pour une propriété d'inexistence d'estimateur non biaisé de la régression dans des contextes non-prarmétriques).

Ainsi, un grande valeur de h se traduit par un estimateur fortement biaisé et alors qu'une trop petite valeur de h entraîne un estimateur à forte variabilité. Comment faut-il donc choisir h ?

D'un point de vu théorique il est facile de répondre à cette à cette question. Concentrons nous par exemple sur le Théorème 3.4.3, et plus particulièrement sur son résultat :

$$MISE(\hat{r}_{NW}) = B^2h^4 + \frac{V}{nh} + o(h^4 + \frac{1}{nh}). \tag{3.35}$$

Il est aisé de minimiser (en h) les termes dominants de ce développement asymptotique , puisque il sagit d'une fonction convexe en h. Le minimum est atteint pour la valeur optimale h_{opt}

$$h_{opt} = \left(\frac{V}{4nB^2}\right)^{\frac{1}{5}}. \tag{3.36}$$

Le même type de raisonnement peut se faire à partir du résultat du Théorème 3.4.2, et on arrive immédiatement aux deux corollaires suivants.

Corollaire 3.1 Convergence optimale en moyenne quadratique ponctuelle. *(Ferraty et Vieu [39])*

Considérons le modèle (3.5) avec $k = 2$ et supposons que les conditions (3.6), (3.10),(3.23) et (3.24) soient réalisées. Supposons en outre que la fenêtre h soit de la forme :

$$h = Cn^{-\frac{1}{5}}, \quad 0 < C < \infty. \tag{3.37}$$

Alors on a

$$E[\hat{r}_{NW}(x) - r(x)]^2 = O(n^{-\frac{4}{5}}). \tag{3.38}$$

Corollaire 3.2 Erreur quadratique moyenne intégrée optimale. *(Ferraty et Vieu [39])*
Considérons le modèle (3.12) avec $k = 2$ et supposons que les conditions (3.13), (3.24),(3.30), (3.31) et (3.37) soient réalisées. On a

$$MISE(\hat{r}_{NW}) = o(n^{-\frac{4}{5}}). \tag{3.39}$$

Le même type d'argument peut être développé à partir des résultats de convergence presque complète donnés dans le Théorème 3.3.1 et 3.3.2. Il s'agit maintenant de minimiser en h des expressions du type

$$O(h^k) + O(\sqrt{\frac{\log n}{nh}}), \tag{3.40}$$

et naturellement la nouvelle hypothèse sur h pour atteindre asymptotiquement le minimum devient

$$h = C\left(\frac{n}{\log n}\right)^{-\frac{1}{2k+1}}, \quad 0 < C < \infty. \tag{3.41}$$

On obtient alors directement les deux résultats suivants comme conséquences directes des Théorèmes 3.3.1 et 3.3.2.

Corollaire 3.3 Convergence presque complète ponctuelle optimale.*(Ferraty et Vieu [39])*
Considérons le modèle (3.5) avec $k > 0$ et supposons que les conditions (3.6), (3.8),(3.9), (3.10) et (3.41) soient réalisées. On a

$$\hat{r}_{NW}(x) - r(x) = O\left((\frac{n}{\log n})^{-\frac{k}{2k+1}}\right), \ \textit{Presque Complète.} \tag{3.42}$$

Corollaire 3.4 Convergence presque complète uniforme optimale.*(Ferraty et Vieu [39])*
Considérons le modèle (3.12) avec $k > 0$ et supposons que les conditions (3.8), (3.9),(3.13), (3.14) et (3.41) soient réalisées. On a

$$\sup_{x \in S} |\hat{r}_{NW}(x) - r(x)| = O\left((\frac{n}{\log n})^{-\frac{k}{2k+1}}\right), \ \textit{Presque .Complète.} \tag{3.43}$$

Ces résultats sont particulièrement intéressant d'un point de vu théorique, puisque l'on sait d'après des résultats généraux de Stone (1981)[94] et (1982) [95] que la vitesse optimale de convergence pour un modèle (3.12) est :

$$n^{-\frac{k}{2k+1}}, \text{ en norme } L_p, \ p < \infty, \tag{3.44}$$

et

$$(\frac{n}{\log n})^{-\frac{k}{2k+1}}, \text{ en norme } L_\infty. \tag{3.45}$$

On a donc l'évidence de ce que les estimateurs à noyau, tout en restant d'une conception initiale relativement simple, atteignent moyennant un bon choix du paramètre de lissage des vitesses de convergence optimales pour des modèles non-paramétriques généraux.

3.5.2 Choix automatique de la fenêtre

Toutefois, ces résultats ne sont pas directement utilisables dans la pratique. Il existe des méthodes automatiques de sélection de ce paramètre. Nous allons décrire une de ces méthodes de choix de fenêtre. Pour bien insister sur le rôle de h nous adopterons pour la mesure d'erreur une notation qui fait apparaître explicitement ce paramètre :

$$MISE(\hat{r}_{NW}) = MISE(h)$$

La méthode que nous allons présenter est une des plus populaires tant sur le plan pratique que du point de vu des résultats asymptotiques dont on dispose à son sujet, et elle s'inspire des idées de validation croisée pour la sélection des modèles. Elle consiste à choisir comme largeur de fenêtre

$$h_{CV} = \arg\min_{h \in H} CV(h), \tag{3.46}$$

où

$$CV(h) = \sum_{i=1}^{n} [Y_i - \hat{r}_{NW}^{-i}(X_i)]^2 w(X_i). \tag{3.47}$$

Dans ces définitions, H est un ensemble de valeurs possibles pour h et r_{NW}^{-i} est l'estimateur à noyau construit à partir de l'échantillon privé de l'observation (X_i, Y_i) :

$$\hat{r}_{NW}^{-i} = \frac{\sum\limits_{j,j\neq i} Y_j K(\frac{x-X_j}{h})}{\sum\limits_{j,j\neq i} K(\frac{x-X_j}{h})}. \tag{3.48}$$

Nous admettons le résultat suivant, qui a été démontré par Hardle et Marron (1985) [54], et qui donne l'optimalité asymptotique de cette largeur de fenêtre en termes d'erreurs quadratiques (voir aussi Rice (1984)).

Théorème 3.5.1 Validation Croisée : optimalité asymptotique. *(Hardle et Marron [54]) Sous les hypothèses du Théorème 3.4.3, et si H ne contient que des largeurs de fenêtres vérifiant (2.95), alors on a la propriété suivante*

$$\frac{\inf_{h\in H} MISE(h)}{MISE(h_{CV})} \to 1, \ \textit{presque. sûre.} \tag{3.49}$$

3.6 Bootstrap dans l'estimation globale de la courbe de régression de la moyenne par la méthode du noyau

Soit $(x_1, y_1),\ (x_2, y_2), ...,\ (x_n, y_n)$ des observations indépendantes issues d'un modèle $r(x) = E(Y/x)$ inconnu. Considérons l'estimateur à noyau de Nadaraya-Watson de $r(x)$ Défini par 3.4.

Supposons que pour l'évaluation de l'efficacité de cette technique on utilise le critère de la minimisation de $MISE(h)$ par rapport à h qui est définie comme suite :

$$MISE(h) = \int E\left(r_h(x) - r(x)\right)^2 dx \tag{3.50}$$

Soit h_0 le paramètre qui minimise ce critère.

Pour calculer la valeur de la fenêtre par la technique de bootstrap on utilise la bootstrap des paires qui est adéquate à ce problème on doit rééchantillonner par cette technique à partir de l'échantillon initial (X, Y) et construisant ensuite l'estimateur de bootstrap sous la forme :

$$\hat{r}_h^j(x) = \begin{cases} \frac{\sum_{i=1}^n Y_i^j K(\frac{x-X_i^j}{h})}{\sum_{i=1}^n K(\frac{x-X_i^j}{h})}, & \text{si } \sum_{i=1}^n K(\frac{x-X_i^j}{h}) \neq 0\,; \\ & \forall x \in \mathbb{R} \text{ et pour } j = 1, 2, ..., B; \\ \frac{1}{n}\sum_{i=1}^n Y_i^j, & \text{sinon.} \end{cases} \tag{3.51}$$

où B est le nombre de réplications de bootstrap.

L'estimateur de la variance de $\hat{r}_h^j(x)$ dans ce cas est donné sous la forme :

$$\frac{1}{B}\sum_{j=1}^{B}\int \left(\hat{r}_h^j(x) - \bar{r}_h^*(x)\right)^2,$$

avec,

$$\bar{r}_h^*(x) = \frac{1}{B}\sum_{j=1}^{B}\hat{r}_h^j(x).$$

Pour estimer maintenant h par la technique de bootstrap nous proposons deux procédures :

1. **Première approche :**

   Nous construisons un estimateur initial de la courbe de régression de la moyenne $\hat{r}_{h_0}$ ensuite rééchantillonné par la technique de bootstrap des paires à partir de l'échantillon initial pour construire les estimateurs $\hat{r}_h^j(x)$ (pour j=1,...,B.).

   Enfin, nous obtenons la fenêtre bootstrapée $\hat{h}_b$ par la minimisation de $BMISE(h, h_0)$ sous h qui s'écrit sous la forme :

   $$BMISE(h, h_0) = \frac{1}{B}\sum_{j=1}^{B}\int \left(\hat{r}_h^j(x) - \hat{r}_{h_0}(x)\right)^2 dx. \tag{3.52}$$

   Remarque 3.1 *On peut définir un autre h_b par :*

   $$h_b = \arg\min_h \left(\int \left(\frac{1}{B}\sum_{j=1}^{B}\hat{r}_h^j(x) - \hat{r}_{h_0}(x)\right)^2 dx\right), \tag{3.53}$$

   si on suppose qu'on s'intéresse à la statistique $\hat{r}_h(x)$.

2. **Deuxième approche**

   La deuxième approche que nous appliquerons dans la partie simulation consiste à se focaliser sur le paramètre h qui minimise le MISE. Pour cela on doit rééchantillonner à partir de l'échantillon initial pour construire les estimateurs $h^j (j = 1, ..., B)$ qui sont définis comme suite :

   $$h^j = \arg\min_h MISE^j, \text{ pour } j = 1, ..., B; \tag{3.54}$$

   avec

   $$MISE_h^j = \int (\hat{r}_h(x^j) - r(x))^2 dx, \text{ pour } j = 1, ..., B. \tag{3.55}$$

   Enfin, nous obtenons la fenêtre bootstrapée $\hat{h}_b$ par la moyenne empirique des h^j $(j = 1, ..., B)$ c'est-à-dire :

   $$\hat{h}_b = \frac{1}{B}\sum_{j=1}^{B} h^j. \tag{3.56}$$

3.7 Simulations et résultats

3.7.1 Introduction

Dans la section 2.8, les résultats de simulation ont été exposés pour étudier l'influence de la technique bootstrap sur la qualité de l'estimateur à noyau de la densité de probabilité. Dans cette section, d'autres résultats vont être exposés, dans le but d'étudier l'influence de la technique bootstrap sur la qualité de l'estimateur à noyau de la courbe de régression de la moyenne au sens d'AMISE. A cet effet, nous avons organisé les étapes de simulation comme suite :

Etape 1 : Générer un échantillon (x_1, y_1),...,(x_n, y_n).

Etape 2 : Construire B-échantillons Bootstrap.

Etape 3 : Estimer h^* et h_{boot} par la méthode de sélection automatique.

Etape 4 : Estimer $r(x, h^*)$ et $r_{boot}(x, h_{boot})$.

Etape 5 : Calculer l'AMISE et l'AMISE$_{boot}$.

Enfin, les exécutions des simulateurs implémentées sous le logiciel R ont été réalisés sur le modèle :

$$y = r(X) + Z, \tag{3.57}$$

tel que :

- $$r(x) = (x+2)e^{-(x-0.5)^2}, \text{(modèle cible)} \tag{3.58}$$
- $X \rightsquigarrow u[0,1]$,
- $Z \rightsquigarrow N(0,1)$,

et pour :
– $n \in \{\ 200, 500,\ 1000,\ 2500,\ 5000,\ 10000\}$.
– $B \in \{5,\ 10,\ 25,\ 50,\ 100\}$,
– K est le noyau gaussien.
– h est estimé par le h^* qui minimise la fonction $CV(h)$ (Cross Validation) définie dans (3.47) .
– h_{boot} estimé par la technique de bootstrap (deuxième approche).

3.7.2 Résultats

n	B	h^*	h_{boot}	$AMISE^*$	$AMISE_{boot}$
200	5	0.1642978	0.0455613	0.2625574	1.7043353
	10	0.1664071	0.0458965	0.4221203	2.4403657
	25	0.1635888	0.0465411	0.2760679	1.0806143
	50	0.1640553	0.0460406	0.2864445	1.6084159
	100	0.1637276	0.0451857	0.4469901	0.9957202
500	5	0.1417100	0.0192890	0.1260718	2.1819843
	10	0.1439495	0.0171721	0.3215928	1.8174972
	25	0.1438466	0.0186703	0.1630689	1.5632174
	50	0.1442689	0.0186729	0.1836023	2.1184196
	100	0.1461801	0.0196747	0.3490090	2.0034755

1000	5	0.1288165	0.0077113	0.1674330	2.2230053
	10	0.1248246	0.0082700	0.1491134	2.4337444
	25	0.1280608	0.0080679	0.1489466	2.0632299
	50	0.1283876	0.0081826	0.2441851	2.1102929
	100	0.1243190	0.0080740	0.1622123	1.5763621
2500	5	0.1088351	0.0039850	0.0570194	1.3916763
	10	0.1087504	0.0039578	0.0931493	1.6880228
	25	0.1092028	0.0040117	0.0673349	1.8100528
	50	0.1097722	0.0040019	0.0860505	1.8872155
	100	0.1095671	0.0040143	0.2791980	1.8824809
5000	5	0.0963329	0.0026055	0.0181784	1.2876793
	10	0.0958822	0.0026510	0.0689601	1.4245756
	25	0.0958175	0.0026468	0.0310401	1.3469604
	50	0.0960417	0.0026557	0.0664974	1.1746108
	100	0.0968828	0.0026291	0.0783425	1.2796163
10000	5	0.0855545	0.0223901	0.0018343	0.9739984
	10	0.0847722	0.0018647	0.0174025	0.9071141
	25	0.0849663	0.0018545	0.0183297	0.8763342
	50	0.0852938	0.0018480	0.0153036	0.8466311
	100	0.0849522	0.0018524	0.0211555	0.9704672

TAB. 3.1: *Résultats des simulations pour déterminer* h^*, h_{boot}, $Amise^*$ *et* $Amise_{boot}$.

Performance des estimateurs :
On constate que les $AMISE_{boot}$ sont toujours supérieurs aux $AMISE^*$ et ceci quelque soit la taille de l'échantillon et le nombre de réplications de bootstrap B. Si on regarde les graphes (graphes fig.5, fig.6, fig.7, fig.8 et fig.9) de l'annexe B, on constate le phénomène de sous-lissage, c'est-à-dire que la technique de bootstrap à échouée. L'échec de cette technique s'explique par la présence d'un biais lors de l'estimation de $r_h(x)$.

Pour remédier à ce problème, on va introduire la correction du biais. L'idée est de remplacer le paramètre de lissage h_{boot} par le h^c_{boot}, qui représente le paramètre de lissage avec correction du biais c'est-à-dire $h^c_{boot} = 2 * h^* - h_{boot}$. Les résultats obtenus sont résumés dans le tableau 3.2.

n	B	h^*	h_{boot}	$AMISE^*$	$AMISE_{boot}$
200	5	0.1660958	0.2855115	0.7155006	0.4951177
	10	0.1636814	0.2791836	1.2387709	0.4853049
	25	0.1616513	0.2778727	0.5531028	0.5357193
	50	0.1675859	0.2888786	0.4644378	0.3570459
	100	0.1638852	0.2815107	0.7126442	0.6851429
500	5	0.1426174	0.2659990	0.1721569	0.4918643
	10	0.1403445	0.2620462	0.1436550	0.3666903
	25	0.1418515	0.2661854	0.1535001	0.3628087
	50	0.1423945	0.2664653	0.1814940	0.4022427
	100	0.1408344	0.2629612	0.1093148	0.3009964

1000	5	0.1301110	0.2520517	0.1027691	0.2021768
	10	0.1287761	0.2497039	0.1270428	0.2375196
	25	0.1248891	0.2417781	0.1911127	0.3971932
	50	0.1291257	0.2501537	0.0985449	0.2909052
	100	0.1287549	0.2494822	0.2030186	0.3035376
2500	5	0.1116084	0.2191609	0.0655949	0.1823262
	10	0.1096591	0.2153304	0.0888674	0.2457290
	25	0.1099558	0.2158608	0.0554455	0.1284211
	50	0.1074324	0.2108814	0.0620819	0.1711894
	100	0.1095072	0.2150558	0.1541611	0.2436490
5000	5	0.0971558	0.1916673	0.0317258	0.1108211
	10	0.0972866	0.1919667	0.0354423	0.1032000
	25	0.0973665	0.1921024	0.0725497	0.1237357
	50	0.0964608	0.1902933	0.0574877	0.0609590
	100	0.0968242	0.1909992	0.0624357	0.1476281
10000	5	0.08568681	0.16951435	0.02029660	0.07005034
	10	0.08604258	0.17020534	0.01981338	0.04978220
	25	0.08532017	0.16877209	0.02608287	0.03701050
	50	0.08546692	0.16908856	0.02312871	0.07274089
	100	0.08460276	0.16734449	0.00897619	0.02443307

TAB. 3.2: *Résultats des simulations pour déterminer* h^*, h^c_{boot}, $Amise^*$ *et* $Amise^c_{boot}$.

Performance des estimateurs :
Les résultats obtenus (tableau 3.2) montrent que la correction du biais par la technique de bootstrap dépend de la taille de l'échantillon n. En effet, on constate que pour n=200, l'$AMISE^c_{boot}$ est toujours inférieur à $AMISE^*$, quelque soit le nombre de réplications. Par contre pour n$\in \{500, 1000, 2500, 5000, 10000\}$ l'$AMISE^c_{boot}$ devient supérieur à l'$AMISE^*$. Cela est dû à la correction du biais qui a engendré une augmentation de la variance qui génère automatiquement une augmentation du $AMISE$. Cette constatation peut se confirmer sur les graphes fig.10, fig.11, fig.12, fig.13, fig.14 et fig.15 de l'annexe B, où l'on constate à partir de $n = 500$, un phénomène de sur-lissage.

3.8 Conclusion

Dans ce chapitre, nous avons présenté l'estimateur à noyau de la courbe de régression de la moyenne ainsi que ses différentes propriétés asymptotiques. Nous avons également présenté deux méthodes de sélection du paramètre de lissage h.

La technique de bootstrap classique dans l'estimation de la courbe de régression de la moyenne par la méthode du noyau nous donne des paramètres de lissage assez petits par rapport au paramètre de lissage optimal, ce qui engendre un phénomène de sous lissage comme on le constate sur les graphes fig.5, fig.6, fig.7, fig.8 et fig.9 de l'annexe B obtenus par simulation. La correction du biais par la technique de bootstrap donne de bons résultats ($AMISE^c_{boot}$ inférieur $AMISE^*$) pour des échantillons de taille n = 200 indépendamment du nombre de réplications B. Cependant, quand n est égale à : 500, 1000, 2500, 5000 et 10 000 $AMISE^c_{boot}$ devient supérieur à $AMISE^*$ et le paramètre de lissage h^c_{boot} est toujours supérieur aussi au paramètre

de lissage optimal, ce qui engendre un phénomène de sur-lissage comme on le constate sur les graphes fig.10, fig.11, fig.12, fig.13, fig.14 et fig.15 de l'annexe B obtenus par simulation.

Chapitre 4

Conclusion

Ce travail est une contribution au problème du choix du paramètre de lissage lors de l'estimation de la densité de probabilité et la courbe de régression de la moyenne par la méthode du noyau. Nous avons utilisé la technique de bootstrap pour estimer le paramètre de lissage et nous avons étudié l'efficacité, la robustesse et la sensitivité temporaire de cette technique pour différentes méthodes de sélection du paramètre du lissage et différents noyaux. Ces performances ont été mesurées numériquement par une étude de simulation.

Ce travail est composé de deux parties principales :

Dans la première partie, après avoir exposé la technique de bootstrap afin d'illustrer son principe, son applicabilité en estimation et les différentes approches de détermination des intervalles de confiance. Nous avons présenté l'estimateur et ses propriétés de la densité de probabilité par la méthode du noyau de Parzen-Rosenblatt. Nous avons terminé cette partie par une étude de simulation dont l'objectif est de tester l'application de la technique de bootstrap dans la sélection du paramètre de lissage. Nous avons simulé des densités de probabilité tests présentant différents aspects (loi normale, loi exponentielle, loi khi-deux). Les résultats obtenus montrent que :

- La technique de bootstrap classique améliore les performances de l'estimateur à noyau de la loi normale au sens du AMISE pour les méthodes de sélection du paramètre de lissage : UCV, SJ, NRD ;

- La technique de bootstrap échoue au sens d'AMISE dans l'estimation de la densité de probabilité de la loi normale pour la méthode de sélection BCV (Biaised Cross Validation), a cause du biais important dans cette méthode ;

- A cause de la présence d'un biais important au voisinage de zéro, la technique de bootstrap échoue au sens du AMISE dans l'estimation des densités à queues lourdes comme la loi exponentielle, ou à convergence lente comme la loi de khi-deux ;

- l'étude montre aussi que le choix du noyau est important pour les densités cibles à support compact comme la loi exponentielle, car il influe sur les performances de l'estimateur. En effet, on obtient un gain plus de 90% de la valeur de l'AMISE lorsque on utilise le noyau gamma (version 2) au lieu du noyau gaussien ;

- Les résultats de simulation confirment l'optimalité du noyau Epanechnikov pour la loi normale ;
- Le choix du noyau dans l'estimation de la densité de la loi de khi-deux n'est pas un problème, mais les résultats sont meilleures lorsque on choisit le noyau Epanechnikov et la méthode SJ pour la sélection du paramètre de lissage.

La deuxième partie est consacrée à l'estimation de la courbe de régression de la moyenne. Nous avons entamé cette partie par une brève représentation de l'estimateur à noyau de la courbe de régression de la moyenne, ses propriétés statistiques et deux méthodes de sélection du paramètre de lissage (h_{CV} et h_{opt}). Une étude de simulation similaire à celle de la première partie a été également réalisée. L'objectif est d'étudier l'intérêt de la technique bootstrap dans l'estimation du paramètre de lissage qui intervient dans l'estimation de la courbe de régression de la moyenne. Les résultats obtenus par simulation pour un modèle test montrent que :

- La technique de bootstrap classique échoue dans l'estimation de la courbe de régression de la moyenne. Elle donne un paramètre trop petit par rapport au paramètre de lissage optimal, ce qui engendre un phénomène de sous-lissage ;

- Une correction du biais, permet à la technique bootstrap de donner de meilleures résultats pour des échantillons de taille $n = 200$. Cependant pour n supérieur à 200, la correction du biais a engendrée une augmentation de la variance, ce qui a généré automatiquement une augmentation du AMISE.

Parmi les perspectives de ce travail, nous pouvons dégager plusieurs axes intéressants, tant sur le plan théorique que pratique :

1. L'échec de la bootstrap ne signifie guère la non applicabilité de cette technique dans l'estimation de la densité de probabilité ou la courbe de régression de la moyenne. Il sera intéressant de revoir ce travail en utilisant la bootstrap lissée, ou même d'envisager des modifications de la bootstrap classique afin d'améliorer les performances des estimateurs ;

2. Il serait intéressant d'effectuer des simulations extensives :
   - Sur des cibles plus complexes, comme par exemple les densités qui possèdent plusieurs modes, ou des densités qui possèdent des discontinuités (par exemple un mélange de lois uniformes) ;
   - Sur d'autres modèles cibles et d'autres méthodes de sélection du paramètre de lissage pour l'estimation de la courbe de régression de la moyenne.
3. Appliquer la technique bootstrap aux :
   - Autres méthodes d'estimation non paramétrique de la densité de probabilité ou de la courbe de régression de la moyenne.
   - Estimateurs à noyau tel que les modèles de prévision des séries temporelles ou à l'estimation du taux de hasard ...
4. Développer d'autres résultats théoriques :
   - Concernant la détermination du nombre de réplications de bootstrap nécessaires pour que le MISE de bootstrap soit inférieur au MISE classique.
   - Pour déterminer le paramètre de lissage optimal qui correspond aux noyaux différents du noyau gaussien.

Annexe A

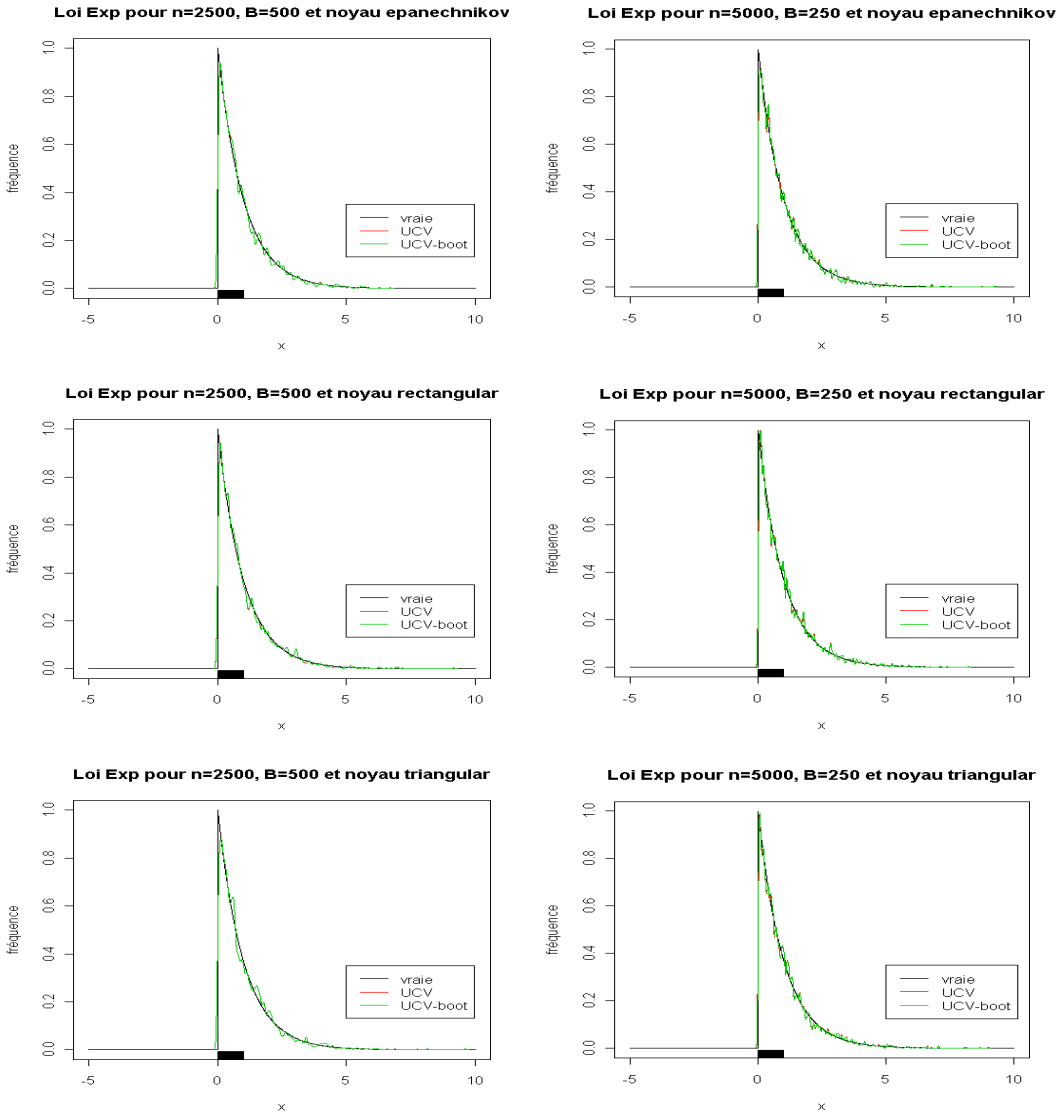


FIG. 1: *Comparaison entre la densité théorique et celle estimée avec le paramètre de lissage h^* et celle estimée avec h^c_{boot}*

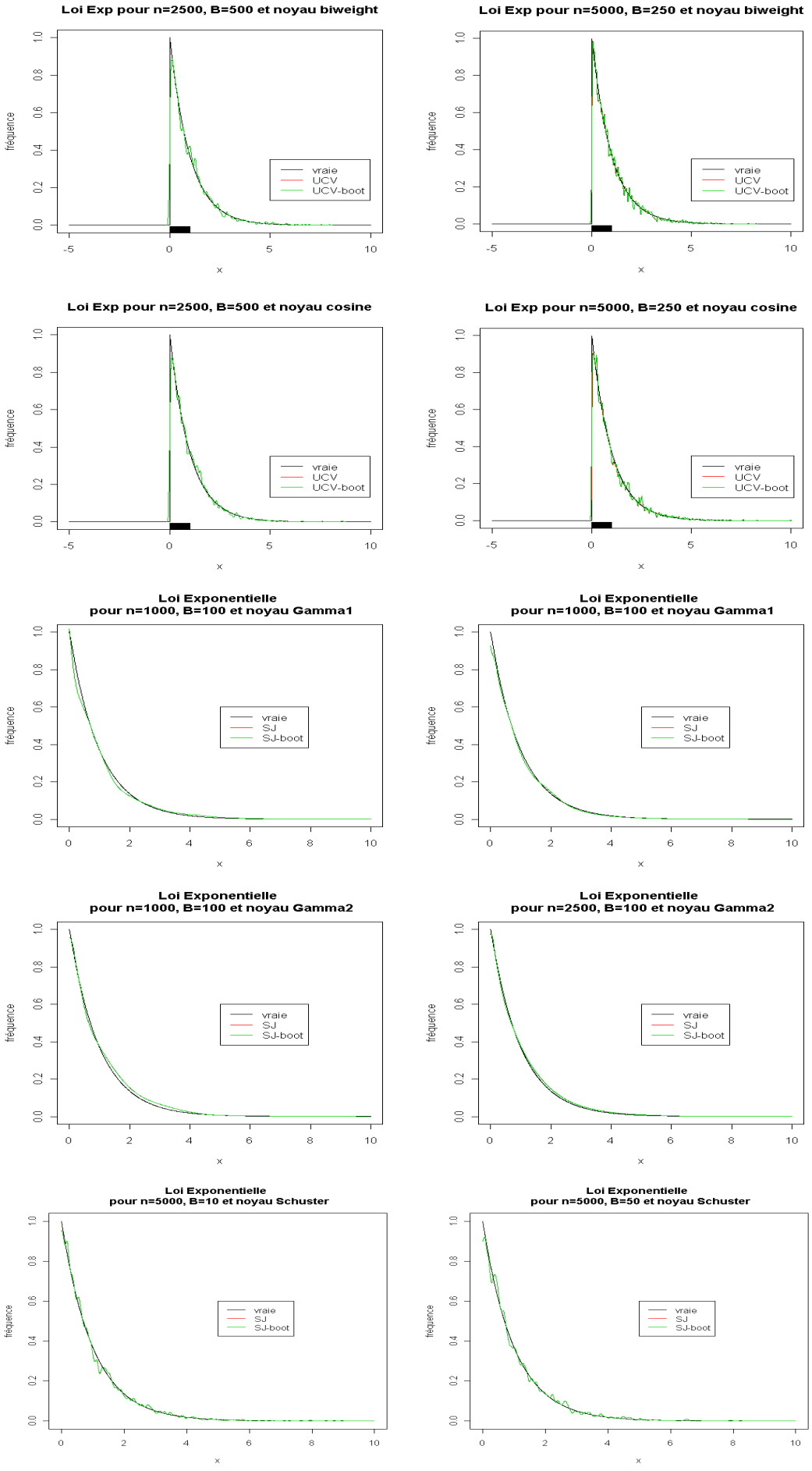


FIG. 2: *Comparaison entre la densité théorique et celle estimée avec le paramètre de lissage h^* et celle estimée avec h^c_{boot}*

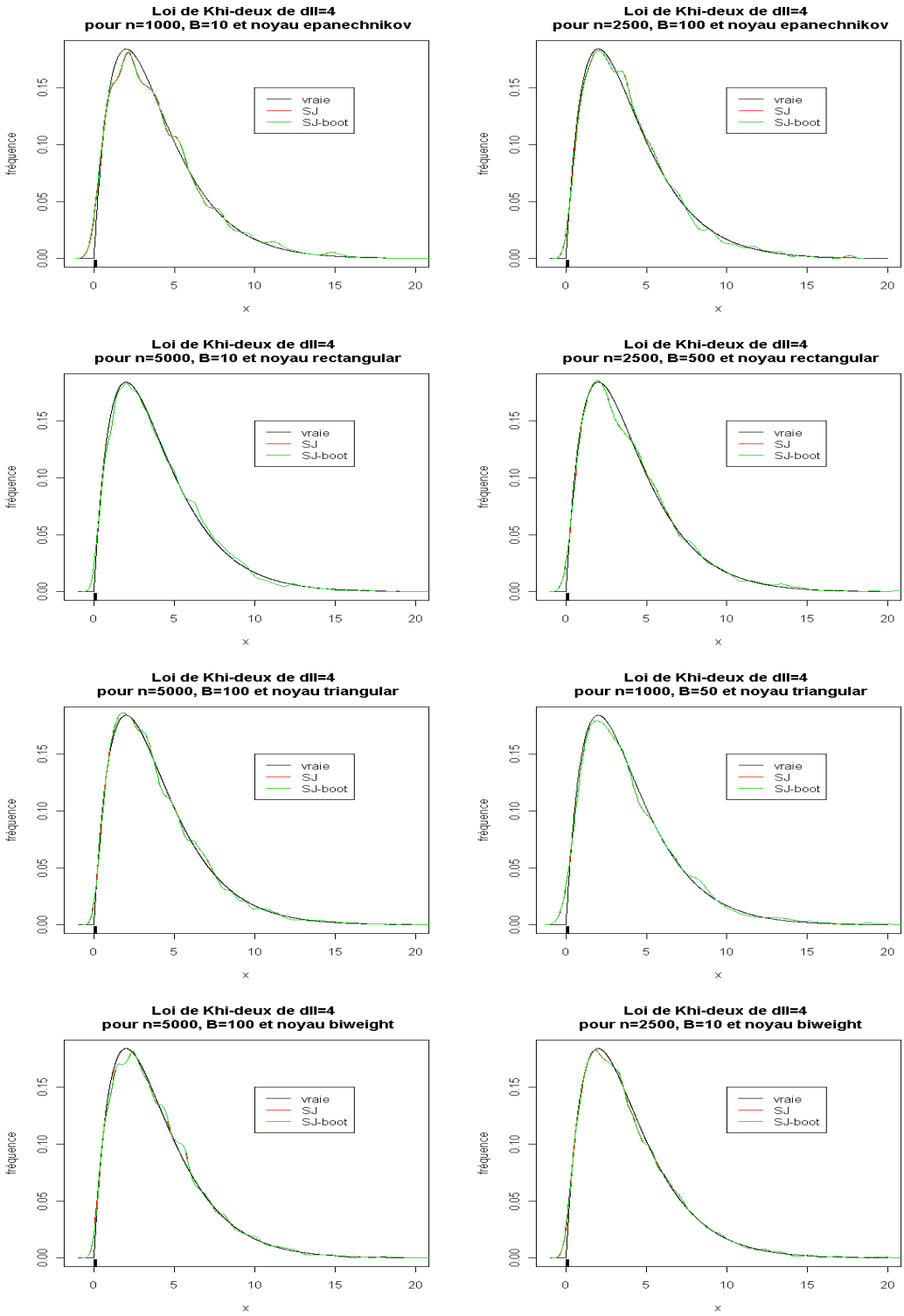


FIG. 3: *Comparaison entre la densité théorique et celle estimée avec le paramètre de lissage h^* et celle estimée avec h^c_{boot}*

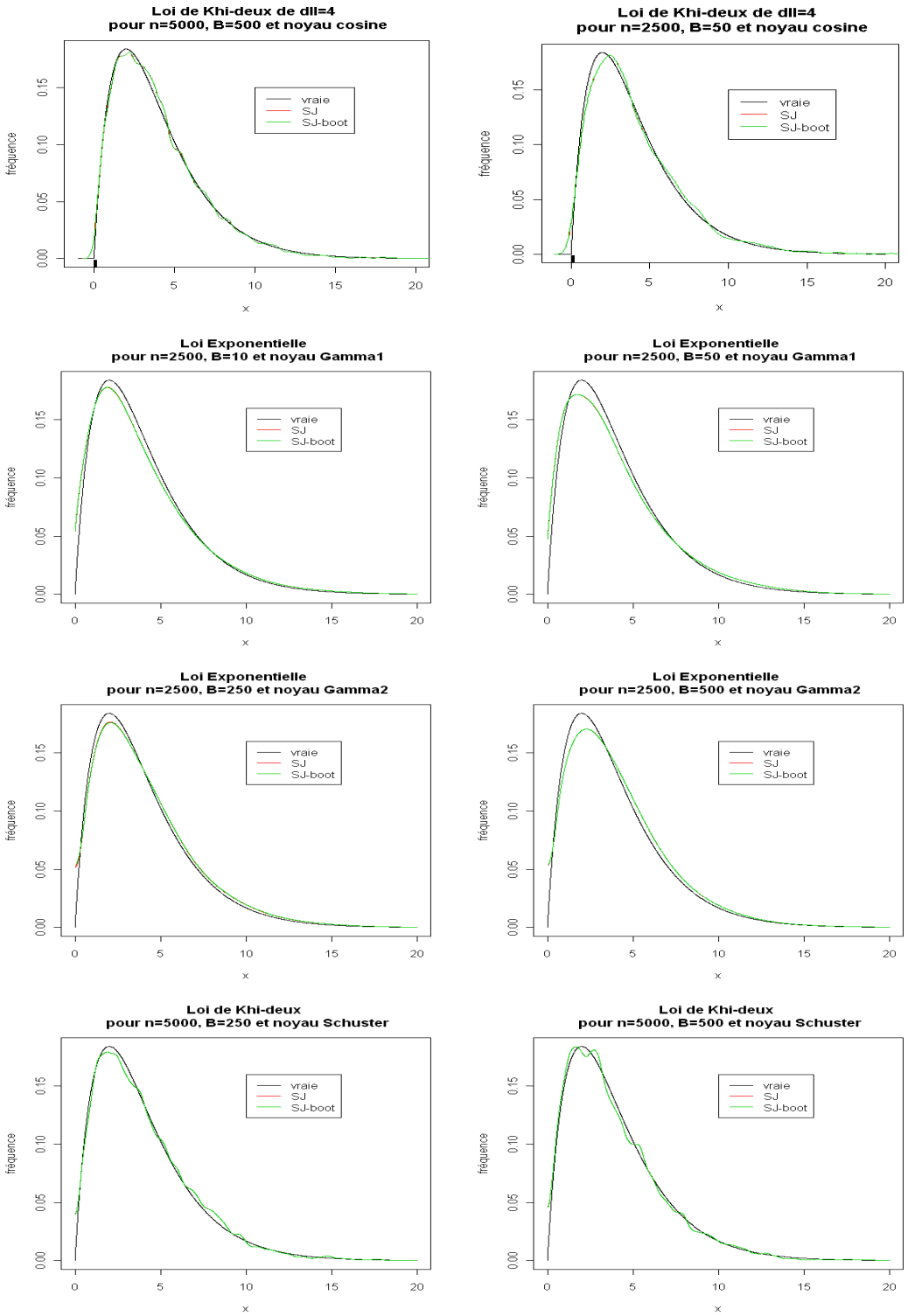


FIG. 4: *Comparaison entre la densité théorique et celle estimée avec le paramètre de lissage h^* et celle estimée avec h^c_{boot}*

Annexe B

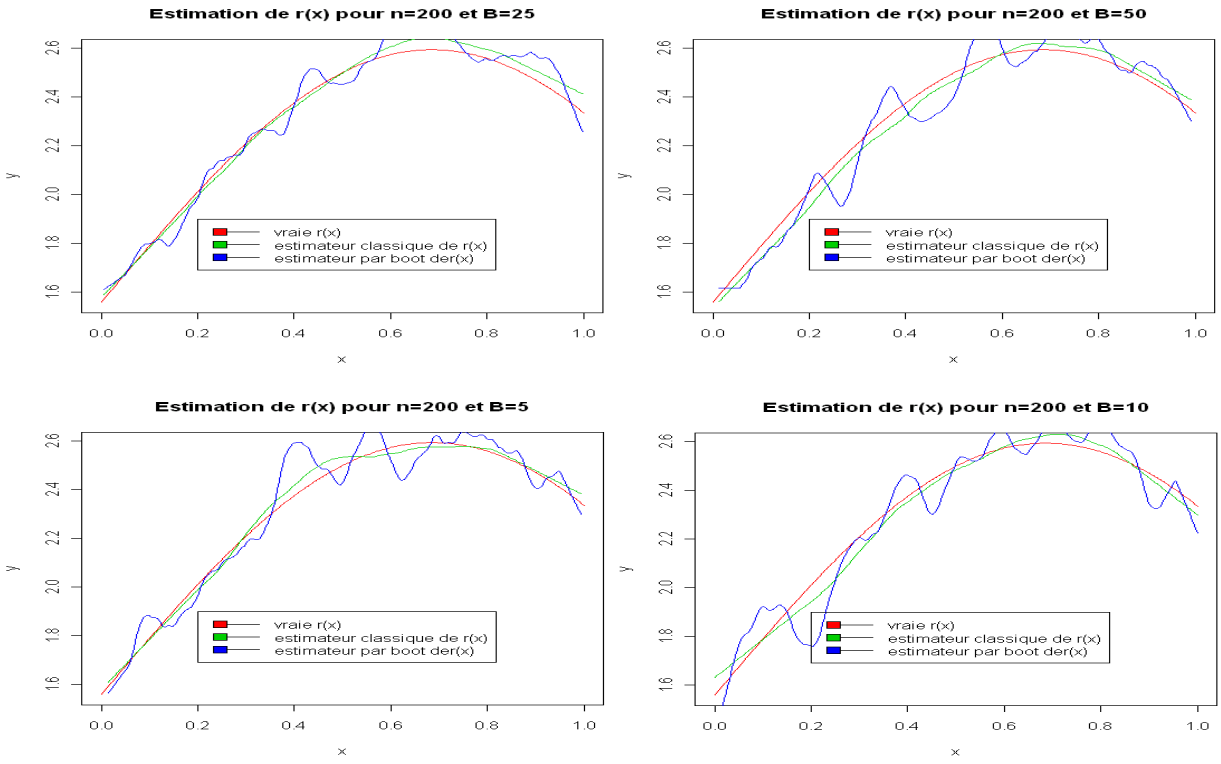


FIG. 5: *Comparaison entre la courbe théorique et celle estimée avec le paramètre de lissage h^* et celle estimée avec h_{boot}*

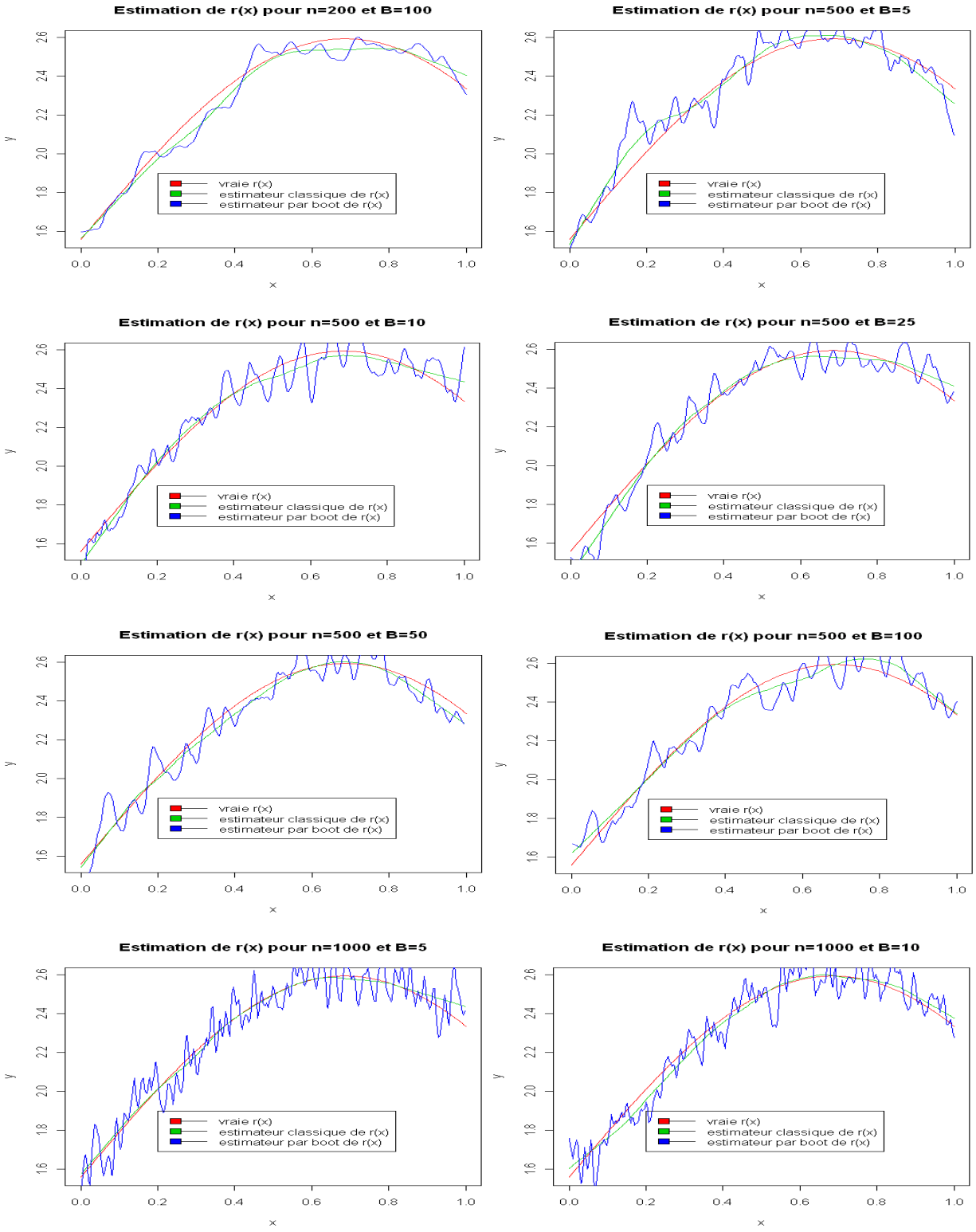


FIG. 6: *Comparaison entre la courbe théorique et celle estimée avec le paramètre de lissage h^* et celle estimée avec h_{boot}*

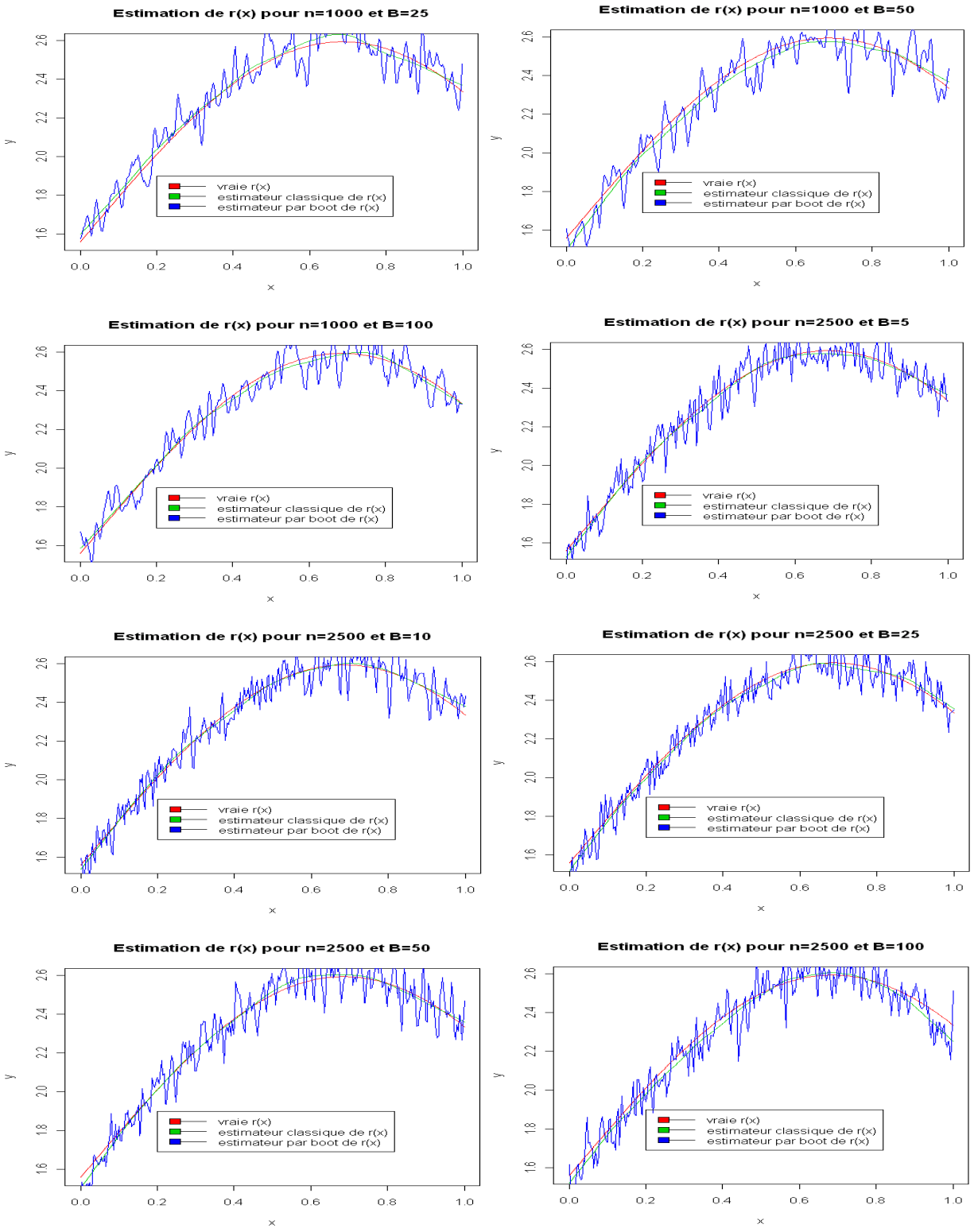


FIG. 7: *Comparaison entre la courbe théorique et celle estimée avec le paramètre de lissage h^* et celle estimée avec h_{boot}*

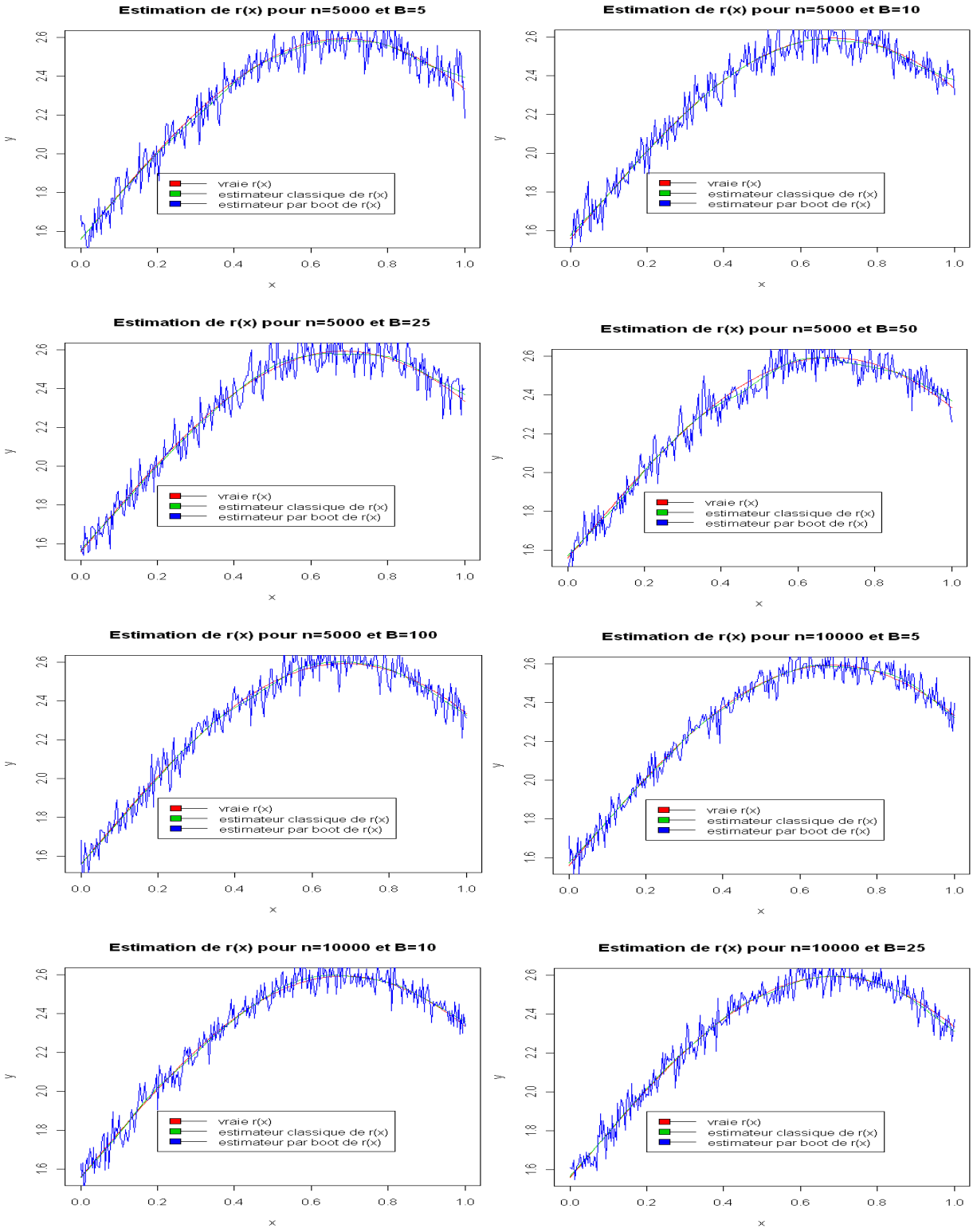


FIG. 8: *Comparaison entre la courbe théorique et celle estimée avec le paramètre de lissage h^* et celle estimée avec h_{boot}*

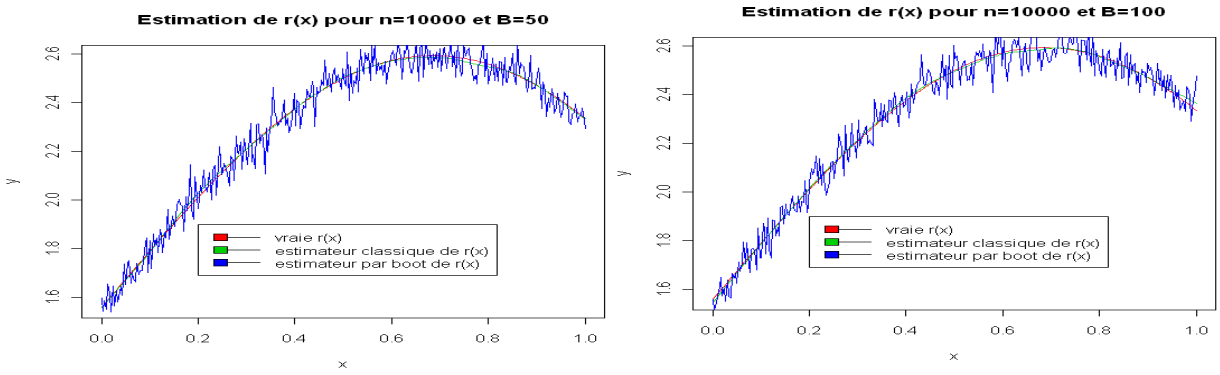


FIG. 9: *Comparaison entre la courbe théorique et celle estimée avec le paramètre de lissage h^* et celle estimée avec h_{boot}*

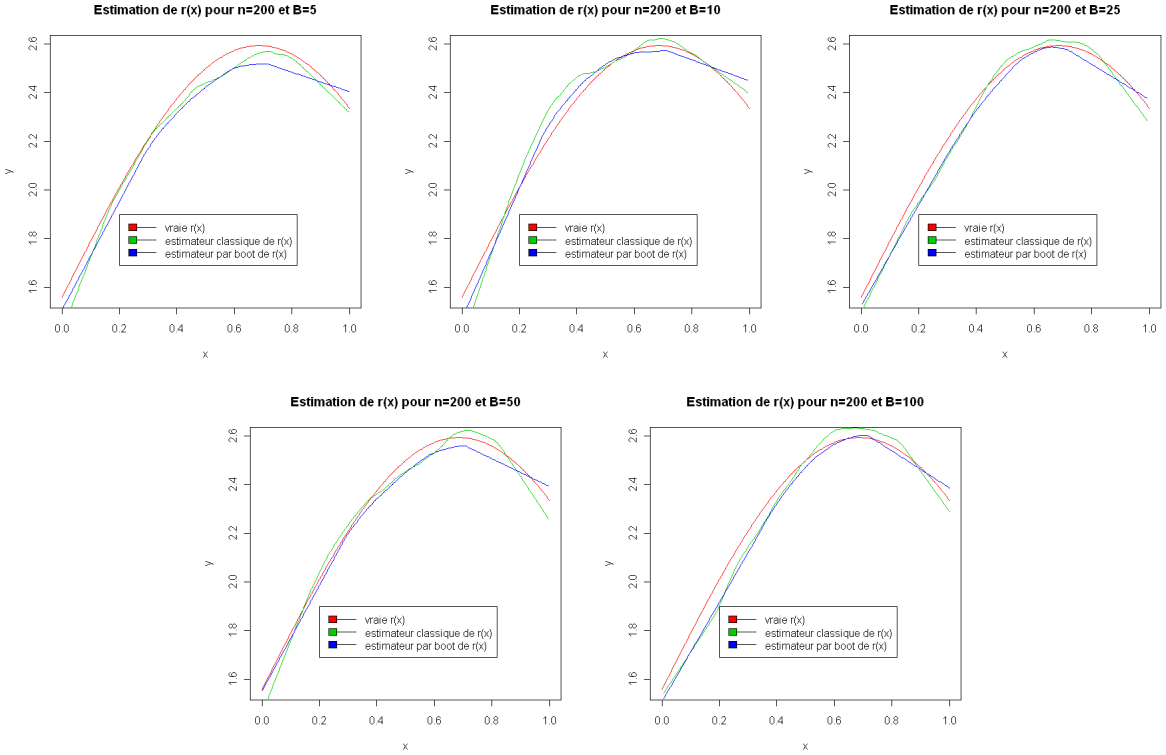


FIG. 10: *Comparaison entre la courbe théorique et celle estimée avec le paramètre de lissage h^* et celle estimée avec h^c_{boot}*

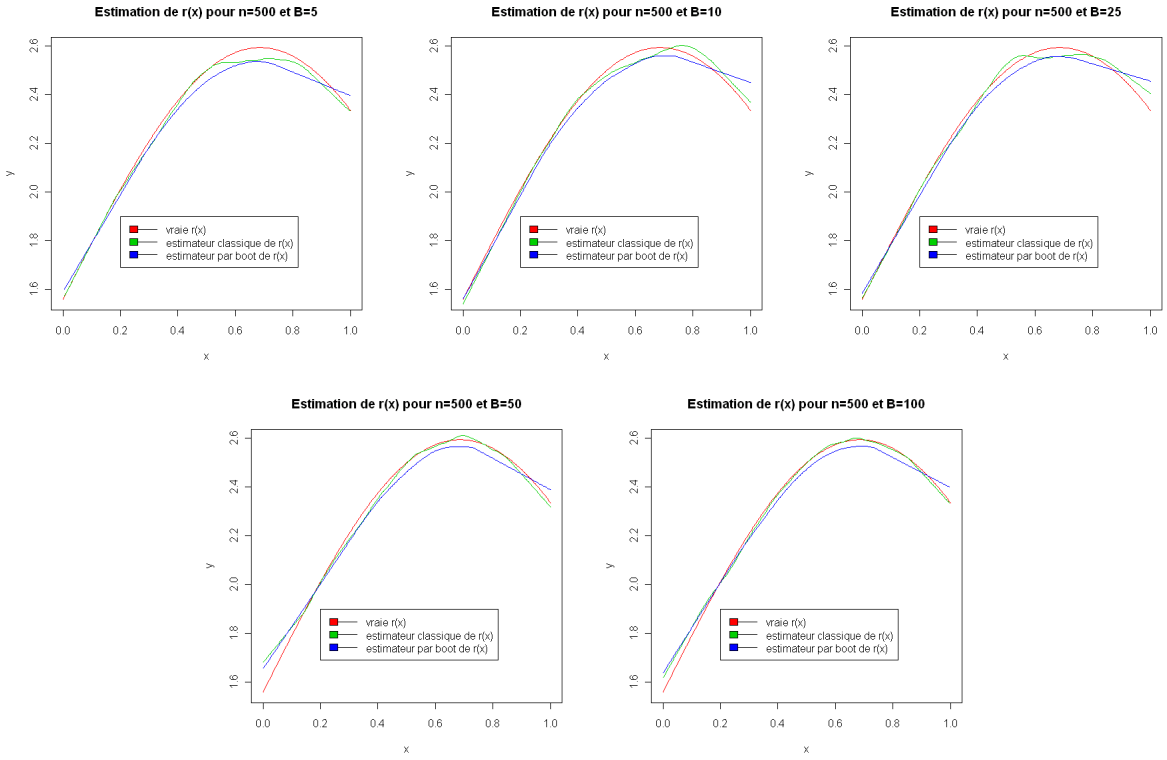


FIG. 11: *Comparaison entre la courbe théorique et celle estimée avec le paramètre de lissage h^* et celle estimée avec h^c_{boot}*

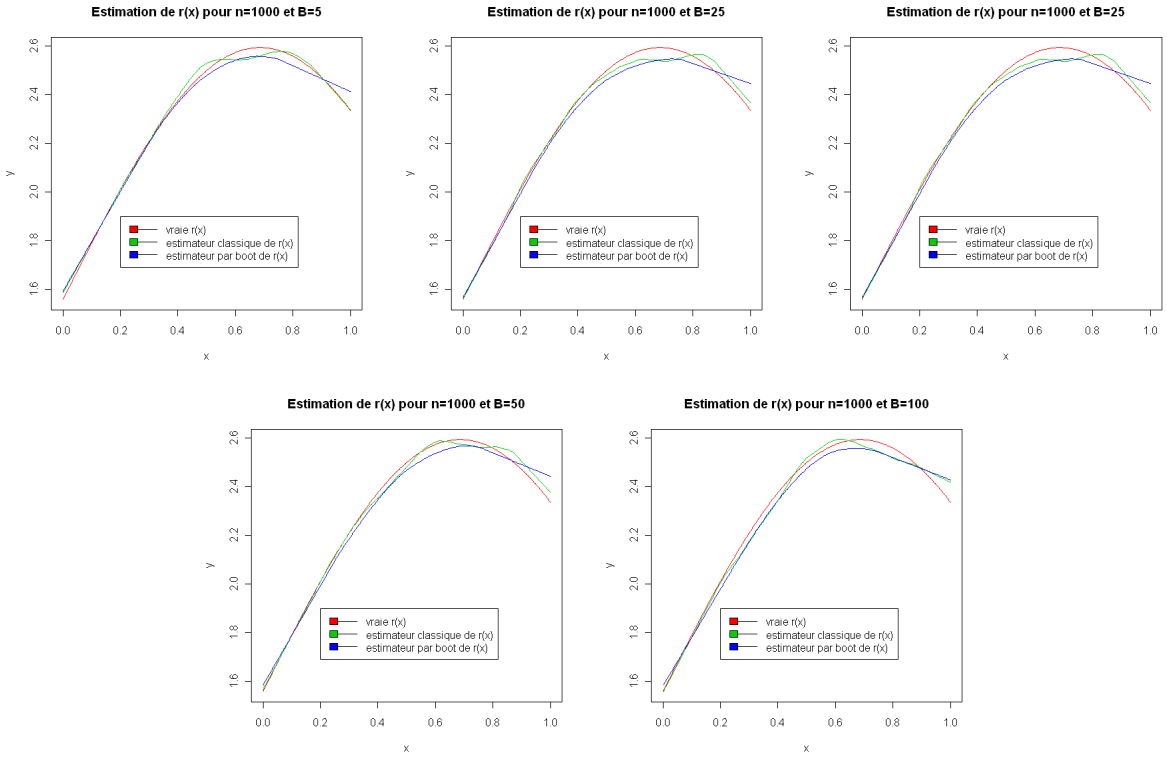


FIG. 12: *Comparaison entre la courbe théorique et celle estimée avec le paramètre de lissage h^* et celle estimée avec h^c_{boot}*

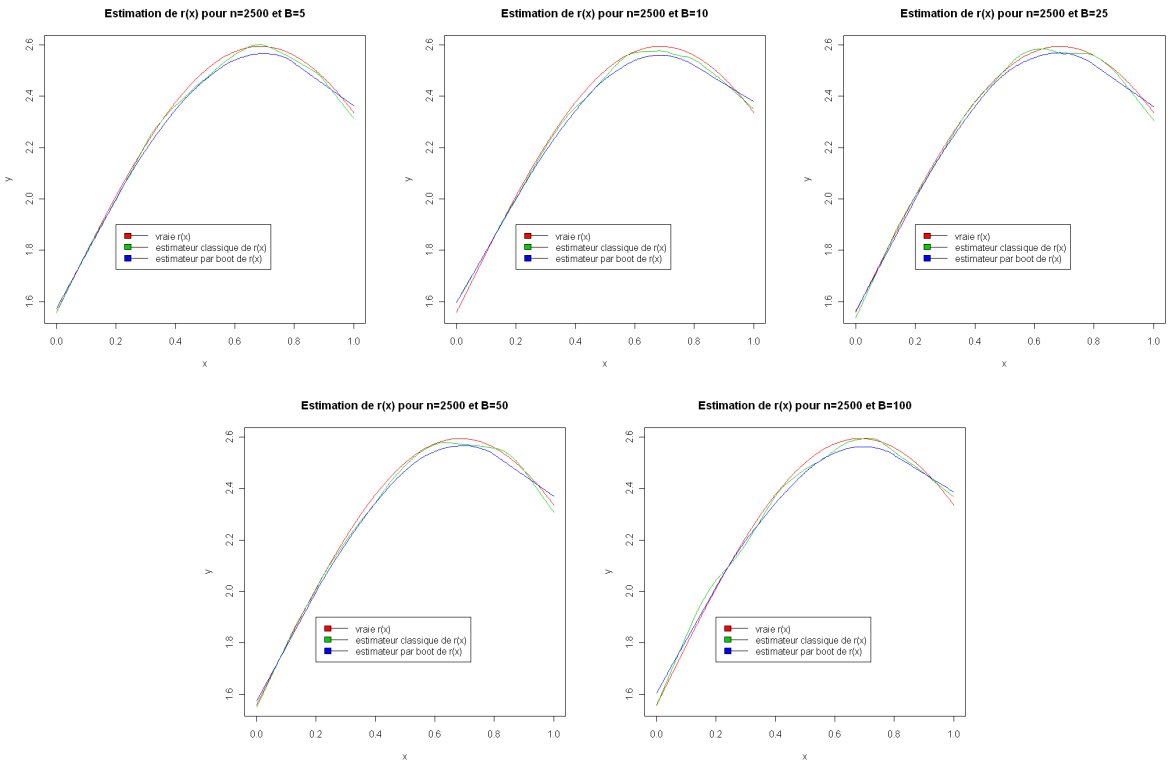


FIG. 13: *Comparaison entre la courbe théorique et celle estimée avec le paramètre de lissage h^* et celle estimée avec h^c_{boot}*

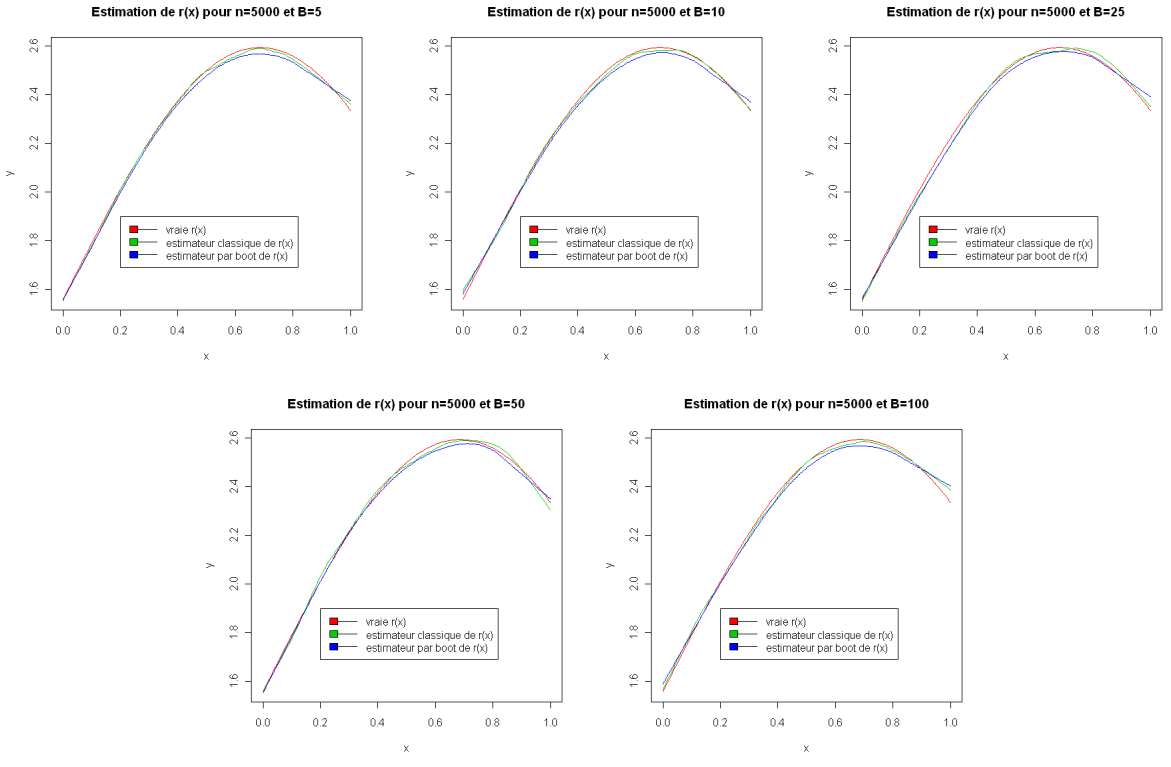


FIG. 14: *Comparaison entre la courbe théorique et celle estimée avec le paramètre de lissage h^* et celle estimée avec h^c_{boot}*

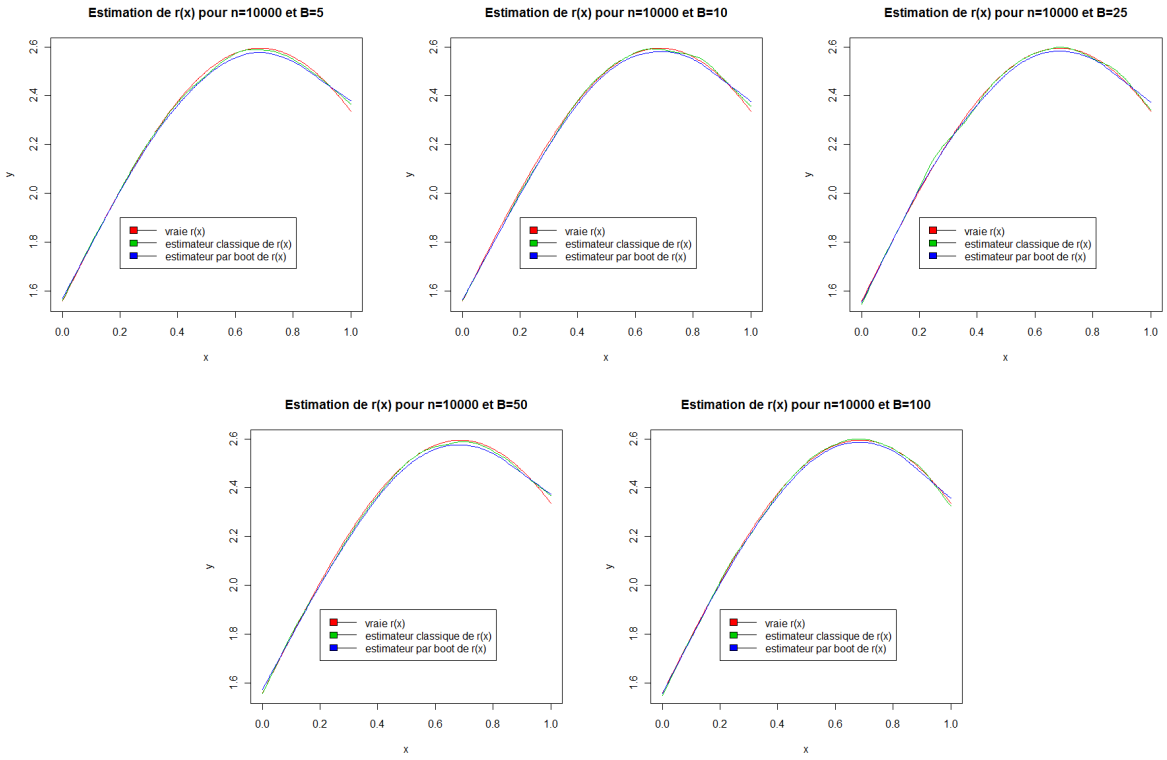


FIG. 15: *Comparaison entre la courbe théorique et celle estimée avec le paramètre de lissage h^* et celle estimée avec h^c_{boot}*

Bibliographie

[1] I. Abramson. Arbitrariness of the pilot estimator in adaptive kernel methods. *J. Multivariate Anal*, (12,562-567,), 1982.

[2] S. Barber and C. Jennison. Bootstrapping the kaplan-meier estimator. *Department of Mathematical Sciences, University of Bath, Claverton Down,Bath, BA2 7AY, UK.*

[3] A. Berlinet. Hierarchies of higher order kernels. *Probability theory and related fields*, (94, 489-504), 1993.

[4] P. Bertail and P. Combris. Bootstrap généralisé d'un sondage. *Annales d'Economie et de Statistique.*, (46), 1997.

[5] D. Bosq. Nonparametric statistics for stochastic processes : estimation and prediction. *Lecture notes in statistics*, (110, Springer-Verlag), 1996.

[6] T. Bouezmarni and J.V.K. Rombouts. Nonparametric density estimation for positive time series. *Econometric Theory*, September 21, 2006.

[7] T. Bouezmarni and O. Scaillet. Consistency of asymmetric kernel density estimators and smoothed histograms with application to income data. *Econometric Theory*, (21, 390-412), 2003.

[8] A. Bowman. A comparative study of some kernel based nonparametric density estimator. *Journal of Statistical Computation and Simulation*, (21,313-327), 1985.

[9] A. W. Bowman. An alternative method of crooss-validation for the smoothing density estimates. *Biometrika*, (71, 553-560), 1984.

[10] B. M. Brown and S. Chen. Beta-bernstein smoothing for regression curves with compact supports. *Scandinavian Journal of Statistics*, (26, 47-59), 1999.

[11] P. Burman. A date dependent approach to density estimation. *Zeitschrift Für Wahrscheinlichkeitstheorie and Verwandte Gebiete*, (69,609-628), 1985.

[12] R. Cao-Abad. Rate of convergence for wild bootstrap in non parametric regression. *Annals Statistics*, 1991.

[13] S. Chen. A beta kernel estimation for density functions. *Computational Statistics and Data Analysis*, (31, 131-145), 1999.

[14] S. Chen. Beta kernel for regression curve. *Statistica Sinica*, (10,73-92), 2000.

[15] S. Chen. Probability density functions estimation using gamma kernels. *Annals of the Institute of Statistical Mathematics*, (52, 471-480), 2000.

[16] S. Chen. Local linear smoothers using asymmetric kernels. *Annals of the Institute of Statistical Mathematics*, (54, 312-323), 2002.

[17] M. Chernick. Bootstrap methods : a practitioner's guide. *New York : Wiley*, 1999.

[18] M. P. Cohen. The bayesian bootstrap and multiple imputation for unequal probability sample designs. *National Center for Education Statistics 555 New Jersey Avenue NW, Washington DC 20208-5654*, 1997.

[19] G. Collomb. Estimation non paramétrique de la régression par la méthode du noyau thèse troisième cycle. *Toulouse 3*, 1976.

[20] G. Collomb. Méthodes non-paramétriques en régression, analyse de séries temporelles, prédictions et discrimination. *Doctorat d'état*, (Toulouse 3), 1983.

[21] G. Collomb. Propriétés de convergences presque complète du prédicteur à noyau. *Z. Wahrscheinlichkeitsthoeie verw. Gebiete*, (66, 441-640), 1984.

[22] P. Dagnelie. Statistique théorique et appliquée. tome 1 : Statistique descriptive et de l'inférence statistiquee. *Bruxelles De Boeck et Larcier*, (508), 1998.

[23] P. Deheuvels, D.M.Mason, and G. R. Shorack. Some results on the influence of extremes on bootstrap. *Annales de l'I.H.P., Section B*, 29(1) :83–103, 1993.

[24] P. Deheuvels and P. Hominal. Estimation non paramétrique de la densité compte tenu d'informations sur le support. *Revue de Statistique Appliquée*, (27, 47–68), 1979.

[25] L. Devroye. The equivalence of weak, strong and complete convergence in l_1 for kernel density estimates. *The Annals of Statistics*, (11,896-904), 1983.

[26] L. Devroye and L. Györfi. Nonparametric density estimation : The l_1. *View, New York ; John Wiley*, 1985.

[27] P. Diggle. A kernel method for smoothing point process data. *Applied Statistics*, (34,138-147), 1985.

[28] Y. Dodge. Some difficulties involving nonparametric estimation of a density function. *Journal of Official Statistics*, (2,193-202), 1986.

[29] B. Efron. Bootstrap methods : Another look at the jackknife. *Annals of Statistics*, (7) :1–26, 1979.

[30] B. Efron. Second thoughts on the bootstrap. *Statistical Science*, 18(2) :135–140, 2003.

[31] B. Efron, E. Halloran, and S. Holmes. Bootstrap confidence levels for phylogenetic trees. *Proc. Natl. Acad. Sci. USA (93,7085–7090)*, (14) :13429–13434, July 1996.

[32] B. Efron and Tibshirani. An introduction to the bootstrap. *Chapman and Hall, New York.*, 1993.

[33] V. A Epanechnikov. Nonparametric estimation of a multidimensional probability density. *Theory Probab. Appl*, (14, 153-158), 1969.

[34] J. Fan and Q. Yao. Nonlinear time series. *Springer-Verlag, New York.*, 2003.

[35] J. J. Faraway and J. Myoungshic. Bootstrap choice of bandwidth for density estimation. *Journal of the American Statistical Association.*, (vol.85, No.412 Theory and methods. 1119-1122.), December 1990.

[36] J.J. Faraway and M. Jhun. Bootstrap choice of bandwidth for density estimation. *J. Amer. Statist. Assoc*, (85, 1119-1122), 1990.

[37] J. D. Fermanian and O. Scaillet. The estimation of copulas :theory and practice arthur charpentier. *Ensae-Crest and Katholieke Universiteit Leuven ; BNP-Paribas and Crest ; HEC Genève and Swiss Finance Institute*, (Revised Proof Ref : 33259e), September 29, 2006.

[38] M. Fernandez and P. Monteiro. Central limit theorem for asymmetric kernel functionals. *Annals of the Institute of Statistical Mathematics*, (57, 425-442), 2005.

[39] F. Ferraty and P. Vieu. Statistique fonctionnelle : Modèles non-paramétriques de régression. *Notes de cours de DEA 2003/2004*.

[40] E. Fix and J.R. Hodges. Discriminatory analysis, nonparametric discrimination : consistency proprieties. *Technical report, Report NUM 4, USAF School of aviation Medicine, Randolph Field, Texas*, 1951.

[41] T. Gasser and H.G. Müller. Kernel estimation of regression functions. in smoothing techniques for curve estimation. *Eds. Gasser and Rosenblatt*, (23-68, Springer Verlag, Heidelberg, Germany), 1979.

[42] L. Györfi, W. Härdle, P. Sarda, and P. Vieu. Nonparametric curve estimation for times series. *Lecture notes in statistics*, (60, Springer-Verlag), 1989.

[43] P. Hall. Cross-validation in density estimation. *Biometrika*, (69, 383-390), 1982.

[44] P. Hall. Large sample optimality of least squares cross validation in density estimation. *The Annals of Statistics*, (11,1156-1174), 1983.

[45] P. Hall. Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *Journal of Multivariate Analysis*, (32, P177-203), 1990.

[46] P. Hall. The bootstrap and edgeworth expansion. *New York : Springer*, 1992.

[47] P. Hall. On plug-in rules for local smoothing of density estimator. *Ann. Statist*, (21,694-710), 1993.

[48] P. Hall and J. S. Marron. Extent to which least squares cross-validation minimises integrate square error in nonparametric density estimation. *Probability Theory and related fields*, (74, 567-581), 1987.

[49] P. Hall and J. S. Marron. Local minima in cross-validation function. *Journal of the royal statistical society*, (90,149-173), 1991.

[50] F. E. Harrell and C. E. Davis. A new distribution-free quantile estimator. *Biometrika*, (69,635-640), 1982.

[51] E. Herrmann. Local bandwidth choice in kernel regression estimation. *J. Comput. Graph. Statists*, (6, 35-54), 1997.

[52] E. Herrmann. Variance estimation and bandwidth selection for kernel regression smoothing and regression : Approaches, computation and application. *Ed. M.G. Schimek*, (71-107 Wiley Series in Probability and Statistics), 2000.

[53] W. Härdle. Applied nonparametric regression. *Cambridge University Press, UK*, 1990.

[54] W. Härdle and J. S. Marron. Optimal bandwidth selection in nonparametric regression function estimation. *Annals Statistics*, (13, 1465-1481), 1985.

[55] D. Ioannides. Integrate square error of non parametric estimators of regression function : the fixed design case. *Stat. and Prob. Let*, (15, 85-94), 1992.

[56] C. Jones, J. S. Marron, and P. Sheather. A brief survey bandwidth selection for density estimation. *Amer. Statist. Assoc.*, (89, 401-407), 1996.

[57] M.C. Jones. Simple boundary correction for kernel density estimation. *Statistical Computing*, (3, 135-146), 1993.

[58] M.C. Jones and P. Foster. A simple nonnegative boundary correction method for kernel density estimation. *Statistica Sinica*, (6, 1005-1013), 1996.

[59] S. Juan and F. Lantz. La mise en oeuvre des techniques de bootstrap pour la prévision econométrique :application à l'industrie automobile. *Oil & Gas Science and Technology*, 56(4) :373–388, 2001.

[60] T.Y. Kim and D.D. Cox. Convergence rates for average square errors for kernel smoothing estimators. *Nonparametric Statistics*, (13, 209-228), 2001.

[61] A. Kozek and E. Shuster. Optimal quantile principle for selecting variable bandwidth in regression estimators. *Proceedings of the computer Sci & Statist. Annual symposium on the interface*, (36), 1990.

[62] M. Lejeune and P. Sarda. Smooth estimators of distribution and density functions. *Computational Statistics and Data Analysis*, (14, 457-471), 1992.

[63] C. Léger and N. Altman. Bootstrap choice of tuning parameters. *Anu. Inst. Statist. Math*, (42, 709-735), 1990.

[64] C. Léger, DHN. Politis, and J.P. Romano. Bootstrap technology and applications. *Technometrics 34.*, pages 378–398, 1992.

[65] Y. Mack and B. Silverman. Weak and strong uniform consistency of kernel regression estimates. *Z. Wahrscheinlichkeitsth*, (61, 405-415), 1982.

[66] E. Mammen. When does bootstrap works : asymptotic results and simulations. *Lect. Notes in statist. 77. Springer Verlag, Berlin*, 1992.

[67] B.F.J. Manly. Randomization, bootstrap and monte carlo methods in biology. *New York : Chapman and Hall*, 1997.

[68] J. S. Marron. Automatic smoothing parameter selection : a survey empirical economics. *J. Multiv. Anal.*, (13, 187-208), 1988.

[69] J. S. Marron and D. Ruppert. Transformations to reduce boundary bias in kernel density estimation. *Journal of the Royal Statistical Society, Series B*, (56, 653-671), 1994.

[70] H.G. Müller. Smooth optimum kernel estimators near endpoints. *Biometrika*, (78, 521-530), 1991.

[71] E. Nadaraya. On estimating regression. *Theory Prob. Appl*, (10, 186-196), 1964.

[72] E. Nadaraya. On nonparametric estimation density function and regression. *Theory Probab P.P.L*, (10,186-190), 1965.

[73] W. Padgett and L. Thombs. Smooth nonparametric quantile estimation under censoring : Simulations and bootstrap methods. *Communication in Statistics, Part B-Simulation and computation*, (5, P1003-1025), 1986.

[74] A. Pagan and A. Ullah. Nonparametric econometrics. *Cambridge University Press, UK*, 1999.

[75] R. Palm. Utilisation du bootstrap pour les problèmes statistiques liés à l'estimation des paramètres. *Biotechnol. Agron. Soc. Environ.*, 29 avril 2002.

[76] B. U. Park and S. J. Marron. Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association*, (85, 66-72), 1990.

[77] E. Parzen. On estimation of a probability density function and mode. *Ann. Math. Statist*, (33, 1065-1076), 1962.

[78] K. Pearson. On the systematic fitting of curves to observations and measurements. *Biometrica*, (1, 265-303. 2, 1-23).

[79] J. Romano. On weak convergence and optimality of kernel density estimates of the mode. *The Annals of Statistics*, (16,P629-647), 1988.

[80] M. Rosenblatt. Remarks in some nonparametric estimates of a density function. *Ann. Math. Statist.*, (27,832-837), 1956.

[81] M. Rudemo. Empirical choice of histogram and kernel density estimators. *Scandinavian Journal of Statistics*, (9,65-78), 1982.

[82] P. Révész. How to apply the method of stochastic approximation in the nonparametric estimation of a regression function. *Math. Operationsforsch. Statist. ser Statist*, (8,119-126), 1977.

[83] P. Sarda and P. Vieu. Kernel regression. smoothing and regression : Approaches, computation and application. *Ed. M.G Schimek*, (43-70, Wiley Series in Probability and Statistics), 2000.

[84] E. Schuster. Incorporating support constraints into nonparametric estimators of densities. *Communications in Statistics - Theory and Methods*, (14, 1123-1136), 1985.

[85] D. W. Scott. Averaged shift histograms : effective nonparametric density estimators in several dimensions. *The annals of statistics*, (Vol. 13 1024-1040), 1985.

[86] D. W. Scott and G.R. Terrell. Oversmoothed nonparametric density estimates. *Journal of the American statistical association*, (80,209-214), 1985.

[87] D. W. Scott and G.R. Terrell. Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association*, (82, P1131-1146), 1987.

[88] J. Shao and D. Tu. The jackknife and bootstrap. *Springer-Verlag, New York*, 1995.

[89] S.J Sheather and M.C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc.*, (B53, 683-690), 1991.

[90] B. Silverman. Weak and strong uniform consistency of the kernel estimate of density function and its derivates. *Ann. Statist*, (6,177-184), 1978.

[91] B. Silverman. Density estimation for statistics and data analysis. *London : Chapman & Hall*, 1986.

[92] B. Silverman and G. Young. The bootstrap : To smooth or not smooth ? *Biometrika*, (12,P469-479), 1987.

[93] B. Silverman and G. Young. The bootstrap : To smooth or not smooth ? *Biometrika*, (74, 469-479), 1987.

[94] C. Stone. Optimal rates of convergence for nonparametric regression. *Annals Statistics*, (9, 1348-1360), 1981.

[95] C. Stone. Optimal global rates of convergence for nonparametric regression. *Annals Statistics*, (10, 1040-1053), 1982.

[96] C. Stone. An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics*, (12,P1285-1297), 1984.

[97] C. Taylor. Bootstrap choice of smoothing parameter in kernel density estimation. *Biometrika*, (76, P705-712), 1989.

[98] J.W. Tukey. Curve as parameters, and touch estimation. proceedings of the 4th symposium on mathematics. *Statistics and probability*, (681-694, Berkeley, CA, USA), 1961.

[99] M. Verleysen and A. Lendasse. Le test des méthodes neuronales : Comment utiliser les techniques de rééchantillonnage ? *ACSEG 2003 proceedings - Connectionist Approaches in Economics and Management Sciences Nantes (France)*, pages 515–534, November 2003.

[100] P. Vieu. Bandwidth selection for kernel regression : a survey. computer intensive methods in statistics. *Eld Härdle, W. et Simar*, (134-149, Physica-Verlag, Heidelberg, Germany), 1993.

[101] P. Vieu. Multiple kernel procedure : an asymptotic support. *Scand. J. of statist*, (26, 61-72), 1999.

[102] G. Wahba. Optimal properties of variable knot, kernel and orthogonal series methods for density estimation. *Ann. Of Statist*, (3,15-29), 1975.

[103] M. Wand. On exact l_1 rates of convergence in nonparametric kernel regression estimation. *Scand. J. of statist*, (17, 251-256), 1990.

[104] G.S Watson. Smooth regression analysis. *Sankhya Ser*, (A, 26, 359-372), 1964.

[105] M. Woodroofe. On choosing a delta-sequence. *Ann. Math. Stat.*, (41, 1665-1671), 1970.

[106] G. Young. Bootstrap : more than a stab in the dark ? *Stat. Sci. 9*, 1994.

[107] K. Ziegler. On local bootstrap bandwidth choice in kernel density estimation. *Technical University of Ilmenau*.

[108] K. Ziegler. On non parametric kernel estimation of the mode of the regression function in random design model. *J.Nonpar. Statist*, (14, 749-774), 2002.

[109] N. Zougab. Etude comparative des méthodes de sélection du paramétrer de lissage dans l'estimation de la densité de probabilité par la méthode du noyau. Thèse de magister, Université de Bejaia, Mai 2007.

RÉSUMÉ

L'estimation de la densité de probabilité par la méthode du noyau à partir d'un échantillon $X_1, X_2, ..., X_n$ nécessite le choix du noyau K et du paramètre de lissage h. Si le choix du noyau n'est pas un problème dans l'estimation de la densité, il n'en est pas de même pour le choix du paramètre de lissage qui ne dépend que de la taille d'échantillon.

La technique de bootstrap qui se base sur le principe de rééchantillonnage a apportée beaucoup de solutions pour des problèmes statistiques divers. Dans ce travail, nous avons proposé l'application de cette technique pour le choix du paramètre de lissage h dans l'estimation de la densité de probabilité et la courbe de régression de la moyenne. Nous avons également étudié l'influence du choix du noyau sur la qualité de l'estimateur à noyau de la densité de probabilité au sens de l'erreur globale asymptotique notée AMISE.

Une étude de simulation sur des échantillons de taille $n = 1000$, 2500, 5000 et 10000 pour l'estimation d'une densité de probabilité par la méthode du noyau et $n = 200$, 500, 1000, 2500, 5000 et 10000 pour l'estimation de la courbe de régression de la moyenne montre que les résultats obtenus par la technique de bootstrap ne sont pas en général meilleurs que ceux obtenus par les méthodes classiques.

L'étude montre aussi, que dans certains cas, le choix du noyau est important particulièrement dans l'estimation des densités de probabilité à support compact, comme la loi exponentielle.

Mots-clés :Bootstrap, densité de probabilité, noyau, paramètre de lissage, courbe de régression de la moyenne, estimation, erreur relative globale, validation croisée, plug-in.

ABSTRACT

The estimate of the probability density by the kernel method from a sample X_1, X_2 ,..., X_n requires the choice of kernel K and smoothing parameter h. If the choice of kernel is not a problem in estimating of the density, it is not the same for choosing the smoothing parameter which depends only on the sample size.

The bootstrap technique which is based on the principle of re-sampling has provided many solutions for various statistical problems. In this work, we proposed the application of this technique for choosing the smoothing parameter h in the estimate of the probability density curve and the regression of the average. We also studied the influence of the choice of the kernel on the quality of the kernel estimator of the probability density within the meaning of global asymptotic error noted AMISE.

A simulation study on samples of size $n =$ 1000, 5000 and 10000 for the estimation of a probability density by the kernel method and $n =$ 200, 500, 1000, 2500, 5000 and 10000 for the estimated regression curve of the average shows that the results obtained by the bootstrap technique are not generally better than those obtained by classical methods. The study also shows that in some cases the choice of kernel is particularly important in estimating the probability densities of compact support, such as the exponential law.

Keywords : Bootstrap, the probability density, kernel, smoothing parameter, regression curve of the average, estimation,error with respect global, cross-validation, plug-in.