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

Université A/MIRA de Béjaia
Faculté des Siences Exactes
Département des Mathématiques/MI

Mémoire de fin de Cycle

En vue de l'obtention d'un Master en Mathématiques

Option

Statistique et Analyse Décisionelle
SAD

Thème

Le bootstrap pour le choix de la fenêtre dans la méthode du noyau.

Présenté par :
M^{elle} **BERKANI Sabrina** M^{elle} **BOUICHE Yasmine**

Devant le jury composé de :

Président	M^{r} **M. BOURAINE**	**MAA**	**Université A/Mira, Béjaïa**
Rapporteur	M^{me} **A. BARECHE**	**MCA**	**Université A/Mira, Béjaïa**
Examinatrice	M^{me} **H. TABTI**	**MAA**	**Université A/Mira, Béjaïa**

Promotion 2014/2015.

Remerciements

Avant tout, nous remercions notre bon dieu, le tout puissant de nous avoir accordé volonté et courage pour mener à bien notre travail.

On tiens a remercier Madame A. Bareche, pour nous avoir proposé ce travail, pour l'avoir dirigé, pour la confiance qu'elle nous a accordée et pour l'aide qu'elle nous a apportée pendant toute la durée de ce mémoire. Son oeuil critique nous a été très précieux pour structurer notre mémoire et pour améliorer sa qualité.

Nous adressons aussi nos sincères remerciements aux membres de jury : Monsieur Bouraine et madame Tabti, pour avoir accepté d'examiner et d'evaluer notre travail et pour le temps et les orientations accordés. Nous remercions également nos familles et amis, pour nous avoir soutenu de près ou de loin.

Dédicaces

Je dédie ce modeste travail qui est le fruit récolté après tant d'années d'efforts : A mes parents : Vous m'avez appris beaucoup de chose, je vous doit ma vie, vous m'avez protégé, supporté, et soutenu durant toute ma vie. Je vous dédie toute ma reconnaissance et mon dévouement.

A toi papa : tu es l'homme que j'aime le plus au monde.

A toi maman : tu m'as tellement donné que je n'ai pas assez de mot pour t'exprimer ma reconnaissance ; Je t'aime.

A la mémoire de mes grands parents et de tonton Abdelkader, que dieu les accueillent dans son vaste paradis.

A mes sœurs, Aida et Celia pour leur soutien moral et leur aide, mes tantes : Dadou, Lila et Wahiba, leurs maris Djamel, Nacer, mes oncles, mes cousines et mes cousins. A ma petite princesses et ange Nélia.

A mes amis et ma binôme Sabrina sans qui ce mémoire ne serait pas ce qu'il est.

Yasmine

Dédicaces

Je dédie ce travail à :

Mon père et ma mère, qui m'ont comblé de leur soutien et m'ont voué un amour inconditionnel. Vous êtes pour moi un exemple de courage et de sacrifice continu, que ce modeste travail témoigne mon affection.

Mes frères et soeur : Houa, Mahdi et Razik qui n'ont cessé d'être pour moi des exemples de persévérance, de courage et de générosité et a mes deux nièces : Lina et Asma.

Sabrina

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

ISE : Integrated Squared Error.

MSE : Mean Squared Error.

MISE : Mean Integrated Squared Error.

AMISE : Asymptotic Mean Integrated Squared Error.

CV : Cross Validation.

UCV : Unbiased Cross Validation.

BCV : Biased Cross Validation.

ROT : Rule Of Thumb.

SJ : Sheather and Jones.

boot : Bootstrap.

Introduction générale

L'estimation non paramétrique [23] est une méthode de la statistique mathématique dans laquelle le prédicteur ne prend pas de forme prédéterminée, mais est construit selon les informations provenant des données. La régression non paramétrique exige des tailles d'échantillons plus importantes que celles de la régression basée sur des modèles paramétriques parce que les données doivent fournir la structure du modèle ainsi que les estimations du modèle.

En statistique, effectuer un ré-échantillonnage [18] c'est utiliser toute méthode permettant d'estimer la précision d'un échantillon statistique (médiane, variance, quantile) en utilisant des sous-ensembles des données disponibles ou en effectuant un tirage aléatoire avec remise, à partir de ce même ensemble de données (bootstrap) ; ou encore de valider des modèles en utilisant des sous-ensembles aléatoires (bootstrap [15], validation croisée [17]) Parmi les techniques de rééchantillonnage les plus utilisées on trouve le bootstrap, la validation croisée et la ré-injection.

L'estimation par noyau (ou encore méthode de Parzen [22] et Rozenblatt [24]) est une méthode non-paramétrique d'estimation de la densité de probabilité d'une variable aléatoire. Elle se base sur un échantillon d'une population statistique et permet d'estimer la densité en tout point du support. Son estimateur dépend des deux paramètre K et h, où K est un noyau (kernel en anglais), (symétrique ou asymétrique) et h un paramètre de lissage aussi nommé fenêtre.

Le problème du choix de la fenêtre dans la méthode du noyau a été largement étudié et plusieurs méthodes ont été proposées. Nous nous intéressons dans ce travail a la méthode bootstrap pour le choix de ce paramètre.

La motivation du bootstrap (Efron, 1982 ; Efron et Tibshirani, 1993) est d'approcher par

simulation (Monte Carlo) la distribution d'un estimateur lorsque l'on ne connaît pas la loi de l'échantillon ou, plus souvent lorsque l'on ne peut pas supposer qu'elle est gaussienne. L'objectif est de remplacer des hypothèses probabilistes pas toujours vérifiées ou même invérifiables par des simulations et donc beaucoup de calcul. Cette méthode consiste à estimer la distribution d'une statistique calculée à partir d'un échantillon provenant d'une certaine loi inconnue en utilisant la distribution de cette même statistique mais calculée à partir d'échantillons provenant de la loi empirique des observations originales. Dans la plupart des cas, on obtient une approximation numérique de cette distribution en faisant une simulation.

L'ojectif de ce travail réside sur l'influence de la technique bootstrap sur le choix du paramètre de lissage dans la méthode du noyau. Pour cela, nous adoptons deux classes de méthodes : validation croisée et ré-injection.
Ce mémoire est divisé en trois chapitres, le premier introduit le principe de bootstrap, le deuxième résume les différentes méthode de séléction de la largeur de la fenêtre, groupées en deux classes.
Les méthodes présentées dans ces chapitres sont appliquées sous le logiciel R dans le troisième chapitre.

Chapitre I

Principe du bootstrap

I.1 Introduction

la technique de bootstrap est une méthode d'inférence statistique datant de la fin des années 1970, époque où devenait abordable des calculs informatiques intensifs. On calculait depuis près d'un siècle des estimations, des mesures de dispersion (variance, écart-type), des intervalles de confiance voire un Test d'hypothèse. Il s'agissait maintenant d'évaluer la sensibilité de ces indications aux particularités de l'échantillon mesuré, en analysant ses sous-échantillons possiblles. Cette méthode est basée sur des simulations, comme les méthodes de Monte-Carlo, les méthodes numériques bayésiennes, à la différence près que le bootstrap ne nécessite pas d'autre information que celle disponible dans l'échantillon : il va même en éliminer des parties entières par roulement.

I.2 Exemple canonique

Un exemple proposé par Efron [10] : À 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 (imitation de médicament sans principe actif, utilisée pour les tests en double aveugle). Leurs durées de vie sont mesurées, en jours, et on donne les résultats dans le tableau (I.1) :

La différence des moyennes est égale à : 30,63.

											Moyenne	Ecart-type
Traitées	X	94	197	16	38	99	141	23	/	/	86,86	25,24
Contrôlées	Y	52	104	146	10	51	30	40	27	46	56,22	14,14

TAB. I.1 – Les données des souris qui reçoivent le traitement et celles qui reçoivent le placebo

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

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

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

$$Se = \sqrt{Se_1 + Se_2} = \sqrt{14,14^2 + 25,24^2} = 28,93.$$

Si on compare directement les deux moyennes 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 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 suffisament grande. Pour pouvoir appliquer l'approximation normale (théorème central limite), on utilisera le fait que :

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

avec

F est la loi empirique de l'échantillon X,

μ est la moyenne empirique de l'échantillon X,

s^2 est 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 a la technique bootstrap.

I.2.1 Problématique

La technique de bootstrap est conçue pour être utilisée dans le contexte du travail empirique, comme le nom le suggère. L'idée du principe original de la méthode est d'utiliser le seul ensemble de données disponibles pour approximer la distribution des aléas ou d'autres quantités du modèle qui est donnée par Bradly Effron [11] pour expliquer le principe de bootstrap, et cela en construisant toutes les combinaisons possibles (toutes les fonctions de répartitions empiriques possibles) de ces données.

I.3 Principe de base du bootstrap

Monde réel		Monde bootstrap
P	$\leftrightarrow$	$P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$ mesure empirique
$\mathbb{X} = (X_1, \ldots, X_n)$ n-échantillon de la loi P	$\leftrightarrow$	$\mathbb{X}^* = (X_1^*, \ldots, X_n^*)$ n-échantillon de la loi P_n, appelé **échantillon bootstrap** associé à $\mathbb{X}$. $P(X_j^* = X_i \mid \mathbb{X}) = \frac{1}{n}\ \forall i, j$
$R_n(\mathbb{X}, P)$	$\leftrightarrow$	$R_n^* = R_n(\mathbb{X}^*, P_n)$ **réplication bootstrap** de la racine

FIG. I.1 – Principe bootstrap (voir [11])

Soit $T(y_1, ..., y_n)$ une statistique basée sur n observations, par exemple un estimateur d'un certain paramètre, ou encore, un estimateur d'un moment des y_i (supposés indépendants et de même loi). Cette méthode a l'avantage de fonctionner même dans un cadre non-paramétrique. Elle est simplement basée sur le principe de base de l'estimation : remplacer une fonction de la vraie distribution (inconnue) par la même fonction appliquée à la distribution empirique. Pour fixer les idées, donnons-nous l'exemple suivant où l'on estime le moment d'ordre 4 des y_i supposés de même loi, $T(Y) = \frac{1}{n}\sum_{i=1}^{n} y_i^4$,

$T^0 = E[y_1^4] = \lim_{n\mapsto\infty} T(Y)$. Pour bien comprendre le bootstrap, il faut interpréter $T(y_1, ..., y_n)$ comme une fonction de répartition empirique F_h des données :

$$F_h(x) = \frac{1}{n}, P(i : y_i \leq x),$$

En effet, la donnée de F_h est équivalente à celle de l'échantillon Y. On utilisera alors l'abus de notations : $T(Y) = T(F_h)$.
Le bootstrap consiste à produire des "échantillons indépendants" de T en faisant comme si la distribution empirique des y_i était la vraie distribution des y_i :
• Générer un échantillon $Y^* = (y_1^*, ..., y_n^*)$ en effectuant n tirages aléatoires avec remise dans $y_1, ..., y_n$.
• $T(Y^*)$ est un échantillon de la statistique $T(y_1, ..., y_n)$.

Ayant ainsi produit un grand nombre d'échantillons $Y^1, ..., Y^B$ on peut par exemple estimer l'écart quadratique moyen entre $T(Y)$ et T^0 par :

$$\hat{\sigma}^2 = \frac{1}{B}\sum_{b=1}^{B}(T(Y^b) - T(Y))^2.$$

Pour une région de confiance, soit ε tel que 95% des $T(Y^b)$ satisfassent

$$T(Y) \in [T(Y^b) - \varepsilon, T(Y^b) + \varepsilon],$$

alors on considèrera que l'intervalle

$$[T(Y) - \varepsilon, T(Y) + \varepsilon],$$

est une région de confiance pour T^0 de niveau approximatif 5% [1].

I.4 Méthodes de ré-échantillonage

I.4.1 Définition

Le terme de ré-échantillonage, 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 échatillon initial. Disposant d'un échantillon destiné à donner une certaine information sur une population, on tire au sort, parmi les sous populations réduites à cet échantillon, un nouvel échantillon

de taille n. On répète cette opération B fois, où B est grand. On analyse ensuite les nouvelles observations initiales.
A priori on peut avoir des doutes sur l'efficacité d'une telle méthode. Cependant, comme on va le voir, si l'on rajoute aucune information cela permet dans certains cas, d'extraire de l'échantillon de base de l'information souhaitée.

I.4.2 Bootstrap des individus

On considère un échantillon de n observations : $x_1, x_2, ..., x_n$ prélevé de manière aléatoire et simple dans une population. Ces observations peuvent concerner une seule variable, ou au contrainte, être relatives à plusieurs variables. Dans ce cas, les x_i représentent des vecteurs de dimension p, p étant le nombre de variables. Le principe de la méthode bootstrap est de prélever une série d'échantillons aléatoires simples avec remise de n observations dans l'échantillon initial, considéré comme une population. Ces échantillons successifs seront notés : $x^{*1}, x^{*2}, ..., x^{*k}, ..., x^{*B}$, B : étant le nombre de ré-échantillonnage effectués.
À titre illustratif [21], nous considérons le problème de l'estimation de diverses caractéristiques de la population des tailles des exploitations agricoles de la région Wallonnie, 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_{iinf} et x_{isup}, il a été généré n_i nombres aléatoires appartenant à une distribution uniforme dans le domaine (x_{iinf}, x_{isup}). Il a été ainsi obtenu les tailles simulées des 24.719 exploitations. La figure (I.2) reprend l'histogramme et les principaux paramètres de cette population dont la caractéristique la plus marquante est la très forte asymétrie à 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, il a été sélectionné, de manière aléatoire et simple, un échantillon de 100 observations. La deuxième colonne du tableau (I.2), 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}, x^{*3}, ceux-ci ayant également été classés par ordre croissant. On constate, par exemple, que pour l'échantillon x^{*1}, la première 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 ré-échantillonnage 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.

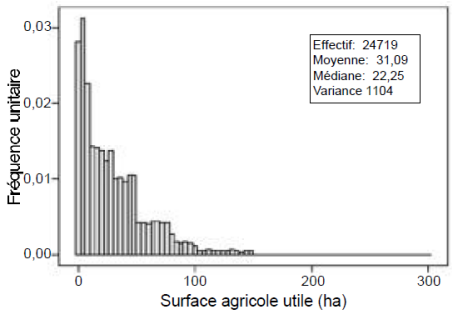

FIG. I.2 – Population simulée des tailles des exploitations de la région Wallonnie en 1995

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
...				
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. I.2 – Echantillon initial x et résultats de trois ré-échantillonage 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 ré-échantillonnage assurant l'égalité de ces proportions sont également proposées. Cette approche porte le nom de ré-échantillonnage balancé (bootstrap bayésien) [8].

I.5 Erreur standard et biais d'un paramètre

I.5.1 Estimation de l'erreur standard

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

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

une estimation de ce paramètre, obtenue à partir des données de l'échantillon initial X. Chaque échantillon obtenu par ré-échantillonage permet de calculer une répétition du bootstrap (bootstrap réplication) de l'estimation $\widehat{\theta}$:

$$\widehat{\theta}^*_k = f(x^*_k), (k = 1....B),$$

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 $\widehat{\theta}$ s'écrit, successivement pour les trois paramètres considérés :

$$\bar{x} = \frac{1}{100}\sum_{i=1}^{100} x_i,$$

$$\tilde{x} = \frac{1}{2}(x_{[50]} + x_{[51]}),$$

$$\widehat{\sigma}^2 = \frac{1}{99}\sum_{i=1}^{100}(x_i - \overline{x})^2.$$

$x_{[50]}$ et $x_{[51]}$ étant les observations de rangs 50 et 51 respectivement de l'échantillon initial. Les valeurs numériques pour ces trois estimations sont données dans la première partie du tableau (I.3) sur la ligne intitulée $\widehat{\theta}$.
Le tableau (I.3) résume le valeurs des paramètres estimés pour l'échantillon initial ($\hat{\theta}$) et pour les trois premiers échantillons obtenus à partir du ré-échantillonnage de ($\hat{\theta}^*_1$,$\hat{\theta}^*_2$, $\hat{\theta}^*_3$) ; moyennes $\hat{\theta}^*$ et écarts-types ($\widehat{\sigma}_{\hat{\theta}^*}$) des paramètres estimés pour 1000 ré-échantillonnages

Les calculs des trois paramètres peuvent être répétés pour les échantillons x^*_1, x^*_2, x^*_3. Les résultats obtenus sont repris dans la seconde partie du tableau qui peut évidemment être complété au fur et à mesure des ré-échantillonnages fournissant $x^*_4; x^*_5; ...x^*_B$. Disposant des

Paramètre	Moyenne	Médiane	Variance
$\hat{\theta}$	28,13	21,56	854,63
$\hat{\theta}_1^*$	27.84	18.91	667.19
$\hat{\theta}_2^*$	26.32	19.95	796.93
.	.	.	.
.	.	.	.
.	.	.	.
$\widehat{\theta}^*$	27.99	20.44	843.58
$\widehat{\sigma}_{\widehat{\theta}^*}$	2.89	2.53	184.25

TAB. I.3 – Paramètres estimés pour l'échantillon initial obtenus à partir du ré-échantillonnage ; moyennes, médianes et écarts-types des paramètres estimés

B répétitions, on peut déterminer la moyenne

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

et l'écart type des $\widehat{\theta}_k^*$

$$\widehat{\sigma}_{\widehat{\theta}^*} = \sqrt{\frac{1}{(B-1)}\sum_{k=1}^{B}(\widehat{\theta}_k^* - \widehat{\theta}^*)^2}.$$

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

Algorithme d'estimation standard des erreurs standards
Tirer B échantillons bootstrap $X_1^*, X_2^*, X_3^*, ..., X_B^*$ à partir de X
Calculer la copie bootstrap $\hat{\theta}_k^* = s(X_k^*)$
Calculer l'erreur standard par les B copies

$$\hat{\sigma}_{\hat{\theta}^*} = \sqrt{\sum_{b=1}^{B}(\hat{\theta}_b^* - \hat{\theta}^*)^2/(B-1)}$$

avec $\hat{\theta}^* = \sum_{b=1}^{B}\frac{\hat{\theta}_b^*}{B}$

Pour l'exemple ci-dessus, 1000 ré-échantillonages ont été réalisés. Les figures qui suivent donnent 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 médianes et des 1000 variances sont donnés dans la troisième partie du tableau précédent (tableau (I.3)), dans les lignes intitulées $\hat{\theta}^*$, $\hat{\sigma}_{\hat{\theta}^*}$.

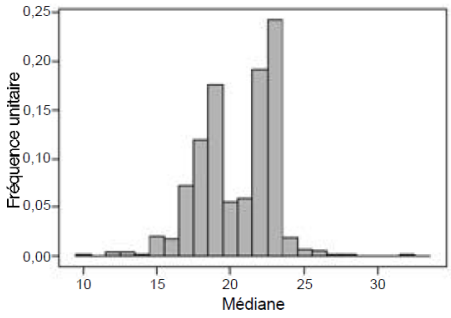

FIG. I.3 – Distribution des médianes de 1000 échantillons obtenus par bootstrap.

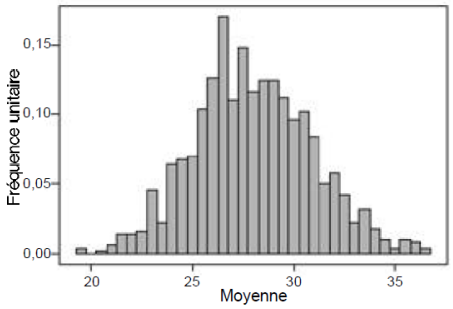

FIG. I.4 – Distribution des moyennes de 1000 échantillons obtenus par bootstrap.

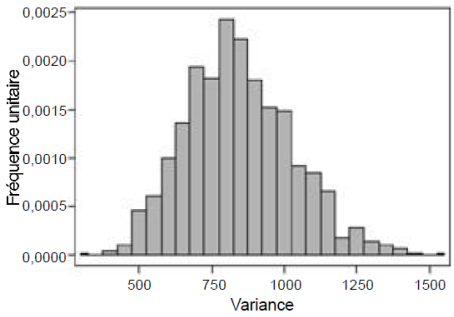

FIG. I.5 – Distribution des variances de 1000 échantillons obtenus par bootstrap.

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 $\hat{\sigma}_{\hat{\theta}^*}$ 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 :

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

L'estimateur $\hat{\sigma}_{plug}$ 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 :

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

En d'autres termes, 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. 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) [13] proposent les règles empiriques suivantes pour le choix de B :
Un nombre réduit de répétitions ($B = 25$, par exemple) 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 fonction de la taille n de l'échantillon.

I.5.2 Estimation du biais

Le biais d'un paramètre peut être éstimé par la méthode du boostrap de la manière suivante :

$$biais_B(\widehat{\theta}) = \widehat{\theta}^* - \widehat{\theta}, \tag{I.1}$$

avec $\widehat{\theta}$ est le paramètre estimé par bootstrap. L'estimation du biais pour la moyenne et la médiane est égale à : 27.99-28.13=-0.14 et 20.44-21.56=-1.12, respectivement. Disposant d'une estimation du biais, on peut éventuellement corriger l'estimation initiale, on obtient alors

$$\widehat{\theta}_c = \widehat{\theta} - biais_B(\widehat{\theta}).$$

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
$\widehat{\theta} = f(x_1, ...x_n)$ est inappropriée pour estimer θ.

I.6 Comportement asymptotique du bootstrap

Tout estimateur raisonnable doit converger vers le paramètre estimé lorsque la taille d'échantillon tend vers $+\infty$. Dans le cas de l'estimateur bootstrap de la loi d'une variable

aléatoire $R(\underline{X}, F)$, avec $\underline{X} = (X_1, X_2, ...X_n) \sim F$ et $\underline{X}^*$ est l'échantillon bootstrap, on souhaite par exemple que :

$$\lim_{n\mapsto+\infty} \sup_x |P_{F_h}(R(\underline{X}^*, F_h) \leq x) - P(R(\underline{X}, F) \leq x)| = \lim_{n\mapsto+\infty} \sup_x |h_{boot}(x) - h_n(x)| = 0.$$

L'équation précedante exprime que la loi de l'estimateur bootstrap $R(\underline{X}^*, F_h)$ est proche de celle de $R(\underline{X}, F)$ quand n est assez grand. (Comme $P_{F_h}(R(\underline{X}^*, F_h) \leq x)$ est une variable aléatoire, on sous-entend ici que la convergence a lieu au sens fort (presque sûre) ou au sens faible (en probabilité)).

I.7 Avantages et inconvénients du bootstrap

Avantages :

Ils sont essentiellement :

1. La simplicité du principe.

2. Des résulats expérimentaux assez bons en divers domaines, par exemple en régression non-paramétrique, ou en calibration d'intervalles de confiance, même sur des échantillons de taille modeste.

3. Des théorèmes démontrant la validité des approximations jusqu'au deuxième ordre par rapport à la taille de l'échantillon. Ceci explique le point précédent.

4. Même si on connaît bien la loi asymptotique de $T(Y) - T^0$ correctement normalisée, il est fréquent qu'il n'y ait pas d'expression explicite pour sa densité ou sa variance, ce qui rend difficile l'exploitation de cette information. Le bootstrap permet de contourner ce problème.

Inconvénients :

1. Le bootstrap n'est sûr que dans un cadre où l'approximation gaussienne est valide ; de plus les valeurs extrêmes dans le monde réel et le monde bootstrappé ont a priori des distributions très différentes.

2. L'inconvénient majeur de la technique de bootstrap est qu'elle nécessite un temps de calcul important [14].

I.8 Conclusion

Il est incontestable que l'utilisation des techniques de ré-échantillonnage a été rendu possible grâce a la génération 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 leurs 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.

Chapitre II

Méthode du noyau

II.1 Introduction

Étant donné un n-échantillon $X_1, X_2, ..., X_n$ de variables aléatoire, indépendantes et de même densité f inconnue, considérons l'estimateur à noyau de Parzen-Rozenblatt de la densité f donné par :

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

où h est le paramètre de lissage et K une fonction de densité définie sur $\mathbb{R}$ apelée noyau. Pour estimer f, il faut choisir le noyau K et le paramètre h. (Voir Scott [27] et Silverman [30] pour de nombreux exemples intéressants, et une bonne introduction aux idées importantes).

Si le choix du noyau n'est pas un problème dans l'estimation des densités symétriques, il n'est pas de même pour le choix de la largeur de la fenêtre h qui dépend essentiellement de la taille n de l'échantillon. En effet, dans l'estimation de la densité par cette méthode, se pose le problème du choix du paramètre de lissage pour lequel il existe de différentes méthodes.

Dans ce chapitre on introduit quelques classes de méthodes telles que :
La méthode de ré-injection (plug-in) et de validation croisée (cross-validation).

II.2 Critères d'erreurs et définitions

II.2.1 Les différents critères d'erreur

Pour mesurer les performances théoriques des estimateurs et identifier le meilleur, il est nécessaire de spécifier un critère d'erreur.
Nous considérons la densité de proabilité f et son estimateur f_h.

L'erreur quadratique intégrée ISE

$ISE(f(x), f_h(x)) = \int [f(x) - f_h(x)]^2 dx.$

L'erreur quadratique moyenne MSE

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

$MSE(f(x), f_h(x)) = [biais f_h(x)]^2 + Var f_h(x).$

L'erreur quadratique moyenne intégrée MISE

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

II.2.2 Quelques définitions

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

Définition II.2.2. On dit qu'un estimateur f_h de f est asymptotiquement sans biais si : $\lim_{n\longmapsto\infty} \mathbb{E}[f_h(x)] = f(x)$, en tout point x pour lequel la densité f est continue.

Définition II.2.3. On dit qu'un estimateur f_h de f est asymptotiquement uniformément sans biais si :
$\lim_{n\longmapsto\infty} \sup_x \mathbb{E}[f_h(x) - f(x)] = 0.$

Définition II.2.4. On dit qu'un estimateur f_h de f est ponctuellement consistant en moyenne quadratique si :
$\lim_{n\longmapsto\infty} MSE(f_h(x) - f(x)) = 0$, en tout point x pour lequel la densité f est continue.

Définition II.2.5. On dit qu'un estimateur f_h de f est ponctuellement consistant en moyenne quadratique intégrée si :
$\lim_{n\longmapsto\infty} MISE(f_h(x) - f(x)) = 0.$

II.3 Propriétés d'un estimateur à noyau

Dans cette section, on présente des propriétés d'un estimateur à noyau (espérance, biais et variance). Ainsi que leurs comportements asymptotiques.

II.3.1 Espérance, biais et variance

L'espérance mathématique de $f_h(x)$:

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

En posant : $y = \frac{x-u}{h}$, on trouve : $\mathbb{E}f_h(x) = \int K(y)f(x-hy)dy$.

Le biais de $f_h(x)$:

$$\begin{aligned} Biais f_h(x) &= \mathbb{E}f_h(x) - f(x) \\ &= \int K(y)f(x-hy)dy - f(x). \end{aligned}$$

La variance de $f_h(x)$:

$$\begin{aligned} Var f_h(x) &= Var \sum_{i=1}^{n} \frac{1}{nh} K(\frac{x-x_i}{h}) \\ &= \frac{1}{n^2h^2} \int [K(\frac{x-y}{h})]^2 f(y)dy - \frac{1}{n}(\frac{1}{h} \int K(\frac{x-y}{h}) f(y)dy)^2, \end{aligned}$$

avec le changement de variable, $y = \frac{x-u}{h}$. Ainsi en faisant le développement de Taylor à l'ordre 2 au point $y = 0$ de $f(x - hy)$, on obtient :

$$Var f_h(x) = \frac{f(x)}{nh} \int K^2(y)dy - \frac{f'(x)}{n} \int yK^2(y)dy - \frac{1}{n}(f(x) + biais f_h(x))^2.$$

II.3.2 Comportement asymptotique

Le biais :

Théorème II.3.1. *Si on a [22] :*

1. $\lim h = 0$ *et* $\lim |yK(y)| = 0$ *quand* $n \to \infty$;
2. $\sup |K(y)| < \infty$ *et* $|K(y)|dy < \infty$;
3. $\int K(y)dy = 1$.

Alors, l'estimateur $f_h(x)$ est asymptotiquement sans biais, c'est à dire :
$\lim_{n\to\infty} \mathbb{E} f_h(x) = f(x)$, *en tout point x pour lequel la densité f est continue.*

La variance :

Théorème II.3.2. *Si on a [22] :*

1. $\lim h = 0$ *et* $\lim |yK(y)| = 0$ *quand* $n \to \infty$;
2. $\sup |K(y)| < \infty$ *et* $|K(y)|dy < \infty$;
3. $\int K(y)dy = 1$.

Alors, $\lim_{n\to\infty} (nh) Var f_h(x) = f(x) \int K^2(y)dy$, *en tout point x pour lequel la densité f est continue.*

II.4 Choix du noyau

II.4.1 Noyaux usuels

Les noyaux les plus utilisés dans l'estimation de la densité de probabilité sont donnés dans le tableau suivant :

Noyau	$K(\mu)$
Uniforme	$\frac{1}{2}, \lvert\mu\rvert \leq 1$
Gaussien	$\frac{1}{\sqrt{2\pi}} exp(\frac{-\mu^2}{2}),\ \mu \in \mathbb{R}$
Triangulaire	$(1 - \lvert\mu\rvert), \lvert\mu\rvert \leq 1$
Epanechnikov	$\frac{3}{4\sqrt{5}}(1 - \frac{\mu^2}{5}),\ \lvert\mu\rvert \leq \sqrt{5}$

TAB. II.1 – Noyaux usuels

II.4.2 Noyaux asymétriques

Quoique les méthodes précédantes diminuent le biais, aux bornes, elles restent peu efficaces car le biais reste concidérable 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, Chen [5] et Chen [6] a proposé d'utiliser des noyaux asymétriques, qui sont 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 $[0, +\infty[$).

Le noyau beta [2]

L'idée est d'utiliser le noyau beta 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 :

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

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

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

avec $B(\alpha,\beta)=\frac{\Gamma(\alpha,\beta)}{\Gamma(\alpha)\Gamma(\beta)}$, et la fonction gamma est définie pour $a>0$ par

$$\Gamma(a)=\int_0^\infty e^{-x}x^{a-1}dx[4].$$

Le noyau gamma [2]

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

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

Ce dernier estimateur reste inefficace au voisinage de zéro, pour cela une autre version du noyau gamma avait été proposée, dont la forme est la suivante :

$$K_{(f_h(x),h)}(t)=\frac{t^{f_h(x)-1}e^{\frac{-t}{h}}}{h^{f_h(x)}\Gamma(f_h(x))},$$

tel que

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

II.5 Choix du paramètre de lissage

Une façon courante d'étudier le comportement de l'estimateur à noyau f_h de f et d'évaluer la largeur de la fenêtre h est de mesurer l'erreur quadratique moyenne intégrée (MISE) :

$$MISE(h)=\mathbb{E}\int(f_h-f)^2. \tag{II.1}$$

L'analyse asymptotique fournit un moyen simple d'évaluer la largeur de le fenêtre h. En particulier, sous des hypothèses techniques standard, le MISE est asymptotiquement (quand $n \mapsto \infty$) évalué par le MISE asymptotique (AMISE) :

$$AMISE(h)=n^{-1}h^{-1}R(K)+h^4R(f'')(\int x^2\frac{K}{2})^2, \tag{II.2}$$

où $R(\varphi)=\int\varphi^2(x)dx$ et $\int x^2K=\int x^2K(x)dx$. Cela évalue l'effet du paramètre de lissage h. En particulier, notons que le premier terme (variance intégrée) est grand lorsque

h est trop petit, et le second terme (biais quadratique intégré) est grand lorsque h est trop grand. Une autre caractéristique utile de AMISE(h) est que sa minimisation est simplement calculée par :

$$h_{AMISE} = [\frac{R(K)}{nR(f'')(\int x^2K)^2}]^{\frac{1}{5}}. \tag{II.3}$$

Cette idée fournit une "bonne" largeur de fenêtre. Par exemple, les petites largeurs, de fenêtre sont mieux quand est trés grand et quand la densité est plus difficile a retrouver (parce que l'effet du biais est plus fort). Dans de nombreuses circonstances h_{AMISE} est une bonne approximation de h_{MISE}, mais parfois ce n'est pas le cas, comme indiqué par Marron et Wand [20]. Plusieurs propositions pour sélectionner la largeur de la fenêtre, ont été faites au cours des années. Pour simplifier la compréhension, ces dernières ont été groupées en classes (première et deuxième). On donne ci-après un bref aperçu des méthodes les plus connues de chaque classe.

II.5.1 Première classe (Validation croisée (Cross validation))

Les méthodes de cette classe ont été proposées par Marron [19], Scott [27] et Silverman [30]. Quelques unes d'entre elles sont discutées si-dessous :

1. Validation croisée non biaisée :

Cette méthode appellée validation croisée non biaisée a été proposée par Rudemo [25] et Bowman [3]. Le critère consiste à choisir le paramètre de lissage qui minimise un estimateur convenable de :

$$\begin{aligned} UCV(h) &= \int_{\mathbb{R}} [f_h(x) - f(x)]^2 dx - \int_{\mathbb{R}} f^2(x)dx \\ &= \int_{\mathbb{R}} f_h^2(x)dx - 2\int_{\mathbb{R}} f_h(x)f(x)dx. \end{aligned} \tag{II.4}$$

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 :

$$\int_{\mathbb{R}} f^2(x)dx - 2\int_{\mathbb{R}} f_h(x)f(x)dx. \tag{II.5}$$

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

$$\int_{\mathbb{R}} f_h(x)f(x)dx = \mathbb{E}[f_h(x)]. \tag{II.6}$$

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 :

$$UCV(h) = \int_{\mathbb{R}} f_h^2(x)dx - \frac{2}{n}\sum_{i=1}^{n} f_{h,i}(x_i), \tag{II.7}$$

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

• UCV(h) est un estimateur sans biais de MISE(h)-R(h), ce qui donne

$$\mathbb{E}[\int_{\mathbb{R}} f_h(x)f(x)dx] = \mathbb{E}\frac{1}{n}\sum_{i=1}^{n} f_{h,i}(x_i). \tag{II.8}$$

Finalement, un estimateur sans biais de MISE(h)-R(h) est donné donc par UCV(h). En utilisant la formule (II.7), le critère UCV(h) devient :

$$UCV(h) = \frac{R(K)}{nh} + \sum_{i=1}^{n}\sum_{j=1}^{n}[\int \frac{1}{n^2h^2}K(\frac{x-x_i}{h})K(\frac{x-x_j}{h})dx - \frac{2}{n(n-1)h}K(\frac{x_i-x_j}{h})]. \tag{II.9}$$

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.

Algorithme de la méthode

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 $i \neq j$
$Som1 = \frac{1}{\sqrt{2\pi}}exp(-(\frac{x_i-x_j}{\sqrt{2\pi}})^2)$,
$Somme1 = Somme1 + Som1$,
$Som2 = \frac{1}{\sqrt{2\pi}}exp(-(\frac{x_i-x_j}{\sqrt{2\pi}})^2)$,
$Somme2 = Somme2 + Som2$,
Fin pour
$UCV(h) = \frac{2}{\sqrt{2\pi}(n-1)h} + \frac{1}{n^2h}Somme1 - \frac{2}{n(n-1)h}Somme2$,
$h_{ucv} = \min_h UCV(h)$.

2. Validation croisée biaisée :

Un critère de validation croisée biaisée, a été introduit par Scott et Terrell [28] 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 :

$$AMISE = \frac{h^4}{4}\delta_K^4 R(f'') + \frac{R(K)}{nh}. \tag{II.10}$$

Le paramètre de lissage basé sur la méthode de validation croisée 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 II.5.1. "Scott et Terrell" [28]
Supposons que le noyau K satisfait aux conditions suivantes :

$$\int K''(\mu)d\mu = 0, \mu_1(K'') = \int \mu K''(\mu) = 0, \mu_2(K'') = \int \mu^2 K''(\mu) = 2; \tag{II.11}$$

On obtient le développement asympotique :

$$\mathbb{E}[R(f_h'')] = R(f'') + \frac{R(K'')}{nh^5} + o(h^2). \tag{II.12}$$

Proposition II.5.1. "Scott et Terrell" [28]
Soit $X_1, X_2, ..., X_n$ un n-échantillon issu d'une variable aléatoire X de fonction de densité f. Pour un noyau K on obtient :

$$BCV(h) = \frac{R(K)}{nh} + h^4 \frac{\mu_{2^2(K)}}{4n^2} \sum_{i=1}^{n}\sum_{j=1}^{n} K_h^2(X_i - X_j). \tag{II.13}$$

Proposition II.5.2. Soit $X_1, X_2, ..., X_n$ un n-échantillon issu d'une variable aléatoire X de fonction de densité f. En choisisant le noyau gaussien on obtient :

$$BCV(h) = \frac{1}{2nh\sqrt{\pi}} + \frac{1}{64n^2h\sqrt{\pi}} \sum_{i=1}^{n}\sum_{j=1}^{n} [(\frac{x_i - x_j}{h})^4 - 12(\frac{x_i - x_j}{h})^2 + 12] exp[-\frac{(x_i - x_j)^2}{4h^2}]. \tag{II.14}$$

Algorithme de la méthode

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 $i \neq j$, $x = (\frac{x_i - x_j}{h})$;
$BCV(h) = BCV(h) + exp(-\frac{x^2}{4})(3 - 3x^2 + \frac{1}{4}x^4)$,
Fin pour
$BCV(h) = \frac{1}{2nh\sqrt{\pi}} + \frac{1}{16n^2h\sqrt{\pi}}BCV(h)$,
$h_{bcv} = \min_h BCV(h)$.

II.5.2 Deuxième classe (Méthodes de ré-injection (plug-in))

De nombreuses méthodes de ré-injection ont été développés. Certains d'entre elles montrent de meilleures propriétés asymptotiques et d'autres de meilleures performances en particulier les petites simulations des échantillons. Ici on présente la sélection par trois approches différentes.

Théorème II.5.1. *(Scott) [26]*
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* $h(n) \to 0$ *et* $nh(n) \to \infty$*, on a le développement asymptotique :*

$$MISE = \frac{h^4}{4}\sigma_K^4 \int (f'')^2 + \frac{\int K^2}{nh} + o(h^5 + \frac{1}{n}). \tag{II.15}$$

Où $\mathbb{L}^2$ *: L'ensemble des fonctions définies sur* $\mathbb{R}$*, telles que* $\int |f(x)|^2 dx < \infty$*. L'erreur quadratique intégrée moyenne asymptotique est alors de la forme :*

$$AMISE = \frac{h^4}{4}\sigma_K^4 R(f'') + \frac{R(K)}{nh}, \tag{II.16}$$

où $R(g) = \int g^2(x)dx$*, pour toute fonction g.*

On remarque que le premier terme du membre à droite du développement (II.16) est un terme de biais, alors que le second est un terme de variance. On constate que dans le AMISE, le terme de biais est fonction croissante en h alors que le terme de 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 provequera une augmentation de la variance et une diminution du biais (phénomène de sous-lissage).

Pour obtenir le paramètre de lissage h^* qui minimise le AMISE, il suffit de résoudre le système suivant :

$$\begin{cases} \frac{dAMISE}{dh} = 0, \\ et \\ \frac{d^2AMISE}{dh^2} > 0. \end{cases}$$

À partir de l'expression (II.16), on a :

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

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

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

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

h^* peut aussi s'écrire :

$$h^* = \psi(K)\varphi(f)n\frac{-1}{5}, \tag{II.17}$$

où $\psi(K) = [\frac{R(K)}{\sigma_K^4}]^{\frac{1}{5}}$ et $\varphi(f) = [\frac{1}{R(f'')}]^{\frac{1}{5}}$ 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 :

$$AMISE^* = \frac{5}{4}[\sigma_K R^4(K)R(f'')]^{\frac{1}{5}} n^{\frac{-4}{5}}. \tag{II.18}$$

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

1. Règle du pouce (Rule of thumb)

L'idée retourne à Deheuvels [9] et à été ensuite dévellopée par Silverman [30]. Dans la formule MISE il suffit d'assigner une valeur au terme $R(f'')$ pour obtenir une estimation de h. Si on choisit f comme étant la loi normale de moyenne 0 et de variance σ^2, on a alors :

$$R(f'') = \int (f''(x))^2 dx = \frac{3}{8}\sqrt{\pi}\sigma^{-5}. \tag{II.19}$$

De plus, si on utilise un noyau gaussien alors la valeur de h^* notée dans ce cas par h_{ROT} est donnée par :

$$\begin{aligned} h_{ROT} &= (4\pi)^{\frac{-1}{10}}[\frac{3}{8}\pi\frac{-1}{2}\sigma]n^{\frac{-1}{5}} \\ &= (\frac{4}{3})^{\frac{1}{5}}\sigma n^{\frac{-1}{5}} \\ &= 1.06\sigma n^{\frac{-1}{5}}. \end{aligned} \tag{II.20}$$

2. Estimateur de Sheather et Jones

On se souvient que le AMISE de l'estimateur à noyau est minimisé en

$$h_{AMISE} = [\frac{R(K)}{nR(f'')(\int x^2K)^2}]^{\frac{1}{5}}.$$

Comme f'' est inconnue, Sheather et Jones [29] proposent d'utiliser la méthode du noyau pour estimer $R(f'')$. Pour ce faire, on note que pour un noyau différentiable deux fois "L" et un paramètre de lissage h_0

$$\begin{aligned} f''_h(x) &= \frac{d^2}{dx^2}\{\frac{1}{nh_0}\sum_{i=1}^{n} L(\frac{x-X_i}{h_0})\} \\ &= \frac{1}{nh_0^3}\sum_{i=1}^{n} L''(\frac{x-X_i}{h_0}). \end{aligned} \tag{II.21}$$

Dans le présent contexte, il est important de noter que l'estimation optimale de f'' ne se fait pas nécessairement avec le même paramètre de lissage optimal que l'estimation de f. Dans le cas où K=L est le noyau gaussien, Sheather et Jones suggèrent une stratègie en deux étapes pour estimer h^*_{AMISE} de f_h.
À la première étape, on choisit un $h_0 \propto n^{-1/7}$ (proportionnel) conduisant à f''_h ; à la seconde étape, on calcule

$$h^*_{AMISE} = [\frac{R(K)}{nR(f''_h)(\int x^2K)^2}]^{\frac{1}{5}}.$$

3. Bootstrap dans l'estimation de la densité de probabilité

a. Bootstrap dans l'estimation globale de la densité de probabilité :

Soit h_0 le paramètre qui minimise l'erreur quadratique moyenne intégrée par l'une des méthodes précédentes. Pour calculer la valeur de la fenêtre par la technique de bootstrap, on doit ré-échantillonner par cette technique à partir de l'échantillon initial et construire ensuite l'estimateur de bootstrap qui s'écrit alors sous la forme :

$$f_h^j(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - X_i^j}{h}), pour j = 1, 2, ..., B; \tag{II.22}$$

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 :

$$\frac{1}{B}\sum_{j=1}^{B}\int (f_h^j(x) - \bar{f}_h^*(x))^2, \tag{II.23}$$

avec,

$$\bar{f}_h^*(x) = \frac{1}{B}\sum_{j=1}^{B} f_h^j(x). \tag{II.24}$$

Nous construisons un estimateur initial de la densité f_{h_0}. Nous ré-échantillonnons ensuite 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 :

$$BMISE(h, h_0) = \frac{1}{B}\sum_{j=1}^{B}\int (f_h^j(x) - f_{h_0(x)})^2 dx. \tag{II.25}$$

b. 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 issu de f. Pour estimer f à un point x, nous utilisons l'estimateur de Parzen-Rozenblatt :

$$f_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - X_i}{h}), \tag{II.26}$$

où K est un noyau tel que $\int_{\mathbb{R}} K(x)dx = 1$ et $h > 0$ le paramètre de lissage.
Le MSE (Mean Squared Error) 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 (Silvermann, [30]), le h optimal au sens de minimisation de MSE(x) par rapport au paramètre h, est asymptotiquement :

$$h^* \sim n^{\frac{-1}{5}}.S^*, \tag{II.27}$$

où : $S^* \equiv S^*(x) = (\frac{f(x)R(K)}{[f''(x)\sigma_K^2]^2})^{\frac{1}{5}}$.
Nous constatons à partir de (II.23) que le choix du paramètre est réduit à un problème de sélection d'un scalaire s dans $h = n^{\frac{-1}{5}}s$. On remplace ensuite ce dernier dans $f_h^j(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x-X_i^j}{h})$, pour $j = 1, 2, ..., B$; on aura :

$$f_s^j(x) = \frac{1}{n^{\frac{4}{5}}s}\sum_{i=1}^{n} K(\frac{x - X_i^j}{n^{\frac{-1}{5}}s}), pour j = 1, 2, ..., B; \tag{II.28}$$

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

$$MSE^*(f_s^*(x), f_h(x)) = \mathbb{E}((f_s^*(x), f_h(x))^2), \tag{II.29}$$

avec :$f_s^* = \frac{1}{B}\sum_{j=1}^{B} f_s^j(x)$ et $f_h(x)$ l'estimateur de f(x) obtenu à partir de l'échantillon initial et h est calculé par la formule du paramètre h :$h = n^{\frac{-1}{5}}S$, avec $S \equiv S(x)$
Enfin le paramètre de lissage est alors sélectionné par :

$$h^* = n^{\frac{-1}{5}}.argmin MSE^*(f_s^*(x), f_h(x)), \tag{II.30}$$

dans le cas d'existence du minimum [7].

II.6 Conclusion

Dans ce chapitre nous avons parlé de l'estimation de la densité par la méthode du noyau. La première partie a été consacrée aux propriétés d'un estimateur a noyau ainsi qu'à son comportement asymptotique. Cet estimateur dépend de deux paramètres : K qui est le noyau (symétrique ou asymétrique) et h qui est le paramètre de lissage. Pour finir, nous avons exposé les deux classes des méthodes du choix du paramètre de lissage.
On a conclu que le choix du noyau n'est pas influent lors de l'estimation d'une densité symétrique sur $\mathbb{R}$. Par contre, il est important de choisir un noyau asymétrique lorsque la densité est définie sur $[0, +\infty[$, pour contourner le problème du biais aux bornes.

Chapitre III

Simulation

III.1 Introduction

R est un logiciel statistique, ou plus exactement un langage, introduit en 1993 par deux chercheurs de l'université d'Auckland (Nouvelle-Zélande) : Robert Gentleman et Ross Ihaka [16]. C'est un logiciel libre, développé à présent par la communauté des participants au sein du R-project et proposé pour les trois systèmes d'exploitation : Unix/Linux, Windows et MacOS. C'est enfin un logiciel modulaire, de nombreux packages complémentaires offrent une grande variété de procédures mathématiques ou statistiques, incluant les méthodes économétriques complexes, le traitement des séries chronologiques et l'analyse des données. Dans ce chapitre, nous allons appliquer les trois méthodes présentées antérieurement (bootstrap, validation croisée et la méthode ré-injéction) à l'aide du logiciel R afin de faire une étude comparative du choix de la fenêtre, entre ces méthodes.

III.2 Plan de simulation

Afin d'illustrer ou de vérifier l'influence de l'application de la technique de bootstrap sur le choix de la fenêtre, on effectue une simulation de différentes lois (normale, exponentielle et gamma). La simulation est organisée selon les étapes suivantes :

Étape1 : Générer un n-échantillon i.i.d d'une densité de probabilité.
Étape2 : Construire B échantillons de bootstrap.
Étape3 : Estimer h_{boot} et h^* par une méthode de sélection.
Étape4 : Estimer $f_{h^*}(x)$ et $f_{h_{boot}}(x)$.

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

- **Densités :**
  1. Normale centrée réduite
  2. Exponentielle de paramètre $\lambda = 1$
  3. Loi gamma
- **Taille d'échantillon d'étape 1 :**
  1. n=100
  2. n=500
  3. n=1000
- **Nombre de réplications de bootstrap :**
  1. B=10
  2. B=50
  3. B=100
- **Méthode de sélection :**
  1. SJ implémente la méthode "Sheather et Jones", l'estimateur de la fenêtre sera noté h_{SJ}.
  2. UCV utilise la méthode de validation croisée biaisée, l'estimateur de la fenêtre sera noté h_{ucv}.

– **Les noyaux**

1. noyau gaussien
2. noyau gamma

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

Étape1 : Générer un n-échantillon i.i.d d'une densité de probabilité.

Étape2 : Construire B échantillons de bootstrap.

Étape3 : Estimer h_b $(b = 1, ..., B)$ pour chaque échantillon de bootstrap.

Étape4 : Calculer h_{boot} par :

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

III.3 Résultats

Dans cette section, nous allons présenter les résultats des simulations (sous forme de tableaux et graphes) obtenus pour l'estimation du paramètre de lissage d'une densité de probabilité (normale, exponentielle, gamma), en utilisant un noyau adéquat (gaussien ou gamma) :

III.3.1 Cas d'une loi normale :

On calcule d'abord le paramètre de lissage h avec la méthode de bootstrap dans l'estimation locale de la densité de probabilité, on le note par ($h_{boot.l}$) puis avec la méthode de bootstrap globale dans l'estimation de la densité de probabilité, on le note ($h_{boot.g}$), cela pour de differents nombres de réplications, on choisit le meilleur d'entre eux.
Voici les résultats :

n	B	$h_{boot.l}$	$h_{boot.g}$
100	10	0.3102889	0.3215789
100	50	0.3734178	0.37456821
100	100	0.3799893	0.3800741
500	10	0.2789367	0.2892011
500	50	0.3155186	0.3224561
500	100	0.3018004	0.304261
1000	10	0.2798014	0.2854212
1000	50	0.2713226	0.2810224
1000	100	0.2925196	0.3015460

TAB. III.1 – Résultats de simulations effectuées sur la loi normale, pour déterminer les largeurs de fenêtres optimales ($h_{boot.l}$) et ($h_{boot.g}$)

D'après les résultats obtenus dans le tableau précédent, on prend :

- **Pour n=100** $h_{boot.l} = 0.3102889$
- **Pour n=500** $h_{boot.l} = 0.2789367$
- **Pour n=1000** $h_{boot.l} = 0.2713226$

- **Pour n=100** $h_{boot.g} = 0.3215789$
- **Pour n=500** $h_{boot.g} = 0.2892011$
- **Pour n=1000** $h_{boot.g} = 0.2810224$

On remarque que la méthode de bootstrap dans l'estimation locale de la densité de probabilité est meilleure pour la loi normale, que celle de bootstrap dans l'estimation globale de la densité de probabilité.
Les resultats de h_{ucv} et h_{SJ} sont résumés dans le tableau suivant :

n	h_{ucv}	h_{SJ}	$h_{boot.l}$	$h_{boot.g}$
100	0.4229666	0.3961440	0.3402889	0.3215789
500	0.3456708	0.2934291	0.2789367	0.2892011
1000	0.3447171	0.2933750	0.2713226	0.2810224

TAB. III.2 – Résultats de simulations effectuées sur la loi normale, pour déterminer les largeurs de fenêtres optimales avec la méthode bootstrap, validation croisée et la méthode de Sheather et Jones

Discussion

À partir des résultats des simulations obtenus pour la loi normale, on constate que :

- L'augmentation de n entraîne la décroissance de h_{ucv}, h_{SJ} et h_{boot}.
- La technique de bootstrap nous donne des estimateurs du paramètre de lissage h plus stables que ceux obtenus par les autres méthodes.
- On constate aussi que l'amélioration apportée par la technique bootstrap est indépendante de de nombre de réplication B de bootstrap

Les graphes suivants nous confirment les résultats précédents :

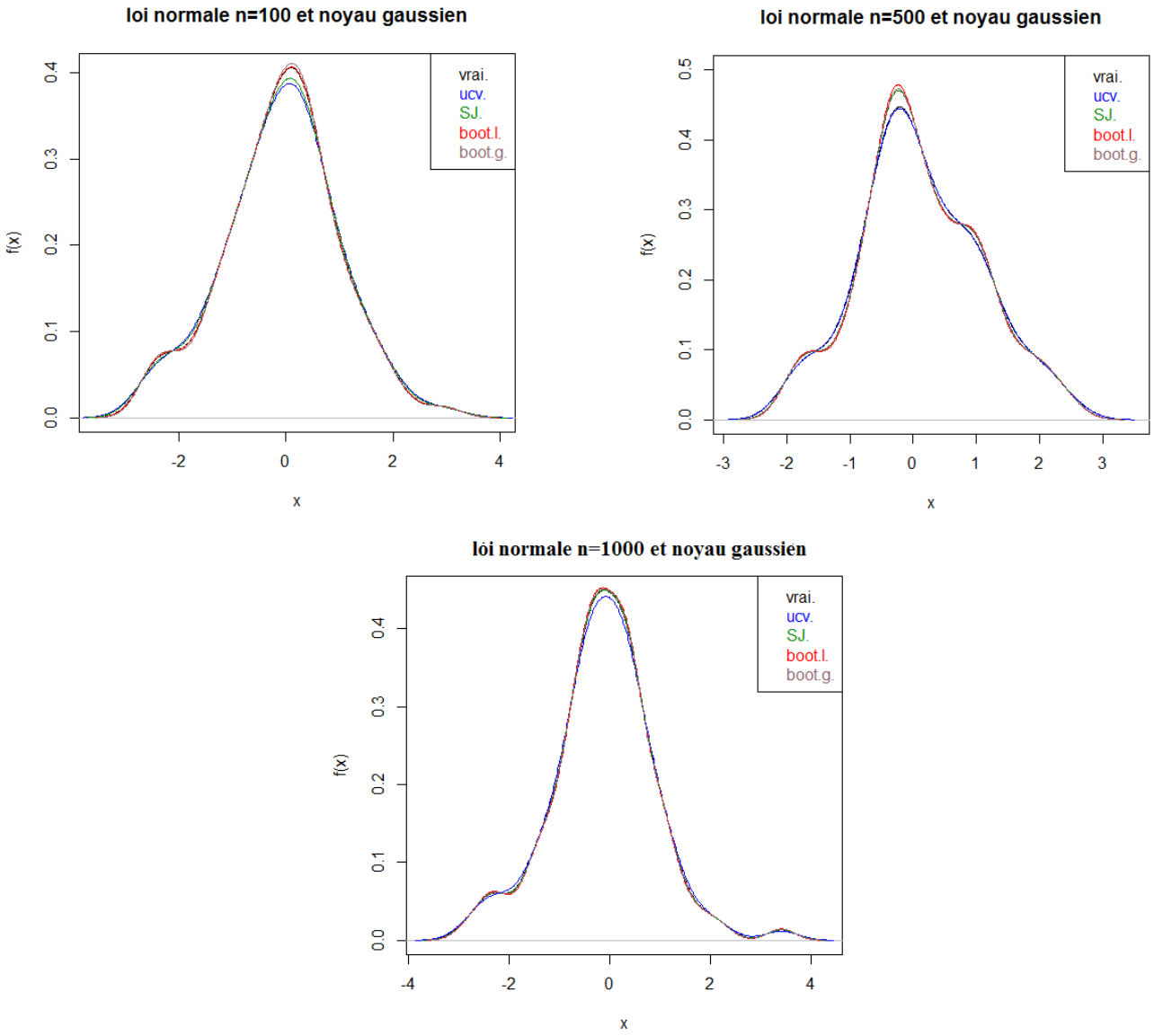

FIG. III.1 – Comparaison entre la densité théorique et celles estimées avec les paramètres de lissage h_{ucv}, h_{SJ} et h_{boot} (loi normale)

III.3.2 Cas d'une loi exponentielle :

a. Avec un noyau gaussien :

En procédant par les mêmes étapes considérées la loi normale, on trouve :

n	B	$h_{boot.l}$	$h_{boot.g}$
100	10	0.3567442	0.3614578
100	50	0.4182518	0.4251780
100	100	0.4305533	0.4470214
500	10	0.3249735	0.3260014
500	50	0.3154154	0.3245701
500	100	0.3106364	0.3154786
1000	10	0.2775839	0.2814570
1000	50	0.2819211	0.28541352
1000	100	0.2862583	0.2901425

TAB. III.3 – Résultats de simulations effectuées sur la loi exponentielle, pour déterminer les largeurs de fenêtres optimales $h_{boot.l}$ et $h_{boot.g}$ (cas d'un noyau gaussien)

D'après les résultats obtenus dans le tableau précédent, on prend :

- **Pour n=100** $h_{boot.l} = 0.3567442$
- **Pour n=500** $h_{boot.l} = 0.3106364$
- **Pour n=1000** $h_{boot.l} = 0.2775839$

- **Pour n=100** $h_{boot.g} = 0.3614578$
- **Pour n=500** $h_{boot.g} = 0.3154786$
- **Pour n=1000** $h_{boot.g} = 0.2814570$

On remarque que la méthode de bootstrap dans l'estimation locale de la densité de probabilité est meilleure pour la loi exponentielle, que celle de bootstrap dans l'estimation globale de la densité de probabilité.

Les résultats h_{ucv} et h_{SJ} sont résumés dans le tableau suivant :

n	h_{ucv}	h_{SJ}	$h_{boot.l}$	$h_{boot.g}$
100	0.19493201	0.1492024	0.3567442	0.3614578
500	0.14280257	0.09674155	0.3106364	0.3154786
1000	0.12915635	0.08646475	0.2775839	0.2814570

TAB. III.4 – Résultats de simulations effectuées sur la loi exponentielle, pour déterminer les largeurs de fenêtres optimales avec la méthode bootstrap, validation croisée et la méthode de Sheather et Jones (cas d'un noyau gaussien)

Discussion

Pour le cas de la loi exponentielle en utilisant un noyau gaussien, les résultats obtenus nous permettent de conclure que :

- La méthode de Sheather et Jones est meilleure car elle donne des estimateurs de paramètre de lissage plus stables.
- Les échantillons suivant une telle loi ne permettent pas une estimation correcte de cette densité. 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.
- L'échec de la technique bootstrap, peut être expliqué par la présence du biais au voisinage de zéro de l'estimateur de la densité exponentielle, car le bootstrap échoue dans l'estimation d'une statistique biaisée.

Les graphes suivants nous confirment les résultats précédents :

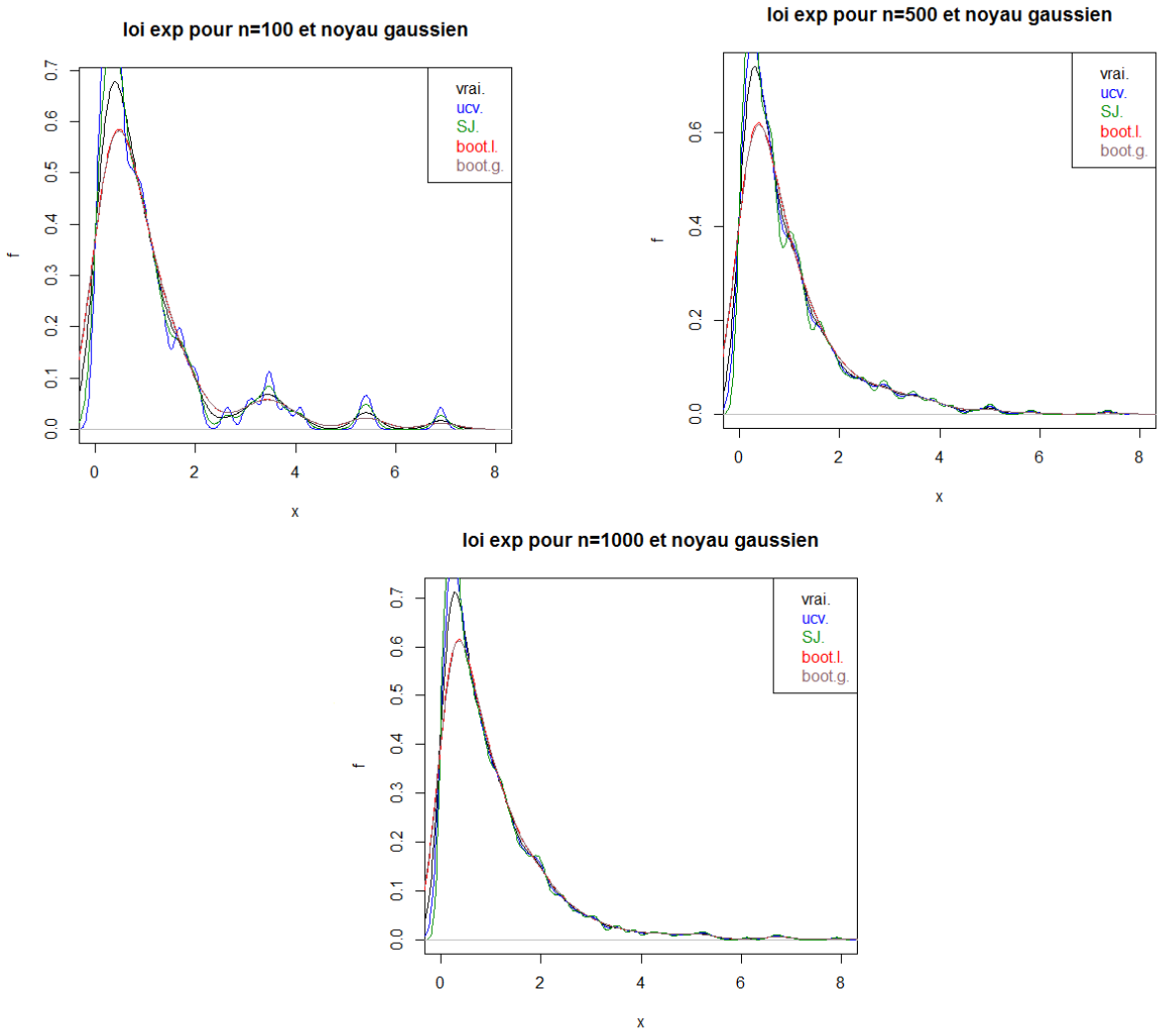

FIG. III.2 – Comparaison entre la densité théorique et celles estimées avec les paramètres de lissage h_{ucv}, h_{SJ} et h_{boot} (loi exponentielle avec un noyau gaussien)

b. Avec un noyau gamma :

En procédants par les même étapes considérées la loi exponentielle on utilisant un noyau gamma, on obtient :

n	B	$h_{boot.l}$	$h_{boot.g}$
100	10	0.0382225	0.0724695
100	50	0.0378311	0.0715426
100	100	0.0359007	0.0690142
500	10	0.0310664	0.0672150
500	50	0.0315154	0.0684210
500	100	0.0301784	0.0670141
1000	10	0.0297153	0.0654120
1000	50	0.0297253	0.0652104
1000	100	0.0282802	0.0649866

TAB. III.5 – Résultats de simulations effectuées sur la loi exponentielle, pour déterminer les largeurs de fenêtres optimales $h_{boot.l}$ et $h_{boot.g}$ (cas d'un noyau gamma)

D'après les résultats obtenus dans le tableau précédant, on choisit :

- **Pour n=100** $h_{boot.l} = 0.0359007$
- **Pour n=500** $h_{boot.l} = 0.0301784$
- **Pour n=1000** $h_{boot.l} = 0.0282802$

- **Pour n=100** $h_{boot.g} = 0.0690142$
- **Pour n=500** $h_{boot.g} = 0.0670141$
- **Pour n=1000** $h_{boot.g} = 0.0649866$

On remarque que la méthode de bootstrap dans l'estimation locale de la densité de probabilité pour la loi exponentielle en utilisant un noyau gamma est meilleure, que celle de bootstrap dans l'estimation globale de la densité de probabilité.

Les resultats de h_{ucv} et h_{SJ} sont résumés dans le tableau suivant :

n	h_{ucv}	h_{SJ}	$h_{boot.l}$	$h_{boot.g}$
100	0.1427809	0.3433508	0.0359007	0.0690142
500	0.13051708	0.3443560	0.0301784	0.0670141
1000	0.13737786	0.3551067	0.0282802	0.0649866

TAB. III.6 – Résultats de simulations effectuées sur la loi exponentielle, pour déterminer les largeurs de fenêtres optimales avec la méthode bootstrap, validation croisée et la méthode de Sheather et Jones (cas d'un noyau gamma)

Ci-après, on donne les graphes correspondant aux résultats précédants :

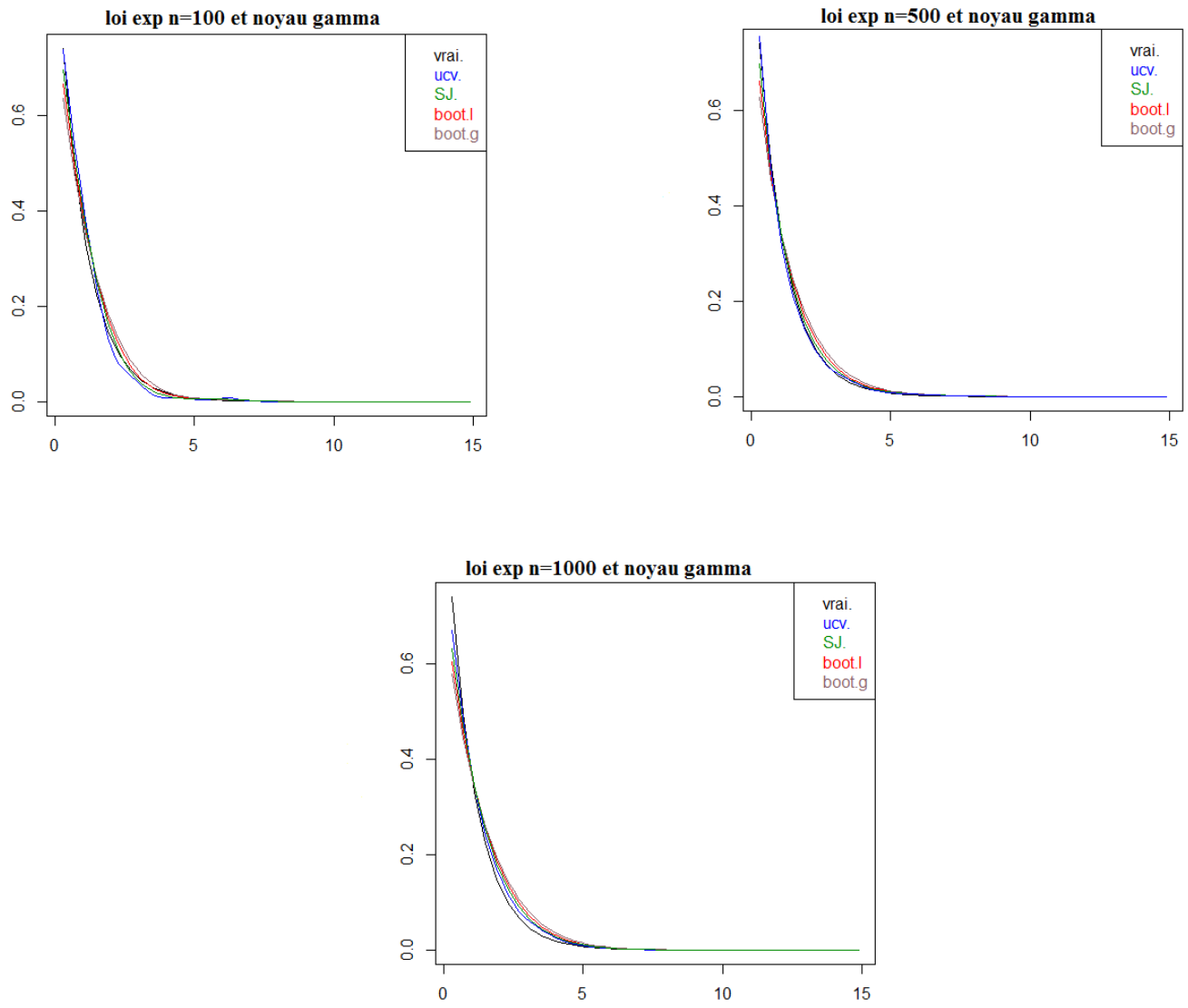

FIG. III.3 – Comparaison entre la densité théorique et celle estimée avec le paramètre de lissage h_{ucv}, h_{SJ} et h_{boot}(loi exponentielle avec un noyau gamma)

Discussion

D'après les graphes, en utilisant un noyau gamma, on remarque que la méthode de bootstrap n'échoue pas pour la loi exponentielle, car le noyau gamma est un noyau de correction aux bornes. Ce qui veut dire que le choix du noyau est important pour les lois définies sur des support fermés ou semi fermés.

III.3.3 Cas d'une loi gamma :

a. Avec un noyau gaussien :

On calcule d'abord $h_{boot.l}$ et $h_{boot.g}$ pour de differents nombres de réplications, on choisit le meilleur d'entre eux.

n	B	$h_{boot.l}$	$h_{boot.g}$
100	10	0.3706433	0.4076540
100	50	0.3590607	0.5127503
100	100	0.2927153	0.4410115
500	10	0.3201944	0.5745138
500	50	0.2954154	0.6083102
500	100	0.3201944	0.6154259
1000	10	0.2798014	0.5126522
1000	50	0.2838451	0.5288761
1000	100	0.2927153	0.5272985

TAB. III.7 – Résultats de simulations effectuées sur la loi gamma, pour déterminer les largeurs de fenêtres optimales $h_{boot.l}$ et $h_{boot.g}$ (cas d'un noyau gaussien)

D'après les résultats obtenus dans le tableau précédent, on prend :

- **Pour n=100** $h_{boot.l} = 0.2927153$
- **Pour n=500** $h_{boot.l} = 0.2854154$
- **Pour n=1000** $h_{boot.l} = 0.2798014$

- **Pour n=100** $h_{boot.g} = 0.4076540$
- **Pour n=500** $h_{boot.g} = 0.5745138$
- **Pour n=1000** $h_{boot.g} = 0.5126522$

On remarque que la méthode de bootstrap dans l'estimation locale de la densité de probabilité pour la loi gamma est meilleure, que celle de bootstrap dans l'estimation globale de la densité de probabilité.
Les résultats de h_{ucv} et h_{SJ} sont résumés dans le tableau suivant :

n	h_{ucv}	h_{SJ}	$h_{boot.l}$	$h_{boot.g}$
100	0.1730615	0.187705	0.2927153	0.4076540
500	0.04634334	0.1026454	0.2854154	0.5745138
1000	0.04030048	0.07508707	0.2798014	0.5126522

TAB. III.8 – Résultats de simulations effectuées sur la loi gamma, pour déterminer les largeurs de fenêtres optimales avec la méthode bootstrap, validation croisée et la méthode de Sheather et Jones (cas d'un noyau gaussien)

Discussion

Les résultats de simulation de la loi gamma avec un noyau gaussien nous permettent de conclure que :

- La méthode de validation croisée est meilleure car elle donne des estimateurs de paramètre de lissage plus stables.
- Les échantillons suivant une telle loi ne permettent pas une estimation correcte de cette densité.
- Contrairement au cas de la densité de la loi normale, l'estimation de la densité de la loi gamma dépend du nombre de réplications B de bootstrap.
- L'échec de la technique bootstrap, peut être expliqué par la présence du biais au voisinage de zéro de l'estimateur de la densité gamma, car le bootstrap échoue dans l'estimation d'une statistique biaisée.

Les graphes suivants nous confirment les résultats précedents :

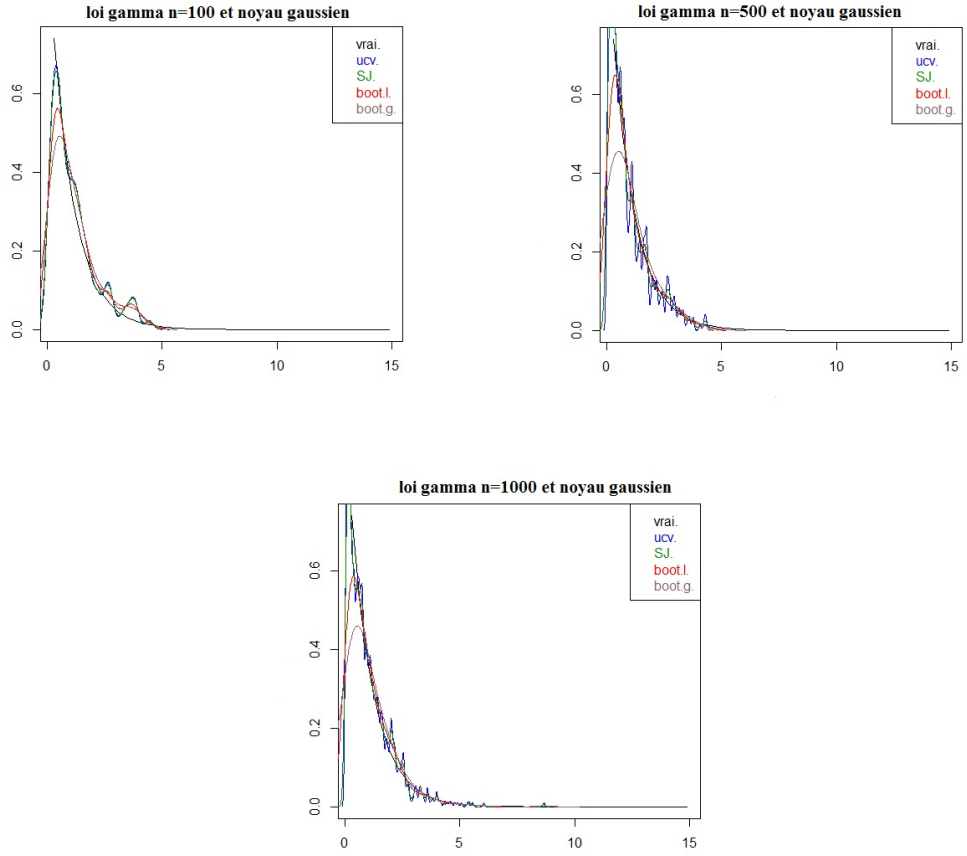

FIG. III.4 – Comparaison entre la densité théorique et celles estimées avec les paramètres de lissage h_{ucv}, h_{SJ} et h_{boot} (loi gamma avec un noyau gaussien)

b. Avec un noyau gamma :

En procédants par les mêmes étapes considérés la loi gamma, on utilisant un noyau gamma, on obtient :

n	B	$h_{boot.l}$	$h_{boot.g}$
100	10	0.2924796	0.3845012
100	50	0.31654497	0.4025610
100	100	0.28447559	0.3015044
500	10	0.2854876	0.322217
500	50	0.2947985	0.3814332
500	100	0.2710784	0.392114
1000	10	0.2754810	0.4024551
1000	50	0.2826981	0.4051447
1000	100	0.2514562	0.312005

TAB. III.9 – Résultats de simulations effectuées sur la loi gamma, pour déterminer les largeurs de fenêtres optimales $h_{boot.l}$ et $h_{boot.g}$ (cas d'un noyau gamma)

D'après les résultats obtenus dans le tableau précédent, on prend :

- **Pour n=100** $h_{boot.l} = 0.28447559$
- **Pour n=500** $h_{boot.l} = 0.2710784$
- **Pour n=1000** $h_{boot.l} = 0.2514562$

- **Pour n=100** $h_{boot.g} = 0.3015044$
- **Pour n=500** $h_{boot.g} = 0.322217$
- **Pour n=1000** $h_{boot.g} = 0.312005$

On remarque que la méthode de bootstrap dans l'estimation locale de la densité de probabilité pour la loi gamma en utilisant un noyau gamma est meilleure, que celle de bootstrap dans l'estimation globale de la densité de probabilité.

Les resultats de h_{ucv} et h_{SJ} sont résumés dans le tableau suivant :

n	h_{ucv}	h_{SJ}	$h_{boot.l}$	$h_{boot.g}$
100	0.325472	0.3152041	0.28447559	0.3015044
500	0.3350142	0.3301245	0.2710784	0.322217
1000	0.3365140	0.3210504	0.2514562	0.312005

TAB. III.10 – Résultats de simulations effectuées sur la loi gamma, pour déterminer les largeurs de fenêtres optimales avec la méthode bootstrap, validation croisée et la méthode de Sheather et Jones (cas d'un noyau gamma)

Ci-après, on donne les graphes correspondant aux résultats précédants :

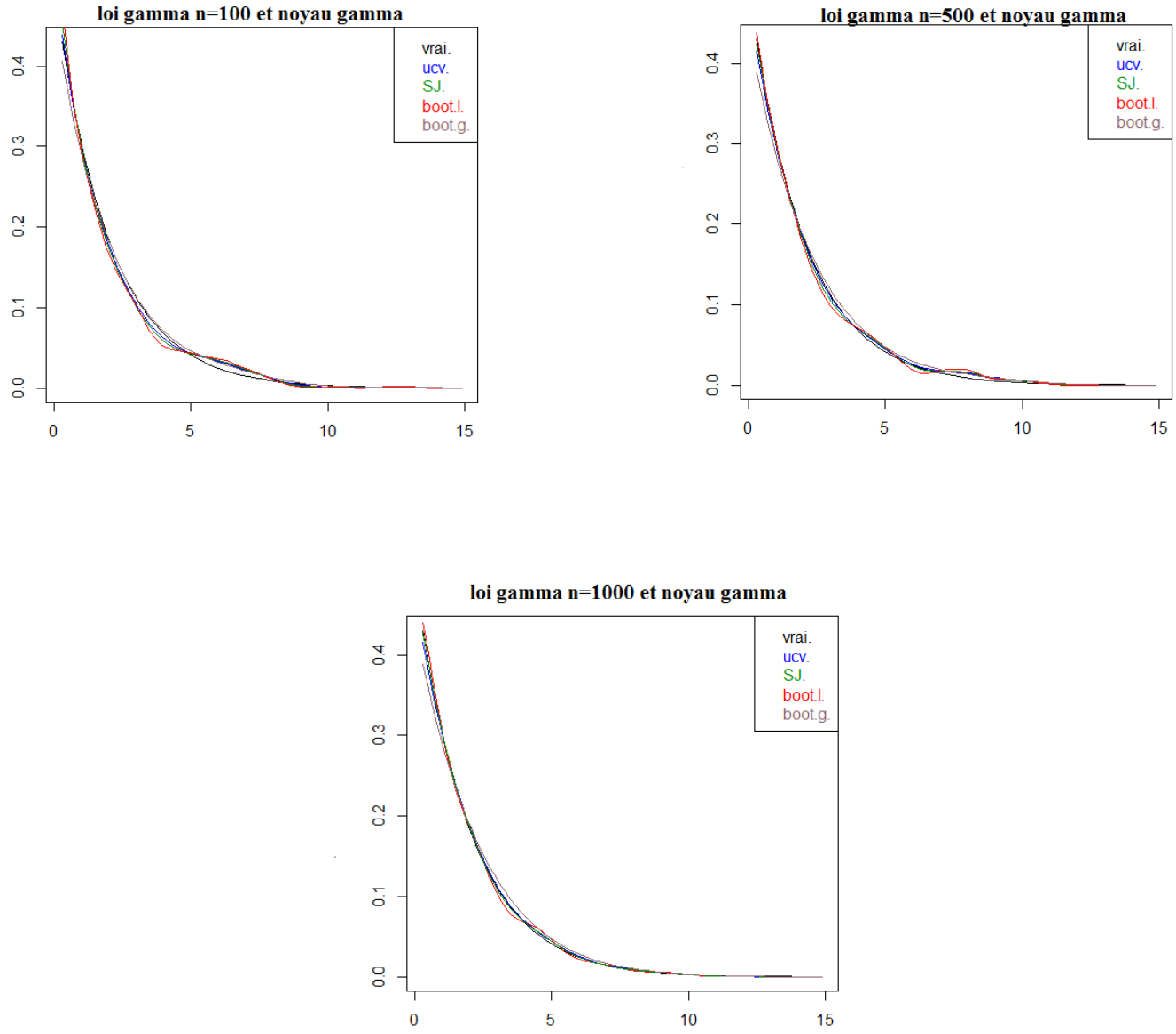

FIG. III.5 – Comparaison entre la densité théorique et celles estimées avec les paramètres de lissage h_{ucv}, h_{SJ} et h_{boot} (loi gamma avec un noyau gamma)

Discussion

D'après les graphes, en utilisant un noyau gamma, on remarque que la méthode de bootstrap n'échoue pas pour la loi gamma, car le noyau gamma est un noyau de correction aux bornes. Ce qui veut dire que le choix du noyau est important pour les lois définies sur des support férmes ou semi fermés.

III.4 conclusion

Les résultats de simulations obtenus nous permettent de conclure que la performance de la technique de bootstrap varie en fonction de la densité à estimer, de la taille de l'échantillon étudié, le choix du noyau, le nombre de réplications de bootstrap et de la procédure de sélection utilisée pour le choix du paramètre de lissage.

La technique bootstrap, lors de l'estimation de la méthode du noyau pour l'estimation de la densité, échoue lorsque l'estimateur contient un biais, comme dans le cas de la loi exponentielle avec un noyau gaussien.

Conclusion générale

Ce travail est une contribution au problème du choix du paramètre de lissage par la technique de bootstrap lors de l'estimation de la densité de probabilité par la méthode du noyau. Nous avons utilisé la technique de bootstrap pour estimer le paramètre de lissage et nous avons étudié l'efficacité et la robustesse de cette technique qui ont été mesurées numériquement par une étude de simulation sur des densités tests présentant differents aspects (loi normale, loi exponentielle et loi gamma).

Dans un premier temps, nous avons exposé la technique de bootstrap afin d'illustrer son principe de base, son intérêt ainsi que son applicabilité en estimation. Cette technique de ré-échantillonage a été rendue possible grâce à la génération des moyan de calculs performants. Le bootstrap est justifié par des propriétés asymptotiques (convergence en loi) lorsque le nombre de réplications (B) croit conjointement avec la taille de l'échantillon (n).

La deuxième partie de ce travail à été consacrée aux propriétés d'un estimateur à noyau ainsi qu'a son comportement asymptotique. Cet estimateur dépend de deux paramètres : K qui est le noyau (symétrique ou assymétrique) et h qui est le paramètre de lissage. Nous avons aussi exposé de deux classes des méthodes du choix du paramètre de lissage, et nous avons parlé de l'influence du choix du noyau lors de de l'estimation d'une densité symétrique et asymétrique.

La dernière partie est consacrée à une étude de simulation dont l'ojectif est de tester l'application de la technique de bootstrap dans la sélection du paramètre de lissage. Nous avons fait une comparaison entre le bootstrap dans l'estimation globale et locale de la densité de probabilité. Nous avons simulé des densités de probabilités présantant différents aspects (loi normale, loi exponentielle, loi gamma) en utilisant deux méthodes de sélection (Sheather et Jones, validation croisée non biaisé). les résultats obtenus montrent que :

- La technique de bootstrap nous donne des estimateurs du paramètre de lissage h plus stables que ceux obtenus par les autres méthodes.
- On constate aussi que l'amélioration apportée par la technique bootstrap est indépendante du nombre de réplications B de bootstrap.
- À cause de la présence d'un biais important au voisinage de zéro, la technique de bootstrap échoue dans l'estimation des densités comme la loi exponentielle, dans le cas d'usage d'un noyau symétrique.
- L'étude montre aussi que le choix du noyau est important pour les densités à support compact, car il influe sur les performances de l'estimateur.
- Le choix du noyau dans l'estimation de la densité de la loi normale n'est pas un problème pour la sélection du paramètre de lissage.

Perspectives :

Parmi les perspectives de ce travail, il serait intérresant de :

- Effectuer des simulations sur des densités complexes par exemple les densités qui possèdent des discontinuités.
- Effectuer des simulations sur des données réelles.
- Appliquer la technique bootstrap aux autres méthodes d'estimation non paramétrique de la densité de probabilité.
- Utiliser d'autres méthodes des deux classes (ré-injection et validation croisée). pour déterminer le paramètre de lissage

Bibliographie

[1] D. Bernard. *Simulation et modélisation, Cours Master 2, Université Rennes I, (4) : 109-115, Sept 2005.*

[2] T. Bouazmarni, O. Scaillet. *Consistency of asymmetric kernel density estimators and smoothed histograms with application to income data. Econometric Theory, (21) : 390-412, 2003.*

[3] A.W. Bowman. *An alternative method of cross-validation for the smoothing density estimates, Biometrika, (71) : 553-560, 1984.*

[4] B.M. Brown, S. Chen. *Beta-bernstein smoothing for regression curves with compact supports. Scandanavian Journal of Statistics,(26) : 47-59, 1999.*

[5] S. Chen. *A beta kernel estimation for density functions. Computational Statistics and Data Analysis, (31) : 131-145, 1999.*

[6] S. Chen. *Probability density functions estimation using gamma kernels. Annals of the institute of Statistical Mathematics, (52) : 471-480, 2002.*

[7] M. Cherfaoui. *Bootstrap dans l'estimation non paramétrique de la densité de probabilité et la courbe de régression de la moyenne. Thèse de magister, Université de Bejaia, 2009.*

[8] M.P. Cohen. *The bayesian bootstrap and multiple imputation for inequal probability sample designs. National Center of Education Statistics, New Jersey, 635-638, 1997.*

[9] P. Deheuvels. *Estimation non paramétrique de la densité par histogrammeS généralisé Revue Statistique Appliquée, (25) : 5-42, 1977.*

[10] B. Efron. *Bootstrap methods : Another look at the jacknife. Annals of Statistics, (7) : 1-26, 1979.*

[11] B. Efron. *Second thoughts on the bootstrap. Statistical science, (18) : 135-140, 2003.*

[12] B. Efron. *The jacknife, the bootstrap and other resampling plans. Technical Report n.63, Stanford University, Departement of statistics. (5) : 35-48, Dec 1981.*

[13] B. Efron, Tibshirani. *An introduction to the bootstrap. Chapman and Hall, New York, 1993.*

[14] W. Hardle. *Applied nonparametric regression, Cambridge University Press, UK, 1990.*

[15] E. Herrmann. *Local bandwidth choice in kernel regression estimation. J. Comput. Graph. Statists, (6) : 35-54, 1997.*

[16] R. Ihaka, R. Gentleman. *R : A language for data analysis and graphics, University of Aukland, New Zealand, 2001*

[17] C. Jones, J.S. Marron and P. Sheather. *A brief survey bandwidth selection for density estimation. Amer. Statist. Assoc,(89) :401-407, 1996.*

[18] C. Leger, N.Altman. *Bootstrap choice of tuning parameters. Anu. Inst. Statist. Math, (42) :709-735, 1990.*

[19] J.S. Marron. *Automatic Smoothing Parameter Selection : A Survey Emperical Economics, (13) :187-208, 1989.*

[20] J.S. Marron, M.P. Wand. *Exact Mean Integrated Squared Error.The Annals of Statistics, (20) : 712-736, 1992.*

[21] R. Palm. *Utilisation du bootstrap pour les problèmes statistiques liés à l'estimation des paramètres. Biotechnol. Agron. Soc, 29 Avril 2002.*

[22] E. Parzen. *On estimation of a probability density function and mode. Ann. Math. Statist, (33) : 1065-1076, 1962.*

[23] K. Pearson. *On the systematic fitting of curves to observations and measurments. Biometrica, (1) : 265-303.*

[24] M. Rosenblatt. *Remarks in some nonparametric estimates of a density function. Ann. Matt. Statist., (27) : 629-647, 1988.*

[25] M. Rudemo. *Empirical choice of histograms and kernel density estimators. Skandanavian Journal of Statistics, (9) : 65-78, 1982.*

[26] D.W. Scott. *Averaged shift histograms : effective nonparametric density estimators in several dimensions. The annals of statistics, (13) : 1024-1040, 1985.*

[27] D.W. Scott. *Multivariate density estimation : Theory, Practice and Visualisation, New York : John Willey, 1992.*

[28] D.W. Scott, G.R. Terell. *Biased and unbiased cross-validation in density estimation , Journal of the American Statistical Association, (82) : 1131-1146, 1987.*

[29] S.J. Sheather, M.C. Jones. *Using Nonstochastic Terms to Advantage in Estimating Integrated Squared Density Derivatives, Statistic and Probability Letters, (11) : 511-514, 1991.*

[30] B.W. Silvermann. *Density estimation for statistics and data analysis, London : Chapman and Hall,1986.*

[31] M.P. Wand, M.C. Jones. *Kernel smoothing, London, Chapman and Hall, 1994.*

[32] N. Zougab. *Etude comparative des méthodes de sélection du paramètre de lissage dans l'estimation de la densité de probabilitépar la méthode du noyau. Thèse de magister, University de Bejaia, Mai 2007.*

RÉSUMÉ

La technique de bootstrap qui se base sur le principe de ré-echantillonnage a apporté beaucoup de solutions pour des problèmes statistiques divers.

Le présent travail fait l'objet de l'estimation de densité pour choisir un paramètre de lissage h par la méthode du noyau, à partir d'un échantillon $X_1, X_2, ..., X_n$. Pour cele nous nous sommes interéssés à l'estimation de certaines lois les plus connues (la loi mormale, la loi exponentielle et la loi de gamma), nous avons proposé l'application de la technique de bootstrap pour le choix du paramètre de lissage h. Nous avons également étudié l'influence du choix du noyau.

La simulation sur des échantillons de taille n=100, 500 et 1000 pour le choix de la fenêtre 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 que dans certains cas le choix du noyau est important particulièrement dans l'estimation des densités de probabilité à support compact ou semi compact, comme la loi exponentielle.

Mots-clés : Bootstrap, densité de probabilité, noyau, paramètre de lissage, estimation, méthodes classiques.

ABSTRACT

The bootstrap technique which is based on the principal of re-sampling has provided many solutions for various statistical problems.

The aim of this work is to estimate a density to choose a smoothing parameter h with the kernel method, from a sample $X_1, X_2, ..., X_n$. For that, we focused on the estimation of the most known laws (exponential, normal and gamma), we proposed the application of the bootstrap technique for choosing the smoothing parameter. We also studied the influence of the choice of the kernel.

the simulation on samples of size n=100, 500 and 1000 for the bandwidth choice 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 the kernel is particularly important in estimating the probability densities of compact or semi compact support, such as the exponential law.

Keywords : Bootstrap, probability density, kernel, smoothing parameter, estimation, classical methods.