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

**Université Mohamed KHIDER - Biskra**

**Faculté des Sciences Exactes et des Sciences de la Nature et de la Vie**

**Département de Mathématiques**

**Thèse présentée en vue de l'obtention du diplôme de**

## Doctorat en Sciences - Mathématiques

**Option : Statistique**

Par
**ZERFAOUI Karima**

**Thème :**

# Sur l'estimation non paramétrique pour les données doublement tronquées

Soutenu Publiquement le 02 /11/2023, devant le Jury composé de :

| | | | |
|---|---|---|---|
| **Mr. MOKHTARI Zouhir** | **Professeur** | **Université Batna 2** | **Président** |
| **Mr. YAHIA Djabrane** | **Professeur** | **Université de Biskra** | **Rapporteur** |
| **Mr. BENATIA Fatah** | **Professeur** | **Université de Biskra** | **Examinateur** |
| **Mr. SAYAH Abdallah** | **Professeur** | **Université de Biskra** | **Examinateur** |
| **Mr. DJEFFAL El-Amir** | **Professeur** | **Université Batna 2** | **Examinateur** |
| **Mr. DJENAIHI Youcef** | **MCA.** | **Université Sétif 1** | **Examinateur** |

# Remerciements

Je remercie tout d'abord ALLAH de m'avoir donné la volonté, la force et le courage pour pouvoir surmonter les moments difficiles et m'avoir aidé à aboutir à ce travail.

Je témoigne une reconnaissance particulière à mon directeur de thèse Monsieur YAHIA Djabrane. Pour son soutien, sa patience et ses conseils avisés tout au long de mon travail de recherche. Grace à lui j'ai pu acquérir de nouvelles compétences, apprendre de nouvelles méthodes de travail et développer une meilleure compréhension de mon domaine d'étude. Je suis très reconnaissante pour son engagement et son dévouement envers mon travail.

Je tiens à remercier infiniment Mr. MOKHTARI Zouhir, Professeur à l'université de Batna et Directeur de l'Office National des Publications pour avoir accepté de présider le Jury de ma soutenance. Mes remerciements vont également au membre de Jury : Prof. BENATIA Fatah et Prof. SAYAH Abdallah de l'université de Biskra, Mr. DJEFFAL El-Amir de l'université Batna 2 et Dr. DJENAIHI Youcef de l'université Sétif 1.

Enfin, je voudrais exprimer ma gratitude envers ma famille pour leur amour, leur soutien et leur encouragement constants. Leur soutien, moral et leur présence ont été une source de force et de motivation pour moi tout au long de cette aventure. Leurs encouragements m'ont aidé à persévérer et à travailler dur pour atteindre mes objectifs

# Résumé

Le phénomène de la double troncature simultanée à gauche et à droite apparaît dans divers domaines, tels que la recherche médicale et l'économie. Le problème de l'estimation de la fonction de mode ou du mode conditionnel pour ce type de données n'a pas été abordé dans la littérature statistique. Dans cette thèse, nous proposons un nouvel estimateur à noyau du mode dans le cadre d'un modèle aléatoire doublement tronqué. Nous établissons la forte consistance avec un taux de convergence pour l'estimation proposée et indiquons sa normalité asymptotique. Une étude de simulation est réalisée pour illustrer et évaluer le comportement sur un échantillon fini de l'estimateur proposé. L'estimation non paramétrique de la fonction du mode conditionnel pour les données sous double troncature est aussi étudier dans cette thèse.

**Mots clés :** Normalité asymptotique ; estimateur du mode ; troncature à droite et à gauche ; taux de convergence.

# Abstract

The phenomenon of simultaneous left and right double truncation appears in various fields, such as medical research and economics. The problem of estimating the mode function or the conditional mode for this type of data has not been mentioned in the statistical literature. In this thesis, we propose a new kernel estimator of the mode in the framework of a doubly truncated random model. We establish the strong consistency with a convergence rate for the proposed estimate and indicate its asymptotic normality. A simulation study is performed to illustrate and evaluate the behavior on a finite sample of the proposed estimator. The nonparametric estimation of the conditional mode function for data under double truncation is also studied in this thesis.

**Keywords:** Asymptotic normality; mode estimator; right and left truncation; convergence rate.

# Table des Matières

# Abréviations et Notations

| | |
|---|---|
| NPMLE | Estimation non paramétrique du maximum de vraisemblance |
| v.a | variable aléatoire |
| i.i.d. | Indépendantes et identiquement distribuées |
| MSE | Erreur quadratique moyenne |
| MISE | Erreur quadratique moyenne intégrée |
| $X$ | Variable d'intérêt |
| $X^*$ | Variable observée (sous troncature) |
| $f(.)$ | Fonction de Densité |
| $F(.)$ | Fonction de distribution |
| $f(./x)$ | Densité conditionnelle |
| $h_n$ | Fenêtre ou bandwith : suite de réels positifs |
| K(.) | Fonction noyau |
| $\theta$ | Mode |
| $\theta(y)$ | Mode conditionnel |
| $F_n$ | Distribution empirique |
| df | Fonction de distribution |
| $I\{A\}$ | Fonction indicatrice d'ensemble A |
| $N(0,1)$ | Loi Normale centrée et réduite |
| $\xrightarrow{L}$ | Convergence en Loi |
| $z_{1-\nu/2}$ | Quantile d'ordre $1-\nu/2$de la distribution normale $N(0,1)$. |
| $Exp(b)$ | Loi Exponentielle de paramètre b |
| N | Taille de l'échantillon globale |
| N | Taille de l'échantillon observé |
| $\|.\|_\infty$ | Norme infinie (sup) |

# Liste des Tableaux

# Liste des Figures

# Introduction générale

Les méthodes non paramétriques pour les données tronquées apparaissent dans une variété de domaines, y compris l'astronomie, la médecine et l'économie. Lynden-Bell (1971) fut le premier à avoir étudié l'estimation non paramétrique pour des données tronquées unilatérales (gauche ou droite), voir également Stute (1993) et Woodroofe (1985). Certains auteurs ont souligné que les informations disponibles sur le temps de troncature permettent de construire des estimateur plus efficaces, voir par exemple Wangn (1989). Lorsque les données sont soumises à une double troncature, la littérature sur les méthodes non paramétriques est beaucoup plus rare, l'une des raisons possibles est l'absence d'estimateurs à valeurs exactes, en effet, les méthodes existantes pour les données doublement tronquées sont itératives et intensives en terme de calculs, et ces problèmes rendent à la fois difficiles les développements théoriques et la mise en œuvre pratique.

Le premier article sur l'estimation non paramétrique du maximum de vraisemblance (NPMLE) de la fonction de répartition en cas de double troncature a été publié en 1999 par (Efron, Petrosian, 1999) et a été motivé par des données sur l'astronomie.

Mentionnant un exemple important dans lesquels la double troncature apparait Bilker et Wang (1996) ont analysé le temps écoulé entre l'infection par le VIH et le diagnostic du SIDA, disons $X^*$, est tronqué à droite par ce que seuls les individus diagnostiqués avec le SIDA avant la fin de l'étude ont été observés. Certes en mettant $V^*$le temps entre le temps de l'infection et la fin de l'étude, on dit que le temps de l'infection par le VIH a été tronqué à droite. En fin d'étude, nous avons $(X^*, V^*)$ qui est observée que lorsque $X^* \leq V^*$. Bilker et Wang (1996) ont reconnu que ces données souffrent également d'une troncature à gauche (et donc d'une double troncature); la raison est que le VIH était inconnu avant 1982 ; tous les cas de SIDA liés à une transfusion avant cette date n'ont pas été correctement classé, ce qui a entrainé la troncature à gauche d'une

observation. Par conséquent, le critère d'éligibilité est reformulé comme suit : $U^* \leq X^* \leq V^*$ ou $U^*$ représente le temps écoulé entre l'infection par le VIH et 1982.

Shen (2010a) a formellement établit la forte consistance et la normalité asymptotique de l'estimateur du maximum de vraisemblance, tandis que les méthodes de bootstrap d'échantillons finis de cet estimateur avec des données doublements tronquées ont été explorées par Moreira, de Uña-Álvarez (2010a) bien que les méthodes de bootstrap soient moins pratique sur le plan technique ils ont également été testé dans les simulations.

Cet estimateur a été étudié plus avant par Stute (1993), qui a donné une représentation presque sure du NPMLE comme une somme de variables aléatoires i.i.d. plus un reste négligeable.

Moreira, de Uña-Álvarez (2012) ont introduit une estimation de la densité pour une variable qui est observée dans le cadre d'une double troncature, y compris une formule de l'erreur quadratique moyenne intégrée (MISE), ils ont constaté que l'estimateur semi paramétrique peut être plus performant que l'estimateur non paramétrique en terme d'erreur quadratique moyenne. En outre, les avantages relatifs de l'utilisation de l'approche semi paramétrique sont clairement visibles même lorsque la taille de l'échantillon est aussi grande. Un progiciel R permettant de calculer le NPMLE d'une fonction doublement tronquée a été présenté dans le package DTDA par Moreira et al., (2010).

Théoriquement, le mode est lié directement à la fonction densité $f$, défini comme la valeur $\theta$ qui la maximise soit $f(\theta) = \sup_{t \in \mathbb{R}} f(t)$.l'estimation du mode repose sur l'estimation de la densité d'abord, et la localisation du mode comme étant un maximum global ou approximativement global. Parzen (1962) a été l'un des premiers à considérer le problème de l'estimation du mode d'une densité de probabilité, il a montré que sous certaines conditions, l'estimateur du mode obtenu en maximisant l'estimateur à noyau était convergeant en probabilité dans le cas indépendant identiquement distribué, il a également établi la normalité asymptotique de $\theta_n$, Nadaraya (1965) a établi des résultats de convergence presque sure dans $\mathbb{R}$ , (Eddy, 1982) a amélioré le résultat de Parzen (1962) en obtenant la loi limite $(\hat{\theta}_n - \theta)$. Romano (1988) s'est affranchi des conditions globales de régularité sur la densité. Il a montré que la convergence presque sure et la distribution limite de $\hat{\theta}_n$ s'obtiennent au moyen d'une hypothèse sur le comportement de $f$ au voisinage du mode. Vieu (1996) a obtenu une vitesse de convergence

presque complète de $(\hat{\theta}_n - \theta)$, Abraham (2004) a étudié la normalité asymptotique du mode à partir de la densité à support dans $\mathbb{R}^d$. Mokkadem et Pelletier (2005) ont étudié sa déviation modérée.

Dans le cas de données incomplètes, pour des variables aléatoires i.i.d. sous censure aléatoire à droite, Louani (1998) a étudié la normalité asymptotique de l'estimateur à noyau du mode. Ould-Saïd, Cai (2005) ont établi, dans le cas iid, la consistance forte uniforme avec des taux d'un estimateur non paramétrique de la fonction du mode conditionnelle censurée. (Dabo-Niang et al. 2014) ont établi l'estimation du mode dans un espace vectoriel semi normé. La normalité asymptotique a été obtenue par Ezzahrioui, Ould-Saïd (2005), dans les cas iid et α-mélange. Ould-Said, Tatachak (2009) ont établi la forte consistance et la normalité asymptotique pour des données tronquées à gauche.

Comme exemples d'estimateurs du mode, on peut citer l'estimateur de (Parzn, 1962). Cet estimateur à noyau introduit en 1962, est défini par :

$$\hat{\theta}_n = inf\left\{t \in \mathbb{R} : \ f_n(t) = \sup_{y\in\mathbb{R}} f_n(y)\right\}$$

ou $f_n$ est l'estimateur à noyau de Parzen-Rosenblatt de la densité $f$ défini par

$$f_n(y) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{y - Y_i}{h}\right)$$

ou $h \coloneqq h_n$ est une suite de réels positifs telle que

$$\lim\ h \to 0 \text{ et } \lim\ nh \to +\infty \text{ lorsque } n \to +\infty$$

et $K$ est une fonction continue vérifiant

$$\int_{-\infty}^{+\infty} K(y)dx = 1;\ lim_{|y|\to\infty} K(y) = 0.$$

Tandis que Khardani et al. (2010) ont établi les mêmes propriétés asymptotiques pour un estimateur à noyau du mode conditionnel sous censure aléatoire. Sous troncature aléatoire à gauche, dans le cas i.i.d., l'estimateur à noyau du mode a été étudié par Ould-Said, Tatachak (2007).

Chaib et al. (2013) ont considéré l'estimation de la fonction de mode lorsque les données sont soumises à une troncature aléatoire à gauche et à une censure à droite, sous des conditions appropriées, ils ont établi une convergence presque sûre et une normalité asymptotique pour un estimateur à noyau du mode, lorsque les données observées sont dépendantes.

En ce qui concerne l'estimateur à noyau de la fonction densité conditionnelle, (Roussas, 1968) fut le premier à établir ses propriétés asymptotiques pour des données markoviennes, ainsi que sa convergence en probabilité. Youndjé (1993) s'est intéressé à l'étude de la densité conditionnelle pour des données complètes indépendantes. Comme application (Rossi, 2004) a utilisé le mode conditionnel, dans le domaine des hautes technologies, pour décrire un procédé de dépollution biologique.

Ainsi le mode conditionnel, de par son importance dans le domine de prévision non paramétrique de beaucoup d'auteurs, citons les travaux de Samanta et Thavaneswaran (1990), qui ont donné les propriétés de convergence et de normalité asymptotique dans le cadre iid alors que les conditions de convergence dans le cas de données alpha mélangeantes avaient été établies par Collomb et al. (1986), enfin Vieu (1996) a estimé le mode conditionnel comme étant le point annulant la dérivée d'ordre un de l'estimateur de la densité conditionnelle et établi la convergence presque complète de cet estimateur sous la condition d'alpha mélange. Louani (1998) a établi la normalité asymptotique dans le cas de données fortement mélangeantes.

Considérons la suite $(X_n, Y_n)_{n\geq 1}$ , pour tout $x \in \mathbb{R}^d$, la densité conditionnelle $f(./x)$ de $Y$ sachant $X = x$ , noté $\theta(x)$, est définie par

$$f(\theta(y)/x) = \sup_{t\in\mathbb{R}} f(y/x).$$

Un estimateur du mode conditionnel $\theta(y)$ est définie comme la v.a. $\theta_n(y)$ maximisant un estimateur de la densité conditionnelle $\hat{f}_n(./x)$ de $f(./x)$, c'est-à-dire

$$\hat{f}_n(\theta_n(y)/x) = \sup_{t\in\mathbb{R}} \hat{f}_n(y/x).$$

Le mode conditionnelle de par son importance dans le domaine de prévision non paramétrique, était les motivations de beaucoup d'auteurs ; qui ont donné les propriétés de convergence et de normalité asymptotique dans le cadre iid alors que les conditions de convergence dans le cas de données $\alpha$ −mélangeantes avaient été établies par Ould-Saïd (1993), Gannoun et Saracco (2002).

(Vieu, 1996) a estimé le mode conditionnel comme étant le point annulant la dérivée d'ordre un de l'estimateur de la densité conditionnelle et établi la convergence presque complète de cet estimateur sous la condition $\alpha$ −mélange.

Il n'est pas rare que les données à traiter soient incomplètes, dans ce cas les techniques statistiques classiques ne s'adaptent pas correctement car l'inférence faite sur l'échantillon observé ne s'étend pas directement à la population mère. Afin de rendre facile la lecture de cette thèse, nous donnons dans le chapitre 2 une présentation de ce que sont les données incomplètes et présentons quelques exemples d'applications.

La modélisation statistique par le biais de données incomplètes est largement employée lors d'étude pratiques sur les durées de vie. Nous avons choisi, au travers de cette thèse, de considérer une durée de vie $Y^*$tronquée à droite par une $U^*$ et tronquée à gauche par une variable $V^*$.

L'estimation pour ce type de variables demande beaucoup de prudence, parmi les pistes explorées pour confronter cette difficulté, on a fait appel à l'estimateur du maximum de vraisemblance non paramétrique (NPMLE) de la fonction de distribution $(df)$ de $Y^*$ voir Efron, Petrosian (1999), qui est d'un emploi très répandu, son importance nous a donc conduits à lui accorder une grande place dans notre travail. Après avoir construit nos estimateurs de la densité et de la densité conditionnelle, du mode et du mode conditionnel dans le cas indépendant, nous avons cherché à déterminer les vitesses de convergence de nos estimateurs. Des techniques probabilistes très délicates (VC- Classes et inégalités exponentielles) ont été les outils principaux qui ont permis la détermination rigoureuse de nos vitesses de convergence.

Cette thèse est organisée en 3 chapitres, suivie d'une brève conclusion et décrit succinctement comme suit :

**Chapitre 1 :** Ce chapitre se concentre sur les données incomplètes, nous présentons les données censurées, les données tronquées et les données doublement tronquées, suivies par des exemples.

**Chapitre 2 :** Ce chapitre aborde l'estimation du mode simple dans le cas de données i.i.d., doublement tronquées. Nous présentons un nouvel estimateur à noyau pour la fonction de mode. Les propriétés asymptotiques (consistance et normalité asymptotique) sont étudiées. Par des

simulations nous montrerons les performances des estimateurs proposés. Ce travail a fait l'objet d'un article paru au journal of Science and Arts.

**Chapitre 3 :** Dans ce chapitre nous supposons qu'une autre source d'information est disponible à travers une covariable $X$ de densité $l(x)$, corrélée avec la variable d'intérêt $Y$ , la loi qui nous intéresse est celle de $Y$ sachant $X = x$ . Nous donnons un résultat sur la convergence uniforme et la normalité asymptotique de l'estimateur à noyau pour le modèle doublement tronqué avec des données i.i.d. des résultats théoriques et des preuves sont mentionnés.

Nous concluons ce manuscrit par une conclusion ainsi que quelques perspectives de recherches.

# Chapitre 1 : Généralités sur les données incomplètes

En médecine ou en biologie, on s'intéresse souvent à des durées comme par exemple la durée de survie de patients ayant eu un infarctus, durée de rémission d'une leucémie aigüe ou la durée de fièvre chez un patient atteint de pneumonie, et on distingue notre variable d'intérêt qui est le temps écoulé avant le décès du malade ou alors le temps écoulé avant la fin de la fièvre. D'un point de vue statistique on note $X$ la variable aléatoire d'intérêt, elle représente un temps, c'est donc une variable aléatoire continue et positive. De manière générale, une durée de vie sera donc le temps écoulé pour passer d'un état A à un état B. lorsque les durées de vies sont observées dans leur totalité.

Lorsque les données sont complètes, l'estimateur de la fonction de répartition de la variable aléatoire $X$ est l'estimateur empirique $F_n$ défini par $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I\{X_i \leq x\}$. Dans le cas contraire, les données sont incomplètes et nécessitent un traitement statistique particulier.

L'analyse des données incomplètes est complexe. Si la durée de vie d'un individu n'est observable que dans une période donnée, on parle de censure. Si l'individu doit survivre longtemps pour être observé dans ce cas il s'agit de troncature.

Notre travail se concentre sur les données incomplètes, nous présentons dans ce qui suit trois cas de données incomplètes, les données tronquées, censurées et les données doublement tronquées.

## 1.1 Données censurées

La censure à droite est l'exemple le plus fréquent d'observations incomplètes en analyse de survie, et a largement été décrit dans la littérature (Andersen, 1993) .

On désigne par $C$ le temps de censure et par $X$ la durée réellement observée.

### 1.1.1 Censure à droite

Une durée de survie $T$ est dite censurée à droite si l'individu n'a pas connu l'évènement d'intérêt à sa dernière visite. Formellement la durée de survie d'un évènement est définie par le couple $(X, \delta)$ avec,

$$X = \inf(T, C) \quad et \quad \delta = \begin{cases} 1 & si \quad T \leq C \\ 0 & si \quad T > C \end{cases}$$

La durée de vie $T$ et le temps de censure $C$ supposés indépendants. C'est à dire on observe le véritable temps de survie que s'il est inférieur à $C$, dans ce cas la donnée n'est pas censurée et $\delta = 1$.

- Si $\delta = 0$, la donnée est dite censurée à droite, au lieu d'observer $T$ on observe, une valeur $C$ avec pour seule information le fait que $T$ soit supérieure à $C$.
- Si l'on ne prend pas en compte la censure, en faisant comme si la donnée censurée est égale à notre variable d'intérêt on aura tendance à sous évaluer les durées.

**Exemple 1.1.1** Imaginons que l'on transplante 4 nanopuces dans les cœurs de 4 bébés tortues marines numérotés de 1 à 4 puis on les lâche dans l'océan. Ces nanopuces mesurent les battements de cœur de ces bébés tortues et sont connectées de façon continue àun outil de mesure dans lequel on reçoit :

— le nombre de battements de cœur du bébé tortue tant que celui-ci est vivant
— le message «Décès» au moment où le cœur du bébé tortue s'arrête de battre
— le message «Erreur» quand le signal avec la nanopuce est perdu

Soit $X_1, ..., X_4$ les durées de vie de ces bébés tortues. On a suivi cette cohorte et au bout d'un mois on a constaté ce qui suit :

— le bébé tortue n°1 est décédé au bout de 9 jours
— On a perdu tout contact avec la nanopuce du bébé tortue n°2 au bout de 5 jours
— le bébé tortue n°3 est décédé au bout de 29 jours
— le bébé tortue n°4 est toujours en vie

Ainsi, ce que l'on sait sur la durée de vie des ces tortues est comme suit :

— Pour le bébé tortue n°1 : on observe $X_1 = 9$ jours qui est sa durée de vie
— Pour le bébé tortue n°2 : on observe $C_2 = 5$ jours qui est la durée pendant laquelle ila survécu jusqu'à perte de tout signal avec la nanopuce. Même si l'on ne la connait pas avec exactitude, on sait que sa durée de vie $X_2$ est forcément plus grande que $C_2$.
— Pour le bébé tortue n°3 : on observe $X_3 = 29$
— Pour le bébé tortue n°4 : on reçoit toujours des signaux de la part de la nanopuce donc on observe $C_4 = 30$ jours et on sait que forcément $X_4 > C_4$

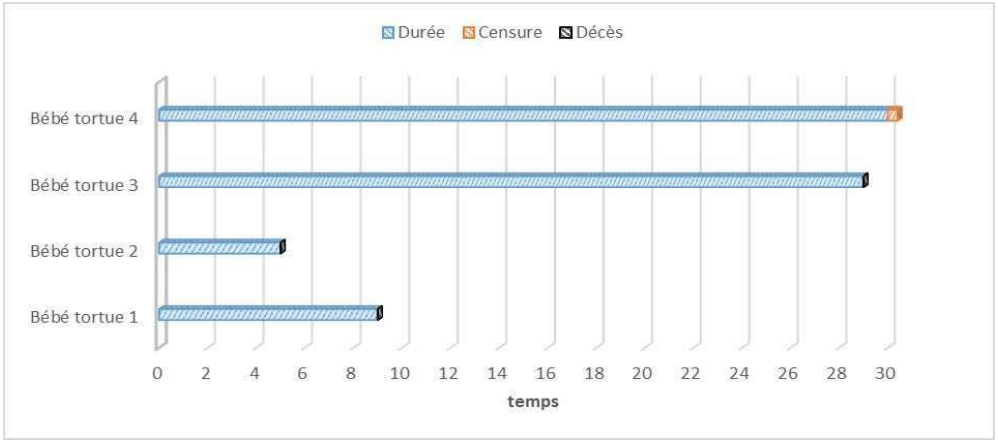


**Fig. 1.1 Illustration de la censure à droite**

### 1.1.2 Censure à gauche

Une durée de survie est dite censurée à gauche si l'individu a déjà connu l'évènement d'intérêt avant l'entrée dans l'étude. La durée de survie est définie par le couple $(X, \delta)$,

où $X = max(T, C)$; $\quad \delta = \begin{cases} 1 & si \quad T > C \\ 0 & si \quad T \leq C \end{cases}$

au lieu d'observer $T$ on observe une valeur $C$.

Dans la censure à gauche on ne connait pas toujours la date d'entrée dans l'étude.

**Exemple 1.1.2** Les babouins de la réserve d'Amboli, au Kenya, dorment dans les arbres et descendent de leurs arbres pour aller se nourrir. L'évènement d'intérêt est l'instant où ils descendent de l'arbre. Des biologistes viennent faire des visites régulières pour voir si les babouins sont descendus de leurs arbres.

L'évènement d'intérêt est observé si le babouin descend de l'arbre après l'arrivée des biologistes. Par contre, la donnée est censurée si le babouin est descendu avant l'arrivée des observateurs. On sait uniquement que l'horaire de descente est inférieur à l'heure d'arrivée des observateurs. Donc on observe le maximum entre l'heure de descente des babouins et l'heure d'arrivée des observateurs.

## 1.2 Données tronquées

Une observation est dite tronquée si elle est conditionnelle à un autre évènement. On dit que la variable $T$ de durée de vie est tronquée si $T$ n'est observable que sous une certaine condition dépendante de la valeur de $T$. Plus précisément, la durée de vie est tronquée à droite (respectivement à gauche) si $T < U$ $(respectivement\ T > U)$ . la variable $U$, appelée variable de troncature droite (ou gauche), est supposée indépendante de la variable $T$.

**Exemple 1.2.1** Le nombre de personnes infectés par le sida est inconnu et l'information est disponible seulement pour ceux qui ont été infectées et développées le SIDA dans un certain laps de temps. Ainsi, les personnes qui n'ont pas encore développées la maladie sont inconnues par l'enquêteur et ne sont pas inclus dans l'échantillon. C'est le cas de troncature à droite.

Voici un exemple simple pour comprendre l'analyse non paramétrique pour des données tronquées à droite. Imaginez que vous voulez étudier la durée de vie d'une ampoule. Vous avez 10 ampoules et vous les allumez toutes en même temps. Vous notez le temps auquel chaque ampoule s'éteint. Cependant, après 1000 heures, vous devez arrêter l'expérience et il reste encore 2 ampoules allumées. Dans ce cas, les données sont tronquées à droite car vous ne connaissez pas la durée de vie exacte des 2 ampoules restantes.

**Exemple 1.2.2** Reprenons l'exemple de la section précédente. Cette fois-ci on transplante les mêmes nanopuces dans les coeurs d'une cohorte plus grande de bébés tortues (40 tortues numérotés de 1 à 40). On suit la cohorte pendant un mois et on décidede mesurer la durée de vie de tortues ayant enregistré des battements de cœur anormaux. Après ce délai, on a constaté ce qui suit :

- On a constaté des battements de cœur anormaux chez les bébés tortues n° 1, 3, 10 et 30.
- On a constaté des battements anormaux chez le bébé tortue n°1 dès le début de l'étude.
  Celui ci est décédé 3 jours après.
- On a constaté des battements anormaux chez le bébé tortue n° 3 quatre jours après ledébut de l'étude. Celui ci est décédé 6 jours après.
- On a constaté des battements anormaux chez la tortue n°10 quinze jours après le débutde l'étude. Le signal avec cette tortue a été perdu 20 jours après.
- On a constaté des battements anormaux chez la tortue n°30 vingt jours après le débutde l'étude. Cette tortue était toujours en vie à la fin de l'étude.

Les bébés tortues que l'on observe ne rentrent dans l'étude qu'après un certain temps. Il y adonc troncature à gauche. Ainsi, après un mois d'étude, on a observé ce qui suit :

- $T_1 = 0$ jour. $X_1 = 3$ jours pour la tortue n°1.
- $T_3 = 4$ jours. $X_3 = 6$ jours pour la tortue n°2.
- $T_{10} = 15$ jours. Cette durée est censurée et l'on observe $X_{10} > C_{10} = 20$ jours.
- $T_{30} = 20$ jours. Cette durée est censurée et l'on observe $X_{30} > C_{30} = 30$ jours.

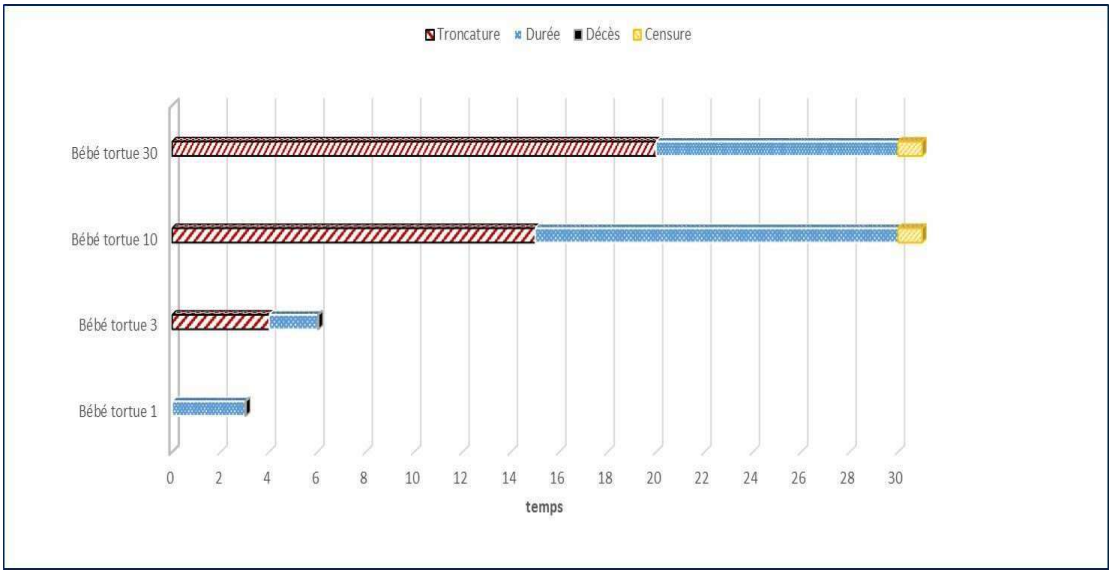


**Fig. 1.2 Illustration de la troncature à gauche**

**Remarque 1.2.1** Si on ne prend pas en compte la troncature à gauche, on va avoir tendance à sous évaluer la mortalité des diabétiques puisqu'on n'a pas les patients qui sont décédés avant d'avoir contacté l'hôpital.

## 1.3 Troncature vs censure

À ce stade, le lecteur peut être curieux de connaître la différence entre la troncature et la censure. La censure à droite est un phénomène très connu dans les domaines suivants l'analyse de survie et les études de fiabilité, entre autres domaines. Elle se produit lorsque le suivi d'un individu donné s'arrête avant que l'événement d'intérêt n'ait eu lieu.

Dans ce cas, l'observateur sait seulement que la variable cible est plus grande que la valeur enregistrée, ce que l'on appelle le temps de censure.

Un échantillon composé de valeurs réelles et censurées est généralement analysé par l'estimateur de Kaplan-Meier (Kaplan El., 1958), qui corrige le fait que certaines des valeurs enregistrées pour $X$ sont plus petites que la vraie valeur. Avec des données tronquées, chaque valeur de l'échantillon correspond à une véritable observation de $X$ ; Cependant, la distribution des valeurs observées peut être décalée par rapport à la vraie valeur en raison de l'événement de troncature.

Cette différence entre la troncature et la censure suggère des méthodes spécifiques devraient être employées pour estimer la distribution cible.

En effet, (Woodroofe, 1985) fournit une analyse approfondie de la troncature unilatérale en introduisant l'idée originale de (Lynden-Bell, 1971) en tant que un estimateur non paramétrique du maximum de vraisemblance (NPMLE) de la distribution de probabilité dans ce cadre. L'estimateur de (Woodroofe, 1985) est un cas particulier de l'estimateur correspondant à des données doublement tronquées.

## 1.4 Les données doublement tronquées

La double troncature est un type de troncature où troncature à gauche et troncature à droite se produisent simultanément. Par exemple, en astronomie, les objets stellaires dans les galaxies ne sont pas détectés s'ils sont trop brillants ou trop sombres. Efron, Petrosian (1999) fournissent le cas où les quasars ne sont observés que lorsque leur luminosité est comprise entre une limite de détection inférieure et une limite de détection supérieure. Si la luminosité est trop basse ou trop élevée, les instruments astronomiques ne la détectent pas.

**Quasars :** est un objet céleste massif et extrêmement éloigné, émettant des quantités d'énergie exceptionnellement grandes, et ayant généralement une image en forme d'étoile dans un télescope.

Cet exemple de quasars a motivé le développement ultérieur de l'analyse de survie avec double troncature, bien que la luminosité ne soit pas une durée de vie au sens littéral. Dans l'analyse des données de survie.

## 1.5 Exemples sur la double troncature

Dans ce qui suit, nous fournissons des exemples de données sujets à la double troncature.

**Exemple 1.4.1** Le terme cancer chez l'enfant fait référence aux cancers qui surviennent entre la naissance et l'âge de 15 ans. Moreira, de Uña-Álvarez (2010b) ont fourni un ensemble de données sur 406 enfants du nord du Portugal qui ont reçu un diagnostic de cancer au cours d'une période de recrutement de 5 ans (entre le 1er janvier 1999 et 31 décembre 2003) la population d'intérêt est un groupe d'enfants du nord du Portugal. Les enfants diagnostiqués avant le 1er janvier 1999 ou après le 31 décembre 2003 n'existent pas dans l'ensemble des données, et par conséquent, les données sont doublement tronquées.

Soit $y^*$ l'âge au moment du diagnostic de cancer pour un enfant sélectionné au hasard dans une population.

Le critère d'inclusion de l'échantillon s'écrit :

$U^* \leq Y^* \leq V^*$, où $U^*$ est l'âge au 1er janvier 1999 et $V^* = U^* + 5$ (ans) est l'âge au 31 décembre 2003 ainsi, la limite de troncature à gauche est $U^*$ et la limite de troncature à droite est $V^*$ :

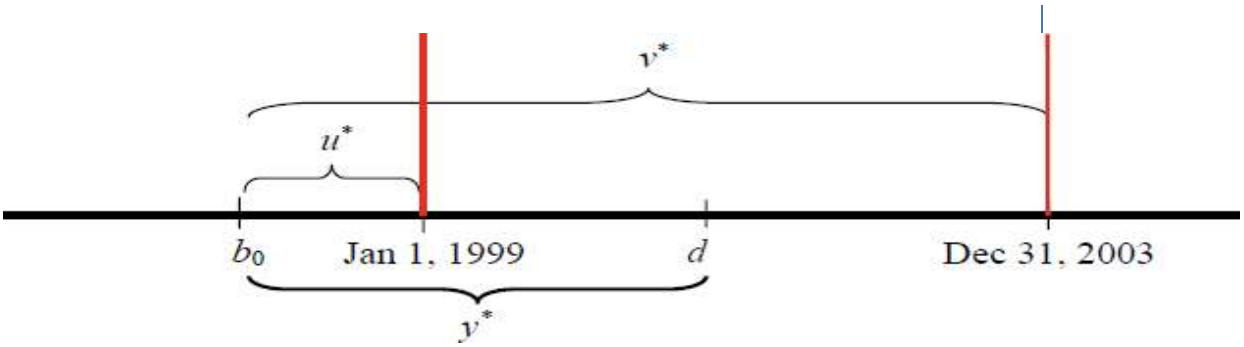


**Fig. 1.3 Données du cancer chez les enfants entre la naissance et l'âge de 15 ans. (1/1/1999 – 31/12/2003) du nord du Portugal. Source : Moreira, de Uña-Álvarez (2010b)**

$b_0$ : Date de naissance de l'enfant

*d :* date du diagnostic (date d'apparition du cancer)

$y^*$ :l'âge au moment du diagnostic de cancer

$u^*$ : l'âge au 1er janvier 1999 ($u^*$ est négative pour les personnes nées après le 1 Jan 1999)

$v^* = u^* + 5$ (ans) est l'âge au 31 décembre 2003

**Exemple 1.4.2** Un exemple pratique pour comprendre comment estimer le mode pour des données doublement tronquées.

Supposons que vous voulez étudier la durée de vie d'une batterie de téléphone portable, vous avez 10 téléphones et vous les utilisez tous en même temps jusqu'à ce que la batterie soit complètement déchargée. Vous notez le temps auquel chaque batterie se décharge entre 5 heures et 10 heures. Les batteries qui se déchargent en moins de 5 heures ou en plus de 10 heures ne sont pas prise en compte.

Dans ce cas, les données sont soumises à une troncature aléatoire. Pour estimer la valeur du temps le plus probable auquel chaque batterie se décharge en fonction des bornes connues, c'est-à-dire entre 5 heures et 10 heures, on va estimer la fonction mode sous un modèle de double troncature, qui nous permettra d'obtenir une estimation plus précise du temps moyen du déchargement.

## 1.6 Estimation de la densité sous double troncature

La littérature s'est principalement concentrée sur le problème de la troncature à gauche ou, plus largement, sur les configurations de troncature unilatérale. Les caractéristiques de l'estimateur non paramétrique du maximum de vraisemblance (NPMLE) de la fonction de distribution (df) avec des données tronquées à gauche ont été examinées par (Woodroofe, 1985) voir également (Lynden-Bell, 1971). (Stute, 1993) a poursuivi ses recherches sur cet estimateur, pour des données tronquées à droite. La littérature sur la double troncature aléatoire est cependant beaucoup plus limitée. L'absence d'estimateurs en forme fermée c'est-à-dire une expression qui peut être calculée en appliquant un nombre fixe d'opérations familières aux arguments, est une cause potentielle ; en fait, la nature itérative et exigeante en termes de calcul des approches actuelles pour les données doublement tronquées rend les développements théoriques et les mises en œuvre pratiques plus difficiles.

Soit $Y^*$ la variable aléatoire d'intérêt avec la fonction de distribution $F$, et supposons qu'elle est doublement tronquée par la paire aléatoire $(U^*,V^*)$.avec la fonction de distribution conjointe $H$, où $U^*$et $V^*(U^* \leq V^*)$ sont les variables de troncature gauche et droite respectivement. Cela signifie que le triplet $(U^*,Y^*,V^*)$ est observé si et seulement si $U^* \leq Y^* \leq V^*$. Aucune information n'est disponible lorsque $Y^* < U^*$ ou $Y^* > V^*$ . Nous supposons que $Y^*$ est indépendant de $(U^*,V^*)$. Soit $(U^*, Y^*, V^*)$, i = 1, ... $n$, désignent les informations d'échantillonnage, il s'agit de données i.i.d. avec la même distribution de $(U^*, Y^*, V^*)$ étant donné que $U^* \leq Y^* \leq V^*$. Introduisons $\alpha = P(U^* \leq Y^* \leq V^*)$, la probabilité de non-troncature. Il est clair que si $\alpha = 0$, aucune donnée ne peut être observée et nous supposons donc tout au long de ce document que $\alpha > 0$.

Pour toute distribution W, notons respectivement les extrémités gauche et droite de son support par

$$a_w = inf\{t: W(t) > 0\} \quad \text{et} \quad b_w = inf\{t: W(t) = 1\}$$

soit $H_1(u) = H(u, \infty)$ et $H_2(v) = H(-\infty, v)$ the marginal distribution functions of $U^*$ and $V^*$, respectivement. Lorsque

$$a_{H_1} \leq a_F \leq a_{H_2} \text{ et } b_{H_1} \leq b_F \leq b_{H_2},$$

$F$ et $H$ sont toutes deux identifiable (Woodroofe, 1985).

Notez par $f(.)$ la fonction de densité de probabilité de $Y^*$. pour définir l'estimateur à noyau non paramétrique de la densité $f(.)$, nous devons d'abord introduire l'estimateur du maximum de vraisemblance non paramétrique (NPMLE) de la fonction de distribution $(df)$ de $Y^*$ (voir Moreira et de Uña-Álvarez, 2010a). Dans le cadre du schéma d'échantillonnage doublement tronqué, l'estimateur NPMLE de la fonction de répartition de $Y^*$est donné par

$$F_n(y) = \alpha_n \int_{-\infty}^{y} \frac{F_n^*(dt)}{G_n(t)}$$

où

$$\alpha_n = \left( \int_{a_F}^{\infty} G_n^{-1}(t) F_n^*(dt) \right)^{-1}$$

est un estimateur de $\alpha$, (voir Shen, 2010b)

$F_n^*(y) = n^{-1} \sum_{i=1}^{n} I_{[Y_i \leq y]}$ est la function de distribution ordinaire de $Y_i$, et

$$G_n(t) = \int_{\{u \leq t \leq v\}} H_n(du, dv)$$

est l'estimateur non paramétrique de $G(t) = P(U^* \leq t \leq V^*)$ qui est la probabilité d'échantillonner une durée de vie $Y^* = t$.

Ici $H_n(u, v)$ est le NPLME de la distribution conjointe H des temps de troncature, voir Moreira, de Uña-Álvarez (2012) pour plus de détails).

$$f(y) = \int K_h(y - t) F_n dt = \frac{\alpha_n}{nh} \sum_{i=1}^{n} \frac{1}{G_n(Y_i)} K\left(\frac{y - Y_i}{h}\right)$$

$$F_n(y) = \frac{\alpha_n}{n} \sum_{i=1}^{n} \frac{1}{G_n(Y_i)} I_{\{Y_i \leq y\}}$$

$K$ est la function noyau (dite la fonction Kernel en anglais) et $h_n$ est la fenêtre qui devient nulle lorsque $n$ tends vers l'infini.

Sous certaines conditions, Moreira et de Uña-Álvarez (2012) ont montrer que :

$$E\left(\hat{f}_n(y)\right) = f(y) + \frac{1}{2}h^2 f''(y)\mu_2(K) + o(h^2)$$

$$Var\left(\hat{f}_n(y)\right) = (nh)^{-1}\alpha G(y)^{-1} f(y) R(K) + o((nh)^{-1})$$

En effet, ils ont démontré que le biais n'est pas affecté par la double troncature, mais que la variance de $\hat{f}_n(y)$ est modifiée. Plus précisément, la variance de l'estimateur est importante aux points $y$ pour lesquels la probabilité relative d'obtenir les valeurs $Y_i$ autour de $y$ (c'est-à-dire $G(y)$) est faible.

En général on s'intéresse à l'erreur globale de l'estimateur, cela peut être mesuré par le MISE (l'erreur quadratique moyen intégré) à savoir :

$$MISE(f_n(y)) = \int MSE(f_n(y))$$

avec

$$MSE\big(f_n(y)\big) = E\big(f_n(y) - f(y)\big)^2$$

$$= Var\big(f_n(y)\big) + Biais^2\big(f_n(y)\big).$$

Sous certaines conditions de régularités, nous avons à partir des résultats précédents l'expression asymptotique du $MISE$ :

$$MISE\big(f_n(y)\big) = \tfrac{1}{4}h^4 R(f'')\mu_2(K)^2 + (nh)^{-1}\alpha R(K)\int G^{-1}f,$$

avec

$$R(f'') = \int f''(t)^2 dt$$

La minimisation de $MISE\big(f_n(y)\big)$ par rapport à la fenêtre $h$ conduit à la fenêtre asymptotique optimale :

$$h_{opt} = \left(\frac{\alpha R(K)\int G^{-1}f}{R(f'')\mu_2^2(K)}\right)^{1/5} n^{-1/5}.$$

**Remarque 1.5.1** Cette dernière dépend des quantités inconnues qui doivent être estimées en pratique. On notera de plus, que c'est la fonction $G$ et le paramètre de lissage $h$ qui influencent la forme de la distribution de troncature. Plutôt que la probabilité de troncature elle-même.

# Chapitre 2 :
# Estimation non paramétrique du mode pour des données doublement tronquées

## 2.1. Introduction

Les données tronquées aléatoirement apparaissent dans divers domaines, notamment l'astronomie, la médecine, l'épidémiologie et l'économie. Un exemple typique de troncature aléatoire à droite est l'analyse des données sur le SIDA, lorsque l'information est restreinte aux personnes ayant développé le SIDA avant une certaine date. Dans certaines applications, la troncature aléatoire bilatérale (plutôt qu'unilatérale) apparaît. Dans Bilker et Wang (1996) il est indiqué que les temps d'induction dans le SIDA sont en fait doublement tronqués puisque le VIH était inconnu avant 1982 et que les patients infectés auraient donc été incorrectement écartés lorsqu'ils ont développé le SIDA avant cette date. En outre, l'étude de Efron et Petrosian (1999) a examiné les luminosités des quasars qui ont été doublement tronquées par certaines limites de détection.

De nombreux auteurs ont introduit des méthodes non paramétriques pour les données tronquées unilatéralement (à gauche ou à droite), voir par exemple Lynden-Bell (1971). Cependant, la littérature sur la double troncature aléatoire est beaucoup plus rare. L'une des raisons possibles est l'absence d'estimateurs de forme fermée. En effet, les méthodes existantes pour les données doublement tronquées sont itératives et intensives en termes de calcul, ce qui complique à la fois les développements théoriques et les mises en œuvre pratiques.

Les auteurs Efron et Petrosian (1999) sont les premiers à avoir introduit la NPMLE de la fonction de distribution en cas de double troncature. La littérature contient également des approches semi paramétriques pour estimer la fonction de distribution sous double troncature. Le problème lorsque la distribution des temps de troncature est supposée appartenir à une famille paramétrique donnée est étudié par Moreira, de Uña-Álvarez (2010b) et (Shen, 2010a).

L'étude de Moreira, de Uña-Álvarez (2012) présente deux estimateurs différents de la densité du noyau, qui sont définis comme une convolution entre une fonction noyau et un estimateur de la fonction de distribution cumulative. Plusieurs procédures de sélection de la largeur de bande pour l'estimation de la densité de noyau de données tronquées deux fois de manière aléatoire sont présentées et comparées; les auteurs Moreira et de Uña-Álvarez (2010a) présentent cinq procédures de sélection de la largeur de la fenêtre et donnent une justification théorique.

À notre connaissance, le problème de l'estimation de la fonction de mode dans le cas de données doublement tronquées n'a pas été abordé dans la littérature statistique. Il s'agit d'un sujet d'intérêt central dans le présent document. Dans ce travail, nous proposons un nouvel estimateur de la fonction du mode et établissons sa convergence uniforme presque sûre et sa normalité asymptotique. Pour ce faire, nous considérons les classes de Vapnik-Cervonenkis (V-C) pour lesquelles des inégalités exponentielles uniformes sont disponibles. En outre, l'estimation fonctionnelle est basée sur la méthode du noyau. Comme application de la normalité asymptotique de notre nouvel estimateur, nous introduisons un intervalle de confiance asymptotique pour le mode. Nos résultats théoriques coïncident avec ceux obtenus dans le cas des données complètes.

Ce chapitre est organisé comme suit : dans la section 2, nous définissons certains résultats importants et utiles dans le modèle de double troncation aléatoire, puis nous définissons l'estimateur du mode du noyau dans le cadre de la double troncation aléatoire. L'hypothèse et les principaux résultats sont donnés dans la section 3 avec la normalité asymptotique de l'estimateur proposé. La section 4 présente une simulation de notre estimateur. Les preuves des principaux résultats sont proposées à la section 5.

## 2.2. Notation et définition de l'estimateur

Nous présentons d'abord certains résultats de la littérature pour les données doublement tronquées, qui seront utilisés pour définir notre estimateur du mode. Soit $Y^*$ la variable aléatoire d'intérêt avec la fonction de distribution $F$, et supposons qu'elle est doublement tronquée par la paire aléatoire $(U^*, V^*)$ avec la fonction de distribution conjointe $H$, où $U^*$et $V^*(U^* \leq V^*)$ sont les variables de troncature gauche et droite respectivement. Cela signifie que le triplet

$(U^*, Y^*, V^*)$ est observé si et seulement si $U^* \leq Y^* \leq V^*$. Aucune information n'est disponible lorsque $Y^* < U^*$ ou $Y^* > V^*$ Nous supposons que $Y^*$ est indépendant de $(U^*, V^*)$.

Soit $(U_i, Y_i, V_i)$ , $i = 1, \ldots . n$, désignent les informations d'échantillonnage, il s'agit de données i.i.d. avec la même distribution de $(U^*, Y^*, V^*)$ étant donné que $U^* \leq Y^* \leq V^*$.

Introduisons $\alpha = P\ (U^* \leq Y^* \leq V^*)$, la probabilité de non-troncature. Il est clair que si $\alpha = 0$, aucune donnée ne peut être observée et nous supposons donc tout au long de ce document que

$\alpha > 0$. Pour toute distribution W, notons respectivement les extrémités gauche et droite de son support par

$$a_w = inf\{t: W(t) > 0\} \quad \text{et} \quad b_w = inf\{t: W(t) = 1\}$$

soit $H_1(u) = H(u, \infty)$ et $H_2(v) = H(-\infty, v)$ les distributions marginales de $U^*$et $V^*$ respectivement. Lorsque $a_{H_1} \leq a_F \leq a_{H_2}$ et $b_{H_1} \leq b_F \leq b_{H_2}$, $F\ et\ H$ sont toutes deux identifiable (voir Woodroofe, 1985).

Notons par $f(.)$ la fonction de densité de probabilité de $Y^*$ et supposons qu'elle a un mode unique défini par

$$\theta = \underset{y \in \mathbb{R}}{\operatorname{argmax}}\, f(y)$$

Comme cela été proposé dans Moreira, de Uña-Álvarez (2012), pour définir l'estimateur non paramétrique à noyau de la densité $f(.)$, nous devons d'abord introduire l'estimateur du maximum de vraisemblance non paramétrique (NPMLE) de la fonction de distribution $(df)$ de $Y^*$ (voir Efron et Petrosian, 1999).

Dans le cadre du schéma d'échantillonnage doublement tronqué, l'estimateur NPMLE de la fonction de répartition de $Y^*$ est donné par

$$F_n(y) = \alpha_n \int_{-\infty}^{y} \frac{F_n^*(dt)}{G_n(t)}$$

où

$$\alpha_n = \left( \int_{a_F}^{\infty} G_n^{-1}(t) F_n^*(dt) \right)^{-1}$$

Est un estimateur de $\alpha$,voir (Shen, 2010b). $F_n^*(y) = n^{-1} \sum_{i=1}^{n} I_{[Y_i \leq y]}$ est la fonction de distribution empirique ordinaire de $Y_i$, et

$$G_n(t) = \int_{\{u \leq t \leq v\}} H_n(du, dv)$$

est l'estimateur non paramétrique de $G(t) = P(U^* \leq t \leq V^*)$ qui est la probabilité d'échantillonner une durée de vie $Y^* = t$. Ici $H_n(u, v)$ est le NPLME de la distribution conjointe H des temps de troncature, voir (Moreira, de Uña-Álvarez, 2012) pour plus de détails.

Notre estimateur non paramétrique du mode θ est définie comme la valeur aléatoire $\hat{\theta}_n$ maximisant l'estimateur de la densité $\hat{f}_n$ ,à savoir

$$f_n(\hat{\theta}_n) = \sup_{a_F \leq y \leq b_F} \hat{f}_n(y) \qquad (\mathbf{2.1})$$

où

$$\hat{f}_n(y) = \int K_h(y - t) F_n dt = \frac{\alpha_n}{nh} \sum_{i=1}^{n} \frac{1}{G_n(Y_i)} K\left(\frac{y - Y_i}{h}\right) \quad , \qquad (\mathbf{2.2})$$

$$F_n(y) = \frac{\alpha_n}{n} \sum_{i=1}^{n} \frac{1}{G_n(Y_i)} I_{\{Y_i \leq y\}}$$

$K$ est la fonction densité de probabilité (appelé la fonction Kernel) et $h_n$ est une suite de nombres réels positifs (appelé fenêtre) qui devient nulle lorsque $n$ tends vers l'infini.

**Remarque 2.2.1.** Rappelons que l'estimateur $\hat{\theta}_n$ n'est pas nécessairement unique et nos résultats sont valables pour n'importe quelle valeur satisfaisant **(2.1)**. Nous soulignons que notre choix peut être spécifié en prenant

$$\hat{\theta}_n = inf\left\{a_F \leq t \leq b_F \ : \ \hat{f}_n(t) = \sup_{a_F \leq y \leq b_F} \hat{f}_n(y)\right\}.$$

## 2.3. Hypothèses et principaux résultats

Tout au long de ce chapitre, nous supposons que $a_{H_1} \leq a_F \leq a_{H_2}$ et $b_{H_1} \leq b_F \leq b_{H_2}$ et soit $\Omega$ un ensemble compact tel que $\Omega \subset \Omega_0 = \{y : a_F \leq y \leq b_F\}$. Considérons maintenant les hypothèses de régularité suivantes :

**(A1)** la fonction kernel $K$ est une densité qui vérifie

$$\int tK(t)dt = 0, \int t^2K(t)dt < \infty \quad \text{and} \quad R(K) = \int K^2(t)dt < \infty$$

**(A2)** la fenêtre $h_n$ satisfait:

$$h \longrightarrow 0 \quad \text{et} \quad nh \longrightarrow \infty \ quand \ n \longrightarrow \infty \qquad et \qquad \frac{nh}{Logn} \longrightarrow \infty \ quand \ n \longrightarrow \ \infty.$$

**(A3)** La fonction kernel $K$ est à support compact, $C^1$densité de probabilité, deux fois différentiable tel que $K, K^{(1)} et \ K^{(2)}$ sont intégrable de plus $K$, $K^{(2)}$ sont lipchitziennes.

**(A4)** La fonction $f(y)$ et $G^{-1}f$ sont différentiable jusqu'à l'ordre 3.

**(A5)** Le mode $\theta$ satisfait la propriété suivante: pour tout $\varepsilon > 0 \ et \ t > 0, il \ exist$

$$\beta > 0 \ tel \ que \ |\theta - t| \geq \varepsilon \ implique \ |f(\theta) - f(t)| \geq \beta$$

**(A6)** $\boldsymbol{h} \coloneqq h_n$ satisfait pour $n \longrightarrow \infty, h \longrightarrow 0; \ nh^3 \longrightarrow 0 \ et \ nh^7 \longrightarrow 0$ .

**(A7)** l'ensemble $\mathcal{F} = \left\{K\left(\frac{x-.}{h}\right) : x \in \mathbb{R}, h \in \mathbb{R} - \{0\}\right\}$ est une classe V-C bornée de fonctions mesurables.

On suppose que $K$ est deux fois différentiables, en dérivant $\hat{f}_n(.)$ on a :

$$\hat{f}_n^{(j)}(y) = \frac{\alpha_n}{nh^{j+1}} \sum_{i=1}^{n} \frac{1}{G_n(Y_i)} K\left(\frac{y - Y_i}{h}\right) \quad j = 1,2$$

De même pour

$$\tilde{f}_n^{(j)}(y) = \frac{\alpha}{nh^{j+1}} \sum_{i=1}^{n} \frac{1}{G(Y_i)} K\left(\frac{y - Y_i}{h}\right)$$

***Discussion des hypothèses :***

- Les hypothèses **(A1) et (A3)** sont classiques dans l'estimation non paramétrique.
- **(A2)** est nécessaire pour la démonstration de la convergence de $\hat{f}_n(.)$
- **(A4)** assure l'existence des dérivées d'ordre supérieur
- **(A5)** implique l'unicité du mode
- **(A7)** a été considérée dans la condition $K_1$ de Giné et Guillou (2001), cette hypothèse est nécessaire pour utiliser l'inégalité de Talagrand (Talagrand, 1996), qui à son tour garantit la convergence uniforme des estimateurs.
- Finalement, il suffit de noter que **(A6)** implique **(A2)**.

## 2.3.1. La consistance

Dans cette section, nous établissons la consistance et la normalité asymptotique de notre estimateur

**Proposition 2.3.1** Sous réserve d'hypothèses **(A1)−(A4)** nous avons

$$\sup_{y \in \Omega} \left|\hat{f}_n(y) - f(y)\right| = O\left( max\left( \left(\frac{logn}{nh}\right)^{\frac{1}{2}}, h^2 \right) \right) \quad p.s \ \ qd \ \ n \longrightarrow \infty$$

**Théorème 2.3.1** Sous réserve d'hypothèses **(A1)−(A5)**, nous avons

$$\hat{\theta}_n - \theta = O\left( max\left( \left(\frac{logn}{nh}\right)^{\frac{1}{4}}, h \right) \right) \quad p.s \ \ qd \ \ n \longrightarrow \infty$$

**Remarque 2.3.2.** si on choisit $h = O\left(\left(\frac{logn}{nh}\right)^{\frac{1}{5}}\right)$, qui est le choix optimal en ce qui concerne le critère de convergence uniforme presque sûr dans l'estimation de la densité, nous obtenons

$$\hat{\theta}_n - \theta = O\left(\left(\frac{logn}{nh}\right)^{\frac{1}{5}}\right) p.s \ \ qd \ \ \ n \longrightarrow \infty$$

C'est le taux optimal obtenu dans le cas des données complètes (Vieu, 1996).

## 2.3.2. La normalité asymptotique

Supposons maintenant que la densité$f(.)$ est unimodal en $\theta$.sous l'hypothèse **(A4)** on a :

$$f^{(1)}(\theta) = 0 \ \ et \ \ f^{(2)}(\theta) < 0$$

De même, nous avons :

$$\hat{f}_n^{(1)}(\hat{\theta}_n) = 0 \ \ et \ \ \hat{f}_n^{(2)}(\hat{\theta}_n) < 0$$

En utilisant le développement de Taylor, nous obtenons

$$\hat{f}_n^{(1)}(\hat{\theta}_n) = \hat{f}_n^{(1)}(\theta) + (\hat{\theta}_n - \theta)\hat{f}_n^{(2)}(\bar{\theta}_n) = 0$$

ou $\bar{\theta}_n$ est entre $\hat{\theta}_n$ et $\theta$, ce qui donne

$$\hat{\theta}_n - \theta = -\frac{\hat{f}_n^{(1)}(\theta)}{\hat{f}_n^{(2)}(\bar{\theta}_n)} \qquad (2.3)$$

Pour établir la normalité asymptotique, nous montrons que le numérateur dans **(2.3)**, est asymptotiquement normalement distribuée et le dénominateur converge en probabilité vers $f^{(2)}(\theta)$.

Le résultat est donné dans le théorème suivant.

**Théorème 2.3.2** Nous supposons que les hypothèses **(A1)−(A6)** sont vérifiées, nous avons alors:

$$\left(\frac{nh^3\left(\hat{f}_n^{(2)}(\theta)\right)^2}{\sigma^2}\right)^{1/2}\left(\hat{\theta}_n-\theta\right)\xrightarrow{L}N(0,1)$$

ou $\xrightarrow{L}$ signifie la convergence en Loi, $N(0,1)$ est la distribution normale standard et

$$\sigma^2=\frac{\alpha f(\theta)}{G(\theta)}\int(K'(r))^2dr$$

**Remarque 2.3.3.** Dans le cas de données complètes (i.e., $\alpha = G(.) = 1$), on a :

$$\sigma^2=f(\theta)\int(K'(r))^2dr$$

C'est ce qui a été obtenu par (Parzen, 1962).

**Corollaire 2.3.1.** L'utilisation d'une méthode d'insertion en remplaçant α et $f$ par leurs estimations, nous permet d'obtenir une estimation convergente $\sigma_n^2$ de $\sigma^2$. Du théorème 2.3.2, que nous obtenons pour chaque $\nu \in (0,1)$, les $(1-\nu)\%$ intervalles de confiance asymptotiques suivants pour $\theta$, à savoir,

$$\theta=\hat{\theta}_n\pm\sigma_n\left(nh^3\left(\hat{f}_n^{(2)}(\theta)\right)^2\right)^{-1/2}z_{1-\nu/2}$$

où $z_{1-\nu/2}$ est le quantile d'ordre $1-\nu/2$ la distribution normale standard $N(0,1)$.

## 2.4. Étude de simulation

Dans cette section, nous illustrons le comportement sur un échantillon fini de notre estimateur $\hat{\theta}_n$ défini dans (2.1) et examiner la normalité asymptotique. On suppose ici que pour la double troncature, on considère le cas $U^*$et $V^*$ sont mutuellement indépendants. Les résultats de cette section ont été obtenus avec R – Software package DATA (Moreira et al., 2010).

Tout d'abord, nous présentons deux modèles simulés qui ont permis de calculer l'estimateur $\hat{\theta}_n$.

**Model 1 (cas de décroissance exponentielle):** La variable Y est distribuée selon une loi normale $N(\mu, \sigma^2)$, qui admet un mode θ égale à la moyenne μ, de densité :

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right\}, y \in \mathbb{R}$$

**Model 2 (Cas de queues lourdes):** la variable $Y$ est distribué sous la forme d'un modèle à trois paramètres Weibull(λ, β, γ) de densité

$$f(y) = \left(\frac{\beta}{\lambda}\right)\left(\frac{x-\gamma}{\lambda}\right)^{\beta-1} e^{-\left(\frac{y-\gamma}{\lambda}\right)^{\beta}} 1_{(y\geq 0)},$$

avec $\gamma \in \mathbb{R}$ est le paramètre de localisation (ou durée de vie sans défaillance), β > 0 est le paramètre de forme et λ > 0 est le paramètre d'échelle ou la durée de vie caractéristique de la distribution et qui admet un mode θ au point $\gamma + \lambda\left(1 - \frac{1}{\beta}\right)^{1/\beta}$.

On prend $U^* \sim Exp(b)$ et $V^* \sim Exp(c)$, ou $b$ et $c$ sont choisis de manière à obtenir les pourcentages de troncature suivants : $70\%, 50\%, 30\%\ et\ 10\%$ correspondant à $(\alpha = 0.3, 0.5, 0.7\ and\ 0.9)$ respectivement. Notons que ce choix de pourcentage de troncature est standard dans ce type d'étude. Il est clair que les valeurs inférieures de $\alpha$ n'ont que peu d'intérêt pratique, car pour $\alpha \approx 0$ presque aucune donnée ne peut être observée.

En utilisant ce schéma, $B = 500$ échantillons indépendants de taille $n$ ont été générés pour chaque modèle. Des tailles d'échantillon de $n = 50, n = 150\ and\ n = 300$ ont été considérées. La troncature se produit lorsque $U^* \leq Y^* \leq V^*$ n'est pas respecté. Cela signifie que, pour chaque essai, le nombre de données simulées est beaucoup plus grand que, en fait $\alpha \approx \frac{n}{N}$ sont nécessaires en moyenne, où rappelons que α représente la proportion d'absence de troncature.

Les cas de petits échantillons (avec n souvent compris entre 20 et 50) sont basés sur l'utilisation d'estimateurs robustes, ce qui n'est pas le cas de cette étude, et pourrait être pris en compte dans nos futurs travaux.

Pour chaque échantillon, nous avons estimé le mode à l'aide de l'estimation plug-in $\sigma^2$ et nous avons calculé le biais, la variance et l'erreur quadratique moyenne (EQM) de l'estimateur proposé. Les résultats sont présentés dans les tableaux **1** et **2**.

Rappelons qu'en estimation non paramétrique, l'optimalité (au sens de l'erreur quadratique moyenne) n'est pas sérieusement influencée par le choix du noyau $K$ mais est affectée par le choix de la fenêtre $h_n$. Dans cette étude, la largeur de $h_n$ est choisie pour satisfaire les hypothèses ci-dessus, et le noyau K est gaussien. La largeur de bande $h_n$ que nous avons utilisée pour estimer le mode est celle utilisée par Moreira et de Uña-Álvarez (2012) basée sur la minimisation de l'AMISE de ($f_n$) (Erreur moyenne intégrée asymptotique), qui conduit à la fenêtre asymptotiquement optimale :

$$h_{opt} = \left(\frac{\alpha R(K) \int G^{-1} f}{R(f'')\mu_2^2(K)}\right)^{1/5} n^{-1/5} = \left(\frac{4}{3}\alpha \int G^{-1} f\right)^{1/5} \sigma n^{-1/5}$$

Comme on peut le voir dans les tableaux **1** et **2**, la qualité de l'estimateur ne semble pas être affectée par les pourcentages de troncature, et l'EQM diminue lorsque la taille de l'échantillon augmente.

De plus, pour illustrer le comportement de l'estimateur, nous avons tracé pour différentes valeurs de n, l'histogramme et le graphique Q-Q plot correspondant par rapport à la distribution gaussienne standard dans les figures 1 à 4 pour le modèle 1 et les figures 5 à 8 par rapport à la distribution de Weibull pour le modèle 2. En outre, le niveau de signification (p-valeur) du test de normalité de Shapiro-Wilk est supérieur à 0,05 dans tous les scénarios simulés. L'hypothèse de normalité est donc fortement conservée.

**Table 2.1. Moyenne estimée du biais, de la variance et de l'EQM, dans le cas d'une décroissance exponentielle, 500 répétitions**

| PT | N | N | Bias | Var | MSE |
|---|---|---|---|---|---|
| 70% | 167 | 50 | 0.01469 | 0.11657 | 0.34174 |
| | 500 | 150 | 0.00097 | 0.06888 | 0.26246 |
| | 1000 | 300 | 0.02786 | 0.04511 | 0.21421 |
| 50% | 100 | 50 | 0.03414 | 0.11857 | 0.11974 |
| | 300 | 150 | 0.01619 | 0.07090 | 0.07116 |
| | 600 | 300 | 0.01491 | 0.04946 | 0.04968 |
| 30% | 72 | 50 | -0.00580 | 0.10191 | 0.31928 |
| | 215 | 150 | -0.00061 | 0.06139 | 0.24778 |
| | 429 | 300 | -0.00509 | 0.04429 | 0.21053 |
| 10% | 56 | 50 | -0.02383 | 0.10695 | 0.32790 |
| | 167 | 150 | 0.01403 | 0.06385 | 0.25307 |
| | 334 | 300 | 0.00483 | 0.04243 | 0.20605 |

**Table 2.2. Moyenne estimée du biais, de la variance et de l'EQM, cas de la queue lourde, 500 répétitions**

| PT | N | n | Bias | Var | MSE |
|---|---|---|---|---|---|
| 70% | 167 | 50 | -0.00903 | 0.00818 | 0.09091 |
| | 500 | 150 | 0.00273 | 0.00476 | 0.06902 |
| | 1000 | 300 | -0.00214 | 0.00456 | 0.06755 |
| 50% | 100 | 50 | -0.01539 | 0.00820 | 0.09185 |
| | 300 | 150 | -0.00770 | 0.00469 | 0.06889 |
| | 600 | 300 | -0.00558 | 0.00331 | 0.05779 |
| 30% | 72 | 50 | -0.01044 | 0.00839 | 0.09221 |
| | 215 | 150 | -0.00252 | 0.00443 | 0.06663 |
| | 429 | 300 | -0.00608 | 0.00348 | 0.05929 |
| 10% | 56 | 50 | 0.00480 | 0.00848 | 0.09223 |
| | 167 | 150 | 0.01018 | 0.00462 | 0.06874 |
| | 334 | 300 | 0.00837 | 0.00337 | 0.05865 |

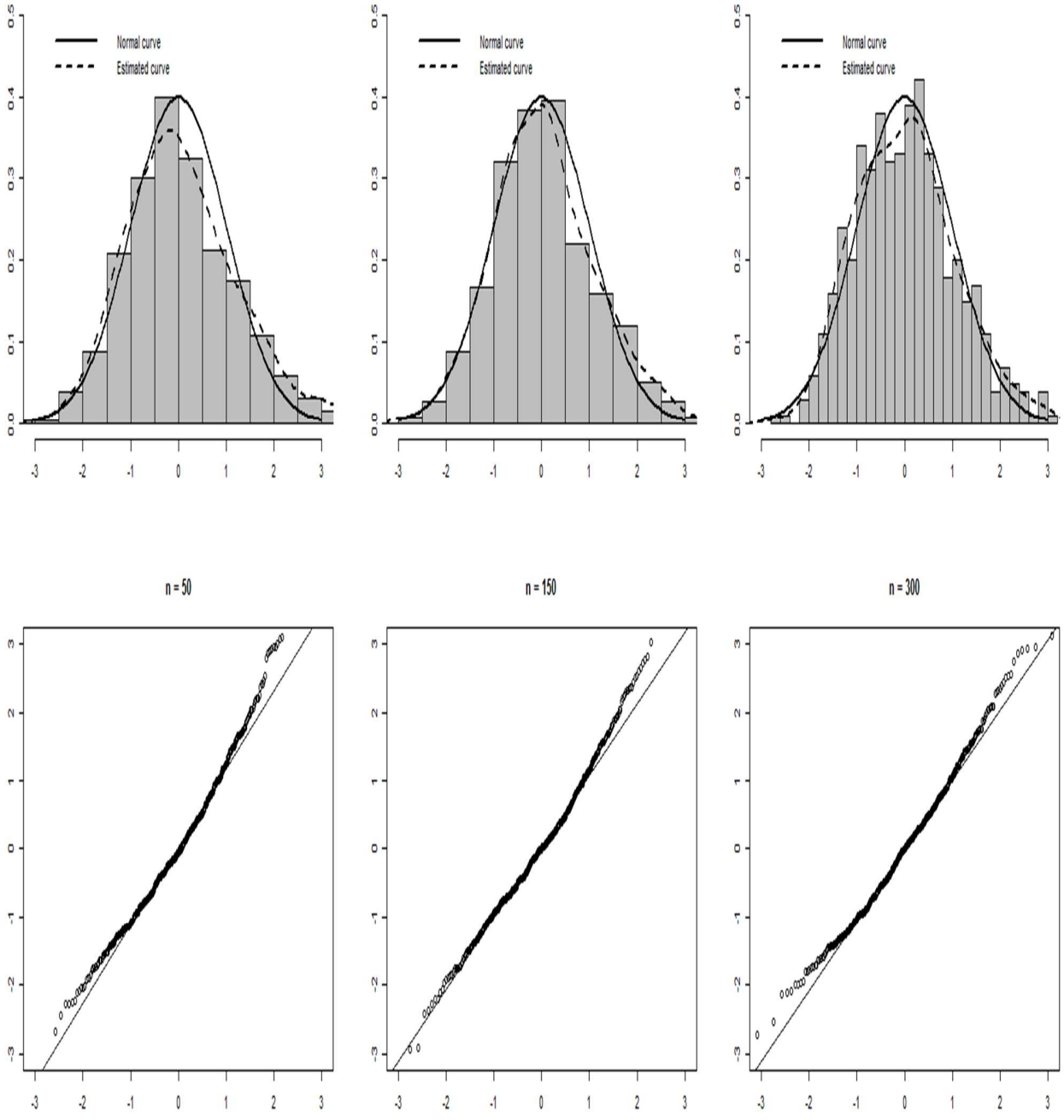


**Figure 1 (Modèle 1) : α = .30, B=500, n=50, 150 et 500 respectivement.**

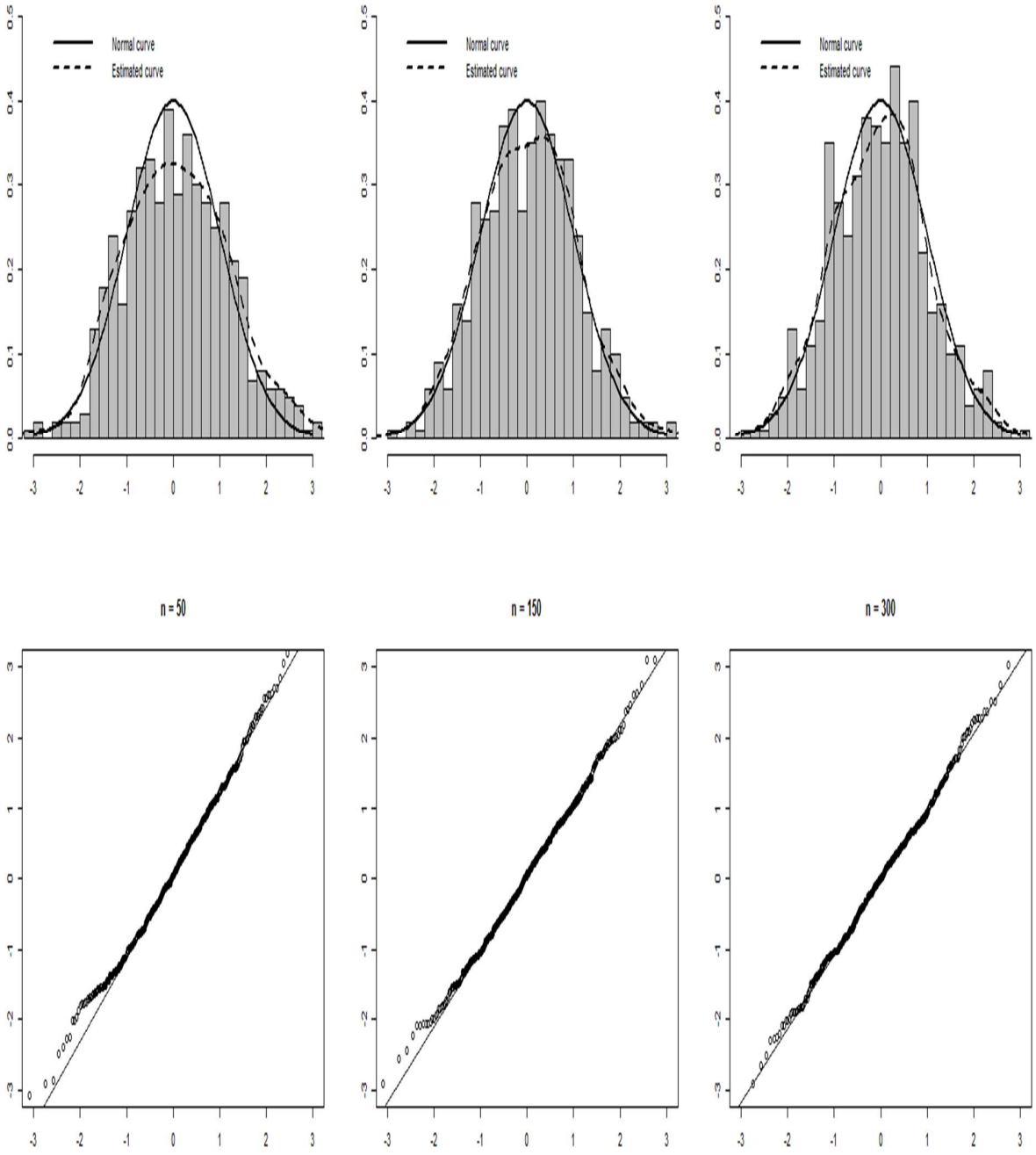


**Figure 2 . (Modèle 1) : α = 0,50, B=500, n=50, 150 et 500 respectivement.**

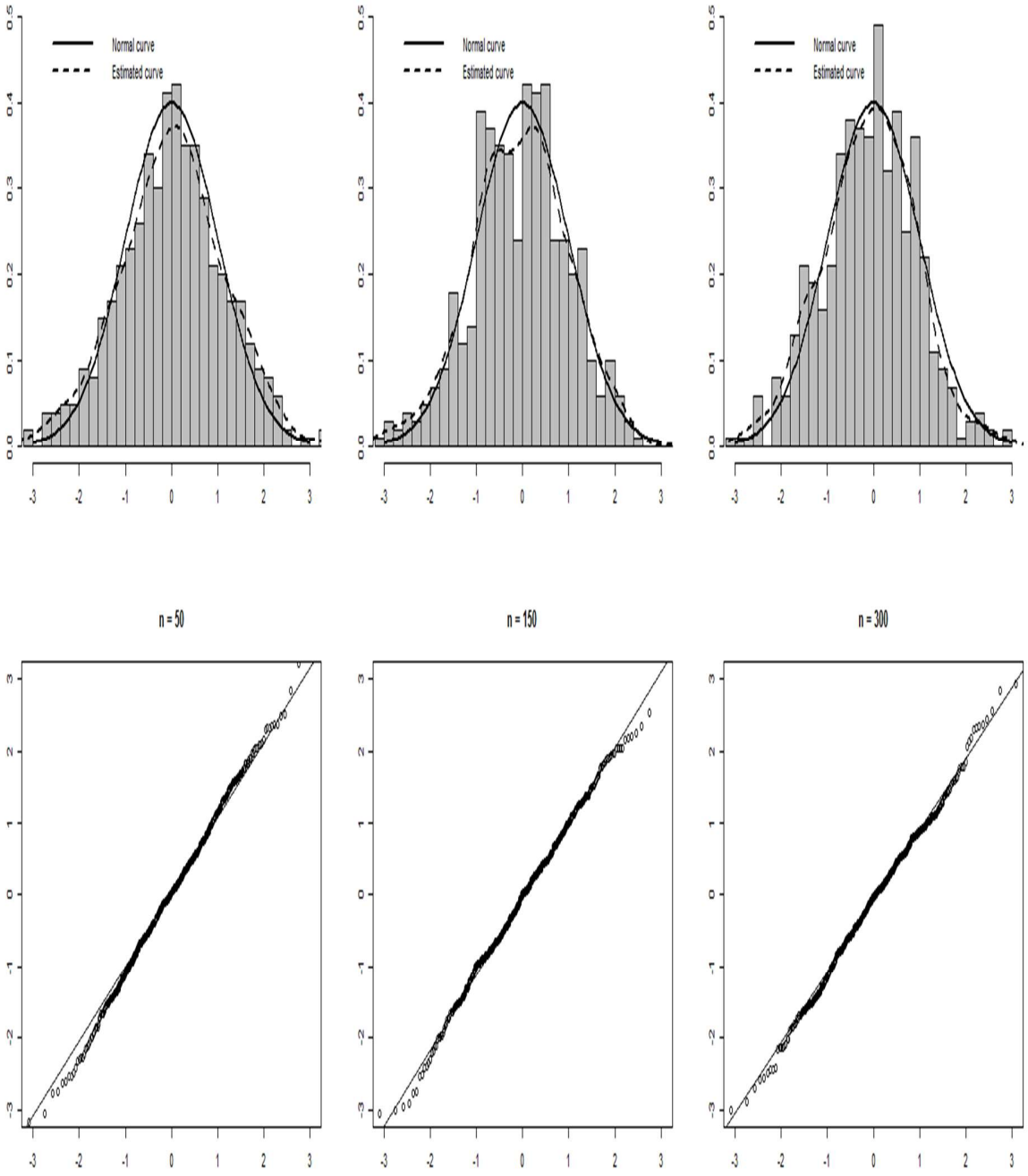


**Figure 3 (Modèle 1) : α = 0,70, B=500, n=50, 150 et 500 respectivement.**

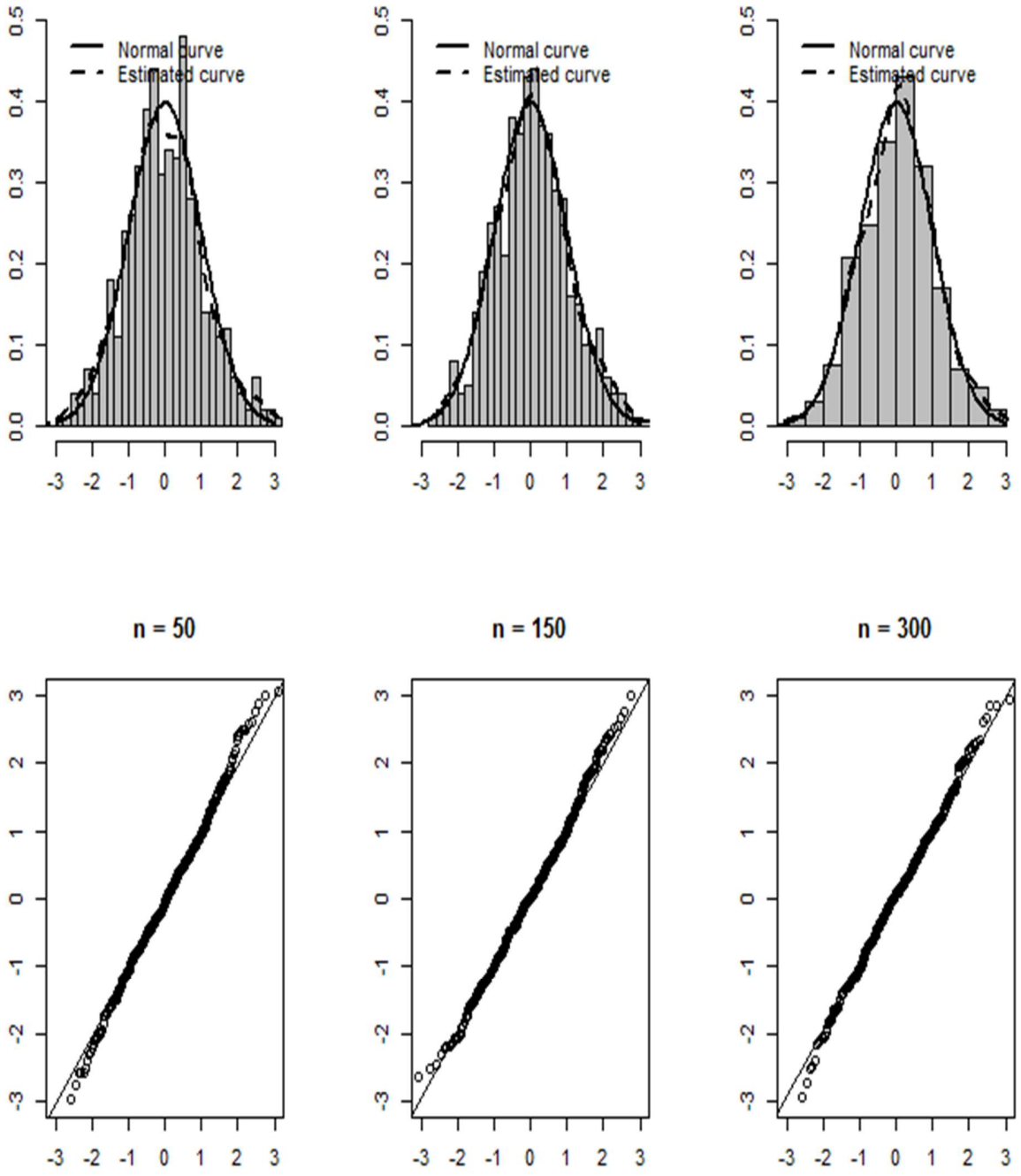


**Figure 4 (Modèle 1) : α = .90, B=500, n=50, 150 et 500 respectivement.**

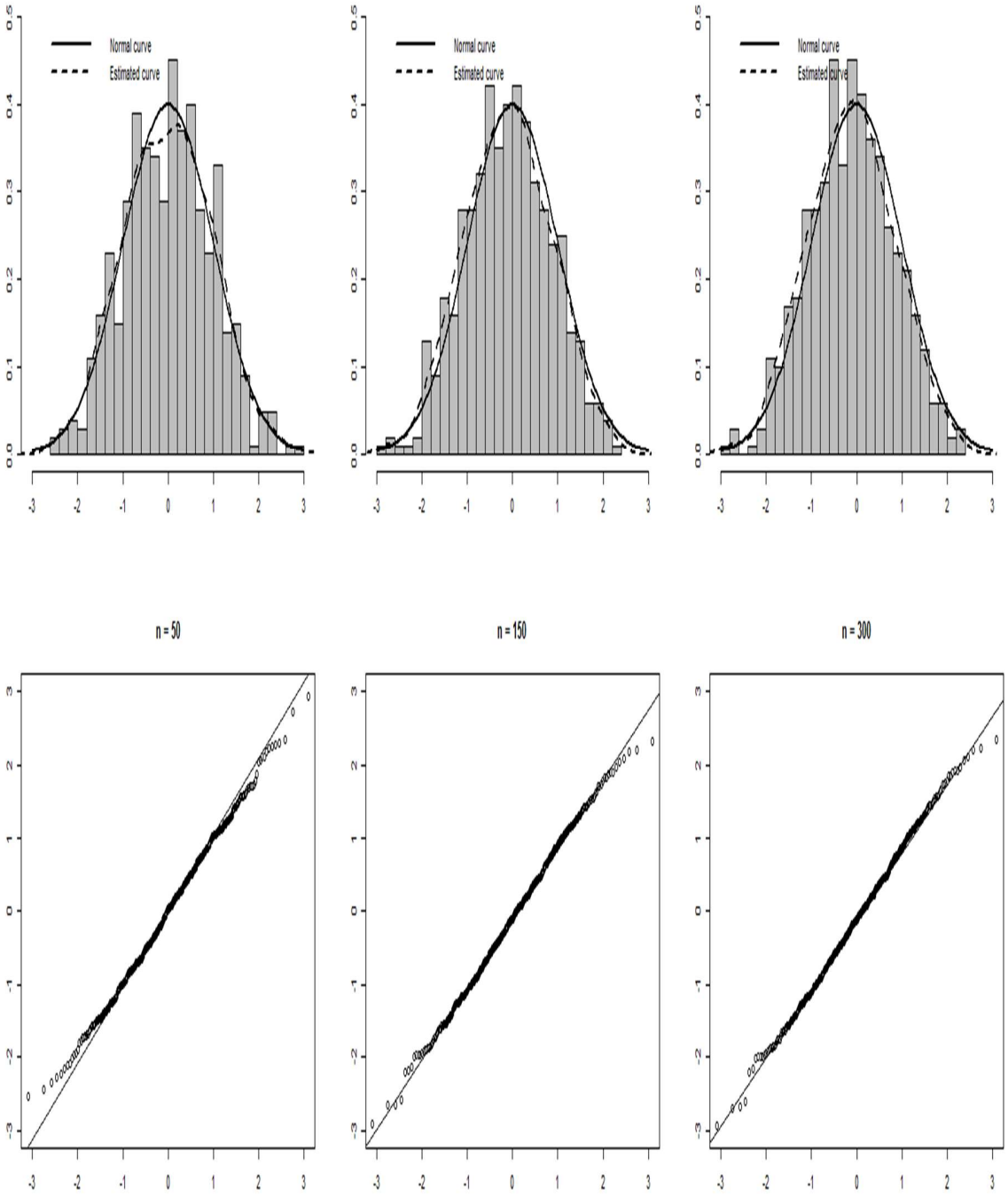


**Figure 5 (Modèle 2) : α = 0,30, B=500, n=50, 150 et 500 respectivement.**

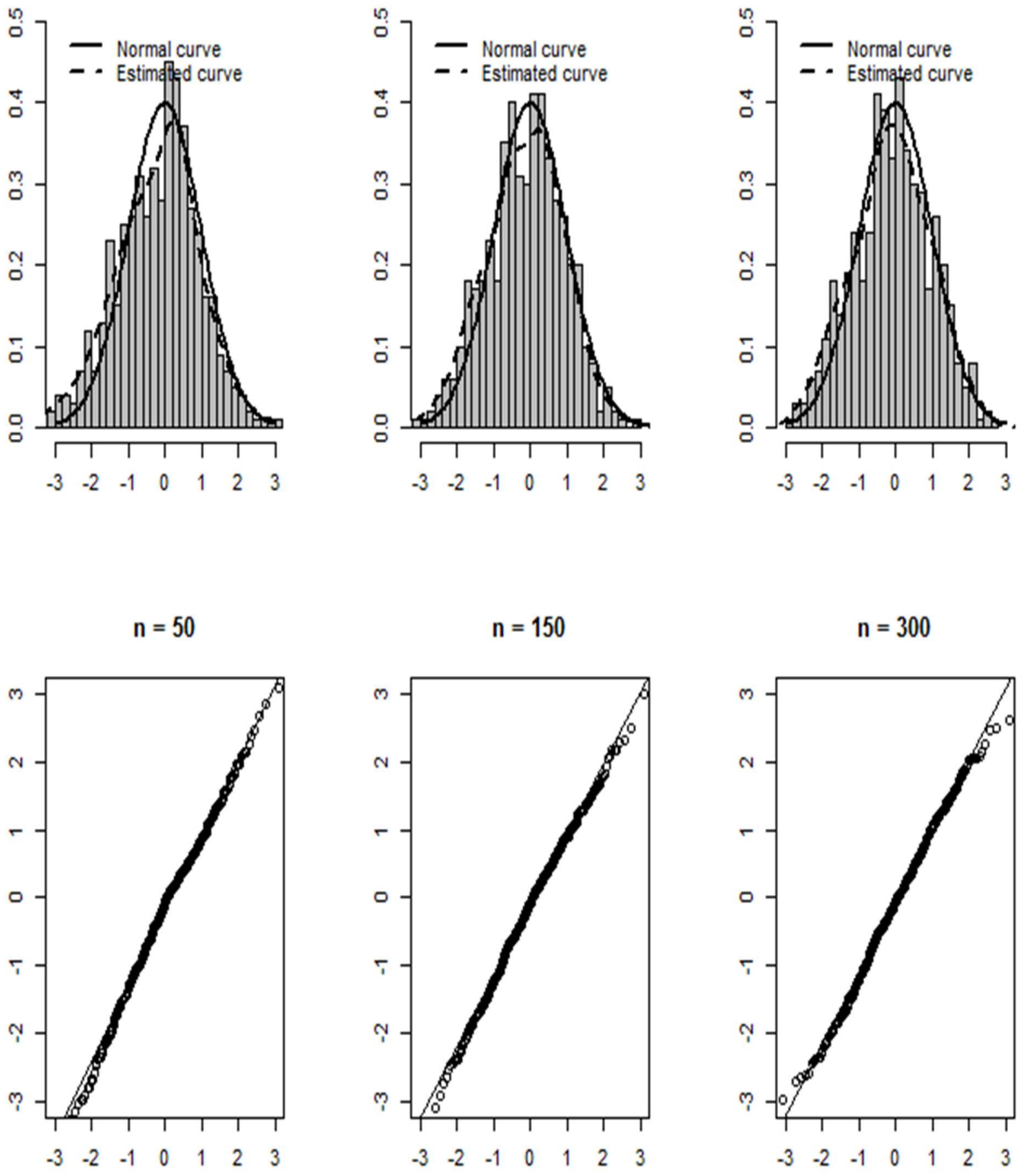


**Figure 6 (modèle 2) : α = 0,50, B=500, n=50, 150 et 500 respectivement.**

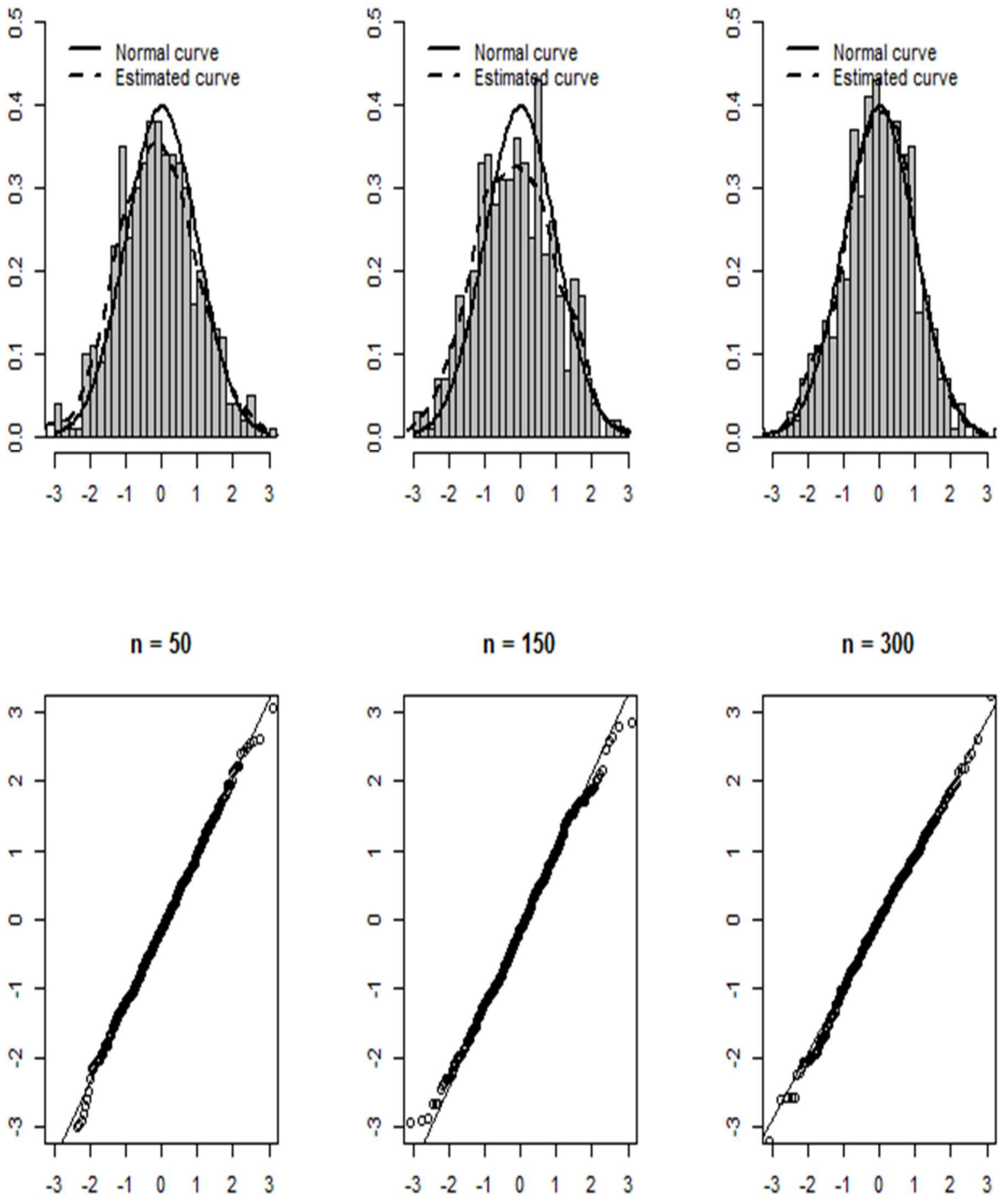


**Figure 7 (Modèle 2) : α = 0,70, B=500, n=50, 150 et 500 respectivement.**

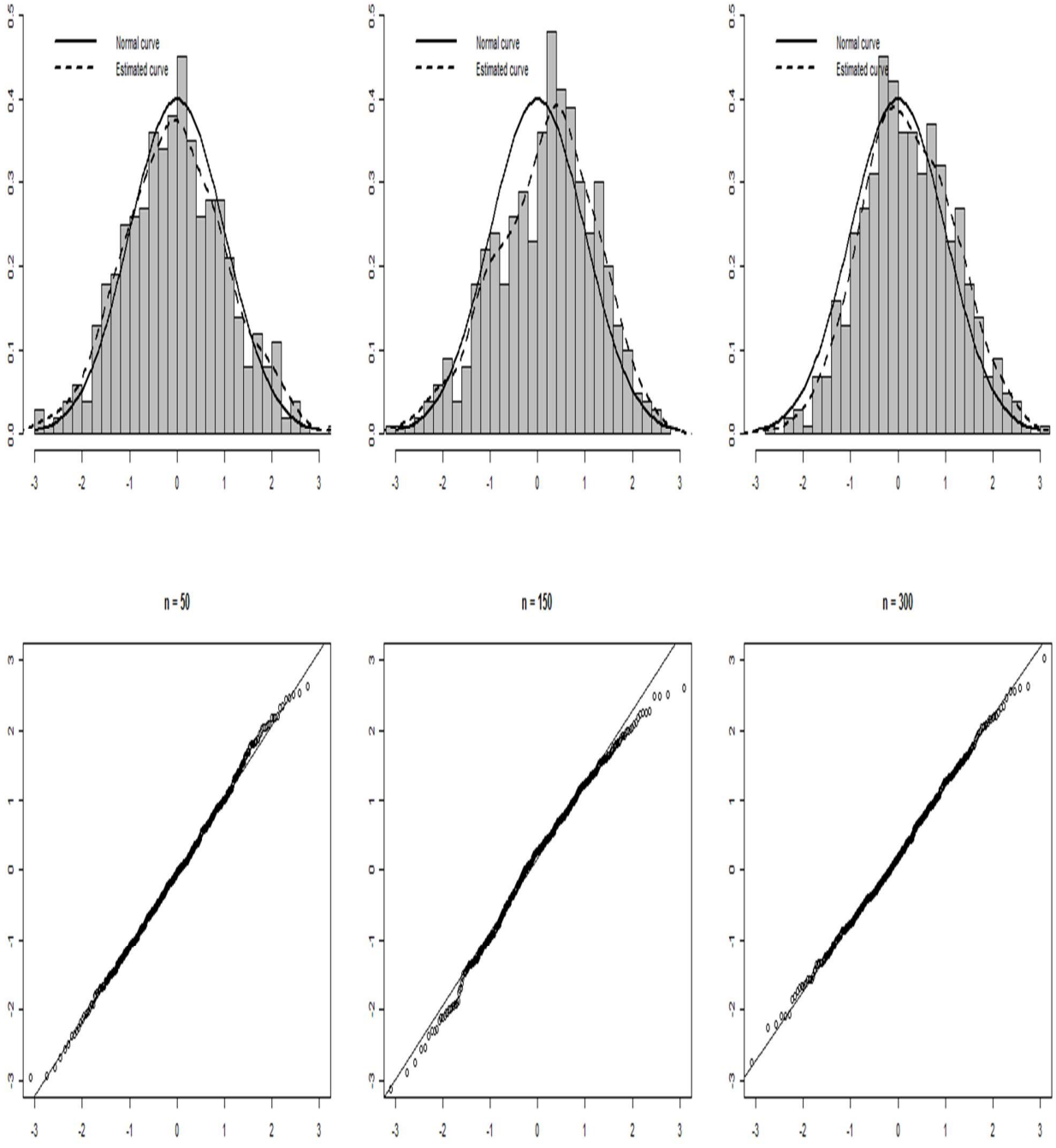


**Figure 8 (Modèle 2) : α = 0,90, B=500, n=50, 150 et 500 respectivement.**

## 2.5. Résultats auxiliaires et preuves

Pour des raisons techniques, nous devons introduire la densité du pseudo-estimateur, notée $\tilde{f}_n$, et par analogie, à (2.2) nous la définissons par

$$\tilde{f}_n(y) = \frac{\alpha}{nh} \sum_{i=1}^{n} \frac{1}{G(Y_i)} K\left(\frac{y-Y_i}{h}\right),$$

sa dérivée est donnée par

$$\tilde{f}_n^{(j)}(y) = \frac{\alpha}{nh^{j+1}} \sum_{i=1}^{n} \frac{1}{G(Y_i)} K\left(\frac{y-Y_i}{h}\right)$$

**Lemme 2.5.1 :** Sous **(A1)**,**(A3)** et **(A4)** on a

$$\sup_{y\in\Omega} \left|E\left[\tilde{f}_n(y)\right] - f(y)\right| = O\left(h^2\right)$$

***Preuve :***

En utilisant un changement de variable et un développement de Taylor, sous les hypothèses **(A1)**,**(A3)** et **(A4)** on a

$$E\left[\tilde{f}_n(y)\right] - f(y) = \frac{1}{h} E\left[\frac{\alpha}{nh} \sum_{i=1}^{n} \frac{1}{G(Y_i)} K\left(\frac{y-Y_i}{h}\right)\right] - f(y)$$

$$= \frac{1}{\boldsymbol{h}} \int \frac{\alpha}{G(t)} K\left(\frac{y-t}{h}\right) f(t)dt - f(y)$$

$$= \frac{1}{\boldsymbol{h}} \int K\left(\frac{y-t}{h}\right) f(t)dt - f(y)$$

$$= \int K(u)\left[f(y) - huf^{(1)}(y) + \frac{h^2u^2}{2} f^{(2)}(y) + O\left(h^2\right)\right] du - f(y)$$

$$= \int K(u) \frac{h^2u^2}{2} f^{(2)}(y)du$$

Ainsi,

$$L_1 = \left|E\left[\tilde{f}_n(y)\right] - f(y)\right|$$
$$\leq \frac{h^2}{2} \sup_{y\in\Omega}\left|f^{(2)}(y)\right| \int u^2 K(u)du$$
$$= O(h^2)$$

Le résultat est valable.

**Lemme 2.5.2 :** Sous les hypothèses **(A2)** et **(A4)**, pour $n$ assez grand, on a

$$\sup_{y\in\Omega}\left|\hat{f}_n(y) - \tilde{f}_n(y)\right| = O\left((nh)^{-1/2}\right)$$

***Preuve*** **:** Nous avons la décomposition suivante

$$\hat{f}_n(y) - \tilde{f}_n(y) = \frac{\alpha_n}{nh}\sum_{i=1}^{n}\frac{1}{G_n(Y_i)}K\left(\frac{y-Y_i}{h}\right) - \frac{\alpha}{nh}\sum_{i=1}^{n}\frac{1}{G(Y_i)}K\left(\frac{y-Y_i}{h}\right)$$

$$L_2 = \left|\hat{f}_n(y) - \tilde{f}_n(y)\right| \leq \frac{1}{nh}\left|\frac{\alpha_n}{G_n(y)} - \frac{\alpha}{G(y)}\right|\sum_{i=1}^{n}K\left(\frac{y-Y_i}{h}\right)$$

$$\leq \frac{1}{nh}\sup_{y\in\Omega}\left|\frac{\alpha_n}{G_n(y)} - \frac{\alpha}{G(y)}\right|$$

Comme indiqué dans Moreira et de Uña-Álvarez (2012) $G_n$ is $\sqrt{n}$ –estimateur consistent de $G_n$. En fait, la consistance $\sqrt{n}$ –$G_n$ est une consequence de $F_n$et $H_n^*$ (voir (Shen, 2010b) pour plus de détails). Par conséquent, en vertu de la régularité, nous obtenons le résultat.

**Lemme 2.5.3 :** Sous les hypothèses **(A2),(A3),(A4)** et **(A7 )**, on a

$$\sup_{y\in\Omega}\left|\tilde{f}_n(y) - E\left[\tilde{f}_n(y)\right]\right| = O\left(\sqrt{\frac{Logn}{nh}}\right)$$

***Preuve :***

Notez que

$$L_3 = \sup_{y\in\Omega}\left|\tilde{f}_n(y) - E\left[\tilde{f}_n(y)\right]\right|$$

Sous **(A3), (A8);** La sequence

$$\xi_n = \left\{ \Phi_y(u) = \frac{\alpha}{nh\,G(y)} K\left(\frac{y-u}{h}\right);\ a_F \le y \le b_F \right\} \quad n \ge 1$$

est V-C classes de fonctions mesurables limitées par leur enveloppe respective

$\boldsymbol{U_n = \frac{\|K\|_\infty}{nhG(y)}}$ Moreover ; sous (**A4)**

$$E\left[\Phi_y^2(Y)\right] \le \frac{1}{nh^2}\ \frac{\|K\|_2^2 \|f\|_\infty}{G(y)} \le \frac{\|f\|_\infty}{nh^2\ G(y)} = \sigma_n^2$$

avec $\sigma_n \le U_n$ for $n$ suffisamment grand.

L'application de l'inégalité de Talagrand (voir proposition 2.2 dans Giné et Guillou (2001)) avec

$\boldsymbol{t = B_3\sqrt{\frac{logn}{nh^2}}}$ pour une constante positive $B_3$, nous obtenons

$$P\left\{ \sup_{\Phi_y \in \xi_n} \left| \sum_{i=1}^{n} \left\{ \Phi_y(Y_i) - E\left[\Phi_y(Y)\right] \right\} \right| \ge\ B_3 \sqrt{\frac{logn}{nh^2}} \right\}$$

$$\le B_2\, exp\left\{ \frac{-1}{B_2}\ \frac{B_3\sqrt{\frac{logn}{nh^2}}}{\|K\|_\infty}\ nh_nG(y)\ log\left[ 1 + \frac{B_3\sqrt{\frac{logn}{nh_n^2}}\ \frac{\|K\|_\infty}{nh_nG(y)}}{\left[ \sqrt{n}\frac{\sqrt{\|f\|_\infty}}{n\sqrt{hG(y)}} + \frac{\|K\|_\infty}{nhG(y)}\ \sqrt{log\,\frac{VU_n}{\sigma_n}} \right]^2} \right] \right\}$$

Sous **(A2)** en utilisant $\log(1+w) \sim w\ (for\ w \longrightarrow\ 0)$ la dernière quantité est de mise

$$B_2 exp\left\{ -\frac{1}{B_2}\ \frac{B_3\sqrt{\frac{logn}{nh^2}}}{\|K\|_\infty}\ nh_nG(y) \frac{B_3\sqrt{\frac{logn}{nh^2}}\ \frac{\|K\|_\infty}{nhG(y)}}{B_2 n \frac{\sqrt{\|f\|_\infty}}{n^2\sqrt{hG(y)}}\ \sigma_n^2} \right\} = B_2\, exp\left\{ -\frac{1}{B_2^2}\ \frac{B_3^2\frac{logn}{nh^2}}{n\frac{\|f\|_\infty}{n^2h^2\ G(y)}} \right\}$$

$$= B_2 n^{-\frac{B_3^2\ G(y)}{B_2^2\|f\|_\infty}}$$

Ce qui, pour n suffisamment grand et par un choix approprié de $B_3$, peut être réalisée à $O(n^{-3/2})$. Ce dernier étant un terme général d'une série sommable, il s'agit alors d'une application directe du lemme de Borel-Cantelli et le résultat est

$$L_3 = O\left(\sqrt{\frac{logn}{nh^2}}\right) \; as \; n \longrightarrow \infty$$

***Preuve de la Proposition 2.3.1***

En utilisant l'inégalité triangulaire, nous avons

$$\sup_{y\in\Omega}\left|\hat{f}_n(y) - f(y)\right| \leq \sup_{y\in\Omega}\left|\hat{f}_n(y) - \tilde{f}_n(y)\right| + \sup_{y\in\Omega}\left|\tilde{f}_n(y) - E\tilde{f}_n(y)\right| + \sup_{y\in\Omega}\left|E\tilde{f}_n(y) - f(y)\right|$$

Alors les lemmes **2.5.1, 2.5.2 et 2.5.3** donnent le résultat.

***Preuve du Théorème 2.3.1***

L'argumentation classique nous donne

$$\begin{aligned}\left|f(\hat{\theta}_n) - f(\theta)\right| &\leq \left|f(\hat{\theta}_n) - \hat{f}_n(\hat{\theta}_n)\right| + \left|\hat{f}_n(\hat{\theta}_n) - f(\theta)\right| \\ &\leq \sup_{y\in\Omega}\left|\hat{f}_n(y) - f(y)\right| + \left|\hat{f}_n(\hat{\theta}_n) - f(\theta)\right| \\ &\leq 2\sup_{y\in\Omega}\left|\hat{f}_n(y) - f(y)\right| \qquad (\mathbf{2.4})\end{aligned}$$

Pour une deuxième partie, un développement de Taylor de $f(.)$ au voisinage de $\theta$ donne

$$f(\hat{\theta}_n) - f(\theta) = \frac{1}{2}\left(\hat{\theta}_n - \theta\right)^2 f''(\bar{\theta})$$

avec $\bar{\theta}$ entre $\hat{\theta}_n$ et $\theta$ . Donc à partir de (**2.4)**, et (**A4)** on obtient :

$$\left(\hat{\theta}_n - \theta\right)^2\left|f''(\bar{\theta})\right| \leq 4\sup_{y\in\Omega}\left|\hat{f}_n(y) - f(y)\right|$$

Ainsi,

$$\left|\hat{\theta}_n - \theta\right| \leq 2\sqrt{\frac{sup\left|\hat{f}_n(y) - f(y)\right|}{\left|f''(\bar{\theta})\right|}}$$

En utilisant la **Proposition 2.3.1,** la preuve est complète.

***Preuve du Théorème 2.3.2*** A partir de **(2.3**), nous obtenons la décomposition suivante

$$\sqrt{nh^3}\left(\hat{\theta}_n - \theta\right) = \sqrt{nh^3}\,\frac{\hat{f}_n^{(1)}(\theta) - \tilde{f}_n^{(1)}(\theta)}{\hat{f}_n''(\bar{\theta}_n)}$$

$$+\sqrt{nh^3}\,\frac{\tilde{f}_n^{(1)}(\theta) - E(\tilde{f}_n^{(1)}(\theta))}{\hat{f}_n^{(2)}(\bar{\theta}_n)} + \sqrt{nh^3}\,\frac{E(\tilde{f}_n^{(1)}(\theta))}{\hat{f}_n^{(2)}(\bar{\theta}_n)}$$

$$= \frac{J_1 + J_2 + J_3}{\hat{f}_n^{(2)}(\bar{\theta}_n)}$$

Pour prouver le résultat, nous établissons que $J_1$ et $J_3$ sont négligeables et que $J_2$ est asymptotiquement normal et que $\hat{f}_n^{(2)}(\bar{\theta}_n)$ converge en probabilité vers $f^{(2)}(\theta)$

$$\hat{f}_n^{(2)}(\bar{\theta}_n) \longrightarrow f^{(2)}(\theta).$$

Pour le premier terme $J_1$, on a

$$J_1 = \hat{f}_n^{(1)}(\theta) - \tilde{f}_n^{(1)}(\theta)$$

$$= \frac{\alpha_n}{nh^2}\sum_{i=1}^{n}\frac{1}{G_n(Y_i)}K^{(1)}\left(\frac{\theta - Y_i}{h}\right) - \frac{\alpha}{nh^2}\sum_{i=1}^{n}\frac{1}{G(Y_i)}K^{(1)}\left(\frac{\theta - Y_i}{h}\right)$$

$$|J_1| \leq \sqrt{nh^3}\,\sup_{y\in\Omega}\left|\frac{\alpha_{\boldsymbol{n}}}{G_n(y)} - \frac{\alpha}{G(y)}\right|\frac{1}{nh^2}\left|\sum_{i=1}^{n}K^{(1)}\left(\frac{\theta - Y_i}{h}\right)\right|$$

Rappelons que

$$\sup_{y}\left|\frac{\alpha_{\boldsymbol{n}}}{G_n(y)} - \frac{\alpha}{G(y)}\right| = O\left((nh)^{-\frac{1}{2}}\right)$$

(voir la preuve du lemme 2.5.2) et puisque

$$\frac{1}{nh^2}\sum_{i=1}^{n} K^{(1)}\left(\frac{\theta - Y_i}{h}\right) = (f_n^*)^{(1)}(\theta), \longrightarrow (f^*)^{(1)}(\theta)$$

ici $(f^*)^{(1)}(.)$représente la dérivée première de la fonction de densité $f^*(.)$des données observées, qui n'est pas nécessairement égale à zéro au point $\theta$, qui nous permettent de conclure que $J_1$est négligeable.

Nous énonçons maintenant les résultats suivants pour $J_3$ :

$$E\left[\tilde{f}_n^{(1)}(\theta)\right] = \frac{\alpha}{nh^2}\int \frac{K^{(1)}\left(\frac{\theta-u}{h}\right)}{G(u)} f(u)du$$

$$= \frac{1}{h^2}\int K^{(1)}\left(\frac{\theta-u}{h}\right) f(u)du$$

$$= \frac{1}{h}\int K^{(1)}(r) f(\theta - rh)dr$$

En intégrant par partie, on obtient

$$E\left[\tilde{f}_n^{(1)}(\theta)\right] = \int K(r) f^{(1)}(\theta - rh)dr$$

Par le développement de Taylor de $f^{(1)}(.)$, **(A1),(A3)** et la définition du mode, nous obtenons

$$\sqrt{nh^3}\; E\left[\tilde{f}_n^{(1)}(\theta)\right] = \sqrt{nh^7}\int r^2 K(r) f^{(3)}(\bar{\theta})dr$$

avec $\bar{\theta}$ entre $\theta$ and $\theta - rh$, par **(A3), (A4)** on conclut que $J_3 \longrightarrow 0 \quad qd \quad n \longrightarrow \infty$

Enfin, pour énoncer la normalité asymptotique de $J_2$, nous devons prouver que

$$Var[J_2] = \frac{\alpha f(\theta)}{G(\theta)}\int \left[K^{(1)}(r)\right]^2 dr \quad as \;\; n \longrightarrow \infty.$$

Notez que

$$Var[J_2] = nh^3\, Var\left[\frac{\alpha}{h^2}\frac{1}{G(y)}K^{(1)}\left(\frac{\theta - y}{h}\right)\right]$$

$$= nh^3\, E\left[\frac{\alpha}{h^2}\frac{1}{G(y)}K^{(1)}\left(\frac{\theta - y}{h}\right)\right]^2 - nh^3\left\{E\left[\frac{\alpha}{h^2}\frac{1}{G(y)}K^{(1)}\left(\frac{\theta - y}{h}\right)\right]\right\}^2$$

$$= I_1 + I_2$$

D'une part, le caractère négligeable de $J_3$ donne $I_2 \longrightarrow 0$ qd $n \longrightarrow \infty$.

D'autre part, en changeant de variable, nous pouvons écrire

$$I_1 = \int \frac{\alpha}{G(\theta)} \left[K^{(1)}(r)\right]^2 f(\theta - rh)\, dr$$

puisque $G(.)$est continue, nous avons sous (A1), (A3), et (A5)

$$I_1 = \frac{\alpha f(\theta)}{G(\theta)} \int \left[K^{(1)}(r)\right]^2 dr + O(1) \qquad as \;\; n \longrightarrow \infty$$

Ce qui donne le résultat.

La deuxième étape, c'est de montrer la condition du théorème de Berry-Esséen qui est :

$$\sum_{i=1}^{n} E\left[\left|\tilde{f}_n^{(1)}(\theta) - E(\tilde{f}_n^{(1)}(\theta))\right|^3\right] < \infty$$

Pour plus de détails voire Chow et Teicher (1997) page 322.

On pose :

$$J_2 = \sum_{i=1}^{n} R_{i,n}(\theta) = \sqrt{nh^3}\, \left(\tilde{f}_n^{(1)}(\theta) - E\left[\tilde{f}_n^{(1)}(\theta)\right]\right)$$

$$R_{i,n}(\theta) = \frac{\alpha}{\sqrt{nh}} \left\{ \frac{1}{G(Y_i)} K^{(1)}\left(\frac{\theta - Y_i}{h}\right) - E\left[\sum_{i=1}^{n} \frac{1}{G(Y_i)} K^{(1)}\left(\frac{\theta - Y_i}{h}\right)\right] \right\}$$

En appliquant la $C_r - in\acute{e}galit\acute{e}$ (Loève, 1977, page 155) on a :

$$\left(nh^3\right)^{-3/2} E\left[\left|R_{i,n}(\theta)\right|^3\right] \leq 4\left(nh^3\right)^{-3/2} \left\{ E\left[\left|\frac{\alpha}{G(Y_i)} K^{(1)}\left(\frac{\theta - Y_i}{h}\right)\right|^3\right] + \left|E\left[\frac{\alpha}{G(Y_i)} K^{(1)}\left(\frac{\theta - Y_i}{h}\right)\right]\right|^3 \right\}$$

Les deux termes à droite de l'inégalité sont finis sous les hypothèses $(\boldsymbol{A_1}) - (\boldsymbol{A_3})$, alors

$$J_2 = \sum_{i=1}^{n} R_{i,n}(\theta) < \infty$$

On conclut que $J_2$ est normalement distribuée.

Complétons maintenant la preuve du théorème **3.2**. en montrant que $\hat{f}_n^{(2)}(\bar{\theta}_n)$ converge en probabilité vers le nombre réel $f^{(2)}(\theta)$.

Nous avons pour *n* suffisamment grand

$$\left|\hat{f}_n^{(2)}(\bar{\theta}_n) - f^{(2)}(\theta)\right| \leq \sup_{y\in\Omega}\left|\hat{f}_n^{(2)}(y) - f^{(2)}(y)\right| + \left|f^{(2)}(\bar{\theta}_n) - f^{(2)}(\theta)\right|$$

avec $\bar{\theta}_n$ est compris entre $\hat{\theta}_n$ et $\theta$. On observe que :

$$\hat{f}_n^{(2)}(y) - f^{(2)}(y) = \left\{\frac{1}{h^2}\int K^{(2)}\left(\frac{y-t}{h}\right)\hat{f}_n(t)dt - \frac{1}{h^2}\int K^{(2)}\left(\frac{y-t}{h}\right)f(t)dt\right\}$$

$$+\left\{\frac{1}{h^2}\int K^{(2)}\left(\frac{y-t}{h}\right)f(t)dt - K\left(\frac{y-t}{h}\right)f^{(2)}(y)dt\right\} = T_1 + T_2$$

Par le développement de Taylor, l'intégration par partie et les hypothèses $(\mathbf{A_1}), (\mathbf{A_2}), (\mathbf{A_6})$ et $(\mathbf{A_7})$ on obtient

$$|T_1| \leq \frac{1}{h}\left|\hat{f}_n(y-rh) - f(y-rh)\right|\int K^{(2)}(r)dr$$

$$\leq \sup_{y\in\Omega}\left|\hat{f}_n(y) - f(y)\right|\int K^{(3)}(r)dr$$

En utilisant la Proposition 2.3.1., $|T_1| = O\left(max\left(\left(\frac{logn}{nh}\right)^{\frac{1}{2}}, h^2\right)\right)$

D'autre part, sous les hypothèses$(\boldsymbol{A3}), (\boldsymbol{A5})$ et en intégrant deux fois par partie, on a

$$T_2 = \frac{1}{h^2}\int K^{(2)}\left(\frac{y-t}{h_n}\right)f(t)dt - \int K\left(\frac{y-t}{h_n}\right)f^{(2)}(t)dt$$

$$= \frac{1}{h^2}\int K\left(\frac{y-t}{h}\right)f^{(2)}(t)dt - \int K\left(\frac{y-t}{h}\right)f^{(2)}(t)dt$$

En utilisant le développement de Taylor, nous avons

$$|T_2| \leq h\int r\ K(r)\left|f^{(3)}(r^*)\right|dr = O(h)$$

où $r^*$ est entre $y - rh$ et $y$. Par conséquent, la continuité de $f^{(2)}(t)$ assure la convergence presque sûre $f^{(2)}(\bar{\theta}_n)$ vers $f^{(2)}(\theta)$. En outre, en raison de la convergence en probabilités de $\hat{f}_n^{(2)}(y)$ vers $f^{(2)}(y)$ sur Ω. On conclut que $\hat{f}_n^{(2)}(\bar{\theta}_n)$ converge en probabilité vers le nombre réel $f^{(2)}(\theta) \neq 0$. Ce qui met fin à la preuve.

# Chapitre 3 :

# Estimation du mode conditionnel pour des données doublement tronquées

## 3.1. Introduction

Soit $X$ une v.a.r. pour tout $x$, on note par $f(./x)$ la probabilité conditionnelle de $Y$ sachant $X = x$. Pour tout $y$,

$$f(y/x) = \frac{f(x,y)}{l(x)},$$

avec $f(x,y)$ est la loi conjointe du couple $(X,Y)$ et $l(x)$ est la densité marginale de $X$.

Nous supposons que $f(y/x)$ à un mode unique $\theta(x)$, et le mode conditionnel de $Y$ sachant $X = x$, est définit par :

$$\theta(x) = \underset{y\in\mathbb{R}}{\operatorname{argmax}} f(y/x) \quad .$$

L'estimation du mode conditionnel a une longue histoire et a été étudiée par de nombreux auteurs dans la littérature statistique. Par exemple, (Parzen, 1962) a établi la consistance faible et la normalité asymptotique pour le cas i.i.d. et la consistance forte a été obtenue par Nadaraya (1965) et Van Ryzin (1969). Eddy (1982) a dérivé la normalité asymptotique sous des conditions plus faibles que celles imposées par (Parzen, 1962), voir aussi Eddy (1982). Chernoff (1964) a étudié l'estimateur du mode défini comme le centre de l'intervalle qui contient le plus d'observations. Collomb et al. (1986) ont étudié la consistance asymptotique de l'estimation du mode conditionnel par le noyau et donner un aperçu des situations où l'estimation du mode conditionnel par le noyau n'est pas possible.

Samata et Thavaneswaran (1990) ont montré que l'estimateur à noyau de la fonction de mode conditionnelle est cohérent et asymptotiquement normalement distribué.

Ezzahrioui et Ould-Saïd (2005) considèrent l'estimation du mode conditionnel lorsque les covariables prennent des valeurs dans un espace de fonctions abstraites et montrent que, sous certaines conditions de régularité, l'estimation du noyau du mode conditionnel est asymptotiquement normalement distribuée, sous le modèle de troncature à gauche Ould-Saïd et Tatachak (2007) ont construit un estimateur à noyau non paramétrique de la fonction de mode conditionnelle pour le modèle de troncature gauche et a établi le taux de cohérence uniforme forte de l'estimation ainsi que la normalité asymptotique du mode conditionnel. Vieu (1996) a obtenu un taux de convergence pour les estimations locales et globales de la fonction de mode. À notre connaissance, le problème de l'estimation de l'estimateur du mode conditionnel dans le cadre d'un modèle à double troncature n'a pas été abordé dans la littérature statistique. C'est l'objet central de ce chapitre.

La seule chose que nous pouvons voir à cause de la double troncature aléatoire est $(X^*, Y^*)$ quand $U^* \leq Y^* \leq V^*$ ou $(U^*, V^*)$ sont également observés. Au contraire, lorsque $U^* \leq Y^* \leq V^*$ n'est pas respecté, on ne voit rien. Nous supposons que les temps de troncature, comme il est d'usage avec la troncature aléatoire, sont indépendants de $(X^*, Y^*)$. Soient $(Y_1, X_1, U_1, V_1), \ldots, (Y_n, X_n, U_n, V_n)$ étant l'échantillon observé, il s'agit de données iid ayant la même distribution que $(X^*, Y^*, U^*, V^*)$ avec $U^* \leq Y^* \leq V^*$.

Dans la configuration doublement tronquée, cette probabilité relative d'observer $(X^*, Y^*) = (x, y)$ est donnée par $G(y) = P(U^* \leq y \leq V^*)$, depuis $(X^*, Y^*)$ et $(U^*, V^*)$ sont indépendantes. Cette fonction $G$ peut être estimée à partir des données par les principes du maximum de vraisemblance voir l'algorithme itératif de Moreira et de Uña-Álvarez (2012).

Pour toute distribution $W$ désignent les extrémités droite et gauche de son support par

$$a_w = inf\{t: W(t) > 0\} \text{ et } b_w = inf\{t: W(t) = 1\}.$$

Soit $H_1(u) = H(u, \infty)$ et $H_2(v) = H(-\infty, v)$ les distributions marginales de $U^*$ et $V^*$ respectivement. Nous avons souligné que $F$ et $H$ sont tous deux complètement identifiables si et seulement si

$$a_{H_1} \leq a_F \leq a_{H_2} \quad \text{and} \quad b_{H_1} \leq b_F \leq b_{H_2}.$$

Soit $F(./x)$ est la distribution conditionnelle de $Y^*$ sachant $X^* = x$, et soit

$$\alpha(x) = P(U^* \leq Y^* \leq V^*/X^* = x) = \int_{-\infty}^{+\infty} G(t)F(dt/x)$$

la probabilité conditionnelle de non troncature. On suppose que $G(t) > 0$. Soit $F^*(./x)$ la fonction de répartition conditionnelle observable, tel que

$$F^*(./x) = P(Y_1 \leq y/X_1 = x)$$

où

$$\hat{F}_n(y/x) = \alpha_n(x) \int_{-\infty}^{y} G_n(t)^{-1} F_n^*(dt/x)$$

est l'estimateur de $F^*(y/x)$ :

$$F^*(y/x) = \alpha(x)^{-1} \int_{-\infty}^{y} G(t)F(dt/x)$$

Rappelons que l'estimateur de la probabilité conditionnelle d'absence de troncature est représenté par

$$\alpha_n = \left( \int_{a_F}^{b_F} G_n(t)^{-1} F_n^*(dt/x) \right)^{-1}$$

est un estimateur de $\alpha$, voir (Shen, 2010b).

$F_n^*(y) = n^{-1} \sum_{i=1}^{n} I_{[Y_i \leq y]}$ est la fonction de distribution empirique ordinaire de $Y_i$, et

$$G_n(t) = \int_{\{u \leq t \leq v\}} H_n(du, dv)$$

est l'estimateur non paramétrique de $G(t) = P(U^* \leq t \leq V^*)$ qui peut être estimé par la méthode du maximum de vraisemblance. $K$ et $H$ sont des densités de probabilités sur $\mathbb{R}$ et $h_n$ est une suite de réels positifs tendant vers zéro quand $n$ tend vers l'infini.

$H_n(u, v)$ est l'estimateur non paramétrique du maximum du vraisemblance (NPMLE) de la distribution conjointe H des temps de troncature, voir (Moreira et de Uña-Álvarez, 2012) pour plus de détails.

Dans ce chapitre, nous présentons un nouvel estimateur et prouvons sa normalité asymptotique et sa convergence uniforme presque certaine. La section 4 contient divers résultats auxiliaires et leurs preuves en plus des preuves des principaux résultats. En outre, l'approche du noyau est le fondement de l'estimation fonctionnelle. La structure de ce chapitre est la suivante : un nouvel estimateur du mode conditionnel à noyau pour le modèle à double troncature est défini à la section 2. Les principaux résultats et hypothèses sont présentés dans la section 3.

## 3.2. Les estimateurs

Dans cette section, nous rappelons quelques résultats et définissons ensuite notre estimateur du mode. Nous désignons par $(U^*, V^*)$ la paire de variables de troncature, avec une fonction de distribution commune $H$, $Y^*$ n'est donc observé que lorsque $U^* \leq Y^* \leq V^*$ En outre, un vecteur de covariables $X^*$ est naturellement attaché à $Y^*$ et, , il ne sera disponible pour le chercheur que si Y n'est pas tronqué. Compte tenu du fait que les données d'échantillonnage sont indiquées par $(X_i, Y_i, U_i, V_i), 1 \leq i \leq n$ ; étant donné que $U^* \leq Y^* \leq V^*$ il s'agit de copies iid avec la distribution conditionnelle de $(X^*, Y^*, U^*, V^*)$ . On estime que $(U^*, V^*)$ est indépendant de $(X^*, Y^*)$.

Dans cet essai, nous supposerons que

$$a_F = inf\{t: F(t) > 0\} \quad \text{et} \quad b_F = inf\{t: F(t) = 1\}$$

avec $a_F$ et $b_F$ désignent les extrémités gauche et droite de la distribution F.

L'estimateur a noyau de la fonction de mode conditionnelle de $Y$ sachant $X = x$ est définie par

$$\hat{\theta}_n(x) = \underset{y \in \mathbb{R}}{\operatorname{argmax}}\, \hat{f}(y/x) \qquad \mathbf{(3.1)}$$

avec

$$\hat{f}_n(y/x) = \frac{\hat{f}_n(x,y)}{\hat{l}_n(x)}, \qquad \mathbf{(3.2)}$$

$$\hat{f}_n(x,y) = \frac{\alpha_n}{nh^2}\sum_{i=1}^{n} G_n(X_i)^{-1} K_h\left(\frac{x-X_i}{h}\right) L_h\left(\frac{y-Y_i}{h}\right) \qquad \mathbf{(3.3)}$$

et

$$\hat{l}_n(x) = \frac{\alpha_n}{nh}\sum_{i=1}^{n} G_n(X_i)^{-1} K\left(\frac{x-X_i}{h}\right) \qquad \mathbf{(3.4)}$$

Il convient de noter que l'estimateur $\widehat{\theta}_n(x)$ n'est pas nécessairement unique et nos résultats sont valables pour n'importe quelle valeur satisfaisant (3.1). Nous pouvons exprimer notre préférence en prenant

$$\widehat{\theta}_n(x) = inf\left\{a_F \le t \le b_F \ : \quad \hat{f}_n(t/x) = \sup_{a_F \le y \le b_F} \hat{f}_n(y/x)\right\}$$

Quelques notations supplémentaires sont nécessaires pour formuler nos résultats, pour tous les noyaux. $K$ ; $K^{(j)}$désigne la dérivée d'ordre $j$ de $K$, pour $(i,j) \in \mathbb{N}^2$, soit

$$f^{(i,j)}(x,y) = \frac{\partial^{(i+j)}}{\partial x^i \partial y^j} f(x,y),$$

et pour $j \ge 1$,

$$\hat{f}_n^{(0,j)}(x,y) = \frac{\partial^j}{\partial y^j}\hat{f}_n(x,y) = \frac{\alpha_n}{nh^{(2+j)}}\sum_{i=1}^{n} G_n^{-1}(Y_i)\, K\left(\frac{x-X_i}{h}\right) L^{(j)}\left(\frac{y-Y_i}{h}\right)$$

Définir $\Omega_0 = \{x \in \mathbb{R} : l(x) > 0\}$ et soit $\Omega$ un sous-ensemble compact de $\Omega_0$ .

Considérons maintenant les hypothèses de régularité suivantes :

## 3.3. Hypothèses et principaux résultats

$(\mathbf{A1})$ la densité jointe $f(.,.)$est différentiable jusqu'à l'ordre 4 et

$$sup_{x\ y}\left|f^{(i,j)}(x,y)\right| < \infty \ , \text{ pour } 1 \le i+j \le 4$$

$(\mathbf{A2})$ $f^{(0,2)}(x,y)$ et $f^{(2,0)}(x,y)$ sont continués.

$(\mathbf{A3})$ la densité marginale $l(.)$ admet une dérivée second continue.

(**A4**) $K$ est une fonction continue et positive, $\int_{\mathbb{R}} tK(t) = 0$ , $\int tL^{(1)}(t)dt = 0$ $et$

$$K(.)L^{(1)}(.) < \infty$$

(**A5**) $K$ est lipchitzienne, différentiable et bornée et L est trois fois différentiable et bornée.

(**A6**) $G(.)$ est continue en , $f(./.)$ $et$ $G^{-1}f(./.)$ sont deux fois continument différentiables.

(**A7**) La fenêtre $h_n := h$ et $h \longrightarrow 0$ satisfait $\frac{Logn}{nh^7} \longrightarrow 0$ $et$ $nh^7 \longrightarrow 0$ $qd\ n \longrightarrow \infty$.

(**A8**)$L\ et\ L^{(2)}\ sont\ lipshitizienne$ , $\mu_2(K) = \int t^2K(t)dt < \infty$, $R(K) = \int K^2(t)\ dt < \infty$,

$$\mu_2(L) = \int t^2L(t)dt < \infty\ ,\ R(L) = \int L^2(t)\ dt < \infty$$

**Remarque 3.3.1.** En gardant à l'esprit que L a des variations limitées, sa première dérivée $L^{(1)}$ est intégrable, et par conséquent, $K(.)L^{(1)}(.)$ est intégrable. En outre, $\left[K(.)L^{(1)}(.)\right]^2$ est également intégrable, ce qui garantit l'existence du terme de variance asymptotique. Voici un exemple de noyau $K$ qui satisfait la condition (**A4**) :

$$K(x) = \frac{1}{2\sqrt{2\pi}}(2 - x)e^{-\frac{x^2}{2}}$$

Notre premier résultat est la convergence uniforme presque sûre avec un taux d'une densité de probabilité conditionnelle, comme indiqué dans la Proposition 3.3.1, et le deuxième résultat concerne la convergence uniforme presque sûre de l'estimateur du mode conditionnel, comme indiqué dans le Théorème 3.3.1.

## 3.3.a. La consistance

Afin d'analyser les propriétés asymptotiques de notre estimateur, nous introduisons l'estimateur artificiel basé sur la vraie valeur de l'échantillon.

$$\widetilde{f_n}(x,y) = \frac{\alpha}{nh^3}\sum_{i=1}^{n} G(Y_i)^{-1} K\left(\frac{x-X_i}{h}\right) L\left(\frac{y-Y_i}{h}\right) \qquad (\mathbf{3.5})$$

**Proposition 3.3.1** Supposons que les hypothèses $(\mathbf{A1})$- $(\mathbf{A5})$ sont vérifiées alors,

$$sup_{x\in\Omega} sup_{a_F\le y\le b_F} \left|\hat{f}_n\,(y/x) - f(y/x)\right| = O\left\{max\left(\sqrt{\frac{logn}{nh^2}}\,, h^2\right)\right\}$$

**Théorème 3.3.1** Sous les hypothèses de la Proposition 3.3.1, si la densité conditionnelle satisfait $sup_{x\in\Omega} f^{(2)}(\theta(x)/x) < 0$, on a

$$sup_{x\in\Omega}\left|\hat{\theta}_n(x) - \theta(x)\right| = O\left\{max\left(\left(\frac{logn}{nh_n^2}\right)^{\frac{1}{4}}, h\right)\right\}$$

## Remarque 3.3.2

L'hypothèse de négativité uniforme sur la dérivée seconde de la densité conditionnelle dans le Théorème **3.3.1** implique l'unicité uniforme du mode conditionnel, c'est-à-dire :

$$\forall\, \varepsilon > 0\ \ \exists\, \lambda > 0, \forall \eta : I \longrightarrow \mathbb{R}, sup_{x\in\Omega}|\theta(x) - \eta(x)| \ge \varepsilon$$

$$\Longrightarrow sup_{x\in\Omega}|f(\theta(x)/x) - f(\eta(x)/x)\,| \ge \lambda$$

## 3.3.b. La normalité asymptotique

Supposons maintenant que la fonction de densité $f(y/x)$ est uni-modale en $\theta(x)$. Alors, par hypothèse $(\mathbf{A1})$ on a $f^{(1)}(\theta(x)/x) = 0$ et nous supposons que $f^{(2)}(\theta(x)/x) < 0$.

De même, nous avons

$$f_n^{(1)}(\theta_n(x)/x) = 0 \text{ et } \hat{f}_n^{(2)}(\theta(x)/x) < 0.$$

Si $\hat{\theta}_n(x)$ est le mode de $\hat{f}_n^{(1)}(./x)$.

Nous obtenons à l'aide d'un développement de Taylor au voisinage de $\theta(x)$

$$\hat{f}_n^{(1)}\big(\hat{\theta}_n(x)/x\big) = \hat{f}_n^{(1)}(\theta(x)/x) + \Big(\hat{\theta}_n(x) - \theta(x)\Big)\hat{f}_n^{(2)}(\bar{\theta}_n(x)/x) = 0$$

où $\bar{\theta}_n(x)$ est entre $\hat{\theta}_n(x)$ et $\theta(x)$ et en utilisant **(3.5)** on peut écrire

$$\hat{\theta}_n(x) - \theta(x) = -\frac{\hat{f}_n^{(0,1)}\big(x,\theta(x)\big)}{\hat{f}_n^{(0,2)}\Big(x,\bar{\theta}_n(x)\Big)} \qquad (\mathbf{3.6})$$

Nous montrons que le numérateur dans **(3.6)**, est asymptotiquement normalement distribué et que le dénominateur converge en probabilité vers $f^{(2)}(x,\theta(x))$. Afin de prouver la normalité asymptotique de $\hat{\theta}_n(x)$ Le théorème suivant énonce le résultat.

**Théorème 3.3.2** Supposons que les hypothèses (**A2**), (**A4**), (**A5**), (**A7**) et (**A8**) sont vérifiées. Alors, on a

$$\big(nh^4\big)^{\frac{1}{2}}\Big(\hat{\theta}_n(x) - \theta(x)\Big) \xrightarrow{L} N(0,\sigma^2(x))$$

$avec \xrightarrow{L} indique\ la\ convergence\ en\ loi,$

$$\sigma^2(x) = \alpha^2 f\big(x,\theta(x)\big)\int_{\mathbb{R}^2}\frac{K^2(r)}{G(r)}\big[L^{(1)}(s)\big]^2 drds$$

**Théorème 3.3.3** Sous réserve d'hypothèses (**A4**), (**A5**), (**A6**) et (**A8**) quand $n \longrightarrow \infty$ on a,

$$E\big(\hat{f}_n\,(y/x)\big) = f(y/x) + \frac{h^2}{2}\mu_2(K)\frac{\partial^2}{\partial x^2}f(y/x) + \frac{h^2}{2}\mu_2(L)\frac{\partial^2}{\partial y^2}f(y/x) + O(h^2) \qquad (\mathbf{3.7})$$

$$Var\big(\hat{f}_n\,(y/x)\big) = \frac{\alpha G(y)^{-1}f(y/x)R(K)R(L)}{nh^2 f(x)} + O\left(\frac{1}{nh^2}\right) \qquad (\mathbf{3.8})$$

Le théorème 3.**3.3** montre que la double troncature n'influence pas la variance et le biais. Nous nous intéresserons maintenant à l'erreur globale de $\hat{f}_n(y/x)$, celle-ci peut être mesuré par l'erreur quadratique moyenne MISE.

En ajoutant la variance **(3.8)** au biais quadratique **(3.7)** on obtient l'erreur quadratique moyenne asymptotique

$$MSE = \left[E\left(\hat{f}_n(y/x)\right) - f(y/x)\right]^2 + Var\left(\hat{f}_n(y/x)\right)$$

$$= \left[\frac{h^2}{2}\mu_2(K)\frac{\partial^2}{\partial x^2}f(y/x) + \frac{h^2}{2}\mu_2(L)\frac{\partial^2}{\partial y^2}f(y/x)\right]^2 + \frac{f(y/x)R(K)R(L)}{nh^3 f(x)} \quad (\mathbf{3.9})$$

L'erreur quadratique moyenne intégrée (MISE) est obtenue en prenant la double intégrale par rapport à x et y de l'erreur quadratique moyenne pondérée (MSE) formée par le produit de **(4.9)** avec $f(x)$, sous régularité, nous avons à partir des résultats précédents l'expression asymptotique suivante :

$$MISE = A_1 h^4 + A_2 h^4 - A_3 h^4 + \frac{1}{nh^3}A_4 \quad (\mathbf{3.10})$$

avec

$$A_1 = \frac{1}{4}\iint \mu_2(K)^2\left[\frac{\partial^2}{\partial x^2}f(y/x)\right]^2 f(x)dxdy;$$

$$A_2 = \frac{1}{4}\iint \mu_2(L)^2\left[\frac{\partial^2}{\partial y^2}f(y/x)\right]^2 f(x)dxdy;$$

$$A_3 = \iint 2\left(\frac{h^2}{2}\mu_2(K)\frac{\partial^2}{\partial x^2}f(y/x)\right)\left(\frac{h^2}{2}\mu_2(L)\frac{\partial^2}{\partial y^2}f(y/x)\right)f(x)dxdy;$$

$$A_4 = \iint \frac{f(y/x)R(K)R(L)}{nh^3 f(x)}f(x)dxdy;$$

En différenciant l'expression **(3.10)** et en la fixant à zéro, on obtient la largeur de la fenêtre optimale asymptotique :

$$h_{MISE} = \left[\frac{f(y/x)R(K)R(L)}{f(x)\left(\mu_2(K)^2\left[\frac{\partial^2}{\partial x^2}f(y/x)\right]^2 + \mu_2(L)^2\left[\frac{\partial^2}{\partial y^2}f(y/x)\right]^2\right)}\right]^{1/7} n^{-1/7}.$$

## 3.4. Résultats auxiliaires et preuves

**Lemme 3.4.1** Sous les hypothèses (**A3**), (**A4**) et (**A5**) on a

$$sup_{x\in\Omega}\left|\hat{l}_n(x) - l(x)\right| = O\left(max\left(\sqrt{\frac{logn}{nh^2}}, h^2\right)\right)$$

***Preuve*** on définit $\hat{l}_n(x) = \frac{\alpha_n}{nh}\sum_{i=1}^{n}\frac{1}{G_n(X_i)}K\left(\frac{x-X_i}{h}\right)$

On a

$$sup_{x\in\Omega}\left|\hat{l}_n(x) - l(x)\right| \leq sup_{x\in\Omega}\left|\hat{l}_n(x) - E\left(\hat{l}_n(x)\right)\right| + sup_{x\in\Omega}\left|E\left(\hat{l}_n(x)\right) - l(x)\right|$$

En utilisant le développement de Taylor, nous avons sous les hypothèses $(\boldsymbol{A3}) - (\boldsymbol{A5})$

$$sup_{x\in\Omega}\left|E\left(\hat{l}_n(x)\right) - l(x)\right| = O(h^2)$$

De plus, l'hypothèse (**A1)** fournit une preuve similaire à celle de la **Proposition 3.3.1** ce qui donne

$$sup_{x\in\Omega}\left|\hat{l}_n(x) - E\left(\hat{l}_n(x)\right)\right| = O\left(\left(\frac{Logn}{nh^2}\right)^{1/2}\right) \quad p.s \quad n \longrightarrow \infty$$

Ce qui permet de conclure la preuve.

**Lemme 3.4.2** Sous les hypothèses (**A4** ) et (**A5**) on a

$$sup_{x\in\Omega}sup_{a_F\leq y\leq b_F}\left|\hat{f}_n(x,y) - f(x,y)\right| = O\left(h^2\right) \quad p.s \quad n \longrightarrow \infty$$

***Preuve***

On définit $V_n = sup_{x\in\Omega}sup_{a_F\leq y\leq b_F}\left|\hat{f}_n(x,y) - f(x,y)\right|$

On a,

$$\hat{f}_n(x,y) - f(x,y) = \frac{1}{h^2}\iint_{-\infty}^{+\infty} K\left(\frac{x-X_i}{h}\right)L\left(\frac{y-Y_i}{h}\right)\left[\hat{F}_n(du,dv) - F(du,dv\right]$$

En utilisant le théorème de Fubini, on obtient par intégration multiple par parties

$$V_n = sup_{x\in\Omega} sup_{a_F\le y\le b_F} \left| \frac{1}{h^2} \iint_{-\infty}^{+\infty} K\left(\frac{x-X_i}{h}\right) L\left(\frac{y-Y_i}{h}\right) \hat{F}_n(du,dv) \right.$$
$$\left. - \frac{1}{h^2} \iint_{-\infty}^{+\infty} K\left(\frac{x-u}{h}\right) L\left(\frac{y-u}{h}\right) F(du,dv) \right|$$
$$\le sup_{x\in\Omega} sup_{a_F\le y\le b_F} \frac{1}{h^2} \iint_{-\infty}^{+\infty} \left|\hat{F}_n(x,y) - F(x,y)\right| \left[ \iint_{-\infty}^{+\infty} |K(r)L(s)| drds \right]$$

En utilisant le résultat de Moreira et al. (2021) et les hypothèses (**A4**) and (**A5**) nous obtenons le résultat.

***Preuve de la ¨Proposition*** $\mathbf{3.3.1}$ il est simple de voir qu'en utilisant la décomposition classique

$$sup_{x\in\Omega} sup_{a_F\le y\le b_F} \left|\hat{f}_n\,(y/x) - f(y/x)\right|$$
$$\le \frac{1}{inf_{x\in\Omega}\hat{f}_n(x)} \left\{ sup_{x\in\Omega} sup_{a_F\le y\le b_F} \left|\hat{f}_n(x,y) - f(x,y)\right| \right.$$
$$\left. + sup_{x\in\Omega} sup_{a_F\le y\le b_F} \left|\hat{f}_n(x) - f(x)\right| \right\}$$

Une application du lemme 3.4.1 et du lemme 3.4.2 et les hypothèses (**A4**) –(**A5**) nous permet de conclure.

***Preuve du Théorème*** $\mathbf{3.3.1}$

On a

$$sup_{x\in\Omega} \left| f\left(\hat{\theta}_n(x)/x\right) - f(\theta(x)/x) \right|$$
$$\le sup_{x\in\Omega} \left| f\left(\hat{\theta}_n(x)/x\right) - \hat{f}_n\left(\hat{\theta}_n(x)/x\right) \right| + sup_{x\in\Omega} \left| \hat{f}_n\left(\hat{\theta}_n(x)/x\right) - f(\theta(x)/x) \right|$$
$$\le sup_{x\in\Omega} sup_{a_F\le y\le b_F} \left|\hat{f}_n\,(y/x) - f(y/x)\right|$$
$$+ sup_{x\in\Omega} \left| sup_{a_F\le y\le b_F} \hat{f}_n\,(y/x) - sup_{a_F\le y\le b_F} f(y/x) \right|$$
$$\le 2 sup_{x\in\Omega} sup_{a_F\le y\le b_F} \left|\hat{f}_n\,(y/x) - f(y/x)\right|$$

Le développement de Taylor $f(./x)$ au voisinage de $\theta(x)$ donne :

$$\left|f\left(\hat{\theta}_n(x)/x\right) - f(\theta(x)/x)\right| = \frac{1}{2}\left(\hat{\theta}_n(x) - \theta(x)\right)^2 f^{(2)}(\bar{\theta}(x)/x)$$

ou $\bar{\theta}(x)$ est entre $\hat{\theta}_n(x)$ et $\theta(x)$.

Ensuite, par **(A2)** on a

$$sup_{x\in\Omega}\left|\hat{\theta}_n(x) - \theta(x)\right| \leq 2\sqrt{\frac{sup_{x\in\Omega} sup_{a_F\leq y\leq b_F}\left|\hat{f}_n\,(y/x) - f(y/x)\right|}{f^{(2)}(\bar{\theta}(x)/x)}}$$

En vertu de la Proposition **3.3.1**, la preuve est complète.

***Preuve du Théorème* 3. 3. 2.**

A partir de **(3.6),** nous obtenons la décomposition suivante

$$\sqrt{nh^4}\left(\hat{\theta}_n(x) - \theta(x)\right) = \sqrt{nh^4}\frac{\hat{f}_n^{(0,1)}\left(x,\theta(x)\right) - \tilde{f}_n^{(0,1)}\left(x,\theta(x)\right)}{f_n^{(0,2)}\left(x,\bar{\theta}_n(x)\right)}$$

$$+\sqrt{nh^4}\frac{\tilde{f}_n^{(0,1)}\left(x,\theta(x)\right) - E\left[\tilde{f}_n^{(0,1)}\left(x,\theta(x)\right)\right]}{\hat{f}_n^{(0,(2)}\left(x,\bar{\theta}_n(x)\right)}$$

$$+\sqrt{nh^4}\frac{E\left[\tilde{f}_n^{(0,1)}\left(\theta(x)\right)\right]}{\hat{f}_n^{(0,2)}\left(x,\bar{\theta}_n(x)\right)}$$

$$= \frac{J_1 + J_2 + J_3}{\hat{f}_n^{(0,2)}\left(x,\bar{\theta}_n(x)\right)} \qquad (\mathbf{3.11})$$

Nous établissons que les numérateurs des termes $J_1$ et $J_3$ sont négligeables, et celle de $J_2$ est normalement distribuée.

Outre la probabilité que le dénominateur finisse par converger en probabilité vers la valeur $f^{(0,2)}\left(x,\theta(x)\right)$.

Pour le premier terme $J_1$ , on a

$$\sqrt{nh^4}\ \left(\hat{f}_n^{(0,1)}\big(x,\theta(x)\big)-\tilde{f}_n^{(0,1)}\big(x,\theta(x)\big)\right)$$
$$\leq \frac{\sqrt{nh^4}}{nh^3}\, sup_{x\in\Omega}\left|\frac{\alpha_n}{G_n(x)}-\frac{\alpha}{G(x)}\right|\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)L^{(1)}\left(\frac{\theta(x)-Y_i}{h}\right)$$

Gardez à l'esprit que $sup_{x\in\Omega}\left|\frac{\alpha_n}{G_n(x)}-\frac{\alpha}{G(x)}\right| = O\left\{\frac{1}{\sqrt{nh_n}}\right\}\ a.s$ (Moreira, de Uña-Álvarez2012).

En utilisant la méthode habituelle du noyau, en ajoutant et en soustrayant l'espérance de

$$(nh^3)^{-1}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)L^{(1)}\left(\frac{\theta(x)-Y_i}{h}\right)$$

Nous pouvons prouver que ce terme converge vers la fonction limitée $f^{(0,1)}\big(x,\theta(x)\big)$ et conclure que $J_1$ est négligeable.

Nous passons maintenant à $J_3$, qui est défini comme suit.

**Lemme 3.4. 3** Sous réserve d'hypothèses (**A1**), (**A4**), (**A5**) et (**A7**), on a

$$J_3 \longrightarrow 0 \quad a.s.\ qd\ n\longrightarrow\infty$$

***Preuve.*** De

$$\tilde{f}_n^{(0,1)}\big(x,\theta(x)\big)=\frac{\alpha}{nh^3}\sum_{i=1}^{n} G(Y_i)^{-1}\,K\left(\frac{x-X_i}{h}\right)L^{(1)}\left(\frac{y-Y_i}{h}\right)$$

nous obtenons

$$J_3 = E\left[\tilde{f}_n^{(0,1)}\big(x,\theta(x)\big)\right]=\frac{\alpha}{nh^3}\int_{\mathbb{R}^2}\frac{1}{G(u)}K\left(\frac{x-u}{h}\right)L^{(1)}\left(\frac{\theta(x)-v}{h}\right)f(u,v)dudv$$

$$=\frac{\sqrt{nh^4}}{nh^3}\int_{\mathbb{R}^2}K\left(\frac{x-u}{h}\right)H^{(1)}\left(\frac{\theta(x)-v}{h}\right)f(u,v)dudv$$

$$=\frac{1}{\sqrt{nh^2}}\int_{\mathbb{R}^2}K\left(\frac{x-u}{h}\right)H^{(1)}\left(\frac{\theta(x)-v}{h}\right)f(u,v)dudv$$

Intégration par parties et utilisation d'hypothèses **(A1)** et (**A4)** et en utilisant le développement de Taylor de $f^{(0,1)}(x - rh_n, \theta(x) - sh_n)$ autours de $\big(x, \theta(x)\big)$ à l'ordre de $h^3$, nous obtenons

$$J_3 = O(nh^7).$$

**Lemme 3.4.4** Sous réserve d'hypothèses **(A1), (A4)** et **(A5)**

$$Var(J_2) \longrightarrow \alpha^2 f\big(x, \theta(x)\big) \int_{\mathbb{R}^2} \frac{K^2(r)}{G^2(r)} \big(L^{(1)}(s)\big)^2 drds \qquad qd \qquad n \longrightarrow \infty$$

***Preuve.***

$$Var(J_2) = Var\left(\frac{\alpha}{nh^2} \sum_{i=1}^{n} G^{-1}(Y_i) K\left(\frac{x - X_i}{h}\right) L^{(1)}\left(\frac{\theta(x) - Y_i}{h}\right)\right)$$

$$= \frac{\alpha^2}{nh^2} E\left\{\frac{1}{G^2(X_1)} K^2\left(\frac{x - X_1}{h}\right) \big(L^{(1)}\big)^2 \left(\frac{\theta(x) - Y_1}{h}\right)\right\}^2$$

$$- \frac{\alpha^2}{nh^2} E^2\left\{\frac{1}{G(X_1)} K\left(\frac{x - X_1}{h}\right) L^{(1)}\left(\frac{\theta(x) - Y_1}{h}\right)\right\}$$

$$= U_{1n} + U_{2n}$$

$$U_{1n} = \frac{\alpha^2}{nh^2} \iint \frac{K^2(r)}{G^2(r)} \big(L^{(1)}\big)^2 f(x - rh,\ \theta(x) - sh) drds$$

$$= \frac{\alpha^2}{nh^2} f\big(x, \theta(x)\big) \iint \frac{K^2(r)}{G^2(r)} \big(L^{(1)}\big)^2 (s) drds + o(1) \qquad as \quad n \longrightarrow \infty$$

De plus, par le lemme 3.4.3

$$U_{2n} = J_3^2 \longrightarrow 0 \quad as \ n \longrightarrow \infty.$$

D'une manière ou d'une autre, cela permet de conclure.

Nous nous concentrons maintenant sur le dénominateur dans (**3.11**). Quand $\bar{\theta}_n(x)$ converge presque sûr vers $\theta(x)$; sa consistance sera établie si l'on prouve

**Lemme 3. 4. 5** Sous réserve d'hypothèses (**A4**) , ( **A5**), **(A7) et (A8)**

$$sup_{a_F\le y\le b_F}\left|\hat{f}_n^{(0,2)}(x,y)-f^{(0,2)}(x,y)\right|\longrightarrow 0 \quad p.s. \quad qd \quad n\longrightarrow\infty.$$

***Preuve.*** On a

$$\left|\hat{f}_n^{(0,2)}(x,y)-f^{(0,2)}(x,y)\right|\le\left|\hat{f}_n^{(0,2)}(x,y)-\tilde{f}_n^{(0,2)}(x,y)\right|+\left|\tilde{f}_n^{(0,2)}(x,y)-f^{(0,2)}(x,y)\right|$$

$$=\gamma_{1,n}(x,y)+\gamma_{2,n}(x,y)$$

Pour $\gamma_{1,n}(x,y)$ on a :

$$sup_{a_F\le y\le b_F}\left|\hat{f}_n^{(0,2)}(x,y)-\check{f}_n^{(0,2)}(x,y)\right|$$

$$\le sup_{a_F\le y\le b_F}\frac{1}{nh^4}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)L^{(2)}\left(\frac{y-Y_i}{h}\right)\left|\frac{\alpha_n}{G_n(y)}-\frac{\alpha}{G(y)}\right|$$

$$\le\frac{1}{nh^4}sup_{a_F\le y\le b_F}\left|\frac{\alpha_n}{G_n(y)}-\frac{\alpha}{G(y)}\right|$$

Comme indiqué dans (Moreira et de Uña-Álvarez, 2012), $\left|\frac{\alpha_n}{G_n(y)}-\frac{\alpha}{G(y)}\right|\longrightarrow 0 \ qd \ n\longrightarrow\infty$, sous les hypothèses les (**A4**) et (**A5**) on obtient $\gamma_{1,n}(x,y)\longrightarrow 0$.

Pour $\gamma_{2,n}(x,y)$ :

$$\left|\tilde{f}_n^{(0,2)}(x,y)-f^{(0,2)}(x,y)\right|=\frac{\alpha}{nh^4}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)L^{(2)}\left(\frac{y-Y_i}{h}\right)-f^{(0,2)}(x,y)$$

en intégrant par parties deux fois par rapport à la deuxième composante et en utilisant un changement de variable, il s'ensuit que

$$\gamma_{2,n}(x,y)=\int_{\mathbb{R}^2}K\left(\frac{x-u}{h}\right)L^{(2)}\left(\frac{y-v}{h}\right)f(u,v)dudv-f^{(0,2)}(x,y)$$

$$=\int_{\mathbb{R}^2}K(r)L(s)\left\{f^{(0,2)}(x-rh,y-sh)-f^{(0,2)}(x,y)\right\}drds$$

En utilisant le développement de Taylor au voisinage de $(x,y)$ on obtient

$$\left|\gamma_{2,n}(x,y)\right|\le\left|K(r)L(s)\left\{rf^{(1,2)}(\overline{x},\overline{y})+sf^{(0,3)}(\overline{x},\overline{y})\right\}\right|drds,$$

$(\overline{x}, \overline{y})$ est entre $(x, y)$ et $(x - rh, y - sh)$ , avec les hypothèses A1 et (**A8)**, il s'ensuit que

$\gamma_{2,n}(x, y) \longrightarrow 0$.

La dernière étape de la démonstration du théorème **3.4.2** consiste à montrer la condition de Berry-Esséen pour $J_2$. Ainsi, compte tenu de $(\mathbf{3.11})$

On pose

$$J_2 = \sum_{i=1}^{n} \Gamma_{i,n}(x, y)$$

où

$$\Gamma_{i,n}(x, y) = \frac{1}{nh^2}\left\{\tilde{f}_n^{(0,1)}\big(x, \theta(x)\big) - E\left[\tilde{f}_n^{(0,1)}\big(x, \theta(x)\big)\right]\right\}$$

$$= \frac{\alpha}{nh^2}\sqrt{nh^4}\left\{G^{-1}(Y_i)K\left(\frac{x - X_i}{h}\right)H^{(1)}\left(\frac{\theta(x) - Y_i}{h}\right)\right.$$

$$\left. - E\left[G^{-1}(Y_1)K\left(\frac{x - X_i}{h}\right)H^{(1)}\left(\frac{\theta(x) - Y_i}{h}\right)\right]\right\} =$$

$$= \frac{\alpha}{nh^2}\left\{G^{-1}(Y_i)K\left(\frac{x - X_i}{h}\right)H^{(1)}\left(\frac{\theta(x) - Y_i}{h}\right)\right.$$

$$\left. - E\left[G^{-1}(Y_1)K\left(\frac{x - X_i}{h}\right)H^{(1)}\left(\frac{\theta(x) - Y_i}{h}\right)\right]\right\} \qquad (\mathbf{3.12})$$

Sous les hypothèses $(\boldsymbol{A_1})$, $(\boldsymbol{A_4})$, $(\boldsymbol{A_5})$ $\boldsymbol{et}$ $(\boldsymbol{A_7})$ nous montrerons que

$$\sum_{i=1}^{n} E\left(\left|\Gamma_{i,n}(x, y)\right|^3\right) < \infty.$$

Application de la $C_r$ −inequality (voir Loève, 2017), on a,

$$E\left(\left|\Gamma_{i,n}(x, y)\right|^3\right) \leq \frac{2^2}{(nh^2)^3} E\left[\left|\alpha G^{-1}(Y_i)K\left(\frac{x - X_1}{h}\right)H^{(1)}\left(\frac{\theta(x) - Y_1}{h}\right)\right|^3\right]$$

$$+ \frac{2^2}{(nh^2)^3} E\left\{\left|E\left[\alpha G^{-1}(Y_1)K\left(\frac{x - X_1}{h}\right)H^{(1)}\left(\frac{\theta(x) - Y_1}{h}\right)\right]\right|^3\right\}. \qquad (\mathbf{3.13})$$

Les deux termes espérances en **(3. 13)** sont limités en vertu de (**A1**), (**A4**) and (**A7**) , on a

$\sum_{i=1}^{n} E\left(\left|\Gamma_{i,n}(x, y)\right|^3\right) = o(1)$. Ceci complète la preuve du **Théorème 3.3.2.**

***Preuve du Théorème* 3.3.3** en appliquant le lemme 2 dans Hyndman et al. (1996)

$$E\left(\hat{f}_n\,(y/x)\right) = E\left[\frac{\frac{\alpha_n}{nh^2}\sum_{i=1}^{n} G_n(Y_i)^{-1}\,K\left(\frac{x-X_i}{h}\right)L\left(\frac{y-Y_i}{h}\right)}{\frac{\alpha_n}{nh}\sum_{i=1}^{n} G_n(Y_i)^{-1}K\left(\frac{x-X_i}{h}\right)}\right]$$

On peut l'estimer par

$$\frac{E\left[\frac{\alpha_n}{nh^2}\sum_{i=1}^{n} G_n(Y_i)^{-1}\,K\left(\frac{x-X_i}{h}\right)L\left(\frac{y-Y_i}{h}\right)\right]}{E\left[\frac{\alpha_n}{nh}\sum_{i=1}^{n} G_n(Y_i)^{-1}K\left(\frac{x-X_i}{h}\right)\right]}$$

Le moment du dénominateur se compose de

$$E\left[\frac{\alpha_n}{nh}\sum_{i=1}^{n} G_n(Y_i)^{-1}K\left(\frac{x-X_i}{h}\right)\right] = \frac{\alpha}{n}\,G_n(Y_i)^{-1}E\left[\frac{1}{h}K\left(\frac{x-X_i}{h}\right)\right] =$$

$$= \frac{\alpha}{n}\,G_n(Y_i)^{-1}\left[f(x) + \frac{h^2}{2}\mu_2(K)f''(x) + o(h^2)\right]$$

Le numérateur est le suivant

$$E\left[\frac{\alpha_n}{nh^2}\sum_{i=1}^{n} G_n(Y_i)^{-1}\,K\left(\frac{x-X_i}{h}\right)L\left(\frac{y-Y_i}{h}\right)\right]$$

$$= \frac{\alpha}{nh^2}\,G_n(Y_i)^{-1}K\left(\frac{x-X_1}{h}\right)\left[f(y/x) + \frac{h^2}{2}\mu_2(L)\frac{\partial^2}{\partial y^2}f(y/x) + O(h^2)\right]$$

En appliquant une seconde fois la formule du moment, on obtient

$$\frac{\alpha}{nh^2}\,G_n(Y_i)^{-1}E\left[K\left(\frac{x-X_1}{h}\right)\left[f(y/x) + \frac{h^2}{2}\mu_2(L)\frac{\partial^2}{\partial y^2}f(y/x) + O(h^2)\right]\right]$$

$$= \frac{\alpha}{nh^2}\,G_n(Y_i)^{-1}E\left[K\left(\frac{x-X_1}{h}\right)Q(x,y)\right]$$

$$\frac{\alpha}{nh^2}\,G_n(Y_i)^{-1}E\left[K\left(\frac{x-X_1}{h}\right)Q(x,y)\right]$$

$$= \frac{\alpha}{nh^2}\,G_n(Y_i)^{-1}\left[f(x)Q(x,y) + \frac{h^2}{2}\mu_2(K)\frac{\partial^2}{\partial x^2}\big(f(x)Q(x,y)\big) + o(h^4)\right]$$

$$= \frac{\alpha}{nh^2} G_n(Y_i)^{-1} f(x) \left[ f(y/x) + \frac{h^2}{2} \mu_2(L) \frac{\partial^2}{\partial y^2} f(y/x) \right] + \frac{h^2}{2} \mu_2(K) f(x) \frac{\partial^2}{\partial x^2} f(y/x) + o(h^4)$$

Utilisation du résultat $\frac{1}{\delta+s} = \frac{1}{s} + \frac{\delta}{s^2} + o(\delta^2)$

Nous obtenons

$$E\left(\hat{f}_n\,(y/x)\right) = f(y/x) + \frac{h^2}{2} \mu_2(K) \frac{\partial^2}{\partial x^2} f(y/x) + \frac{h^2}{2} \mu_2(L) \frac{\partial^2}{\partial y^2} f(y/x) + O(h^2)$$

Hyndman et al. (1996) permet d'approximer la variance de $\hat{f}_n\,(y/x)$ par

$$Var\left(\hat{f}_n\,(y/x)\right) = \frac{Var\left[\frac{\alpha_n}{nh^2} \sum_{i=1}^{n} G_n(Y_i)^{-1} K\left(\frac{x - X_i}{h}\right) L\left(\frac{y - Y_i}{h}\right)\right]}{E^2\left[\frac{\alpha_n}{nh} \sum_{i=1}^{n} G_n(Y_i)^{-1} K\left(\frac{x - X_i}{h}\right)\right]}$$

Le numérateur donne

$$Var\left[\frac{\alpha_n}{nh^2} \sum_{i=1}^{n} G_n(Y_i)^{-1} K\left(\frac{x - X_i}{h}\right) L\left(\frac{y - Y_i}{h}\right)\right] = \frac{\alpha^2 G_n(Y_i)^{-2} f(x) f(y/x) R(K) R(L)}{n^2 h^3}$$

L'approximation du dénominateur donne

$$E^2\left[\frac{\alpha_n}{nh} \sum_{i=1}^{n} G_n(Y_i)^{-1} K\left(\frac{x - X_i}{h}\right)\right] = \frac{\alpha^2 G_n(Y_i)^{-2} f^2(x)}{n} + O(h^2)$$

Alors,

$$Var\left(\hat{f}_n\,(y/x)\right) = \frac{f(y/x) R(K) R(L)}{nh^3 f(x)} + O\left(\frac{1}{nh^2}\right).$$

Ce qui met fin à la preuve du Théorème.

# Conclusion et Perspectives

L'objectif principal de cette thèse était de proposer un nouvel estimateur de la fonction mode simple et conditionnel, et qui sont observée lors d'une double troncature.

Nous avons présenté deux modèles simulés (cas de décroissance exponentielle et cas de queue lourde) et étudié le comportement sur un échantillon fini. Il a été montré qu'en général, la normalité asymptotique est fortement conservée, qui est connue pour être une question importante mais délicate.

Le présent document ne traite pas des procédures de sélection pour le choix du paramètre de lissage. Il s'agit d'un sujet susceptible de faire l'objet d'études futures. De même, les cas de petits échantillons reposent sur l'utilisation d'estimateurs robustes, ce qui n'est pas le cas de la présente étude, et pourrait être envisagé dans nos travaux futurs. La littérature contient également des approches semi-paramétriques pour estimer la fonction de densité en cas de double troncature, voir par exemple Moreira et de Uña-Álvarez (2012). Ainsi, dans nos recherches ultérieures, nous étudierons le problème de l'estimation du mode lorsque la distribution est supposée appartenir à une famille paramétrique donnée. On peut également envisager d'étendre cette étude au cas de données alpha-mélangeante.

Evidemment, il reste beaucoup de questions sans réponse. D'autres questions ouvertes peuvent être abordées suite à notre travail, comme par exemple, faire une étude comparative avec les travaux sur le mode et mode conditionnel, qui permet de choisir le meilleur prédicteur.

# Bibliographie


[1] Abraham, C. E. (2004). On the asymptotic properties of a simple estimate of the mode. *ESAIM Probab. Statist*, 8-11.

[2] Andersen, P. B. (1993). *Statistical Models Based on Counting Processes.* New york: Springer Verlag.

[3] Austin, M. D., Simon, D. K., Betensky, R. A. (2014). Computationally simple estimation and improved efficiency for special cases of double truncation. *Lifetime data analysis*, *20*, 335-354.

[4] Bickel, D. R., Frühwirth, R. (2006). On a fast, robust estimator of the mode: Comparisons to other robust estimators with applications. *Computational Statistics & Data Analysis*, *50*(12), 3500-3530.

[5] Bilker, W. B., Wang, M. C. (1996). A semiparametric extension of the Mann-Whitney test for randomly truncated data. *Biometrics*, 10-20.

[6] Chaib, Y., Boutabia, H., & Sadki, O. (2013). A nonparametric mode estimate under LTRC model and dependent data. *South African Statistical Journal*, *47*(2), 91-109.

[7] Chaieb, L. L., Rivest, L. P., Abdous, B. (2006). Estimating survival under a dependent truncation. *Biometrika*, *93*(3), 655-669.

[8] Chernoff, H. (1964). Estimation of the mode. *Annals of the Institute of Statistical Mathematics*, *16*(1), 31-41.

[9] Cheng Wang, M. (1987). Product limit estimates: a generalized maximum likelihood study. *Communications in Statistics-Theory and Methods*, *16*(11), 3117-3132.

[10] Chow, Y. S., Teicher, H. (1997). *Encyclopaedia of Mathematical Sciences: Independence, Interchangeability, Martingales. Probability Theory*. Springer.

[11] Collomb, G., Härdle, W., Hassani, S. (1986). A note on prediction via estimation of the conditional mode function. *Journal of Statistical Planning and Inference*, *15*, 227-236.

[12] Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187-202.

[13] Dabo-Niang, S., Ferraty, F., & Vieu, P. (2004). Estimation du mode dans un espace vectoriel semi-normé. *Comptes Rendus Mathematique*, *339*(9), 659-662.

[14] Dörre, A., Emura, T. (2019). *Analysis of doubly truncated data: an introduction*. Singapore: Springer Singapore.

[15] Dörre, A. (2020). Bayesian estimation of a lifetime distribution under double truncation caused by time-restricted data collection. *Statistical Papers*, *61*(3), 945-965.

[16] Eddy, W. F. (1982). The asymptotic distributions of kernel estimators of the mode. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, *59*(3), 279-290.

[17] Efron, B., Petrosian, V. (1999). Nonparametric methods for doubly truncated data. *Journal of the American Statistical Association*, *94*(447), 824-834.

[18] Emura, T., Konno, Y. (2012a). A goodness-of-fit test for parametric models based on dependently truncated data. *Computational Statistics & Data Analysis*, *56*(7), 2237-2250.

[19] Emura, T., Konno, Y. (2012b). Multivariate normal distribution approaches for dependently truncated data. *Statistical Papers*, *53*, 133-149.

[20] Emura, T., Konno, Y., Michimae, H. (2015). Statistical inference based on the nonparametric maximum likelihood estimator under double-truncation. *Lifetime Data Analysis*, *21*, 397-418.

[21] Emura, T., Wang, W. (2016). Semiparametric inference for an accelerated failure time model with dependent truncation. *Annals of the Institute of Statistical Mathematics*, *68*, 1073-1094.

[22] Ezzahrioui, M., Ould Saïd, E. (2005). *Asymptotic normality of nonparametric estimators of the conditional mode for functional data* (No. 249). Technical report. *Univ. du Littoral Cote d'Opale.*

[23] Ezzahrioui, M. (2010). On the asymptotic properties of a nonparametric estimator of the conditional mode for functional dependent data. *Statistica Neerlandika*, *64*, 171-201.

[24] Frank, G., Chae, M., Kim, Y. (2019). Additive time-dependent hazard model with doubly truncated data. *Journal of the Korean Statistical Society*, *48*(2), 179-193.

[25] Gannoun, A., Saracco, J. (2002). A new proof of strong consistency of kernel estimation of density function and mode under random censorship. *Statistics & probability letters*, *59*(1), 61-66.

[26] Giné, E., Guillou, A. (1999). Laws of the iterated logarithm for censored data. *The Annals of Probability*, *27*(4), 2042-2067.

[27] Giné, E., Guillou, A. (2001). On consistency of kernel density estimators for randomly censored data: rates holding uniformly over adaptive intervals. In *Annales de l'IHP Probabilités et statistiques* (Vol. 37, No. 4, pp. 503-522).

[28] Gross, S. T., Lai, T. L. (1996). Nonparametric estimation and regression analysis with left-truncated and right-censored data. *Journal of the American Statistical Association*, *91*(435), 1166-1180.

[29] Hyndman, R. J., Bashtannyk, D. M., Grunwald, G. K. (1996). Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics*, *5*(4), 315-336.

[30] Khardani, S., Lemdani, M., Ould Saïd, E. (2010). Some asymptotic properties for a smooth kernel estimator of the conditional mode under random censorship. *Journal of the Korean Statistical Society*, *39*(4), 455-469.

[31] Khardani, S., Lemdani, M., Ould Saïd, E. (2012). On the strong uniform consistency of the mode estimator for censored time series. *Metrika*, *75*(2), 229-241.

[32] Loève, M. *(*1977*). Probability Theory.* New York : Springer-Verlag.

[33] Loève, M. (2017). *Probability theory*. Courier Dover Publications.

[34] Louani, D. (1998). On the asymptotic normality of the kernel estimators of the density function and its derivatives under censoring. *Communications in Statistics-Theory and Methods*, *27*(12), 2909-2924.

[35] Lynden-Bell, D. (1971). A method of allowing for known observational selection in small samples applied to 3CR quasars. *Monthly Notices of the Royal Astronomical Society*, *155*(1), 95-118.

[36] Mandel, M., de Uña-Álvarez, J., Simon, D. K., Betensky, R. A. (2018). Inverse probability weighted Cox regression for doubly truncated data. *Biometrics*, *74*(2), 481-487.

[37] Mokkadem, A., Pelletier, M. (2005). Moderate deviations for the kernel mode estimator and some applications. *Journal of statistical planning and inference*, *135*(2), 276-299.

[38] Moreira, C., de Uña-Álvarez, J. (2010a). Bootstrapping the NPMLE for doubly truncated data. *Journal of Nonparametric Statistics*, *22*(5), 567-583.

[39] Moreira, C., de Uña-Álvarez, J. (2010b). A semiparametric estimator of survival for doubly truncated data. *Statistics in Medicine*, *29*(30), 3147-3159.

[40] Moreira, C., de Uña-Álvarez, J., Crujeiras, R. M. (2010). DTDA: An R package to analyze randomly truncated data. *Journal of Statistical Software*, *37*, 1-20.

[41] Moreira, C., de Uña-Álvarez, J. (2012). Kernel density estimation with doubly truncated data. *Electronic Journal of Statistics*, *6*, 501-521.

[42] Moreira, C., Van Keilegom, I. (2013). Bandwidth selection for kernel density estimation with doubly truncated data. *Computational Statistics & Data Analysis*, *61*, 107-123.

[43] Moreira, C., de Uña-Álvarez, J., Meira-Machado, L. (2016). Nonparametric regression with doubly truncated data. *Computational Statistics & Data Analysis*, *93*, 294-307.

[44] Moreira, C., Uña-Álvarez, J. D., Braekers, R. (2021). Nonparametric estimation of a distribution function from doubly truncated data under dependence: Double truncation under dependence. *Computational Statistics*, *36*(3), 1693-1720.

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

[46] Ould-Saïd, E. (1993). Estimation non paramétrique du mode conditionnel. Application à la prévision. *Comptes rendus de l'Académie des sciences. Série 1, Mathématique*, *316*(9), 943-947.

[47] Ould-Saïd, E., Cai, Z. (2005). Strong uniform consistency of nonparametric estimation of the censored conditional mode function. *Nonparametric Statistics*, *17*(7), 797-806.

[48] Ould-Saïd, E., Tatachak, A. (2007). Asymptotic properties of the kernel estimator of the conditional mode for the left truncated model. *Comptes Rendus Mathematique*, *344*(10), 651-656.

[49] Ould-Said, E., Tatachak, A. (2009). On the non-parametric estimation of the simple mode under random lefttruncation model. *Rev. Roum. Math. Pure. A*, *54*(3), 243-266.

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

[51] Rennert, L., Xie, S. X. (2018). Cox regression model with doubly truncated data. *Biometrics*, *74*(2), 725-733.

[52] Romano, J. P. (1988). On weak convergence and optimality of kernel density estimates of the mode. *The Annals of Statistics*, 629-647.

[53] Rossi, V. (2004). *Filtrage non linéaire par noyaux de convolution: application à un procédé de dépollution biologique* (Doctoral dissertation, Montpellier, ENSA).

[54] Roussas, G. G. (1968). On some properties on nonparametric estimates of probability density functions. *Bull. Soc. Math. Grèce (N.S.), 9*(09A), 29-43.

[55] Samanta, M. (1973). Nonparametric estimation of the mode of a multivariate density. *South African Statistical Journal*, *7*(2), 109-117.

[56] Samanta, M., Thavaneswaran, A. (1990). Nonparametric estimation of the conditional mode. *Communications in Statistics-Theory and Methods*, *19*(12), 4515-4524.

[57] Shen, P. S. (2010a). Semiparametric analysis of doubly truncated data. *Communications in Statistics Theory and Methods*, *39*(17), 3178-3190.

[58] Shen, P. S. (2010b). Nonparametric analysis of doubly truncated data. *Annals of the Institute of Statistical Mathematics*, *62*(5), 835-853.

[59] Shen, P. S. (2011). Empirical likelihood ratio with doubly truncated data. *Journal of Applied Statistics*, *38*(10), 2345-2353.

[60] Shen, P. S. (2013). Regression analysis of interval censored and doubly truncated data with linear transformation models. *Computational Statistics*, *28*, 581-596.

[61] Shen, P. S. (2016). Analysis of transformation models with doubly truncated data. *Statistical Methodology*, *30*, 15-30.

[62] Shen, P. S. (2017). Semiparametric analysis of transformation models with dependently left-truncated and right-censored data. *Communications in Statistics-Simulation and Computation*, *46*(3), 2474-2487.

[63] Shen, P. S., Liu, Y. (2019a). Pseudo maximum likelihood estimation for the Cox model with doubly truncated data. *Statistical papers*, *60*, 1207-1224.

[64] Shen, P. S., Liu, Y. (2019b). Pseudo maximum likelihood estimation for the Cox model with doubly truncated data. *Statistical papers*, *60*, 1207-1224.

[65] Shen, P. S., Liu, Y. (2019d). Pseudo MLE for semiparametric transformation model with doubly truncated data. *Journal of the Korean Statistical Society*, *48*(3), 384-395.

[66] Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society: Series B (Methodological)*, *43*(1), 97-99.

[67] Silverman, B. W. (1986). *Density estimation for statistics and data analysis* (Vol. 26). London: Chapman and Hall. London.

[68] Støvring, H., & Wang, M. C. (2007). A new approach of nonparametric estimation of incidence and lifetime risk based on birth rates and incident events. *BMC medical research methodology*, 7(1), 1-11.

[69] Stute, W. (1982). A law of the logarithm for kernel density estimators. *The annals of Probability*, 414-422.

[70] Stute, W. (1993). Almost sure representations of the product-limit estimator for truncated data. *The Annals of Statistics*, 146-156.

[71] Talagrand, M. (1996). New concentration inequalities in product spaces. *Inventiones mathematicae*, *126*(3), 505-563.

[72] Van der Vaart, A. W. (2000). *Asymptotic statistics* (Vol. 3). Cambridge University Press.

[73] Van Ryzin, J. (1969). On strong consistency of density estimates. *The Annals of Mathematical Statistics*, *40*(5), 1765-1772.

[74] Vieu, P. (1996). A note on density mode estimation. *Statistics & probability letters*, *26*(4), 297-307.

[75] Wand, M. P., & Jones, M. C. (1995). Kernel Smoothing, London: Chapman&Hall.

[76] Wang, M. C. (1989). A semiparametric model for randomly truncated data. *Journal of the American Statistical Association*, *84*(407), 742-748.

[77] Wang, M. C. (1991). Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association*, *86*(413), 130-143.

[78] Wegman, E. J. (1971). A note on the estimation of the mode. *The Annals of Mathematical Statistics*, 1909-1915.

[79] Woodroofe, M. (1985). Estimating a distribution function with truncated data. *The Annals of Statistics*, *13*(1), 163-177.

[80] Youndje, E. (1993). *Estimation non paramétrique de la densité conditionnelle par la méthode du noyau* (Doctoral dissertation, Rouen).

[81] Zerfaoui, K., Zahnit, A., Yahia, D. (2023). A nonparametric mode estimate under doubly truncated model. *Journal of Science and Arts,* 23(1), 161-176.

[82] Zhu, H., Wang, M. C. (2012). Analysing bivariate survival data with interval sampling and application to cancer epidemiology. *Biometrika*, *99*(2), 345-361.

# Annexe A : Quelques outils de probabilités

Nous rappelons ici quelques définitions et théorèmes utilisés tout au long de cette thèse.

## A.1. Convergence en probabilité :

**Définition A.1 :**Soit $(X, X_n), n \geq 1$ une suite de variables aléatoires réelle, définies sur le même espace de probabilité $(\Omega, \mathcal{F}, P)$. La suite $(X_n)$ converge en probabilité vers $X$ si :

$$\forall \varepsilon > 0, \lim_{n \longrightarrow \infty} P(|X_n - X| \geq \varepsilon) = 0.$$

## A.2. Limite d'un produit

**Proposition A.1 :** Soient $X_n, Y_n$ deux suites de variables aléatoires réelle.

Si $X_n \longrightarrow X$ , $Y_n \longrightarrow Y$ en probabilité, alors $X_n . Y_n \longrightarrow X . Y$ en probabilité.

## A.3. Convergence presque sûre

**Définition A.2 :** La suite $(X_n)$ converge presque sûrement (p.s) vers $X$, si

$$P\left\{w : \lim_{n \longrightarrow \infty} X_n(w) = X(w)\right\} = 1$$

## A.4. Convergence en loi

**Définition A.3 :** La suite $(X_n)$ converge presque sûrement (p.s) vers $X$, si et seulement si, pour toute $\phi$ mesurable, bornée,

$$\lim_{n \longrightarrow \infty} E\big(\phi(X_n)\big) = E\big(\phi(X)\big).$$

## A.5. Théorème de la limite centrale

**Théorème A.1** : Soit $(\boldsymbol{X_n})$ un échantillon iid d'une loi de moyenne $\boldsymbol{m}$ et de variance $\boldsymbol{\sigma^2}$. Alors :

$$\sqrt{\boldsymbol{n}}\frac{\bar{\boldsymbol{X}}-\boldsymbol{m}}{\boldsymbol{\sigma}} = \frac{\boldsymbol{S_n}-\boldsymbol{nm}}{\boldsymbol{\sigma\sqrt{n}}} \overset{\boldsymbol{Loi}}{\longrightarrow} \boldsymbol{N(0,1)}.$$

## A.6. Inégalités exponentielles

**Théorème A.2** ***(Talagrand, 1996) :*** *Si* $\boldsymbol{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n}$ *sont n v.a.r.i.i.d et si* $\boldsymbol{\mathcal{H}}$ *est une VC-Classe mesurable et uniformément bornées de fonctions telle que*

$$\boldsymbol{U \geq \sup_{f \in \mathcal{F}} \|f\|_\infty} \text{ et } \boldsymbol{\sigma^2 \geq \sup_{f \in \mathcal{F}} Var(f(\varepsilon_1))}$$

où $\boldsymbol{\sigma}$ et $U$ sont des nombres réelles vérifiant $0 < \boldsymbol{\sigma} <$ U. *Alors, il existe des constantes* $C$ *et* $K_0$ *ne dépendant que des caractéristiques de A de la VC- Classe telle que :*

$$P\left\{\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} [f(\varepsilon_i)] - E\, f(\varepsilon_1) \right| > t \right\} \leq K_0 \, exp\left\{ -\frac{t}{K_0 U} Log\left(1 + \frac{tU}{K_0 V_n^2}\right)\right\}$$

$$\forall\, t \geq C\sqrt{Log\left(\frac{AU}{\sigma}\right)} V_n \quad ou \;\; V_n = \sigma\sqrt{n} + U\sqrt{Log\left(\frac{AU}{\sigma}\right)}.$$

**Lemme A.1 :** *(Borel-Cantelli)* Soit $\boldsymbol{A_n}$ une suite d'évènements Si :

$$\sum_{n=1}^{\infty} P(\boldsymbol{A_n}) < \infty \;\; alors \;\; P(\limsup \boldsymbol{A_n}) = 0.$$

On suppose maintenant que les évènements $\boldsymbol{A_n}$ sont indépendants. Si :

$$\sum_{n=1}^{\infty} P(\boldsymbol{A_n}) = +\infty \;\; alors \;\; P(\limsup \boldsymbol{A_n}) = 1.$$

# Résumé

Le phénomène de la double troncature simultanée à gauche et à droite apparaît dans divers domaines, tels que la recherche médicale et l'économie. Le problème de l'estimation de la fonction de mode ou du mode conditionnel pour ce type de données n'a pas été abordé dans la littérature statistique. Dans cette thèse, nous proposons un nouvel estimateur à noyau du mode dans le cadre d'un modèle aléatoire doublement tronqué. Nous établissons la forte consistance avec un taux de convergence pour l'estimation proposée et indiquons sa normalité asymptotique. Une étude de simulation est réalisée pour illustrer et évaluer le comportement sur un échantillon fini de l'estimateur proposé. L'estimation non paramétrique de la fonction du mode conditionnel pour les données sous double troncature est aussi étudier dans cette thèse.

# Abstract

The phenomenon of simultaneous left and right double truncation appears in various fields, such as medical research and economics. The problem of estimating the mode function or the conditional mode for this type of data has not been mentioned in the statistical literature. In this thesis, we propose a new kernel estimator of the mode in the framework of a doubly truncated random model. We establish the strong consistency with a convergence rate for the proposed estimate and indicate its asymptotic normality. A simulation study is performed to illustrate and evaluate the behavior on a finite sample of the proposed estimator. The nonparametric estimation of the conditional mode function for data under double truncation is also studied in this thesis.

# ملخص

تظهر ظاهرة الاقتطاع المزدوج الأيمن والأيسر المتزامن في مجالات مختلفة، مثل البحث الطبي والاقتصاد. لم يتم تناول مشكلة تقدير وظيفة المنوال أو المنوال الشرطي لهذا النوع من البيانات في الدراسات الإحصائية. في هذه الأطروحة، نقترح مقدر نواة جديدة للمنوال في إطار نموذج عشوائي مبتور بشكل مضاعف. نؤسس تناسقًا قويًا مع معدل تقارب للتقدير المقترح ونشير إلى طبيعته المقاربة. يتم إجراء دراسة محاكاة لتوضيح وتقييم السلوك على عينة محدودة من المقدر المقترح. كما تمت دراسة التقدير اللامعلمي لوظيفة المنوال الشرطي للبيانات تحت الاقتطاع المزدوج في هذه الأطروحة.