

RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE



UNIVERSITÉ FRÈRES MENTOURI CONSTANTINE 1
FACULTÉ DES SCIENCES EXACTES

=====

DÉPARTEMENT DE MATHÉMATIQUES

N° d'ordre : 34/ DS/2024

N° de série : 02/ Math /2024

THÈSE

présentée pour l'obtention du diplôme de

Doctorat en Sciences en Mathématiques

« Estimation non-paramétrique de la médiane et du mode conditionnels dans un modèle de censure »

présentée par

Hadjer Benchoulak

OPTION : Probabilités et Statistiques

Devant le jury :

Présidente :	M ^{me}	S. Belaloui	Prof	Université Frères Mentouri, Constantine 1.
Encadreur :	M.	M. Boukeloua	M.C.A	École Nationale Polytechnique de Constantine.
Examineur :	M.	Z. Mohdeb	Prof	Université Salah Boubnider, Constantine 3.
Examinatrice :	M ^{me}	S. Kharfouchi	Prof	Université Salah Boubnider, Constantine 3.
Examinatrice :	M ^{me}	I. Laroussi	M.C.A	Université Frères Mentouri, Constantine 1.
Examinatrice :	M ^{me}	K. A. Mezhoud	M.C.A	Université Frères Mentouri, Constantine 1.
	M ^{me}	F. Messaci	Prof	Invitée d'honneur.

Soutenue le : 01/07/2024.

Remerciements

Ce document est le fruit d'années de recherche, qui n'aurait pas été fructueux sans l'aide et le soutien de nombreuses personnes.

Je tiens à remercier particulièrement mes directeurs de thèse la professeure F. Messaci et le maître de conférence M. Boukloua pour leurs encadrements. Je suis très heureuse d'exprimer ici tous mes remerciements, mes profondes gratitude et reconnaissance à ma directrice de thèse principale depuis le démarrage de cette recherche jusqu'à son aboutissement et jusqu'à la fin décembre 2023, sans qui ce travail n'aurait pu voir le jour. Merci de m'avoir proposée ce sujet de thèse aussi intéressant. Merci d'avoir pris le temps de m'enseigner avec patience les nécessaires bases de la statistique mathématique et de m'avoir insufflée la rigueur mathématique et je la remercie une autre fois pour autant de gentillesse et de générosité.

Je remercie pour la deuxième fois mon co-cadreur Boukeloua Mohamed devenu encadreur de cette thèse depuis le début de l'année 2024 à cause de l'arrêt du travail de madame F. Messaci. Je le remercie pour sa confiance en m'offrant la possibilité de réaliser ce travail de recherche et toute l'aide qu'il m'a apportée, pour tout le temps qu'il m'a consacré et l'extraordinaire dévouement qu'il a dans son travail avec moi et avec les autres.

Je remercie madame la professeure S. Belaloui de me faire l'honneur de présider mon jury de soutenance de thèse.

Je tiens à remercier monsieur le professeur Mohdeb, madame la professeure Kharfouchi, madame Laroussi et madame Kenza Assia Mezhoud qui ont bien accepté d'examiner ce travail.

Je remercie grandement de manière générale tous les membres du Laboratoire de Mathématiques et Sciences de la Décision (LAMASD) de l'université frères Mentouri Constantine 1.

Dédicaces

Je dédie ce modeste travail à :

- Mes parents et mon mari Aouine Ahmed Chaouki pour leur soutien et d'être toujours à mes cotés pour m'encourager.
- Mes chers enfants : Hanine, Mohamed Ouail, Nouha Hibet El Rahmanne et Ahmed Ouassime.
- Mes frères et ma chère soeur décédée Karima.
- Toute ma famille sans exception.
- Mes amis sans exception.

Article

H. Benchoulak, M. Boukeloua et F. Messaci (2024). Asymptotic properties of a kernel estimate of the conditional mode in the presence of doubly censored data. *Communications in Statistics - Theory and Methods*, 53 (4) :1143-1161, 2024.

Communications

- H. Benchoulak, N. Nemouchi. Comparaison des bandes de confiance et des intervalles de confiance. .WPS2013. Biskra. janvier 2013 .
- H. Benchoulak, F. Messaci. Estimation non paramétrique de la médiane conditionnelle dans un modèle de censure .CIAA2014. Constantine. décembre 2014.
- H. Benchoulak, F. Messaci. Normalité asymptotique d'un estimateur non paramétrique du quantile conditionnel dans un modèle de censure mixte. CMA2016. Batna. Novembre 2016.
- H. Benchoulak, M. Boukeloua, F. Messaci .Strong consistency of a conditional mode estimator in the presence of doubly censored data. ICMA2021. Blida. décembre 2021.
- H. Benchoulak, M. Boukeloua, F. Messaci . International e-conference on pure and applied mathematical sciences. ICPAMS2022. Tunisie. Mai 2022.

Table des matières

Introduction	3
1 Données censurées et les classes V-C	7
1.1 Introduction aux données censurées	7
1.1.1 L'estimateur de Kaplan-Meier	12
1.1.2 Les estimateurs self-consistants	13
1.1.3 L'estimateur de Patilea et Rolin	14
1.2 Rappels sur les classes (V-C) et l'inégalité de Talagrand . . .	15
1.2.1 Les classes (V-C) d'ensembles	16
1.2.2 Les classes Vapnik-Červonenkis de fonctions	18
1.2.3 Inégalités de type Talagrand	20
2 Estimation à noyau du quantile Conditionnel	21
2.1 La convergence presque complète et l'inégalité de Bernstein	22
2.1.1 Définitions et propriétés	22
2.1.2 Inégalités de type Bernstein	24
2.2 Introduction de l'estimateur du quantile conditionnel	25
2.2.1 Dans le cas des données complètes	25
2.2.2 Dans le cas des données censurées à droite	27
2.3 Estimation du quantile conditionnel dans le modèle de cen- sure mixte	28
2.3.1 Le modèle	28
2.3.2 Estimation et hypothèses	29

2.3.3	Résultats et preuves	34
3	Les propriétés asymptotiques de l'estimateur à noyau du mode conditionnel dans la présence des données doublement censurées	41
3.1	Définition de l'estimateur	42
3.2	Hypothèses et principaux résultats	43
3.2.1	Convergence de $\theta_n(x)$	48
3.2.2	La normalité asymptotique de $\theta_n(x)$	56
4	Simulation et données réelles	64
4.1	Etude numérique I	65
4.2	Etude numérique II	68
4.2.1	La convergence	68
4.2.2	Normalité asymptotique	71
4.3	Application et données réelles	73
	Perspectives	75

Introduction

En 1632 Galilei (1632) a introduit pour la première fois en géométrie, l'idée principale d'ajustement d'un nuage de points par une droite qui a été interprétée par la suite comme un outil de prédiction.

En statistiques cette idée à été interprétée à expliquer une variable aléatoire à travers une autre. Autrement dit on cherche à étudier ou à trouver une relation entre deux variables aléatoires (v.a) réelles Y (variable réponse) et X (variable explicative / co-variable). Ce lien peut être modélisé par plusieurs manières citons par exemple la régression, la régression modale (le mode conditionnel) et le quantile conditionnel. Il existe plusieurs approches pour estimer ces modèles. Nous renvoyons le lecteur à Zhou et Liang (2003) qui ont utilisé l'approche à noyau de l'opérateur de régression pour estimer la médiane conditionnelle. Ils ont montré la normalité asymptotique de l'estimateur construit lorsque les observations sont α mélangeantes. Une généralisation de cet estimateur a été proposée par Gannoun *et al.* (2003). Ces derniers ont établi la convergence presque complète et la normalité asymptotique de ces estimateurs. Gannoun (1991) ont construit deux estimateurs non paramétriques de la médiane conditionnelle par les méthodes du médianogramme et du noyau pour des observations α -mélangeantes, pour faire de la prévision à différents horizons.

Demongeot *et al.* (2010) ont proposé un estimateur de la densité conditionnelle et ils ont établi la convergence ponctuelle et uniforme presque

complète ensuite ils ont utilisé ces résultats pour étudier les propriétés asymptotiques de l'estimateur du mode conditionnel par la méthode des polynômes locaux.

L'étude de modèles non paramétriques par la méthode du noyau liés au quantile conditionnelle et à la régression modale fait l'objet d'une littérature abondante. Les premiers résultats sur ces modèles ont été obtenus par Gannoun *et al.* (2002) qui ont proposé l'estimateur non paramétrique à noyau de la fonction répartition en suite dans le même article, ils ont proposé l'estimateur non paramétrique à double noyau de la fonction du quantile conditionnel .

Collomb *et al.* (1987)) ont montré la convergence uniforme de l'estimateur à noyau de la fonction du mode conditionnel lorsque les observations sont ϕ -mélangeantes.

Pour la dimension infinie, les premiers résultats sur la pévision par le mode conditionnel et les quantiles conditionnels ont été traités par Ferraty et Vieu (2006). Ils ont précisé la vitesse de convergence presque complète des estimateurs à noyau pour ces fonctions. D'autres auteurs ont travaillé dans ce cadre. Nous pouvons citer par exemple, Ezzahrioui et Ould Saï (2005), Dabo-Niang et Laksaci (2007), Ezzahrioui et Ould Saï (2008) et Messaci *et al.* (2015).

La modélisation et l'approche non paramétriques à noyau ne sont pas l'unique difficulté rencontrée en statistique, une autre complication souvent rencontré est l'existence de données incomplètes. Cette incomplétude est essentiellement due à deux phénomènes : la censure et la troncature. Lemdani *et al.* (2009) ont introduit un estimateur du quantile conditionnel pour des données aléatoirement tronquées à gauche et ils ont étudié sa convergence presque sûre avec vitesse de convergence ainsi que la normalité asymptotique. Dans le cas des données censurées à droite, Ould Saïd (2006) a introduit un estimateur du quantile conditionnel et il a étudié sa convergence presque sûre. Ce dernier en collaboration avec

Ould-Saïd et Sadki (2008a) ont proposé un estimateur non paramétrique à double noyau du quantile conditionnel et ils ont étudié ses propriétés asymptotiques. D'autres auteurs comme Khardani *et al.* (2010) et Ould Saïd et Cai (2005) ont traité les propriétés asymptotiques de l'estimateur d'un autre modèle de prévision qui est le mode conditionnel.

Bien que la censure à droite soit la plus courante dans la pratique, d'autres types de censure, généralisent cette censure. La littérature est beaucoup plus riche en censure double que de censure mixte qui est plus récente. Dans cette thèse, on va s'intéresser particulièrement aux deux types de censure, en considérant des modèles non paramétriques où la variable explicative est complètement observée et la variable réponse est censurée. Plus précisément, les points étudiés sont la prévision par le quantile conditionnel où la variable d'intérêt Y est soumise à une censure mixte et la prévision par le mode conditionnel où la variable d'intérêt Y est doublement censurée. Turnbull (1974) étudie ce genre de censure, il a proposé des estimateurs, pour les fonctions de survie des variables latentes, vérifiant des équations de self-consistance et plusieurs travaux sont basés sur ce modèle. Citons, sans prétendre à l'exhaustivité, Chang et Yang (1987), Chang (1990), Ren (1997) et Boukeloua et Messaci (2016). Pour la prévision aux données réelles du contexte de censure double, se référer à Lin *et al.* (2002) et Miller et Halpern (1982).

Dans le contexte de censure mixte, les premiers résultats ont été obtenus par Patilea et Rolin (2006) qui ont généralisé l'estimateur de Kaplan et Meier (1958) et ils ont établi sa convergence uniforme presque sûre ainsi que sa normalité asymptotique. Tandis que l'estimation par la méthode des poids a été utilisée pour la première fois en 2010 par Messaci (2010) qui a étudié un autre outil de la prévision, c'est la prévision par le modèle de la régression. Cette méthode d'estimation a connu un développement continu. En effet, Kebabi et Messaci (2012) ont donné une vitesse de convergence presque complète aussi bien ponctuelle qu'uniforme de l'estimateur à noyau étudié par Messaci (2010).

Aouicha et Messaci (2019) ont montré la convergence en moyenne quadratique avec taux de l'estimateur à noyau de la densité conditionnelle et pour l'estimateur à noyau de la densité (inconditionnelle) Boukeloua (2015) a montré sa convergence en moyenne quadratique. Boukeloua et Messaci (2016) ont établi sa normalité asymptotique. D'autres auteurs comme Messaci et Nemouchi (2011) et Messaci et Nemouchi (2013) ont exposé la loi du logarithme itéré de l'estimateur de la fonction de survie.

Cette thèse, comprend donc deux idées :

La première contribution de cette thèse porte sur un résultat de convergence uniforme presque sûre ainsi qu'un résultat de normalité asymptotique pour l'estimateur de la fonction du mode conditionnel dans un modèle doublement censuré et une application sur les données de transplantation cardiaque de Stanford. C'est l'objet du Chapitre 3.

En ce qui concerne la deuxième contribution qui est dans le Chapitre 2, nous établissons la convergence presque complète et nous donnons avec précision un taux de convergence de l'estimateur non-paramétrique de la fonction du quantile conditionnel dans un cadre de censure mixte avec des notions préliminaires sur la convergence presque complète et l'inégalité de Bernstein.

On peut considérer le premier Chapitre 1 comme une boîte à outils où se regroupe deux sections. La première section 1.1 est dédiée aux rappels des notions essentielles sur les différents modèles et les différents types de censures utilisées tout au long de notre travail. La présentation de la technique des classes de Vapnik-Chervonenkis (les V-C classes) est l'objet de la deuxième section 1.2.

Enfin, nous précisons que le traitement des données (simulations et applications réelles) est réalisé à l'aide de logiciel d'analyse statistique R, au Chapitre 4.

Chapitre 1

Données censurées et les classes V-C

1.1 Introduction aux données censurées

La notion de la censure est liée à ce qu'on appelle le temps de défaillance ou la durée de vie qui s'écoule jusqu'à la réalisation d'un certain événement attendu mais sans avoir accès à toute l'information, et pour laquelle on a perdu certaines données pour une raison ou une autre. Ces données sont appelées des données censurées. Le cas non observé le plus courant est le cas de la censure à droite et il y a une autre censure c'est la censure à gauche mais ce ne sont pas les seules.

Dans ce travail , nous allons mettre en lumière deux types de données incomplètes, l'une concerne la censure mixte et l'autre concerne les données doublement censurées qui peuvent aussi empêcher l'observation de la vraie durée et qui génèrent d'un mélange de données censurées à droite et à gauche selon les modèles de Patilea et Rolin (2006) et de Turnbull (1974).

Maintenant, nous présentons les différents modèles de censure et un peu plus loin, nous détaillerons des estimateurs auxquels nous nous intéresserons.

Censure à droite

Dans cette censure, l'évènement d'intérêt échappera à l'observation c'est-à-dire la durée d'observation peut se terminer et que l'évènement étudié n'a pas encore eu lieu. On peut l'expliquer par un exemple médical, le succès de l'intervention n'aura lieu qu'à la fin de la période observationnelle. De façon formelle, le succès d'une chirurgie est décrite par la variable d'intérêt Y et la durée d'observation est décrite par la variable R , dans ce cas la, on ne connaît qu'une borne inférieure de Y , c'est-à-dire $Y \geq R$.

Censure à gauche

Dans cette censure, on sait uniquement que la variable d'intérêt Y est inférieure ou égale à une variable connue L . Autrement dit on ne connaît qu'une borne supérieure de Y .

L'individu a déjà subi l'évènement avant qu'il ne soit observé.

Prenons par exemple L'apprentissage de la marche d'un bébé. L'âge auquel un bébé commence à marcher est à partir de 10 mois, cette période est la durée observée.

Certains parents peuvent se rappeler les causes qui permettent au bébé de marcher très tôt, avant l'âge de 10 mois comme :

- Le bébé a déjà développé les muscles qui permettent la marche.
- Le cerveau du bébé est capable d'envoyer les messages appropriés aux muscles.

Mais ils ne peuvent pas se rappeler à quel l'âge le bébé a commencé exactement la marche ? Ils donnent seulement une durée inférieure de 10 mois, c'est le cas où la durée observée pour ce bébé est alors censurée à gauche, c'est-à-dire $Y \leq L$.

Censure double ou mixte

Les deux phénomènes de censure mixte et double conduisent au cas où l'échantillon contient des données censurées à droite et à gauche en même temps.

- On parle de censure mixte lorsque une information est censurée au sens du modèle I qui est proposé par Patilea et Rolin (2006). Dans ce modèle, les trois variables aléatoires Y, L et R sont positives indépendantes où Y représente la durée d'intérêt et L et R sont les durées de censure à gauche et à droite respectivement. Au lieu d'observer un échantillon de Y on observe un échantillon de variables indépendantes et de même loi que $(Z = \max(\min(Y, R), L), A)$ où

$$A = \begin{cases} 0, & \text{si } L < Y < R, \\ 1, & \text{si } L < R \leq Y, \\ 2, & \text{si } \min(Y, R) \leq L. \end{cases}$$

est l'indicateur de censure.

- Par contre, dans un modèle de censure double, on considère les censures à droite et à gauche comme deux phénomènes liés et agissant indépendamment de la valeur exacte qui est irrémédiablement perdue.

Autrement dit la variable d'intérêt Y est indépendante du couple (R, L) et $P(L \leq R) = 1$. Au lieu d'observer un échantillon de Y on observe un échantillon de variables indépendantes et de même loi que $(Z = \max(\min(Y, R), L), \delta)$

où

$$\delta = \begin{cases} 1 & \text{si } L < Y \leq R \\ 2 & \text{si } Y > R, \\ 3 & \text{si } Y \leq L, \end{cases}$$

est l'indicateur de censure.

Dans la littérature d'autres modèles ont été proposés notamment la censure par intervalle qui est présentée par Turnbull (1976), avec plus de détails.

Censure par intervalle

Dans la pratique, il n'est pas toujours possible de disposer d'un échantillon de données censurée à droite, une observation censurée par intervalle, comme son nom l'indique, ne fournit aucune information sur la vraie variable d'intérêt et fournit des informations concernant un intervalle la contenant. On sait que ce type de censure permet de présenter les données censurées à gauche ou à droite par des intervalles du type $[0, b]$ et $[b, +\infty[$, respectivement.

Types de censure

I. La censure de type I (fixe) Ce type de censure décrit une expérience pendant une certaine période de temps, dans ce cas là, on a fixé le temps de censure par une constante " T". En général, ce mécanisme est rencontré dans les applications industrielles.

Prenons l'exemple de tester 200 ampoules électriques au bout d'un certain temps "T", au lieu d'observer les 200 ampoules qui nous intéressent, on observe uniquement "Y_i" le nombre d'ampoules ayant défaille avant le temps terminal $Y_i \leq T$ sinon $Y_i > T$. Une autre façon consiste à observer

une variable Z_i telle que $Z_i = \min(Y_i, T)$, $\delta_i = 1_{\{Y_i \leq T\}}$ pour $i = 1, \dots, n$ où δ_i l'indicateur de censure, qui détermine si Y a été censuré ou non .

II. La censure de type II (attente) Ce type se caractérise par la fixation de nombre d'événements à observer et où la durée de l'expérience n'est pas limitée. Par exemple, si on a $n=300$ ampoules, nous laissons la durée de l'expérience ouverte (c'est-à-dire elle est aléatoire) jusqu'à ce que les 70 premières ampoules sont tombées en panne. Donc le temps de censure ou le temps de défaillance est aléatoire (Y_1, \dots, Y_n) . L'observation dure jusqu'à ce qu'un nombre d'ampoules $k=70$ parmi les $n=300$ sources soient détectées (c'est-à-dire non aléatoire).

Dans cette censure, on observe $(Z_1, \delta_1), \dots, (Z_n, \delta_n)$ au lieu d'observer un échantillon de durées de survie (Y_1, \dots, Y_n) où $Z_i = \min(Y_i, Y_k)$ et $\delta_i = 1_{\{Y_i \leq Y_k\}}$.

III. La censure de type III (ou censure aléatoire de type I)

Ce type est basé sur deux échantillons aléatoires, l'une est la durée observée totale qui est tirée aléatoirement et peut être une borne supérieure ou inférieure à la durée de survie d'intérêt qui n'est pas maîtrisée (c'est-à-dire elle est aussi aléatoire). Autrement dit on peut considérer le temps de censure ou le nombre d'événements observés comme une variable aléatoire Y dont l'échantillon correspondant est (Y_1, \dots, Y_n) et on exprime la vraie durée de l'expérience par une autre variable aléatoire C dont l'échantillon correspondant est (C_1, \dots, C_n) , donc au lieu d'observer directement (Y_1, \dots, Y_n) on observe les couples $(Z_1, \delta_1), \dots, (Z_n, \delta_n)$ avec $Z_i = \min(Y_i, C_i)$, $\delta_i = 1_{\{Y_i \leq C_i\}}$.

Dans ce modèle qualifié de censure aléatoire, l'indépendance est une hypothèse importante entre la variable aléatoire de la censure et l'événement étudié. Cela signifie que les perdus de vue ne le sont pas pour des raisons liées à l'événement étudié.

Dans la sous section qui va suivre nous nous intéressons à la construction

des principaux estimateurs qui jouent un rôle essentiel dans notre travail.

1.1.1 L'estimateur de Kaplan-Meier

Dans le cas où on a des données complètes, la fonction de répartition de la variable d'intérêt Y s'estime de manière très simple en utilisant la fonction de répartition empirique qui est une fonction en escalier, continue à droite, admettant une limite à gauche et mettant un poids $1/n$ sur chaque point Y_i d'un échantillon de n v. a. i. d. Y_1, \dots, Y_n de même loi que Y . Autrement dit pour tout $t \in \mathbb{R}$ on a

$$F_n(t) = \frac{1}{n} \sum 1_{\{X_i \leq t\}}.$$

Mais, dans la pratique, lorsque Y est censurée, il est impossible d'utiliser l'estimateur empirique puisqu'il fait intervenir des quantités non observées. Nous nous intéressons donc dans cette partie à un autre estimateur de la fonction de répartition mais pour les données censurées à droite. Dans ce cas là nous disposons de deux échantillons Y_1, \dots, Y_n et R_1, \dots, R_n

représentant les durées d'intérêt et les temps de censure de même loi que Y et R respectivement qui sont indépendantes. Alors au lieu d'observer Y_i on observe le couple $Z_i = \min(Y_i, R_i)$, ainsi que l'indicateur de censure δ_i qui vaut 1 si la durée d'intérêt est observée, et 0 si elle est censurée, i.e $\delta_i = 1_{\{Y_i \leq R_i\}}$. Dans ce contexte de censure, Kaplan et Meier (1958) ont étudié l'estimateur de la fonction de répartition de Y , donné par la forme suivante,

pour $z < Z_{(n)}$ où $Z_{(n)} = \max\{Z_1, \dots, Z_n\}$

$$F_n(z) = 1 - \prod_{i: Z_i \leq z} \left(\frac{B_n(Z_i) - 1}{B_n(Z_i)} \right)^{\delta_i}$$

où $B_n(x) = \sum_{i=1}^n 1_{\{Z_i \geq x\}}$.

1.1.2 Les estimateurs self-consistants

Plusieurs modèles non paramétriques ont été proposés pour l'étude de la censure double. Le modèle de Turnbull (1974) est le plus utilisé. Dans ce modèle, au lieu d'observer un échantillon de Y on observe un échantillon $(Z_i = \max(\min(Y_i, R_i), L_i), \delta_i)$ de variables indépendantes et de même loi que $(Z = \max(\min(Y, R), L), \delta)$ où R et L sont respectivement des variables de censure à droite et à gauche, telles que R, L et Y sont positives et indépendantes où $L \leq R$ p.s et

$$\delta = \begin{cases} 1 & \text{si } L \leq Y \leq R \\ 2 & \text{si } Y > R, \\ 3 & \text{si } Y \leq L, \end{cases}$$

est l'indicateur de censure.

Les estimateurs non paramétriques pour les fonctions de survie S_Y, S_R et S_L notés respectivement par S_Y^n, S_R^n et S_L^n qui sont proposés dans ce modèle ne sont pas pour autant faciles à utiliser, car ils sont donnés par des équations intégrales dont les solutions ne sont pas connues explicitement. Ces équations sont appelées les équations de self-consistance et elles sont définies dans Ren (1997) pour tout $t \geq 0$ par les formules suivantes

$$S_Y^{(n)} = Q^{(n)} - \int_{u \leq t} \frac{S_Y^{(n)}(t)}{S_Y^{(n)}(u)} dQ_2^{(n)}(u) + \int_{t < u} \frac{1 - S_Y^{(n)}(t)}{1 - S_Y^{(n)}(u)} dQ_3^{(n)}(u),$$

avec $\int_{u \leq t} = 0$ si $S_Y^{(n)}(t) = 0$ et $\int_{t < u} = 0$ si $S_Y^{(n)}(t) = 1$

$$S_R^{(n)}(t) = 1 + \int_{u \leq t} \frac{dQ_2^{(n)}(u)}{S_Y^{(n)}(u)}, \quad t < B_n$$

$$S_L^{(n)}(t) = - \int_{t < u} \frac{dQ_3^{(n)}(u)}{1 - S_Y^{(n)}(u)}, \quad t \geq A_n,$$

où

$$Q_j(t) = P(Z > t, \delta = j), \quad Q_j^{(n)}(t) = \frac{1}{n} \sum_{i=1}^n 1_{\{Z_i > t, \delta_i = j\}}, \quad j = 1, 2, 3,$$

$$Q^{(n)}(t) = \sum_{i=0}^2 Q_j^{(n)}(t), \quad A_n = \min \{Z_i; S_Y^n(Z_i) < 1\} \text{ et } B_n = \max \{Z_i; S_Y^n(Z_i) > 0\}.$$

1.1.3 L'estimateur de Patilea et Rolin

Dans le cas où les données sont soumises à la fois à la censure à gauche et à la censure à droite, au sens du modèle I Patilea et Rolin (2006), au lieu d'observer un échantillon de Y on observe un échantillon du couple (Z, A) où $Z = \max(\min(Y, R), L)$ voir (1.1), auquel nous nous référons par modèle de censure mixte. Avant de donner la construction de l'estimateur produit limite S_n de la fonction de survie de la variable d'intérêt Y , rappelons les quantités et les notations suivantes.

Soit $M(t) = \sum_{k=0}^2 M^{(k)}(t)$ la fonction de répartition de Z où

$$M^{(k)}(t) = P(Z \leq t, A = k), \quad \text{pour } k = \{0, 1, 2\}, \quad (1.1)$$

sont les sous-distributions de la fonction de répartition de Z qui peuvent s'écrire comme suit

$$M^{(0)}(t) = \int_0^t F_L(u_-) S_R(u_-) dF_X(u), \quad (1.2)$$

$$M^{(1)}(t) = \int_0^t F_L(u_-) S_X(u) dF_R(u), \quad (1.3)$$

$$M^{(2)}(t) = \int_0^t \{1 - S_X(u) S_R(u)\} dF_L(u), \quad (1.4)$$

En s'inspirant des étapes suivante, Patilea et Rolin (2006) ont proposé leur estimateur qui est donné par (1.5).

Ils ont considéré d'abord deux variables $Y' = \min(X, R)$ et L dans un modèle de censure à gauche et ils ont estimé la fonction de répartition de Y' . Puis ils ont utilisé l'estimateur de la fonction de répartition de Y' pour estimer la fonction répartition de la variable d'intérêt Y en considérant un modèle de censure à droite. à partir d'un échantillon $(Z_i, A_i)_{1 \leq i \leq n}$, du couple (Z, A) et en remplaçant les relations (1.2),(1.3),(1.4) par leurs estimateurs empiriques, qui jouent un rôle important pour obtenir l'estimateur de la fonction de survie S_Y , donné par

$$S_n(Z'_j) = 1 - F_n(Z'_j) = \prod_{1 \leq l \leq j} \left\{ 1 - \frac{D_{0l}}{U_{l-1} - N_{l-1}} \right\}, \quad (1.5)$$

où $(Z'_j)_{1 \leq j \leq B}$ sont les valeurs distinctes des Z_i prises dans l'ordre croissant, et

$$D_{kj} = \sum_{1 \leq i \leq n} 1_{\{Z_i=Z'_j, A_i=k\}}, \quad N_j = \sum_{1 \leq i \leq n} 1_{\{Z_i \leq Z'_j\}},$$

$$U_{j-1} = n \prod_{j \leq l \leq B} \left\{ 1 - \frac{D_{2l}}{N_l} \right\},$$

pour $0 \leq l \leq 2$ et $1 \leq j \leq B$.

1.2 Rappels sur les classes (V-C) et l'inégalité de Talagrand

Cette section est consacrée à la présentation d'un outil nécessaire à notre recherche, qui nous a permis d'étudier en détail certaines propriétés de notre estimateur 3.7. Cet outil permet d'obtenir des bornes supérieures

pour certaines probabilités concernant la norme du supremum du processus empirique indexé par des classes de fonctions de Vapnik-Červonenkis. C'est l'inégalité exponentielle de type Talagrand (1996).

Il a bénéficié d'une large utilisation et sous différentes versions selon les conditions imposées aux classes de Vapnik-Červonenkis dans la recherche, nous citons Ould Saïd et Lemdani (2006), Khardani *et al.* (2011), Giné et Guillou (2001), Blondin (2004) et Bouzebda et El-hadjali (2020).

Avant de donner l'écriture formelle de cette inégalité, nous commençons par introduire des notions et quelques propriétés sur les classes de Vapnik-Červonenkis.

Nous rappelons d'abord le concept très puissant des classes V-C d'ensembles et de fonctions voir Van Der Vaart et Wellner (1996).

.

1.2.1 Les classes (V-C) d'ensembles

Définition 1.1. Soit \mathfrak{C} une classe de sous-ensembles d'un ensemble \mathcal{X} et soit $\{x_1, \dots, x_n\} \subset \mathcal{X}$. On dit que \mathfrak{C} prélève un sous-ensemble $A \subset \{x_1, \dots, x_n\}$ s'il existe $C \in \mathfrak{C}$ tel que $A = C \cap \{x_1, \dots, x_n\}$. On dit que \mathfrak{C} pulvérise $\{x_1, \dots, x_n\}$ si tous les sous-ensembles de ce dernier sont prélevés par \mathfrak{C} .

Définition 1.2. On appelle taille VC ou dimension VC de l'ensemble \mathfrak{C} le plus petit $n \in \mathbb{N}^*$ pour laquelle aucun sous-ensemble de \mathcal{X} de cardinal n n'est pulvérisé par \mathfrak{C} . Si \mathfrak{C} pulvérise tous les ensembles de tailles arbitraires, sa taille VC est infinie. Cette dimension est notée en général $\mathcal{V}(\mathfrak{C})$ ou $\text{VCD}(\mathfrak{C})$. De manière plus formelle, on introduit la taille des prélèvements $\Delta_n(\mathfrak{C}; x_1, \dots, x_n)$ qui correspond au nombre de sous-ensembles de $\{x_1, \dots, x_n\}$ que \mathfrak{C} peut prélever, et la taille VC d'une classe \mathfrak{C} par

$$\begin{aligned}\Delta_n(\mathfrak{C}; x_1, \dots, x_n) &= \text{card}\{C \cap \{x_1, \dots, x_n\} : C \in \mathfrak{C}\} \\ \mathcal{V}(\mathfrak{C}) &= \inf \left\{ n \in \mathbb{N}^* : \max_{x_1, \dots, x_n \in \mathcal{X}} \Delta_n(\mathfrak{C}; x_1, \dots, x_n) < 2^n \right\}.\end{aligned}$$

L'ensemble \mathfrak{C} est appelé classe de Vapnik-Červonenkis ou classe V-C si sa taille VC est finie : $\mathcal{V}(\mathfrak{C}) < +\infty$.

Exemple 1. La classe des intervalles $\mathcal{C} = \{] - \infty, t] : t \in \mathbb{R} \}$ est une classe V-C de dimension VC égale à 1.

La classe \mathcal{C} arrive à expulser un point mais pas deux points. En effet si $t_1, t_2 \in \mathbb{R}$ avec $t_1 < t_2$ alors l'ensemble $\{t_1, t_2\}$ n'est pas pulvérisé par \mathcal{C} car \mathcal{C} n'arrive pas à prélever $\{t_2\}$ puisque $\forall t \geq t_2,] - \infty, t] \cap \{t_1, t_2\} \neq \{t_2\}$ et $\forall t < t_2,] - \infty, t] \cap \{t_1, t_2\}$ ne contient pas t_2 .

Pour d'autres exemples on peut voir Van Der Vaart et Wellner (1996) page 135 et Kosorok (2008) page 156.

Les propriétés des classes V-C d'ensembles

Maintenant on rappelle les propriétés de stabilité des classes V-C . Autrement dit, des opérations des classes V-C (comme l'union, l'intersection, ...) donnent encore une classe V-C.

Lemme 1.1. Soit \mathfrak{C} et \mathfrak{D} des classes V-C d'un même espace \mathcal{X} et soient

$\phi : \mathcal{X} \rightarrow \mathcal{Y}$ et $\psi : \mathcal{Z} \rightarrow \mathcal{X}$ des fonctions fixées.

$\mathfrak{C}^c = \{C^c : C \in \mathfrak{C}\}$ est une classe V-C ,

$\mathfrak{C} \cap \mathfrak{D} = \{C \cap D : C \in \mathfrak{C}, D \in \mathfrak{D}\}$ est une classe V-C ,

$\mathfrak{C} \cup \mathfrak{D} = \{C \cup D : C \in \mathfrak{C}, D \in \mathfrak{D}\}$ est une classe V-C ,

$\phi(\mathfrak{C})$ est une classe V-C si ϕ est injective

$\psi^{-1}(\mathfrak{C})$ est une classe V-C.

Si $\mathfrak{C}, \mathfrak{D}$ sont des classes V-C respectivement des ensembles \mathcal{X} et \mathcal{Y} alors

$\mathfrak{C} \times \mathfrak{D}$ est une classe V-C de $\mathcal{X} \times \mathcal{Y}$.

Démonstration. Voir (le lemme 2.6.17 page 147) Van Der Vaart et Wellner (1996) \square

1.2.2 Les classes Vapnik-Červonenkis de fonctions

La classe des fonctions indicatrices $\mathfrak{F} = \{1_{]-\infty, t]} : t \in \mathbb{R}\}$ de dimension 1 a été introduite pour la première fois par Vapnik et Červonenkis (1981) et en s'inspirant de son nom.

Définition 1.3. Soit \mathfrak{F} une classe de fonctions mesurables définies sur \mathcal{X} et à valeurs réelles. Si la classe des sous-graphes des fonctions de \mathfrak{F} (le sous-graphe de $f \in \mathfrak{F}$ est $C_f = \{(x, t) : x \in \mathcal{X} \times \mathbb{R}, t \leq f(x)\}$) forme une classe V-C d'ensembles de $(\mathcal{X} \times \mathbb{R})$, alors \mathfrak{F} est appelée classe V-C (ou classe V-C de sous-graphes).

Les propriétés des classes V-C de fonctions

Soient \mathfrak{F} et \mathfrak{G} des classes V-C de fonctions d'un espace \mathcal{X} et $g : \mathcal{X} \rightarrow \mathbb{R}$, $\phi : \mathbb{R} \rightarrow \mathbb{R}$ et $\psi : \mathcal{Z} \rightarrow \mathcal{X}$ des fonctions.

$\mathfrak{F} \wedge \mathfrak{G} = \{f \wedge g : f \in \mathfrak{F}, g \in \mathfrak{G}\}$ est une classe V-C de fonctions (rappel :

$$a \wedge b = \min(a, b);$$

$\mathfrak{F} \vee \mathfrak{G} = \{f \vee g : f \in \mathfrak{F}, g \in \mathfrak{G}\}$ est une classe V-C de fonctions (rappel :

$$a \vee b = \max(a, b);$$

$\mathfrak{F}^+ = \{\{f > 0\} : f \in \mathfrak{F}\}$ est une classe V-C d'ensemble ,

$-\mathfrak{F}$ est une classe V-C de fonctions ,

$\mathfrak{F} + g = \{f + g : f \in \mathfrak{F}\}$ est une classe V-C de fonctions ,

$\mathfrak{F} \cdot g = \{fg : f \in \mathfrak{F}\}$ est une classe V-C de fonctions ,

$\mathfrak{F} \circ \psi = \{f(\psi) : f \in \mathfrak{F}\}$ est une classe V-C de fonctions ,

$\phi \circ \mathfrak{F}$ est une classe V-C de fonctions si ϕ est monotone.

Démonstration. Voir Van Der Vaart et Wellner (1996) (le lemme 2.6.18 page 147). \square

On se donne un espace métrique (E, d) et un $\epsilon > 0$. Le nombre de ϵ -recouvrement de l'espace métrique (E, d) noté $N(E, d, \epsilon)$ est défini comme le nombre minimal de boules ouvertes de centres dans E et de rayon ϵ requis pour couvrir l'ensemble E . Une classe de fonctions mesurables \mathfrak{F} est une classe V-C de fonctions par rapport à l'enveloppe M s'il existe une fonction mesurable M presque partout finie avec $|f| \leq M$ pour toute fonction $f \in \mathfrak{F}$, et des nombres réels K et v tels que :

$$N(\mathfrak{F}, \|\cdot\|_\infty, \epsilon \|\mathfrak{F}\|_\infty) \leq \left(\frac{K}{\epsilon}\right)^v, \quad 0 < \epsilon < 1.$$

Par ailleurs, nous avons besoin du lemme suivant dans la suite.

Lemme 1.2. *Giné et Guillou (1999)*

(a). Si \mathfrak{F} est finie alors \mathfrak{F} est une classe V-C par rapport à l'enveloppe $\max\{|f|/f \in \mathfrak{F}\}$.

(b). Si $\mathfrak{F} = \{f_x : x \in J\}$ où J est une partie de \mathbb{R} et $0 \leq f_x(s) \leq f_y(s)$ pour tout $x, y \in J, x < y$ et $s \in \mathcal{X}$ alors \mathfrak{F} est une classe V-C par rapport à $M = \sup\{|f|/f \in \mathfrak{F}\}$.

(c). Si \mathfrak{F}_1 et \mathfrak{F}_2 sont deux classes V-C par rapport à M_1 et M_2 respectivement, alors $\{f_1 + f_2/f_1 \in \mathfrak{F}_1, f_2 \in \mathfrak{F}_2\}$ et $\{f_1 - f_2/f_1 \in \mathfrak{F}_1, f_2 \in \mathfrak{F}_2\}$ sont des classes V-C par rapport à $(M_1^2 + M_2^2)^{1/2}$.

Pour la preuve de ce lemme, veuillez consulter l'article de Giné et Guillou (1999).

1.2.3 Inégalités de type Talagrand

Sous certaines conditions, Giné et Guillou (2001) ont pu trouver des nouvelles bornes supérieures des inégalités de Talagrand (1994) et Talagrand (1996). Dans notre travail nous nous intéressons à l' inégalité suivante montrée par Giné et Guillou (2001) (voir proposition (2.2)).

Proposition 1.1. *Giné et Guillou (2001)*

Soit $\{X_n \quad n \geq 1\}$ une suite de variables aléatoires réelles indépendantes et identiquement distribuées et soit \mathfrak{F} une classe de fonctions V-C mesurable uniformément bornée. Nous considérons σ_n^2 et U_n tels que $\sigma_n^2 \geq \sup_{f \in \mathfrak{F}} \text{Var} f(X_1)$, $U_n \geq \sup_{f \in \mathfrak{F}} \|f\|_\infty$ et $0 < \sigma_n \leq U_n$. Alors, il existe des constantes C et A . telles que

$$P \left[\sup_{f \in \mathfrak{F}} \left| \sum_{i=1}^n (f(X_i) - E(f(X))) \right| > \eta \right] \\ \leq C \exp \left\{ -\frac{1}{C} \frac{\eta}{U_n} \log \left(1 + \frac{\eta U_n}{C \left(\sqrt{n} \sigma_n + U_n \sqrt{\log(AU_n/\sigma_n)} \right)^2} \right) \right\}$$

pour

$$\eta \geq B \left[U_n \log \left(\frac{AU_n}{\sigma_n} \right) + \sqrt{n} \sigma_n \sqrt{\log \left(\frac{AU_n}{\sigma_n} \right)} \right]. \quad (1.6)$$

Chapitre 2

Estimation à noyau du quantile Conditionnel

Dans ce chapitre, nous nous intéressons aux estimateurs non-paramétriques de la fonction répartition conditionnelle et du quantile conditionnel dans un modèle de censure mixte.

Nous commençons par rappeler quelques outils et résultats relatifs à la convergence presque complète et à l'inégalité de Bernstein, Nous présentons également la définition de l'estimateur du quantile conditionnel introduit par Gannoun *et al.* (2002) en absence de censure. Dans le cadre de la censure à droite, on est passé par la définition de Ould Saïd (2006) et Ould-Saïd et Sadki (2008a). Enfin, nous proposons d'estimer la fonction du quantile Conditionnel par un estimateur dans le cas où la variable réponse est soumise à une censure mixte (relation 2.9) et nous montrons sa convergence presque complète uniforme avec un taux.

2.1 La convergence presque complète et l'inégalité de Bernstein

La convergence presque complète que nous avons utilisée est plus forte que la convergence presque sûre et la convergence en probabilité mais elle n'implique pas la convergence en moyenne quadratique. Dans les dernières années elle se répand largement dans la statistique non-paramétrique des données fonctionnelles. nous citons par exemple Ferraty et Vieu (2006), Ezzahrioui et Ould Saï (2008) et Messaci *et al.* (2015) et pour les données censurées voir Kebabi et Messaci (2012), Kitouni *et al.* (2015).

Ci-dessous nous donnons la définition formelle de la convergence presque complète et ses propriétés puis nous présentons une version de l'inégalité exponentielle de type Bernstein, dont nous avons besoin pour l'établissement des résultats que nous avons choisis de reprendre.

2.1.1 Définitions et propriétés

Nous commençons à rappeler quelques définitions et propriétés de ce type de convergence.

Définition 2.1. Soient (U_n) et (V_n) deux suites de nombres réels. Nous supposons que (V_n) ne s'annule pas à partir d'un certain rang. On dit que (U_n) est dominée par (V_n) s'il existe un nombre réel M et un entier n_0 , tel que, pour tout $n \geq n_0$, on obtient

$$|U_n| \leq M|V_n|,$$

et nous notons $U_n = O(V_n)$.

Définition 2.2. La définition de la convergence presque complète est apparue pour la première fois dans les travaux de Hsu et Robbins (1947).

On dit que la suite de variables aléatoires réelles $(X_n)_{n \in \mathbb{N}}$ converge presque complètement vers une variable aléatoire X lorsque $n \rightarrow \infty$ si

$$\forall \varepsilon > 0, \quad \sum_{n \in \mathbb{N}} P[|X_n - X| > \varepsilon] < \infty,$$

et on note par :

$$\lim_{n \rightarrow \infty} X_n = X, p.co.$$

Définition 2.3. La vitesse de convergence presque complète

On dit que la vitesse de convergence presque complète de la suite de variables aléatoires réelles $(X_n)_{n \in \mathbb{N}}$ vers X est d'ordre (U_n) ((U_n) étant une suite numérique déterministe), si et seulement si :

$$\exists \varepsilon_0 > 0, \quad \sum_{n \in \mathbb{N}} P[|X_n - X| > \varepsilon_0 U_n] < \infty,$$

et on écrit

$$X_n - X = O_{p.co.}(U_n).$$

Nous passons maintenant aux propriétés de ce mode de convergence.

Proposition 2.1. Soient l_x et l_y deux nombres réels et (U_n) une suite denombres réels tels que $\lim_{n \rightarrow \infty} U_n = 0$, nous avons

- i) Si $\lim_{n \rightarrow \infty} X_n = l_x$ p.co. et $\lim_{n \rightarrow \infty} Y_n = l_y$ p.co, alors
 - a) $\lim_{n \rightarrow \infty} (X_n + Y_n) = l_x + l_y$ p.co,
 - b) $\lim_{n \rightarrow \infty} (X_n Y_n) = l_x l_y$ p.co.,
 - c) $\lim_{n \rightarrow \infty} \frac{1}{X_n} = \frac{1}{l_x}$ p.co. lorsque $l_x \neq 0$.
- ii) Si $X_n - l_x = O_{a.co.}(U_n)$ et $Y_n - l_y = O_{p.co.}(U_n)$, alors
 - a) $(X_n + Y_n) - l_x - l_y = O_{p.co.}(U_n)$,
 - b) $(X_n Y_n) - l_x l_y = O_{p.co.}(U_n)$,
 - c) $\frac{1}{X_n} - \frac{1}{l_x} = O_{p.co.}(U_n)$ lorsque $l_x \neq 0$.
- iii) Si $X_n = O_{p.co.}(U_n)$ et $\lim_{n \rightarrow \infty} Y_n = l_y$ p.co, alors

- a) $X_n Y_n = O_{p.co.}(U_n)$,
 b) $\frac{X_n}{Y_n} = O_{p.co.}(U_n)$, lorsque $l_y \neq 0$.

Démonstration. voir Ferraty et Vieu (2006). □

2.1.2 Inégalités de type Bernstein

Nous allons donner une version des inégalités exponentielles de type Bernstein qui nous facilite notre recherche. Soit $\{X_n, n \geq 1\}$ une suite de variables aléatoires réelles, indépendantes et centrées.

Proposition 2.2. *Si*

$$\forall m \geq 2, |E(X_i^m)| \leq \left(\frac{m!}{2}\right) (a_i)^2 b^{m-2},$$

alors

$$\forall \varepsilon \geq 0, P \left[\sum_{i=1}^n |X_i| > \varepsilon A_n \right] \leq 2 \exp \left\{ \frac{-\varepsilon^2}{2(1 + \frac{b\varepsilon}{A_n})} \right\},$$

où $(a_i)_{1 \leq i \leq n}$ sont des réels positifs, $b \in \mathbb{R}^+$ et $A_n^2 = a_1^2 + a_2^2 + a_3^2 + \dots + a_n^2$.

Démonstration. cf. Bernstein (1946). □

Corollaire 2.1. *Ferraty et Vieu (2006)*

a) *S'il existe une constante positive $M < \infty$, telle que $|X_1| \leq M$, alors on a*

$$\forall \varepsilon \geq 0, P \left(\left| \sum_{i=1}^n X_i \right| > \varepsilon n \right) \leq 2 \exp \left\{ \frac{-\varepsilon^2 n}{2\sigma^2(1 + \frac{M\varepsilon}{\sigma^2})} \right\},$$

où $\sigma^2 = E(X_i^2)$.

b) Supposons que $(X_i)_{1 \leq i \leq n}$ dépendent de n et posons $\sigma_n^2 = E(X_i^2)$, s'il existe $M = M_n < \infty$ tel que $|X_1| \leq M$ et si $\frac{M}{\sigma_n^2} \leq C < \infty$ et $U_n = n^{-1}\sigma_n^2 \log n$ vérifie $\lim_{n \rightarrow \infty} U_n = 0$, nous trouvons

$$\frac{1}{n} \sum_{i=1}^n X_i = O_{p.co}(\sqrt{U_n}).$$

Démonstration. a) En utilisant la proposition 2.2 avec $a_i^2 = \sigma^2$, $A_n^2 = n\sigma^2$ et $b = M$ nous obtenons a).

b) Comme $\frac{MU_n}{\sigma_n^2} \rightarrow 0$, il suffit d'appliquer le résultat a) pour $\varepsilon = \varepsilon_0 \sqrt{U_n}$, donc il existe une constante C' telle que

$$\begin{aligned} P \left[\frac{1}{n} \left| \sum_{i=1}^n X_i \right| > \varepsilon_0 \sqrt{U_n} \right] &\leq 2 \exp \left\{ \frac{-\varepsilon_0^2 \log n}{2 \left(1 + \varepsilon_0 \sqrt{\frac{MU_n}{\sigma_n^2}} \right)} \right\} \\ &\leq 2n^{-C'} \varepsilon_0^2. \end{aligned}$$

Pour ε_0 bien choisi le terme de droite est le terme général d'une série convergente. Le corollaire est donc démontré. \square

2.2 Introduction de l'estimateur du quantile conditionnel

2.2.1 Dans le cas des données complètes

En absence de censure, plusieurs approches ont été développées pour l'estimation des quantiles conditionnels, citons sans prétendre à l'exhaustivité, Samanta (1989), Honda (2000), Ying Liang et Ua-Àlvarez (2009), Gannoun *et al.* (2003), Gannoun (1991). Gannoun *et al.* (2002), ont proposé un

estimateur à noyau de la fonction de répartition conditionnelle

$F^x(y) := P(Y \leq y | X = x)$, à l'aide de l'indicatrice $1_{\{Y_i \leq y\}}$, défini pour $y \in \mathbb{R}$ par

$$F_n^x(y) = \frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right) 1_{\{Y_i \leq y\}}}{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)}, \quad (2.1)$$

où (X_i, Y_i) est un échantillon de (X, Y) . Notons $f(x, y)$ la densité de probabilité conjointe de la dernière paire de variables. La fonction K est une densité de probabilité, appelée noyau et h_n est une suite de nombres réels strictement positifs tendant vers 0 et appelée fenêtre.

Il en découle naturellement un estimateur du quantile conditionnel t_p^x d'ordre p ($p \in]0, 1[$), défini par

$$t_{p,n}^x = \inf\{y \in \mathbb{R}, F_n^x(y) \geq p\}. \quad (2.2)$$

où t_p^x est donné par

$$t_p^x = \inf\{y \in \mathbb{R}, F^x(y) \geq p\}. \quad (2.3)$$

Comme l'estimateur (2.1) n'est pas dérivable, ils ont suggéré une généralisation de cet estimateur, en remplaçant l'indicatrice par une fonction de répartition admettant une densité de probabilité.

On obtient alors l'estimateur de la fonction répartition conditionnelle par le noyau produit (double noyau) $F_n^x(y)$:

$$F_n^x(y) = \frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right) H\left(\frac{y-Y_i}{g_n}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)}, \quad (2.4)$$

où g_n représente la fenêtre de lissage qui tend avec une certaine vitesse vers 0 lorsque n tend vers l'infini.

2.2.2 Dans le cas des données censurées à droite

Dans la suite de cette section, nous passons à une classe d'estimateurs du quantile conditionnel (2.3) où les observations Y_i sont soumises à un mécanisme de censure à droite c'est à dire qu'on a seulement des observations du triplet $(X_i, Z_i = \min(X_i, R_i), \delta_i = 1_{\{Y_i \leq R_i\}})$, qui a été défini dans la section 1.1. Dans ce modèle, l'estimation non paramétrique du quantile conditionnel a suscité beaucoup d'intérêt. Le premier qui a proposé cet estimateur à l'aide de l'indicatrice est Ould Saïd (2006) d'abord il a estimé la fonction de répartition conditionnelle $F^x(y)$ en se basant sur l'idée de Carbonez *et al.* (1995) (qui a été reprise par Kohler *et al.* (2002), alors l'estimateur à noyau de $F^x(y)$ est donné par

$$\begin{aligned} F_n^{x,d_1}(y) &= \sum_{i=1}^n W_{i,n}(x) \frac{\delta_i 1_{\{Z_i \leq y\}}}{S_n(Z_i)}, \\ &= \frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right) \frac{\delta_i 1_{\{Z_i \leq y\}}}{S_n(Z_i)}}{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)}, \end{aligned} \quad (2.5)$$

où S_n est l'estimateur de Kaplan-Meier de la fonction de survie S_R de R et les poids $W_{i,n}$ sont des fonctions mesurables de x dépendant de X_1, X_2, \dots, X_n .

En fin Ould Saïd (2006) a énoncé des résultats sur la convergence uniforme presque sûr de l'estimateur du quantile conditionnel dans le cas des

observations i.i.d, mais cet estimateur n'est pas dérivable ce qui a rendu difficile l'établissement de sa normalité asymptotique, danc Ould-Saïd et Sadki (2008a) ont construit un nouvel estimateur de la fonction de répartition conditionnelle par la méthode des noyaux qui est donné par

$$F_n^{x,d_2}(y) = \frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right) H\left(\frac{y-Z_i}{g_n}\right) \delta_i}{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right) S_n(Z_i)}, \quad (2.6)$$

et ils ont établi ses propriétés asymptotique (la vitesse de convergence uniforme presque sûr et la normalité asmpototique). Au final, ils ont déduit facilement la formule de l'estimateur du quantile conditionnel non-paramétrique 2.3 et ses propriétés.

2.3 Estimation du quantile conditionnel dans le modèle de censure mixte

2.3.1 Le modèle

Notre travail rentre donc dans le cadre de la censure mixte correspondant au modèle I de Patilea et Rolin (2006) et abordé dans la sous section (1.1.3) Autrement dit, nous observons l'échantillon $(X_i, Z_i, A_i)_{\{1 \leq i \leq n\}}$ du couple $(X, Z, L), A$ où $Z = \max(\min(Y, R), L)$, avec les variables aléatoires positives et indépendantes Y, R et L représentant, respectivement, la variable d'intérêt, la variable de censure à droite et la variable de censure à gauche. L'indicateur de censure A est de la forme suivante :

$$A = \begin{cases} 0 & \text{si } L < Y < R, \\ 1 & \text{si } L < R \leq Y, \\ 2 & \text{si } \min(Y, R) \leq L. \end{cases}$$

2.3.2 Estimation et hypothèses

Dans cette section, nous proposons un estimateur non paramétrique du quantile conditionnel et nous montrons sa convergence uniforme presque complète, avec un taux. Avant de développer le détail de construction de notre estimateur, nous allons introduire des quantités et des notations qui nous seront indispensables.

Dans ce qui suit, pour toute variable aléatoire U , nous notons $T_U = \sup\{t : F_U(t) < 1\}$ et $I_U = \inf\{t : F_U(t) > 0\}$ où $F_U(u) = P(U \leq u)$ et $S_U(u) = 1 - F_U(u)$ sont respectivement la fonction de répartition et la fonction de survie. Sous l'hypothèse A_7 ci-dessous et l'indépendance de Y , R et L , on a

$$E(1_{\{A_i=0\}}/X, Y) = S_R(Y)F_L(Y). \quad (2.7)$$

On peut la montrer facilement. Pour tout B dans $\sigma(X, Y)$ (tribu engendrée par le couple (X, Y)) il existe un borélien C tel que $B = (X, Y)^{-1}(C)$.

L'indépendance de (X, Y) et (L, R) permet d'écrire

$$\begin{aligned}
\int_B (1_{A=0}) dP &= \int_B (1_{L<Y<R}) dP \\
&= \int_{C \times \mathbb{R}_+^2} (1_{l<y<r}) dP_{(X,Y,L,R)} \\
&= \int_{C \times \mathbb{R}_+^2} (1_{l<y<r}) dP_{(X,Y)} \otimes dP_{(L,R)}.
\end{aligned}$$

Sous l'hypothèse A_7 ci-dessous puis on applique le théorème de Fubini et l'indépendance de R et L , nous avons

$$\begin{aligned}
\int_B (1_{A=0}) dP &= \int_C \left(\int_{\mathbb{R}_+^2} (1_{l<y<r}) dP_{(L,R)} \right) dP_{(X,Y)} \\
&= \int_C \left(\int_{\mathbb{R}_+} (1_{l<y}) dP_L \times \int_{\mathbb{R}_+} (1_{y<r}) dP_R \right) dP_{(X,Y)} \\
&= \int_C (F_L(y)S_R(y)) dP_{(X,Y)} \\
&= \int_B F_L(Y_1)S_R(Y_1)dP,
\end{aligned}$$

car F est continue. De plus, $F_L(Y)S_R(Y)$ étant clairement mesurables par rapport à $\sigma(X, Y)$, le résultat en découle.

Dans ce modèle, rappelons que Messaci (2010) a proposé une estimation de $E(g'_1(X, Y))$, donnée par :

$$E(g'_1(X, Y)) = \frac{1}{n} \sum_{i=1}^n \frac{1_{\{A_i=0\}} g'_1(X_i, Z_i)}{S_R(Z_i)F_L(Z_i)}, \quad (2.8)$$

où $g'_1 : \mathbb{R}^2 \rightarrow \mathbb{R}$.

En effet, en utilisant la relation (2.7) dans l'expression précédente, nous trouvons que

$$\begin{aligned} E \left(\frac{1_{\{A=0\}} g'_1(X, Z)}{S_R(Z) F_L(Z)} \right) &= E \left(E \left(\frac{1_{\{A_i=0\}} g'_1(X, Y)}{S_R(Z) F_L(Z)} \right) / (X, Y) \right) \\ &= E \left(\frac{g'_1(X, Y)}{S_R(Y) F_L(Y)} E (1_{\{A_i=0\}} / (X, Y)) \right) \\ &= E (g'_1(X, Y)), \end{aligned}$$

car F est continue. De plus, $F_L(Y)S_R(Y)$ étant clairement mesurable par rapport à $\sigma(X, Y)$, le résultat en découle.

Comme S_R, F_L sont généralement inconnues, nous les remplacerons par leurs estimations données respectivement par S_n, F_n . S_n est introduit par Patilea et Rolin (2006) et F_n est obtenu par Kaplan et Meier (1958). Puis, en suivant les relations (2.4), (2.8), il est facile de construire un estimateur non paramétrique de la fonction de répartition conditionnelle comme suit :

$$F_n^{x,m}(y) := \sum_{i=1}^n W_{n,i}(x) \frac{1_{\{A_i=0\}} H_i(y)}{S_{(n)}(Z_i) F_n(Z_i) + V_n} = \frac{F_{N,n}^m(x, y)}{f_n(x)}, \quad (2.9)$$

où $H_i(y) := H\left(\frac{y-Z_i}{g_n}\right)$, $K_i(x) := K\left(\frac{x-X_i}{h_n}\right)$ et $W_{n,i}(x) = \frac{K_i(x)}{\sum_{i=1}^n K_i(x)}$,

$$f_n(x) = \frac{1}{nh_n} \sum_{i=1}^n K_i(x), \quad (2.10)$$

est l'estimateur du noyau de la densité $f_X(x)$ de X ,

$$F_{N,n}^m(x, y) = \frac{1}{nh_n} \sum_{i=1}^n \frac{1_{\{A_i=0\}} K_i(x) H_i(y)}{S_n(Z_i) F_n(Z_i) + V_n}, \quad (2.11)$$

et V_n est une suite de réels strictement positifs, permettant d'obtenir une borne inférieure non aléatoire appropriée pour le dénominateur.

A partir de l'estimation de la fonction de répartition conditionnelle, on peut attaquer le problème de prédiction en définissant un estimateur à noyau du quantile conditionnel t_p^x comme suit :

$$t_{p,n}^{x,m} = \inf\{y \in \mathbb{R}, F_n^{x,m}(y) \geq p\}, \quad (2.12)$$

où t_p^x est donnée dans la relation (2.3). Rappelons que $t_{1/2}^x$ est la médiane conditionnelle.

Nous aurons aussi besoin de poser

$$\tilde{F}_{N,n}^m(x, y) = \frac{1}{nh_n} \sum_{i=1}^n \frac{1_{\{A_i=0\}} K_i(x) H_i(y)}{S_R(Z_i) F_L(Z_i)}, \quad (2.13)$$

$$F(x, y) = F^x(y) f_X(x). \quad (2.14)$$

Maintenant nous considérons l'hypothèse suivante

$$\mathbf{A}_0 : \max(I_Y, I_L) < I_R.$$

On peut appliquer le théorème 1 de Kitouni *et al.* (2015) qui donne pour

$$\theta < \min(T_R, T_Y),$$

$$\sup_{u \leq \theta} |S_n(u) - S_R(u)| = O_{p.co.} \left(\sqrt{\frac{\log n}{n}} \right). \quad (2.15)$$

Grâce à la relation (3) dans Kitouni *et al.* (2015), pour tout $\theta > \min(I_Y, I_R)$,

$$\sup_{t \geq \theta} |F_n(t) - F_L(t)| = O_{p.co.} \left(\sqrt{\frac{\log n}{n}} \right). \quad (2.16)$$

Le reste de nos hypothèses (énoncées ci-dessous) sont, grosso modo, standard dans le cadre de l'estimation du noyau.

- A₁** : K est une densité bornée à support compact et $\int tK(t)dt = 0$.
- A₂** : H est une fonction de distribution avec une dérivée H' vérifiant $\int H'(t)dt = 1$, $\int tH'(t) = 0$ et $\int t^2H'(t)dt < \infty$ a un support compact.
- A₃** : La fonction de distribution conditionnelle $F^x(y)$ a une dérivée première continue par rapport à y , notée $f^x(y)$. De plus, l'application F définie dans (2.14) est deux fois continûment dérivable autour de (x, y) .
- A₄** : h_n, g_n satisfont : $\lim_{n \rightarrow \infty} h_n = 0$, $\lim_{n \rightarrow \infty} g_n = 0$, $\lim_{n \rightarrow \infty} nh_n = \infty$, $\lim_{n \rightarrow \infty} nh_n^3 = 0$, $\lim_{n \rightarrow \infty} \frac{ng_n^4}{h_n} = 0$.
- A₅** : La densité marginale f_X de la variable X est deux fois continûment dérivable autour de x , $f_X(x) > 0$.
- A₆** : Pour tout $v \in \mathbb{R}$, l'application partielle $f(\cdot, v)$ est différentiable et $\exists M > 0, \exists N_x$ tel que $\forall y \in N_x, \forall v \in \mathbb{R}, \left| \frac{\partial f(y, v)}{\partial y} \right| \leq M$.
Maintenant, dans ce cadre, nous utilisons les hypothèses utilisées dans Messaci (2010).
- A₇** : $(R, L), (X, Y)$ sont indépendants et F_L est continue.

\mathbf{A}_8 : $\exists T < T_R$ et $I > I_L$ tel que $\forall n \in \mathbb{N}$, $\forall i$ ($1 \leq i \leq n$), $A_i = 0 \implies I \leq Z_i \leq T$ p.s.

2.3.3 Résultats et preuves

Le but de la présente partie est de donner un comportement asymptotique de l'estimateur à noyau du quantile conditionnel, $t_{p,n}^{x,m}$ que nous avons déjà introduit par la formule (2.9).

Ce comportement asymptotique est décrit dans le théorème suivant.

Théorème 2.1. *Sous les hypothèses $A_0 - A_8$, nous avons*

$$|t_{p,n}^{x,m} - t_p^x| = O(h_n^2) + O(g_n^2) + O(h_n g_n) + O_{p.co} \left(\sqrt{\frac{\log n}{nh_n}} \right).$$

Pour cela nous avons besoin de la proposition suivante qui joue un rôle important :

Proposition 2.3. *Sous les hypothèses : $A_0 - A_8$, Nous avons :*

$$|F_n^{x,m}(y) - F^x(y)| = O(h_n^2) + O(g_n^2) + O(h_n g_n) + O_{p.co} \left(\sqrt{\frac{\log n}{nh_n}} \right).$$

Démonstration. Nous devons utiliser la décomposition suivante pour prouver la proposition.

$$F_n^{x,m}(y) - F^x(y) = F_n^{x,m}(y) - \hat{F}_n^{x,m}(y) + \frac{\tilde{F}_{N,n}^m(x, y) - F(x, y)}{f_n(x)} - \frac{f_n(x) - f(x)}{f_n(x)} F^x(y),$$

$$\text{où } \hat{F}_n^{x,m}(y) = \frac{\tilde{F}_{N,n}^m(x,y)}{f_n(x)}.$$

Dans ce qui suit, C désigne une constante générique.

La preuve de la proposition 2.3 découle des lemmes suivants.

Lemme 2.1. *Sous les hypothèses A_0 , A_2 , A_4 et A_8 nous avons*

$$\left| F_n^{x,m}(y) - \hat{F}_n^{x,m}(y) \right| = O_{p.co} \left(\sqrt{\frac{\log n}{nh_n}} \right).$$

Démonstration. On obtient à partir des hypothèses A_8 et A_2 .

$$\begin{aligned} \left| F_n^{x,m}(y) - \hat{F}_n^{x,m}(y) \right| &= \left| \sum_{i=1}^n W_{n,i}(x) 1_{\{A_i=0\}} \frac{H\left(\frac{y-Z_i}{g_n}\right)}{S_n(Z_i)F_n(Z_i) + V_n} - \sum_{i=1}^n W_{n,i}(x) 1_{\{A_i=0\}} \frac{H\left(\frac{y-Z_i}{g_n}\right)}{S_R(Z_i)F_L(Z_i)} \right| \\ &\leq \sum_{i=1}^n W_{n,i}(x) 1_{\{A_i=0\}} \\ &\quad \times \left| \frac{F_L(Z_i)S_R(Z_i) - F_n(Z_i)S_R(Z_i) + F_n(Z_i)S_R(Z_i) - F_n(Z_i)S_n(Z_i) - V_n}{(F_n(I)S_n(T) + V_n)F_L(I)S_R(T)} \right|, \\ &\leq \sum_{i=1}^n W_{n,i}(x) 1_{\{A_i=0\}} \frac{\sup_{t \geq I} |F_n(t) - F_L(t)| + \sup_{t \leq T} |S_n(t) - S_R(t)| + V_n}{(F_n(I)S_n(T) + V_n)F_L(I)S_R(T)}. \end{aligned}$$

Sous l'hypothèse A_0 , les relations (2.15) et (2.16) permettent d'écrire

$$\sup_{\{t \leq T\}} |S_n(t) - S_R(t)| + \sup_{\{t \geq I\}} |F_L(t) - F_L(t)| = O_{p.co} \left(\sqrt{\frac{\log n}{n}} \right).$$

et $S_n(T) \xrightarrow{p.co} S_R(T)$ et $F_n(I) \xrightarrow{p.co} F_L(I)$ quand $n \rightarrow \infty$, ainsi que nous avons $\sum_{i=1}^n W_{n,i}(x) = 1$. Alors, on peut conclure que

$$|F_n^{x,m}(y) - \hat{F}_n^{x,m}(y)| = O_{p.co} \left(\sqrt{\frac{\log n}{n}} \right).$$

Enfin, la preuve est terminée en utilisant le dernier résultat et l'hypothèse A_4 , on obtient

$$F_n^{x,m}(y) - \hat{F}_n^{x,m}(y) = O_{p.co} \left(\sqrt{\frac{\log n}{nh_n}} \right). \quad (2.17)$$

□

Lemme 2.2. *Sous les hypothèses $A_1 - A_3$ et A_7 on a*

$$\left| E\tilde{F}_{N,n}^m(x, y) - F(x, y) \right| = O(h_n^2) + O(g_n^2) + O(h_n g_n).$$

Démonstration 2.1. *comme (X_i, Z_i, δ_i) sont identiquement distribués, on a*

$$E\tilde{F}_{N,n}^m(x, y) = E \left[\frac{K_1(x)}{h_n} E \left(\frac{1_{\{\delta_1=1\}} H_1(y)}{S_R(Y_1) F_L(Y_1)} \middle| X_1 \right) \right]. \quad (2.18)$$

Nous allons commencer par le calcul de ce terme $E \left(\frac{1_{\{\delta_1=1\}} H_1(y)}{S_R(Y_1) F_L(Y_1)} \middle| X_1 \right)$.

En utilisant (3.3), une intégration par parties et l'hypothèse N_2 , on obtient

$$\begin{aligned} E \left(\frac{1_{\{\delta_1=1\}} H_1(y)}{S_R(Y_1) F_L(Y_1)} \middle| X_1 \right) &= E \left[\frac{H_1(y)}{S_R(Y_1) F_L(Y_1)} E (1_{\{\delta_1=1\}} | (X_1, Y_1)) \middle| X_1 \right], \\ &= \int H \left(\frac{y-v}{g_n} \right) f^{X_1}(v) dv, \\ &= \int F^{X_1}(y - wg_n) H'(w) dw. \end{aligned} \quad (2.19)$$

Le dernier résultat avec (2.18), nous permet d'écrire

$$\begin{aligned} E\tilde{F}_{N,n}^m(x, y) &= \frac{1}{h_n} \int K\left(\frac{x-u}{h_n}\right) \left(\int F^u(y-wg_n)H'(w)dw \right) f(u)du, \\ &= \int \int K(z)F(x-zh_n, y-wg_n)H'(w)dwdz. \end{aligned} \quad (2.20)$$

Sous A_3 , le développement de Taylor combiné avec $A_1 - A_2$ implique que

$$E\tilde{F}_{N,n}^m(x, y) - F(x, y) = O(h_n^2) + O(g_n^2) + O(h_n g_n). \quad (2.21)$$

Lemme 2.3. Sous les hypothèses $A_1 - A_3, A_8$ nous avons

$$\left| E\tilde{F}_{N,n}^m(x, y) - \tilde{F}_{N,n}^m(x, y) \right| = O_{p.co} \left(\sqrt{\frac{\log n}{nh_n}} \right).$$

Démonstration. Posons $\Delta_i(x, y) = \gamma_i(x, y) - E\gamma_i(x, y)$,

où $\gamma_i(x, y) = \frac{1_{\{A_i=0\}}H_1(y)K_1(x)}{h_n S_R(Z_i)F_L(Z_i)}$. Sous les hypothèses A_1, A_2 et A_8 , nous pouvons montrer que

$$|\gamma_i(x, y)| \leq \frac{C}{h_n S_R(T)F_L(I)} := M. \quad (2.22)$$

Nous devons maintenant borner $\sigma^2 = E\Delta_1^2(x, y) \leq E\gamma_1^2(x, y)$.

$$E\gamma_1^2(x, y) = \frac{1}{h_n^2} E \left[K_1^2(x) E \left(\frac{1_{\{A_1=0\}}H_1^2(y)}{S_R^2(Z_1)F_L^2(Z_1)} \middle| X_1 \right) \right].$$

Sous les hypothèses A_1, A_2 et A_8 on peut voir que

$$E\gamma_1^2(x, y) \leq \frac{C}{h_n S_R^2(T) F_L^2(I)} E \left[\frac{K_1(x)}{h_n} E(H_1(y)|X_1) \right],$$

où la dernière égalité est dérivée de la même manière que la relation (2.20).

Les hypothèses A_3, A_1 et A_2 combinées à un développement de Taylor autour de (x, y) et à la borne de l'application F définie dans (2.14), permettent d'écrire $E \left[\frac{K_1(x)}{h_n} E(H_1(y)|X_1) \right] \leq C$, donc on peut conclure que

$$E\gamma_1^2(x, y) \leq \frac{C}{h_n}. \quad (2.23)$$

Maintenant, on est en mesure d'appliquer l'inégalité exponentielle de type Bernstein par le corollaire A.9-i Ferraty et Vieu (2006). Cette inégalité combinée aux relations (2.22) et (2.23) donne directement : pour tout $\epsilon < \frac{\sigma^2}{M}$.

$$\mathbb{P} \left[|E\tilde{F}_{N,n}^m(x, y) - \tilde{F}_{N,n}^m(x, y)| > \epsilon \right] \leq 2e^{-\frac{n\epsilon^2 h_n}{4C}}.$$

La condition A_4 , avec $\epsilon = \epsilon_0 \sqrt{\frac{\log n}{nh_n}}$ et pour ϵ_0 suffisamment grand, montre que le terme de droite est le terme général d'une série convergente.

Lemme 2.4. *Sous les conditions A_1, A_4 et A_5 on a :*

- i) $E f_n(x) - f_X(x) = O(h_n^2)$, ii) $E f_n(x) - f_n(x) = O_{p.co}(\sqrt{\frac{\log n}{nh_n}})$.*
et iii) $\exists \eta_1 > 0, \sum_{i=1}^{\infty} P(f_n(x) < \eta_1) < \infty$.

Démonstration. La preuve découle de la même façon que le lemme 4 de Kebabi et Messaci (2012) \square

Passons maintenant à la démonstration du théorème 2.1.

Démonstration. Montrons d'abord que $\lim_{n \rightarrow \infty} t_{p,n}^{x,m} = t_p^x$ p.co
La propriété de continuité de $F^{-1}(y)$ au point $F^x(t_p^x)$ peut s'écrire :

$$\forall \epsilon > 0, \exists q(\epsilon) > 0, |F^x(y) - F^x(t_p^x)| \leq q(\epsilon) \Rightarrow |y - t_p^x| < \epsilon.$$

En prenant $y = t_{p,n}^{x,m}$, on arrive finalement au résultat suivant

$$\exists p(\epsilon), \mathbb{P}[|t_{p,n}^{x,m} - t_p^x| > \epsilon] \leq \mathbb{P}[|F_n^x(t_{p,n}^{x,m}) - F^x(t_p^x)| > p(\epsilon)].$$

Parce que

$$F_n^x(t_{p,n}^{x,m}) = F^x(t_p^x) = p. \quad (2.24)$$

Depuis

$$y = t_{p,n}^{x,m}, \quad (2.25)$$

l'expression précédente peut être réécrite comme suit :

$$\exists p(\epsilon), \mathbb{P}[|t_{p,n}^{x,m} - t_p^x| > \epsilon] \leq \mathbb{P}[|F^x(y) - F_n^x(y)| > p(\epsilon)],$$

ce qui implique

$$\lim_{n \rightarrow \infty} t_{p,n}^{x,m} = t_p^x \quad p.co. \quad (2.26)$$

Le développement de Taylor de la fonction $F^x(y)$, conduit à l'existence de $t_p^{x,*}$ entre $t_{p,n}^{x,m}$ et t_p^x tel que :

$$F^x(t_{p,n}^{x,m}) - F^x(t_p^x) = (t_{p,n}^{x,m} - t_p^x) f^x(t_p^{x,*}). \quad (2.27)$$

En combinant la relation (2.26) et l'hypothèse A_3 , ensemble, il s'ensuit que

$$\lim_{n \rightarrow \infty} f^x(t_p^{x,*}) = f^x(t_p^x), \quad p.co$$

ce qui implique que

$$\exists k > 0, \sum_{n=1}^{\infty} \mathbb{P}[f^x(t_p^{x,*}) < k] < \infty. \quad (2.28)$$

Revenons à l'expression du développement limité (2.27) et en utilisant les relations (2.24), nous trouvons que

$$(t_{p,n}^{x,m} - t_p^x) = [F^x(t_{p,n}^{x,m}) - F_n^x(t_{p,n}^{x,m})] \frac{1}{f^x(t_p^{x,*})}.$$

□

La combinaison du dernier résultat avec (2.25), (2.28) et la proposition 2.3 permet de terminer la preuve du théorème 2.1.

□

Chapitre 3

Les propriétés asymptotiques de l'estimateur à noyau du mode conditionnel dans la présence des données doublement censurées

Dans ce chapitre, la problématique abordée est l'estimation non paramétrique de la fonction du mode conditionnel. Nous considérons un modèle de censure aléatoire double et nous construisons un estimateur à double noyau du mode conditionnel. Nous étudions sa convergence uniforme presque sûre sur un compact ainsi que sa normalité asymptotique. Ce travail a fait l'objet d'une publication dans la revue "Journal of Communications in Statistics - Theory and Methods" Benchoulak *et al.* (2024).

3.1 Définition de l'estimateur

Soit X une variable aléatoire réelle et soit Y une variable aléatoire non négative telle que la densité $f(x, y)$ de (X, Y) existe. Tout le long de ce chapitre, f_X (resp. $f^{(0)}(\cdot|x)$), cette notation est introduite pour faciliter l'étude représente la densité marginale de X (resp. la densité conditionnelle de Y donné $X = x$). Donc, pour tout x tel que $f_X(x) \neq 0$, on a $f^{(0)}(y|x) = \frac{f(x,y)}{f_X(x)}$. De plus, pour toute v. a. r. T , $S_T(t) = P(T > t)$ désigne la fonction de survie de T .

Dans le cas de données complètes, en se basant sur l'échantillon $(X_i, Y_i)_{1 \leq i \leq n}$ de n copies indépendantes et identiquement distribuées (i.i.d.) de (X, Y) , il est bien connu qu'un estimateur à double noyau $f_{n,c}(y|x)$ de $f^{(0)}(y|x)$ est donné par

$$f_{n,c}(y|x) = \frac{\sum_{i=1}^n \frac{1}{h_n^2} K\left(\frac{x-X_i}{h_n}\right) H\left(\frac{y-Y_i}{h_n}\right)}{\sum_{i=1}^n \frac{1}{h_n} K\left(\frac{x-X_i}{h_n}\right)}, \quad (3.1)$$

où K et H sont des fonctions noyau appropriées et h_n est une suite de nombres réels strictement positifs tendant vers zéro (voir Collomb *et al.* (1987)).

Mais, dans notre contexte, Y est doublement censuré, nous ne pouvons donc observer qu'un échantillon d'observations i.i.d. $(X_i, Z_i, \delta_i)_{1 \leq i \leq n}$ de $(X, Z := \max(\min(Y, R), L), \delta)$, où R et L sont des v.a.r. de censure non négatives telles que $L \leq R$ p. s. et

$$\delta = \begin{cases} 1 & \text{if } L \leq Y \leq R \\ 2 & \text{if } Y > R \\ 3 & \text{if } Y < L, \end{cases}$$

est l'indicateur de la censure. Notre objectif est d'estimer $f^{(0)}(y|x)$ à partir de l'échantillon $(X_i, Z_i, \delta_i)_{1 \leq i \leq n}$ et d'en déduire un estimateur pour le mode conditionnel sur un ensemble compact $G' \subset]0, \infty[$. On suppose que le mode conditionnel $\theta(x)$ est unique sur le compact G' , il est défini par

$$\theta(x) = \arg \max_{y \in G'} f^{(0)}(y|x). \quad (3.2)$$

3.2 Hypothèses et principaux résultats

Suivant Chang et Yang (1987) et Chang (1990), nous utilisons les hypothèses suivantes.

H_1 : S_Y, S_R et S_L sont des fonctions continues de t pour $t \geq 0$ et $0 < S_Y(t) < 1$ pour $t > 0$.

H_2 : $0 < P(L < t < R) = S_R(t) - S_L(t)$ pour tout $t > 0$.

H_3 : (X, Y) et (L, R) sont indépendants.

H_4 : Il existe $\alpha, \beta, 0 < \alpha < \beta < \infty$, tels que $P(L \in]0, \alpha]) = 0$ et $P(L \leq \beta) = 1$.

H_5 : $S_R(0) = S_Y(0) = 1, S_L(\infty) = S_R(\infty) = S_Y(\infty) = 0$.

Sous H_1 et H_3 , on peut facilement montrer que

$$E(1_{\{\delta=1\}} | (X, Y)) = S_R(Y) - S_L(Y). \quad (3.3)$$

Dans ce cas, pour tout B'_2 dans $\sigma(X, Y)$ (tribu engendrée par le couple (X, Y)) il existe un borélien A'_2 tel que $B'_2 = (X, Y)^{-1}(A'_2)$.

Les hypothèses H_1, H_3 et le théorème de Fubini permettent d'écrire

$$\begin{aligned}
 \int_{B'_2} 1_{\{\delta=1\}} dP &= \int_{B'_2} (1_{L < Y < R}) dP \\
 &= \int_{A'_2 \times \mathbb{R}_+^2} (1_{l < y < r}) dP_{(X,Y,L,R)} \\
 &= \int_{A'_2 \times \mathbb{R}_+^2} (1_{l < y < r}) dP_{(X,Y)} \otimes dP_{(L,R)} \\
 &= \int_{A'_2} \left(\int_{\mathbb{R}_+^2} (1_{l < y < r}) dP_{(L,R)} \right) dP_{(X,Y)} \\
 &= \int_{A'_2} P(L < y \leq R) dP_{(X,Y)} \\
 &= \int_{A'_2} (S_R(y) - S_L(y)) dP_{(X,Y)}.
 \end{aligned}$$

Ainsi, pour toute fonction $h'_2 : \mathbb{R}^2 \rightarrow \mathbb{R}$ (telle que les espérances suivantes existent), nous avons

$$\begin{aligned}
 E \left(\frac{1_{\{\delta_i=1\}} h'_2(X, Z)}{S_R(Z) - S_L(Z)} \right) &= E \left(E \left(\frac{1_{\{\delta_i=1\}} h'_2(X, Y)}{S_R(Z) - S_L(Z)} \middle/ (X, Y) \right) \right) \\
 &= E \left(\frac{h'_2(X, Y)}{S_R(Y) - S_L(Y)} E(1_{\{\delta_i=1\}} / (X, Y)) \right) \\
 &= E(h'_2(X, Y)),
 \end{aligned}$$

sous $H_1 - H_3$.

Comme $S_R - S_L$ est généralement inconnu, nous proposons d'estimer $E(h'_2(X, Y))$ par

$$\frac{1}{n} \sum_{i=1}^n \frac{1_{\{\delta_i=1\}} h'_2(X_i, Z_i)}{|S_R^{(n)}(Z_i) - S_L^{(n)}(Z_i)| + V_n}, \quad (3.4)$$

où la suite $V_n := 1/n$ sert à éviter la division par zéro et $S_R^{(n)}$ (resp. $S_L^{(n)}$) est l'estimateur self consistant de S_R (resp. S_L) introduit dans Turnbull (1974) et défini dans Ren (1997) pour tout $t \geq 0$ par les équations suivantes :

$$S_Y^{(n)} = Q^{(n)} - \int_{u \leq t} \frac{S_Y^{(n)}(t)}{S_Y^{(n)}(u)} dQ_2^{(n)}(u) + \int_{t < u} \frac{1 - S_Y^{(n)}(t)}{1 - S_Y^{(n)}(u)} dQ_3^{(n)}(u),$$

avec $\int_{u \leq t} = 0$ si $S_Y^{(n)}(t) = 0$ et $\int_{t < u} = 0$ si $S_Y^{(n)}(t) = 1$,

$$S_R^{(n)}(t) = 1 + \int_{u \leq t} \frac{dQ_2^{(n)}(u)}{S_Y^{(n)}(u)}, \quad t < B_n,$$

$$S_L^{(n)}(t) = - \int_{t < u} \frac{dQ_3^{(n)}(u)}{1 - S_Y^{(n)}(u)}, \quad t \geq A_n,$$

où $Q_j(t) = P(Z > t, \delta = j)$, $Q_j^{(n)}(t) = \frac{1}{n} \sum_{i=1}^n 1_{\{Z_i > t, \delta_i = j\}}$, $j = 1, 2, 3$,
 $Q^{(n)}(t) = \sum_{i=0}^2 Q_j^{(n)}(t)$, $A_n = \min \{Z_i; S_Y^n(Z_i) < 1\}$ et $B_n = \max \{Z_i; S_Y^n(Z_i) > 0\}$.

Il découle de Chang et Yang (1987), que sous les hypothèses $H_1 - H_3$ on a

$$P \left[\sup_{t \in [0, \infty)} |S_R^{(n)}(t) - S_R(t)| \rightarrow 0, \text{ as } n \rightarrow \infty \right] = 1 \quad (3.5)$$

et

$$P \left[\sup_{t \in [0, \infty)} |S_L^{(n)}(t) - S_L(t)| \rightarrow 0, \text{ as } n \rightarrow \infty \right] = 1. \quad (3.6)$$

D'après (3.1) et (3.4) et en fixant $H_i(y) := H\left(\frac{y-Z_i}{h_n}\right)$ et $K_i(x) := K\left(\frac{x-X_i}{h_n}\right)$, on propose comme estimateur de $f^{(0)}(y|x)$

$$f_n^{(0)}(y|x) := \frac{1}{h_n} \sum_{i=1}^n \frac{K_i(x)}{\sum_{i=1}^n K_i(x)} \frac{1_{\{\delta_i=1\}} H_i(y)}{|S_R^{(n)}(Z_i) - S_L^{(n)}(Z_i)| + V_n} = \frac{\hat{f}_{n,0}(x, y)}{f_n(x)},$$

où

$$f_n(x) := \frac{1}{nh_n} \sum_{i=1}^n K_i(x),$$

est l'estimateur par noyau de la densité $f_X(x)$ de X et

$$\hat{f}_{n,0}(x, y) := \frac{1}{nh_n^2} \sum_{i=1}^n \frac{1_{\{\delta_i=1\}} K_i(x) H_i(y)}{|S_R^{(n)}(Z_i) - S_L^{(n)}(Z_i)| + V_n}.$$

Ainsi, compte tenu de (3.2) un estimateur non paramétrique naturel $\theta_n(x)$ de $\theta(x)$ est défini comme suit

$$\theta_n(x) = \arg \max_{y \in G'} f_n^{(0)}(y|x). \quad (3.7)$$

Posons

$$\tilde{f}_{n,0}(x, y) := \frac{1}{nh_n^2} \sum_{i=1}^n \frac{1_{\{\delta_i=1\}} K_i(x) H_i(y)}{S_R(Z_i) - S_L(Z_i)}.$$

Il faut aussi introduire pour $p = 1, 2$ la p -ième dérivée partielle par rapport à y de $f^{(0)}(y|x)$ (resp. $f_n^{(0)}(y|x)$), notée $f^{(p)}(y|x)$ (resp. $f_n^{(p)}(y|x)$). Posons

$H_i^{(p)}(y)$ comme la p -ième dérivée de H au point $\frac{y-Y_i}{h_n}$ pour $p = 1, 2$, $f_n^{(p)}(y|x)$ qui est donnée par

$$f_n^{(p)}(y|x) = \frac{\hat{f}_{n,p}(x, y)}{f_n(x)}, \quad (3.8)$$

où

$$\hat{f}_{n,p}(x, y) := \frac{1}{nh_n^{p+2}} \sum_{i=1}^n \frac{1_{\{\delta_i=1\}} K_i(x) H_i^{(p)}(y)}{|S_R^{(n)}(Z_i) - S_L^{(n)}(Z_i)| + V_n}. \quad (3.9)$$

De la même manière, la p -ième dérivée partielle de $\tilde{f}_{n,0}(x, y)$ par rapport à y est donnée par

$$\tilde{f}_{n,p}(x, y) := \frac{1}{nh_n^{p+2}} \sum_{i=1}^n \frac{1_{\{\delta_i=1\}} K_i(x) H_i^{(p)}(y)}{S_R(Z_i) - S_L(Z_i)}. \quad (3.10)$$

Nous avons également besoin des hypothèses suivantes qui sont standard dans le cadre non paramétrique (voir par exemple Khardani *et al.* (2010)).

N_0 : Le noyau K est une densité bornée à support compact et telle que $\int xK(x)dx = 0$.

N_1 : Le noyau H est une densité deux fois continûment dérivable à support compact et telle que $\int xH(x)dx = 0$.

N_2 : Les ensembles $\left\{ \varphi_{x,y}(u, v) = K\left(\frac{x-u}{h_n}\right) H^{(p)}\left(\frac{y-v}{h_n}\right), x \in G, y \in G' \right\}$, où G est un ensemble compact de \mathbb{R} , sont des classes bornées Vapnik-Červonenkis (V-C) de fonctions mesurables (voir Giné et Guillou (1999)).

N_3 : $\frac{\log n}{nh_n^6} \rightarrow 0$ et $nh_n^8 \rightarrow 0$, si $n \rightarrow \infty$.

N_4 : La densité marginale $f_X(x)$ de la variable X est continûment dérivable et elle vérifie $\inf_{x \in G} f_X(x) > 0$.

N_5 : La densité jointe $f(\cdot, \cdot)$ est bornée et trois fois continûment dérivable.

Les remarques suivantes sont utiles dans la suite du chapitre.

Remarque 3.1. Il est clair que dans les expressions de $\hat{f}_{n,p}$ et $\tilde{f}_{n,p}$, $Z_i = Y_i$. Donc, comme H a un support compact (sous N_1) et $h_n \rightarrow 0$ (sous N_3), pour n assez grand il existent deux constantes $T' > 0$ et $T > T'$ telles que $\forall y \in G', \forall i = 1, n, Z_i \in [T', T]$ dans les expressions de $\hat{f}_{n,p}$ et $\tilde{f}_{n,p}$ (pour $p = 0, 2$).

Remarque 3.2. L'hypothèse N_1 implique que les dérivées $H^{(p)}$ sont bornées pour $p = 1, 2$.

Maintenant, nous énonçons dans la proposition 3.1 la convergence presque sûre uniforme de l'estimateur de la densité conditionnelle que nous avons introduit et de ses dérivées. Ensuite, la même convergence de l'estimateur du mode conditionnel $\theta_n(x)$ est donné dans le Théorème 3.1.

3.2.1 Convergence de $\theta_n(x)$

Proposition 3.1. *Sous les conditions $N_0 - N_5$ et $H_1 - H_3$, pour $p = 0 - 2$, nous avons*

$$\lim_{n \rightarrow \infty} \sup_{x \in G} \sup_{y \in G'} |f_n^{(p)}(y|x) - f^{(p)}(y|x)| = 0 \quad p.s.$$

Pour dériver le résultat suivant, nous devons ajouter une hypothèse standard qui stipule la propriété d'unicité uniforme du mode conditionnel.

$$\forall \alpha > 0, \exists b > 0 \text{ tel que pour toute fonction } \eta : G \rightarrow \mathbb{R}, \text{ on a}$$

$$\sup_{x \in G} |\theta(x) - \eta(x)| \geq \alpha \Rightarrow \sup_{x \in G} |f^{(0)}(\theta(x)|x) - f^{(0)}(\eta(x)|x)| \geq b. \quad (3.11)$$

Théorème 3.1. *Sous la condition (3.11) et les conditions de la Proposition 3.1, on obtient*

$$\lim_{n \rightarrow \infty} \sup_{x \in G} |\theta_n(x) - \theta(x)| = 0 \quad p. s.$$

Remarque 3.3. Pour le cas particulier des données censurées à droite, Ould Saïd et Cai (2005) et Khardani *et al.* (2010) ont établi des taux de convergence pour leurs estimations du mode conditionnel. Mais, à notre connaissance, il n'y a pas de taux de convergence presque sûre pour les estimations $S_R^{(n)}$ et $S_L^{(n)}$. Ainsi, nous ne nous attendons pas à dériver un taux de convergence pour notre estimateur $\theta_n(x)$.

Dans toute la suite, C représente une constante générique.

Démonstration de la Proposition 3.1

En utilisant (3.8) avec les hypothèses N_0 et $N_3 - N_5$, on peut écrire

$$\begin{aligned} \sup_{x \in G} \sup_{y \in G'} |f_n^{(p)}(y|x) - f^{(p)}(y|x)| &\leq \frac{\sup_{x \in G} \sup_{y \in G'} \{|\hat{f}_{n,p}(x, y) - \tilde{f}_{n,p}(x, y)|\}}{\inf_{x \in G} f_n(x)} \\ &+ \frac{\sup_{x \in G} \sup_{y \in G'} \{|\tilde{f}_{n,p}(x, y) - E\tilde{f}_{n,p}(x, y)| + |E\tilde{f}_{n,p}(x, y) - \frac{\partial^p f(x, y)}{\partial y^p}|\}}{\inf_{x \in G} f_n(x)} \\ &+ \frac{\sup_{x \in G} \sup_{y \in G'} \frac{\partial^p f(x, y)}{\partial y^p} \{|f_n(x) - Ef_n(x)| + |Ef_n(x) - f_X(x)|\}}{\inf_{x \in G} f_X(x) \inf_{x \in G} f_n(x)}. \end{aligned}$$

Remarquons que pour n suffisamment grand $\inf_{x \in G} f_n(x) > 0$ (voir le lemme 3.4 ci-dessous).

La preuve de cette proposition est alors une conséquence des lemmes suivants.

Lemme 3.1. *Supposons que N_0, N_1, N_3, N_5 et $H_1 - H_3$ sont réalisées. Nous avons*

$$\lim_{n \rightarrow \infty} \sup_{x \in G} \sup_{y \in G'} |\hat{f}_{n,p}(x, y) - \tilde{f}_{n,p}(x, y)| = 0 \quad \text{p. s.}$$

Démonstration. Remarquons que sous H_1 et H_2 , nous avons

$$\inf_{t \in [T', T]} (S_R(t) - S_L(t)) > 0, \quad (3.12)$$

où T et T' sont définis dans la Remarque 3.1.

De plus, compte tenu de (3.5) et de (3.6), on a pour n assez grand

$$\forall t \in [T', T], S_R^{(n)}(t) - S_L^{(n)}(t) + V_n > \frac{S_R(t) - S_L(t)}{2} \quad (\text{rappeler que } \lim_{n \rightarrow \infty} V_n = 0).$$

Ce terme, combiné avec (3.9), (3.10), (3.12) et Remarque 3.1, implique que

$$\begin{aligned} \sup_{x \in G} \sup_{y \in G'} |\hat{f}_{n,p}(x, y) - \tilde{f}_{n,p}(x, y)| &\leq 2 \frac{\sup_{t \in [T', T]} |(S_R^{(n)}(t) - S_R(t)) + (S_L^{(n)}(t) - S_L(t))| + V_n}{(\inf_{t \in [T', T]} (S_R(t) - S_L(t))^2} \\ &\quad \times \sup_{x \in G} \sup_{y \in G'} \left| \sum_{i=1}^n \frac{K_i(x) H_i^{(p)}(y)}{nh_n^{p+2}} \right|. \end{aligned} \quad (3.13)$$

En vertu des relations (3.5), (3.6) et du fait que $\lim_{n \rightarrow \infty} V_n = 0$ et que

$$\sup_{x \in G} \sup_{y \in G'} \left| \frac{\partial^p f(x, y)}{\partial y^p} \right| < \infty \quad (\text{sous } N_5), \text{ il suffit de prouver que}$$

$$\lim_{n \rightarrow \infty} \sup_{x \in G} \sup_{y \in G'} |f_{n,p}(x, y) - \frac{\partial^p f(x, y)}{\partial y^p}| = 0 \quad \text{p.s, où}$$

$$f_{n,p}(x, y) := \frac{1}{nh_n^{p+2}} \sum_{i=1}^n K_i(x) H_i^{(p)}(y). \text{ A cette fin, nous écrivons}$$

$$f_{n,p}(x, y) - \frac{\partial^p f(x, y)}{\partial y^p} = [f_{n,p}(x, y) - E f_{n,p}(x, y)] + [E f_{n,p}(x, y) - \frac{\partial^p f(x, y)}{\partial y^p}] := W_1^{(p)} + W_2^{(p)}, \text{ où } W_1^{(p)} = f_{n,p}(x, y) - E f_{n,p}(x, y) \text{ et } W_2^{(p)} = E f_{n,p}(x, y) - \frac{\partial^p f(x, y)}{\partial y^p}.$$

• D'une part, nous avons

$$\lim_{n \rightarrow \infty} \sup_{x \in G} \sup_{y \in G'} |W_2^{(p)}| = 0, \quad (3.14)$$

En vertu du Lemme 3.2(ci-dessous) si on prend la constante 1 au lieu de la variable $\frac{1_{\{\delta_i=1\}}}{S_R(Z_i)-S_L(Z_i)}$.

• D'autre part, le fait que $\lim_{n \rightarrow \infty} \sup_{x \in G} \sup_{y \in G'} |W_1^{(p)}| = 0$ p.s.découle de la preuve du lemme 3.3(ci-dessous), en utilisant la constante 1 au lieu de la variable $\frac{1_{\{\delta_i=1\}}}{S_R(Z_i)-S_L(Z_i)}$. \square

Lemme 3.2. *Sous les hypothèses N_0, N_1, N_3, N_5, H_1 et H_3 , nous avons pour $p = 0 - 2$.*

$$\lim_{n \rightarrow \infty} \sup_{x \in G} \sup_{y \in G'} \left| E \tilde{f}_{n,p}(x, y) - \frac{\partial^p f(x, y)}{\partial y^p} \right| = 0.$$

Démonstration. Comme (X_i, Z_i, δ_i) sont identiquement distribués, $\sup_{y \in G'} \left| \frac{\partial^p f(x, y)}{\partial y^p} \right| < \infty$ (En vertu de N_5), nous avons

$$\left| E \tilde{f}_{n,p}(x, y) - \frac{\partial^p f(x, y)}{\partial y^p} \right| = \left| E \left[\frac{K_1(x)}{h_n^{p+2}} E \left(\frac{1_{\{\delta_1=1\}} H_1^{(p)}(y)}{S_R(Y_1) - S_L(Y_1)} \middle| X_1 \right) \right] - \frac{\partial^p f(x, y)}{\partial y^p} \right|. \quad (3.15)$$

Il faut d'abord calculer le terme $\frac{1}{h_n^p} E \left(\frac{1_{\{\delta_1=1\}} H_1^{(p)}(y)}{S_R(Y_1) - S_L(Y_1)} \middle| X_1 \right)$.

En utilisant (3.3) et en faisant p intégrations successives par parties, on obtient

$$\begin{aligned} \frac{1}{h_n^p} E \left(\frac{1_{\{\delta_1=1\}} H_1^{(p)}(y)}{S_R(Y_1) - S_L(Y_1)} \middle| X_1 \right) &= \frac{1}{h_n^p} E \left[\frac{H_1^{(p)}(y)}{S_R(Y_1) - S_L(Y_1)} E(1_{\{\delta_1=1\}} | (X_1, Y_1)) \middle| X_1 \right] \\ &= \frac{1}{h_n^p} \int H^{(p)} \left(\frac{y-v}{h_n} \right) f(v | X_1) dv \\ &= \int H \left(\frac{y-v}{h_n} \right) f^{(p)}(v | X_1) dv. \end{aligned}$$

En combinant cela avec la relation (3.15) et en utilisant un changement de variable, on obtient

$$|E\tilde{f}_{n,p}(x, y) - \frac{\partial^p f(x, y)}{\partial y^p}| = \left| \int \int K(z)H(w) \frac{\partial^p f(x-zh_n, y-wh_n)}{\partial y^p} dzdw - \frac{\partial^p f(x, y)}{\partial y^p} \right|.$$

Ainsi, sous N_0, N_1 et N_5 , par le développement de Taylor, il exist (x^*, y^*) à l'intérieur du segment qui relie (x, y) et $(x - zh_n, y - wh_n)$ tel que

$$\begin{aligned} & \sup_{x \in G} \sup_{y \in G'} \left| E\tilde{f}_{n,p}(x, y) - \frac{\partial^p f(x, y)}{\partial y^p} \right| \\ &= \sup_{x \in G} \sup_{y \in G'} \left| \int \int K(z)H(w) \left(\frac{\partial^p f(x - zh_n, y - wh_n)}{\partial y^p} - \frac{\partial^p f(x, y)}{\partial y^p} \right) dzdw \right| \\ &\leq \sup_{x \in G} \sup_{y \in G'} \left| \int \int K(z)H(w) \left(-zh_n \frac{\partial^{p+1} f(x^*, y^*)}{\partial x \partial y^p} - wh_n \frac{\partial^{p+1} f(x^*, y^*)}{\partial y^{p+1}} \right) dzdw \right| \\ &= O(h_n) \rightarrow 0, \quad \text{as } n \rightarrow \infty, \end{aligned} \tag{3.16}$$

en vertu de N_3 . □

Lemme 3.3. Sous $N_0, N_1, N_2, N_3, N_5, H_1$ et H_2 , nous avons pour $p=0-2$

$$\sup_{x \in G} \sup_{y \in G'} |\tilde{f}_{n,p}(x, y) - E\tilde{f}_{n,p}(x, y)| = O \left(\sqrt{\frac{\log n}{nh_n^{2p+2}}} \right) \quad p.s.$$

Démonstration. Considérons pour $p = 0, 2$ la famille d'ensembles suivante

$$\mathfrak{F}_n = \left\{ \psi_{x,y}(u, v, l, r) = \frac{1}{nh_n^{p+2}} \frac{1_{\{\ell \leq v \leq r\}}}{S_R((v \wedge r) \vee \ell) - S_L((v \wedge r) \vee \ell)} \times K\left(\frac{x-u}{h_n}\right) H^{(p)}\left(\frac{y-v}{h_n}\right), \right. \\ \left. x \in G, y \in G' \right\}.$$

\mathfrak{F}_n sont des classes V-C bornées grâce à l'hypothèse N_2 et au lemme 2.6.18 de Van Der Vaart et Wellner (1996).

De plus, nous avons

$$\sup_{\psi_{x,y} \in \mathfrak{F}_n} \|\psi_{x,y}\|_\infty \leq \frac{\|K\|_\infty \|H^{(p)}\|_\infty}{nh_n^{p+2} \inf_{t \in [T', T]} (S_R(t) - S_L(t))} =: U_n$$

(où $\|\psi\|_\infty := \sup_{x \in \mathbb{R}} |\psi(x)|$ est la norme du sup)

et

$$E(\psi_{x,y}^2(X, Y, L, R)) \leq \frac{1}{n^2 h_n^{2p+4}} \int \int K^2\left(\frac{x-u}{h_n}\right) \left[H^{(p)}\left(\frac{y-v}{h_n}\right) \right]^2 \frac{f(u, v) du dv}{(S_R(v) - S_L(v))^2} \\ = \frac{1}{n^2 h_n^{2p+2}} \int \int \frac{K^2(z) [H^{(p)}(w)]^2 f(x - zh_n, y - wh_n)}{(S_R(y - wh_n) - S_L(y - wh_n))^2} dz dw \\ \leq \frac{1}{n^2 h_n^{2p+2}} \frac{\|K\|_2^2 \|H^{(p)}\|_2^2 \|f\|_\infty^2}{(\inf_{t \in [T', T]} (S_R(t) - S_L(t))^2} =: \sigma_n^2$$

avec $\sigma_n \leq U_n$ pour n assez grand.

Donc l'inégalité de Talagrand (voir Giné et Guillou (2001), proposition 2.2) donne

$$P \left[\sup_{\psi_{x,y} \in \mathfrak{F}_n} \left| \sum_{i=1}^n (\psi_{x,y}(X_i, Y_i, L_i, R_i) - E(\psi_{x,y}(X, Y, L, R))) \right| > \eta \right] \\ \leq C \exp \left\{ -\frac{1}{C} \frac{\eta}{U_n} \log \left(1 + \frac{\eta U_n}{C (\sqrt{n} \sigma_n + U_n \sqrt{\log(AU_n/\sigma_n)})^2} \right) \right\}$$

pour

$$\eta \geq B \left[U_n \log \left(\frac{AU_n}{\sigma_n} \right) + \sqrt{n} \sigma_n \sqrt{\log \left(\frac{AU_n}{\sigma_n} \right)} \right], \quad (3.17)$$

où A, B et C sont des constantes positives.

Ainsi, en choisissant $\eta = D \sqrt{\frac{\log n}{nh_n^{2p+2}}}$, avec D est une constante positive (ce choix de η vérifie (3.17) pour n assez grand), on a

$$\begin{aligned} & P \left(\sup_{x \in G} \sup_{y \in G'} |\tilde{f}_{n,p}(x, y) - E\tilde{f}_{n,p}(x, y)| > D \sqrt{\frac{\log n}{nh_n^{2p+2}}} \right) \\ & \leq C \exp \left\{ \frac{-D \sqrt{\frac{\log n}{nh_n^{2p+2}}} nh_n^{p+2} \inf_{t \in [T', T]} (S_R(t) - S_L(t))}{C \|K\|_\infty \|H^{(p)}\|_\infty} \times \right. \\ & \quad \left. \log \left[1 + \frac{\frac{D \sqrt{\frac{\log n}{nh_n^{2p+2}}} \|K\|_\infty \|H^{(p)}\|_\infty}{nh_n^{p+2} \inf_{t \in [T', T]} (S_R(t) - S_L(t))}}{C \left(\frac{\|K\|_2 \|H^{(p)}\|_2 \sqrt{\|f\|_\infty}}{\sqrt{nh_n^{2p+2}} \inf_{t \in [T', T]} (S_R(t) - S_L(t))} + \frac{\|K\|_\infty \|H^{(p)}\|_\infty \times \sqrt{\log \left(\frac{A \|K\|_\infty \|H^{(p)}\|_\infty}{h_n \|K\|_2 \|H^{(p)}\|_2 \sqrt{\|f\|_\infty} \right)}}{nh_n^{p+2} \inf_{t \in [T', T]} (S_R(t) - S_L(t))} \right)^2} \right] \right\}. \end{aligned}$$

Puisque $\log(1 + u) \sim u$ quand $u \rightarrow 0$, le terme à droite est de l'ordre de

$$\begin{aligned}
 & C \exp \left\{ \frac{-D\sqrt{n \log n} h_n \inf_{t \in [T', T]} (S_R(t) - S_L(t))}{C \|K\|_\infty \|H^{(p)}\|_\infty} \times \frac{D\sqrt{\frac{\log n}{n^3 h_n^{p+6}}} \|K\|_\infty \|H^{(p)}\|_\infty}{\inf_{t \in [T', T]} (S_R(t) - S_L(t))} \right. \\
 & \left. \times \frac{nh_n^6 \left(\inf_{t \in [T', T]} (S_R(t) - S_L(t)) \right)^2}{C \|K\|_2^2 \|H^{(p)}\|_2^2 \|f\|_\infty} \right\} \\
 & = Cn^{-\frac{D^2 (\inf_{t \in [T', T]} (S_R(t) - S_L(t)))^2}{C^2 \|K\|_2^2 \|H^{(p)}\|_2^2 \|f\|_\infty}}.
 \end{aligned}$$

Par un choix approprié de D , on peut faire de ce terme $O(n^{-3/2})$ qui est un terme général d'une série convergente, donc par le lemme de Borel-Cantelli, on en déduit le résultat revendiqué. \square

Lemme 3.4. *Sous les hypothèses N_0, N_3 et N_4 , on a*

- i) $\lim_{n \rightarrow \infty} \sup_{x \in G} |Ef_n(x) - f_X(x)| = 0$, ii) $\lim_{n \rightarrow \infty} \sup_{x \in G} |f_n(x) - Ef_n(x)| = 0$ p.s et*
- iii) $\exists \theta > 0$ tel que $\inf_{x \in G} f_n(x) > \theta$ pour n assez grand.*

Démonstration. La preuve de *i)* (resp. *ii)*) est semblable à celle du lemme 3.2 (resp. Le lemme 3.3), en utilisant la constante 1 au lieu de la variable $\frac{1_{\{\delta_i=1\}} H_i(y)}{h_n S_R(Y_i) - S_L(Y_i)}$ et en utilisant N_4 au lieu de N_5 .

Il reste à prouver *iii)*.

De *i)* et *ii)*, on déduit immédiatement que $\lim_{n \rightarrow \infty} \sup_{x \in G} |f_n(x) - f_X(x)| = 0$. p.s. Donc, $\exists n_0 \in \mathbb{N}$ tel que $\forall n \geq n_0 \sup_{x \in G} |f_n(x) - f_X(x)| < \frac{\inf_{x \in G} f_X(x)}{2}$. Cela implique que $\forall n \geq n_0 \inf_{x \in G} f_n(x) > \frac{\inf_{x \in G} f_X(x)}{2}$. \square

Démonstration du Theorem 3.1. Cette preuve est classique mais, du fait de sa brièveté, nous choisissons de la rappeler. Nous avons

$$\begin{aligned} & \sup_{x \in G} |f^{(0)}(\theta_n(x)|x) - f^{(0)}(\theta(x)|x)| \leq \sup_{x \in G} |f^{(0)}(\theta_n(x)|x) - f_n^{(0)}(\theta_n(x)|x)| \\ & + \sup_{x \in G} |f_n^{(0)}(\theta_n(x)|x) - f^{(0)}(\theta(x)|x)| \\ & \leq 2 \sup_{x \in G} \sup_{y \in G'} |f_n^{(0)}(y|x) - f^{(0)}(y|x)|. \end{aligned}$$

Il reste à appliquer (3.11) et la proposition 3.1 avec $p = 0$ pour obtenir le résultat annoncé. □

3.2.2 La normalité asymptotique de $\theta_n(x)$

Dans cette sous section, nous établissons la normalité asymptotique de l'estimateur à noyau du mode conditionnel par ce théorème

Théorème 3.2. Soit $x \in \mathbb{R}$, si $f^{(2)}(\theta(x)|x) < 0$, sous les conditions $H_1 - H_5$, $N_0 - N_3$ et N_5 on a

$$\sqrt{nh_n^4}(\theta_n(x) - \theta(x)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2(x)),$$

où $\xrightarrow{\mathcal{D}}$ désigne la convergence en loi et

$$\sigma^2(x) = \frac{f(x, \theta(x))}{[S_R(\theta(x)) - S_L(\theta(x))] \left(\frac{\partial^2}{\partial y^2} f(x, \theta(x)) \right)^2} \|K\|_2^2 \|H^{(1)}\|_2^2,$$

avec $\|\psi\|_2 := (\int \psi^2(x) dx)^{1/2}$ pour toute fonction ψ de carré intégrale.

Ce résultat généralise le théorème (3.7) de Khardani *et al.* (2010) au cas des données doublement censurées.

Démonstration. Un développement de Taylor de $f_n^{(1)}(\cdot|x)$ en $\theta(x)$ autour de $\theta_n(x)$, combiné avec la relation (3.8), permet d'écrire la décomposition suivante.

$$\begin{aligned} \sqrt{nh_n^4}(\theta_n(x) - \theta(x)) &= \sqrt{nh_n^4} \frac{\hat{f}_{n,1}(x, \theta(x)) - \tilde{f}_{n,1}(x, \theta(x))}{\hat{f}_{n,2}(x, \tilde{\theta}_n(x))} \\ &+ \sqrt{nh_n^4} \frac{\tilde{f}_{n,1}(x, \theta(x)) - E\tilde{f}_{n,1}(x, \theta(x))}{\hat{f}_{n,2}(x, \tilde{\theta}_n(x))} \\ &+ \sqrt{nh_n^4} \frac{E\tilde{f}_{n,1}(x, \theta(x))}{\hat{f}_{n,2}(x, \tilde{\theta}_n(x))}, \end{aligned}$$

où $\tilde{\theta}_n(x)$ est entre $\theta_n(x)$ et $\theta(x)$.

Alors la preuve de ce théorème est une conséquence directe des étapes suivantes.

Etape 1 Nous prouvons que $\sqrt{nh_n^4}|\hat{f}_{n,1}(x, \theta(x)) - \tilde{f}_{n,1}(x, \theta(x))| \xrightarrow{\mathcal{P}} 0$,

où $\xrightarrow{\mathcal{P}}$ désigne la convergence en probabilité.

En procédant comme dans la relation (3.13), pour $p = 1$, on obtient

$$\begin{aligned} \sqrt{nh_n^4}|\hat{f}_{n,1}(x, \theta(x)) - \tilde{f}_{n,1}(x, \theta(x))| &\leq 2\sqrt{h_n^3} \left| \frac{1}{nh_n^3} \sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right) H^{(1)}\left(\frac{\theta(x) - Y_i}{h_n}\right) \right| \\ &\times \frac{\sqrt{n}(\sup_{t \in [T', T]} |(S_R^{(n)}(t) - S_L^{(n)}(t)) - (S_R(t) - S_L(t))| + V_n)}{(\inf_{t \in [T', T]} (S_R(t) - S_L(t))^2)}. \end{aligned} \quad (3.18)$$

Sous les hypothèses H_1, H_2, H_4 et H_5 , d'après le Théorème 3.1 de Chang (1990)

$$\sqrt{n} \sup_{t \in [T', T]} |(S_R^{(n)}(t) - S_L^{(n)}(t)) - (S_R(t) - S_L(t))| = O_p(1) \quad (3.19)$$

et nous avons $\sqrt{n}V_n = \frac{1}{\sqrt{n}} \rightarrow 0$, si $n \rightarrow \infty$.

Il reste à traiter le terme $\frac{1}{nh_n^3} \sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right) H^{(1)}\left(\frac{\theta(x)-Y_i}{h_n}\right)$.

La relation (3.14) pour $p=1$ donne

$$E \left[\frac{1}{nh_n^3} \sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right) H^{(1)}\left(\frac{\theta(x)-Y_i}{h_n}\right) \right] \rightarrow \frac{\partial f(x, \theta(x))}{\partial y} = 0.$$

De plus, nous avons

$$\begin{aligned} & Var \left[\frac{1}{nh_n^3} \sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right) H^{(1)}\left(\frac{\theta(x)-Y_i}{h_n}\right) \right] \\ &= \frac{1}{nh_n^6} Var \left[K\left(\frac{x-X_1}{h_n}\right) H^{(1)}\left(\frac{\theta(x)-Y_1}{h_n}\right) \right] \\ &\leq \frac{1}{nh_n^6} E \left[K^2\left(\frac{x-X_1}{h_n}\right) \left[H^{(1)}\left(\frac{\theta(x)-Y_1}{h_n}\right) \right]^2 \right] \\ &= \frac{1}{nh_n^6} \int \int K^2\left(\frac{x-u}{h_n}\right) \left[H^{(1)}\left(\frac{\theta(x)-v}{h_n}\right) \right]^2 f(u, v) dudv \\ &= \frac{1}{nh_n^6} \int \int K^2\left(\frac{z}{h_n}\right) \left[H^{(1)}\left(\frac{w}{h_n}\right) \right]^2 f(x-z, \theta(x)-w) dzdw. \end{aligned}$$

On peut montrer comme dans la preuve du théorème de Bochner (voir théorème 1A de Parzen (1962)) que

$$\begin{aligned} & \lim_{n \rightarrow \infty} \frac{1}{h_n^2} \int \int K^2\left(\frac{z}{h_n}\right) \left[H^{(1)}\left(\frac{w}{h_n}\right) \right]^2 f(x-z, \theta(x)-w) dzdw \\ &= f(x, \theta(x)) \int \int K^2(z) [H^{(1)}(w)]^2 dzdw. \end{aligned}$$

et puisque $nh_n^4 \rightarrow \infty$, on obtient

$$\lim_{n \rightarrow \infty} \text{Var} \left[\frac{1}{nh_n^3} \sum_{i=1}^n K \left(\frac{x-X_i}{h_n} \right) H^{(1)} \left(\frac{\theta(x)-Y_i}{h_n} \right) \right] = 0.$$

Ainsi $\frac{1}{nh_n^3} \sum_{i=1}^n K \left(\frac{x-X_i}{h_n} \right) H^{(1)} \left(\frac{\theta(x)-Y_i}{h_n} \right)$ converge en moyenne quadratique vers 0. Par conséquent

$$\frac{1}{nh_n^3} \sum_{i=1}^n K \left(\frac{x-X_i}{h_n} \right) H^{(1)} \left(\frac{\theta(x)-Y_i}{h_n} \right) \xrightarrow{\mathcal{P}} 0. \text{ En combinant cela avec (3.18) et (3.19), on obtient que } \sqrt{nh_n^4} |\hat{f}_{n,1}(x, \theta(x)) - \tilde{f}_{n,1}(x, \theta(x))| \xrightarrow{\mathcal{P}} 0.$$

Etape 2 Nous prouvons que $\lim_{n \rightarrow \infty} \sqrt{nh_n^4} E\tilde{f}_{n,1}(x, \theta(x)) = 0$.

En procédant comme dans (3.16) on obtient

$$\begin{aligned} \sqrt{nh_n^4} E\tilde{f}_{n,1}(x, \theta(x)) &= \sqrt{nh_n^4} \left(E\tilde{f}_{n,1}(x, \theta(x)) - \frac{\partial f}{\partial y}(x, \theta(x)) \right) \\ &= \sqrt{nh_n^4} \int \int K(z)H(w) \left[\frac{\partial f}{\partial y}(x - zh_n, \theta(x) - wh_n) - \frac{\partial f(x, \theta(x))}{\partial y} \right] dzdw \\ &= \sqrt{nh_n^4} \int \int K(z)H(w) \left[-zh_n \frac{\partial^2 f(x, \theta(x))}{\partial x \partial y} - wh_n \frac{\partial^2 f(x, \theta(x))}{\partial y^2} + \frac{z^2 h_n^2}{2} \frac{\partial^3 f(x^*, y^*)}{\partial x^2 \partial y} \right. \\ &\quad \left. + \frac{w^2 h_n^2}{2} \frac{\partial^3 f(x^*, y^*)}{\partial y^3} + zwh_n^2 \frac{\partial^3 f(x^*, y^*)}{\partial x \partial y^2} \right] dzdw, \end{aligned}$$

où (x^*, y^*) compris entre (x, y) et $(x - zh_n, y - wh_n)$.

Donc les hypothèses N_0, N_1, N_3 et N_5 donnent

$$\sqrt{nh_n^4} E\tilde{f}_{n,1}(x, \theta(x)) = O(\sqrt{nh_n^8}) \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Etape 3 Nous prouvons que $\sqrt{nh_n^4} [\tilde{f}_{n,1}(x, \theta(x)) - E\tilde{f}_{n,1}(x, \theta(x))]$ est asymptotiquement normal.

Nous avons $\sqrt{nh_n^4}[\tilde{f}_{n,1}(x, \theta(x)) - E\tilde{f}_{n,1}(x, \theta(x))] = \sum_{i=1}^n L_{i,n}$,

où

$$L_{i,n} = \frac{1}{\sqrt{nh_n^2}} \left(\frac{1_{\{\delta_i=1\}} H_i^{(1)}(\theta(x)) K_i(x)}{S_R(Z_i) - S_L(Z_i)} \right) - \frac{1}{\sqrt{nh_n^2}} E \left(\frac{1_{\{\delta_i=1\}} H_i^{(1)}(\theta(x)) K_i(x)}{S_R(Z_i) - S_L(Z_i)} \right).$$

Il suffit donc d'appliquer le théorème central limite de Lindeberg, pour cela il faut vérifier que pour tout $\epsilon > 0$

$$\lim_{n \rightarrow \infty} \frac{1}{\text{Var}(\sum_{i=1}^n L_{i,n})} \sum_{i=1}^n E \left(L_{i,n}^2 1_{\{|L_{i,n}| > \epsilon \sqrt{\text{Var}(\sum_{i=1}^n L_{i,n})}\}} \right) = 0.$$

En vertu de la bornitude de $H^{(1)}$ et K (voir remarque 3.2 et l'hypothèse N_0) et la relation (3.12), on a

$$|L_{1,n}| \leq \frac{C}{\sqrt{nh_n^2} \inf_{t \in [T', T]} (S_R(t) - S_L(t))}.$$

Ainsi, l'inégalité de Chebyshev-Bienayme permet d'écrire

$$\begin{aligned} & \frac{1}{\text{Var}(\sum_{i=1}^n L_{i,n})} \sum_{i=1}^n E \left(L_{i,n}^2 1_{\{|L_{i,n}| > \epsilon \sqrt{\text{Var}(\sum_{i=1}^n L_{i,n})}\}} \right) \\ & \leq \frac{C}{\epsilon^2 (\inf_{t \in [T', T]} (S_R(t) - S_L(t)))^2 nh_n^2 \text{Var}(\sqrt{nh_n^4} \tilde{f}_{n,1}(x, \theta(x)))}. \end{aligned}$$

Il reste à calculer $\lim_{n \rightarrow \infty} \text{Var}(\sqrt{nh_n^4} \tilde{f}_{n,1}(x, \theta(x)))$. Nous avons

$$\begin{aligned}
 \text{Var}(\sqrt{nh_n^4} \tilde{f}_{1,n}(x, \theta(x))) &= n \text{Var}(L_{1,n}) \\
 &= \frac{1}{nh_n^2} \text{Var} \left(\frac{1_{\{\delta_1=1\}} K_1(x) H_1^{(1)}(\theta(x))}{S_R(Z_1) - S_L(Z_1)} \right) \\
 &= \frac{1}{h_n^2} E \left(\frac{1_{\{\delta_1=1\}} K_1^2(x) (H_1^{(1)})^2(\theta(x))}{(S_R(Z_1) - S_L(Z_1))^2} \right) \\
 &\quad - \frac{1}{h_n^2} E^2 \left(\frac{1_{\{\delta_1=1\}} K_1(x) H_1^{(1)}(\theta(x))}{S_R(Z_1) - S_L(Z_1)} \right) =: A_1 - A_2.
 \end{aligned}$$

D'une part, la relation (3.16) pour $p = 1$, l'hypothèse N_3 et le fait que $\frac{\partial f(x, \theta(x))}{\partial y} = 0$ permettant d'écrire

$$A_2 = \frac{1}{h_n^2} E^2 \left(\frac{1_{\{\delta_1=1\}} K_1(x) H_1^{(1)}(\theta(x))}{S_R(Z_1) - S_L(Z_1)} \right) = h_n^4 E^2 \tilde{f}_{n,1}(x, \theta(x)) \rightarrow 0.$$

De plus, en utilisant (3.3) et un changement de variable, on obtient

$$\begin{aligned}
 A_1 &= E \left[\frac{K_1(x)^2}{h_n^2} E \left(\frac{1_{\{\delta_1=1\}} \left(H_1^{(1)} \left(\frac{\theta(x) - Y_1}{h_n} \right) \right)^2}{(S_R(Y_1) - S_L(Y_1))^2} \middle| X_1 \right) \right] \\
 &= \frac{1}{h_n^2} \int K^2 \left(\frac{x - u}{h_n} \right) \int \frac{\left(H^{(1)} \left(\frac{\theta(x) - v}{h_n} \right) \right)^2}{(S_R(v) - S_L(v))^2} f(v|X_1) f_X(u) du dv \\
 &= \frac{1}{h_n^2} \int \int \frac{K^2 \left(\frac{z}{h_n} \right) \left(H^{(1)} \left(\frac{w}{h_n} \right) \right)^2}{S_R(\theta(x) - w) - S_L(\theta(x) - w)} f(x - z, \theta(x) - w) dz dw.
 \end{aligned}$$

Soit $a = \min(G')$ et $b = \max(G')$ ($a > 0$ car $G' \subset]0, \infty[$), puisque $H^{(1)}$ a un support compact, il existe $\nu > 0$ tel que $-\nu \leq \frac{w}{h_n} \leq \nu$, donc $-\nu h_n \leq w \leq \nu h_n$. Alors, pour n assez grand, on a $-1 \leq w \leq \frac{a}{2}$ et $(\theta(x) - w) \in [\frac{a}{2}, b + 1]$.

Ainsi, $A_1 = \frac{1}{h_n^2} \int \int K^2\left(\frac{z}{h_n}\right) \left(H_1^{(1)}\left(\frac{w}{h_n}\right)\right)^2 \phi(x-z, \theta(x)-w) dz dw$, où $\phi(u, v) = \frac{f(u, v)}{S_R(v) - S_L(v)} 1_{[\frac{a}{2} \leq v \leq b+1]}$.

À partir de là nous pouvons montrer comme dans le théorème de Bochner (voir le théorème 1A de Parzen (1962)) que

$$\lim_{n \rightarrow \infty} A_1 = \frac{f(x, \theta(x))}{S_R(\theta(x)) - S_L(\theta(x))} \int \int K^2(z) (H^{(1)}(w))^2 dz dw.$$

Etape 4 Dans cette étape, montrons que $\hat{f}_{n,2}(x, \tilde{\theta}_n(x)) \xrightarrow{\mathcal{P}} \frac{\partial^2 f(x, \theta(x))}{\partial y^2}$.

En effet, notons que

$$\begin{aligned} \left| \hat{f}_{n,2}(x, \tilde{\theta}_n(x)) - \frac{\partial^2 f(x, \theta(x))}{\partial y^2} \right| &\leq \left| \hat{f}_{n,2}(x, \tilde{\theta}_n(x)) - \frac{\partial^2 f(x, \tilde{\theta}_n(x))}{\partial y^2} \right| \\ &\quad + \left| \frac{\partial^2 f(x, \tilde{\theta}_n(x))}{\partial y^2} - \frac{\partial^2 f(x, \theta(x))}{\partial y^2} \right|. \end{aligned}$$

Pour le deuxième terme, puisque $\theta_n(x) \xrightarrow{\mathcal{P}} \theta(x)$ (grâce au théorème 3.1 pour $G = \{x\}$) et puisque $\tilde{\theta}_n(x)$ est entre $\theta_n(x)$ et $\theta(x)$, nous avons $\tilde{\theta}_n(x) \xrightarrow{\mathcal{P}} \theta(x)$.

Donc, $\frac{\partial^2 f(x, \tilde{\theta}_n(x))}{\partial y^2} \xrightarrow{\mathcal{P}} \frac{\partial^2 f(x, \theta(x))}{\partial y^2}$ par la continuité de $\frac{\partial^2 f}{\partial y^2}$.

Pour le premier terme, nous avons

$$\begin{aligned} \sup_{y \in G'} \left| \hat{f}_{n,2}(x, y) - \frac{\partial^2 f(x, y)}{\partial y^2} \right| &\leq \sup_{y \in G'} \left| \hat{f}_{n,2}(x, y) - \tilde{f}_{n,2}(x, y) \right| + \sup_{y \in G'} \left| \tilde{f}_{n,2}(x, y) - E \tilde{f}_{n,2}(x, y) \right| \\ &\quad + \sup_{y \in G'} \left| E \tilde{f}_{n,2}(x, y) - \frac{\partial^2 f(x, y)}{\partial y^2} \right|. \end{aligned}$$

i. En vertu du Lemme 3.1 pour $p=2$, on en déduit que

$$\lim_{n \rightarrow \infty} \sup_{y \in G'} |\hat{f}_{n,2}(x, y) - \tilde{f}_{n,2}(x, y)| = 0 \text{ p.s.} \quad (3.20)$$

ii. En vertu du Lemme 3.3 avec $p=2$, on obtient

$$\lim_{n \rightarrow \infty} \sup_{y \in G'} |\tilde{f}_{n,2}(x, y) - E\tilde{f}_{n,2}(x, y)| = 0 \text{ p.s.} \quad (3.21)$$

iii. Enfin, en appliquant le Lemme 3.2 avec $p = 2$, on obtient

$$\lim_{n \rightarrow \infty} \sup_{y \in G'} \left| E\tilde{f}_{n,2}(x, y) - \frac{\partial^2 f(x, y)}{\partial y^2} \right| = 0, \text{ p.s.}$$

En combinant cela avec (3.20) et (3.21), on peut voir que

$\sup_{y \in G'} \left| \hat{f}_{n,2}(x, y) - \frac{\partial^2 f(x, y)}{\partial y^2} \right| \xrightarrow{\mathcal{P}} 0$ quand $n \rightarrow \infty$. En combinant cette convergence avec les étapes 1, 2 et 3, nous obtenons le résultat voulu.

□

Chapitre 4

Simulation et données réelles

Dans ce chapitre nous nous intéressons à la présentation de nos études des prévisions mais cette fois sur des données générées et finies pour les deux phénomènes de censure mixte et double .

Dans un premier temps, nous commençons par étudier le comportement de l'estimateur de la fonction de la médiane conditionnelle qui est un cas particulier de l'estimateur de la fonction du quantile conditionnel (on met $p=1/2$ dans la relation (2.12)) et par comparer sa qualité avec la vraie fonction de médiane pour certains modèles .

Dans un second temps, nous nous concentrons sur la prévision via le mode conditionnel sur des données générées finies afin d' examiner la bonne performance de son estimateur. En outre nous réalisons des simulations sur la normalité asymptotique sur des échantillons finis pour illustrer nos études. En fin nous appliquons notre estimateur (3.7) sur des données réelles pour confirmer nos résultats théoriques.

4.1 Etude numérique I

Dans cette section, nous allons illustrer nos résultats théoriques par une étude de simulation sur des données générées qui est menée pour évaluer la performance de l'estimateur non paramétrique de la fonction de la médiane conditionnelle comme annoncé avant. On génère n points selon les deux modèles connus suivants.

Le modèle linéaire **M1** : $Y = X + 0,2\epsilon$,

Le modèle non linéaire **M2** : $Y = \exp(0,5X - 0,9) + 0,2\epsilon$ où les v.a.r. X et ϵ sont générés suivant une loi exponentielle de paramètre $1/3$ et une loi normale standard respectivement.

Nous prenons un noyau d'Epanechnikov pour K et la fonction de distribution $H(x) = \int_{-\infty}^x \frac{3}{4}(1 - z^2)1_{[0,1]}(z)dz$. En ce qui concerne la fenêtre, cette dernière est sélectionnée par minimisation de la distance maximale entre l'estimateur et la fonction de la médiane conditionnelle théorique sur un ensemble de valeurs de fenêtre dans $h_n \in [0, 0.3]$.

Les variables de censures sont générés suivant des lois exponentielle, les paramètres de ces lois γ_1 et γ_2 sont choisis de façon à ajuster le pourcentage de censure (CR) et on a ajouté 0.01 à la variable de censure à droite pour confirmer l'hypothèse N_0 .

1) Effet de la taille d'échantillon : Grâce aux figures (4.1) et (4.2), il est possible de vérifier les performances de l'estimateur non paramétrique de la fonction médiane conditionnelle pour un taux de censure fixe, nous pouvons observer une amélioration de la qualité d'ajustement à la droite théorique lorsque n augmente. Et la même remarque pour le modèle non linéaire.

2) Effet du taux de censure : Les figures 4.3 et 4.4 illustrent la simulation de l'estimateur non paramétrique de la fonction médiane conditionnelle pour une taille d'échantillon fixe et une variation du taux de censure, nous pou-

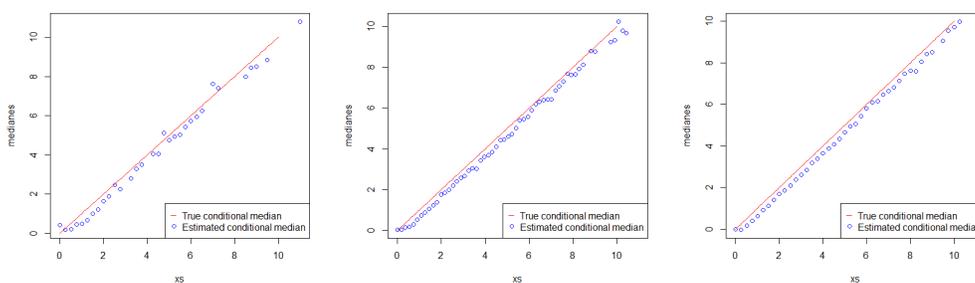


FIGURE 4.1 – **M1**, $CR \simeq 30\%$ et de gauche à droite $n = 100$, $n = 500$, et $n = 1000$.

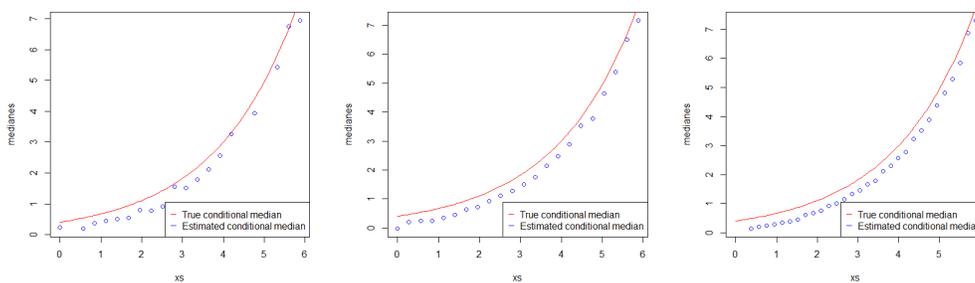


FIGURE 4.2 – **M2**, $CR \simeq 30\%$ et de gauche à droite $n = 100$, $n = 500$, et $n = 1000$

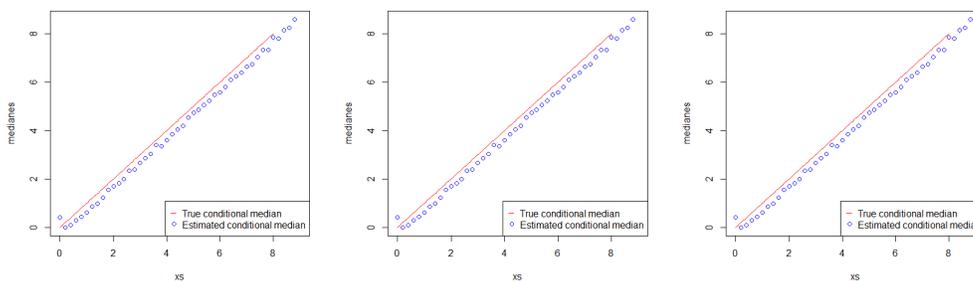


FIGURE 4.3 – **M1**, $n = 500$ et de gauche à droite $CR = 0\%$, $CR \simeq 30\%$ et $CR \simeq 60\%$

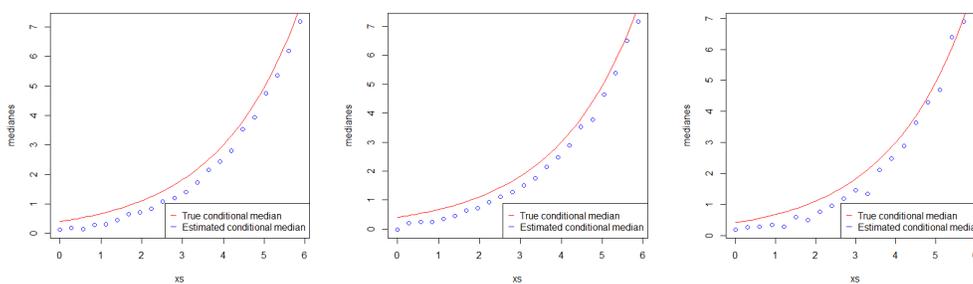


FIGURE 4.4 – **M2**, $n = 500$ et de gauche à droite $CR = 0\%$, $CR \simeq 30\%$ et $CR \simeq 60\%$

vous voir que lorsque le taux de censure augmente, la qualité d'ajustement se détériore légèrement pour les deux modèles.

4.2 Etude numérique II

4.2.1 La convergence

Nous examinons d'abord les performances pour des échantillons finis de l'estimation du mode conditionnel $\theta_n(x)$, introduit dans (3.7), pour les modèles usuels suivants.

M1 : $Y = X + 0,2\epsilon$, **M2** : $Y = \exp(0,5X - 0,9) + 0,2\epsilon$; et **M3** : $Y = \cos(1,5X) + 0,2\epsilon$, où les v.a.r. X et ϵ sont indépendantes et suivent respectivement la loi exponentielle de paramètre $1/3$ et la loi normale standard. La variable de censure à droite est choisie comme $R = L + V$, où la variable de censure à gauche L et la variable V suivent la distribution exponentielle avec les paramètres λ_1 et λ_2 respectivement. Cela garantit que L est plus petit que R p.s. puisqu'il s'agit de modèles de données doublement censurés. Nous prenons différentes valeurs de λ_1 et λ_2 pour obtenir différents taux de censure (CR). On choisit le même noyau d'Epanechnikov pour K et H . Enfin, on utilise une fenêtre optimale obtenue en minimisant la norme supremum entre la densité conditionnelle estimée et la vraie sur un ensemble de valeurs de la fenêtre sur $[0, 0.3]$.

Nous évaluons d'abord l'effet des taux de censure en fixant la taille d'échantillon (Figures 1 – 3). Ensuite, nous fixons le taux de censure pour vérifier l'effet de la taille d'échantillon les figures 4-6. En conclusion, comme on peut s'y attendre, la qualité de l'ajustement s'améliore lorsque la taille d'échantillon augmente ou que le taux de censure diminue. Aussi, la qualité de l'estimation semble être acceptable à travers tous nos résultats.

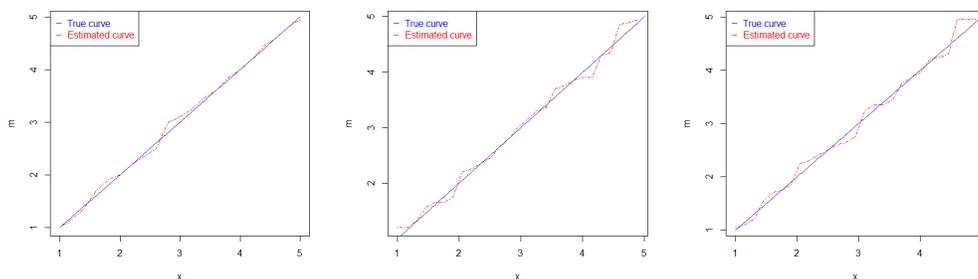


FIGURE 4.5 – **M1**, $n = 100$ et de gauche à droite $CR = 0\%$, $CR \simeq 30\%$ et $CR \simeq 60\%$

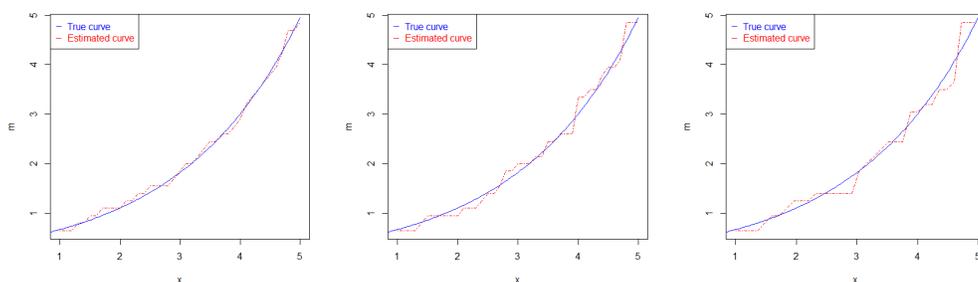


FIGURE 4.6 – **M2**, $n = 300$ et de gauche à droite $CR = 0\%$, $CR \simeq 30\%$ et $CR \simeq 60\%$

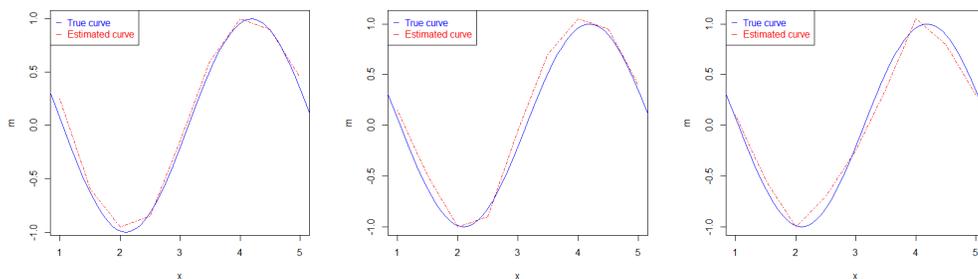


FIGURE 4.7 – **M3**, $n = 500$ et de gauche à droite $CR = 0\%$, $CR \simeq 30\%$ et $CR \simeq 60\%$

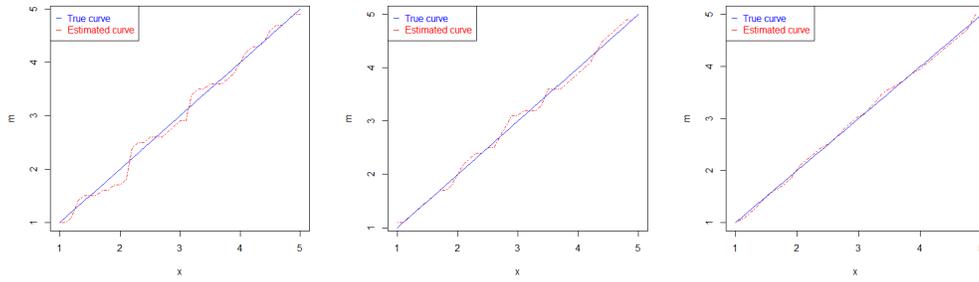


FIGURE 4.8 – **M1**, $CR \simeq 30\%$ et de gauche à droite $n = 50$, $n = 300$, et $n = 500$

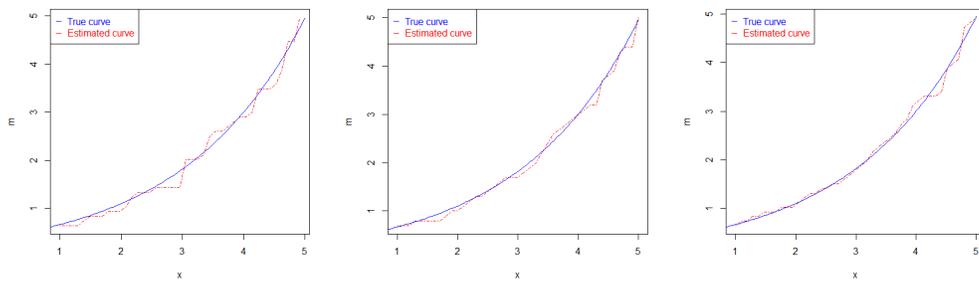


FIGURE 4.9 – **M2**, $CR \simeq 30\%$ et de gauche à droite $n = 50$, $n = 300$, et $n = 500$

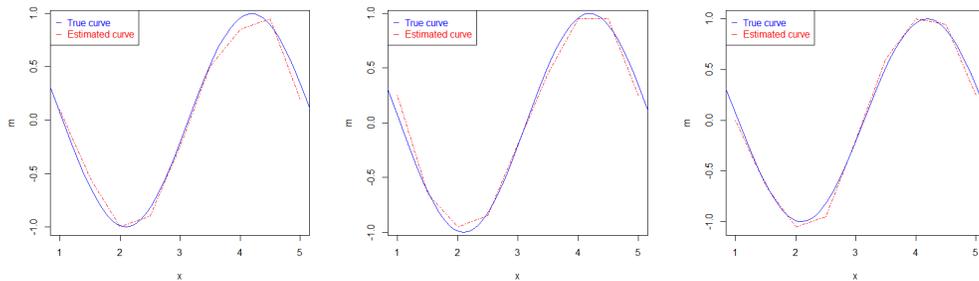
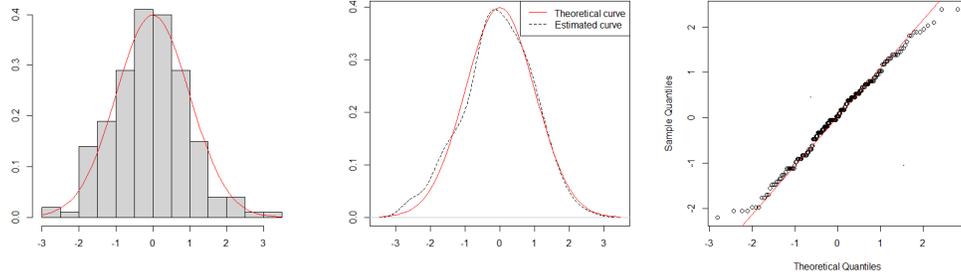


FIGURE 4.10 – **M3**, $CR \simeq 30\%$ et de gauche à droite $n = 300$, $n = 500$, et $n = 800$.

FIGURE 4.11 – Les résultats obtenus pour le modèle M1 si $x=2$

4.2.2 Normalité asymptotique

Dans cette partie, nous montrons le comportement gaussien l'estimateur du mode conditionnel $\theta_n(x)$. On utilise les mêmes modèles de régression M1, M2 et M3 du paragraphe précédent et on considère pour chaque modèle, le cas correspondant à un taux de censure d'environ 30%. A un point fixé $x \in \mathbb{R}$ (qui sera précisé pour chaque modèle), on génère 200 échantillons de taille 300 des variables X , Y , R et L et on calcule la valeur de $\theta_n(x)$ qui correspond à chaque échantillon. Ensuite, on centre ces valeurs en supprimant $\theta(x)$ et on les normalise en divisant par l'écart-type. Pour comparer la distribution d'échantillon obtenu avec la distribution gaussienne, nous traçons l'histogramme de cet échantillon et son estimateur de densité à noyau par rapport à la densité de la distribution gaussienne standard. Nous utilisons également le graphique Q-Q plot comme outil supplémentaire pour vérifier la normalité. Les résultats obtenus sont donnés dans les figures 7 – 9 ci-dessous. Ces résultats sont en faveur d'une bonne approximation par la distribution normale standard de la distribution asymptotique de $\theta_n(x)$ convenablement normalisée.

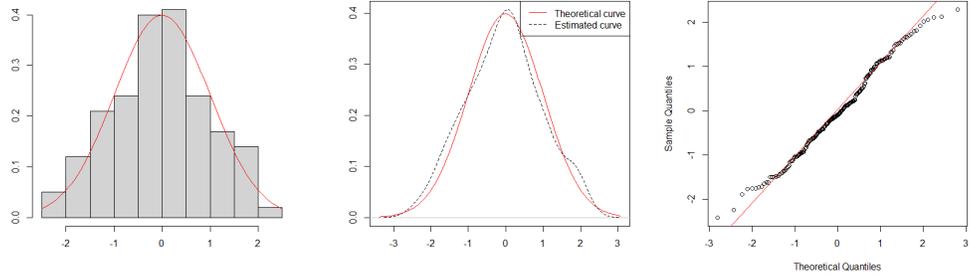


FIGURE 4.12 – Les résultats obtenus pour le modèle M2 si $x=3$

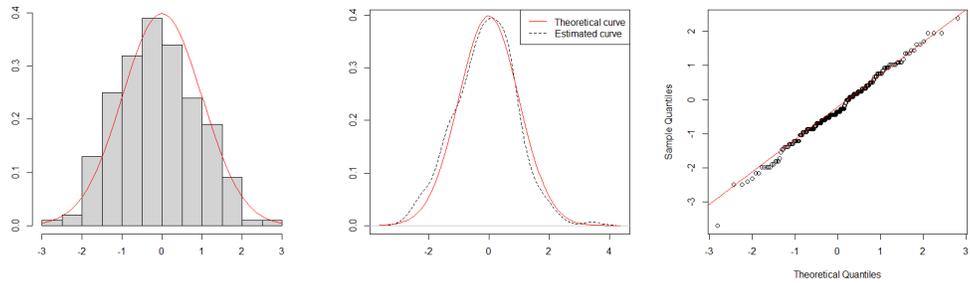


FIGURE 4.13 – Les résultats obtenus pour le modèle M3 si $x=0$

4.3 Application et données réelles

Dans cette section, nous appliquons notre méthodologie sur les données de transplantation cardiaque de Stanford. Cet ensemble de données se trouve dans Miller et Halpern (1982) ainsi que quelques explications à son sujet. Il contient les durées de survie des patients ayant reçu une transplantation cardiaque. Suivant Miller et Halpern (1982) nous ne traitons que $n = 157$ patients pour les quels le score d'inadéquation T_5 est connu. Comme dans Miller et Halpern (1982) et Shen (2012), la variable réponse dans notre étude est le \log_{10} du temps de survie (en jours). Par ailleurs, la variable explicative est l'âge au moment de la greffe. Le temps de survie est censuré à droite et nous avons 55 observations censurées à droite et 102 observations complètes. On procède comme dans Shen (2012) pour inclure la censure à gauche. Plus précisément, on génère 102 variables $(L_i)_{1 \leq i \leq 102}$ à partir de la distribution uniforme sur l'intervalle $[1, 3]$ et on prend le maximum entre les \log_{10} du temps de survie (les observations complètes) et L_i . Cela donne un échantillon doublement censuré. Nous obtenons 55 de données complètes, 55 de données censurées à droite et 47 censurées à gauche.

Soit $(Z_i)_{1 \leq i \leq n}$ l'échantillon doublement censuré obtenu, $(\delta_i)_{1 \leq i \leq n}$ l'indicateur de censure correspondant et $(X_i)_{1 \leq i \leq n}$ désigne l'échantillon de la variable explicative (l'âge au moment de la greffe). Nous procédons comme suit. On divise l'échantillon $(X_i, Z_i, \delta_i)_{1 \leq i \leq n}$ en deux parties : l'échantillon d'apprentissage et l'échantillon de test. L'échantillon d'apprentissage est composé de 79 observations et l'échantillon de test est composé des 78 restants. En utilisant le même noyau d'Epanechnikov pour K et H et la technique de validation croisée pour choisir la fenêtre optimale, nous nous basons sur l'échantillon d'apprentissage pour calculer l'estimateur du mode conditionnel $\theta_n(X_i)$ ($80 \leq i \leq 157$) pour les valeurs X_i de l'échantillon test. Ensuite, nous représentons les valeurs prédites $(\theta_n(X_i))_{80 \leq i \leq 157}$ par rap-

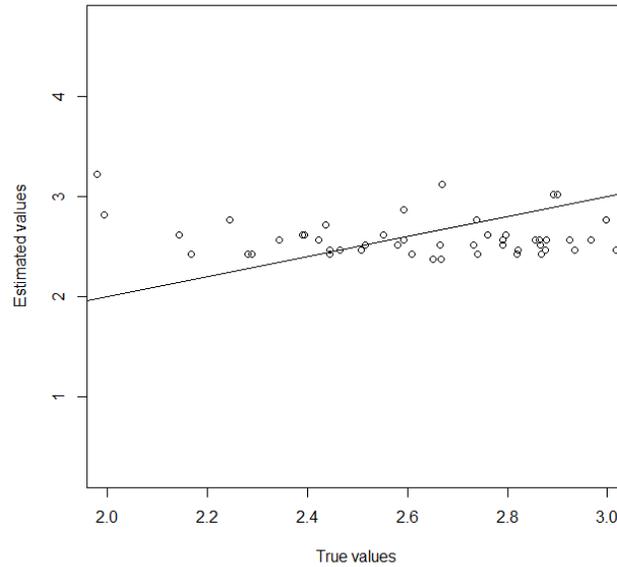


FIGURE 4.14 – Prédiction du mode conditionnel

port aux vraies valeurs $(Z_i)_{80 \leq i \leq 157}$ ce qui donne la figure 10. Nous traçons également la première bissectrice pour montrer la qualité de la prédiction. On remarque que le nuage de points $(Z_i, \theta_n(X_i))_{80 \leq i \leq 157}$ est proche de la première bissectrice ce qui montre que l'estimation du mode conditionnel θ_n donne une bonne prédiction des valeurs Z_i . Ceci est confirmé par le calcul de l'erreur moyenne quadratique suivante :

$$MSE = \frac{1}{78} \sum_{i=80}^{157} (\theta_n(X_i) - Z_i)^2 = 0,3337.$$

Extensions et perspectives

Il reste beaucoup de questions sans réponses :

- Dans le chapitre 2, nous n'avons obtenu qu'un résultat de convergence. Il serait alors intéressant d'étudier la normalité asymptotiques de l'estimateur de la fonction du quantile conditionnel par la méthode du double noyaux. On pourra aussi penser à étudier ce dernier dans le cadre dépendant (sous divers types de dépendances : ϕ -mélange, α -mélange, association, etc).
- Il serait aussi intéressant d'étudier le comportement de l'estimation de la fonction du quantile conditionnel dans le cadre de données doublement censurées dans un contexte non-paramétrique.
- Une question importante concernant l'étude des propriétés asymptotiques des différents outils de prévision pour des données dépendantes. Il serait donc intéressant d'étudier la loi du logarithme itéré pour le processus empirique dans ce contexte
- Une autre piste consiste à étendre nos travaux au cas d'un espace de dimension infinie.

Bibliographie

- L. AOUICHA et F. MESSACI : Kernel estimation of the conditional density under a censorship model. *Statistics & Probability Letters*, 145:173–180, 2019.
- L. AOUICHA et F. MESSACI : Estimation non paramétrique du mode conditionnel dans un modèle de censure. Thèse du Doctorat en Science, Université Les Frères Mentouri Constantine1, 2019.
- H. BENCHOUAK, M. BOUKELOUA et F. MESSACI : Asymptotic properties of a kernel estimate of the conditional mode in the presence of doubly censored data. *Communications in Statistics - Theory and Methods*, 53 (4):1143–1161, 2024 .
- S. BERNSTEIN : *Probability Theory, 4th ed. (in russian)*. M. L. Gostechizdat, 1946.
- D. BITOUZÉ, B. LAURENT et P. MASSART : A Dvoretzky-Kiefer-Wolfowitz type inequality for the Kaplan-Meier estimator. *Ann. Inst. Henri Poincaré, Probab. Stat.*, 35(6):735–763, 1999. ISSN 0246-0203.
- D. BLONDIN : Estimation nonparamétrique multidimensionnelle des dérivées de la régression. *C. R. Acad. Sci. Paris*, 339:713-716, 2004.
- M. BOUKELOUA : Rates of mean square convergence of density and failure rate estimators under twice censoring. *Statistics & Probability Letters*, 106: 121-128, 2015.

- M. BOUKELOUA et F. MESSACI : Asymptotic normality of kernel estimators based upon incomplete data. *Journal of Nonparametric Statistics*, 28 (3) :469–486, 2016.
- S. BOUZEBDA et T. EL-HADJALI : Uniform convergence rate of the kernel regression estimator adaptive to intrinsic dimension in presence of censored data. *Journal of Nonparametric Statistics* 1–51, 2020.
- M. N. CHANG : Weak convergence of a self-consistent estimator of the survival function with doubly censored data. *The Annals of Statistics* , 18 (1):391–404, 1990.
- M. N. CHANG : G. L. YANG : Strong consistency of a nonparametric estimator of the survival function with doubly censored data. *The Annals of Statistics* , 15(4):1536–1547, 1987.
- A. CARBONEZ, L. GYÖRFI et E. C. van der MEULEN : Partition-estimates of a regression function under random censoring. *Statist. Decisions*, 13:21–37, 1995.
- G. COLLOMB, W. HÄRDLE et S. HASSANI : A note on prediction via estimation of the conditional mode function. *Journal of Statistical Planning and Inference*, 15:227–236, 1987.
- S. DABO-NIANG et A. LAKSACI , Estimation non paramétrique du mode conditionnel pour variable explicative fonctionnelle. . *Pub. Inst. Stat. Univ. Paris*, 3 :27–42, 2007.
- J. DEMONGEOT , A. LAKSACI , F. MADANI et M. RACHDI : Functional data : Local linear estimation of the conditional density and its application. *C. R., Math., Acad. Sci. Paris*, 348, Issues 15-16, :931–934, 2010.
- M. EZZAHRIOUI et E. OULD SAÏD : Asymptotic normality of nonparametric estimators of the conditional mode for functional data. *Technical report, No.249, LMPA, Univ. Littoral Côte d’Opale*.
- M. EZZAHRIOUI, E. OULD SAÏD : Asymptotic results of a nonparametric

- conditional quantile estimator for functional time series. *Communications in Statistics- Theory and Methods*, 37(17): 2735–2759, 2008.
- M. KOHLER, K. MÁTHÉ et M. PINTÉR : Prediction from randomly right censored data. *J. Multivariate Anal.*, 80:73–100, 2002.
- F. FERRATY et P. VIEU : Statistique fonctionnelle : Modèles non-paramétriques de régression. Notes de cours de DEA, Université Paul Sabatier, Toulouse. France, 2002.
- F. FERRATY et P. VIEU : *Nonparametric functional data analysis : Theory and practice*. Springer, 2006.
- A. FÖLDES et L. REJTŐ : A LIL type result for the product limit estimator. *Probability Theory and Related Fields*, 56(1):75–86, 1981.
- A. GANNOUN : Prediction non paramétrique : medianogramme et méthode du noyau en estimation de la médiane conditionnelle . *Statistique et analyse des données*, 16(1):23–42, 1991.
- A. GANNOUN, S. GIRARD, C. GUINOT et J. SARACCO : Trois méthodes non paramétriques pour l'estimation de courbes de référence-application à l'analyse de propriétés biophysiques de la peau. *Statistique appliquée*, 50 (1):65–89, 2002.
- A. GANNOUN, J. SARACCO et K. YU : Nonparametric prediction by conditional median and quantiles. *Journal of statistical Planning and inference*, 117(2):207–223, 2003.
- G. GALILEI Dialogue concerning the two chief world systems, 2001. *Modern library. Translated by Stillman Drake*, 1632.
- E. GINÉ et A. GUILLOU : Laws of the iterated logarithm for censored data. *The Annals of Probability*, 27(4):2042–2067, 1999.
- E. GINÉ et A. GUILLOU : On consistency of kernel density estimators for

- randomly censored data : rates holding uniformly over adaptive intervals. *Annales de l'institut Henri Poincaré*, 37(4):503–522., 2001.
- Z. GUESSOUM, M. A.MANSOURI et E. OULD SAÏD : Asymptotic properties of the kernel mode estimator under twice censorship model. *Communications in Statistics - Theory and Methods*, 47(9):2195-2212, 2018.
- T. HONDA : Nonparametric estimation of a conditional quantile for α -mixing processes. *Ann. Inst. Statist. Math*, 52:459-470, 2000.
- P. L. HSU et H. ROBBINS : Complete convergence and the law of large numbers. *Proceedings of the national academy of sciences*, 33(2), 1947.
- E. L. KAPLAN et P. MEIER : Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481, 1958.
- K. KEBABI, I. LAROUCSI et F. MESSACI : Least squares estimators of the regression function with twice censored data. *Statistics & Probability Letters*, 81(11):1588–1593, 2011.
- K. KEBABI et F. MESSACI : Rate of the almost complete convergence of a kernel regression estimate with twice censored data. *Statistics & Probability Letters*, 82(11):1908–1913, 2012.
- S. KHARDANI , M. LEMDANI et E. OULD SAÏD : 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, 2010.
- S. KHARDANI, M. LEMDANI et E. OULD SAÏD : Uniform rate of strong consistency for a smooth kernel estimator of the conditional mode for censored time series. *Journal of Statistical Planning and Inference*, 141(11): 3426–3436, 2011.
- A. KITOUNI, M. BOUKELOUA et F. MESSACI : Rate of strong consistency

- for nonparametric estimators based on twice censored data. *Statistics & Probability Letters*, 96:255–261, 2015.
- J. KIEFER : On large deviations of the empiric df of vector chance variables and a law of the iterated logarithm. *Pacific J. Math*, 11(3):649–660, 1961.
- M. R. KOSOROK : Introduction to Empirical Processes and Semiparametric Inference. *Springer Verlag, New York*, 2008.
- M. LEMDAN, E. OULD SAÏD et N. POULIN :
Asymptotic properties of a conditional quantile estimator with randomly truncated data. *J. Multivariate Anal*, 100(1):546–449, 2009.
- G. LIN, X. HEB et S. PORTNOY : Quantile regression with doubly censored data. *Computational Statistics and Data Analysis*, 56:797–812, 2012.
- F. MESSACI : Local averaging estimates of the regression function with twice censored data. *Statistics & Probability Letters*, 80(19-20):1508–1511, 2010.
- F. MESSACI et N. NEMOUCHI : A law of the iterated logarithm for the product limit estimator with doubly censored data. *Statistics & Probability Letters*, 81(8):1241–1244, 2011.
- F. MESSACI et N. NEMOUCHI : Erratum to a law of the iterated logarithm for the product limit estimator with doubly censored data" [Statist. Probab. Lett. 81 (2011) 1241–1244]. *Statistics & Probability Letters*, 83(9):2142, 2013.
- F. MESSACI, N. NEMOUCHI, I. OUASSOU et M. RACHDI : Local polynomial modelling of the conditional quantile. *Stat Methods Appl*, 83..(9).. 2015.
- R. MILLER et J. HALPERN : Regression with censored data. *Biometrika*, 69:521–531, 1982.
- E. OULD SAÏD : A strong uniform convergence rate of kernel conditio-

- nal quantile estimator under random censorship. *Statistics & Probability Letters*, 76: 579–586, 2006.
- E. OULD SAÏD et Z. CAI : Strong uniform consistency of nonparametric estimation of the censored conditional mode function. *Journal of Nonparametric Statistics*, 17(7):797–806, 2005.
- E. OULD SAÏD et M. LEMDANI Asymptotic properties of a nonparametric regression function estimator with randomly truncated data. *Ann. Inst. Statist. Math*, 58: 357–378, 2006.
- E. OULD-SAÏD et O. SADKI : Prediction via the conditional quantile for right censorship model. *Far East Theoretical Statistics*, 2(2):145–179, 2008a.
- E. PARZEN : On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962.
- V. PATILEA et J. M. ROLIN : Product-limit estimators of the survival function with twice censored data. *Ann. Statist.*, 34(2):925–938, 2006.
- J. J. REN : On Self-Consistent Estimators and Kernel Density Estimators With Doubly Censored Data. *Statistical Planning and Inference*, 64:27–43, 1997.
- M. SAMANTA : Non-parametric estimation of conditional quantiles. *Statist. Probab. Lett.*, 7:407-412, 1989.
- M. SAMANTA et A. THAVANESWARAN : Non-parametric estimation of the conditional mode. *Communications in Statistics-Theory and Methods*, 19(12):4515–4524, 1990.
- P. S. SHEN : Estimation in multiple linear regression model with doubly censored data. *Journal of Statistical Computation and Simulation*, 82(4):503–514, 2012.

- M. TALAGRAND : Sharper bounds for gaussian and empirical processes. *The Annals of Probability*, 22(1):28–76, 1994.
- M. TALAGRAND : New concentration inequalities in product spaces. *Invent. Math*, 126:505–563, 1996.
- B. W. TURNBULL : Nonparametric estimation of a survivorship function with doubly censored data. *Journal of the American Statistical Association*, 69(345):169–173, March 1974.
- B. W. TURNBULL : The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38(3):290–295, 1976.
- H. YING LIANG et J. UÑA-ÁLVAREZ : Asymptotic properties of conditional quantile estimator for censored dependent observations. *Statistical Mathematics*, 63:267-289, 2009.
- A. W. VAN DER VAART et J. A. WELLNE : Weak Convergence and Empirical Processes. *Springer, New York*, 1996.
- V. N. VAPNIK et A. JA. ČERVONENKIS : Necessary and sufficient conditions for the uniform convergence of means to their expectations. *Theory Probab. App*, 26 :532–553,1981.
- Y. ZHOU et H. LIANG : properties for L1 norm kernel estimator of conditional median under dependance. *J. Nonparametr. Stat*, 15:205-219, 2003.

Résumé

Dans cette thèse, nous nous intéressons à la prévision non-paramétrique via l'estimation du mode conditionnel et du quantile conditionnel.

Les modèles considérés ici sont les modèles censurés aléatoirement, le premier selon Turnbull (1974), on observe dans ce contexte le $\max(\min(X, R), L)$ avec $L \leq R$ presque sûrement et qui est utilisé dans différents domaines pratiques. Le second modèle selon Patilea et Rolin (2006) est semblable au modèle doublement censuré mais X, R et L sont indépendantes.

Nous réservons un chapitre à construire un estimateur du quantile conditionnel basé sur des données soumises à la censure mixte, ensuite nous étudions la convergence uniforme presque complète en précisant la vitesse de convergence sous des conditions usuelles.

Un autre intérêt est de proposer et d'étudier un estimateur à noyau de la fonction du mode conditionnel pour un échantillon constitué de variable i.i.d et que la variable d'intérêt est sujette à une censure double. Nous montrons la convergence presque sûre de cet estimateur. Enfin, nous établissons sa normalité asymptotique.

Des études de simulation sont conduites afin de tester les comportements de nos estimateurs pour un échantillon de taille finie et pour valider nos résultats de la normalité asymptotique. Enfin nous appliquons notre estimateur du mode conditionnel sur des données de transplantation car-

diaque de Stanford , en se basant sur les travaux de Miller et Halpern (1982) et Shen (2012) .

Mots clés : la censure mixte, la censure double, la convergence presque sûre, non-paramétrique, taux de convergence presque complète, la normalité asymptotique, estimateur du quantile conditionnel, estimateur du mode conditionnel, simulation, application.

Abstract

In this thesis, we are interested in non-parametric prediction according to the conditional mode and conditional quantile estimation.

The models considered here are the randomly censored models, the first according to Turnbull (1974), we observe in this context the $\max(\min(X, R), L)$ with $L \leq R$ almost surely and which is used in different practical fields. The second model according to Patilea et Rolin (2006) is similar to the double censored model but X, R and L are independent.

We reserve a chapter to construct a conditional quantile estimator when the variable of interest is twice censored, then we study the uniform almost complete convergence by specifying the rate of convergence, under usual conditions.

Another interesting topic is to propose and study a kernel estimator of the conditional mode function for a sample consisting of r.r.v. variables and such that the variable of interest is subject to double censoring. We prove the almost sure convergence of this estimator. Finally, we establish its asymptotic normality.

Simulation studies are conducted in order to the behavior of our estimators for a sample of finite size and to validate our results of asymptotic normality. Finally we apply our conditional mode estimator on heart transplant data from Stanford, basing on the work of Miller et Halpern (1982)

and Shen (2012).

Keywords : Twice censoship, double censorship, non-parametric, almost sure convergence, almost complete convergence rate, the asymptotic normality, conditional quantile estimator, conditional mode estimator , simulation, application.

ملخص

في هذه الأطروحة، سنهتم بالتنبؤ غير الوسيطى بواسطة تقدير المنوال الشرطي و الوسيط الشرطي.

النماذج المعتبرة هنا هي نماذج خاضعة للرقابة، العشوائية النموذج الاول حسب (Turnbull 1974) ، في هذا السياق يلاحظ $\max(\min(X; R); L)$ مع $L \leq R$ شبه مؤكد والذي استخدم في مجالات عملية مختلفة.

النموذج الثاني حسب (Patilea et Rolin 2006) وهو مشابه للنموذج الاول الخاضع للرقابة المزدوجة

لكن متغيراته X ، R و L مستقلة.

نخصص فصل لبناء تقدير للوسيط الشرطي معتمدين على بيانات خاضعة للرقابة المختلطة ثم ندرس التقارب المنتظم الشبه كامل مع تحديد سرعة التقارب تحت شروط اعتيادية.

اهتمام اخر هو اقتراح ودراسة مقدر دالة المنوال الشرطي باستعمال طريقة النواة من اجل عينة متكونة من متغيرات مستقلة و متماثلة التوزيع في الحالة أين يكون المتغير المدروس خاضع للرقابة المزدوجة. كما أننا نبرهن على التقارب شبه الأكيد و التقارب الطبيعي لهذا المقدر.

تجرى دراسات المحاكاة لاختبار سلوكيات مقدراتنا من اجل عينة منتهية وللتحقق من صحة نتائجنا الخاصة بالتقارب الطبيعي. وفي النهاية نطبق مقدر المنوال الشرطي على معطيات زرع القلب لستانفورد بالاعتماد على نتائج بحث (Shen و Halpern و Miller 1982).

الكلمات المفتاحية :

الرقابة المختلطة، الرقابة المزدوجة ،التقارب الشبه اكيد ،غير وسيطي، سرعة التقارب الشبه كامل، التقارب الطبيعي، مقدر الوسيط الشرطي، مقدر المنوال الشرطي، المحاكاة، الدراسة التطبيقية.