Sur l'estimation non paramétrique de la densité et du mode dans les modèles de données incomplètes et associées

Yacine Ferrani

THÈSE

Présentée pour l'obtention du diplôme de Doctorat

En : Mathématiques

Spécialité : Probabilités et Statistique

Par : FERRANI Yacine

Sujet :

Sur l'estimation non paramétrique de la densité et du mode dans les modèles de données incomplètes et associées

soutenue publiquement le 23 Novembre 2014, devant le jury composé de :

Mme.	Z.	GUESSOUM	MC.A	U.S.T.H.B, Algérie	Présidente
Mr.	E.	OULD SAÏD	MCF-HDR	Univ. du Littoral, France	Dir. de thèse
Mr.	A.	TATACHAK	MC.A	U.S.T.H.B, Algérie	Dir. de thèse
Mr.	J.F.	DUPUY	Professeur	Univ. de Rennes, France	Rapporteur
Mme.	F.	MESSACI	Professeur	Univ. de Constantine, Algérie	Rapporteur
Mme.	A.	LECLERCQ SAMSON	Professeur	Univ. de Grenoble, France	Examinatrice

Remerciements

Je désire tout d'abord, témoigner de ma reconnaissance envers mes directeurs de thèse, Abdelkader Tatachak et Elias Ould Saïd.

A Abdelkader Tatachak, Docteur à l'USTHB, l'instigateur de cette étude, qui m'a remis le pied à l'étrier de la recherche, pour sa présence, la constance de son soutien, son aide incommensurable et ses hautes compétences scientifiques.
Sans toi, ce travail aurait été plus difficile, voire impossible.

A Elias Ould Saïd, MCF-HDR à l'Université du Littoral, pour la justesse de ses orientations, ses remarques et conseils pertinents, ma bibliothèque ambulante.
Ce fût pour moi un immense privilège de travailler sous l'égide du laboratoire LMPA de Lille, où tu m'as initié dans le cadre de la co-tutelle. J'espère que notre collaboration se prolongera bien au-delà de cette thèse.

Je suis très heureux que Madame Zohra Guessoum, Docteur à l'USTHB, ait accepté de me faire l'honneur de présider le jury de cette thèse.

Je suis également très reconnaissant envers Madame Fatiha Messaci, Professeur à l'Université de Constantine, et Monsieur Jean-François Dupuy, Professeur à l'Université de Rennes, pour avoir bien voulu consacrer de leur précieux temps à l'expertise de cette thèse, en acceptant d'en être les rapporteurs.

Je suis de plus particulièrement honoré que Madame Adeline Leclercq-Samson, Professeur à l'Université de Grenoble, ait accepté de faire partie de mon jury.

Je désire adresser une marque particulière de ma gratitude à Messieurs les Professeurs B. Benzaghou, Recteur de l'Université des Sciences et de la Technologie Houari Boumediene (USTHB), D.E. Akretche, Vice Recteur Chargé de la Post-Graduation et de la Recherche

Scientifique (USTHB), K. Boukhetala, Doyen de la Faculté de Mathématiques (USTHB), R. Durand, Président de L'Université du Littoral Côte d'Opale (ULCO), M. Bendjelloun, responsable du Domaine SPI de l'Ecole Doctorale (ULCO), H. Sadok, Directeur du LMPA (ULCO), responsables du partenariat entre l'USTHB et l'ULCO concrétisé par la convention de co-tutelle de cette thèse.

Je ne peux oublier l'immense contribution du Professeur K. Baddari, Doyen de la Faculté des Sciences (FS) de l'Université M'Hamed Bougara de Boumerdes (UMBB)(actuellement Recteur de l'Université Mohand Akli Ouelhadj de Bouira), pour son accord de financement et sa sollicitude permanente.

Mes remerciements sont également adressés à tous mes professeurs, collègues et amis de la Faculté de Mathématiques de l'USTHB, de la Faculté des Sciences de l'UMBB, les membres du laboratoire MSTD-USTHB dont je fais partie, ainsi que ceux du projet de recherche 'Modélisation Stochastique et Applications-UMBB'.

Je ne pourrais oublier toutes les personnes que j'ai eu le plaisir de côtoyer durant mes quinze années d'enseignement au Centre Bio-Médical de Dergana.

Résumé

Cette thèse porte sur l'étude des propriétés asymptotiques d'un estimateur non paramétrique de la densité de type Parzen-Rosenblatt, sous un modèle de données censurées à droite, vérifiant une structure de dépendance de type associé.

Dans ce cadre, nous rappelons d'abord les résultats existants, avec détails, dans les cas i.i.d. et fortement mélangeant (α-mélange). Sous des conditions de régularité classiques, il est établi que la vitesse de convergence uniforme presque sûre de l'estimateur étudié, est optimale.

Dans la partie dédiée aux résultats de cette thèse, deux résultats principaux et originaux sont présentés :

• Le premier résultat concerne la convergence uniforme presque sûre de l'estimateur étudié sous l'hypothèse d'association. L'outil principal ayant permis l'obtention de la vitesse optimale est l'adaptation du Théorème de Doukhan et Neumann (2007), dans l'étude du terme des fluctuations (partie aléatoire) de l'écart entre l'estimateur considéré et le paramètre étudié (densité). Comme application, la convergence presque sûre de l'estimateur non paramétrique du mode est établie.

Les résultats obtenus ont fait l'objet d'un article accepté pour publication dans Communications in Statistics-Theory & Methods.

• Le deuxième résultat établit la normalité asymptotique de l'estimateur étudié sous le même modèle et constitue ainsi une extension au cas censuré, du résultat obtenu par Roussas (2000). Ce résultat est soumis pour publication.

Mots-clés. Alpha-mélange, Association, Censure droite, Convergence uniforme presque sûre, Estimation non paramétrique, Kaplan-Meier, Mode, Normalité asymptotique.

Abstract

This thesis deals with the study of asymptotic properties of a kernel (Parzen-Rosenblatt) density estimate under associated and censored model.

In this setting, we first recall with details the existing results, studied in both i.i.d. and strong mixing condition (α-mixing) cases. Under mild standard conditions, it is established that the strong uniform almost sure convergence rate, is optimal.

In the part dedicated to the results of this thesis, two main and original stated results are presented :

• The first result concerns the strong uniform consistency rate of the studied estimator under association hypothesis. The main tool having permitted to achieve the optimal speed, is the adaptation of the Theorem due to Doukhan and Neumann (2007), in studying the term of fluctuations (random part) of the gap between the considered estimator and the studied parameter (density). As an application, the almost sure convergence of the kernel mode estimator is established.

The stated results have been accepted for publication in Communications in Statistics-Theory & Methods.

• The second result establishes the asymptotic normality of the estimator studied under the same model and then, constitute an extension to the censored case, the result stated by Roussas (2000). This result is submitted for publication.

Keywords. Alpha-mixing, Association, Asymptotic normality, Censoring, Kaplan-Meier, Mode, Non-parametric estimation, Strong uniform consistency.

Articles et Communications

1. Ferrani, Y., Ould Saïd, E., Tatachak, A. (2014) (*To appear*) On kernel density and mode estimates for associated and censored data. DOI♯ 10.1080/03610926.2013.867996. *Communication in Statistics-Theory and Methods*.

2. Ferrani, Y., Ould Saïd, E., Tatachak, A. (*Submitted for publication*) Asymptotic Normality for a Kernel Density Estimate Under Censored and Associated Model.

3. Ferrani, Y., Ould Saïd, E., Tatachak, A. (2013) Asymptotic behavior of kernel density and mode estimates for censored and associated data. 15^{th} *Applied Stochastic Models and Data Analysis (ASMDA 2013)*, Mataró (Barcelona), Spain, June 2013.

4. Ferrani, Y., Ould Saïd, E., Tatachak, A. (2014) Asymptotic Normality for a Kernel Density Estimate Under Censored and Associated Model. 9^{th} *International Statistic Days Symposium-ISDS'2014 (IGS 2014)*, Antalya, Turkey, May 2014.

Table des matières

Chapitre 1

Introduction générale

1.1 Introduction

L'estimation est un élément fondamental de la statistique. Elle permet de généraliser, autant que faire se peut, des résultats observés. On y distingue

• l'approche paramétrique, qui considère que les modèles sont connus, avec des paramètres inconnus. La loi de la variable étudiée est supposée appartenir à une famille de lois pouvant être caractérisée par une forme fonctionnelle connue (fonction de répartition F, densité f,...) qui dépend d'un ou plusieurs paramètres inconnus à estimer.

• l'approche non paramétrique, qui ne fait aucune hypothèse sur la loi, ni sur ses paramètres. Nos connaissances sur le modèle ne sont pas précises, ce qui est souvent le cas dans la pratique. Dans cette situation, il est naturel de vouloir estimer une des fonctions décrivant le modèle, soit généralement la fonction de répartition ou la densité (pour le cas continu) : c'est l'objectif de l'estimation fonctionnelle. Si la fonction de répartition empirique F_n résout le problème statistique fondamental de la distribution de probabilité (associée à un échantillon $(T_1, T_2, \ldots . T_n)$ de variables aléatoires réelles indépendantes et de même loi) en fonction des valeurs numériques observées, elle est par contre limitée pour décrire visuellement les caractères de l'échantillon. Pour Deheuvels (1980), c'est l'estimation la plus naturelle dans le cas où on ne fait, extérieurement à l'échantillon,

aucune hypothèse restreignant le choix de cette distribution à une famille particulière.
L'usage de F_n est justifié par divers résultats comme le théorème de Glivenko-Cantelli assurant que

$$\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \longrightarrow 0, \; p.s.,$$

où F est la fonction de répartition de la loi recherchée, et le théorème de Donsker pour la normalité asymptotique. De plus, elle est simple, à la fois dans sa formulation et dans son calcul.
Cependant, elle n'en est pas moins limitée pour décrire les caractères de l'échantillon (ce qui occasionne une perte d'information sur la loi recherchée).

1.1.1 La densité

La densité, lorsqu'elle existe, est plus appropriée pour caractériser la distribution et permet de mieux la visualiser : son graphe permet de voir le ou les modes, la symétrie, la dispersion, l'aplatissement,...
Pour décrire une loi, on se sert particulièrement de la courbe représentative de la densité. De plus, les fluctuations sont plus apparentes dans les courbes de densités. D'autre part, l'outil informatique a permis de lever les difficultés calculatoires pour déterminer la densité, à partir des valeurs numériques observées d'un échantillon ; d'où l'intérêt accordé durant ce dernier demi-siècle à l'estimation de la densité, particulièrement depuis les travaux fondateurs de Rosenblatt (1956) et Parzen (1962) qui ont, en généralisant la notion d'estimation par histogramme (estimateur naïf), donné naissance à la méthode du noyau (grâce à laquelle a été aussi développée l'estimation de la fonction de régression par Nadaraya & Watson (1964)).
Soit $(T_1, T_2, \ldots, T_n)$, un échantillon de variables aléatoires indépendantes et identiquement distribuées (i.i.d.) de densité inconnue f. On définit l'estimateur à noyau de f par :

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

où $h := h(n)$ est une suite de nombres réels positifs (dépendant de n), appelés fenêtres ou largeurs de fenêtre, qui contrôlent le lissage de la courbe estimée, et K est une fonction

bornée, intégrable, d'intégrale égale à 1, appelée noyau. Si de plus $\lim_{|u|\longrightarrow\infty.} uK(u) = 0$, K est appelé noyau de Parzen-Rosenblatt (et f_n l'estimateur à noyau de Parzen-Rosenblatt). Un noyau est dit d'ordre r $(r \geq 1)$ si

- les fonctions $t \longrightarrow t^j K(t), j = 1, ..., r$, sont intégrables et vérifient
- $\int K(t)dt = 1$, et $\int t^j K(t)dt = 0, j = 1, ..., r$.

L'estimation de la densité est devenue un problème statistique classique. Plusieurs types d'estimateurs ont été proposés (approches par les plus proches voisins, séries orthogonales, maximum de vraisemblance pénalisé, histo-splines, ondelettes,...), mais d'une manière générale, les résultats qui en découlent ne sont pas significativement meilleurs que par la méthode du noyau. Les résultats établissant les propriétés des estimateurs proposés sont, naturellement, obtenus d'abord pour des variables indépendantes. Le plus célèbre des estimateurs reste l'estimateur à noyau, qui a été largement étudié, particulièrement pour les données complètes, indépendantes et identiquement distribuées (Prakasa Rao (1983), Silverman (1986),...).

Les performances de l'estimateurs à noyau dépendent principalement du choix de h, le choix du noyau n'ayant pas une grande influence (voir la notion d'efficacité au sens d'Epanechnikov).

L'écart quadratique moyen (Mean Square Error, MSE en abrégé) est un critère très répandu dans la littérature comme mesure d'erreur locale (pour évaluer la précision d'une valeur estimée) :

$$MSE(t) = \mathbb{E}\left(f_n(t) - f(t)\right)^2.$$

On peut décomposer le MSE en somme de la variance et du carré du biais de l'estimateur :

$$MSE(t) = \mathbb{E}(f_n(t) - \mathbb{E}(f_n(t))^2 + (\mathbb{E}f_n(t) - f(t))^2 = \sigma^2(t) + b^2(t).$$

En intégrant sur $\mathbb{R}$, on obtient le risque intégré (Mean Integrated Square Error ou MISE) comme mesure globale de l'erreur : $MISE(f_n) = \int MSE(t)dt$.

Pour le cas d'un noyau d'ordre r, on a sous certaines conditions,

$b^2(t_0) = \mathcal{O}(h^r)$, et $\sigma^2(t_0) = \mathcal{O}(\frac{1}{nh})$, (si $nh \longrightarrow \infty$), pour tout $t_0 \in \mathbb{R}$, et donc $MSE(t_0) \leq C(h^r + \frac{1}{nh})$.

Le minimum en h du membre de droite est $h^{MSE} = \mathcal{O}(n^{-\frac{1}{2r+1}})$ et par conséquent, pour cette fenêtre, uniformément en t_0, $MSE(t_0) = \mathcal{O}(n^{-\frac{2r}{2r+1}})$, lorsque $n \longrightarrow \infty$. (Tsybakov

(2003)).
On retrouve les mêmes résultats optimaux pour le risque intégré (fenêtre et MISE). Pour un noyau d'ordre 2, la fenêtre optimale h^{MSE} ou h^{MISE} tend vers 0 à la vitesse $n^{-\frac{1}{5}}$, tandis que le MSE et le MISE à la vitesse $n^{-\frac{4}{5}}$.
Pour la convergence dans L^∞ (la norme infinie est définie par $\|f\|_\infty = \sup_{t\in\mathbb{R}} |f(t)|$), sous la condition que $h = \mathcal{O}\left(\frac{n}{\log n}\right)^{-\frac{1}{2r+1}}$, on a

$$\sup_{t\in\mathbb{R}} |f_n(t) - f(t)| = \mathcal{O}\left(\frac{n}{\log n}\right)^{-\frac{r}{2r+1}},$$

et plus généralement si $f^{(j)}$ désigne la dérivée $j^{ème}$ de f,

$$\sup_{t\in\mathbb{R}} |f_n^{(j)}(t) - f^{(j)}(t)| = \mathcal{O}\left(\frac{n}{\log n}\right)^{-\frac{r}{2r+1}}.$$

(Voir Györfi & *al.* (1989), Vieu (1996)).
Le cas de données complètes i.i.d. représente pour les spécialistes, le cas "idéal", un cas qui ne reflète pas toujours la réalité : que se passe-t-il si l'on n'a pas les conditions classiques, pas toujours vérifiables, d'indépendance ou de complétude ? Que deviennent les théories élaborées sous ces hypothèses "parfaites", lorsque l'on s'en écarte ? Au lieu d'observer des réalisations i.i.d. de la variable d'intérêt T (des durées de vie, par exemple), on observe la réalisation de T soumise à des perturbations, indépendantes ou non du phénomène.
A ce titre, l'objet de cette thèse est d'estimer la densité, ainsi que le mode, dans le cas où aucune des deux hypothèses précédentes n'est vérifiée : nos données seront supposées incomplètes (du fait qu'on n'a pas accès à toute l'information) et présentant une structure de dépendance (appelée association). Nous développerons par la suite ces deux concepts.

1.1.2 Le mode

Estimer le mode est souvent une conséquence directe de l'estimation de densité. Son importance est dûe au fait que c'est une mesure naturelle de tendance centrale, qui n'est pas influencée par les queues des distributions. Le mode est la valeur la plus probable : pour une densité de probabilité f, c'est la valeur pour laquelle f admet un maximum (global ou local). Pour une distribution symétrique, il coincide avec deux autres paramètres de

position, la moyenne et la médiane. On peut distinguer plusieurs estimateurs du mode : on en citera quelques uns qui sont définis selon deux approches :

1.1.2.1 Méthode indirecte

Cete approche consiste à obtenir dans un premier temps une estimation de la densité f (pour le cas non paramétrique), et à prendre pour mode la valeur de t pour laquelle $f(t)$ est maximale. Parzen (1962) a été l'un des premiers à s'intéresser au problème de l'estimation du mode θ dans le cas d'une densité univariée. Il définit un estimateur comme la variable aléatoire qui maximise l'estimateur à noyau f_n de la densité :

$$\hat{\theta}_n = \arg\max_{t\in\mathbb{R}} f_n(t).$$

Il démontre que cet estimateur est uniformément convergent (en probabilité), asymptotiquement normal et donne une évaluation de l'erreur quadratique moyenne (MSE). Nadaraya (1965) en établit la consistence forte. D'autres études ont affaiblit les conditions suffisantes de normalité asymptotique (Eddy (1980), (1982), Romano (1988)), ou établi la convergence dans L^p (Devroye & Wagner (1980), Grund & Hall (1995)). La liste des travaux est longue, on citera ceux de Van Ryzin (1969), Mokkadem & Pelletier (2003), Herrmann & Ziegler (2004) parmi d'autres. Plus récemment, Shi & al (2009) améliorent le taux de convergence de l'estimateur du mode, exprimé en fonction de la fenêtre h.
D'autres auteurs se sont penchés sur le sujet, parmi lesquels on peut citer :

• Vieu (1996) qui propose l'étude de quatre estimateurs à noyau (globaux et locaux) du mode, basés sur des estimateurs à noyau de la densité et de sa dérivée.
- Deux estimateurs globaux, $\theta_{n,1}$ défini par Parzen, et $\theta_{n,3}$ obtenu en annulant $f'_n(t)$, l'estimateur de la dérivée $f'(t)$,
- Deux estimateurs locaux $\theta_{n,2}$ et $\theta_{n,4}$ basés respectivement, sur l'estimateur

$$f_L(t) = \frac{1}{nh(t)} \sum_{i=1}^{n} K\left(\frac{t-t_i}{h(t)}\right)$$

et sa dérivée, qui s'appuient sur une fenêtre locale, et non plus globale.

Cet auteur montre que les estimateurs locaux améliorent les deux premiers.

• Bickel (2003) qui introduit une autre approche indirecte. Il étudie deux estimateurs paramétriques, en transformant les observations $t_1, t_2, ..., t_n$ en données approximativement normales $y_i = (t_i)^\alpha$, $\alpha \in \mathbb{R}$. Ensuite, on annule la dérivée de la densité normale ainsi obtenue (les paramètres de la loi normale sont estimés par la moyenne $\bar{y}$ et l'écart type σ des données transformées), et obtient pour estimateur

$$M = \left[\frac{1}{2}\left(\bar{y} + \sqrt{\bar{y}^2 + \frac{4\sigma^2(\alpha-1)}{\alpha}}\right)\right]^{1/\alpha}.$$

Pour $\alpha = 1$, $M = \bar{y}$, ce qui correspond au mode dans le cas d'une distribution symétrique comme la loi normale.
Le deuxième estimateur, plus robuste, est obtenu en remplaçant les paramètres respectivement par la médiane et l'écart médian absolu standardisé ($\Delta(y_i) = C.med\,|y_i - med(y_i)|$) égal à σ dans le cas normal.

1.1.2.2 Méthode directe

On procède d'entrée à l'estimateur du mode θ en considérant que dans l'échantillon, on doit observer un groupement de valeurs dans le voisinage du mode (Chernoff, Dalenius, Grenander, Venter, ...) :

• Chernoff (1964) présente l'estimateur noté $\hat{t}_{a_n}$ comme le milieu d'un intervalle de longueur $2a_n$ contenant le maximum d'observations $t_1, t_2, ..., t_n$, $(a_n)_n$ étant une suite de réels positifs décroissant lentement vers zéro. Cet estimateur est inspiré de l'estimateur naïf de noyau $K_a(t) = \frac{1}{2a}$ si $|t| \leq a$, dont le mode est le milieu de l'intervalle $[-a, a]$.

• Wegman (1971) montre la consistence forte de cet estimateur.

• Grenander (1965) définit les estimateurs

$$M_{p,k} = \left[\frac{1}{2}\sum_{i=1}^{n-k}\frac{(t_{i+k}+t_i)}{(t_{i+k}-t_i)^p}\right]\left[\sum_{i=1}^{n-k}\frac{1}{(t_{i+k}-t_i)^p}\right]^{-1}, \; 1 < p < k,$$

pour un échantillon d'observations $(t_1, t_2, ..., t_n)$ ordonnées par ordre croissant, $t_1 \leq t_2 \leq ... \leq t_n$.

• Venter (1967) élabore un estimateur à partir d'une suite $(k_n)_n$ d'entiers naturels : c'est le milieu du plus petit intervalle contenant k_n observations parmi $t_1, t_2, ..., t_n$.

• Hall (1982) établit la normalité asymptotique de $M_{p,k}$ pour $k > 2p$. Ces estimateurs ont suscité peu d'intérêt, car hautement influencés par les valeurs extrêmes.

• Bickel (2002) propose deux estimateurs plus robustes appelés :
- HSM (Half-Sample Mode) obtenu à partir d'algorithmes basés sur des demi-échantillons successifs.
- HRM (Half-Range Mode), obtenu en cherchant un "intervalle modal" contenu dans d'autres intervalles modaux (le mode sera alors l'étendue du petit intervalle modal).

Pour Hedges & Shah (2003), HRM est obtenu par une "méthode simple et rapide et produit moins de biais dans les simulations". Ces deux derniers auteurs utilisent cet estimateur dans le cadre d'une étude (à l'Institut d'Astrobiologie - NASA) sur l'horloge moléculaire indiquant une distribution asymétrique, la moyenne étant plus élevée que la médiane dans la plupart des cas, justifiant l'usage du mode, dont les vraies valeurs sont inconnues. (*l'horloge moléculaire est un phénomène décrivant l'empreinte laissée par le temps dans les molécules du vivant. Ce phénomène est exploité par les biologistes pour remonter le cours de l'évolution des organismes*).

Cette liste d'estimateurs fait partie des plus cités dans la littérature, mais n'est pas exhaustive.

1.1.3 Données incomplètes

Nous allons d'abord introduire le domaine de l'analyse de survie.

1.1.3.1 Durée de vie

Un modèle de survie est basé sur les durées de vie. Ce terme (durée de vie) est une variable aléatoire positive et continue désignant le temps écoulé jusqu'à l'apparition d'un évènement précis, appelé communément "décès", "panne" ou "échec", selon les multiples domaines d'application que ce soit en médecine, ingénierie, finance, sciences sociales, industrie, économie ou autre (durée de survie après un infarctus, délai de rémission d'une maladie, durée de fonctionnement d'un composant, temps entre 2 pannes successives d'un appareil, temps d'attente dans un guichet, durée de chomage,...).
Deux spécificités des durées de vie retiennent l'attention :

• La première est que, outre la densité f, la fonction de répartition F et la fonction caractéristique, l'analyse de survie utilise d'autres notions, à savoir :
- La fonction de survie $S(t) = 1-F(t) = \mathbb{P}(T > t) = \int_t^{+\infty} F(u)du$, qui décrit la probabilité de "survivre" au moins jusqu'au temps t,
- La fonction (ou taux) de hasard $\lambda(t) = \frac{f(t)}{S(t)} = -\frac{d \log S(t)}{dt}$,
- La taux de hasard cumulé $\Lambda(t) = \int_0^t \lambda(u)du = -\log S(t)$.

Remarque 1.1. *Le taux de hasard (aussi appelé fonction de risque instantané de "mort") s'interprète comme suit :*
$\lambda(t) = \lim_{dt\to 0} \frac{\mathbb{P}(T\in[t,t+dt]|T>t)}{dt} = \lim_{dt\to 0} \frac{\mathbb{P}(\text{"mourir" entre t et t+dt, sachant que l'on est "vivant" à l'instant t})}{dt}$,
la probabilité de "mourir" à un instant t donné étant nulle ($\mathbb{P}(T = t) = 0$).

Le taux de hasard , tout comme la densité, les fonctions de répartition, caractéristique et de survie, caractérise entièrement la loi de T, et on a :

$$S(t) = \exp\left(-\int_0^t \lambda(u)du\right).$$

• La seconde est la présence de données incomplètes : La variable d'intérêt T n'est pas complètement observée pour toutes les données de l'échantillon $\{T_i, 1 \leq i \leq n\}$.

Deux types de données incomplètes intéressent le statisticien :

1.1.3.2 La censure

C'est le phénomène le plus couramment rencontré en analyse de survie. La variable d'intérêt T, n'est pas observée (l'individu n'a pas subi l'évènement), et est majorée ou

minorée par une variable ou une valeur (de censure), notée C qui, elle, a été observée.
On considère une variable d'intérêt T (une durée de vie, par exemple). Au lieu d'observer les variables $T_1, T_2, ..., T_n$, qui nous intéressent, on n'observe T_i que lorsque $\{T_i < C\}$, sinon on sait seulement que $\{T_i > C\}$. On parle alors de censure à droite, la plus fréquente.

Remarque 1.2. *Si l'individu a déjà subi l'évènement avant d'être observé (on observe C et non T_i), et que l'on sait que $\{T_i < C\}$, il y a censure à gauche.*
L'évènement est dit censuré par intervalle, si au lieu de l'observer, on sait seulement qu'il est observé entre 2 dates $\{C_1 < T_i < C_2\}$.

Si Y est la variable réellement observée, on utilise la notation $Y_i = T_i \wedge C = \min(T_i, C)$.
Ce type de censure (à droite) est souvent rencontré en fiabilité (pour les durées de vie des pièces fabriquées pendant une période précise), médecine (pour tester l'efficacité d'un traitement), biologie,...
On en distingue trois types :

1.1.3.2.a Censure de type I

Nous sommes dans le cas décrit précédemment où C est une valeur constante fixée.

1.1.3.2.b Censure de type II

Si on observe $Y_i = T_i \wedge T_{(r)}$, où r est un entier fixé, on parle de censure de type II ou de censure jusqu'au $r^{ième}$ "décès", $T_{(r)}$ étant la $r^{ième}$ statistique d'ordre.

Exemple 1.1. *(Droesbeke et al. (1989)) : Pour tester la fiabilité d'un système complexe, on met en état de fonctionnement n systèmes du même type, et on s'arrête lorsque la $r^{ième}$ panne est observée.*

1.1.3.2.c Censure de type III (ou censure aléatoire de type I)

Soient $C_1, ..., C_n$ des v.a. On observe les variables $Y_i = T_i \wedge C_i = \min(T_i, C_i)$.
L'information disponible peut être résumée par :
- la variable Y_i réellement observée,
- l'indicatrice $\delta_i = \mathbf{1}_{\{T_i \leq C_i\}}$

$\delta_i = 1$ si on observe les 'vraies' durées (l'évènement est observé), i.e. $Y_i = T_i$, $\delta_i = 0$ si on observe les durées incomplètes (l'individu est censuré), i.e. $Y_i = C_i$.

Exemple 1.2. *(Koziol & Green (1976)) : Un essai clinique est réalisé sur 211 individus atteints du cancer de la prostate (phase 4) traités par oestrogène (hormone). A la fin de l'étude, 90 meurent du cancer de la prostate, 105 meurent d'autres causes et 16 sont encore vivants. Les censurés à droite sont les 105 + 16 (= 121) individus qui ne sont pas morts du cancer de la prostate (objet de l'étude).*

1.1.3.3 La troncature

Ce mécanisme empêche l'observation de la variable T entièrement (en général les valeurs extrêmes), et engendre une perte d'information (on n'étudie qu'un sous-échantillon). Il y a troncature si l'observation de la variable d'intérêt T n'a lieu que conditionnellement à un évènement A. On dit qu'il y a troncature

- droite lorsque T n'est observable que si elle est inférieure à un seuil C positif fixé ou aléatoire,

- gauche lorsque T n'est observable que si elle est supérieure à C. En général, on observe le couple (T, C), avec $T \geq C$ ou inversement, selon le cas. La troncature gauche apparait par exemple, dans la détection de réserves pétrolières (voir Woodroofe (1985)) : On ne connaît pas le nombre total de gisements, mais on peut observer les n gisements suffisamment grands (supérieurs à C) pour être exploités.

1.1.3.4 Effets de la censure (et la troncature)

1.1.3.4.a Identifiabilité

Le mécanisme de censure est généralement supposé indépendant de l'évènement étudié : on parle de censure non-informative. Cette hypothèse est importante d'un point de vue mathématique et règle le problème d'identifiabilité : Autrement dit, si on connaît la loi des observations on peut déterminer de façon unique (d'identifier) la loi de T. Elle est vérifiée dans le cas d'une censure causée par la fin de l'étude, ou par un individu ayant quitté l'étude (perdu de vue) à cause d'un changement de résidence, par exemple. Ce n'est

pas le cas pour l'arrêt ou le changement de traitement à cause de son inefficacité ou des effets secondaires.

1.1.3.4.b Estimation de la fonction de survie

Soient, $\{T_i, 1 \leq i \leq n\}$, un échantillon de v.a. positives i.i.d., de fonction de répartition F inconnue,
$\{C_i, 1 \leq i \leq n\}$, un échantillon de v.a. de censure positives i.i.d. et indépendantes des $T_i, i = 1, 2, ..., n$, de fonction de répartition G inconnue.
Dans le modèle de censure à droite on observe les couples $(Y_1, \delta_1), (Y_2, \delta_2), ..., (Y_n, \delta_n)$. où $Y_i = \min(T_i, C_i)$, et $\delta_i = \mathbf{1}_{\{T_i \leq C_i\}}$. Les v.a. Y_i sont de fonction de répartition H définie par :

$$H(y) = 1-\mathbb{P}(Y > y) = 1-\mathbb{P}(T \wedge C > y) = 1-\mathbb{P}(T > y)\mathbb{P}(C > y) = 1-\overline{F}(y)\overline{G}(y), \quad y \in R.$$

Dans le cas de données complètes, un estimateur naturel de la survie de la variable d'intérêt T est la survie empirique

$$S_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{T_i > t\}}.$$

Dans le cas de données censurées, on n'observe plus T mais la variable Y. On fait appel alors à l'estimateur de Kaplan-Meier (EKM) qui, lorsqu'il n'y a pas d'ex-aequo (les temps d'évènements - "décès" et censure - sont distincts) est défini par :

$$\widehat{S}_n(y) = \overline{F}_n(y) = \begin{cases} \prod_{i=1}^{n} \left(1 - \frac{1}{n-i+1}\right)^{\delta_{(i)}} & \text{si } y < Y_{(n)} \\ 0 & \text{si } y \geq Y_{(n),} \end{cases}$$

où $Y_{(1)} \leq Y_{(2)} \leq ... \leq Y_{(n)}$ représentent les statistiques d'ordre associées à Y_i.
L'EKM pour la survie de la variable de censure est défini de la même façon par :

$$\overline{G}_n(y) = \begin{cases} \prod_{i=1}^{n} \left(1 - \frac{1}{n-i+1}\right)^{1-\delta_{(i)}} & \text{si } y < Y_{(n)} \\ 0 & \text{si } y \geq Y_{(n).} \end{cases}$$

L'EKM a des propriétés analogues à celles de la fonction de répartition empirique : les théorèmes de convergence (de Glivenko-Cantelli pour la convergence uniforme presque sûre, de Donsker pour la normalité asymptotique,...), ont été étendus à l'EKM (Breslow & Crowley (1974), Peterson (1977), Gill (1980),...), voir le livre de Shorack & Wellner (1986) pour plus de détails.

Remarque 1.3. *Lorsqu'il n'y a pas de censure, L'EKM se réduit à la fonction de survie empirique.*

1.1.4 Mesure de Dépendance

En exploitant les diverses caractérisations de l'indépendance, on peut construire plusieurs formes de dépendance, parmi lesquelles la notion de mélange (qui est un phénomène de dépendance faible) sous ses divers aspects, bénéficie d'un intérêt particulier.

1.1.4.1 Mélangeance

La notion de forte mélangeance a été introduite par Rosenblatt (1956), comme structure de dépendance.

Définition 1.1. *Soit* $\{X_i, i \geq 1\}$ *une suite de variables aléatoires. Pour tout entier naturel non nul n, on définit le coefficient d'alpha-mélange par*

$$\alpha(n) = \sup_k \sup\{|\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|, A \in \mathcal{F}_1^k(X) \text{ et } B \in \mathcal{F}_{k+n}^{\infty}(X), k \in \mathbb{N}\}$$

où $\mathcal{F}_i^k(X)$ *désigne la tribu des évènements engendrés par les* $X_j, i \leq j \leq k$.
La suite est dite alpha-mélangeante ou fortement mélangeante si le coefficient d'alpha-mélange vérifie $\lim_{n\to\infty} \alpha(n) = 0$.

On peut définir deux types de mélanges forts :

• à décroissance géométrique, s'il existe $c > 0$ et $\rho \in]0, 1[$ tels que $\alpha(n) \leq c\rho^n$. Les processus AR et ARMA sont des exemples de processus géométriquement fortement mélangeants.

• à décroissance arithmétique d'ordre $s > 0$, s'il exite un nombres réel c strictement positif tels que $\alpha(n) \leq \frac{c}{n^s}$.

Remarque 1.4. *Le coefficient de mélange* α *est tel que* $0 \leq \alpha \leq \frac{1}{4}$.
Ce coefficient est notamment plus faible que d'autres coefficients de mélange notés β, ϕ, ρ *et* ψ *(Voir Doukhan (1994), Rio (2000)).*

Les résultats obtenus dans le cas de l'α-mélange vont donc concerner une classe plus large de processus.

1.1.4.2 Association

Le concept d'association a été introduit par Esary, Proschan et Walkup (1967), qui s'inspirent de la dépendance positive (plus précisément la dépendance positive par quadrant, en anglais Positive Quadrant Dependence ou PQD) définie par Lehmann (1966) comme suit :

Définition 1.2. *Deux variables aléatoires X et Y sont dites PQD si pour tous nombres réels x et y*

$$\mathbb{P}\left(X \leq x, Y \leq y\right) \geq \mathbb{P}(X \leq x).\mathbb{P}(Y \leq y).$$

Esary *et al.* proposent, à travers l'association, une notion plus forte.
On considère une famille $X = \{X_i, i \in \mathbb{N}\}$ de variables aléatoires réelles, définie sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$.

Définition 1.3. *La famille finie $\{X_1, X_2, ..., X_n\}$ est dite associée si pour tout sous ensemble $I \in \{1, 2, ..., n\}$*

$$Cov\left(\psi_1(X_i, i \in I), \psi_2(X_j, j \in I)\right) \geq 0,$$

pour toutes fonctions non décroissantes ψ_1, ψ_2 de $\mathbb{R}^{|I|} \longrightarrow \mathbb{R}$ pour lesquelles la covariance existe.

Pour rappel, une application h de $\mathbb{R}^n$ dans $\mathbb{R}$ est dite croissante si, pour tout $i = 1, 2, ..., n$, l'application $t \longrightarrow h_i(t) := h(x_1, .., x_{i-1}, t, x_{i+1}, .., x_n)$ est croissante sur $\mathbb{R}$.
Une famille infinie de variables aléatoires est associée si toute sous famille finie est associée.
Les variables indépendantes sont un exemple de variables associées, de même que des variables gaussiennes positivement corrélées (Pitt 1982).

L'association décrit la structure de dépendance positive de modèles rencontrés en analyse de survie, fiabilité, physique statistique (à travers les inégalités FKG du nom de

leurs auteurs Fortuin, Kasteleyn & Ginibre (1971), qui sont des inégalités de corrélation utilisées dans l'étude de modèles de graphes aléatoires et en théorie de la percolation).

Propriétés 1.1. *(Esary & al. (1967) et Suquet (2001))*

1. *Tout sous ensemble d'un ensemble fini de variables aléatoires réelles associées est encore associé.*
2. *Si deux ensembles de variables associées sont indépendants l'un de l'autre, leur union est un ensemble associé.*
3. *Tout singleton formé d'une variable aléatoire réelle X est associé.*
4. *Si $X = (X_1, ..., X_n)$ est associé et si $f_1, ..., f_k$ sont des fonctions toutes croissantes de $\mathbb{R}^n$ dans $\mathbb{R}$ (ou toutes décroissantes), alors le vecteur $Y = (f_1(X), ..., f_k(X))$ est associé.*
5. *Si $X^{(k)} = (X_1^{(k)}, ..., X_n^{(k)})$ est associé pour tout k, et si $X^{(k)}$ converge en loi vers $X = (X_1, ..., X_n)$ lorque k tend vers $+\infty$, alors X est associé.*

Exemple 1.3. *Quelques exemples importants :*

1. ***Famille de variables indépendantes***
   Tout vecteur aléatoire dont les composantes sont indépendantes est associé.
2. ***Processus gaussien*** *(Pitt (1982))*
   Tout vecteur gaussien $(X_1, ..., X_n)$ est associé si et seulement si $Cov(X_i, X_j) \geq 0$, pour tous $i, j \in \{1, ..., n\}$.
3. ***Variables aléatoires binaires***
   Un vecteur aléatoire (X_1, X_2) de variables binaires (qui prennent les valeurs 0 ou 1) est associé si et seulement si sa covariance $Cov(X_1, X_2) \geq 0$.
4. ***Processus linéaire*** *(Doukhan & Louhichi (1999))*
   Soit $(\epsilon_i)_{i\in\mathbb{Z}}$, une suite de variables aléatoires indépendantes prenant les 2 valeurs $\{-1/2, 1/2\}$ avec équiprobabilité. Soit $(X_n)_{n\in\mathbb{Z}}$, une suite aléatoire vérifiant l'équation d'auto-régression
   $$X_n = \frac{1}{2} X_{n-1} + \epsilon_n, \ n \in \mathbb{Z}.$$
   Alors, $X_n = \sum_{i=0}^{\infty} 2^{-i} \epsilon_{n-i}$ presque sûrement, pour tout $n \in \mathbb{Z}$ (qui est une série de termes indépendants qui converge en moyenne quadratique et presque sûrement).
   Le processus linéaire $(X_n)_{n\in\mathbb{Z}}$ est associé.

Durant les années 1980, l'intérêt des variables associées s'est porté sur plusieurs aspects probabilistes et statistiques, à l'image de :

• Newman (1980) qui établit le théorème central limit (TCL) pour des variables strictement stationnaires, après avoir développé une inégalité pour les fonctions caractéristiques qui a longtemps servi dans le cadre de la normalité asymptotique pour le cas associé . Pour des suites stationnaires, Louhichi & Soulier (2002) obtiennent un TCL dans le cas où la variance est infinie. La stationnarité ne sera pas requise par Birkel (1988b) qui propose une taux de convergence dans le TCL.

• Newman & Wright (1981) généralisent le TCL au cas fonctionnel, avant que Cox & Grimmett (1984) ne remplacent la condition de stationnarité par des conditions sur les moments des variables associées.

• Newman (1984) établit la loi forte des grands nombres (SLLN) pour des suites $(X_n, n \in \mathbb{N})$ strictement stationnaires sous l'hypothèse que

$$\frac{1}{n}\sum_{i=1}^{n} Cov(X_1, X_i) \longrightarrow 0,$$

tandis que Birkel (1989) prolonge l'étude à des suites non stationnaires. Cependant, aucun des deux auteurs ne propose un taux de convergence, ce que réaliseront Azevedo (2010) qui utilise l'inégalité exponentielle de Ioannides & Roussas (1999), et Oliveira (2010) qui obtient un taux de convergence optimal (d'ordre $\frac{\log\log n}{\sqrt{n}}$) pour des variables bornées.

• Dabrowski (1985) développe une loi du logarithme itéré fonctionnel (FLIL).

• Birkel (1988a) serait le premier à établir des conditions pour obtenir des inégalités sur les moments.
Pour le cas de données complètes, l'estimateur de la densité f par la méthode du noyau, pour un échantillon $(X_1, X_2, ..., X_n)$ défini par

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

peut être étendu au cas associé. Bagai & Prakasa Rao (1995) montrent que, dans ce cas, l'estimateur est fortement consistant aussi bien ponctuellement qu' uniformément, pour certains ensembles (intervalles fermés de $\mathbb{R}$), alors que Roussas (2000) établit sa normalité asymptotique. Roussas (1991) étudie la consistance uniforme forte des estimateurs à noyau des dérivées d'ordre r de f (pour des suites strictement stationnaires), sous certaines conditions de régularité du noyau et de la fenêtre. Il obtient aussi des taux de convergence. Les deux précédents articles étudient aussi les estimateurs à noyau des fonctions de hasard et de survie.

Masry (2002) prolonge l'étude de Roussas au cas multivarié. Dewan & Prakasa Rao (1999) élaborent un estimateur non paramétrique de la densité, qui est une généralisation de plusieurs estimateurs connus (histogramme, à noyau, séries orthogonales,...). Dewan & Prakasa Rao (2005) donnent une borne (de type Berry-Esseen) pour la déviation moyenne intégrée de f_n par rapport à f. Henriques & Oliveira (2005) proposent un taux de décroissance exponentielle pour l'estimateur à noyau de la densité. Récemment, Douge (2007) établit une nouvelle inégalité exponentielle, qui conduit à un taux de convergence uniforme presque sûre (d'ordre $\left(\frac{\log^2 n}{n}\right)^{\rho/(2\rho+1)}, 0 < \rho < 1$) de f_n sur des ensembles compacts, sous une condition de décroissance géométrique des covariances. Il améliore sensiblement les résultats de Doukhan & Louhichi (2001) et de Masry (2002). Sous la même hypothèse de décroissance des covariances, Henriques & Oliveira (2004) proposent des inégalités exponentielles et obtiennent $\sqrt{\frac{\log n}{n^{1-\delta}h}}$, $\delta \in (0,1)$ comme taux de convergence.

Roussas (1997) étend certains résultats relatifs aux estimateurs des quantiles au cas associé, tandis que Cai & Roussas (1998) établissent des propriétés asymptotiques (convergence uniforme et normalité asymptotique) de l'estimateur de Kaplan-Meier de la fonction de répartition sous association, et Ioannides & Roussas (1999) développent une inégalité exponentielle du type Bernstein-Hoeffding pour des variables bornées et obtiennent un taux de convergence dans la loi forte des grands nombres de l'ordre de $\left(\frac{n}{(\log n)^2}\right)^{1/3}$. Une extension pour des variables associées non bornées est due à Oliveira (2005) avec taux (plus lent) d'ordre $\left(\frac{n}{(\log n)^5}\right)^{1/3}$.

Lin & Li (2007) établissent la normalité asymptotique de l'estimateur à noyau de la médiane conditionnelle pour la norme L^1.

Plus récemment, Guessoum *et al.* (2012) établissent la consistance uniforme forte de l'es-

timateur de Lynden-Bell sous l'association. On peut citer aussi les travaux de Li & Wang (2008) qui développent une notion plus générale que l'association, la LPQD (Linear Positive Quadrant Dependence) due à Newman (1984), Xing & Yang (2010) qui établissent des inégalités exponentielles ainsi que des taux de convergence pour le SLLN, Fu *et al.* (2012), ...
L'intérêt porté à cette structure de dépendance, a permis d'en développer d'autres formes appelées associations positive (ou faible), négative et quasi-association.

1.1.4.2.a Association Positive

Burton, Dabrowski & Dehling (1986) définissent une classe strictement plus large de variables aléatoires incluant l'association : l'association positive ou faible association.

Définition 1.4. *La famille finie* $\{X_1, X_2, ..., X_n\}$ *est dite positivement (ou faiblement) associée si pour tout sous ensembles disjoints* I *et* $J \in \{1, 2, ..., n\}$

$$Cov\left(\psi_1(X_i, i \in I), \psi_2(X_j, j \in J)\right) \geq 0$$

pour toutes fonctions non décroissantes $\psi_1 : \mathbb{R}^{|I|} \longrightarrow \mathbb{R}$, $\psi_2 : \mathbb{R}^{|J|} \longrightarrow \mathbb{R}$ *pour lesquelles la covariance existe.*

Une famille infinie de variables aléatoires est positivement associée (PA) si toute sous famille finie est positivement associée.

Remarque 1.5. *L'association positive est souvent confondue avec l'association : les deux définitions sont ressemblantes, mais pas équivalentes. La définition de la PA est strictement plus faible que celle de l'association, dans le sens où l'association implique la PA. La réciproque est en général fausse. Pour le cas bivarié, Esary et al (1967) montrent les implications strictes suivantes :*
X *et* Y *associées* $\Longrightarrow Cov\left(f(X), g(Y)\right) \geq 0$ *pour toutes fonctions non décroissantes* f *et* g $(\Longrightarrow Cov\left(X, Y\right) \geq 0)$.
Autrement dit, Il existe des variables PA qui ne sont pas associées, comme le montre l'exemple d'un cas binaire dû à Esary et al. *(1967), puis repris et développé par Bulinski & Shashkin (2007, page 09). Pitt (1982) montre qu'une famille* $X = \{X_1, X_2, ..., X_n\}$ *de*

variables aléatoires gaussiennes (centrées) est associée si et seulement elle est positivement correlée (i.e. $Cov(X_i, X_j) \geq 0$). Bulinski et al. (2012) ajoutent que "pour de telles familles, les concepts d'association et de PA coincident".

Parmi les travaux relatifs à la PA, on peut citer ceux de Burton *et al.* (1986) qui démontrent un TCL, Dabrowski & Dehling (1988) qui proposent un LIL ainsi qu'une inégalité de type Berry-Esseen, et parmi les plus récents, Xing & Yang (2010a,2010b,2011).

1.1.4.2.b Association négative

L'association négative a été introduite par Alam & Saxena (1981) et développée par Joag-Dev & Proschan (1983).

Définition 1.5. *La famille finie $\{X_1, X_2, ..., X_n\}$ est dite négativement associée (NA) si pour tout sous ensembles disjoints I et $J \in \{1, 2, ..., n\}$*

$$Cov\left(\psi_1(X_i, i \in I), \psi_2(X_j, j \in J)\right) \leq 0$$

pour toutes fonctions non décroissantes $\psi_1 : \mathbb{R}^{|I|} \longrightarrow \mathbb{R}$, $\psi_2 : \mathbb{R}^{|J|} \longrightarrow \mathbb{R}$ pour lesquelles la covariance existe.

Joag-Dev & Proschan (1983) montrent, en plus des propriétés fondamentales des variables NA, que certaines distributions multivariées (les lois gaussiennes négativement correlées, les lois multinomiales, les lois de Dirichlet,...) possèdent la propriété de NA. L'étude de l'association négative a, depuis, connu son essor (avec les travaux de Newman (1984), Matula (1992), Roussas (1994, 2000)), Bulinski (1996), et plus particulièrement ces dernières années avec les contributions de Su et al (1997) qui proposent des inégalités pour les moments et étudient la convergence faible, Zhang (2001), qui établit un LIL, Baek *et al.* (2005), Jabbari & Azarnoosh (2006), Xing *et al.* (2009), Kuczmaszewska (2009), Sun *et al.* (2010), Doosti & Dewan (2010), Mi-Hwa Ko (2011) pour la convergence complète, Guang & Hui Cai (2011) pour un principe d'invariance fort,...

1.1.4.2.c Quasi-Association (QA)

C'est la plus récente des formes d'association, dûe à Bulinski & Sabanovitch (1998).

Définition 1.6. *Soit $X = \{X_n, n \in \mathbb{N}\}$ une famille de variables aléatoire réelles telle que $EX_j^2 < \infty$ pour tout $j \in \mathbb{N}$. Soient I et J deux sous-ensembles finis disjoints de $\mathbb{N}$. Alors, la famille X est dite quasi-associée si pour toutes fonctions lipschitziennes $f : \mathbb{R}^{|I|} \longrightarrow \mathbb{R}$, $g : \mathbb{R}^{|J|} \longrightarrow \mathbb{R}$, on a*

$$|Cov\left(f(X_i, i \in I), g(X_j, j \in J)\right| \leq \sum_{i \in I} \sum_{j \in J} Lip_i(f) Lip_j(g) \left|Cov(X_i, X_j)\right| \tag{1.1}$$

où les constantes de Lipschitz $Lip_i(f)$ sont telles que pour tous $x = (x_i, i \in I), y = (y_i, i \in I)$ dans $\mathbb{R}^{|I|}$

$$|f(x) - f(y)| \leq \sum_{i \in I} Lip_i(f) \left|x_i - y_i\right|$$

avec

$$Lip_i(f) = \sup_{x_i \neq y_i} \frac{\left|f(x_1, ..., x_{i-1}, x_i, x_{i+1}, ..., x_{|I|}) - f(x_1, ..., x_{i-1}, y_i, x_{i+1}, ..., x_{|I|})\right|}{|x_i - y_i|},$$

le sup étant pris pour $x_1, x_2, ..., x_{|I|}, y_i \in \mathbb{R}$.

La définition de la quasi-association (QA) "*unifie les systèmes aléatoires PA et NA, sous l'hypothèse que la fonction de covariance est sommable (ce qui est habituellement requis pour les théorèmes limites)*" (Bulinski & Shashkin (2007)).

Remarque 1.6. *L'inégalité (1.1) est vérifiée pour des variables PA ou NA (voir Bulinski & Sabanovitch (1998)). Autrement dit, la PA ou la NA impliquent la quasi-association. Ce qui fait de la quasi-association la plus faible forme de dépendance dans l'association. Shashkin (2002) montre que toute suite aléatoire gaussienne, dont la fonction de covariance prend des valeurs positives et négatives, est QA.*

Remarque 1.7. *La même année (1998), une autre définition de la QA a été proposée par Khoshnevisan & Lewis. Cette définition s'apparente plutôt à la PA.*

1.1.4.2.d Faible dépendance et association

Doukhan & Louhichi (1999) proposent d'unifier la condition de faible dépendance à travers la définition suivante :

Définition 1.7. *La suite $(X_n)_{n\in\mathbb{N}}$ de v.a. est dite $(\theta, \mathcal{F}, \psi)-$faiblement dépendante, s'il existe une classe $\mathcal{F}$ de fonctions réelles, une suite $\theta = (\theta_r)_{r\in\mathbb{N}}$ décroissant vers 0 à l'infini, et une fonction ψ d'arguments $(h, k, u, v) \in \mathcal{F}^2 \times \mathbb{N}^2$, telles que pour tous $u-$uple $(i_1, ..., i_u)$ et $v-$uple $(j_1, ..., j_v)$, avec $i_1 \leq ... \leq i_u < i_u + r \leq j_1 \leq ... \leq j_v$, on a*

$$|Cov\left(h(X_{i_1}, ..., X_{i_u}), k(X_{j_1}, ..., X_{j_v})\right)| \leq \psi(h, k, u, v)\theta_r,$$

pour toutes fonctions $h, k \in \mathcal{F}$ définies respectivement sur $\mathbb{R}^u$ et $\mathbb{R}^v$.

Le lemme suivant va spécifier l'association.
Soit $\mathcal{L}$, l'ensemble des fonctions lipschitziennes bornées de $\mathbb{R}^u$ dans $\mathbb{R}, u \in \mathbb{N}^*$.

Lemme 1.1. *Si $(X_n)_{n\in\mathbb{N}}$ est une suite de v.a. associées et centrées, alors $(X_n)_{n\in\mathbb{N}}$ est $(\theta, \mathcal{L}, \psi)-$faiblement dépendante, avec*

$$\theta_r = \sup_i \sum_{j:|i-j|\geq r} Cov(X_i, X_j)$$

et $\psi(h, k, u, v) = min(u, v)Lip(h)Lip(k)$,
où

$$Lip(h) = \sup_{x\neq y} \frac{|h(x) - h(y)|}{\|x - y\|_1},$$

représente le module de Lipschitz de la fonction $h : \mathbb{R}^u \longrightarrow \mathbb{R}$, et $\mathbb{R}^u$ étant muni de la norme L^1.

D'autres formes (mélange fort, processus gaussiens, modèles a représentation markovienne,...) sont aussi spécifiées par des lemmes les rattachant à la définition précédente.

1.1.4.2.e Autres formes de dépendance positive et négative

Il existe d'autres formes de dépendance positive et négative, certaines en relation avec l'association. On peut citer (les abréviations données entre parenthèses correspondent aux initiales des expressions en anglais) :

- la dépendance par quadrant, positive (PQD) et négative (NQD),
- la dépendance linéaire par quadrant, positive (LPQD) et négative (LNQD),
- la dépendance en régression, positive (PRD) et négative (NRD),

- la dépendance conditionnellement décroissante en séquence (CDS) et la dépendance négative en séquence (NDS),...

Remarque 1.8. *A titre indicatif, nous avons les relations suivantes :*
- PQD $\Rightarrow$ Association $\Rightarrow$ LPQD (Li & Wang (2008)).
- Pour un couple de variables aléatoires, PQD $\Longleftrightarrow$ PA (NQD $\Longleftrightarrow$ NA).

Pour plus de détails, nous renvoyons le lecteur aux références suivantes : Lehman (1966), Fortuin & Kasteleyn (1971), Newman & Wright (1981), Block, Savits & Shaked (1982), Li & Wang (2008), Doukhan (2010).

1.1.5 Organisation de la thèse

Cette thèse est organisée en 5 chapitres, suivis d'une brève conclusion et décrits succintement comme suit :

• **Chapitre 1.** On y présente une introduction générale dans laquelle certaines notions intervenant dans la suite, et nécessaires pour une meilleure compréhension de la thèse, sont rappelées ou développées.

• **Chapitre 2.** Ce chapitre aborde le cas de données indépendantes affectées d'une censure à droite. On y étudie la consistance forte de l'estimateur à noyau de la densité.

• **Chapitre 3.** Le chapitre 3 est consacré aux travaux concernant le même estimateur de la densité, toujours pour le cas de variables censurées à droite, mais affectées d'une structure d'alpha-mélange.

• **Chapitre 4.** Ce chapitre est une extension des résultats des chapitres 2 et 3 au cas de variables associées, tout en prolongeant l'étude à l'estimation du mode.

• **Chapitre 5.** Le chapitre 5 constitue une suite du chapitre précédent. On y développe l'étude de la normalité asymptotique de l'estimateur de la densité sous l'hypothèse d'association.

Enfin, nous terminons de manière classique par un mini chapitre pour conclure et énumérer quelques perspectives de recherche, alors que la dernière partie présentera les différentes références citées dans le document.

Il est utile de rappeler que l'étude du mode, relative aux cas indépendant et alpha-mélangeant, n'a pas été abordée afin d'éviter la redondance : les étapes du raisonnement, ainsi que les résultats, sont exactement les mêmes que pour le cas associé, traité au chapitre 4.

Chapitre 2

Cas de données i.i.d. censurées à droite

2.1 Introduction

Soient $T_1, T_2, ..., T_n$, des variables aléatoires indépendantes et identiquement distribuées, de fonction de répartition F et de densité f inconnues, et des temps de censure $C_1, C_2, ..., C_n$, indépendants et identiquement distribués (i.i.d.) et indépendants des $T_i, i = 1, 2, ..., n$, de fonction de répartition G inconnue. Dans le modèle de censure à droite on observe les couples $(Y_1, \delta_1), (Y_2, \delta_2), ..., (Y_n, \delta_n)$. où $Y_i = \min(T_i, C_i)$, et $\delta_i = \mathbf{1}_{\{T_i \leq C_i\}}$. Les v.a. Y_i sont de fonction de répartition H définie par :

$$H(y) = 1 - (1 - F(y))(1 - G(y)) = 1 - \overline{F}(y)\overline{G}(y), \quad y \in R.$$

On pose $\tau_F = \inf\{t \text{ tels que } F(t) = 1\} \leq \infty$. τ_F est la plus petite borne supérieure du support de F. On définit un pseudo-estimateur de f par :

$$\widetilde{f}_n(t) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{t - Y_i}{h}\right) \frac{\delta_i}{\overline{G}(Y_i)},$$

où K est un noyau positif d'intégrale 1, et $h = h(n)$ une suite de nombres réels positifs, telle que $h \longrightarrow 0$ quand $n \longrightarrow \infty$.
Le coefficient de pénalisation $\frac{\delta_i}{\overline{G}(Y_i)}$ a été introduit par Koul, Susarla & Van Ryzin (1981) dans le cadre d'un nouvel estimateur des paramètres du modèle de régression linéaire.
Lorsque G est connue, $\widetilde{f}_n$ estime la densité commune des durées de vie. Mais dans la

pratique, G est en général inconnu. L'estimateur de Kaplan-Meier qui lui est associé a pour expression :

$$G_n(y) = \begin{cases} 1 - \prod_{i=1}^{n} \left(1 - \frac{1-\delta_i}{n-i+1}\right)^{\mathbf{1}\left\{Y_{(i)} \leq y\right\}} & \text{si } y < Y_{(n)} \\ 1 & \text{si } y \geq Y_{(n)} \end{cases}$$

où $Y_{(1)} \leq Y_{(2)} \leq ... \leq Y_{(n)}$ représentent les statistiques d'ordre associées à Y_i. Ainsi l'estimateur de f associé sera défini par :

$$\widehat{f}_n(t) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{t - Y_i}{h}\right) \frac{\delta_i}{\overline{G}_n(Y_i)}.$$

2.2 Hypothèses

Dans toute la suite, c désignera une constante générique positive qui prendra différentes valeurs selon le contexte, mais ne dépendra ni de n, ni des distributions considérées. D'autre part, toutes les limites sont obtenues pour $n \longrightarrow \infty$, sauf indication.

On considère un compact $[0, \tau]$ où $\tau < \tau_F$ ainsi que les hypothèses suivantes :

M1. $\{T_i, i \geq 1\}$ est une suite de v.a., i.i.d., de fonction de répartition F admettant une densité f et des moments d'ordre deux finis,

M2. Les variables de censure $\{C_i, i \geq 1\}$ sont i.i.d., de fonction de répartition G, et indépendantes des $\{T_i, i \geq 1\}$,

K. K est une densité lipschitzienne à support compact, vérifiant $\int uK(u)du = 0$,

D1. La densité $f(.)$ est deux fois continûment différentiable sur $[0, \tau]$,

H. Le paramètre de lissage $h := h(n)$ est tel que $h \longrightarrow 0$, et $\frac{\log n}{nh^3} \longrightarrow 0$.

Théorème 2.1. *Sous les hypothèses M1, M2, K, D1 et H, on a :*

$$\sup_{t\in[0,\tau]} |\widetilde{f}_n(t) - \mathbb{E}\widetilde{f}_n(t)| = \mathcal{O}\left(\sqrt{\frac{\log n}{nh}}\right) \quad p.s., \quad lorsque\ n \to \infty$$

Théorème 2.2. *Sous les hypothèses M1, M2, K, D1 et H, on a :*

$$\sup_{t\in[0,\tau]} |\widehat{f}_n(t) - f(t)| = \mathcal{O}\left\{\max\left(\sqrt{\frac{\log n}{nh}}, h^2\right)\right\} \quad p.s., \quad lorsque\ n \to \infty.$$

2.3 Preuves

Preuve du Théorème 2.1. Pour démontrer ce théorème, on fait appel à l'inégalité de Bernstein qui stipule que, si $(Z_i)_i$ est une suite de variables réelles indépendantes centrées, et s'il existe $M < \infty$, $|Z_1| \leq M$ et $S^2 = E(Z_1^2)$, alors

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

Pour tout $t \in [0,\tau]$ on pose :

$$\tilde{K}(t, Y_i) = \frac{\delta_i}{\overline{G}(Y_i)} K\left(\frac{t-Y_i}{h}\right) - \mathbb{E}\left[\frac{\delta_1}{\overline{G}(Y_1)} K\left(\frac{t-Y_1}{h}\right)\right] =: \tilde{K}_i(t).$$

On a

$$\mathbb{E}(\widetilde{K}_i(t)) = 0 \quad \text{et} \quad \left|\tilde{K}_i(t)\right| \leq \left\|\tilde{K}_i\right\|_\infty \leq \frac{2\|K\|_\infty}{\overline{G}(\tau)} =: M$$

et

$$\widetilde{f}_n(t) - \mathbb{E}\widetilde{f}_n(t) = \frac{1}{nh}\sum_{i=1}^{n} \tilde{K}_i(t).$$

Calculons la variance de $\tilde{K}_i(t)$ désignée par S^2 :

$$\begin{aligned}
S^2 &= Var\left[\frac{\delta_1}{\overline{G}(Y_1)}K\left(\frac{t-Y_1}{h}\right)\right] = \mathbb{E}\left[\frac{\delta_1}{\overline{G}(Y_1)}K\left(\frac{t-Y_1}{h}\right)\right]^2 - \mathbb{E}^2\left[\frac{\delta_1}{\overline{G}(Y_1)}K\left(\frac{t-Y_1}{h}\right)\right] \\
&= \mathbb{E}\left(K^2\left(\frac{t-T_1}{h}\right)\frac{1}{\overline{G}(T_1)}\right) - \mathbb{E}^2K\left(\frac{t-T_1}{h}\right) \\
&= \int \frac{1}{\overline{G}(t_1)}K^2\left(\frac{t-t_1}{h}\right)f(t_1)dt_1 - \left[\int K\left(\frac{t-t_1}{h}\right)f(t_1)dt_1\right]^2 \\
&= h\int K^2(u)\frac{f(t-hu)}{\overline{G}(t-hu)}du - h^2\left[\int K(u)f(t-hu)du\right]^2, \text{ en posant } u = \frac{t-t_1}{h} \\
&= : S_1 - S_2.
\end{aligned}$$

Un développement de Taylor de f au voisinage de t, conjointement avec **K** et **D1** permettent d'avoir

$$\begin{aligned}
S_2 &= h^2\left[\int K(u)\left(f(t) - huf'(t) + \frac{h^2u^2}{2}f''(t^*)\right)du\right]^2, \ t^* \text{ étant compris entre } t \text{ et } t-hu \\
&= h^2\left[f(t)\int K(u)du - hf'(t)\int uK(u)du + \frac{h^2}{2}\int u^2K(u)f''(t^*)du\right]^2 \\
&= h^2\left[f(t) + \frac{h^2}{2}\int u^2K(u)f''(t^*)du\right]^2 \leq h^2\left[f(t) + \frac{h^2}{2}\sup_{t\in[0,\tau]}|f''(t)|\int u^2K(u)du\right]^2 \\
&= \mathcal{O}(h^2).
\end{aligned}$$

D'autre part

$$\frac{S_1}{h} = \int K^2(u)\frac{f(t-hu)}{\overline{G}(t-hu)}du \leq \frac{\|f\|_\infty}{\overline{G}(\tau)}\int K^2(u)du,$$

et donc d'après le théorème de convergence dominée

$$\frac{S_1}{h} \longrightarrow \frac{f(t)}{\overline{G}(t)}\int K^2(u)du \ \text{ lorsque } n \to \infty.$$

On en déduit que $S_1 = \mathcal{O}(h)$ et que

$$S^2 = \mathcal{O}(h) + \mathcal{O}(h^2) = \mathcal{O}(h).$$

Revenons à la preuve du théorème. Aussi, on recouvre l'ensemble compact $[0,\tau]$ en un nombre fini q_n d'intervalles I_j de centres t_j , $j = 1, 2, ..., q_n$ et de même rayon $\frac{\tau}{2q_n}$. Ensuite, on décompose le terme de fluctuations comme suit :

$$\widetilde{f}_n(t) - \mathbb{E}\widetilde{f}_n(t) = \left(\widetilde{f}_n(t) - \widetilde{f}_n(t_j)\right) + \left(\mathbb{E}\widetilde{f}_n(t_j) - \mathbb{E}\widetilde{f}_n(t)\right) + \left(\widetilde{f}_n(t_j) - \mathbb{E}\widetilde{f}_n(t_j)\right),$$

de sorte que

$$\begin{aligned}\sup_{t\in[0,\tau]}\left|\widetilde{f}_n(t)-\mathbb{E}\widetilde{f}_n(t)\right| &\leq \max_{1\leq j\leq q_n}\sup_{t\in I_j}\left|\widetilde{f}_n(t)-\widetilde{f}_n(t_j)\right|+\max_{1\leq j\leq q_n}\sup_{t\in I_j}\left|\mathbb{E}\widetilde{f}_n(t_j)-\mathbb{E}\widetilde{f}_n(t)\right| \\ &+ \max_{1\leq j\leq q_n}\left|\widetilde{f}_n(t_j)-\mathbb{E}\widetilde{f}_n(t_j)\right| \\ &= : I_1+I_2+I_3,\end{aligned}$$

avec

$$\widetilde{f}_n(t)-\widetilde{f}_n(t_j)=\frac{1}{nh}\sum_{i=1}^{n}\frac{\delta_i}{\overline{G}(Y_i)}\left[K\left(\frac{t-Y_i}{h}\right)-K\left(\frac{t_j-Y_i}{h}\right)\right],$$

et

$$\begin{aligned}\left|\widetilde{f}_n(t)-\widetilde{f}_n(t_j)\right| &\leq \frac{1}{nh\overline{G}(\tau)}\sum_{i=1}^{n}\left|K\left(\frac{t-Y_i}{h}\right)-K\left(\frac{t_j-Y_i}{h}\right)\right|=\frac{\left|K\left(\frac{t-Y_1}{h}\right)-K\left(\frac{t_j-Y_1}{h}\right)\right|}{h\overline{G}(\tau)} \\ &\leq \frac{\mu\left|t-t_j\right|}{h^2\overline{G}(\tau)}.\end{aligned}$$

K étant un noyau lipschitzien, et μ la constante de Lipschitz. De plus, $t\in I_j$, implique que

$$|t-t_j|\leq\frac{\tau}{2q_n},$$

et donc

$$\sup_{t\in I_j}\left|\widetilde{f}_n(t)-\widetilde{f}_n(t_j)\right|\leq\frac{\mu\tau}{2q_nh^2\overline{G}(\tau)}.$$

En choisissant $q_n=\mathcal{O}\left(\sqrt{\frac{n}{h^3}}\right)$, la quantité I_1 est ainsi majorée par,

$$I_1=\mathcal{O}\left(\frac{1}{\sqrt{nh}}\right).$$

Par un raisonnement similaire on obtient la même majoration pour I_2 $\left(I_2 \leq \frac{1}{\sqrt{nh}} \frac{\mu\tau}{2\overline{G}(\tau)}\right)$. Concernant I_3, on applique l'inégalité de Bernstein comme suit,

$$\begin{aligned}
&\mathbb{P}\left(\max_{1\leq j\leq q_n}\left|\widetilde{f}_n(t_j)-\mathbb{E}\widetilde{f}_n(t_j)\right|>\varepsilon\right)\\
&\qquad\leq \sum_{j=1}^{q_n}\mathbb{P}\left(\left|\widetilde{f}_n(t_j)-\mathbb{E}\widetilde{f}_n(t_j)\right|>\varepsilon\right)\\
&\qquad\leq q_n\sup_{t\in[0,\tau]}\mathbb{P}\left(\left|\widetilde{f}_n(t)-\mathbb{E}\widetilde{f}_n(t)\right|>\varepsilon\right)\\
&\qquad= q_n\sup_{t\in[0,\tau]}\mathbb{P}\left(\frac{1}{nh}\left|\sum_{i=1}^{n}\tilde{K}_i(t)\right|>\varepsilon\right)\\
&\qquad= q_n\sup_{t\in[0,\tau]}\mathbb{P}\left(\left|\sum_{i=1}^{n}\tilde{K}_i(t)\right|>nh\varepsilon\right)\\
&\qquad\leq 2q_n\exp\left\{-\frac{n(h\varepsilon)^2}{2S^2\left(1+M\frac{h\varepsilon}{S^2}\right)}\right\}\\
&\qquad\approx 2q_n\exp\left\{-\frac{nh\varepsilon^2}{2\left(c+M\varepsilon\right)}\right\}
\end{aligned}$$

avec $S^2 \leq ch$. Pour $\varepsilon = \varepsilon_0\sqrt{\frac{\log n}{nh}}$, il vient

$$\begin{aligned}
\mathbb{P}\left(\max_{1\leq j\leq q_n}\left|\widetilde{f}_n(t_j)-\mathbb{E}\widetilde{f}_n(t_j)\right|>\varepsilon_0\sqrt{\frac{\log n}{nh}}\right) &\leq 2c\sqrt{\frac{n}{h^3}}\exp\left\{-\frac{\varepsilon_0^2\log n}{2\left(c+M\varepsilon_0\sqrt{\frac{\log n}{nh}}\right)}\right\}\\
&= 2c\sqrt{\frac{n}{h^3}}n^{-\frac{\varepsilon_0^2}{2\left(c+M\varepsilon_0\sqrt{\frac{\log n}{nh}}\right)}}\\
&= \frac{2c}{\sqrt{h^3}}n^{\frac{1}{2}\left(1-\frac{\varepsilon_0^2}{c+M\varepsilon_0\sqrt{\frac{\log n}{nh}}}\right)} \approx \frac{2c}{\sqrt{h^3}}n^{\frac{1}{2}\left(1-\frac{\varepsilon_0^2}{c}\right)}
\end{aligned}$$

sous la condition $\frac{\log n}{nh} \longrightarrow 0$. Pour un choix approprié de ε_0, cette dernière quantité peut être considérée comme le terme général d'une série convergente. En appliquant le lemme de Borel-Cantelli, il vient que

$$\max_{1\leq j\leq q_n}\left|\widetilde{f}_n(t_j)-E\widetilde{f}_n(t_j)\right|>\varepsilon=\mathcal{O}\left(\sqrt{\frac{\log n}{nh}}\right) \quad p.s. \text{ lorsque } n\to\infty.$$

Par les majorations obtenues pour les 3 expressions I_1, I_2 et I_3, le Théorème 2.1 est démontré. □

Preuve du Théorème 2.2. On utilise la décomposition standard suivante :

$$\widehat{f}_n(t) - f(t) = \left(\widehat{f}_n(t) - \widetilde{f}_n(t)\right) + \left(\widetilde{f}_n(t) - \mathbb{E}\widetilde{f}_n(t)\right) + \left(\mathbb{E}\widetilde{f}_n(t) - f(t)\right), \quad t \in [0,\tau]$$

qui donne,

$$\begin{aligned}\sup_{t\in[0,\tau]} |\widehat{f}_n(t) - f(t)| &\leq \sup_{t\in[0,\tau]} \left|\mathbb{E}\widetilde{f}_n(t) - f(t)\right| + \sup_{t\in[0,\tau]} \left|\widehat{f}_n(t) - \widetilde{f}_n(t)\right| + \sup_{t\in[0,\tau]} \left|\widetilde{f}_n(t) - \mathbb{E}\widetilde{f}_n(t)\right| \\ &= : J_1 + J_2 + J_3,\end{aligned}$$

avec

$$\mathbb{E}\widetilde{f}_n(t) = \mathbb{E}\left[\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{t-Y_i}{h}\right)\frac{\delta_i}{\overline{G}(Y_i)}\right] = \frac{1}{h}\mathbb{E}\left[K\left(\frac{t-Y_1}{h}\right)\frac{\delta_1}{\overline{G}(Y_1)}\right]$$

par stationnarité des T_i et des C_i. Le calcul d'espérance par conditionnement donne,

$$\mathbb{E}\widetilde{f}_n(t) = \frac{1}{h}\mathbb{E}\left[K\left(\frac{t-T_1}{h}\right)\right]$$

En effet,

$$\begin{aligned}\mathbb{E}\left[K\left(\frac{t-Y_i}{h}\right)\frac{\delta_i}{\overline{G}(Y_i)}\right] &= \mathbb{E}\left[\mathbb{E}\left[\left(K\left(\frac{t-Y_i}{h}\right)\frac{\delta_i}{\overline{G}(Y_i)}\right) \mid T_i\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\left(K\left(\frac{t-Y_i}{h}\right)\frac{\mathbf{1}_{\{T_i\leq C_i\}}}{\overline{G}(Y_i)}\right) \mid T_i\right]\right] \\ &= \mathbb{E}\left[K\left(\frac{t-T_i}{h}\right)\frac{1}{\overline{G}(T_i)}\mathbb{E}\left[\left(\mathbf{1}_{\{T_i\leq C_i\}}\right) \mid T_i\right]\right] \\ &= \mathbb{E}\left[K\left(\frac{t-T_i}{h}\right)\frac{1}{\overline{G}(T_i)}\overline{G}(T_i)\right] = \mathbb{E}\left[K\left(\frac{t-T_1}{h}\right)\right]\end{aligned}$$

ceci du fait que,

$$\mathbb{E}\left[\left(\mathbf{1}_{\{T_i\leq C_i\}}\right) \mid T_i\right] = \mathbb{P}\left((T_i \leq C_i) \mid T_i\right) = \mathbb{P}(T_i \leq C_i) = \overline{G}(T_i)$$

T_i et C_i étant indépendantes. De plus

$$\mathbb{E}\left[K\left(\frac{t-T_1}{h}\right)\right] = \int_R \left(\frac{t-t_1}{h}\right) f(t_1)dt_1 = h\int_{-\infty}^{+\infty} K(u)f(t-hu)du$$

Si **D1** est vérifiée,

$$f(t-hu) = f(t) - huf'(t) + \frac{h^2u^2}{2}f''(t^*)$$

t^* étant compris entre $t-hu$ et t, alors,

$$\mathbb{E}\widetilde{f}_n(t) = \frac{1}{h}\mathbb{E}(K(\frac{t-T_1}{h})) = f(t) - h.f'(t)\int_{-\infty}^{+\infty} uK(u)du + \frac{h^2}{2}\int_{-\infty}^{+\infty} f''(t^*)u^2K(u)du$$

de sorte que le terme de biais

$$\mathbb{E}\widetilde{f}_n(t) - f(t) \leq \frac{h^2}{2}\sup_{t\in[0,\tau]}|f''(t)|\int_{-\infty}^{+\infty} u^2K(u)du,$$

et si $\sup_{t\in[0,\tau]}|f''(t)| < \infty$,

$$J_1 = \sup_{t\in[0,\tau]}|\mathbb{E}\widetilde{f}_n(t) - f(t)| = \mathcal{O}(h^2).$$

Pour J_2, en utilisant le fait que $\overline{G}_n(\tau) \to \overline{G}(\tau)$, on a

$$\begin{aligned}
|\widehat{f}_n(t) - \widetilde{f}_n(t)| &\leq \frac{|G_n(Y_i) - G(Y_i)|}{\overline{G}_n(Y_i)\overline{G}(Y_i)}\frac{1}{nh}\sum\nolimits_{i=1}^{n}\delta_i K\left(\frac{t-Y_i}{h}\right) \\
&\leq \frac{|G_n(Y_i) - G(Y_i)|}{\left(\overline{G}(\tau)\right)^2}\frac{1}{nh}\sum\nolimits_{i=1}^{n} K\left(\frac{t-T_i}{h}\right) \\
&\leq \frac{\sup_{0<y<\tau}|G_n(y) - G(y)|}{\left(\overline{G}(\tau)\right)^2}\frac{1}{nh}\sum\nolimits_{i=1}^{n} K\left(\frac{t-T_i}{h}\right),
\end{aligned}$$

où $\sup_{y\in[0,\tau]}\left|\overline{G}(y) - \overline{G}_n(y)\right| = \mathcal{O}\left(\sqrt{\frac{\log\log n}{n}}\right)$ *p.s.* d'après la loi du logarithme itéré (LIL) pour des données censurées, dans le cas i.i.d. (voir Deheuvels & Einmahl (2000)). D'autre part, la Loi Forte des Grands Nombres nous indique que,

$$\frac{1}{nh}\sum\nolimits_{i=1}^{n} K\left(\frac{t-T_i}{h}\right) = \frac{1}{n}\sum\nolimits_{i=1}^{n}\frac{1}{h}K\left(\frac{t-T_i}{h}\right) \longrightarrow \mathbb{E}\left(\frac{1}{h}K\left(\frac{t-T_1}{h}\right)\right) = \mathcal{O}(1),$$

d'où

$$J_2 = \sup_{t\in[0,\tau]}|\widehat{f}_n(t) - \widetilde{f}_n(t)| = \mathcal{O}\left(\sqrt{\frac{\log\log n}{n}}\right) \quad p.s.$$

De plus, d'après le Théorème 2.1

$$J_3 = \sup_{t\in[0,\tau]}|\widetilde{f}_n(t) - \mathbb{E}\widetilde{f}_n(t)| = \mathcal{O}\left(\sqrt{\frac{\log n}{nh}}\right).$$

On en déduit le résultat recherché. □

Chapitre 3

Cas de données alpha-mélangeantes censurées à droite

3.1 Introduction et Position du problème

On considère des temps de survie $T_1, T_2, ..., T_n$ strictement stationnaires, non négatifs, satisfaisant la propriété d'alpha-mélange, de fonction de répartition F et de densité f inconnues, et des temps de censure $C_1, C_2, ..., C_n$ indépendants et identiquement distribués (i.i.d.) et indépendants des $T_i, i = 1, 2, ..., n$, de fonction de répartition G inconnue. Dans le modèle de censure à droite on observe les couples $(Y_1, \delta_1), (Y_2, \delta_2), ..., (Y_n, \delta_n)$ où $Y_i = \min(T_i, C_i)$, et $\delta_i = \mathbf{1}_{\{T_i \leq C_i\}}$. Les v.a. Y_i sont de fonction de répartition H définie par :

$$H(y) = 1 - (1 - F(y))(1 - G(y)) = 1 - \overline{F}(y)\overline{G}(y), \ y \in R.$$

On pose $\tau_F = \inf\{t \ tq \ F(t) = 1\} \leq \infty$. τ_F est la plus petite borne supérieure du support de F. On définit un pseudo-estimateur de f par :

$$\widetilde{f}_n(t) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{t - Y_i}{h}\right) \frac{\delta_i}{\overline{G}(Y_i)},$$

où K est un noyau positif d'intégrale 1, et $h := h(n)$ une suite de nombres réels positifs, telle que $h \longrightarrow 0$ quand $n \longrightarrow \infty$.
Lorsque G est connue, $\widetilde{f}_n$ estime la densité commune des durées de vie. Mais dans la

pratique, G est en général inconnu. L'estimateur de Kaplan-Meier qui lui est associé a pour expression :

$$G_n(y) = \begin{cases} 1 - \prod_{i=1}^{n} \left(1 - \frac{1-\delta_i}{n-i+1}\right)^{\mathbf{1}\left\{Y_{(i)} \leq y\right\}} & \text{si } y < Y_{(n)} \\ 1 & \text{si } y \geq Y_{(n)} \end{cases}$$

où $Y_{(1)} \leq Y_{(2)} \leq ... \leq Y_{(n)}$ représentent les statistiques d'ordre associées à Y_i. Ainsi l'estimateur de f associé sera défini par :

$$\widehat{f}_n(t) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{t - Y_i}{h}\right) \frac{\delta_i}{\overline{G}_n(Y_i)}.$$

Dans toute la suite, on suppose que les variables d'intérêt T_i, satisfont la condition d'alpha-mélange définie dans l'introduction générale (voir définition 1.1). On s'intéressera particulièrement à des variables arithmétiquement alpha-mélangeantes. Notons que le coefficient d'alpha-mélange à décroissance géométrique est aussi à décroissance arithmétique d'ordre s pour tout $s > 0$.

3.2 Hypothèses

On considère un compact $[0, \tau]$ où $\tau < \tau_F$ ainsi que les hypothèses suivantes :

M1. $\{T_i, i \geq 1\}$ est une suite strictement stationnaire de v.a. alpha-mélangeantes, de fonction de répartition F admettant une densité f et des moments d'ordre deux finis,

M2. Les variables de censure $\{C_i, i \geq 1\}$ sont i.i.d., de fonction de répartition G, et indépendantes des $\{T_i, i \geq 1\}$,

K. K est une densité lipschitzienne à support compact, vérifiant $\int uK(u)du = 0$,

D1. La densité $f(.)$ est deux fois continûment différentiable sur $[0, \tau]$,

D2. La densité jointe $f_{1,j}(.,.)$ de (T_1, T_{j+1}) vérifie :

$$\sup_{j>1} \sup_{u,v \in [0,\tau]} |f_{1,j}(u, v) - f(u)f(v)| < c,$$

D3. Le coefficient d'alpha-mélange des T_i vérifie $\alpha(n) = \mathcal{O}(n^{-\upsilon})$ pour tout $\upsilon > 4$.

H. Le paramètre de lissage $h = h(n)$ est telle que $h \longrightarrow 0$, $nh \longrightarrow \infty$ et

$$\frac{nh^{\frac{\upsilon+4}{\upsilon-4}}}{(\log n)^{\frac{\upsilon+1}{\upsilon-4}} \left[\log(\log n)\right]^{\frac{6}{\upsilon-4}}} \longrightarrow +\infty$$

lorsque $n \longrightarrow \infty$, pour tout $\upsilon > 4$.
Le théorème suivant établit la convergence uniforme presque sûre de $\widehat{f}_n$ vers f.

Théorème 3.1. *Sous les hypothèses **M1**, **M2**, **K**, **D1**, **D2**, **D3** et **H**, nous avons*

$$\sup_{t\in[0,\tau]} |\widehat{f}_n(t) - f(t)| = \mathcal{O}\left\{\max\left(\sqrt{\frac{\log n}{nh}}, h^2\right)\right\} \quad p.s., n \to \infty.$$

3.3 Preuves

La preuve du Thèorème 3.1, est basée sur les lemmes suivants :

Lemme 3.1. *[(Fuk-Nagaev)] Soit $\{X_i, i \in \mathbb{N}\}$ une suite de variables aléatoires réelles centrées, de coefficient d'alpha-mélange $\alpha(n) = \mathcal{O}(n^{-\upsilon})$ pour $\upsilon > 1$, vérifiant pour tout $n \in \mathbb{N}, |X_i| < \infty, 1 \leq i \leq n$. Alors, pour tout $r > 1$*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i\right| > \varepsilon\right) \leq c\left(1 + \frac{\varepsilon^2}{16rS_n^2}\right)^{-r/2} + \frac{nc}{r}\left(\frac{2r}{\varepsilon}\right)^{\upsilon+1},$$

où $S_n^2 = \sum_{1\leq i,j\leq n} |cov(X_i, X_j)|$.

Lemme 3.2. *Sous les hypothèses **K** et **D1**, on a :*

$$\sup_{t\in[0,\tau]} |\mathbb{E}\widetilde{f}_n(t) - f(t)| = \mathcal{O}\left(h^2\right) \quad p.s. \ \ lorsque\ n \to \infty.$$

Preuve. On montre d'abord que

$$\mathbb{E}\widetilde{f}_n(t) = \frac{1}{h}\mathbb{E}\left[K\left(\frac{t - T_1}{h}\right)\right].$$

En effet,

$$\mathbb{E}\widetilde{f}_n(t) = \mathbb{E}\left[\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{t - Y_i}{h}\right)\frac{\delta_i}{\overline{G}(Y_i)}\right] = \frac{1}{h}\mathbb{E}\left[\frac{\delta_1}{\overline{G}(Y_1)}K\left(\frac{t - Y_1}{h}\right)\right],$$

avec

$$\begin{aligned}
\mathbb{E}\left[\frac{\delta_1}{\overline{G}(Y_1)}K\left(\frac{t-Y_1}{h}\right)\right] &= \mathbb{E}\left[\mathbb{E}\left(\frac{\delta_1}{\overline{G}(Y_1)}K\left(\frac{t-Y_1}{h}\right) \mid T_1\right)\right] \\
&= \mathbb{E}\left[\left(\frac{\mathbf{1}_{\{T_1\leq C_1\}}}{\overline{G}(Y_1)}K\left(\frac{t-Y_1}{h}\right)\right) \mid T_1\right] \\
&= \mathbb{E}\left[\frac{1}{\overline{G}(T_1)}K\left(\frac{t-T_1}{h}\right)\mathbb{E}\left(\mathbf{1}_{\{T_1\leq C_1\}} \mid T_1\right)\right] \\
&= \mathbb{E}\left[\frac{1}{\overline{G}(T_1)}K\left(\frac{t-T_1}{h}\right)\mathbb{P}(T_1\leq C_1)\right], \ T_1 \text{ et } C_1 \text{ étant indépendants} \\
&= \frac{1}{\overline{G}(T_1)}\mathbb{E}\left[K\left(\frac{t-T_1}{h}\right)\right]\overline{G}(T_1) \\
&= \mathbb{E}\left[K\left(\frac{t-T_1}{h}\right)\right].
\end{aligned}$$

Ainsi,

$$\begin{aligned}
\mathbb{E}\widetilde{f}_n(t) &= \frac{1}{h}\mathbb{E}\left[K\left(\frac{t-T_1}{h}\right)\right] \\
&= \frac{1}{h}\int K\left(\frac{t-t_1}{h}\right)f(t_1)dt_1 \\
&= \int K(u)f(t-hu)du,
\end{aligned}$$

en posant $u=\frac{t-t_1}{h}$. On utilise un développement de Taylor de f à l'ordre 2 ($f(t-hu) = f(t)-hu\ f'(t)+\frac{h^2u^2}{2}f''(t^*)$, t^* étant compris entre t et $t-hu$) pour obtenir

$$\begin{aligned}
\mathbb{E}\widetilde{f}_n(t) &= \int K(u)\left[f(t)-huf'(t)+\frac{h^2u^2}{2}f''(t^*)\right]du \\
&= \int K(u)f(t)du - hf'(t)\int uK(u)du + \frac{h^2}{2}\int f''(t^*)u^2K(u)du \\
&= f(t)+\frac{h^2}{2}\int f''(t^*)u^2K(u)du \\
&\leq f(t)+\frac{h^2}{2}\sup_{t\in[0,\tau]}|f''(t)|\int u^2K(u)du,
\end{aligned}$$

par l'hypothèse **K**. D'où le biais de $\widetilde{f}_n$

$$\mathbb{E}\widetilde{f}_n(t)-f(t)=\mathcal{O}(h^2).$$

La preuve du Lemme 3.2 est ainsi achevée. □

Lemme 3.3. *Sous les hypothèses **D1** à **D3**, **K** et **H**,*

$$\sup_{t\in[0,\tau]}|\widetilde{f}_n(t)-\mathbb{E}\widetilde{f}_n(t)|=\mathcal{O}\left(\sqrt{\frac{\log n}{nh}}\right)\quad p.s.\ \ lorsque\ n\to\infty.$$

Preuve. L'intervalle $[0,\tau]$ étant compact, on peut le recouvrir par un nombre fini q_n d'intervalles I_j de centres t_j^* , $1\leq j\leq q_n$, et de demi-longueur $a_n=\sqrt{\frac{h^3}{n}}$. $[0,\tau]$ étant borné, il existe une constante $c_1>0$ telle que $q_n\leq c_1\sqrt{\frac{n}{h^3}}$ (en effet, $l([0,\tau])=2q_na_n=2q_n\sqrt{\frac{h^3}{n}}\leq 2c_1\sqrt{\frac{n}{h^3}}\sqrt{\frac{h^3}{n}}=2c_1$).
Pour tout $t\in[0,\tau]$ on pose

$$Z_i(t)=Z(t,Y_i)=\frac{1}{nh}\left[\frac{\delta_i}{\overline{G}(Y_i)}K\left(\frac{t-Y_i}{h}\right)-E\left(\frac{\delta_1}{\overline{G}(Y_1)}K\left(\frac{t-Y_1}{h}\right)\right)\right].$$

On a

$$\sum_{i=1}^{n}Z_i(t)=\widetilde{f}_n(t)-\mathbb{E}\widetilde{f}_n(t),$$

qu'on décompose comme suit

$$\begin{aligned}\sum_{i=1}^{n}Z_i(t) &= \left\{\left[\widetilde{f}_n(t)-\widetilde{f}_n(t_j^*)\right]-\left[\mathbb{E}\widetilde{f}_n(t)-\mathbb{E}\widetilde{f}_n(t_j^*)\right]\right\}+\left[\widetilde{f}_n(t_j^*)-\mathbb{E}\widetilde{f}_n(t_j^*)\right]\\ &= :\sum_{i=1}^{n}\tilde{Z}_i(t)+\sum_{i=1}^{n}Z_i(t_j^*),\end{aligned}$$

d'où

$$\sup_{t\in[0,\tau]}\left|\sum_{i=1}^{n}Z_i(t)\right|\leq\max_{1\leq j\leq q_n}\sup_{t\in I_j}\left|\sum_{i=1}^{n}\tilde{Z}_i(t)\right|+\max_{1\leq j\leq q_n}\left|\sum_{i=1}^{n}Z_i(t_j^*)\right|=:S_1+S_2.$$

D'autre part

$$\begin{aligned}\left|\sum_{i=1}^{n}\tilde{Z}_i(t)\right| &= \left|\widetilde{f}_n(t)-\widetilde{f}_n(t_j^*)-\mathbb{E}\left[\widetilde{f}_n(t)-\widetilde{f}_n(t_j^*)\right]\right|\\ &\leq \frac{1}{nh}\sum_{i=1}^{n}\frac{\delta_i}{\overline{G}(Y_i)}\left|K\left(\frac{t-Y_i}{h}\right)-K\left(\frac{t_j^*-Y_i}{h}\right)\right|\\ &+ \frac{1}{h}\mathbb{E}\left[\frac{\delta_i}{\overline{G}(Y_i)}\left|K\left(\frac{t-Y_i}{h}\right)-K\left(\frac{t_j^*-Y_i}{h}\right)\right|\right]\\ &= :\Sigma_1(t)+\Sigma_2(t),\end{aligned}$$

avec

$$\begin{aligned}\sup_{t\in I_j}\Sigma_1(t) &= \sup_{t\in I_j}\left|\widetilde{f}_n(t)-\widetilde{f}_n(t_j^*)\right| \leq \frac{1}{h\overline{G}(\tau)}\sup_{t\in I_j}\left|K\left(\frac{t-Y_i}{h}\right)-K\left(\frac{t_j^*-Y_i}{h}\right)\right| \\ &\leq \frac{1}{h\overline{G}(\tau)}\frac{\mu\left|t-t_j\right|}{2h} \leq \frac{\mu a_n}{h^2\overline{G}(\tau)},\ K \text{ étant lipschitzienne} \\ &= \frac{\mu}{h^2\overline{G}(\tau)}\sqrt{\frac{h^3}{n}} = \frac{\mu}{\overline{G}(\tau)\sqrt{nh}} = \mathcal{O}\left(\frac{1}{\sqrt{nh}}\right),\end{aligned}$$

μ étant la constante de Lipschitz. De la même manière,

$$\sup_{t\in I_j}\Sigma_2(t) = \mathcal{O}\left(\frac{1}{\sqrt{nh}}\right),$$

ce qui entraine que

$$S_1 = \max_{1\leq j\leq q_n}\sup_{t\in I_j}\left|\sum_{i=1}^{n}\tilde{Z}_i(t)\right| = \mathcal{O}\left(\frac{1}{\sqrt{nh}}\right).$$

Pour l'étude de S_2 on applique le Lemme 3.1.

Pour rappel, Cai(1998, Lemme 1) montre que les variables $Y_1, ..., Y_n$ sont alpha-mélangeantes de coefficient d'alpha-mélange égal à $4\alpha(n)$.

Soit

$$U_i = U_{i,k} := nhZ_i(t_k^*) = \frac{\delta_i}{\overline{G}(Y_i)}K\left(\frac{t_k^*-Y_i}{h}\right) - \mathbb{E}\left[\frac{\delta_1}{\overline{G}(Y_1)}K\left(\frac{t_k^*-Y_1}{h}\right)\right].$$

Calculons d'abord

$$\begin{aligned}S_n^2 &= \sum_{1\leq i,j\leq n}|cov(U_i,U_j)| = \sum_{i\neq j}|cov(U_i,U_j)| + nVar(U_1) \\ &:= \mathcal{V} + nVar(U_1).\end{aligned}$$

Par conditionnement on a

$$\begin{aligned}
Var(U_1) &= Var\left[\frac{\delta_1}{\overline{G}(Y_1)}K\left(\frac{t_k^*-Y_1}{h}\right)\right] \\
&= \mathbb{E}\left[\frac{\delta_1}{(\overline{G}(Y_1))}K\left(\frac{t_k^*-Y_1}{h}\right)\right]^2 - \mathbb{E}^2\left[\frac{\delta_1}{\overline{G}(Y_1)}K\left(\frac{t_k^*-Y_1}{h}\right)\right] \\
&\leq \mathbb{E}\left[\frac{\delta_1}{(\overline{G}(Y_1))}K\left(\frac{t_k^*-Y_1}{h}\right)\right]^2 \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(K\left(\frac{t_k^*-Y_1}{h}\right)\frac{\mathbf{1}_{\{T_1\leq C_1\}}}{\overline{G}(Y_1)}\right)^2 \mid T_1\right]\right] \\
&= \mathbb{E}\left[\left(K\left(\frac{t_k^*-T_1}{h}\right)\frac{1}{\overline{G}(T_1)}\right)^2 \mathbb{E}\left[\mathbf{1}_{\{T_1\leq C_1\}} \mid T_1\right]\right] \\
&= \mathbb{E}\left[\left(K\left(\frac{t_k^*-T_1}{h}\right)\frac{1}{\overline{G}(T_1)}\right)^2 P(T_1\leq C_1)\right] = \mathbb{E}\left[\frac{1}{\overline{G}(T_1)}K^2\left(\frac{t_k^*-T_1}{h}\right)\right] \\
&\leq \frac{1}{\overline{G}(\tau)}\int K^2\left(\frac{t_k^*-y}{h}\right)f(y)dy \\
&= \frac{h}{\overline{G}(\tau)}\int K^2(u)f(t_k^*-hu)du = \mathcal{O}(h),
\end{aligned}$$

c'est-à-dire que

$$nVar(U_1) = \mathcal{O}(nh).$$

D'autre part,

$$\begin{aligned}
|cov(U_i,U_j)| &= |\mathbb{E}(U_iU_j)| \\
&= \left|\mathbb{E}\left(\frac{\delta_i\delta_j}{\overline{G}(Y_i)\overline{G}(Y_j)}K\left(\frac{t_k^*-Y_i}{h}\right)K\left(\frac{t_k^*-Y_j}{h}\right)\right) - \mathbb{E}^2\left[\frac{\delta_1}{\overline{G}(Y_1)}K\left(\frac{t_k^*-Y_1}{h}\right)\right]\right| \\
&= M_1 - M_2,
\end{aligned}$$

avec

$$\begin{aligned}
M_1 &= \mathbb{E}\left(K\left(\frac{t_k^*-Y_i}{h}\right)K\left(\frac{t_k^*-Y_j}{h}\right)\frac{\delta_i\delta_j}{\overline{G}(Y_i)\overline{G}(Y_j)}\right) \\
&= \mathbb{E}\left[\mathbb{E}\left(\left(K\left(\frac{t_k^*-Y_i}{h}\right)K\left(\frac{t_k^*-Y_j}{h}\right)\frac{\delta_i\delta_j}{\overline{G}(Y_i)\overline{G}(Y_j)}\right) \mid T_i,T_j\right)\right] \\
&= \mathbb{E}\left[K\left(\frac{t_k^*-Y_i}{h}\right)K\left(\frac{t_k^*-Y_j}{h}\right)\left(\frac{\mathbb{E}\left(\delta_i\delta_j \mid T_i,T_j\right)}{\overline{G}(T_i)\overline{G}(T_j)}\right)\right],
\end{aligned}$$

où

$$\begin{aligned}
\mathbb{E}\left(\delta_i\delta_j \mid T_i, T_j\right) &= \mathbb{E}\left[\mathbf{1}_{\{T_i\leq C_i\}}\mathbf{1}_{\{T_j\leq Cj\}} \mid T_i, T_j\right] = \mathbb{E}\left[\mathbf{1}_{\{T_i\leq C_i\}\cap\{T_j\leq Cj\}} \mid T_i, T_j\right] \\
&= \mathbb{P}\left(\{T_i \leq C_i\}\cap\{T_j \leq C_j\} \mid T_i, T_j\right) = \mathbb{P}\left(\{T_i \leq C_i\}\cap\{T_j \leq C_j\}\right) \\
&= \mathbb{P}\left(T_i \leq C_i\right)\mathbb{P}\left(\{T_j \leq C_j\} \mid \{T_i \leq C_i\}\right) = \overline{G}(T_i)\mathbb{P}\left(\{T_j \leq C_j\} \mid \{T_i \leq C_i\}\right) \\
&\leq \overline{G}(T_i)\mathbb{P}\left(T_j \leq C_j\right) = \overline{G}(T_i)\overline{G}(T_j),
\end{aligned}$$

de sorte que sous les hypothèses **K** et **D2**,

$$\begin{aligned}
|cov(U_i, U_j)| &= \left|\mathbb{E}\left[K\left(\frac{t_k^*-T_i}{h}\right)K\left(\frac{t_k^*-T_j}{h}\right)\frac{\mathbb{P}\left(\{T_j\leq c_j\} \mid \{T_i\leq c_i\}\right)}{\overline{G}(T_j)}\right]\right. \\
&\quad \left. -\mathbb{E}^2\left[K\left(\frac{t_k^*-T_1}{h}\right)\right]\right| \\
&\leq \left|\mathbb{E}\left[K\left(\frac{t_k^*-T_i}{h}\right)K\left(\frac{t_k^*-T_j}{h}\right)\right] - \mathbb{E}^2\left[K\left(\frac{t_k^*-T_1}{h}\right)\right]\right| \\
&= \left|\int\int K\left(\frac{t_k^*-w_1}{h}\right)K\left(\frac{t_k^*-w_2}{h}\right)f_{i,j}(w_1,w_2)dw_1dw_2\right. \\
&\quad \left. -\int\int K\left(\frac{t_k^*-w_1}{h}\right)K\left(\frac{t_k^*-w_2}{h}\right)f(w_1)f(w_2))dw_1dw_2\right| \\
&= h^2\int\int K(t_1)K(t_2)\left|f_{i,j}(t_k^*-t_1h, t_k^*-t_2h) - f(t_k^*-t_1h)f(t_k^*-t_2h)\right|dt_1dt_2 \\
&\leq h^2\sup_{j>1}\left\|f_{1,j}-f\otimes f\right\|_\infty.
\end{aligned}$$

Autrement dit :

$$|cov(U_i, U_j)| = \mathcal{O}(h^2).$$

De plus, en utilisant la décomposition de Masry (1986), définissons

$$\mathcal{B}_1 = \{(i,j); 1\leq|i-j|\leq\eta_n\} \text{ et } \mathcal{B}_2 = \{(i,j); \eta_n+1\leq|i-j|\leq n-1\},$$

où $\eta_n = o(n)$. On décompose $\mathcal{V}$ comme suit :

$$\begin{aligned}
\mathcal{V} &= \sum_{i=1}^{n}\sum_{\mathcal{B}_1}|cov(U_i,U_j)| + \sum_{i=1}^{n}\sum_{\mathcal{B}_2}|cov(U_i,U_j)| \\
&:= \mathcal{V}_1 + \mathcal{V}_2,
\end{aligned}$$

avec, d'après (3.3)

$$\mathcal{V}_1 = \mathcal{O}(nh^2\eta_n).$$

Pour $\mathcal{V}_2$ on applique l'inégalité de Davydov modifiée pour le cas mélangeant (voir Rio (2000)). Aussi, pour $i \neq j$,

$$|cov(U_i, U_j)| \leq c\alpha(|i-j|),$$

ce qui entraine, sous la condition **D3**, que

$$\begin{aligned}
\mathcal{V}_2 &\leq c\sum_{i=1}^{n}\sum_{\mathcal{B}_2}\alpha(|i-j|) \leq cn \sum_{\eta_n+1\leq|i-j|\leq n-1}\alpha(|i-j|) \\
&\leq cn\int_{\eta_n}^{n} u^{-\upsilon}du = \frac{cn}{1-\upsilon}\left[u^{1-\upsilon}\right]_{\eta_n}^{n} \\
&= \frac{cn}{\upsilon-1}\left[(\eta_n)^{1-\upsilon}-(n)^{1-\upsilon}\right] \leq \frac{cn(\eta_n)^{1-\upsilon}}{\upsilon-1},
\end{aligned}$$

qui se traduit par

$$\mathcal{V}_2 = \mathcal{O}(n\eta_n^{1-\upsilon}).$$

En choisissant $\eta_n = h^{-2/\upsilon}, \upsilon > 4$, on aura

$$\mathcal{V} = \mathcal{V}_1 + \mathcal{V}_2 = \mathcal{O}(nh^2).$$

(En effet, $\mathcal{V}_1 = \mathcal{O}(nh^2\eta_n) \leq cnh^{2(1-1/\upsilon)} = \mathcal{O}(nh^2)$,
et
$\mathcal{V}_2 = \mathcal{O}(n\eta_n^{1-\upsilon}) \leq cnh^{-2\frac{1-\upsilon}{\upsilon}} = cnh^{2(1-1/\upsilon)} = \mathcal{O}(nh^2)$).
Donc

$$S_n^2 = \sum_{1\leq i,j\leq n}|cov(U_i,U_j)| = \mathcal{O}(nh) + \mathcal{O}(nh^2) = \mathcal{O}(nh).$$

A ce stade de la démonstration, on applique l'inégalité du Lemme 3.1

$$\begin{aligned}
\mathbb{P}\left(|\widetilde{f}_n(t_k^*) - E\widetilde{f}_n(t_k^*)| > \varepsilon\right) &= \mathbb{P}\left(\left|\sum_{i=1}^{n} Z_i(t_k^*)\right| > \varepsilon\right) \\
&= \mathbb{P}\left(\frac{1}{nh}\left|\sum_{i=1}^{n} U_i\right| > \varepsilon\right) = \mathbb{P}\left(\left|\sum_{i=1}^{n} U_i\right| > nh\varepsilon\right) \\
&\leq c\left(1+\frac{(nh\varepsilon)^2}{16rS_n^2}\right)^{-r/2} + ncr^{-1}\left(\frac{2r}{nh\varepsilon}\right)^{\upsilon+1} \\
&= : c\left(\mathcal{E}_1+\mathcal{E}_2\right).
\end{aligned}$$

Pour $\varepsilon = \varepsilon_0\sqrt{\frac{\log n}{nh}}$ et en utilisant (3.3), on aura

$$\mathcal{E}_1 = \left(1 + \frac{(nh\varepsilon)^2}{16rS_n^2}\right)^{-r/2} = \left[1 + \frac{\varepsilon_0^2}{16c\left[\log\left(\log n\right)\right]^{1/\upsilon}}\right]^{-\frac{c}{2}},$$

et

$$\mathcal{E}_2 = nr^{-1}\left(\frac{2r}{\varepsilon_0\sqrt{nh\log n}}\right)^{\upsilon+1} = nc_2\frac{r^\upsilon}{\varepsilon_0^{\upsilon+1}}\left(nh\log n\right)^{-(r+1)/2}.$$

En choisissant $r = c\log n\left[\log\left(\log n\right)\right]^{1/\upsilon}$,

$$\begin{aligned}\mathcal{E}_1 &= \left[1 + \frac{(nh\varepsilon_0)^2\log n}{16cS_n^2 nh\log n\left[\log\left(\log n\right)\right]^{1/\upsilon}}\right]^{-\frac{c}{2}\log n[\log(\log n)]^{1/\upsilon}} \\ &= \left[1 + \frac{\varepsilon_0^2}{16c\left[\log\left(\log n\right)\right]^{1/\upsilon}}\right]^{-\frac{c}{2}\log n[\log(\log n)]^{1/\upsilon}}.\end{aligned}$$

En passant au logarithme,

$$\log\mathcal{E}_1 = -\frac{c}{2}\log n\left[\log\left(\log n\right)\right]^{1/\upsilon}\log\left(1 + \frac{\varepsilon_0^2}{16c\left[\log\left(\log n\right)\right]^{1/\upsilon}}\right)$$

puis en utilisant un développement de Taylor de $\log(1+x)$,

$$\begin{aligned}\log\mathcal{E}_1 &\simeq -\frac{c}{2}\log n\left[\log\left(\log n\right)\right]^{1/\upsilon}\left(\frac{\varepsilon_0^2}{16c\left[\log\left(\log n\right)\right]^{1/\upsilon}}\right) \\ &= -\frac{\varepsilon_0^2}{32}\log n = \log n^{-\frac{\varepsilon_0^2}{32}}.\end{aligned}$$

On en déduit que

$$\mathcal{E}_1 = n^{-\frac{\varepsilon_0^2}{32}}.$$

Pour le même choix de ε et r, on a

$$\begin{aligned}\mathcal{E}_2 &= nr^{-1}\left(\frac{2r}{\varepsilon_0\sqrt{nh\log n}}\right)^{\upsilon+1} = \frac{n}{c\log n\left[\log\left(\log n\right)\right]^{1/\upsilon}}\left(\frac{2c\log n\left[\log\left(\log n\right)\right]^{1/\upsilon}}{\varepsilon_0\sqrt{nh\log n}}\right)^{\upsilon+1} \\ &= 2(2c)^\upsilon\varepsilon_0^{-(\upsilon+1)}h^{-(\frac{\upsilon+1}{2})}n^{\frac{1-\upsilon}{2}}(\log n)^{\frac{\upsilon-1}{2}}\log\left(\log n\right).\end{aligned}$$

En reprenant l'inégalité de Fuk-Nagaev on peut écrire que

$$\begin{aligned}
\mathbb{P}\left(\max_{1\leq k\leq q_n}\left|\sum_{i=1}^{n} Z_i(t_k^*)\right| > \varepsilon_0\sqrt{\frac{\log n}{nh}}\right) &\leq \sum_{k=1}^{q_n}\mathbb{P}\left(\left|\sum_{i=1}^{n} Z_i(t_k^*)\right| > \varepsilon_0\sqrt{\frac{\log n}{nh}}\right)\\
&\leq cq_n\left(n^{-\frac{\varepsilon_0^2}{32}} + \varepsilon_0^{-(\upsilon+1)}h^{-(\frac{\upsilon+1}{2})}n^{\frac{1-\upsilon}{2}}(\log n)^{\frac{\upsilon-1}{2}}\log(\log n)\right)\\
&\leq cc_1\sqrt{\frac{n}{h^3}}\left(n^{-\frac{\varepsilon_0^2}{32}} + \varepsilon_0^{-(\upsilon+1)}h^{-(\frac{\upsilon+1}{2})}n^{\frac{1-\upsilon}{2}}\right.\\
&\quad\left.\times(\log n)^{\frac{\upsilon-1}{2}}\log(\log n)\right)\\
&\leq cc_1\left(n^{\frac{1}{2}-\frac{\varepsilon_0^2}{32}}h^{-3/2} + \varepsilon_0^{-(\upsilon+1)}h^{-(\frac{\upsilon+4}{2})}n^{\frac{2-\upsilon}{2}}\right.\\
&\quad\left.\times(\log n)^{\frac{\upsilon-1}{2}}\log(\log n)\right)\\
&= : cc_1\left(\Delta_1+\Delta_2\right).
\end{aligned}$$

L'hypothèse **H** peut être transformée comme suit

$$\frac{1}{n}\left[h^{-(\upsilon+4)}(\log n)^{\upsilon+1}(\log(\log n))^6\right]^{1/(\upsilon-4)} \longrightarrow 0 \text{ lorsque } n\longrightarrow\infty, \text{pour tout } \upsilon>4$$

ou encore, en élevant à la puissance $(\upsilon-4)/2$

$$h^{-\frac{(\upsilon+4)}{2}}n^{-\frac{(\upsilon-4)}{2}}(\log n)^{\frac{\upsilon+1}{2}}(\log(\log n))^3 \longrightarrow 0,$$

et donc

$$h^{-\frac{(\upsilon+4)}{2}} = o\left(n^{\frac{(\upsilon-4)}{2}}(\log n)^{-\frac{\upsilon+1}{2}}(\log(\log n))^{-3}\right).$$

On peut alors réecrire Δ_2 sous la forme

$$\begin{aligned}
\Delta_2 &= \varepsilon_0^{-(\upsilon+1)}\left(n^{\frac{(\upsilon-4)}{2}}(\log n)^{-\frac{\upsilon+1}{2}}(\log(\log n))^{-3}\right)n^{\frac{2-\upsilon}{2}}(\log n)^{\frac{\upsilon-1}{2}}\log(\log n)\\
&= o\left(\frac{1}{n\log n(\log(\log n))^2}\right),
\end{aligned}$$

qui s'apparente au terme général d'une série de Bertrand convergente. □

Lemme 3.4. *Sous les hypothèses **D1**, **D2**, **D3**, **K** et **H**,*

$$\sup_{t\in[0,\tau]}\left|\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{t-T_i}{h}\right) - f(t)\right| = \mathcal{O}\left\{\max\left(\sqrt{\frac{\log n}{nh}}, h^2\right)\right\} \quad p.s.$$

Preuve. On décompose l'expression précédente comme suit :

$$\begin{aligned}\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{t-T_i}{h}\right)-f(t) &= \sum_{i=1}^{n}\left[\frac{1}{nh}K\left(\frac{t-T_i}{h}\right)-\mathbb{E}\left(\frac{1}{nh}K\left(\frac{t-T_1}{h}\right)\right)\right]\\ &+ \left[\mathbb{E}\left(\frac{1}{h}K\left(\frac{t-T_1}{h}\right)\right)-f(t)\right]\\ &= : A_1+A_2,\end{aligned}$$

avec

$$\begin{aligned}A_2 &= \mathbb{E}\left(\frac{1}{h}K\left(\frac{t-T_1}{h}\right)\right)-f(t)=\int\frac{1}{h}K\left(\frac{t-t_1}{h}\right)f(t_1)dt_1-f(t)\\ &= \int K(u)f(t-hu)du-f(t)=\mathbb{E}\widetilde{f}_n(t)-f(t)=Biais(\widetilde{f}_n)\\ &= O(h^2),\ p.s.,\end{aligned}$$

d'après le Lemme 3.2.

Pour A_1 on raisonne de la même façon que pour le Lemme 3, à la différence qu'ici,

$$Z_i(t)=\frac{1}{nh}\left(K\left(\frac{t-Y_i}{h}\right)-E\left[K\left(\frac{t-Y_1}{h}\right)\right]\right)$$

et que les q_n boules I_j qui recouvrent le compact $[0,\tau]$ sont de rayon $r_n=a_n\overline{G}(\tau)$. Ainsi, tout comme pour le Lemme 3.3, $A_1=\mathcal{O}\left(\sqrt{\frac{\log n}{nh}}\right)$, et par conséquent le Lemme 3.4 est démontré. □

Lemme 3.5. *Sous les hypothèses **D1**, **D3**, **K** et **H**,*

$$\sup_{t\in[0,\tau]}|\widehat{f}_n(t)-\widetilde{f}_n(t)|=\mathcal{O}\left(\sqrt{\frac{\log\log n}{n}}\right)\quad p.s.$$

Preuve.

$$\widehat{f}_n(t)-\widetilde{f}_n(t)=\frac{1}{nh}\sum_{i=1}^{n}\delta_iK\left(\frac{t-Y_i}{h}\right)\left(\frac{1}{\overline{G}_n(Y_i)}-\frac{1}{\overline{G}(Y_i)}\right).$$

Ce qui implique que

$$\begin{aligned}\left|\widehat{f}_n(t)-\widetilde{f}_n(t)\right| &\leq \frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{t-Y_i}{h}\right)\frac{\left|\overline{G}(Y_i)-\overline{G}_n(Y_i)\right|}{\overline{G}(Y_i)\overline{G}_n(Y_i)}\\ &\leq \frac{\sup\limits_{y\in[0,\tau]}\left|\overline{G}(y)-\overline{G}_n(y)\right|}{\overline{G}(\tau)\overline{G}_n(\tau)}\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{t-Y_i}{h}\right),\end{aligned}$$

où $\sup_{y\in[0,\tau]}\left|\overline{G}(y)-\overline{G}_n(y)\right| = \mathcal{O}\left(\sqrt{\frac{\log\log n}{n}}\right)$ *p.s.*, d'après la loi du logarithme itéré (LIL) pour des données censurées, dans le cas i.i.d. (voir Deheuvels & Einmahl (2000)). D'autre part, selon la loi forte des grands nombres pour le cas alpha-mélangeant établie par Cai & Roussas (1992) (Théorème 2.3), pour n assez grand on a :

$$\frac{1}{nh}\sum\nolimits_{i=1}^{n} K\left(\frac{t-Y_i}{h}\right) \longrightarrow E\left(\frac{1}{h}K\left(\frac{t-Y_i}{h}\right)\right) = \mathcal{O}(1) \ \ p.s.,$$

(sous réserve que la condition $\frac{1}{n}\sum_{n=1}^{\infty}\log n.(\log\log n)^{1+\delta}\alpha(n) < \infty$, $\delta > 0$, soit assurée). Le résultat est alors une conséquence du Lemme 3.4.
Les Lemmes 3.1, 3.3 et 3.5 permettent de démontrer le Théorème 3.1. □

Preuve du Théorème 3.1. On décompose l'expression comme suit

$$\widehat{f}_n(t)-f(t) = \left(\widehat{f}_n(t)-\widetilde{f}_n(t)\right)+\left(\widetilde{f}_n(t)-\mathbb{E}\widetilde{f}_n(t)\right)+\left(\mathbb{E}\widetilde{f}_n(t)-f(t)\right), \quad t\in[0,\tau]$$

Par conséquent,

$$\sup_{t\in[0,\tau]}|\widehat{f}_n(t)-f(t)| \leq \sup_{t\in[0,\tau]}\left|\widehat{f}_n(t)-\widetilde{f}_n(t)\right| + \sup_{t\in[0,\tau]}\left|\widetilde{f}_n(t)-\mathbb{E}\widetilde{f}_n(t)\right| + \sup_{t\in[0,\tau]}\left|\mathbb{E}\widetilde{f}_n(t)-f(t)\right|$$

avec

$$\sup_{t\in[0,\tau]}|\mathbb{E}\widetilde{f}_n(t)-f(t)| = \mathcal{O}\left(h^2\right) \ \ p.s.,$$

d'après le Lemme 3.2. En appliquant le Lemme 3.3 on obtient

$$\sup_{t\in[0,\tau]}|\widetilde{f}_n(t)-\mathbb{E}\widetilde{f}_n(t)| = \mathcal{O}\left(\sqrt{\frac{\log n}{nh}}\right) \ \ p.s.,$$

et

$$\sup_{t\in[0,\tau]}\left|\widehat{f}_n(t)-\widetilde{f}_n(t)\right| = \mathcal{O}\left(\sqrt{\frac{\log\log n}{n}}\right) \ \ p.s.,$$

selon le Lemme 3.5. D'où le résultat. □

Chapitre 4

Cas de données censurées-associées : consistance

Estimateur à noyau de la densité et du mode pour des données associées et censurées *

Soit $\{T_i, i \geq 1\}$ une suite strictement stationnaire de variables aléatoires associées, de même loi que T. Ce travail a pour but d'établir la consistence uniforme forte sur un ensemble compact, avec taux de convergence, de l'estimateur à noyau de la densité f, losque la variable d'intérêt T est censurée à droite par une autre variable C. Comme conséquence, on montre la convergence presque sûre avec vitesse d'un nouvel estimateur à noyau $\hat{\theta}_n$ du vrai mode θ de f.

4.1 Introduction et motivation

On considère une suite de variables aléatoires (v.a.) $\{T_n, n \geq 1\}$ définies sur un même espace de probabilité $(\Omega, \mathscr{F}, \mathbb{P})$. Rappelons que les v.a. $\{T_i; 1 \leq i \leq n\}$ sont dites associées si, pour toute paire de fonctions ψ_1 et ψ_2 de $\mathbb{R}^n$ dans $\mathbb{R}$, non décroissantes (composante par composante), on a :

$$\mathrm{Cov}[\psi_1(T_i, 1 \leq i \leq n), \psi_2(T_j, 1 \leq j \leq n)] \geq 0,$$

∗. Ce chapitre a fait l'objet d'une publication dans Communications in Statistics-Theory & Methods.

lorsque cette covariance existe. Une suite infinie de v.a. est dite associée, si tout sous ensemble fini de ces v.a. est associé.

En inférence statistique classique, les v.a. d'intérêt observées sont généralement supposées indépendantes et identiquement distribuées (i.i.d.). Cependant, dans plusieurs systèmes de durées de vie réels, il est plus commun que les composants soient dépendants. Par exemple, en fiabilité ou en analyse de survie, les v.a. des durées de vie ne sont pas indépendantes mais associées.

Le concept d'association a été introduit et défini par Esary et al. (1967), principalement pour des raisons pratiques. Par exemple, l'association survient souvent dans certains problèmes en théorie de la fiabilité, aussi bien que dans plusieurs modèles importants employés en mécanique statistique, dans certains problèmes de tests d'hypothèses, ou dans les études d'essais cliniques. Pour plus de détails sur le concept d'association, nous renvoyons le lecteur à Bulinski & Shashkin (2007). Dans ce livre, on peut trouver bon nombre de nouveaux résultats et exemples, aussi bien qu'une revue sur l'association et les concepts relatifs sur plus de 30 années. Il est utile de mentionner aussi que ce livre ne traite pas uniquement de suites aléatoires associées, mais aussi de champs aléatoires associés.

Lors d'un suivi médical ou dans les études de survie dans les sciences de l'ingénieur, on ne peut pas observer complètement la variable d'intérêt (durée de vie). Parmi les différentes formes où les données incomplètes apparaissent, la censure et la troncature en sont les deux principales. Notons que dans les essais cliniques, les temps de survie sont souvent dépendants et censurés, ceci étant dû à la nature des expériences.

Ici, on s'intéresse à l'estimation de la densité pour un modèle de censure à droite, donné par une collection de v.a. associées $\{T_i, 1 \leq i \leq n\}$ (temps de survie), strictement stationnaires, de même fonction de répartition (f.r.) F admettant une densité de probabilité (d.p.) f relativement à la mesure de Lebesgue , et $\{C_i, 1 \leq i \leq n\}$ (temps de censure) de densité G. En outre, on suppose que les temps de censure sont i.i.d. et indépendants des temps de survie.

Dans le modèle de censure à droite, on n'observe que les n paires (Y_i, δ_i) avec $Y_i = \min(T_i, C_i)$ et $\delta_i = \mathbf{1}_{\{T_i \leq C_i\}}$, où $\mathbf{1}_{\{A\}}$ est la fonction indicatrice de l'évènement A. L'indépendance des T_i par rapport aux C_i permet d'assurer l'identifiabilité du modèle et ainsi, la fonction de répartition H de Y_1 vérifie $\bar{H} := 1 - H = (1 - G)(1 - F)$. Le problème

actuel consiste en la mise en oeuvre des méthodes d'estimation et d'inférence dans un cadre non paramétrique pour f et son mode θ, basées sur les n paires observées (Y_i, δ_i). Dans ce genre de modèle, il est bien connu que la fonction de répartition empirique n'estime pas convenablement F et G. C'est pourquoi Kaplan et Meier (1958) ont proposé des estimateurs consistants F_n et G_n pour F et G, respectivement définis par

$$F_n(t) = 1 - \prod_{i=1}^{n} \left[1 - \frac{\delta_{(i)}}{n-i+1}\right]^{\mathbf{1}_{\{Y_{(i)} \leq t\}}} \text{ et } G_n(t) = 1 - \prod_{i=1}^{n} \left[1 - \frac{1-\delta_{(i)}}{n-i+1}\right]^{\mathbf{1}_{\{Y_{(i)} \leq t\}}},$$

où $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$ sont les statistiques d'ordre de $Y_1, Y_2, \ldots, Y_n$ et $\delta_{(i)}$ est le concomitant de $Y_{(i)}$. Ces estimateurs ont été étudiés en profondeur (voir Stute & Wang, 1993 ; Deheuvels & Einmahl, 2000) dans le cas i.i.d., alors que Cai & Roussas (1998) établissent la consistance uniforme forte et la normalité asymptotique pour F_n sous l'hypothèse d'association.

Supposons maintenant que f possède un maximum unique au point (mode) θ, défini par

$$\theta := \arg\sup_{t \in \mathbb{R}} f(t).$$

Basé sur des techniques de lissage, un quasi-estimateur à noyau pour f est donné par

$$\tilde{f}_n(t) = \frac{1}{nh_n} \sum_{i=1}^{n} \frac{\delta_i}{1 - G(Y_i)} K\left(\frac{t - Y_i}{h_n}\right), \tag{4.1}$$

où K est une densité de probabilité (appelée noyau) et $h_n =: h > 0$ est une suite de nombres réels (appelés fenêtres) tendant vers zéro lorsque n tend vers l'infini. En pratique, l'estimateur dans (4.1) ne peut pas être calculé vu que G est inconnue, mais joue un grand rôle quant à l'établissement de nos résultats. Par conséquent, un estimateur approprié pour f est obtenu en remplaçant (plug-in) $G(\cdot)$ par son estimateur de Kaplan-Meier $G_n(\cdot)$, c'est-à-dire

$$\hat{f}_n(t) = \frac{1}{nh} \sum_{i=1}^{n} \frac{\delta_i}{1 - G_n(Y_i)} K\left(\frac{t - Y_i}{h}\right). \tag{4.2}$$

Un estimateur à noyau naturel du mode θ est alors

$$\hat{\theta}_n := \arg\max_{t \in \mathbb{R}} \hat{f}_n(t).$$

Notons qu'en l'absence de censure ($\delta_i = 1$), l'estimateur $G_n(\cdot)$ est partout nul et ainsi (4.2) devient l'estimateur à noyau classique de la densité, $f_n(\cdot)$ (voir Parzen, 1962). Ce

dernier estimateur a été largement étudié dans les deux cas i.i.d. et mélangeant.
Dans le cas de données complètes associées, l'estimateur f_n a fait l'objet de plusieurs travaux, comme par exemple ceux de Bagai & Prakasa Rao (1995), Dewan & Prakasa Rao (1999,2005) et Roussas (1991,2000).
Il est utile de souligner que Rezapour et al. (2013), mettent en avant l'intérêt d'incorporer des notions générales de dépendance comme l'association, les concepts de dépendance positive, les mesures de dépendance, dans l'analyse des données pour l'estimation des densités et fonctions de répartition.
Jusqu'ici, et à notre connaissance, il n'y a pas de résultats asymptotiques disponibles pour les estimateurs à noyau de la densité pour des modèles de données associées sous censure aléatoire à droite. C'est la raison pour laquelle, par une contribution modeste, on établit un résultat de consistance uniforme forte avec vitesse, sur un ensemble compact, pour l'estimateur à noyau de la densité f défini dans (4.2), lorsque les durées de vie sont aléatoirement censurées à droite par les temps de censure et forment une suite associée strictement stationnaire. En guise d'application, on exhibe un taux de convergence presque sûre de l'estimateur à noyau $\hat{\theta}_n$ du vrai mode θ, défini en (4.3) précédemment.
La suite de ce travail est organisée comme suit : Les notations requises ainsi que les hypothèses sont introduites à la Section 2 ; quelques commentaires sur les hypothèses y figurent également. Les résultats principaux sont établis dans ce même paragraphe, mais leurs preuves sont reportées à la Section 3.

4.2 Hypothèses et principaux résultats

Dans la suite, pour toute f.r. L, on considère que τ_L est défini par

$$\tau_L = \inf\{t : L(t) = 1\} \leq \infty.$$

Ainsi pour la fonction de répartition marginale H des Y_i, il vient que $\tau_H = \tau_F \wedge \tau_G$ (voir Stute & Wang, 1993).
On considère aussi que $\mathscr{C} := [0, \tau] \subset [0, \tau_H[$ est un ensemble compact tel que $\theta \in \mathscr{C}$. A cause de ces restrictions et sans perte en généralités, nous réduisons notre définition du mode au réel $\theta := \arg\max_{t \in \mathscr{C}} f(t)$, ce qui implique la nécessité de modifier la définition

précédente de l'estimateur à noyau du mode en

$$\hat{\theta}_n := \arg\max_{t\in\mathscr{C}} \hat{f}_n(t). \tag{4.3}$$

Tout au long de cette étude, c désignera une constante positive qui peut prendre une valeur différente d'une expression à l'autre.
Nous regroupons les hypothèses ci-après pour faciliter les renvois.

M1. $\{T_i; i \geq 1\}$ est une suite strictement stationnaire de variables aléatoires associées, de f.r. F et de densité de probabilité f, admettant des seconds moments finis ,

M2. Les variables temps de censure $\{C_i; i \geq 1\}$ sont i.i.d. de f.r. G, et sont indépendantes des $\{T_i; i \geq 1\}$,

M3. Le terme de covariance vérifie : $\rho(r) := \sup_{j:|\ell-j|\geq r} Cov(T_j, T_\ell)$ pour tous $\ell \geq 1$ et $r > 0$, où $\rho(r) \leq \gamma_0 e^{-\gamma r}$ pour toutes constantes positives γ_0 and γ.

K. K est une densité lipschitzienne à support compact vérifiant $\int uK(u)du = 0$.

D1. La densité $f(\cdot)$ est deux fois continûment différentiable sur $\mathscr{C}$, de dérivée seconde $f^{(2)}(\cdot)$ telle que $f^{(2)}(\theta) < 0$,

D2. La densité jointe $f_{1,j}(\cdot\,,\cdot)$ de (T_1, T_j) vérifie : $\sup_{j>1}\sup_{u,v\in\mathscr{C}} |f_{1,j}(u,v) - f(u)f(v)| \leq c$,

D3. le mode θ satisfait à la propriété : pour tous $\varepsilon > 0$ et t, il existe $\delta > 0$ tel que $|\theta - t| \geq \varepsilon \Rightarrow |f(\theta) - f(t)| \geq \delta$.

H. La fenêtre h vérifie : $h \to 0$, $nh \to \infty$ et $\frac{\log^5 n}{nh} \to 0$, lorsque $n \to +\infty$.

Remarque 4.1. *(Quelques commentaires sur les hypothèses) : Les conditions* **M1-M2** *décrivent le modèle étudié ici. L'hypothèse* **M3** *mesure la tendance progressive vers l'indépendance asymptotique entre le "passé" et le "futur". Cette condition a été utilisée dans Doukhan & Neumann (2007) pour établir une inégalité exponentielle qui sera appliquée dans la preuve du Théorème 4.1. La condition* **K** *est assez faiblement restrictive et est fréquemment utilisée dans l'étude de la consistance uniforme d'estimateurs. Les conditions* **D1** *et* **D3** *sont standards dans l'estimation à noyau du mode alors que* **D2** *est souvent utilisée dans l'estimation non paramétrique lorsque les données présentent une structure de dépendance. Si* $\sup_{j,u,v} f_{1,j}(u,v) < \infty$*, il n'est pas difficile de vérifier que* **D2** *est satisfaite*

puisque $\sup_u f(u) = f(\theta)$. *Enfin, la condition* **H** *est principalement utilisée pour établir des résultats asymptotiques, en particulier, la dernière partie de* **H** *est utile pour montrer que le taux de convergence optimal établi dans le Théorème 4.1 est préservé dans le contexte d'association.*

Pour établir nos résultats, on définit

$$\tilde{K}_i(t,h) = \frac{\delta_i}{\bar{G}(Y_i)} K\left(\frac{t-Y_i}{h}\right) - \mathbb{E}\left[\frac{\delta_1}{\bar{G}(Y_1)} K\left(\frac{t-Y_1}{h}\right)\right],$$

pour tout $i = 1, \ldots, n$ et chaque nombre réel t. Evidemment, on a

$$\tilde{f}_n(t) - \mathbb{E}\left[\tilde{f}_n(t)\right] = \frac{1}{nh}\sum_{i=1}^{n} \tilde{K}_i(t,h). \tag{4.4}$$

Proposition 4.1. *Soient* $\tilde{K}_1(t,h), \tilde{K}_2(t,h), \ldots, \tilde{K}_n(t,h)$ *définis comme précédemment. Alors il existe des constantes* M, L_1, $L_2 < \infty$, μ, $\lambda \geq 0$ *et une suite non décroissante de coefficients réels* $(\phi(n))_{n\geq 1}$, *telles que pour tout* u*-uplet* $(s_1, \ldots, s_u)$ *et tout* v*-uplet* $(w_1, \ldots, w_v)$ *with* $1 \leq s_1 \leq \cdots \leq s_u \leq w_1 \leq \cdots \leq w_v \leq n$, *on a*

(a) $Cov\left(\prod_{i=s_1}^{s_u} \tilde{K}_i(t,h), \prod_{j=w_1}^{w_v} \tilde{K}_j(t,h)\right) \leq c^{u+v} huv\phi(w_1 - s_u)$,

(b) $\sum_{s=0}^{\infty} (s+1)^{k_0}\phi(s) \leq L_1 L_2^{k_0}(k_0!)^{\mu}$, $\forall k_0 \geq 0$,

(c) $\mathbb{E}\left[\left|\tilde{K}_i(t,h)\right|^{k_0}\right] \leq (k_0!)^{\lambda} M^{k_0}$.

Remarque 4.2. *Les points de la Proposition 4.1 se rapprochent des conditions du Théorème 1 de Doukhan & Neumann (2007). Ils nous permettront d'utiliser leur inégalité exponentielle afin de justifier les résultats à venir.*

Théorème 4.1. *Sous les hypothèses* **M1**-**M3**, **K**, **D1**, **D2** *et* **H**, *on a*

$$\sup_{t\in\mathscr{C}} \left|\tilde{f}_n(t) - \mathbb{E}\left[\tilde{f}_n(t)\right]\right| = O\left(\sqrt{\frac{\log n}{nh}}\right) \text{ p.s., lorsque } n \longrightarrow \infty.$$

Théorème 4.2. *Sous les hypothèses du Théorème 4.1, on a*

$$\sup_{t\in\mathscr{C}} |\hat{f}_n(t) - f(t)| = O\left\{\max\left(\sqrt{\frac{\log n}{nh}}, h^2\right)\right\} \text{ p.s., lorsque } n \longrightarrow \infty.$$

Remarque 4.3. *Notons que si on choisit* $h = O\left(\left(\frac{\log n}{n}\right)^{1/5}\right)$*, qui est le choix optimal relativement au critère de convergence uniforme presque sûre dans l'estimation à noyau de la densité (voir Stute, 1982), alors on obtient*

$$\sup_{t\in\mathscr{C}} |\hat{f}_n(t) - f(t)| = O\left(\left(\frac{\log n}{n}\right)^{2/5}\right) \text{ p.s., lorsque } n \longrightarrow \infty,$$

qui est optimal au sens minimax et établi par Bertail et al. (2006, p.350) pour des données complètes sous association (sous notre hypothèse **D1** *avec leur* $\rho = 2$ *et* $d = 1$*).*

Corollaire 4.1. *Sous les conditions du Théorème 4.1 et* **D3***, on a*

$$|\hat{\theta}_n - \theta| = O\left\{\max\left(\left(\frac{\log n}{nh}\right)^{1/4}, h\right)\right\} \text{ p.s., lorsque } n \longrightarrow \infty.$$

De plus, si on choisit $h = O\left(\left(\frac{\log n}{n}\right)^{1/5}\right)$*, alors*

$$|\hat{\theta}_n - \theta| = O\left(\left(\frac{\log n}{n}\right)^{1/5}\right) \text{ p.s., lorsque } n \longrightarrow \infty.$$

4.3 Etude de simulation

L'objet de cette section est de mettre en évidence les résultats théoriques obtenus dans le Théorème 4.2 et Corollaire 4.1, en étudiant par simulation le comportement, à distance finie, des estimateurs $\hat{f}_n$ et $\hat{\theta}_n$. Pour ce faire, les variables T, C et Y ont été générées selon le procédé suivant :

- *Etape 1.* génération de $(n+2)$ variables aléatoires i.i.d. Z de loi $\mathcal{N}(0,1)$.
- *Etape 2.* génération de n variables aléatoires associées $T_i = \exp((Z_{i-1} + Z_{i-2})/2)$, de loi $\log\mathcal{N}(0, 1/2)$.
- *Etape 3.* génération de n variables aléatoires C_i de loi $\mathcal{E}(\lambda = 7.5)$.
- *Etape 4.* génération de n variables aléatoires associées $Y_i = \min(T_i, C_i)$.

Pour le calcul du $MSE(\hat{\theta}_n)$, le procédé est répété $B = 300$ fois et l'estimation de $\hat{\theta}_n$ est faite à l'aide de la méthode de dichotomie pour approcher $\arg zero\hat{f}_n'$ sur l'intervalle $]0, 6]$

$\tau(C)$	n	$GMSE(\hat{f}_n, h)$	$MISE(\hat{f}_n, h)$	$MSE(\hat{\theta}_n, h)$	h
$\simeq$ **10**%	**50**	4.60×10^{-3}	2.76×10^{-2}	3.54×10^{-5}	0.34
	100	3.50×10^{-3}	2.06×10^{-2}	2.03×10^{-5}	0.29
	300	1.80×10^{-3}	1.06×10^{-2}	1.26×10^{-5}	0.24
$\simeq$ **20**%	**50**	5.70×10^{-3}	3.43×10^{-2}	6.24×10^{-5}	0.34
	100	3.50×10^{-3}	2.10×10^{-2}	2.38×10^{-5}	0.29
	300	1.90×10^{-3}	1.11×10^{-2}	1.29×10^{-5}	0.24
$\simeq$ **30**%	**50**	6.60×10^{-3}.	3.97×10^{-2}	8.07×10^{-5}	0.34
	100	4.30×10^{-3}	2.55×10^{-2}	4.07×10^{-5}	0.29
	300	2.20×10^{-3}	1.31×10^{-2}	1.77×10^{-5}	0.24
$\simeq$ **20**%	**1000**	9.10×10^{-4}	5.40×10^{-3}	6.90×10^{-6}	0.19

TABLE 4.1 – Illustration des performances de $\hat{f}_n$ et $\hat{\theta}_n$

.

avec une erreur d'approximation de l'ordre 10^{-7}.
Le critère $GMSE(\hat{f}_n)$ (MSE global) qui est une moyenne du $MSE(\hat{f}_n)$) a été évalué aux points $t_k = 0.02k; k = 1, \cdots, B.$ (discrétisation de l'intervalle $]0, 6]$).
Enfin, la méthode des trapèzes a été utilisée pour évaluer le critère $MISE(\hat{f}_n)$.
Ce procédé de simulation a été appliqué pour différentes tailles n et différents taux de censure ($\tau(C)$) comme illustré dans la Table 4.1.

Les résultats obtenus montrent une dégradation des critères de mesure considérés lorsque le taux de censure augmente et s'améliorent lorsque la taille de l'échantillon augmente.
Enfin, comme les performances des estimateurs ne sont pas très sensibles aux variations du taux de censure, différents ajustements pour $\tau(C) \simeq 20\%$ et $n = 50, 100$ et 300 sont illustrés. Les graphiques montrent l'existence d'un effet de bord au voisinage de $t_1 = 0.02$ dans tous les cas et s'atténue progressivement quand n augmente. Cet effet de bord disparaît complètement lorsque $n \longrightarrow \infty$ comme le montre la Figure 4.2.

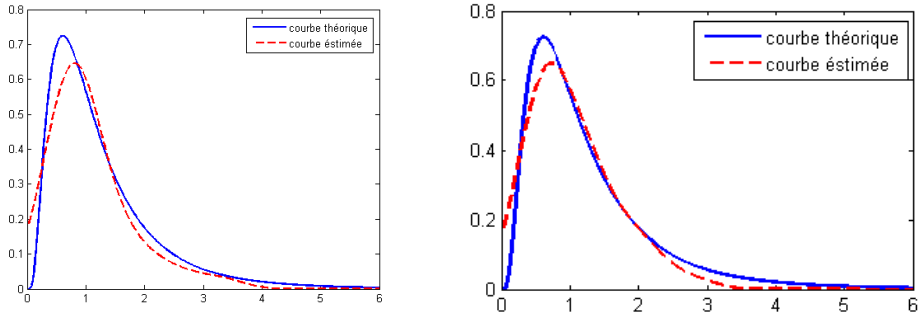

FIGURE 4.1 – $n = 50$ et 100, $\tau(C) \simeq 20$

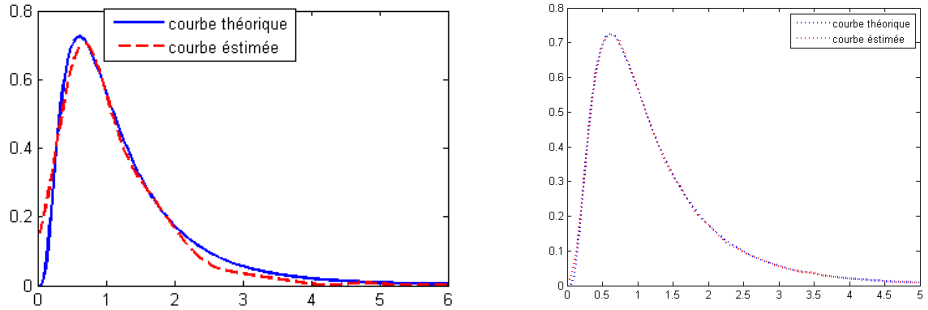

FIGURE 4.2 – $n = 300$ et 10^5, $\tau(C) \simeq 20$

4.4 Preuves

Preuve de la Proposition 4.1 : La partie (a) sera établie en combinant (i) et (ii) du lemme suivant :

Lemme 4.1. *Si les hypothèses du Théorème 4.1 sont vérifiées, alors*

$$(i)\ Cov\left(\prod_{i=s_1}^{s_u}\tilde{K}_i(t,h),\prod_{j=w_1}^{w_v}\tilde{K}_j(t,h)\right) =: Cov_1 \leq c^{u+v}h^{-2}uv\rho(w_1-s_u),$$

$$(ii)\ Cov\left(\prod_{i=s_1}^{s_u}\tilde{K}_i(t,h),\prod_{j=w_1}^{w_v}\tilde{K}_j(t,h)\right) =: Cov_2 \leq c^{u+v}h^{2}.$$

Preuve du Lemme 4.1 : Le terme de covariance dans la partie gauche est borné de différentes façons, aussi bien en utilisant le Théorème 5.3 dans Bulinski & Shashkin

(2007) que par majoration directe.
Commençons d'abord par le point (i). En utilisant le fait que $Cov(Y_i, Y_j) \leq Cov(T_i, T_j)$ (voir Cai & Roussas, 1998, résultat 2.13) et le Théorème 5.3 dans Bulinski & Shashkin (2007), on a

$$Cov\left(\prod_{i=s_1}^{s_u}\tilde{K}_i(t,h), \prod_{j=w_1}^{w_v}\tilde{K}_j(t,h)\right) \leq Lip\left(\prod_{i=s_1}^{s_u}\tilde{K}_i(t,h)\right) Lip\left(\prod_{j=w_1}^{w_v}\tilde{K}_j(t,h)\right) \times \sum_{i=s_1}^{s_u}\sum_{j=w_1}^{w_v} Cov(T_i, T_j),$$

où $Lip(\Phi)$ désigne le module de continuité de Lipschitz de Φ, défini par,

$$Lip(\Phi) = \sup_{x\neq y} \frac{|\Phi(x) - \Phi(y)|}{\|x - y\|_1}$$

with $\|(z_1, \ldots, z_n)\|_1 = |z_1| + \cdots + |z_n|$.
Maintenant, puisque $Lip\left(\prod_{i=1}^{k}\tilde{K}_i(t,h)\right) \leq \frac{1}{h}\left(\frac{2}{\bar{G}(\tau)}\right)^k \|K\|_\infty^{k-1} Lip(K)$, alors par l'hypothèse **M1** on obtient

$$Cov\left(\prod_{i=s_1}^{s_u}\tilde{K}_i(t,h), \prod_{j=w_1}^{w_v}\tilde{K}_j(t,h)\right) \leq \frac{2^{u+v}}{h^2(\bar{G}(\tau))^{u+v}} \|K\|_\infty^{u+v-2} \left(Lip(K)\right)^2 uv\rho(w_1 - s_u).$$

D'autre part, par les hypothèses **M1**, **K** et **D2**, le résultat dans (ii) est immédiat, ce qui complète la preuve du Lemme 4.1. □
Maintenant, en posant $\phi(\cdot) = \rho^{\frac{1}{4}}(\cdot)$, la borne supérieure dans le point (a) devient $Cov_1^{1/4} Cov_2^{3/4}$. Les preuves de (b) et (c) sont similaires à ceux de la Proposition 8 (voir Doukhan & Neumann, 2007) en choisissant $\lambda = 0, \mu = 1$ et $L_1 = L_2 = \frac{1}{1-e^{-\gamma/4}}$ et donc seront omises. Ceci termine la preuve de la Proposition 4.1. □

Preuve du Théorème 4.1 : Puisque les T_i sont indépendants des C_i, nous utiliserons par la suite, et sans le préciser, le fait que $\mathbb{E}\left[\frac{\delta_i}{\bar{G}(Y_i)} K\left(\frac{t - Y_i}{h}\right)\right] = \mathbb{E}\left[K\left(\frac{t - T_1}{h}\right)\right]$. Rappelons que le résultat principal permettant de majorer le terme de fluctuation dans (4.4) est une inégalité exponentielle due à Doukhan & Neumann (2007), qui stipule que pour toutes fonctions aléatoires centrées $\tilde{K}_1(t,h), \tilde{K}_2(t,h), \ldots, \tilde{K}_n(t,h)$ satisfaisant les conditions de

la Proposition 4.1 et pour tout $\varepsilon > 0$,

$$\mathbb{P}\left(\sum_{i=1}^{n} \tilde{K}_i(t,h) \geq \varepsilon\right) \leq \exp\left(-\frac{\varepsilon^2/2}{A_n + B_n^{1/(\mu+\lambda+2)}\varepsilon^{(2\mu+2\lambda+3)/(\mu+\lambda+2)}}\right), \tag{4.5}$$

où A_n peut être choisi tel que $A_n \leq \sigma_n^2$ avec

$$\sigma_n^2 := Var\left(\sum_{i=1}^{n} \tilde{K}_i(t,h)\right),$$

et

$$B_n = 2cL_2\left(\frac{2^{4+\mu+\lambda}cnhL_1}{A_n} \vee 1\right).$$

Avant de continuer, nous devons calculer $\sigma_n^2 = (nh)^2 Var(\tilde{f}_n(t))$. En effet

$$\begin{aligned} nhVar(\tilde{f}_n(t)) &= \frac{1}{h}Var\left(\frac{\delta_i}{\bar{G}(Y_i)}K\left(\frac{t-Y_i}{h}\right)\right) \\ &+ \frac{1}{nh}\sum_{i=1}^{n}\sum_{j=1:j\neq i}^{n} Cov\left(\frac{\delta_i}{\bar{G}(Y_i)}K\left(\frac{t-Y_i}{h}\right), \frac{\delta_j}{\bar{G}(Y_j)}K\left(\frac{t-Y_j}{h}\right)\right) \\ &=: V + \frac{S}{nh}. \end{aligned}$$

D'une part, en utilisant les techniques de l'espérance conditionnelle , on a

$$\begin{aligned} V &= \frac{1}{h}\mathbb{E}\left[\frac{1}{\bar{G}(T_1)}K^2\left(\frac{t-T_1}{h}\right)\right] - \frac{1}{h}\mathbb{E}^2\left[K\left(\frac{t-T_1}{h}\right)\right] \\ &= \int K^2(u)\frac{f(t-hu)}{\bar{G}(t-hu)}du - h\left[\int K(u)f(t-hu)du\right]^2. \end{aligned}$$

De plus, par le Théorème de Convergence Dominée, un développement de Taylor et l'hypothèse **D1**, on obtient

$$V \longrightarrow \frac{f(t)}{\bar{G}(t)}\int K^2(u)du.$$

D'autre part, grâce à l'hypothèse **D2** on a

$$\begin{aligned} Cov\left(\frac{\delta_i}{\bar{G}(Y_i)}K\left(\frac{t-Y_i}{h}\right), \frac{\delta_j}{\bar{G}(Y_j)}K\left(\frac{t-Y_j}{h}\right)\right) &= \mathbb{E}\left[\frac{\mathbb{E}\left[\delta_i\delta_j|T_i,T_j\right]}{\bar{G}(T_i)\bar{G}(T_j)}K\left(\frac{t-T_i}{h}\right)K\left(\frac{t-T_j}{h}\right)\right] \\ &- \mathbb{E}^2\left[\frac{\mathbb{E}\left[\delta_1|T_1\right]}{\bar{G}(T_1)}K\left(\frac{t-T_1}{h}\right)\right] \\ &= O(h^2). \end{aligned} \tag{4.6}$$

Remarque 4.4. *On notera que ce dernier résultat peut être facilement déduit du Lemme 4.1 (ii). On a rajouté l'étape intermédiaire uniquement pour montrer au lecteur comment l'obtenir.*

De plus, en suivant la technique de Masry (1986), on définit

$$\mathcal{B}_1 = \{(i,j); 1 \leq |i-j| \leq \eta_n\} \text{ and } \mathcal{B}_2 = \{(i,j); \eta_n + 1 \leq |i-j| \leq n-1\}$$

où $\eta_n = o(n)$. Alors on écrit

$$\begin{aligned} S &= \sum_{i=1}^{n} \sum_{\mathcal{B}1} Cov\left(\frac{\delta_i}{\bar{G}(Y_i)} K\left(\frac{t-Y_i}{h}\right), \frac{\delta_j}{\bar{G}(Y_j)} K\left(\frac{t-Y_j}{h}\right)\right) \\ &+ \sum_{i=1}^{n} \sum_{\mathcal{B}_2} Cov\left(\frac{\delta_i}{\bar{G}(Y_i)} K\left(\frac{t-Y_i}{h}\right), \frac{\delta_j}{\bar{G}(Y_j)} K\left(\frac{t-Y_j}{h}\right)\right) \\ &=: S_1 + S_2. \end{aligned}$$

Et, par (4.6) et l'hypothèse **M3** on obtient

$$\frac{S_1}{nh} = O\left(\eta_n h\right), \tag{4.7}$$

et

$$\begin{aligned} \frac{S_2}{nh} &\leq \frac{c}{nh} \sum_{i=1}^{n} \sum_{\mathcal{B}2} h e^{-\frac{\gamma|i-j|}{4}} \leq c \int_{\eta_n}^{n} e^{-\frac{\gamma u}{4}} du \\ &= O\left(e^{-\frac{\gamma \eta_n}{4}}\right). \end{aligned} \tag{4.8}$$

En choisissant $\eta_n = O(h^{\nu-1})$ avec $0 < \nu < 1$, les termes dans (4.7) et (4.8) sont de l'ordre de $o(1)$. Par conséquent

$$nhVar(\tilde{f}_n(t)) \longrightarrow \frac{f(t)}{\bar{G}(t)} \int K^2(u)du,$$

et donc

$$\sigma_n^2 = nh \frac{f(t)}{\bar{G}(t)} \int K^2(u)du + o\left(nh\right).$$

Ainsi, on peut choisir

$$A_n = nh \frac{f(t)}{\bar{G}(t)} \int K^2(u)du = O(nh) \text{ and } B_n = O(1).$$

On peut maintenant appliquer (4.5) et revenir à la preuve de notre théorème principal. A ce propos, soit $q_n = O\left(\sqrt{\frac{n}{h^3}}\right)$ un entier (en prenant sa partie entière). On divise le compact $\mathscr{C}$ en q_n segments égaux (plus précisément, un segment et $q_n - 1$ demi-intervalles "half intervals"). Soit I_k le k-ième segment et t_k son centre, $k = 1, ..., q_n$. On a alors

$$\begin{aligned}\sup_{t\in\mathscr{C}}\left|\tilde{f}_n(t) - \mathbb{E}\left[\tilde{f}_n(t)\right]\right| &\leq \max_{1\leq k\leq q_n}\sup_{t\in \mathrm{I}_k}\left|\tilde{f}_n(t) - \tilde{f}_n(t_k)\right| + \max_{1\leq k\leq q_n}\sup_{t\in \mathrm{I}_k}\left|\mathbb{E}\left[\tilde{f}_n(t_k) - \tilde{f}_n(t)\right]\right| \\ &\quad + \max_{1\leq k\leq q_n}\left|\tilde{f}_n(t_k) - \mathbb{E}\left[\tilde{f}_n(t_k)\right]\right| \\ &=: I_1 + I_2 + I_3.\end{aligned}$$

Pour traiter I_1, on remarque d'abord que $|t - t_k| \leq \frac{\tau}{2q_n}$, ensuite on utilise les hypothèses **K** et **H** pour obtenir

$$I_1 \leq \frac{c\tau}{2\bar{G}(\tau)\sqrt{nh}} = O\left(\frac{1}{\sqrt{nh}}\right).$$

Des arguments similaires à ce qui précède donnent $I_2 = O\left(\frac{1}{\sqrt{nh}}\right)$. Pour I_3, on pose $\varepsilon = \varepsilon_0\sqrt{\frac{\log n}{nh}}$ et $\mu^2 = \int K^2(u)du$. Alors

$$\begin{aligned}\mathbb{P}\left(\max_{1\leq j\leq q_n}\left|\tilde{f}_n(t_j) - \mathbb{E}\left[\tilde{f}_n(t_j)\right]\right| > \varepsilon_0\sqrt{\frac{\log n}{nh}}\right) &\leq \sum_{j=1}^{q_n}\mathbb{P}\left(\left|\tilde{f}_n(t_j) - \mathbb{E}\left[\tilde{f}_n(t_j)\right]\right| > \varepsilon_0\sqrt{\frac{\log n}{nh}}\right) \\ &\leq q_n\sup_{t\in\mathscr{C}}\mathbb{P}\left(\left|\tilde{f}_n(t) - \mathbb{E}\left[\tilde{f}_n(t)\right]\right| > \varepsilon_0\sqrt{\frac{\log n}{nh}}\right) \\ &\leq 2q_n\exp\left\{\frac{-\frac{\varepsilon_0^2}{2}\log n}{\frac{f(t)}{\bar{G}(t)}\mu^2 + (\varepsilon_0)^{5/3}\left(\frac{\log^5 n}{nh}\right)^{1/6}}\right\} \\ &\approx \frac{2}{\sqrt{h^3}}n^{-\frac{\bar{G}(\tau)\varepsilon_0^2}{2f(\theta)\mu^2}+\frac{1}{2}}. \qquad (4.9)\end{aligned}$$

Pour un choix approprié de ε_0, cette dernière quantité peut être considérée comme le terme général d'une série convergente. A titre d'exemple, pour tout $\beta\,;\, 0 < \beta < 1$, si on choisit $\varepsilon_0 > \sqrt{3(1+\beta)\frac{f(\theta)\mu^2}{\bar{G}(\tau)}}$, alors la convergence de la série est assurée aussi bien pour $h = \left(\frac{\log n}{n}\right)^\beta$ que pour $h = n^{-\beta}$. En appliquant le lemme de Borel-Cantelli à (4.9) il vient que

$$\max_{1\leq j\leq q_n}\left|\tilde{f}_n(t_j) - \mathbb{E}\left[\tilde{f}_n(t_j)\right]\right| = O\left(\sqrt{\frac{\log n}{nh}}\right) \text{ p.s., lorsque } n \longrightarrow \infty.$$

Cette dernière expression permet d'obtenir le résultat escompté. Le Théorème 4.1 est ainsi démontré. □

Preuve du Théorème 4.2 : on utilise la décomposition standard suivante :

$$\begin{aligned}\sup_{t\in\mathscr{C}}|\hat{f}_n(t)-f(t)| &\leq \sup_{t\in\mathscr{C}}\left|\mathbb{E}\left[\tilde{f}_n(t)\right]-f(t)\right|+\sup_{t\in\mathscr{C}}\left|\hat{f}_n(t)-\tilde{f}_n(t)\right|+\sup_{t\in\mathscr{C}}\left|\tilde{f}_n(t)-\mathbb{E}\left[\tilde{f}_n(t)\right]\right| \\ &=: J_1+J_2+J_3.\end{aligned}$$

On remarquera de prime abord que J_1 ne dépend pas de la structure de dépendance. Par conséquent, et sous les hypothèses **D1** et **K**, il est aisé de vérifier que $J_1 = O(h^2)$. Pour ce faire, il suffit d'utiliser les techniques standard de calcul d'espérance conditionnelle, ainsi qu'un développement de Taylor à l'ordre 2. Pour J_2 on a

$$J_2 \leq \sup_{u\in\mathscr{C}}\frac{|G_n(u)-G(u)|}{\left(\bar{G}(\tau)\right)^2}\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{t-T_i}{h}\right)=O\left(\sqrt{\frac{\log\log n}{n}}\right) \text{ p.s., lorsque } n\longrightarrow\infty.$$

Le résultat est obtenu en utilisant la Loi du Logarithme Itéré pour des variables aléatoires censurées à droite (voir Deheuvels & Einmahl, 2000), qui donne $\sup_{u\in\mathscr{C}}|G_n(u)-G(u)| = O\left(\sqrt{\frac{\log\log n}{n}}\right)$ p.s., lorsque $n\longrightarrow\infty$. De plus, sous les hypothèses **M1**, **M2**, en utilisant un changement de variables et la définition du mode, pour $\varepsilon>0$ fixé, il vient

$$\begin{aligned}\mathbb{P}\left(\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{t-T_i}{h}\right)\geq\varepsilon\right) &\leq \frac{1}{\varepsilon h}\mathbb{E}\left[K\left(\frac{t-T_i}{h}\right)\right] \\ &\leq c\varepsilon^{-1}f(\theta)=O_P(1).\end{aligned}$$

Finalement, en appliquant le Théorème 4.1 pour J_3, la preuve du Théorème 4.2 est établie. □

Preuve du Corollaire 4.1 : On a

$$\begin{aligned}\left|f(\hat{\theta}_n)-f(\theta)\right| &\leq \left|\hat{f}_n(\hat{\theta}_n)-f(\hat{\theta}_n)\right|+\left|\hat{f}_n(\hat{\theta}_n)-f(\theta)\right| \\ &\leq \sup_{t\in\mathscr{C}}\left|\hat{f}_n(t)-f(t)\right|+\sup_{t\in\mathscr{C}}\left|\hat{f}_n(t)-f(t)\right| \\ &\leq 2\sup_{t\in\mathscr{C}}\left|\hat{f}_n(t)-f(t)\right|. \qquad (4.10)\end{aligned}$$

Un développement de Taylor à l'ordre 2 de $f(\cdot)$ au voisinage de θ, associé à la définition du mode donne

$$f(\hat{\theta}_n) = f(\theta) + \frac{(\hat{\theta}_n - \theta)^2}{2!} f^{(2)}(\theta_n^*),$$

où θ_n^* est compris entre $\hat{\theta}_n$ et θ. Aussi, de (4.10), **D1** et **D3** on aboutit à

$$\left|\hat{\theta}_n - \theta\right| \leq 2\sqrt{\frac{\sup\limits_{t\in\mathscr{C}} \left|\hat{f}_n(t) - f(t)\right|}{|f^{(2)}(\theta_n^*)|}}.$$

Ainsi, la preuve du Corollaire 4.1 est immédiate à partir du Théorème 4.2. □

Chapitre 5

Cas de données censurées-associées : normalité asymptotique

Normalité asymptotique de l'estimateur à noyau de la densité sous le modèle censuré-associé *

5.1 Introduction

L'estimation de densité est un problème statistique classique. Au début, les résultats établissant les propriétés des estimateurs proposés font de l'indépendance une hypothèse forte. Cependant, cette condition n'est pas toujours recevable : dans certaines situations, une structure de dépendance s'impose aux données observées. La forme de dépendance considérée dans cet article est l'association, un concept introduit par Esary *et al.* (1967) et développé depuis par plusieurs auteurs, parmi lesquels Newmann (1980,1984), Birkel (1988a,1988b,1989), Roussas (1991,1993,1995,1997,2000), Dewan & Prakasa Rao (1999,2005), Cai & Roussas (1997,1998,1999a,1999b), Doukhan & Louhichi (2001), Guessoum *et al.* (2012), Ferrani *et al.* (2014), et la liste est non exhaustive. Il est intéressant de se référer à l'excellent livre de Bulinski & Shashkin (2007), pour plus de détails concernant une multitude de résultats établis dans le cadre de l'association, ainsi que pour d'autres

∗. Ce chapitre a fait l'objet d'un article soumis pour publication.

types de dépendances. Rappelons que les variables aléatoires (v.a.) $\{T_i; 1 \leq i \leq n\}$ sont dites associées si, pour toute paire de fonctions ψ_1 et ψ_2 de $\mathbb{R}^n$ dans $\mathbb{R}$, non décroissantes (composante par composante), on a :

$$\mathrm{Cov}[\psi_1(T_i, 1 \leq i \leq n), \psi_2(T_j, 1 \leq j \leq n)] \geq 0,$$

lorsque cette covariance existe. Une famille infinie de v.a. est associée si toute sous-famille finie est associée. L'association se rencontre, entre autres domaines, dans certains problèmes d'analyse de survie (elle est appropriée pour modéliser certains essais cliniques), en fiabilité ou en mécanique statistique dans la théorie de la percolation (voir les inégalités FKG, initiales de leurs auteurs Fortuin, Kasteleyn & Ginibre 1971). Le modèle décrit ci-après, objet de notre intérêt, est également assujetti à une censure aléatoire droite.
Soit $\{T_i, i \geq 1\}$ une suite strictement stationnaire de variables aléatoires associées et de carré intégrable, de fonction de répartition F et de densité f, et $\{C_i, i \geq 1\}$ une suite de v.a. (positives) de censure (à droite) indépendantes et de même loi de fonction de répartition G inconnue. On suppose que les v.a. C_i sont indépendantes des T_i.
Dans le modèle de censure à droite, on observe n paires $(Y_1, \delta_1), \cdots, (Y_n, \delta_n)$, où

$$Y_i = (T_i \wedge C_i), \ et \ \delta_i = \mathbf{1}_{\{T_i \leq C_i\}} = \begin{cases} 1 & \text{si non censure} \\ 0 & \text{si censure} \end{cases}$$

Les v.a. Y_i sont de fonction de répartition H définie par :

$$H(y) = 1 - (1 - F(y))(1 - G(y)) = 1 - \overline{F}(y)\overline{G}(y), \quad y \in R.$$

On pose $\tau_F = \inf\{y \ tq \ F(y) = 1\} \leq \infty$. τ_F est la plus petite borne supérieure du support de F.
On définit un estimateur de f par :

$$\widetilde{f}_n(t) = \frac{1}{nh} \sum_{j=1}^{n} \frac{\delta_i}{\overline{G}(Y_j)} K\left(\frac{t - Y_j}{h}\right).$$

où K est un noyau positif d'intégrale 1, et $h = h(n)$ une suite de nombres réels positifs, telle que $h \longrightarrow 0$ quand $n \longrightarrow \infty$.
Lorsque G est connue, $\widetilde{f}_n$ estime la densité commune des durées de vie. Mais dans la

pratique, G est en général inconnu. L'estimateur de Kaplan-Meier qui lui est associé a pour expression :

$$G_n(y) = \begin{cases} 1 - \prod_{i=1}^{n} \left(1 - \frac{1-\delta_i}{n-i+1}\right)^{\mathbf{1}\left\{Y_{(i)} \leq y\right\}} & \text{si} \quad y < Y_{(n)} \\ 1 & \text{si} \quad y \geq Y_{(n)} \end{cases}$$

où $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$ représentent les statistiques d'ordre associées à Y_i.
Ainsi l'estimateur de f associé sera défini par :

$$\widehat{f}_n(t) = \frac{1}{nh} \sum_{j=1}^{n} \frac{\delta_j}{\overline{G}_n(Y_j)} K\left(\frac{t - Y_j}{h}\right) := \frac{1}{nh} \sum_{j=1}^{n} K_j(t),$$

où

$$K_j = K_j(t) := \frac{\delta_j}{\overline{G}_n(Y_j)} K\left(\frac{t - Y_j}{h}\right).$$

Ferrani *et al.* (2014) ont étudié la convergence uniforme presque sûre de $\widehat{f}_n$ avec vitesse. On se propose, pour compléter l'étude, d'établir la normalité asymptotique de cet estimateur. Il est à noter que cette propriété de normalité asymptotique de l'estimateur à noyau de la densité a été traitée dans d'autres situations parmi lesquelles : Parzen (1962) pour le cas de données complètes indépendantes, Eddy (1980) pour le même cas sous des hypothèses plus faibles, Mielniczuk (1986) dans le cas de l'estimateur défini sous forme intégrale de la densité et de ses dérivées, dans le cas de censure (à droite). Louani (1998) établit les mêmes résultats en affaiblissant les conditions. Roussas (2000) pour des données complètes associées. Benrabah *et al.* (2014) pour des données tronquées alpha-mélangeantes.
Cet chapitre est organisé comme suit : la Section 5.2 est consacrée aux notations et hypothèses, le résultat principal ainsi que son corollaire y figurent également, alors que les preuves et les différents résultats auxiliaires sont reportés à la Section 5.3.

5.2 Hypothèses et résultats

Dans toute la suite et pour des commodités de notation, on pose

$$\begin{aligned} \widetilde{K}_j(t) &= \widetilde{K}(t, Y_j) := K_j(t) - E\left(K_j(t)\right) \\ &= \frac{\delta_j}{\overline{G}(Y_j)} K\left(\frac{t - Y_j}{h}\right) - E\left[\frac{\delta_1}{\overline{G}(Y_1)} K\left(\frac{t - Y_1}{h}\right)\right], \end{aligned}$$

et

$$Z_j = Z_j(t) := \frac{1}{\sqrt{nh}}\widetilde{K}_j(t).$$

Alors

$$\frac{1}{\sqrt{nh}}\left(\widetilde{f}_n(t) - E\widetilde{f}_n(t)\right) = \frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\widetilde{K}_j(t) = \sum_{i=1}^{n} Z_j(t).$$

On veut montrer que

$$\frac{\widetilde{f}_n(t) - E\widetilde{f}_n(t)}{\sigma\left(\widetilde{f}_n(t)\right)} = \frac{\sqrt{nh}\left(\widetilde{f}_n(t) - E\widetilde{f}_n(t)\right)}{\sqrt{nh}\sigma\left(\widetilde{f}_n(t)\right)},$$

converge vers une variable aléatoire normale centrée réduite. Pour celà, on établit les conditions suivantes :

On considère un compact $[0, \tau]$ où $\tau < \tau_F$ ainsi que les hypothèses suivantes :

M1. $\{T_i, i \geq 1\}$ est une suite strictement stationnaire de v.a. associées, de fonction de répartition F admettant une densité f et des moments d'ordre deux finis,

M2. Les variables de censure $\{C_i, i \geq 1\}$ sont i.i.d., de fonction de répartition G, et indépendantes des $\{T_i, i \geq 1\}$,

M3. Le terme de covariance défini par $\rho(r) := \sup_{j:|l-j|\geq r} cov(T_j, T_l)$ pour tous $l \geq 1$ et $r > 0$, vérifie $\rho(r) \leq \gamma_0 e^{-\gamma r}$ pour toutes constantes positives γ_0 et γ,

K. K est une densité lipschitzienne à support compact, vérifiant $\int uK(u)du = 0$,

D1. La densité $f(.)$ est deux fois continûment différentiable sur $[0, \tau]$.

D2. La densité jointe $f_{1,j+1}(.,.)$ de (T_1, T_{j+1}) vérifie :

$$\sup_{j\geq 1}\sup_{u,v\in[0,\tau]} |f_{1,j+1}(u,v) - f(u)f(v)| < c,$$

H. Le paramètre de lissage h est tel que $\frac{nh^4}{\log\log n} \longrightarrow \infty$, et $nh^5 \longrightarrow 0$ lorsque $n \longrightarrow \infty$,

N. Il existe des suites de nombres entiers $(p_n)_n$, $(q_n)_n$ et $(k_n)_n$ définie par $k_n := \left[\frac{n}{p+q}\right]$,(où [.] désigne la partie entière), tendant vers l'infini avec n, de sorte que $k(p+q) \leq n$ et $\frac{k(p+q)}{n} \longrightarrow 1$, telles que :
$(i)\frac{kq}{n} \underset{n\longrightarrow\infty}{\longrightarrow} 0$,
$(ii)\ ph \underset{n\longrightarrow\infty}{\longrightarrow} 0$, et $\frac{p}{\sqrt{nh}} \underset{n\longrightarrow\infty}{\longrightarrow} 0$,
$(iii)\frac{1}{h^3}\sum_{j=q}^{\infty}|Cov(T_1, T_{j+1})| \longrightarrow 0$,
où p, q et k désignent respectivement p_n, q_n et k_n.

Remarque 5.1. *Les hypothèses H et N(i) impliquent que* $h\log\log n \longrightarrow 0$ *et* $\frac{kp}{n} \underset{n\longrightarrow\infty}{\longrightarrow} 1$, *respectivement.*

Théorème 5.1. *Sous les hypothèses M1, M2, K, D1, H, N, on a*

$$\sqrt{nh}\left(\widetilde{f}_n(t) - E\widetilde{f}_n(t)\right) \xrightarrow{d} N(0, \sigma^2(t)), \text{ lorsque } n \longrightarrow \infty,$$

où

$$\sigma^2(t) := \sigma^2 = \lim_{n\longrightarrow\infty} nhvar(\widetilde{f}_n(t)) = \frac{f(t)}{\overline{G}(t)}\int K^2(u)du.$$

Corollaire 5.1. *Sous les conditions du Théorème 5.1 et si de plus, la condition H est satisfaite, alors on a :*

$$\sqrt{nh}\left(\widehat{f}_n(t) - f(t)\right) \xrightarrow{d} N(0, \sigma^2(t)).$$

5.3 Preuves

Remarque 5.2. *Etant donné que*

$$\sqrt{nh}\left(\widehat{f}_n(t) - f(t)\right) = \sqrt{nh}(\widehat{f}_n(t) - \widetilde{f}_n(t)) + \sqrt{nh}\left(\widetilde{f}_n(t) - E\widetilde{f}_n(t)\right) + \sqrt{nh}\left(E\widetilde{f}_n(t) - f(t)\right),$$

les hypothèses supplémentaires dans le corollaire permettent d'assurer que

$$\sqrt{nh}\left(E\widetilde{f}_n(t) - f(t)\right) \longrightarrow 0,$$

de même que

$$\sqrt{nh}(\widehat{f}_n(t) - \widetilde{f}_n(t)).$$

Remarque 5.3. *Pour démontrer le théorème, on partage la somme* $\sum_{i=1}^{n} Z_j(t)$ *en grands blocs et petits blocs comme suit :*
Pour $m = 1, 2, \cdots, k$, *on partage l'ensemble* $\{1, 2, \cdots, n\}$ *en* k *grands* p*-blocs* I_m *et* k *petits* q*-blocs* J_m *où :*
$I_m = \{i \text{ tels que } i = (m-1)(p+q)+1, \cdots, (m-1)(p+q)+p\}$ *contient* p *éléments pour chaque* $m = 1, 2, \cdots, k$,
$J_m = \{j \text{ tels que } j = (m-1)(p+q)+p+1, \cdots, m(p+q)\}$ *contient* q *éléments pour chaque* $m = 1, 2, \cdots, k$. *Les points restants vont constituer l'ensemble* $\{l \ ; k(p+q)+1 \leq l \leq n\}$ *qui peut être vide.*
Pour $m = 1, 2, \cdots, k$, *on définit les variables aléatoires suivantes :*

$$V_m = \sum_{j=(m-1)(p+q)+1}^{(m-1)(p+q)+p} Z_j, \ W_m = \sum_{j=(m-1)(p+q)+p+1}^{m(p+q)} Z_j, \ S_{n3} = \sum_{l=k(p+q)+1}^{n} Z_l$$

Alors,

$$S_n = \sum_{i=1}^{n} Z_j(t) = \sum_{m=1}^{k} V_m + \sum_{m=1}^{k} W_m + S_{n3} := S_{n1} + S_{n2} + S_{n3}$$

et donc la convergence du Théorème 5.1 devient

$$S_n \xrightarrow{d} N(0, \sigma^2),$$

qui sera établie en montrant que $S_{n1} \xrightarrow{d} N(0, \sigma^2)$ *et que* $E(S_{n2}^2) + E(S_{n3}^2) \longrightarrow 0$.
Au préalable, on étudie la variance asymptotique de $\widetilde{f}_n(t)$.

Lemme 5.1. *Sous les conditions M1-M3, K et D2, nous avons*

$$\lim_{n \longrightarrow \infty} nhvar(\widetilde{f}_n(t)) = \frac{f(t)}{\overline{G}(t)} \int K^2(u)du.$$

Preuve. La preuve a été déjà établie dans la partie traitant la consistance. □

Lemme 5.2. *Si les conditions M1, K, D1-D2, N(i) et (ii) sont satisfaites, alors :*
(*a*) $\frac{1}{h} var(K_1) \longrightarrow \frac{f(t)}{\overline{G}(t)} \int K^2(u)du$,
(*b*) $|cov(K_i, K_j)| \leq ch^2$,
(*c*) $\frac{k}{nh} \sum_{1 \leq i < j \leq p} |cov(K_i, K_j)| \longrightarrow 0$.

Preuve. Les assertions (a) et (b) s'obtiennent par conditionnement et l'application du Théorème de Convergence Dominée (TCD). Pour plus de détails concernant le point (a), on se réfère aux résultats sur la consistance.
Pour (c), d'après (b), on a :

$$\begin{aligned}\frac{k}{nh}\sum_{1\leq i<j\leq p}|cov(K_i,K_j)| &\leq \frac{k}{nh}p^2\,|cov(K_i,K_j)|\\ &\leq \frac{k}{nh}p^2ch^2 = c\frac{pk}{n}ph\longrightarrow 0,\end{aligned}$$

par les hypotèses N(i) et (ii). □

Lemme 5.3. *Si les hypothèses M1, K, D1-D2 et N sont vérifiées, alors pour n assez grand on a :*
(a) $kvar(W_1)\longrightarrow 0$,
(b) $|cov(W_1,W_{l+1})|\leq\left[\frac{LipK}{\overline{G}(\tau)}\right]^2\frac{q}{nh^3}\sum_{r=l(p+q)-(q-1)}^{l(p+q)+(q-1)}|cov(T_1,T_{r+1})|$,
(c) $\sum\limits_{1\leq i<j\leq k}|cov(W_i,W_j)|\longrightarrow 0$.

Preuve. Concernant le point (a), grâce aux conditions M1, K et D2 on a :

$$\begin{aligned}var(W_1) &= var\left(\frac{1}{\sqrt{nh}}\sum_{i=1}^{q}\tilde{K}_i(t)\right)\\ &= \frac{q}{nh}var\,(K_1)+\frac{2}{nh}\sum_{1\leq i<j\leq q}|cov(K_1,K_j)|\\ &\leq \frac{q}{n}\frac{1}{h}var\,(K_1)+\frac{2q^2}{nh}ch^2,\end{aligned}$$

et si D1, N(i) et (ii) sont satisfaites, on obtient

$$kvar(W_1)\leq\frac{kq}{n}\frac{1}{h}var\,(K_1)+2c\frac{kq}{n}qh\longrightarrow 0.$$

Pour le point (b)

$$\begin{aligned}
|cov(W_1, W_{l+1})| &= \left| cov\left(\sum_{i=p+1}^{p+q} Z_i, \sum_{j=l(p+q)+p+1}^{(l+1)(p+q)} Z_j \right) \right| = \left| \sum_{i=p+1}^{p+q} \sum_{j=l(p+q)+p+1}^{(l+1)(p+q)} cov(Z_i, Z_j) \right| \\
&= \sum_{r=1}^{q} \sum_{j=l(p+q)+p+1}^{(l+1)(p+q)} |cov(Z_{p+r}, Z_j)| \\
&= \sum_{j=l(p+q)+p+1}^{(l+1)(p+q)} |cov(Z_{p+1}, Z_j)| + \cdots + \sum_{j=l(p+q)+p+1}^{(l+1)(p+q)} |cov(Z_{p+q}, Z_j)| \\
&= : C_1 + \cdots + C_q,
\end{aligned}$$

où

$$\begin{aligned}
C_1 &= \sum_{j=l(p+q)+p+1}^{(l+1)(p+q)} |cov(Z_{p+1}, Z_j)| \\
&= \left|cov(Z_{p+1}, Z_{l(p+q)+p+1})\right| + \left|cov(Z_{p+1}, Z_{l(p+q)+p+2})\right| + \cdots + \left|cov(Z_{p+1}, Z_{l(p+q)+p+q})\right| \\
&= \left|cov(Z_1, Z_{l(p+q)+1})\right| + \left|cov(Z_1, Z_{l(p+q)+2})\right| + \cdots + \left|cov(Z_1, Z_{l(p+q)+q})\right|
\end{aligned}$$

$$\begin{aligned}
C_2 &= \sum_{j=l(p+q)+p+1}^{(l+1)(p+q)} |cov(Z_{p+2}, Z_j)| \\
&= \left|cov(Z_{p+2}, Z_{l(p+q)+p+1})\right| + \left|cov(Z_{p+2}, Z_{l(p+q)+p+2})\right| + \cdots + \left|cov(Z_{p+2}, Z_{l(p+q)+p+q})\right| \\
&= \left|cov(Z_1, Z_{l(p+q)})\right| + \left|cov(Z_1, Z_{l(p+q)+1})\right| + \cdots + \left|cov(Z_1, Z_{l(p+q)+q-1})\right|
\end{aligned}$$

$$\begin{aligned}
C_3 &= \sum_{j=l(p+q)+p+1}^{(l+1)(p+q)} |cov(Z_{p+3}, Z_j)| \\
&= \left|cov(Z_{p+3}, Z_{l(p+q)+p+1})\right| + \left|cov(Z_{p+3}, Z_{l(p+q)+p+2})\right| + \cdots + \left|cov(Z_{p+3}, Z_{l(p+q)+p+q})\right| \\
&= \left|cov(Z_1, Z_{l(p+q)-1})\right| + \left|cov(Z_1, Z_{l(p+q)})\right| + \cdots + \left|cov(Z_1, Z_{l(p+q)+q-2})\right|.
\end{aligned}$$

$\cdots$

$$\begin{aligned}
C_q &= \sum_{j=l(p+q)+p+1}^{(l+1)(p+q)} |cov(Z_{p+q}, Z_j)| \\
&= \left|cov(Z_{p+q}, Z_{l(p+q)+p+1})\right| + \left|cov(Z_{p+q}, Z_{l(p+q)+p+2})\right| + \cdots + \left|cov(Z_{p+q}, Z_{l(p+q)+p+q})\right| \\
&= \left|cov(Z_1, Z_{l(p+q)-(q-2)})\right| + \left|cov(Z_1, Z_{l(p+q)-(q-1)})\right| + \cdots + \left|cov(Z_1, Z_{l(p+q)+1})\right|.
\end{aligned}$$

Donc

$$\begin{aligned}
|cov(W_1, W_{l+1})| &= q\left|cov(Z_1, Z_{l(p+q)+1})\right| + (q-1)\left|cov(Z_1, Z_{l(p+q)+2})\right| + \\
&+ (q-2)\left|cov(Z_1, Z_{l(p+q)+3})\right| + \cdots + \left|cov(Z_1, Z_{l(p+q)+q})\right| + \\
&+ (q-1)\left|cov(Z_1, Z_{l(p+q)})\right| + (q-2)\left|cov(Z_1, Z_{l(p+q)-1})\right| \\
&+ \cdots + \left|cov(Z_1, Z_{l(p+q)-q})\right|,
\end{aligned}$$

de sorte que

$$\begin{aligned}
|cov(W_1, W_{l+1})| &= q\left|cov(Z_1, Z_{l(p+q)+1})\right| + (q-1)\left[\left|cov(Z_1, Z_{l(p+q)})\right| + \left|cov(Z_1, Z_{l(p+q)+2})\right|\right] + \\
&\quad +(q-2)\left[\left|cov(Z_1, Z_{l(p+q)-1})\right| + \left|cov(Z_1, Z_{l(p+q)+3})\right|\right] + \cdots + \\
&\quad + \left[\left|cov(Z_1, Z_{l(p+q)-q})\right| + \left|cov(Z_1, Z_{l(p+q)+q})\right|\right] \\
&\leq q\left[\left|cov(Z_1, Z_{l(p+q)-q})\right| + \left|cov(Z_1, Z_{l(p+q)-(q-1)})\right| + \cdots + \left|cov(Z_1, Z_{l(p+q)})\right|\right. \\
&\quad \left. + \left|cov(Z_1, Z_{l(p+q)+1})\right| + \cdots + \left|cov(Z_1, Z_{l(p+q)+q})\right|\right],
\end{aligned}$$

soit

$$\begin{aligned}
|cov(W_1, W_{l+1})| &\leq q \sum_{r=l(p+q)-q}^{l(p+q)+q} |cov(Z_1, Z_r)| \overset{r=j+1}{=} q \sum_{j+1=l(p+q)-q}^{l(p+q)+q} |cov(Z_1, Z_{j+1})| \\
&= q \sum_{j=l(p+q)-(q+1)}^{l(p+q)+q-1} |cov(Z_1, Z_{j+1})| \\
&= \frac{q}{nh} \sum_{j=l(p+q)-(q+1)}^{l(p+q)+q-1} |cov(K_1, K_{r+1})| \\
&\leq \left[\frac{LipK}{\overline{G}(\tau)}\right]^2 \frac{q}{nh^3} \sum_{r=l(p+q)-(q-1)}^{l(p+q)+(q-1)} |cov(T_1, T_{r+1})|,
\end{aligned}$$

par association et en utilisant Cai & Roussas (1998).

Pour le point (c), on a :

$$\begin{aligned}\sum_{1\leq i<j\leq k}|cov(W_i,W_j)| &= \sum_{l=1}^{k-1}(k-1)\,|cov(W_1,W_{l+1})| \quad \text{par stationnarité}\\ &\leq k\sum_{l=1}^{k-1}|cov(W_1,W_{l+1})|\\ \text{d'après } (b) \qquad &\leq \left[\frac{LipK}{\overline{G}(\tau)}\right]^2\frac{kq}{nh^3}\sum_{r=l(p+q)-(q-1)}^{l(p+q)+(q-1)}|cov(T_1,T_{r+1})|\longrightarrow 0,\end{aligned}$$

d'après les hypothèses N(i) et (iii). □

Lemme 5.4. *Sous le hypothèses M1, K, D1-D2, N(i) et (iii), on a :*
(a) $kvar(V_1)\longrightarrow\sigma^2$,
(b) $|cov(V_1,V_{l+1})|\leq\left[\frac{LipK}{\overline{G}(\tau)}\right]^2\frac{p}{nh^3}\sum_{r=l(p+q)-p}^{l(p+q)+p}|cov(T_1,T_{r+1})|$,
(c) $\sum\limits_{1\leq i<j\leq k}|cov(V_i,V_j)|\longrightarrow 0$,
(d) $var(S_{n_1})\longrightarrow\sigma^2$.

Preuve. La preuve du point (a) se fait sous les conditions M1, K et D2, comme suit :

$$\begin{aligned}var(V_1) &= var\left(\frac{1}{\sqrt{nh}}\sum_{i=1}^{p}\widetilde{K}_i(t)\right)\\ &= \frac{p}{nh}var\,(K_1)+\frac{2}{nh}\sum_{1\leq i<j\leq p}|cov(K_1,K_j)|\\ &\leq \frac{p}{n}\frac{1}{h}var\,(K_1)+\frac{2p^2}{nh}ch^2.\end{aligned}$$

Ce qui implique que

$$kvar(V_1)=\frac{kp}{n}\frac{1}{h}var\,(K_1)+\frac{2k}{nh}\sum_{1\leq i<j\leq p}|cov(K_1,K_j)|.$$

Le résultat vient par l' hypothèses N(i) et du fait que

$$\frac{1}{h}var\left(\frac{\delta_1}{\overline{G}(Y_1)}K\left(\frac{t-Y_1}{h}\right)\right)\longrightarrow\sigma^2,$$

et

$$\left|cov\left(\frac{\delta_i}{\overline{G}(Y_i)}K\left(\frac{t-Y_i}{h}\right),\frac{\delta_j}{\overline{G}(Y_j)}K\left(\frac{t-Y_j}{h}\right)\right)\right| = o(h^2).$$

La preuve du point (b) s'obtient par le même raisonnement que pour le Lemme 5.3 (b). En effet, on a :

$$\begin{aligned}|cov(V_1,V_{l+1})| &\leq \frac{p}{nh}\sum_{r=l(p+q)-p}^{l(p+q)+p}\left|cov(K_1,K_{r+1})\right| \\ &\leq \left[\frac{LipK}{\overline{G}(\tau)}\right]^2\frac{p}{nh}\frac{1}{h^2}\sum_{r=l(p+q)-p}^{l(p+q)+p}|cov(T_1,T_{r+1})|.\end{aligned}$$

Concernant le point (c), la preuve est obtenue de manière analogue à celle du Lemme 5.3 (c),

$$\begin{aligned}\sum_{1\leq i<j\leq k}|cov(V_i,V_j)| &= \sum_{l=1}^{k-1}(k-1)\,|cov(V_1,V_{l+1})|, \quad \text{par stationnarité} \\ &\leq k\sum_{l=1}^{k-1}|cov(V_1,V_{l+1})| \leq k\left[\frac{LipK}{\overline{G}(\tau)}\right]^2\frac{p}{nh^3}\sum_{l=1}^{k-1}\sum_{r=l(p+q)-p}^{l(p+q)+p}|cov(T_1,T_{r+1})| \\ &\leq \left[\frac{LipK}{\overline{G}(\tau)}\right]^2\frac{kp}{nh^3}\sum_{r=q}^{\infty}|cov(T_1,T_{r+1})| \longrightarrow 0,\end{aligned}$$

d'après les hypothèses N(i) et (iii).
Enfin, pour le point (d), on a :

$$var(S_{n1}) = var\left(\sum_{m=1}^{k}V_m\right) = kvar(V_1) + 2\sum_{1\leq i<j\leq k}|cov(V_i,V_j)| \longrightarrow \sigma^2,$$

d'après (a) et (c) . □

Lemme 5.5. *D'après les hypothèses M1, K, D1-D2 et N, on a :*
(a) $var(S_{n2}) \longrightarrow 0$,
(b) $var(S_{n3}) \longrightarrow 0$.

Preuve. La preuve du point (a) se fait sous les conditions M1, D2, K, N, comme suit :

$$\begin{aligned} var(S_{n2}) &= var\left(\sum_{m=1}^{k} W_m\right) \\ &= kvar(W_1) + 2\sum_{1\leq i<j\leq k} |cov(W_i, W_j)| \longrightarrow 0, \end{aligned}$$

par le Lemme 5.3 (a) et (c).
Pour le point (b), sous la condition D1 on a :

$$\begin{aligned} var(S_{n3}) &= var\left(\sum_{l=k(p+q)+1}^{n} Z_l\right) = var\left(\frac{1}{\sqrt{nh}}\sum_{l=k(p+q)+1}^{n} \widetilde{K}_l(t)\right) \\ &= var\left(\frac{1}{\sqrt{nh}}\sum_{l=k(p+q)+1}^{n} K_l\right) \\ &\leq \frac{n-k(p+q)}{nh} var\left(K_1\right) + \frac{2}{nh}\sum_{k(p+q)+1\leq i<j\leq n} |cov(K_i, K_j)| \\ &= \frac{n-k(p+q)}{nh} var\left(K_1\right) \\ &\quad + \frac{2}{nh}\sum_{+1\leq i<j\leq n-k(p+q)} |cov(K_i, K_j)|, \quad \text{par stationnarité} \\ &\leq \frac{p}{n}\frac{1}{h} var\left(K_1\right) + \frac{2}{nh}\sum_{1\leq i<j\leq p} |cov(K_i, K_j)|, \end{aligned}$$

$n-k(p+q)$ étant inférieur à p.
Le résultat découle du Lemme 5.2 (a) et (c). □

Preuve du Théorème 5.1. Le Lemme précédent permet d'établir que $E(S_{n2}^2)+E(S_{n3}^2) \longrightarrow 0$.
Il reste à montrer que $S_{n_1} = \sum_{m=1}^{k} V_m \xrightarrow{d} N(0, \sigma^2)$.
On prodède en 2 étapes :
1. On montrera d'abord que la valeur absolue de la différence entre la fonction caractéristique de $S_{n_1} = \sum_{m=1}^{k} V_m$ et le produit des fonctions caractéristiques des $V_m, m = 1, \cdots, k$ tend vers 0,

2. On montrera ensuite que la loi du produit des fonctions caractéristiques des $V_m, m = 1, \cdots, k$ converge vers la loi normale $N(0, \sigma^2)$.

1. Soit

$$\begin{aligned}
I_k(t) &: = \left|\mathbb{E}e^{it\sum_{m=1}^{k} V_m} - \prod_{m=1}^{k}\mathbb{E}e^{itV_m}\right| = \left|\mathbb{E}e^{it\sum_{m=1}^{k} V_m} - \prod_{m=1}^{k-1}\mathbb{E}e^{itV_m}.\mathbb{E}e^{itV_k}\right| \\
&= \left|\left(\mathbb{E}e^{it\sum_{m=1}^{k} V_m} - \mathbb{E}e^{itV_k}.\mathbb{E}e^{it\sum_{m=1}^{k-1} V_m}\right) + \left(\mathbb{E}e^{itV_k}.\mathbb{E}e^{it\sum_{m=1}^{k-1} V_m} - \prod_{m=1}^{k-1}\mathbb{E}e^{itV_m}.\mathbb{E}e^{itV_k}\right)\right| \\
&\leq \left|\mathbb{E}e^{it\sum_{m=1}^{k} V_m} - \mathbb{E}e^{itV_k}.\mathbb{E}e^{it\sum_{m=1}^{k-1} V_m}\right| + \left|\mathbb{E}e^{itV_k}.\mathbb{E}e^{it\sum_{m=1}^{k-1} V_m} - \prod_{m=1}^{k-1}\mathbb{E}e^{itV_m}.\mathbb{E}e^{itV_k}\right| \\
&\leq \left|\mathbb{E}e^{itV_k}\right|\left|\mathbb{E}e^{it\sum_{m=1}^{k-1} V_m} - \prod_{m=1}^{k-1}\mathbb{E}e^{itV_m}\right| + \left|\mathbb{E}e^{it\sum_{m=1}^{k} V_m} - \mathbb{E}e^{itV_k}.\mathbb{E}e^{it\sum_{m=1}^{k-1} V_m}\right| \\
&= \left|\mathbb{E}e^{it\sum_{m=1}^{k-1} V_m} - \prod_{m=1}^{k-1}\mathbb{E}e^{itV_m}\right| + \left|\mathbb{E}e^{it\sum_{m=1}^{k} V_m} - \mathbb{E}e^{itV_k}.\mathbb{E}e^{it\sum_{m=1}^{k-1} V_m}\right| \\
&= I_{k-1}(t) + \left|cov\left(e^{it\sum_{m=1}^{k-1} V_m}, e^{itV_k}\right)\right|.
\end{aligned}$$

Par un raisonnement analogue, on a

$$I_{k-1}(t) \leq I_{k-2}(t) + \left|cov\left(e^{it\sum_{m=1}^{k-2} V_m}, e^{itV_{k-1}}\right)\right|,$$

et ainsi de suite, de sorte que

$$\begin{aligned}
I_k(t) &:= \left|\mathbb{E}e^{it\sum_{m=1}^{k} V_m} - \prod_{m=1}^{k}\mathbb{E}e^{itV_m}\right| \\
&\leq I_{k-1}(t) + I_{k-2}(t) + I_{k-3}(t) + \cdots + I_2(t),
\end{aligned}$$

autrement dit

$$\begin{aligned}
\left|\mathbb{E}e^{it\sum_{m=1}^{k} V_m} - \prod_{m=1}^{k}\mathbb{E}e^{itV_m}\right| &\leq \left|cov\left(e^{it\sum_{m=1}^{k-1} V_m}, e^{itV_k}\right)\right| + \left|cov\left(e^{it\sum_{m=1}^{k-2} V_m}, e^{itV_{k-1}}\right)\right| \\
&+ \cdots + \left|cov\left(e^{itV_2}, e^{itV_1}\right)\right|.
\end{aligned}$$

Pour chaque terme du second membre de cette dernière inégalité, on applique le Corollaire suivant :

Corollaire 5.2. *[Bulinski & Shashkin (2007), page 90)]*
Supposons que $X = (X_1, X_2, \cdots, X_n)$ *est un vecteur aléatoire associé (PA ou NA) tel que* $E\|X\|^2 < \infty$. *Alors,* $\forall t_1, t_2, \cdots, t_n \in R$, *on a :*

$$\left| \mathbb{E} e^{i \sum_{j=1}^{n} t_j X_j} - \prod_{j=1}^{n} \mathbb{E} e^{i t_j X_j} \right| \leq 4 \sum_{1 \leq j,k \leq p, j \neq k} |t_j t_k| \ |cov(X_j, X_k)| .$$

Ainsi on a :

$$\left| cov\left(e^{itV_2}, e^{itV_1}\right) \right| = \left| \mathbb{E} e^{it(V_1+V_2)} - \mathbb{E} e^{itV_1} \mathbb{E} e^{itV_2} \right| \leq 4t^2 \left| cov(V_1, V_2) \right| ,$$

d'après le Lemmme 5.4 (b), de même

$$\begin{aligned} \left| cov\left(e^{it(V_1+V_2)}, e^{itV_3}\right) \right| &= \left| \mathbb{E} e^{it(V_1+V_2+V_3)} - \mathbb{E} e^{it(V_1+V_2)} \mathbb{E} e^{itV_3} \right| \leq 4t^2 \left| cov(V_1+V_2, V_3) \right| \\ &= 4t^2 \left[|cov(V_1, V_3)| + |cov(V_2, V_3)| \right] = 4t^2 \left[|cov(V_1, V_3)| + |cov(V_1, V_2)| \right], \end{aligned}$$

par stationnarité.
Le troisième terme donne

$$\begin{aligned} \left| cov\left(e^{it(V_1+V_2+V_3)}, e^{itV_4}\right) \right| &\leq 4t^2 \left| cov(V_1+V_2+V_3, V_4) \right| \\ &= 4t^2 \left[|cov(V_1, V_4)| + |cov(V_2, V_4)| + |cov(V_3, V_4)| \right] \\ &= 4t^2 \left[|cov(V_1, V_2)| + |cov(V_1, V_3)| + |cov(V_1, V_4)| \right], \end{aligned}$$

et ainsi de suite jusqu'au dernier terme

$$\begin{aligned} \left| cov\left(e^{it \sum_{m=1}^{k-1} V_m}, e^{itV_k} \right) \right| &= \left| \mathbb{E} e^{it \sum_{m=1}^{k} V_m} - \mathbb{E} e^{it \sum_{m=1}^{k-1} V_m} . \mathbb{E} e^{itV_k} \right| \\ &\leq 4t^2 \sum_{i=1}^{k-1} |cov(V_i, V_k)| = 4t^2 \sum_{i=1}^{k-1} |cov(V_1, V_{k-i+1})| , \text{ par stationnarité} \\ &= 4t^2 \sum_{l=1}^{k-1} |cov(V_1, V_{l+1})| , \text{ en posant } k - i = l. \end{aligned}$$

Ainsi,

$$\begin{aligned}
\left|\mathbb{E}e^{it\sum_{m=1}^{k}V_m}-\prod_{m=1}^{k}\mathbb{E}e^{itV_m}\right| &\leq 4t^2\left[(k-1)\left|cov(V_1,V_2)\right|+(k-2)\left|cov(V_1,V_3)\right|\right.\\
&\quad \left.+\cdots+(k-j)\left|cov(V_1,V_{j+1})\right|+\cdots+\left|cov(V_1,V_k)\right|\right]\\
&\leq 4t^2k\sum_{m=2}^{k}\left|cov(V_1,V_m)\right|\\
&\leq 4t^2k\left[\frac{LipK}{\overline{G}(\tau)}\right]^2\frac{p}{nh^3}\sum_{m=2}^{k}\sum_{r=(m-1)(p+q)-p}^{(m-1)(p+q)+p}\left|cov(T_1,T_{r+1})\right|,\\
&\leq 4t^2\left[\frac{LipK}{\overline{G}(\tau)}\right]^2\frac{pk}{nh^3}\sum_{r=q}^{\infty}\left|cov(T_1,T_{r+1})\right|\longrightarrow 0,
\end{aligned}$$

d'après le Lemme 5.4 (b) et les hypothèses N(i) et (iii). En effet,

$$\begin{aligned}
\sum_{m=2}^{k}\sum_{r=(m-1)(p+q)-p}^{(m-1)(p+q)+p}\left|cov(T_1,T_{r+1})\right| &= \sum_{m=2}^{k}\sum_{r=(m-2)p+(m-1)q}^{(m-2)p+(m-1)q+2p}\left|cov(T_1,T_{r+1})\right|\\
&= \sum_{r=q}^{2p+q}\left|cov(T_1,T_{r+1})\right|+\sum_{r=p+2q}^{3p+2q}\left|cov(T_1,T_{r+1})\right|+\\
&\quad \sum_{r=2p+3q}^{4p+3q}\left|cov(T_1,T_{r+1})\right|+\cdots+\sum_{r=(k-1)p+kq}^{(k-1)p+kq+2p}\left|cov(T_1,T_{r+1})\right|\\
&\leq \sum_{r=q}^{\infty}\left|cov(T_1,T_{r+1})\right|.
\end{aligned}$$

2. On considère les v.a. $V_m, m=1,\cdots,k$, et soient $v_m, m=1,\cdots,k$, des v.a. indépendantes, de même loi que V_1, centrées. On peut réecrire la dernière relation :

$$\left|\mathbb{E}e^{it\sum_{m=1}^{k}V_m}-\prod_{m=1}^{k}\mathbb{E}e^{itv_m}\right|\longrightarrow 0,$$

autrement dit, les v.a. U_m étant indépendantes

$$\left|\mathbb{E}e^{it\sum_{m=1}^{k}V_m}-\mathbb{E}e^{it\sum_{m=1}^{k}v_m}\right|\longrightarrow 0.$$

Montrons maintenant que

$$\sum_{m=1}^{k}v_m\xrightarrow{d}N(0,\sigma^2).$$

Soient $s^2 := kvar(V_1)$ et $U_m := \frac{v_m}{s}$.
Alors les v.a. U_m, $m = 1, \cdots, k$, sont i.i.d., de fonction de répartition F_n, centrées ($E(U_1) = 0$) et de variance

$$var(U_1) = \frac{var(v_1)}{s^2} = \frac{var(V_1)}{kvar(V_1)} = \frac{1}{k},$$

il s'ensuit que

$$\sum_{m=1}^{k} var(U_m) = 1,$$

d'autre part, par le Lemme 5.3 (i),

$$s^2 \longrightarrow \sigma^2.$$

Ainsi

$$\sum_{m=1}^{k} v_m \xrightarrow{d} N(0, \sigma^2) \iff \sum_{m=1}^{k} U_m \xrightarrow{d} N(0, 1).$$

Par le critère de Feller-Lindeberg, il suffit de montrer que, pour tout $\varepsilon > 0$

$$g_n(\varepsilon) := k \int_{\{|t|>\varepsilon\}} t^2 dF_n \longrightarrow 0.$$

De la relation

$$v_m = \sum_{j=(m-1)(p+q)+1}^{(m-1)(p+q)+p} Z_j = \frac{1}{\sqrt{nh}} \sum_{j=(m-1)(p+q)+1}^{(m-1)(p+q)+p} \widetilde{K}_j(t),$$

on déduit que

$$|v_m| \leq \frac{2\|K\|_\infty}{\overline{G}(\tau)} \frac{p}{\sqrt{nh}},$$

de sorte que

$$|U_1| = \frac{|v_m|}{s} \leq \frac{2\|K\|_\infty}{\overline{G}(\tau)} \frac{p}{s\sqrt{nh}}, \quad \text{avec probabilité 1.}$$

Ainsi,

$$\begin{aligned} g_n(\varepsilon) &= k\mathbb{E}\left(U_1 \mathbb{I}_{\{|U_1|>\varepsilon\}}\right) \leq \left[\frac{2\|K\|_\infty}{\overline{G}(\tau)}\right]^2 \frac{kp^2}{nhs^2} \mathbb{P}\left(|U_1| > \varepsilon\right) \\ &\leq \left[\frac{2\|K\|_\infty}{\overline{G}(\tau)}\right]^2 \frac{kp^2}{nhs^2} \frac{var(U_1)}{\varepsilon^2}, \text{ d'après l'inégalité de Tchebychev,} \\ &= \left[\frac{2\|K\|_\infty}{s\varepsilon\overline{G}(\tau)}\right]^2 \frac{p^2}{nh} \longrightarrow 0, \qquad (kvar(U_1) = 1) \end{aligned}$$

du fait de la condition N(ii). □

Preuve du Corollaire 5.1.

$$\begin{aligned}\left(\widehat{f}_n(t) - f(t)\right) &= \left(E\widetilde{f}_n(t) - f(t)\right) + (\widehat{f}_n(t) - \widetilde{f}_n(t)) + \left(\widetilde{f}_n(t) - E\widetilde{f}_n(t)\right) \\ &: = J_1 + J_2 + J_3,\end{aligned}$$

où $\sqrt{nh}J_1$ et $\sqrt{nh}J_2$ sont négligeables, ce qui permet d'achever la preuve du Théorème 5.1.

En effet sous les hypothèses K et D1, en utilisant les techniques de calcul d'espérance par conditionnement, et un développement de Taylor à l'ordre 2, il vient que

$$\left|E\widetilde{f}_n(t) - f(t)\right| = \mathcal{O}(h^2) \quad p.s. \text{ losque } n \longrightarrow \infty,$$

et donc, sous la condition H,

$$\sqrt{nh}J_1 = \sqrt{nh}\left|E\widetilde{f}_n(t) - f(t)\right| = \mathcal{O}(\sqrt{nh^5}) = o(1).$$

de même, sous les conditions M1 et M2 et en utilisant, la Loi du Logarithme Itéré (LIL) pour des données censurées à droite (voir Deheuvels & Einmahl (2000)) et un changement de variable on obtient

$$\left|\widehat{f}_n(t) - \widetilde{f}_n(t)\right| = \mathcal{O}\left(\sqrt{\frac{\log\log n}{n}}\right) \quad p.s. \text{ losque } n \longrightarrow \infty.$$

puis, sous l'hypothèse H,

$$\begin{aligned}\sqrt{nh}J_2 &= \sqrt{nh}\left|\widehat{f}_n(t) - \widetilde{f}_n(t)\right| \\ &= \mathcal{O}\left(\sqrt{h\log\log n}\right) = o(1).\end{aligned}$$

La preuve est maintenant achevée. □

Conclusion

Les résultats proposés dans cette thèse sont spécifiques à des données censurées à droite, le cas de censure le plus fréquent. L'estimation à noyau de la densité et du mode ont été considérés dans les situations d'indépendance et de dépendance (faible). L'apport essentiel de notre étude concerne une forme de dépendance caractérisée par les covariances : l'association (notons que le cas indépendant, résolu au Chapitre 2, n'a pas fait l'objet d'étude, à notre connaissance).
Nous avons montré que la vitesse de convergence obtenue pour l'estimateur de la densité, dans le cas de l'association (tout comme dans le cas alpha-mélangeant) reste la même que pour le cas indépendant. Cela a été réalisé grâce à l'usage de l'inégalité de Doukan-Neumann dans la majoration du terme de fluctuations, moyennant la stricte stationnarité et une hypothèse de décroissance des covariances, des hypothèses classiques dans l'étude de variables associées (ou mélangeantes).
Il nous semble également intéressant de modifier l'étude avec un terme de covariance à décroissance polynomiale, et voir son influence sur le résultat.
En outre, la dernière partie du travail a permis de s'assurer que l'estimateur (de la densité) considéré est asymptotiquement normal.

Perspectives de recherche

Enfin, il serait judicieux d'explorer certaines pistes dans des recherches à venir :

• Cai & Roussas (1998) ont établi les propriétés asymptotiques de l'estimateur de Kaplan-Meier dans le cas associé. On pourrait étudier cet estimateur pour des données associées en présence de variables expliquatives.

• Guessoum et al. (2012) ont étudié la consistance forte uniforme avec vitesse de l'estimateur de Lynden-Bell (cas tronqué) sous l'hypothèse d'association. On pourrait prolonger le travail à l'estimateur de Lynden-Bell conditionnel.

• Un autre sujet d'investigation à prendre en compte serait la représentation de Bahadur, pour des variables incomplètes et associées.

• Il nous semble possible également d'étendre les résultats de cette thèse à l'étude du mode conditionnel.

• On peut aussi étudier le prolongement de nos résultats au cas de données fonctionnelles (lorsque la covariable est dans un espace de dimension infinie).

• Il serait intéressant d'établir un résultat de type Berry-Esseen pour la densité conditionnelle dans un modèle de données incomplètes et associées.

Bibliographie

Alam, K., Saxena, K. M. L. (1981). Positive dependence in multivariate distributions. *Comm. Statist.* A 10 : 1183–1196.

Azevedo, C. (2010). A note on convergence rates in the strong law of large numbers for associated sequences.

Proceedings of the 13th WSEAS. International Conference on Applied Mathematics (MATH'08) : 98–100.

Baek, J. I., Park, S. T., Chung, S. M., Seo, H. Y. (2005). On the almost sure convergence of weighted sums of negatively associated random variables. *Commun. korean Math. soc.* 20 : 539–546.

Bagai, I., Prakasa Rao, B. L. S. (1995). Kernel-type density and failure rate estimation for associated sequences. *Ann. Inst. Statist. Math.* 47 : 253–266.

Benrabah, O., Ould Saïd, E., Tatachak, A. (2014). A kernel mode estimate under random left troncation and time series model : Asymptotic normality.

Bertail, P., Doukhan, P., Soulier, P.(2006). Dependence in Probability and Statistics. *Lecture Notes in Statistics, Springer.*

Bickel, D.R. (2002). Robust estimators of the mode and skewness of continuous data. *Computational Statistics and Data Analysis* 39 : 153–163.

Bickel, D.R. (2003). Robust and efficient estimation of the mode of continuous data : The mode is a viable measure of central tendency. *Journal of Statistical Computation and Simulation* 73 : 899-912.

Birkel, T. (1988). Moment bounds for associated sequences. *Ann. Probab.* 16 (3) : 1184–1193.

Birkel, T. (1988). On the convergence rate in the central limit theorem for associated processes. *Ann. Probab.* 16 (4) : 1685–1698.

Birkel, T. (1989). A note on strong law of large numbers for positively dependent random variables. *Statist. Probab. Lett.* 7 : 17–20.

Block, H.W., Savits, T.H., Shaked, M. (1982). Some concepts of negative dependence. *Ann. Probab.* 10 (3) : 765–772.

Breslow, N. and Crowley, J. (1974). A large sample study of the life table and product-limit estimates under random censorship. *Ann. Statist.* 2 : 437–443.

Bulinski, A., Shashkin, A. (2007). Limit theorems for associated random fields and related systems. Vol. 10. *Advanced series on statistical science & applied probability.*

Bulinski, A. V. (1996). On the convergence rates in the CLT for positively and negatively dependent random fields. In :*Ibragimov, I. A.*

Bulinski, A., Sabanovitch, E. (1998). Asymptotical behaviour of some functionals of positively and negatively random fields. *Fundam. Appl. Math.* 4 : 479–492. (*en russe*)

Bulinski, A., Spodarev, E., Timmermann, F. (2012). Central limit theorem for the excursion set volumes of weakly dependent random fields. *Bernoulli* 18(1) : 100–118.

Bulinski, A., Suquet, C. (2001). Normal approximation for quasi-associated random fields. *Statist. Probab. Lett.* 54 : 215–226.

Burton, R.M., Dabrowski, A. R., Dehling, H. (1986). An invariance principle for weakly associated random vectors. *Stochastic Processes and their applications* 23 : 301–306.

Cai, G. H. (2011). A strong invariance principle for negatively associated random fields. *Czec. Math. J.* 61 (1) : 27–40.

Cai, Z., Roussas, G. G. (1997). Smooth estimate of quantiles under association. *Statist. Probab. Lett.* 36 : 275–287.

Cai, Z., Roussas, G. G. (1998). Kaplan-Meier estimator under association. *J. Multivariate Anal.* 67 : 318–348.

Cai, Z., Roussas, G. G. (1999a). Weak convergence for a smooth estimator of a distribution function under association. *Stochastic Anal. Appl.* 17 : 145–168.

Cai, Z., Roussas, G. G. (1999b). Berry-Esseen bounds for smooth estimator of a distribution function under association. *J. Nonparametric Statist.* 10 : 79–106.

Chernoff, H. (1964). Estimation of the mode. *Ann. Instit. Statist. Math.* 16 : 31–41.

Cox, J. T., Grimmett, G. (1984). Central limit theorem for associated random variables and the percolation model. *Ann. Probab.* 12 : 514–528.

Dabrowski, A. R. (1985). A funtional law of the iterated logarithm for associated sequences. *Statist. Probab. Lett.* 3 : 209–212.

Dabrowski, A. R., Dehling, H. (1988). A Berry-Esséen theorem and a functional law of the iterated logarithm for weakly associated random vectors. *Stochastic Processes and their applications* 30 : 277–289.

Deheuvels, P., Einmahl, J. (2000). Functional limit laws for the increments of Kaplan-Meier product limit processes and applications. *Ann. Probab.* 28 : 1301–1335.

Devroye, L.P., Wagner, T.J. (1980). The strong uniform consistency of kernel density estimates. *In Multivariate Analysis-V, North-Holland Publishing Company* : 59–77.

Dewan, I., Prakasa Rao, B.L.S. (2005). Non-uniform and uniform Berry-Esseen type bounds for stationary associated sequences. *J. Nonparametric Statist.* 17 : 217–235.

Dewan, I., Prakasa Rao, B. L. S. (1999). A general method of density estimation for associated random variables. *J. Nonparametric Statist.* 10 : 405–420.

Dewan, I., Prakasa Rao, B.L.S. (1997). Remarks on the strong law of large numbers for a triangular array of associated random variables. *Metrika* 45 : 225–234.

Doosti, H., Dewan, I. (2010). Wavelet linear density estimation for associated stratified size-biased sample.

Doukhan, P. (1994). Mixing : Properties and Examples. *Lecture Notes in Statistics, Springer-Verlag.*

Doukhan, P., Lang, G., Surgailis, G., Teissière, G. (2010). Dependence in probability and statistics. *Lecture notes in statistics. Springer.*

Doukhan, P., Louhichi, S. (1999). A new weak dependence condition and applications to moments inequalities. *Stochastic Processes and their applications* 84 : 313–342.

Doukhan, P., Louhichi, S. (2001). Functional estimation of a density under a new weak dependence condition. *Scandinavian Journal of Statist.* 28 : 325–341.

Doukhan, P., Neumann, M. (2007). Probability and moment inequalities for sums of weakly dependent random variables, with applications. *Stochastic Processes and their Applications* 117 : 878–903.

Droesbeke, J.J., Fichet, B., Tassi, P. (1989). Analyse statistique des durées de vie. *Economica.*

Eddy, W.F. (1980). Optimal kernel estimators of the mode. *Ann. Statist* 8 : 870–882.

Eddy, W.F. (1982). The asymptotic distributions of kernel density estimators of the mode. *Z. Wahrsch. verw. Gebiete* 59 : 279–290.

Esary, J., Proschan, F., Walkup, D. (1967). Association of random variables with applications. *Ann. Math. Stat.* 38 : 1466–1476.

Ferrani, Y., Ould Saïd, E., Tatachak, A. (2014) (*To appear*) On kernel density and mode estimates for associated and censored data. DOI♯ 10.1080/03610926.2013.867996. *Communication in Statistics-Theory and Methods.*

Ferrani, Y., Ould Saïd, E., Tatachak, A. (2014) Asymptotic Normality for a Kernel Density Estimate Under Censored and Associated Model. 9^{th} *International Statistic Days Symposium-ISDS'2014*, Antalya, Turkey.

Ferrani, Y., Ould Saïd, E., Tatachak, A. (2013) Asymptotic behavior of kernel density and mode estimates for censored and associated data. 15^{th} *Applied Stochastic Models and Data Analysis (ASMDA 2013)*, Mataró (Barcelona), Spain.

Fortuin, C., Kasteleyn, P., Ginibre, J. (1971). Correlation inequalities on some partially ordered sets. *Comm. Math. Phys.* 22 : 89–103.

Gill, R. D. (1980). Censoring and stochastic integrals. *Mathematical center tracts*, 124, *Mathematics centrum, Amesterdam.*

Grenander, U. (1965). Some direc estimates of the mode. *Ann. Math. Statist.* 36 : 131–138.

Grund, B., Hall, P. (1995). On the minimization of L^p error in mode estimation. *Ann. Statist.* 9 : 2264–2284.

Guessoum, Z., Ould Saïd, E., Sadki, O., Tatachak, A. (2012). A note on the Lynden-Bell estimator under association. *Statist. probab. Lett.* 82 : 1994–2000.

Györfi, L., Härdle, W., Sarda, P., Vieu, P. (1989). Nonparametric Estimation from Time Series. *Lecture Notes in Statistics* 60, Springer, Berlin.

Hall, P. (1982). Limit theorems for estimators based on inverses of spacings of order statistics. *Ann. Probab.* 10 : 992–1003.

Hedges, S.B., Shah, P. (2003). Comparison of mode estimation methods and application in molecular clock analysis. *BMC Bioinformatics* 4 : 1–11.

Henriques, C., Oliveira, P. E. (2004). Almost optimal convergence rates for kernel density estimation under association.*Pré-Publicaçoes do Departamento de Matemàtica, Universidade de Coimbra, Preprint Number* 04–06

Henriques, C., Oliveira, P. E. (2005). Exponential rates for kernel density estimation under association. *Statist. Neerlandica* 59 (4) : 448–466

Hermann, E., Ziegler, K. (2004). Rates of consistency for nonparametric estimation of the mode in absence of smoothness asumptions. *Statist. probab. Lett.* 68 : 359–368.

Ioannides, D. A., Roussas, G. G. (1999). Exponential inequality for associated random variables. *Statist. Probab. Lett.* 42 : 423–431.

Jabbari, H., Azarnoosh, H. A. (2006). Almost sure convergence rates for the estimation of a covariance operator for negatively associated samples. *JIRSS* 5 : 53–67.

Joag-Dev, K., Proschan, F. (1983). Negative association of random variables, with applications. *Ann. Statist.* 11 : 286–295.

Kaplan, E. L., Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* 53 : 457–481.

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

Khoshnevisan, D., Lewis, T. M. (1998). A law of the iterated logarithm for stable processes in random scenery. *Stochastic Processes and their applications* 74 : 89–121.

Ko, M. (2011). On the complete convergence for negatively associated random fields. *Taiwanese Journal of Math.* 15 : 171–179.

Koul, H., Susarla, V., Van Ryzin, J. (1981). Regression Analysis with randomly right-censored data. *Ann. Statist.* 9 : 1276–1288.

Koziol, J.A., Green, S.B. (1976). A Cramer-Von Mises statistic for randomly censored data. *Biometrika* 63 : 465–474.

Kuczmaszewska, A. (2009). On complete convergence for arrays of rowwise negatively associated random variables. *Statist. Probab. Lett.* 79 : 116–124.

Lehmann, E. L. (1966). Some concepts of dependence. *Ann. Math. Statist.* 37 : 1137–1153.

Li, Y. X., Wang, J. F. (2008). The law of the iterated logarithm for positively dependent random variables. *J. Math. Anal. Appl.* 33 : 259–265.

Lin, Z., Li, D. (2007). Asymptotic normality for L_1-norm kernel estimator of conditional median under association dependence. *J. Multivariante Anal.* 98 : 1214–1230.

Liu, A., Wang, X. (2009). A remark on the exponential inequality for negatively associated random variables. *J. Korean Statist. Society*, 38 : 53–57.

Louhichi, S., Soulier, P. (2002). The central limit theorem for stationary associated sequences. *Acta Math. Hungar.* 97 (1-2) : 15–36.

Masry, E. (1986). Recursive probability density estimation for weakly dependent process. *IEEE Trans. Inform. Theory* 32 : 254–267.

Masry, E. (2002). Multivariate probability density estimation for associated processes : strong consistency and rates. *Statist. Probab. Lett.* 58 : 205-219

Matula, P. (1992). A note on the almost sure convergence of sums of negatively dependent random variables. *Statist. Probab. Lett.* 15 : 209–213.

Mokkadem, A., Pelletier, M. (2003). The law of iterated logarithm for the multivariate kernel mode estimator. *ESAIM : Probab. Statist* 7 : 1–21.

Nadaraja, E.A. (1965). On nonparametric estimation of density function and regression. *Theory of Probability and its applications* 10 : 186–190.

Newman, C. M. (1980). Normal fluctuations and the FKG inequalities. *Comm. Math. Phys.* 74 : 119–128.

Newman, C. M. (1984). Asymptotic independence and limit theorems for positively and negatively dependent random variables, in :*Tong, Y. L. (Ed.) Inequalities in statistics and probability, IMS Lecture Notes-Monograph Ser., vol. 5 Inst. Math. Statist., Hayward, CA* : 127–140.

Newman, C. M. , Wright, A. L. (1981). An invariance principle for certain dependent sequences. *Ann. Probab.* 9 : 671–675.

Oliveira, P. E. (2005). An exponential inequality for associated variables. *Statist. Probab. Lett.* 73(2) : 189-197.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics* 33 : 1065–1076.

Peterson, A. V. (1977). Expressing the Kaplan-Meier estimator as a function of subsurvival function. *J. Amer. statist. Assoc.*, 72 : 854–858.

Pitt, L. D. (1982). Positively correlated normal variables are associated. *Ann. Probab.* 10 (3) : 496–499.

Rezapour, M., Alamatsaz, M.H., Balakrishnan, N., Cramer, E. (2013). On properties of progressively Type-II censored order statistics arising from dependent and non-identical random variables. *Statist. Method.* 10 : 58–71.

Rio, E. (2000). Théorie asymptotique des processus aléatoires faiblement dépendants, *Springer ISMAI.*

Romano, J. (1988). On weak convergence and optimality of kernel density estimates of the mode. *Ann. Statist.* 16 : 629–647.

Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proc. Natl. Acad. Sci. USA* 42 : 43–47.

Roussas, G. G. (1991). Kernel estimates under association : strong uniform consistency. *Statist. Probab. Lett.* 12 : 393–403.

Roussas, G. G. (1993). Curve estimation in random fields of associated processes. *J. Nonparametric Statist.* 2 : 215–224.

Roussas, G. G. (1994). Asymptotic normality of a smooth estimate of random fields of positively or negatively associated processes. *J. Multivariate Anal.* 50 : 152-173.

Roussas, G. G. (1995). Asymptotic normality of a smooth estimate of a random field distribution function under association. *Statist. Probab. Lett.* 24 : 77–90.

Roussas, G. G. (2000). Asymptotic normality of the kernel estimate of a probability density function under association. *Statist. Probab. Lett.* 50 : 1–12.

Shashkin, A. P. (2002). Quasi-associatedness of a gaussian system of random vectors. *Russ. Math. Surv.* 57 : 1243–1244.

Shi, X., Wu, Y., Miao, B. (2009). A note on the convergence rate of the kernel density estimator. *Statist. probab. Lett.* 79 : 1866–1871.

Shorack, G. R., Wellner, J.A . (1986). Empirical processes with applications to statistics. *Wiley, New York.*

Stute, W. (1982). A law of the logarithm for kernel density estimators. *Ann. Probab.* 10 : 414–422.

Stute, W., Wang, J. L. (1993). The strong law under random censorship. *Ann. Statist.* 21 : 1591–1607.

Su, C., Zhao, L., Wang, Y. (1996). Moment inequalities for negatively associated sequences and weak convergence. *Sci. China* (A) 26 : 1091–1099.

Su, C., Zhao, L., Wang, Y. (1997). Moment inequalities and weak convergence for negatively associated sequences. *Sci. China* (A) 40 : 172–182.

Tsybakov, A. B. (2003). Introduction à l'estimation non paramétrique. *Mathématiques et applications*, 41, Springer.

Van Ryzin, J. (1969). On strong consistency of density estimates. *Ann. Math. Statist.* 40 : 1765–1772.

Venter, J.H. (1967). On estimation of the mode. *Ann. Math. Statist.* 38 : 1446–1455.

Vieu, P. (1996). A note on density mode estimation. *Statist. probab. Lett.* 26 : 297–307.

Wegman, E.J. (1971). A note on the estimation of the mode. *Ann. Math. Statist.* 42 : 1909–1915.

Woodroofe, M. (1985). Estimating a distribution function with truncated data.*Ann. Statist.* 13 : 163–177.

Xing, G., Yang, S. (2010a). An exponential inequality for strictly stationary and negatively associated random variables. *Comm. Statist. Theory and Methods* 39 : 340-349.

Xing, G., Yang, S. (2010b). Some exponential inequalities for positively associated random variables and rates of convergence of the Strong Law of Large Numbers.*J. Theo. Probab.* 23 : 169–192.

Xing, G., Yang, S. (2011). On the strong convergence rate for positively associaterd random variables. *J. Math. Anal. Appl.* 373 : 422–431.

Zaitsev, A. Yu. (Ed.), Proceedings of the Kolmogorov semester in the Euler Math. Inst., March 1993, Probability Theory and Mathematical Statistics. Gordon and Breach, London : 3–14.

Zhang, L. (2001). A law of the iterated logarithm for negatively associated random fields, (Technical report).