
UNIVERSITÉ GASTON BERGER DE SAINT-LOUIS

ÉCOLE DOCTORALE DES SCIENCES ET DES TECHNOLOGIES

U.F.R. de Sciences Appliquées et de Technologie



THÈSE DE DOCTORAT

présentée par

PATHÉ NDAO

pour obtenir le grade de

Docteur de l'Université Gaston Berger

Spécialité: Probabilités-Statistique

**MODÉLISATION DE VALEURS EXTRÊMES
CONDITIONNELLES EN PRÉSENCE DE CENSURE**

soutenue publiquement le (date de soutenance), devant le jury ci-dessous :

Jury :

PRÉSIDENT :	M. Gane Samb LO	Professeur	- UGB (Saint-Louis, Sénégal)
RAPPORTEURS :	M. Abdelhakim NÉCIR	Professeur	- UB (Biskra, Algérie)
	M. Gilles DURRIEU	Professeur	- UBS (Vannes, France)
DIRECTEURS :	M. Papa NGOM	Maitre de Conférences	- UCAD (Dakar, Sénégal)
	M. Aliou DIOP	Professeur	- UGB (Saint-Louis, Sénégal)
EXAMINATEURS :	M. Jean-François DUPUY	Professeur	- INSA (Rennes, France)
	M. Ali Souleymane DABYE	Professeur	- UGB (Saint-Louis, Sénégal)
	M. Abdou Kâ DIONGUE	Maitre de Conférences	- UGB (Saint-Louis, Sénégal)

Dédicace

À ma famille.

Remerciements

Le temps est venu pour moi d'exprimer toute ma grande reconnaissance envers tous celles et ceux qui m'ont aidé et encouragé dans les moments difficiles dont le but était l'obtention du grade de Docteur de l'Université Gaston Berger de Saint-Louis.

Je tiens d'abord à remercier chaleureusement, je dis bien chaleureusement mes Directeurs de thèse, Les Professeurs Aliou Diop et Jean-François Dupuy sans eux ce travail n'aurait pas pu voir le jour qui ont accepté de superviser mes premiers pas dans la recherche. Être leur étudiant fut un honneur et plaisir, et je suis très reconnaissant de la disponibilité sans faille et de la confiance dont ils ont fait preuve en dépit leurs emplois du temps trop chargés. Mes Professeurs, j'ai beaucoup appris à vos côtés et vos leçons intéressantes qui m'ont permis d'avancer dans la thèse resteront toujours gravées dans ma mémoire. J'ai acquis pas mal de connaissances en leur côtoyant grâce à leur générosité sans limite tant pédagogiques que sociales.

Je souhaite également adresser mes sincères remerciements au Professeur Gane Samb Lo pour ses conseils et ses encouragements et pour m'avoir fait l'honneur de présider le jury de la thèse; qu'il trouve ici l'expression de ma plus grande profonde gratitude.

Je remercie également Les Professeurs Abdelhakim N'cir, Gilles Durrieu et Papa Ngom d'avoir accepté d'être rapporteurs malgré leurs emplois du temps chargés et le temps si précieux. Merci pour vos relectures attentives, vos remarques, vos conseils et pour l'intérêt que vous avez porté à mon travail. Qu'ils reçoivent mes vifs remerciements et l'expression de ma très haute reconnaissance.

Je tiens aussi à remercier les examinateurs, Les Professeurs Ali Souleymane Dabye et Abdou Kâ Diongue d'avoir accepté d'examiner la thèse et d'être parti du jury. Je suis profondément reconnaissant.

Toute ma reconnaissance aux programmes "Horizons Francophones" de l'Agence Universitaire de la Francophonie, "Aires-Sud" de l'Appuis Intégrés pour le Renforcement des Equipes Scientifiques du Sud et Centre de Mathématiques de l'Institut National des Sciences Appliquées de Rennes qui ont contribué financièrement à ma thèse.

Mention spéciale à tous les membres du laboratoire M.i.a. (Université de La Rochelle) là où j'ai effectué mon premier stage de recherche. Le moment est venu pour remercier chaleureusement à tous les membres et personnel du Centre de Mathématiques de l'Insa de Rennes pour leurs conseils, leurs aides. Je n'oublierai jamais le savoir être, le sens de la générosité sans limite qu'il y a au café au centre du Centre de Mathématiques.

Merci à tous mes amis ou camarades de l'école primaire à l'université qui m'ont aidé de près ou bien de loin à réaliser cette tâche. Je remercie le corps Professoral, le Personnel de l'U.f.r de Sciences Appliquées et de Technologie et particulièrement les membres du L.e.r.s.t.a.d. et L.a.n.i. trouvent ici mes remerciements.

Enfin je remercie tous les membres de ma famille plus particulièrement Mon père Mamadou Ndao, Ma mère Fatou Ndao et Mon épouse Sira Ndiaye qui m'ont soutenu dans tous les domaines. Et le dernier remerciement est pour mon fils Mouhamed Ndao.

Abréviations & Notations

Abréviation et notations

$v.a.$: variable aléatoire.
$p.s.$: Presque sûrement.
GEV	: « Generalized Extreme Value »
GPD	: « Generalized Pareto Distribution »
POT	: « Peaks-Over-Threshold »
$\mathbb{P}(A)$: La probabilité de l'événement A .
$\mathbb{E}(X)$: L'espérance mathématique de la $v.a.$ X .
$Var(X)$: La variance de la variable aléatoire X .
$Cov(X, Y)$: La covariance entre X et Y .
\mathbb{N}	: Ensemble des entiers naturels.
\mathbb{N}^*	: Ensemble des entiers naturels non nuls
\mathbb{R}	: Ensemble des réels et $\mathbb{R}^p = \underbrace{\mathbb{R} \times \dots \times \mathbb{R}}_{p \text{ fois}}, p \in \mathbb{N}^*$.
$X \stackrel{\mathcal{D}}{=} Y$: Les deux $v.a.$ X et Y ont la même loi de probabilité.
$X \stackrel{d}{=} Y$: Égalité en distribution entre X et Y .
$X \xrightarrow{\mathcal{D}} Y$: La $v.a.$ X converge en loi vers Y .
$X_n \xrightarrow{p.s.} Y$: La suite de $v.a.$ $(X_n)_{n \in \mathbb{N}}$ converge presque sûrement vers Y .
$o_{\mathbb{P}}(\cdot)$: Converge vers 0 en probabilité.
$O_{\mathbb{P}}(\cdot)$: Être borné en probabilité.
$\mathcal{N}(\mu, \sigma^2)$: Loi normale de moyenne μ et de variance σ^2 .
$X \wedge Y$: $\inf(X, Y)$.
$[a]$: Partie entière de a .
i.i.d.	: indépendantes et identiquement distribuées.
$\mathbb{1}_{\{A\}}$: fonction indicatrice de l'événement A .
\overline{F}	: $1 - F$, fonction de survie.

Table des matières

Dédicace	ii
Remerciements	iv
Abréviations & Notations	vi
Publications et Conférences	xvi
Introduction générale	xviii
1 Quelques rappels sur la théorie des valeurs extrêmes et sur la censure	2
1.1 Introduction	3
1.2 Rappels sur la théorie des valeurs extrêmes	3
1.2.1 Lois des valeurs extrêmes	6
1.2.2 Domaines d'attraction	8
1.2.2.1 Les fonctions à variations régulières	10
1.2.2.2 Domaine d'attraction de Fréchet	12
1.2.2.3 Domaine d'attraction de Weibull	20
1.2.2.4 Domaine d'attraction de Gumbel	20
1.2.2.5 Représentations graphiques	21
1.2.3 Quelques rappels sur la modélisation des valeurs extrêmes conditionnelles	22
1.2.3.1 Introduction	22
1.2.3.2 Estimation de l'indice et de quantiles extrêmes conditionnels	23
1.3 Quelques généralités sur la censure	24
1.3.1 Fonction de survie et de risque	24

1.3.2	La notion de censure	25
1.3.2.1	Définition	25
1.3.2.2	Caractéristiques	25
1.3.3	Estimation de la fonction de survie	27
1.3.3.1	Estimateur de Kaplan-Meier	28
1.3.3.2	Estimateur de Kaplan-Meier généralisé	28
1.3.4	Représentations graphiques en présence de données censurées	33
2	Estimation des quantiles extrêmes conditionnels en design fixe en présence de censure	35
2.1	Introduction	36
2.2	Modèle et notations	36
2.3	Définitions des estimateurs	37
2.4	Propriétés asymptotiques	41
2.4.1	Propriété asymptotique de l'estimateur de l'indice	43
2.4.2	Propriété asymptotique de l'estimateur des quantiles	45
2.5	Simulation et illustration sur des données réelles	47
2.5.1	Expériences numériques par simulation	47
2.5.2	Exemple sur des données réelles	60
2.6	Discussion et perspectives	66
2.6.1	Taux de censure sur les estimations	67
2.6.2	Effet de l'estimateur standard	67
2.6.3	Propriétés asymptotiques	67
2.6.4	Perspectives	68
2.7	Preuves des théorèmes	68
3	Estimation des quantiles extrêmes conditionnels en design aléatoire en présence de censure	80
3.1	Introduction	81
3.2	Construction de l'estimateur et propriétés asymptotiques	83
3.2.1	Définition de l'estimateur	83
3.2.2	Résultats asymptotiques	85
3.3	Estimation de quantiles extrêmes conditionnels	87
3.4	Expériences numériques	88
3.4.1	Génération des données	88
3.4.2	Résultats des simulations	89
3.4.3	Estimation des quantiles extrêmes conditionnels	90
3.4.4	Comparaison	91
3.5	Conclusion et Perspectives	91

3.6	Démonstrations	107
3.6.1	Résultats préliminaires	107
3.6.2	Démonstrations des principaux résultats	113
4	Loi de Paréto généralisée conditionnelle en présence de censure	120
4.1	Introduction	121
4.2	Modèle d'étude	121
4.3	Estimation des paramètres	122
4.3.1	Estimateurs de Beirlant et al. (2007) adaptés	123
4.3.2	Estimation par l'approche P.O.T.	124
4.3.3	Estimation du quantile extrême conditionnel	124
4.4	Résultats théoriques des estimateurs	125
4.5	Simulations	126
4.5.1	Données pour les simulations	127
4.5.2	Résultats	128
4.6	Discussion et perspectives	128
4.7	Preuve du Théorème	129
	Conclusions et perspectives	134
	A Quelques outils	136
	B Calculs de covariances	138
	Bibliographie	149

Table des figures

1.1	Explication graphique de quantiles extrêmes et queue de distribution. . .	5
1.2	Illustration graphique du théorème 1.2.1 sur une loi normale centrée réduite en comparant $\mathcal{H}_0(y)$ (rouge) et $\mathbb{P}\left[\frac{Y_{(n)}-b_n}{a_n} \leq y\right]$ avec $n = 10$ (bleu)	7
1.3	Représentation de la densité et fonction de répartition : Gumbel ($\gamma = 0$), Fréchet ($\gamma = 1$) et Weibull ($\gamma = -1$).	9
1.4	Comportement asymptotique de l'estimateur de Hill avec des intervalles de confiance au niveau de confiance 0.95 sur les données d'assurance incendie Danoise pour la période allant de 01/01/1980 au 31/12/1990.	16
1.5	Comportement asymptotique de l'estimateur de Pickands avec des intervalles de confiance au niveau de confiance 0.95 sur les données d'assurance incendie Danoise pour la période allant de 01/01/1980 au 31/12/1990.	17
1.6	Comportement asymptotique de l'estimateur des moments avec des intervalles de confiance au niveau de confiance 0.95 sur les données d'assurance incendie Danoise pour la période allant de 01/01/1980 au 31/12/1990.	18
1.7	Comportement asymptotique de l'estimateur UH avec des intervalles de confiance au niveau de confiance 0.95 sur les données d'assurance incendie Danoise pour la période allant de 01/01/1980 au 31/12/1990. . .	19
2.1	Les différentes étapes de sélection des données. En ordonnée on a la variable Z et en abscisse la covariable X	38
2.2	Comportement graphique de la fonction γ_1 sur $[0, 1]$	48

2.3	Comportement graphique de l'estimateur de l'indice en fonction du pourcentage de censure et du type d'estimateur ($n = 1500$): <i>hill</i> , <i>moments</i> et <i>uh</i>	49
2.4	Comportement graphique de l'estimateur des quantiles en fonction du pourcentage de censure et du type d'estimateur : ($n = 1500$) <i>hill</i> , <i>moments</i> et <i>uh</i>	50
2.5	Histogramme pour $R = 1000$ Hill (1 ^{re} ligne), Moments (2 ^e ligne) et UH (3 ^e ligne) estimation de l'indice des valeurs extrêmes au point $x = .5$ ($\gamma_1(.5) = .35$), pour $c = 10\%$ (colonne de gauche), $c = 25\%$ (milieu), $c = 40\%$ (droite). La taille de l'échantillon est 500.	51
2.6	Histogramme pour $R = 1000$ Hill (1 ^{re} ligne), Moments (2 ^e ligne) et UH (3 ^e ligne) estimation de l'indice des valeurs extrêmes au point $x = .5$ ($\gamma_1(.5) = .35$), pour $c = 10\%$ (colonne de gauche), $c = 25\%$ (milieu), $c = 40\%$ (droite). La taille de l'échantillon est 1500.	52
2.7	Histogramme pour $R = 1000$ de l'estimateur de $q(1/5000, .5) \approx 19.70786$, basés sur les estimateurs de Hill (1 ^{re} ligne), des Moments (2 ^e ligne) et UH (3 ^e ligne) estimateurs de l'indice de queue conditionnel, pour $c = 10\%$ (colonne de gauche), $c = 25\%$ (milieu), $c = 40\%$ (droite). La taille de l'échantillon est 500.	53
2.8	Histogramme pour $R = 1000$ de l'estimateur de $q(1/5000, .5) \approx 19.70786$, basés sur les estimateurs de Hill (1 ^{re} ligne), des Moments (2 ^e ligne) et UH (3 ^e ligne) estimateurs de l'indice de queue conditionnel, pour $c = 10\%$ (colonne de gauche), $c = 25\%$ (milieu), $c = 40\%$ (droite). La taille de l'échantillon est 1500	54
2.9	Évolution de la charge virale et du système immunitaire à l'absence de traitement (Source : Wikipédia).	60
2.10	Choix de (h_x^*, k_x^*) . Graphe de la fonction $k \mapsto d_{h,k}(x) $ (pour $h \in \mathcal{H}$) quand $x = 27$ (à gauche), $x = 37$ (à droite) et $x = 47$ (en bas) : notre cas.	62
2.11	Choix de (h_x^*, k_x^*) . Graphe de la fonction $k \mapsto d_{h,k}(x) $ (pour $h \in \mathcal{H}$) quand $x = 27$ (à gauche), $x = 37$ (à droite) et $x = 47$ (en bas) : Cas-Complet.	63
2.12	Choix de (h_x^*, k_x^*) . Graphe de la fonction $k \mapsto d_{h,k}(x) $ (pour $h \in \mathcal{H}$) quand $x = 27$ (à gauche), $x = 37$ (à droite) et $x = 47$ (en bas) : Cas-Ignoré.	64
2.13	Application sur les données du SIDA. Estimateur de Hill $\gamma_1(27)$ (estimateurs proposés dans (2.1), Cas-Complet et Cas-Ignoré) (à gauche) et $q(1/1000, 27)$ (estimateurs proposés dans (2.3), Cas-Complet et Cas-Ignoré) (à droite).	65

2.14 Application sur les données du SIDA. Estimateur de Hill $\gamma_1(37)$ (estimateurs proposés dans (2.1), Cas-Complet et Cas-Ignoré) (à gauche) et $q(1/1000, 37)$ (estimateurs proposés dans (2.3), Cas-Complet et Cas-Ignoré) (à droite).	65
2.15 Application sur les données du SIDA. Estimateur de Hill $\gamma_1(47)$ (estimateurs proposés dans (2.1), Cas-Complet et Cas-Ignoré) (à gauche) et $q(1/1000, 47)$ (estimateurs proposés dans (2.3), Cas-Complet et Cas-Ignoré) (à droite).	66
2.16 Estimation de la proportion des données non censurées pour les k plus grandes valeurs de Z qui sont proches de x ($x = 27, 37, 47$).	66
3.1 La fonction de $\gamma_1(\cdot)$ sur $[0, 1]$	89
3.2 Boxplots des R estimations de $\gamma_1(x)$ pour $n = 200$ (1 ^{ière} ligne), $n = 400$ (2 ^{ième} ligne), $n = 600$ (3 ^{ième} ligne), $n = 800$ (4 ^{ième} ligne). A gauche : $c = 10\%$, Au Centre : $c = 25\%$, A droite : $c = 40\%$	95
3.3 Boxplots des R estimations de $q(1/1000 x)$ pour $n = 200$ (1 ^{ière} ligne), $n = 400$ (2 ^{ième} ligne), $n = 600$ (3 ^{ième} ligne), $n = 800$ (4 ^{ième} ligne). A gauche : $c = 10\%$, Au Centre : $c = 25\%$, A droite : $c = 40\%$	98

Liste des tableaux

1.1	Quelques exemples de lois classées selon leurs domaines d'attraction. . .	9
2.1	Résultats des simulations pour l'estimateur (2.1) de $\gamma_1(x)$. Pour chaque scénario, la première ligne donne la moyenne pour $R = 1000$, estimation de $\gamma_1(x)$. (\cdot) : Racine carrée de l'erreur quadratique moyenne (RMSE). $[\cdot]$: Erreur absolue moyenne (MAE). $[\cdot, \cdot]$: intervalle de confiance au niveau de confiance 0.95 pour $\gamma_1(x)$ (* la borne inférieure de l'intervalle de confiance est négative et donc tronquée à 0). \dagger : probabilité de couverture.	55
2.2	Résultats des simulations pour l'estimateur (2.3) de $q(1/5000, .5)$. Pour chaque scénario, la première ligne donne la moyenne pour $R = 1000$, estimation de $q(1/5000, .5)$. (\cdot) : Racine carrée de l'erreur quadratique moyenne (RMSE). $[\cdot]$: Erreur absolue moyenne (MAE). $[\cdot, \cdot]$: intervalle de confiance au niveau de confiance 0.95 pour $q(1/5000, .5)$ (* la borne inférieure de l'intervalle de confiance est négative et donc tronquée à 0). \dagger : probabilité de couverture.	56
2.3	Estimation de $\gamma_1(x)$: le Cas-Complet et Cas-Ignoré. Pour chaque situation de simulation, la première ligne donne l'estimation de $\gamma_1(x)$. (\cdot) : Racine carrée de l'erreur quadratique moyenne (RMSE). $[\cdot]$: Erreur absolue moyenne (MAE).	58
2.4	Estimation de $q(1/5000, .5)$: le CC et CI. Pour chaque scénario de simulation, la première ligne donne l'estimation de $q(1/5000, .5)$. (\cdot) : Racine carrée de l'erreur quadratique moyenne (RMSE). $[\cdot]$: Erreur absolue moyenne (MAE).	59
2.5	Application sur les données du SIDA. Nous utilisons notre méthode pour estimer $\gamma_1(x)$ et $q(1/1000, x)$ avec les estimateurs proposés dans (2.1) et (2.3).	62

3.1	Résultats des simulations pour l'estimateur proposé dans (3.2) de $\gamma_1(x)$. Pour chaque scénario, la première ligne donne la moyenne de $\gamma_1(x)$ pour $R = 100$. (\cdot) : Racine carrée de l'erreur quadratique moyenne (RMSE). $[\cdot]$: Erreur absolue moyenne (MAE). $[\cdot, \cdot]$: intervalle de confiance au niveau de confiance 0.95 pour $\gamma_1(x)$. \dagger : Amplitude des intervalles.	93
3.2	Résultats de simulation pour $q(1/1000 x)$. Pour toute configuration avec les différents paramètres (n, c, x) , première ligne l'estimation de (3.6) de $q(1/1000 x)$ pour $R = 100$. (\cdot) : RMSE. $[\cdot]$: MAE.	96
3.3	Résultats des simulations pour l'estimateur proposé dans (3.2)(Cas- Complet-Noyau) de $\gamma_1(x)$. Pour chaque scénario, la première ligne donne la moyenne de $\gamma_1(x)$ pour $R = 100$. (\cdot) : Racine carrée de l'erreur quadratique moyenne (RMSE). $[\cdot]$: Erreur absolue moyenne (MAE).	99
3.4	Résultats de simulation pour $q(1/1000 x)$ (Cas-Complet-Noyau). Pour toute configuration avec les différents paramètres (n, c, x) , première ligne l'estimation de $q(1/1000 x)$ par l'estimateur proposé dans (3.6) pour $R = 100$. (\cdot) : RMSE. $[\cdot]$: MAE.	101
3.5	Résultats des simulations pour l'estimateur proposé dans (3.2)(Cas- Ignoré-Noyau) de $\gamma_1(x)$. Pour chaque scénario, la première ligne donne la moyenne de $\gamma_1(x)$ pour $R = 100$. (\cdot) : Racine carrée de l'erreur qua- dratique moyenne (RMSE). $[\cdot]$: Erreur absolue moyenne (MAE).	103
3.6	Résultats de simulation pour $q(1/1000 x)$ (Cas-Ignoré-Noyau). Pour toute configuration avec les différents paramètres (n, c, x) , première ligne donne l'estimation de $q(1/1000 x)$ par l'estimateur proposé dans (3.6) pour $R = 100$. (\cdot) : RMSE. $[\cdot]$: MAE.	105
4.1	Résultats des simulations pour les estimateurs de $\gamma_1(0.5)$ et du quantile $q(.5, 1/1000)$. Pour chaque scénario, la première ligne donne la moyenne avec $R = 1000$, estimations de $\gamma_1(0.5)$ et $q(.5, 1/1000)$. (\cdot) : La racine carrée de l'erreur quadratique moyenne (RMSE). $[\cdot]$: Erreur absolue moyenne (MAE).	128

Publications et Conférences

Articles

- ✍ Ndao, P., Diop, A., Dupuy, J.-F. Nonparametric estimation of the conditional tail and extreme quantiles under random censoring. *Computational Statistics & Data Analysis* 79, 63-79, 2014.
- ✍ Ndao, P., Diop, A., Dupuy, J.-F. Nonparametric estimation of the conditional extreme-value index with random covariates and censoring, accepté pour publication dans *Journal of Statistical Planning and Inference*. <http://dx.doi.org/10.1016/j.spi.2015.06.004>.
- ✍ Conditional Peaks-Over-Threshold modeling under random Censoring (en cours).

Conférences

- ✍ Ndao P., Diop A. Modélisation des valeurs extrêmes non stationnaires: Application sur le Bassin du fleuve Sénégal. Workshop « Modélisation & Environnement », Mai 2010, Université Cheikh Anta Diop, Dakar, Sénégal.
- ✍ Ndao P., Diop A. Modélisation des valeurs extrêmes non stationnaires: Application sur le Bassin du fleuve Sénégal. Conférence « Protections des ressources naturelles », Février 2011, Université Moulay Ismaël, Meknès, Maroc.
- ✍ Ndao P., Diop A., Dupuy J.-F. Estimation non-paramétrique des quantiles extrêmes conditionnels en présence de censure. Conférence « Journées Internationales Analyse Statistique: Théorie et Applications », Oujda, Maroc, 2012 (Communication sur sélection).
- ✍ Ndao P., Diop A., Dupuy J.-F. Estimation non-paramétrique des quantiles extrêmes conditionnels en présence de censure. Conférence « Statistique Appliquée pour le Développement en Afrique (S.A.D.A.) », Cotonou, Bénin, 2013 (Communication sur sélection).
- ✍ Ndao P., Diop A., Dupuy J.-F. Estimating conditional extreme quantiles under random censoring. Conférence « 29th International Workshop on Statistical Modelling », Göttingen, Allemagne, 2014 (Communication sur sélection).

Introduction générale

Une revue de la littérature

AU cours des dernières années, nous avons pu observer dans la recherche scientifique, une modélisation des événements rares. Ces événements rares sont des événements dont la probabilité d'apparition est trop faible c'est-à-dire se trouve dans les queues des distributions. Ils apparaissent en général dans les contextes physiques nombreux et variés en particulier les catastrophes naturelles: en hydrologie (crues décennales ou centennales et hauteur des barrages et digues susceptibles de les contenir (cf [3]), tempêtes occasionnant d'importants dommages matériels et environnementaux (cf [14]), dans les grands incendies (cf [4]), dans les tremblements de terre (cf [70], [2]), dans les risques financiers (les kraks boursiers, les crises financières, [18]), dans les records sportifs (cf [43], [45]), mais aussi dans l'étude de la résistance d'un matériau fibreux (cf [29]), etc. La théorie des valeurs extrêmes est une branche de la statistique qui essaie d'amener une solution face à ces phénomènes. Elle se repose principalement sur des distributions limites des extrêmes et leurs domaines d'attraction. Cependant, on y retrouve deux modèles: loi généralisée des extrêmes (GEV: « Generalized Extreme Value ») et loi de Paréto généralisée (GPD: « Generalized Pareto Distribution »). Ainsi, tout a commencé avec les auteurs Fisher et Tippett (1928, [49]) quand ils étudiaient la résistance des fils de coton puis plus tard Gnedenko (1943, [57]) s'est intéressé à ces distributions. Ils ont énoncé un théorème fondamental avec la création de trois domaines d'attraction: domaine d'attraction de Fréchet, Gumbel et Weibull. Ce théorème intéressant fait référence à un paramètre appelé l'indice de queue qui donne la forme de la queue de distribution. En effet, si l'indice de queue est positif nous sommes en présence du domaine d'attraction de Fréchet; puis si c'est négatif, domaine d'attraction de Weibull par contre si l'indice est nul alors domaine de Gumbel. Von Mises (1954, [94])

puis Jenkinson (1955, [68]) ont rassemblé les distributions de ces trois domaines en une seule écriture. C'est en ce moment que plusieurs auteurs se sont focalisés aux estimations de l'indice des valeurs extrêmes. Nous pouvons citer Hill (cf 1975, [67]) dans le cas où l'indice est positif. Puis Pickands (cf [78]) dans la même année a proposé un estimateur de l'indice des valeurs extrêmes dans le cas général. Par contre Dekkers et al. ont généralisé l'estimateur de Hill, dénommé l'estimateur des Moments (1989, [40]). Beirlant et al. (1996, [13]) ont présenté à leur tour, l'estimateur de l'indice des valeurs extrêmes généré à partir de l'estimateur de Hill et de la fonction quantile. Puis récemment Beirlant et al. dans leur livre ([8], 2004, chapitre 7: « Regression analysis ») ont entamé l'estimation de l'indice de queue en présence de covariables en utilisant la méthode du maximum de vraisemblance avec les deux modèles: G.E.V. et G.P.D. Ils ont utilisé cet estimateur de l'indice des valeurs extrêmes pour calculer les quantiles extrêmes conditionnels. Enfin, Gardes et Girard (2008, [51]), ont commencé à leur tour, la statistique non-paramétrique des valeurs extrêmes en présence de covariables à design fixe ou aléatoire. La stratégie utilisée par ces derniers découlent de celle de Ferrati et Vieu (cf [47]) c'est-à-dire la méthode de la fenêtre mobile sur la covariable de la variable d'intérêt. Parmi les résultats de Gardes et al., nous pouvons citer en particulier la thèse de Lékina (cf [72]) résultante de Gardes et al.(2010, [54]), Daouia et al.(2011, [30]).

L'étude des données de survie est née au XVII^e siècle, dans le domaine de la démographie (biologie, épidémiologie, biostatistique). L'objectif des analystes de cette époque était l'estimation, à partir des registres de décès, de diverses caractéristiques de la population, son effectif, sa longévité, etc. Ces analyses, très générales, ne sont réalisées qu'à partir du XIX^e siècle, avec l'apparition de catégorisations suivant des variables exogènes (sexe, nationalité, catégories socioprofessionnelles, etc.), détermination de la probabilité de mourir à un âge donné. C'est par la suite que l'étude des données de survie commence à déborder le cadre strict de la démographie au XX^e siècle au profit de toutes les disciplines susceptibles d'avoir recours à de tels types de données: finance (défaillance de crédit, intervalles entre cotations), la physique (avec l'apparition de la théorie de la fiabilité), l'industrie (pharmaceutique, biomédicale), sciences sociales (économie, sociologie, science politique), etc. C'est en ce moment que plusieurs auteurs se sont focalisés sur l'analyse des données de survie. Ainsi, en 1951, Weibull présente un modèle paramétrique dans le domaine de la fiabilité en proposant une loi dénommée « loi de Weibull » très utilisée par les spécialistes de données de survie. En 1958, Kaplan et Meier (cf [69]) proposent un estimateur jusque là inconnue de la fonction de répartition dans le cas où les données sont censurées. Ils en profitent pour déterminer ses propriétés asymptotiques. Par contre en 1972, Cox

y développe des travaux importants avec l'approche semi-paramétrique en introduisant des variables exogènes. Beran (1981, [15]) a quant à lui donné un estimateur de la fonction de répartition conditionnelle dans le cas où les données sont censurées appelé « l'estimateur de Kaplan-Meier généralisé ou conditionnel ».

La modélisation des valeurs extrêmes censurées voit le jour en première fois en 1997 dans la littérature des extrêmes avec la sortie du livre Reiss et Thomas (cf [80]). Il a fallu qu'en 2007 Beirlant et al. (cf [9]) abordent réellement la statistique non paramétrique des valeurs extrêmes avec des données censurées. Leur estimateur est basé sur un estimateur standard de l'indice de queue divisé par l'estimateur de la proportion de données non censurées dépassant un certain seuil donné. Ils ont appliqué cette théorie sur des données du SIDA. Puis, Einmahl et al. (2008, [44]) ont utilisé le même concept pour proposer un estimateur de l'indice de queue sur les k -plus grandes valeurs ensuite déterminer ses propriétés asymptotiques et enfin illustrer son comportement sur ces mêmes données du SIDA. Puis, la recherche sur la théorie des valeurs extrêmes censurées est devenue une actualité.

Problématique et contribution

CES événements rares que nous avons énumérés au début sont souvent entachés d'observations censurées aléatoirement à droite avec une covariable à design fixe ou aléatoire. Soit Y de la variable d'intérêt censurée aléatoirement à droite par une variable aléatoire C en présence d'une covariable X fixe ou aléatoire. Nous essayons de proposer pour chaque $X = x$, un estimateur du quantile extrême $q(\alpha_n|x)$ vérifiant

$$\mathbb{P}(Y > q(\alpha_n|x)|X = x) = \alpha_n \quad \text{si} \quad \alpha_n \rightarrow 0 \quad \text{quand} \quad n \rightarrow \infty$$

et aussi calculer la probabilité α_n d'avoir le quantile extrême $q(\alpha_n|x)$ définie par

$$\alpha_n = \mathbb{P}(Y > q(\alpha_n|x)|X = x) \quad \text{quand} \quad q(\alpha_n|x) \rightarrow \infty \quad \text{lorsque} \quad n \rightarrow \infty.$$

Exemple 1 (*Etude de crue : hydrologie*)

La hauteur d'un fleuve peut être modélisée par une variable Y censurée aléatoirement à droite par la variable aléatoire C mesurée conjointement sur des positions géographiques (altitude, latitude, longitude) qui représente la covariable X . Soit $(Y_i)_{1 \leq i \leq n}$ un échantillon de taille n , de hauteurs d'eau annuelles avec comme statistique d'ordre

$$Y_{(1)} \leq \dots \leq Y_{(n)}.$$

▮ Nous pouvons envisager de calculer pour tout $X = x$, le niveau d'eau h_x qui est dépassé une seule fois sur T années où $T > n$ vérifiant:

$$\mathbb{P}(Y > h_x | X = x) = \frac{1}{T}$$

appelé le niveau de retour.

▮ Calculer pour tout $X = x$, la probabilité p_x d'une hauteur d'eau $h_x > Y_{(n)}$ extrême définie

$$p_x = \mathbb{P}(Y > h_x | X = x)$$

appelé période de retour.

Le travail dans cette thèse consiste à résoudre ce genre de problème. Notre tâche, revient donc à modéliser des données censurées aléatoirement à droite avec une covariable c'est-à-dire raccommoder les travaux de Beirlant et al., Einmahl et al. et Gardes et al. pour proposer des estimateurs de l'indice des valeurs extrêmes conditionnel et de son quantile extrême conditionnel correspondant. Cet estimateur du quantile sera basé sur l'estimateur de Weissman (cf [95]) adapté avec l'estimateur de Kaplan-Meier simple ou conditionnel selon la nature de la covariable. Pour cela, nous allons scinder le problème en sous-problèmes selon la nature de la covariable. Ainsi, pour mieux comprendre cette thèse, nous présentons dans le chapitre 1 quelques rappels sur la statistique des valeurs extrêmes et aussi sur la notion de censure. Dans le chapitre 2, nous élaborons quelques estimateurs de l'indice de queue conditionnel et aussi des quantiles extrêmes conditionnels par la méthode non-paramétrique en présence d'une covariable à design fixe. Et dans le chapitre 3, nous proposons un estimateur de l'indice des valeurs extrêmes conditionnel et son quantile extrême conditionnel par la méthode non-paramétrique avec les noyaux où la covariable est de nature aléatoire. Par contre dans le chapitre 4 nous utilisons le modèle P.O.T. appliqué à la méthode du maximum de vraisemblance adaptée à la censure avec les résultats de Balkema et de Haan (1974, [5]) pour estimer l'indice de queue et le quantile extrême correspondant.

Quelques rappels sur la théorie des valeurs extrêmes et sur la censure

Résumé

Dans ce chapitre, nous énonçons quelques rappels essentiels sur la théorie des valeurs extrêmes dans le cas unidimensionnel et sur la notion de censure.

Sommaire

1.1 Introduction	3
1.2 Rappels sur la théorie des valeurs extrêmes	3
1.2.1 Lois des valeurs extrêmes	6
1.2.2 Domaines d'attraction	8
1.2.2.1 Les fonctions à variations régulières	10
1.2.2.2 Domaine d'attraction de Fréchet	12
1.2.2.3 Domaine d'attraction de Weibull	20
1.2.2.4 Domaine d'attraction de Gumbel	20
1.2.2.5 Représentations graphiques	21
1.2.3 Quelques rappels sur la modélisation des valeurs extrêmes conditionnelles	22
1.2.3.1 Introduction	22
1.2.3.2 Estimation de l'indice et de quantiles extrêmes conditionnels	23
1.3 Quelques généralités sur la censure	24
1.3.1 Fonction de survie et de risque	24

1.3.2	La notion de censure	25
1.3.2.1	Définition	25
1.3.2.2	Caractéristiques	25
1.3.3	Estimation de la fonction de survie	27
1.3.3.1	Estimateur de Kaplan-Meier	28
1.3.3.2	Estimateur de Kaplan-Meier généralisé	28
1.3.4	Représentations graphiques en présence de données censurées	33

1.1 Introduction

DANS cette partie, nous rappelons quelques notions essentielles sur la théorie des valeurs extrêmes et sur la notion de censure qui permettent de faciliter la lecture de la thèse. Ainsi, il s'agira de présenter brièvement les résultats essentiels rencontrés dans la littérature. Nous définissons rapidement les notions de domaine d'attraction, fonctions à variations régulières puis nous les caractérisons dans le cas unidimensionnel. Quant à la censure nous présenterons quelques définitions liées à la statistique des durées de survie. La censure est fondée avec quelques fonctions telles que la **fonction de répartition**, la **fonction de survie**, la **fonction de risque**. Beaucoup d'auteurs se sont intéressés sur la notion notamment Kaplan et Meier (1958, [69]), qui ont proposé un estimateur de la fonction de survie dont Beran a généralisé (cf, 1981, [15]) dans le cas conditionnel appelé l'estimateur de Kaplan-Meier généralisé ou conditionnel.

1.2 Rappels sur la théorie des valeurs extrêmes

La théorie des valeurs extrêmes communément appelée « **Extreme Value Theory** » (**EVT**) en anglais, est une vaste théorie dont le but est d'étudier les événements rares c'est-à-dire les événements dont la probabilité d'apparition est faible. Autrement dit elle essaie d'amener des éléments de réponses aux intempéries, aux inondations, aux catastrophes naturelles, aux problèmes financiers, etc. en prédisant leurs occurrences dans les années à venir. En d'autres termes on veut estimer des petites probabilités ou des quantités dont la probabilité d'observation est très faible c'est-à-dire proche de zéro. Ces quantités sont appelées quantiles extrêmes car l'ordre de ces quantiles tend vers zéro lorsque la taille de l'échantillon, n , tend vers l'infini.

On considère, pour cela, n variables aléatoires réelles $(Y_i)_{1 \leq i \leq n}$ indépendantes et identiquement distribuées de fonction de répartition F non nécessairement continue. Et soit $Y_{(1)} \leq Y_{(2)} \leq \dots \leq Y_{(n)}$ l'échantillon ordonné associé. Nous voulons déterminer, à partir de l'échantillon, le quantile extrême d'ordre $\alpha_n \rightarrow 0$ quand $n \rightarrow \infty$ défini par :

$$q(\alpha_n) = \bar{F}^{\leftarrow}(\alpha_n) = \inf\{y : \bar{F}(y) \leq \alpha_n\}, \quad (1.1)$$

avec la convention $\inf\{\emptyset\} = \infty$, où \bar{F}^{\leftarrow} est l'inverse généralisée de $\bar{F} = 1 - F$. Si la taille de l'échantillon tend vers l'infini, nous avons

$$\begin{aligned} \mathbb{P}\left(Y_{(n)} < q(\alpha_n)\right) &= \mathbb{P}\left(Y_i < q(\alpha_n), \forall i = 1, \dots, n\right) \\ &= \mathbb{P}\left(\bigcap_{i=1}^n \{Y_i < q(\alpha_n)\}\right) \\ &= \prod_{i=1}^n \mathbb{P}(Y_i < q(\alpha_n)) \\ &= F^n(q(\alpha_n)) \\ &= (1 - \alpha_n)^n \\ &= \exp(n \log(1 - \alpha_n)) \\ &= \exp(-n\alpha_n(1 + o(1))) \text{ quand } \alpha_n \rightarrow 0. \end{aligned}$$

Cette probabilité dépend donc du comportement asymptotique de $n\alpha_n$. Estimer les quantiles extrêmes revient à étudier la limite de $n\alpha_n$ lorsque n tend vers l'infini. Ainsi, on distingue trois situations selon la vitesse de convergence de α_n vers 0 :

Première situation : Si $n\alpha_n \rightarrow \infty$ alors $\mathbb{P}\left(Y_{(n)} < q(\alpha_n)\right) \rightarrow 0$.

Dans cette situation, le quantile à estimer se trouve avec une grande probabilité dans l'échantillon disponible. Elle correspond au cas où α_n converge lentement vers 0 autrement dit le quantile à estimer converge lentement vers l'infini lorsque n tend vers l'infini. Dans ce cas estimer le quantile revient à interpoler à l'intérieur de l'échantillon et donc on obtient $Y_{(n - \lfloor n\alpha_n \rfloor + 1)}$ comme estimateur.

Deuxième situation : Si $n\alpha_n \rightarrow c \in [1, \infty[$ alors $\mathbb{P}\left(Y_{(n)} < q(\alpha_n)\right) \rightarrow e^{-c}$.

Dans ce deuxième cas, nous avons une faible probabilité que le maximum de l'échantillon soit supérieur au quantile. Cela signifie que l'estimation du quantile extrême se repose sur les grandes observations situées au voisinage de la frontière de l'échantillon et toujours dans l'ensemble des données. De ce fait, l'estimateur proposé dans la première situation est celui envisagé.

Troisième situation : Si $n\alpha_n \rightarrow c \in [0, 1[$ alors $\mathbb{P}(Y_{(n)} < q(\alpha_n)) \rightarrow e^{-c}$.

Dans ce cas, le quantile à estimer est supérieur au maximum des observations disponibles. Proposer un estimateur du quantile est impossible avec la **fonction de répartition empirique** définie :

$$F_n(y) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{Y_i \leq y\}} = \begin{cases} 0 & \text{si } y < Y_{(1)}, \\ \frac{i-1}{n} & \text{si } Y_{(i-1)} \leq y < Y_{(i)}, 2 \leq i \leq n \\ 1 & \text{si } y \geq Y_{(n)}. \end{cases}$$

Cependant pour estimer le quantile extrême il faut extrapoler au delà du maximum des observations dont on dispose car $F_n(y) = 1$ si $y \geq Y_{(n)}$. Ce phénomène peut être expliqué en utilisant le schéma suivant :

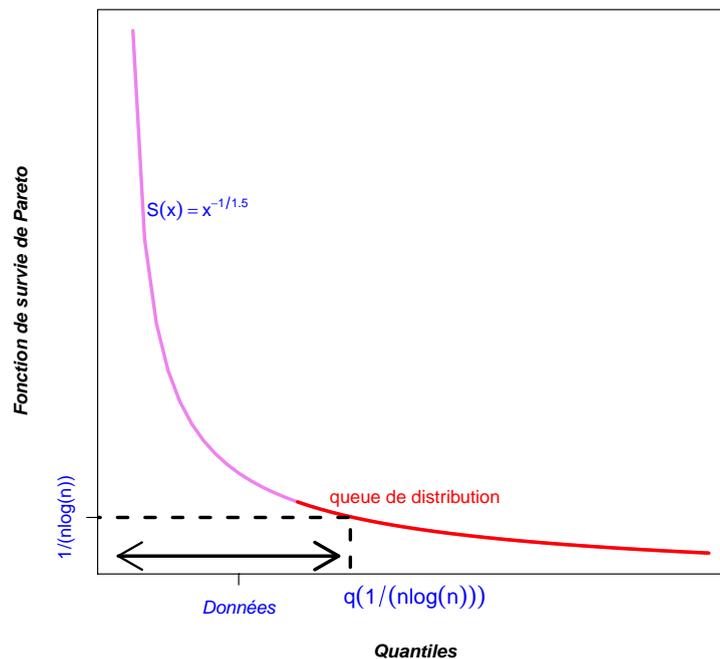


FIGURE 1.1 – Explication graphique de quantiles extrêmes et queue de distribution.

La figure 1.1 révèle que le quantile d'ordre $1/(n \log(n))$ qui tend strictement vers 0 quand n tend vers l'infini, qu'on veut déterminer, se situe au delà du maximum des observations disponibles. Il faut donc une information sur la forme de la distribution en étudiant la loi du maximum. Remarquons d'abord

$$\lim_{n \rightarrow \infty} F_{Y_{(n)}}(y) = \lim_{n \rightarrow \infty} [F(y)]^n = \mathbb{1}_{\{y \geq y_F\}} = \begin{cases} 1 & \text{si } y \geq y_F \\ 0 & \text{si } y < y_F \end{cases} \quad (1.2)$$

où $y_F = \sup \{y \in \mathbb{R}, F(y) < 1\}$ est le point terminal de la loi F . L'équation (1.2) montre que la loi du maximum, $Y_{(n)}$, est dégénérée donc fournit peu d'information.

Ainsi, ce qui suit donne plus d'information sur la forme de la queue de distribution avec les constantes de normalisation.

1.2.1 Loïs des valeurs extrêmes

La Théorie des Valeurs Extrêmes (**EVT**) repose principalement sur deux résultats. Ces deux résultats nous donnent le comportement asymptotique de la variable aléatoire $Y_{(n)}$ ou des dépassements d'un seuil u . L'intérêt de ces résultats provient du fait qu'il n'est pas nécessaire de connaître la loi, F du processus Y que nous souhaitons prédire. Cependant, l'EVT est analogue au Théorème Central Limite (**TCL**) mais pour les phénomènes extrêmes. Ainsi, là où le TCL montre que la moyenne empirique de la variable Y converge vers une loi normale (indépendamment de la loi de la variable d'intérêt et lorsque les moments d'ordre 1 et 2 existent); l'EVT établit des résultats analogues mais pour les valeurs extrêmes de Y . Voici le théorème qui donne la loi du maximum.

Théorème 1.2.1 (Fisher-Tippett-Gnedenko-Von Mises-Jenkinson)

Sous certaines conditions de régularités sur la fonction F , il existe un paramètre $\gamma \in \mathbb{R}$ et deux suites $(a_n)_{n \geq 1}$ positive et $(b_n)_{n \geq 1}$ réelle telles que pour tout $y \in \mathbb{R}$,

$$\lim_{n \rightarrow \infty} \mathbb{P} \left[\frac{Y_{(n)} - b_n}{a_n} \leq y \right] = \mathcal{H}_\gamma(y)$$

avec

$$\mathcal{H}_\gamma(y) = \begin{cases} e^{-(1+\gamma y)^{-1/\gamma}} \mathbb{1}_{\{1+\gamma y > 0\}} & \text{si } \gamma \neq 0, \\ e^{-e^{-y}} & \text{sinon.} \end{cases} \quad (1.3)$$

- γ l'indice des valeurs extrêmes.
- \mathcal{H}_γ est la loi des valeurs extrêmes.

Les suites $(a_n)_{n \geq 1}$ et $(b_n)_{n \geq 1}$ sont appelées **suites de normalisation**. Plusieurs auteurs se sont intéressés à la détermination de ces deux suites. Ainsi, Embrechts et al. ([46], exemple 3.3.29) ont proposé une estimation de ces constantes de normalisation dans le cas où la variable d'intérêt a pour loi, la loi normale centrée réduite.

Elles sont ainsi définies,

$$a_n = (2 \log n)^{-1/2} \text{ et } b_n = (2 \log n)^{1/2} - \frac{\log \log n + \log 4\pi}{2(2 \log n)^{1/2}} + o\left((\log n)^{-1/2}\right).$$

La figure 2.4 est obtenue en utilisant ces résultats. Elle compare graphiquement la loi du maximum des observations à celle de \mathcal{H}_0 c'est-à-dire $\gamma = 0$ et lorsque $n = 10$.

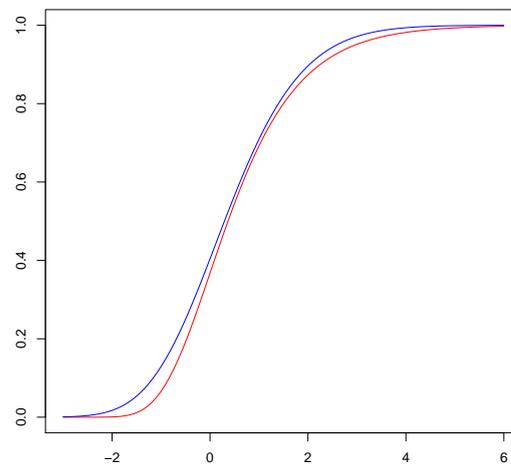


FIGURE 1.2 – Illustration graphique du théorème 1.2.1 sur une loi normale centrée réduite en comparant $\mathcal{H}_0(y)$ (rouge) et $\mathbb{P}\left[\frac{Y_{(n)} - b_n}{a_n} \leq y\right]$ avec $n = 10$ (bleu).

Ce théorème 1.2.1 admet un résultat similaire pour la modélisation des dépassements au dessus d'un seuil.

Théorème 1.2.2 (Balkema-de Haan-Pickands)

S'il existe des constantes $\{a_n\}_{n \geq 0}$ positives et $\{b_n\}_{n \geq 0}$ telles que

$$\mathbb{P} \left[\frac{Y_{(n)} - b_n}{a_n} \leq y \right] \rightarrow \mathbb{G}(y), n \rightarrow +\infty \quad (1.4)$$

où \mathbb{G} a la même forme que \mathcal{H}_γ . Alors,

$$\mathbb{P}[Y \leq y | Y > u] \rightarrow \mathbb{H}(y), u \rightarrow u_f \quad (1.5)$$

où $u_f = \sup\{u \in \mathbb{R} : \mathbb{P}[Y \leq u] < 1\}$ et

$$\mathbb{H}(y) = 1 - \left[1 + \xi \left(\frac{y - u}{\sigma} \right) \right]_+^{-1/\gamma} \quad (1.6)$$

où $z_+ = \max(0, z)$ et $\sigma > 0$. σ et γ correspondent respectivement aux paramètres d'échelle et de **forme**. Le paramètre u sera indifféremment appelé **seuil** ou paramètre de **position**.

De même que le Théorème 1.2.1, ce théorème affirme que : « La loi de $Y | Y > u$ converge vers une unique loi non dégénérée appelée la loi de Pareto Généralisée (G.P.D) ». La modélisation des valeurs extrêmes consiste à exploiter les deux théorèmes qui donnent des informations sur la queue de distribution. Ces deux théorèmes admettent un paramètre qui révèle la forme de la fonction de répartition. Selon le signe de ce paramètre on classe les lois de la statistique en trois familles appelées **domaines d'attraction**.

1.2.2 Domaines d'attraction

Selon le signe de γ , on distingue trois cas de domaines d'attraction (D.A.) :

- Si $\gamma > 0$, F appartient au **D.A. de Fréchet**, et l'on note $\mathbf{F} \in \mathbf{D.A. (Fréchet)}$. Il contient toutes les lois dont la fonction de survie décroît comme une fonction puissance. Ce sont les lois à « queue lourde ». Les distributions du domaine de Fréchet sont beaucoup utilisées en fiabilité mécanique, dans les phénomènes climatiques tels que la météorologie, l'hydrologie, la vitesse du vent enregistrée en continu dans les aéroports et en finance dans les études de risque.
- Si $\gamma = 0$, F appartient au **D.A. de Gumbel**, et l'on note $\mathbf{F} \in \mathbf{D.A. (Gumbel)}$. Ce sont les lois dont la fonction de survie décroît vers zéro à une vitesse exponentielle. Ces distributions sont souvent utilisées pour faire des prévisions dans les événements environnementaux tels que le séisme (le tremblement de

terre), l'hydrologie (les inondations, la destruction des barrages), etc.

- $\gamma < 0$, F appartient au **D.A. de Weibull**, et l'on note $\mathbf{F} \in \mathbf{D.A. (Weibull)}$. Ce domaine regroupe toutes les lois dont le point terminal, $x_F = \inf \{x, F(x) \geq 1\}$ est fini. Les distributions de type de Weibull sont souvent utilisées pour décrire la résistance mécanique d'un matériau ou encore le temps de fonctionnement d'un appareil électronique ou mécanique.

Dans ce tableau 1.1, nous avons regroupé quelques lois et leur domaine d'attraction comme dans les Tableaux 3.4.2 à 3.4.4 dans Embrechts et al. (1997, [46]).

Domaines d'attraction	Gumbel $\gamma = 0$	Fréchet $\gamma > 0$	Weibull $\gamma < 0$
Lois	Gumbel Weibull Gamma Normale Logistique Log-normale Exponentielle Benktander-type-I Benktander-type-II	Burr Pareto Cauchy Student Log-gamma Log-logistique α -Stable avec $\alpha < 2$	Beta Uniforme ReverseBurr

Tableau 1.1 – Quelques exemples de lois classées selon leurs domaines d'attraction.

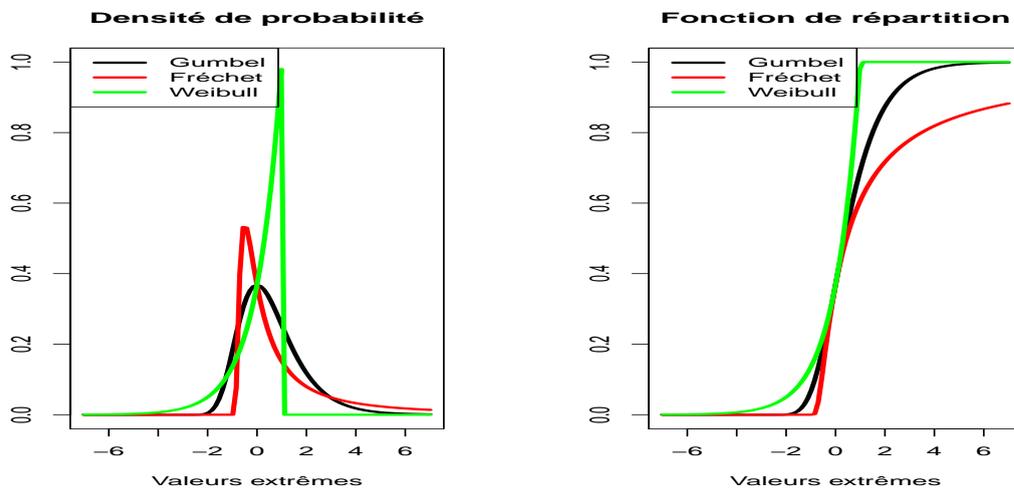


FIGURE 1.3 – Représentation de la densité et fonction de répartition : **Gumbel** ($\gamma = 0$), **Fréchet** ($\gamma = 1$) et **Weibull** ($\gamma = -1$).

Ces domaines d'attraction sont caractérisés via les fonctions à variations régulières.

1.2.2.1 Les fonctions à variations régulières

Dans cette partie, nous allons définir quelques notions essentielles de la théorie des valeurs extrêmes permettant de caractériser les domaines d'attraction. Cependant, ces fonctions à variations régulières permettent d'avoir une écriture unique pour chaque domaine d'attraction.

Définition 1 On dit qu'une fonction $U(\cdot)$ est à **variations régulières à l'infini** d'indice $\rho \in \mathbb{R}$ à l'infini, que l'on notera $U(\cdot) \in \mathcal{RV}_\rho$, si U est positive à l'infini (i.e. s'il existe A tel que pour tout $x > A$, $U(x) > 0$) et si pour tout $\lambda > 0$,

$$\lim_{x \rightarrow \infty} \frac{U(\lambda x)}{U(x)} = \lambda^\rho.$$

Si $\rho = 0$ c'est-à-dire $U(\cdot) \in \mathcal{RV}_0$ alors, la fonction $U(\cdot)$ est appelée **fonction à variations lentes à l'infini**, notée pour la suite $L(\cdot)$.

Parmi les fonctions à variations lentes, on peut citer :

- les fonctions possédant une limite strictement positive à l'infini;
- les fonctions l telles que

$$\exists M > 0, \forall x \geq M \quad l(x) = c + dx^{-\beta}(1 + o(1))$$

où $c, \beta > 0$ et $d \in \mathbb{R}$.

L'ensemble de ces fonctions l est appelé **classe de Hall**.

Par contre si $\rho = \infty$, on parle de fonction à **variations rapides à l'infini**. On peut montrer facilement que toute fonction $U(\cdot)$ à variations régulières d'indice $\rho \in \mathbb{R}$ s'écrit,

$$U(x) = x^\rho L(x), L(\cdot) \in \mathcal{RV}_0.$$

Toutes les fonctions $L(\cdot)$ à variations lentes s'écrivent sous la forme :

$$L(x) = c(x) \exp \left\{ \int_1^x \frac{\Delta(u)}{u} du \right\} \quad (x \geq 1),$$

où $c(x) \rightarrow c > 0$ et $\Delta(x) \rightarrow 0$ lorsque $x \rightarrow \infty$. Cette formule des fonctions à variations lentes s'appelle **représentation de Karamata**. Dans le cas où la fonction $c(\cdot)$ est constante, la fonction $L(\cdot)$ correspondante est dite **normalisée**. Si la fonction L est **normalisée** alors elle est dérivable de dérivée L' définie pour tout x par

$$L'(x) = \frac{\Delta(x)L(x)}{x}.$$

Proposition 1.2.1 (Resnick (1987, [81]), **Proposition 0.5**)

Si U est une fonction à variations régulières d'indice ρ à l'infini alors, pour tout $0 < a < b$,

$$\sup_{\lambda \in [a,b]} \left| \frac{U(\lambda x)}{U(x)} - \lambda^\rho \right| \rightarrow 0$$

lorsque $x \rightarrow \infty$.

Proposition 1.2.2 (de Haan et Ferreira (2006, [35]), **Proposition B.1.9**)

Soient L, L_1 et L_2 des fonctions à variations lentes à l'infini

i) Pour tout $\gamma > 0$,

$$\lim_{n \rightarrow \infty} n^\gamma L(n) = +\infty \text{ et } \lim_{n \rightarrow \infty} n^{-\gamma} L(n) = 0.$$

ii) Si a_n et b_n sont deux suites qui convergent vers l'infini lorsque n tend vers l'infini et telles que

$$\liminf_{n \rightarrow \infty} \frac{a_n}{b_n} > 0 \text{ et } \limsup_{n \rightarrow \infty} \frac{a_n}{b_n} < \infty,$$

alors

$$L(a_n) \sim L(b_n) \text{ i.e. } \frac{L(a_n)}{L(b_n)} \rightarrow 1$$

lorsque $n \rightarrow \infty$.

iii) $(\log L(x)) / \log x \rightarrow 0$ quand $x \rightarrow \infty$.

iv) L^α est une fonction à variations lentes à l'infini pour tout $\alpha \in \mathbb{R}$.

v) $L_1 L_2$ et $L_1 + L_2$ sont des fonctions à variations lentes à l'infini et en plus $L_1 \circ L_2$ l'est aussi si $L_2(x) \rightarrow \infty$ quand $x \rightarrow \infty$.

Inégalité de Potter

Un résultat important de l'analyse des valeurs extrêmes portant sur les fonctions à variations régulières à l'infini est l'inégalité de Potter.

Proposition 1.2.3 Soit g une fonction à variations régulières d'indice α à l'infini . Alors pour tout réel A strictement supérieur à 1 et pour tout ε strictement positif, il existe M tel que :

$$\forall x \geq M \quad \forall y \geq M \quad \frac{g(y)}{g(x)} \leq A \max \left\{ \left(\frac{y}{x} \right)^{\alpha+\varepsilon}, \left(\frac{y}{x} \right)^{\alpha-\varepsilon} \right\}.$$

Cette inégalité peut également s'écrire :

$$\forall \varepsilon > 0 \quad \exists t_0 \quad \forall y \geq 1 \quad \forall t \geq t_0 \quad (1 - \varepsilon)y^{\alpha-\varepsilon} \leq \frac{g(ty)}{g(t)} \leq (1 + \varepsilon)y^{\alpha+\varepsilon}.$$

Théorème 1.2.3 (Théorème Taubérien)

Une variable aléatoire Y de fonction de répartition F et de transformée de Laplace \mathcal{L}_F est à variations régulières d'indice $-\alpha$ ($\alpha \leq 0$) si les conditions équivalentes suivantes sont vérifiées ($L_F, L_{F^\leftarrow}, L_{\mathcal{L}}$ et L_f désignent des fonctions à variations lentes à l'infini) :

- (a) \bar{F} est à variations régulières d'indice $-\alpha$ ie $\bar{F}(y) = y^{-\alpha} L_F(y)$.
- (b) La fonction quantile est à variations régulières :

$$F^\leftarrow \left(1 - \frac{1}{y} \right) = y^{1/\alpha} L_{F^\leftarrow}(y).$$

- (c) La transformée de Laplace de F vérifie $\mathcal{L}_F(t) = t^\alpha L_{\mathcal{L}}(1/t)$.
- (d) La densité existe et vérifie $y f(y) / \bar{F}(y) \rightarrow \alpha$ quand $y \rightarrow \infty$, alors la densité f , est à variations régulières d'indice $-(\alpha + 1)$: $f(y) = y^{-(\alpha+1)} L_f(y)$.

Ce théorème révèle que la condition relative à la transformée de Laplace permet d'établir que la propriété de variations régulières à l'infini à paramètre fixé est stable par convolution. Pour plus détails sur ce théorème, le lecteur pourra consulter le livre de Bingham et al. (cf [17]).

1.2.2.2 Domaine d'attraction de Fréchet

Théorème 1.2.4 (Caractérisation de D.A. (Fréchet))

Une fonction de répartition $F(\cdot)$ appartient au domaine d'attraction de Fréchet, $F(\cdot) \in \mathbf{D.A. (Fréchet)}$, avec un indice des valeurs extrêmes $\gamma > 0$ si et seulement si $y_F = \sup \{x : F(x) < 1\} = +\infty$ et sa fonction de survie $\bar{F}(\cdot) \in \mathcal{RV}_{-1/\gamma}$ c'est-à-dire

$$\bar{F}(x) = x^{-1/\gamma} L(x) \Leftrightarrow q(\alpha) = \alpha^{-\gamma} l(\alpha^{-1}), (L, l) \in \mathcal{RV}_0^2$$

où $\alpha \in]0, 1[$ est appelé **ordre du quantile** $q(\alpha)$.

Dans ce cas un choix possible des suites de normalisation a_n et b_n du théorème 1.2.1 est :

$$a_n = F^\leftarrow \left(1 - \frac{1}{n} \right) = \bar{F}^\leftarrow \left(\frac{1}{n} \right) = n^\gamma L(n) \text{ et } b_n = 0, L \in \mathcal{RV}_0$$

où F^\leftarrow est l'inverse généralisée de F . Pour plus de détails le lecteur pourra consulter les ouvrages d'Embrechts et al. (1997, [46]), de Bingham et al. (1987, [17]), de Coles (2001, [21]).

Toutes les fonctions définies précédemment sont indicées par un unique paramètre

positif γ inconnu. Dans la littérature plusieurs chercheurs se sont intéressés à l'estimation de cet indice des valeurs extrêmes et son quantile extrême correspondant (voir Hill [67], 1975; Pickands [78], 1975; de Haan et Resnick [37], 1980; Hall [64], 1982; Mason [74], 1982; Csörgő et al. [22], 1985; Hall et Welsh [66], 1985; Dekkers et al. [40], 1989; Davis et Resnick [32], 1989; Grubel et Wolf [63], 1994). Et l'estimateur de l'indice le plus célèbre est

☞ Estimateur de Hill (1975, [67])

$$\begin{aligned}\hat{\gamma}_{k,n}^H &= \frac{1}{k} \sum_{i=1}^k \log Y_{(n-i+1)} - \log Y_{(n-k)} \\ &= \frac{1}{k} \sum_{i=1}^k i \left(\log Y_{(n-i+1)} - \log Y_{(n-i)} \right)\end{aligned}$$

où $Y_{(1)} \leq \dots \leq Y_{(n)}$ est l'échantillon ordonné associé aux variables aléatoires Y_1, \dots, Y_n . La construction de cet estimateur est basée sur la méthode du Maximum de Vraisemblance où on se sert des statistiques d'ordre supérieur à un certain seuil u , pour ne garder que les observations les plus grandes, de façon à ce que qu'elles suivent approximativement une distribution de Paréto. Après sa construction, plusieurs chercheurs ont essayé de déterminer ses propriétés asymptotiques. Ainsi, Mason (1982, [74]) a prouvé la consistance faible de l'estimateur de Hill pour toute suite $k = k(n)$ satisfaisant $k \rightarrow \infty$ et $k/n \rightarrow 0$ quand $n \rightarrow \infty$ appelée suite intermédiaire d'entiers. La condition $k \rightarrow \infty$ assure que la taille de statistiques d'ordre k est assez grande afin d'obtenir des estimateurs stables. Par contre, la condition $k/n \rightarrow 0$ permet de rester dans la queue de distribution. Davis et Resnick (1984, [31]) ont proposé sa normalité asymptotique sous les conditions de Von Mises ; Csörgő et Mason (1985, [23]) ont présenté sa normalité asymptotique en introduisant l'approximation des processus empiriques par les ponts browniens. Dans cette même lancée Resnick et de Haan (1998, [38]) ont montré cette propriété asymptotique.

Remarque 1 Cet estimateur de Hill est un cas particulier de l'estimateur de Csörgő et al. (1985, [22]) défini :

$$\hat{\gamma}_{n,k}^K = k^{-1} \sum_{j=1}^k j K(j/k) \left(\log Y_{(n-j+1)} - \log Y_{(n-j)} \right) \Big/ k^{-1} \sum_{j=1}^k K(j/k)$$

où K est un noyau d'intégrale égale à 1. Selon le choix du noyau K , plusieurs estimateurs peuvent en résulter notamment celui de Hill avec $K(x) = \mathbb{1}_{]0,1]}(x)$.

Nous pouvons aussi citer :

☞ Estimateur de Pickands (1975, [78])

$$\widehat{\gamma}_{k,n}^P = \frac{1}{\log 2} \log \left(\frac{Y_{(n-k+1)} - Y_{(n-2k+1)}}{Y_{(n-2k+1)} - Y_{(n-4k+1)}} \right), k = 1, \dots, \lfloor n/4 \rfloor$$

où $\lfloor x \rfloor$ est la partie entière de x .

☞ Estimateur des Moments de Dekkers et al. (1989, [40]) :

$$\begin{aligned} \widehat{\gamma}_{k,n}^M &= \mathcal{M}_{k,n}^{(1)} + 1 + \frac{1}{2} \left(1 - \frac{(\mathcal{M}_{k,n}^{(1)})^2}{\mathcal{M}_{k,n}^{(2)}} \right)^{-1} \\ &= \widehat{\gamma}_{k,n}^H + 1 + \frac{1}{2} \left(1 - \frac{(\widehat{\gamma}_{k,n}^H)^2}{\mathcal{M}_{k,n}^{(2)}} \right)^{-1} \end{aligned}$$

où

$$\mathcal{M}_{k,n}^{(r)} = \frac{1}{k} \sum_{i=1}^k \left(\log Y_{(n-i+1)} - \log Y_{(n-k)} \right)^r, \quad r = 1, 2.$$

Il est appelé l'estimateur des Moments car les $\mathcal{M}_{k,n}^{(r)}$ peuvent être considérés comme des moments empiriques d'ordre r .

☞ Estimateur UH de Beirlant et al. (1996, [13]) :

$$\begin{aligned} \widehat{\gamma}_{k,n}^{UH} &= \frac{1}{k} \sum_{j=1}^k \log UH_{(j)} - \log UH_{(k+1)} \\ &= \frac{1}{k} \sum_{j=1}^k \log \left(Y_{(n-j)} \widehat{\gamma}_{j,n}^H \right) - \log \left(Y_{(n-k-1)} \widehat{\gamma}_{k+1,n}^H \right) \end{aligned}$$

où

$$UH_{(j)} = Y_{(n-j)} \left(\frac{1}{j} \sum_{i=1}^j \log Y_{(n-i+1)} - \log Y_{(n-j)} \right) = Y_{(n-j)} \widehat{\gamma}_{j,n}^H.$$

Rappelons à présent les propriétés asymptotiques de ces estimateurs. Pour cela, nous allons commencer par les conditions du premier et du second ordre avec les fonctions quantiles définies ainsi :

$$Q(s) = F^{\leftarrow}(s) := \inf \{x \in \mathbb{R} : F(x) \geq s\}, \quad 0 < s < 1$$

et

$$U(t) := Q(1 - 1/t) = \left(1/\overline{F}\right)^{\leftarrow}(t), \quad 1 < t < \infty.$$

Proposition 1.2.4 (*Conditions du premier ordre, de Haan et Ferreira (2006, [35])*)

Les assertions suivantes sont équivalentes :

(a) F est à queue lourde

$$F \in D.A.(\text{Fréchet}), \quad \gamma > 0.$$

(b) \bar{F} est une fonction à variations régulières à l'infini d'indice $-1/\gamma$

$$\lim_{t \rightarrow \infty} \frac{1 - F(ty)}{1 - F(t)} = y^{-1/\gamma}, \quad y > 0. \quad (1.7)$$

(c) $Q(1 - s)$ est une fonction à variations régulières à zéro d'indice $-\gamma$

$$\lim_{s \rightarrow 0} \frac{Q(1 - sy)}{Q(1 - s)} = y^{-\gamma}, \quad y > 0. \quad (1.8)$$

(d) U est une fonction à variations régulières à l'infini d'indice γ

$$\lim_{t \rightarrow \infty} \frac{U(ty)}{U(t)} = y^\gamma, \quad y > 0. \quad (1.9)$$

Proposition 1.2.5 (Conditions du second ordre de Haan et Ferreira (2006, [35]))

Une fonction de répartition $F(\cdot) \in D.A.(\text{Fréchet})$, $\gamma > 0$, admet une condition du second ordre à l'infini si elle satisfait à l'une des assertions suivantes :

(a) Il existe un paramètre $\rho \leq 0$ et une fonction $A_1(\cdot)$ qui tend vers 0 (ne change pas de signe à l'infini) définie par, $\forall y > 0$

$$\lim_{t \rightarrow \infty} \frac{(1 - F(ty))/(1 - F(t)) - y^{-1/\gamma}}{A_1(t)} = y^{-1/\gamma} \frac{y^\rho - 1}{\rho}.$$

(b) S'il existe un paramètre $\rho \leq 0$ et une fonction $A_2(\cdot)$ qui tend vers 0 (ne change pas de signe à zéro) définie par, $\forall y > 0$

$$\lim_{s \rightarrow 0} \frac{Q(1 - sy)/Q(1 - s) - y^{-\gamma}}{A_2(s)} = y^{-\gamma} \frac{y^\rho - 1}{\rho}.$$

(c) S'il existe un paramètre $\rho \leq 0$ et une fonction $A(\cdot)$ qui tend vers 0 (ne change pas de signe à l'infini) définie par, $\forall y > 0$

$$\lim_{t \rightarrow \infty} \frac{U(ty)/U(t) - y^\gamma}{A(t)} = y^\gamma \frac{y^\rho - 1}{\rho}.$$

si $\rho = 0$, on remplace $(y^\rho - 1)/\rho$ par $\log y$.

Les fonctions $A(\cdot)$, $A_1(\cdot)$ et $A_2(\cdot)$ sont à variations régulières à l'infini d'indices respectifs ρ , ρ/γ et $-\rho$ avec $A_1(t) = A(1/(1 - F(t)))$ et $A_2(s) = A(1/s)$. Leur but est de contrôler la vitesse de convergence dans (1.7), (1.8) et (1.9) respectivement.

Ces deux conditions ont permis de déterminer les propriétés asymptotiques de certains estimateurs de l'indice des valeurs extrêmes.

Théorème 1.2.5 (de Haan et Ferreira (2006, [35]), Théorème 3.2.5)

Supposons $F \in \text{D.A.}(\text{Fréchet})$ de paramètre de forme $\gamma > 0$, $k \rightarrow \infty$ et $k/n \rightarrow 0$ si $n \rightarrow \infty$.

(a) Consistance faible :

$$\widehat{\gamma}_{k,n}^H \xrightarrow{\mathbb{P}} \gamma \quad n \rightarrow \infty$$

(b) Consistance forte :

Supposons que $k/\log \log n \rightarrow \infty$ si $n \rightarrow \infty$

$$\widehat{\gamma}_{k,n}^H \xrightarrow{p.s.} \gamma \quad n \rightarrow \infty$$

(c) Normalité asymptotique :

Supposons que $F(\cdot)$ vérifie la condition (c) de la proposition 1.2.5 et $\sqrt{k}A(n/k) \rightarrow \lambda$ si $n \rightarrow \infty$ alors,

$$\sqrt{k} \left(\widehat{\gamma}_{k,n}^H - \gamma \right) \xrightarrow{d} \mathcal{N} \left(\frac{\lambda}{1-\rho}, \gamma^2 \right) \quad n \rightarrow \infty.$$

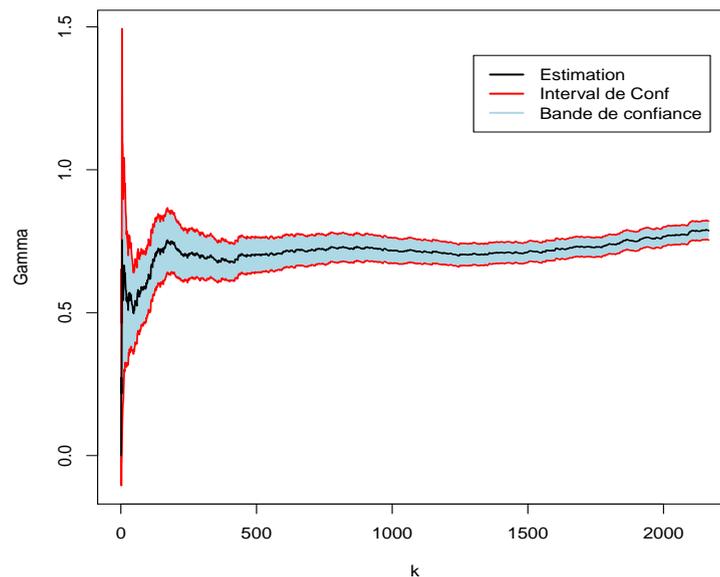


FIGURE 1.4 – Comportement asymptotique de l'estimateur de Hill avec des intervalles de confiance au niveau de confiance 0.95 sur les données d'assurance incendie Danoise pour la période allant de 01/01/1980 au 31/12/1990.

Théorème 1.2.6 (Pickands (1975, [78]), Dekkers et de Haan (1989, [39]))

Supposons que $F \in D.A.(\mathcal{H}_\gamma)$, $\gamma \in \mathbb{R}$ et $k/n \rightarrow 0$ si $n \rightarrow \infty$.

(a) Consistance faible :

$$\hat{\gamma}_{k,n}^P \xrightarrow{\mathbb{P}} \gamma \quad n \rightarrow \infty$$

(b) Consistance forte

Supposons que $k/\log \log n \rightarrow \infty$ si $n \rightarrow \infty$

$$\hat{\gamma}_{k,n}^P \xrightarrow{p.s.} \gamma \quad n \rightarrow \infty$$

(c) Normalité asymptotique :

$$\sqrt{k} \left(\hat{\gamma}_{k,n}^P - \gamma \right) \xrightarrow{d} \mathcal{N} \left(0, \sigma^2 \right) \quad n \rightarrow \infty$$

$$\text{où } \sigma^2 = \frac{\gamma^2(2^{2\gamma+1}+1)}{(2(2^\gamma-1)\log 2)^2}.$$

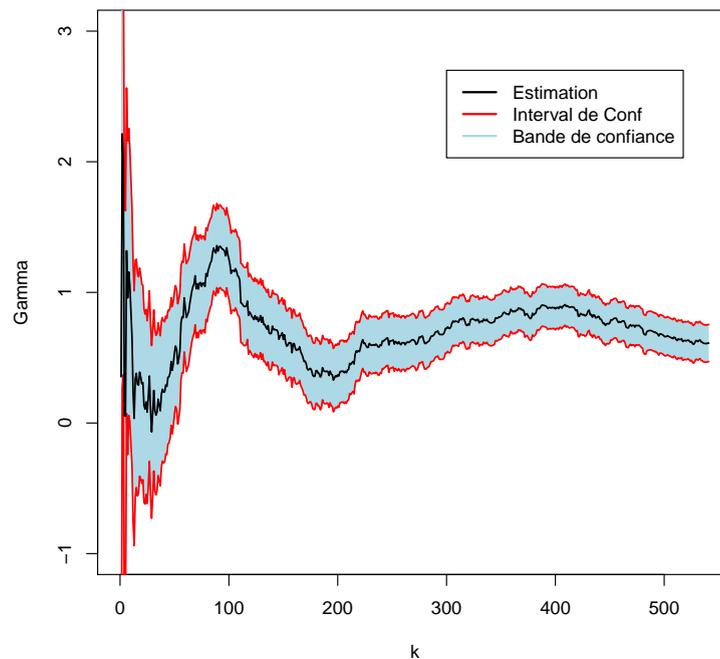


FIGURE 1.5 – Comportement asymptotique de l'estimateur de Pickands avec des intervalles de confiance au niveau de confiance 0.95 sur les données d'assurance incendie Danoise pour la période allant de 01/01/1980 au 31/12/1990.

Théorème 1.2.7 (Dekkers al. (1989, [40]))

Supposons que $F \in D.A.(\mathcal{H}_\gamma)$, $\gamma \in \mathbb{R}$ et $k/n \rightarrow 0$ si $n \rightarrow \infty$.

(a) Consistance faible :

$$\widehat{\gamma}_{k,n}^M \xrightarrow{\mathbb{P}} \gamma \quad n \rightarrow \infty$$

(b) Consistance forte Supposons que $k/(\log n)^\delta \rightarrow \infty$ si $n \rightarrow \infty$, avec $\delta > 0$

$$\widehat{\gamma}_{k,n}^M \xrightarrow{p.s.} \gamma \quad n \rightarrow \infty$$

(c) Normalité asymptotique : (voir Théorème 3.1 et corollaire 3.2 de [40])

$$\sqrt{k} (\widehat{\gamma}_{k,n}^M - \gamma) \xrightarrow{d} \mathcal{N}(0, \sigma^2) \quad n \rightarrow \infty$$

où

$$\sigma^2 = \begin{cases} 1 + \gamma^2 & \text{si } \gamma \geq 0, \\ (1 - \gamma^2)(1 - 2\gamma) \left(4 - 8 \frac{1-2\gamma}{1-3\gamma} + \frac{(5-11\gamma)(1-2\gamma)}{(1-3\gamma)(1-4\gamma)} \right) & \text{si } \gamma < 0. \end{cases}$$

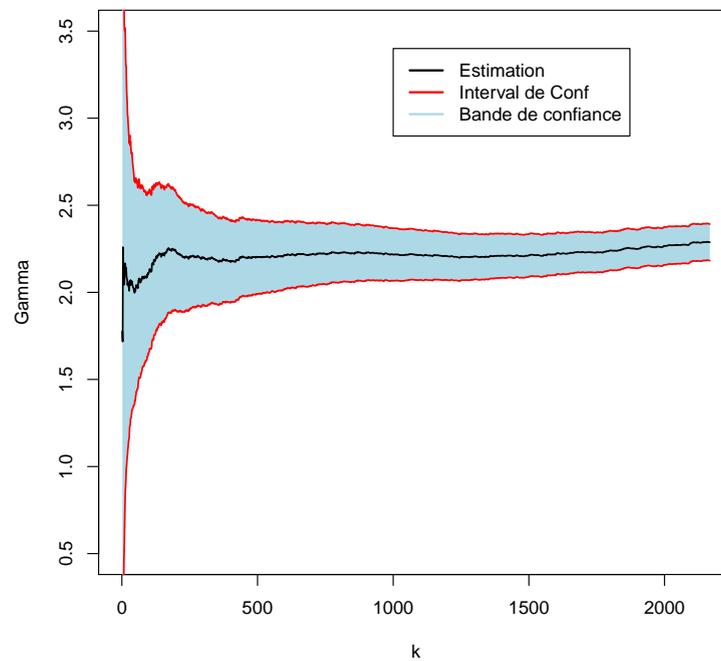


FIGURE 1.6 – Comportement asymptotique de l'estimateur des moments avec des intervalles de confiance au niveau de confiance 0.95 sur les données d'assurance incendie Danoise pour la période allant de 01/01/1980 au 31/12/1990.

Théorème 1.2.8 (Beirlant et al. (2005, [6]), Théorème 1)

Supposons que $F(\cdot)$ vérifie la condition (c) de la proposition 1.2.5 et $\sqrt{k}A(n/k) \rightarrow \lambda$ si $n \rightarrow \infty$ alors,

$$\sqrt{k} \left(\hat{\gamma}_{k,n}^{UH} - \gamma \right) \xrightarrow{d} \mathcal{N} \left(\frac{\lambda}{1-\rho}, \sigma^2 \right) \quad n \rightarrow \infty$$

où

$$\sigma^2 = \begin{cases} 1 + \gamma^2 & \text{si } \gamma \geq 0, \\ \frac{(1-\gamma)(1+\gamma+2\gamma^2)}{1-2\gamma} & \text{si } \gamma < 0. \end{cases}$$

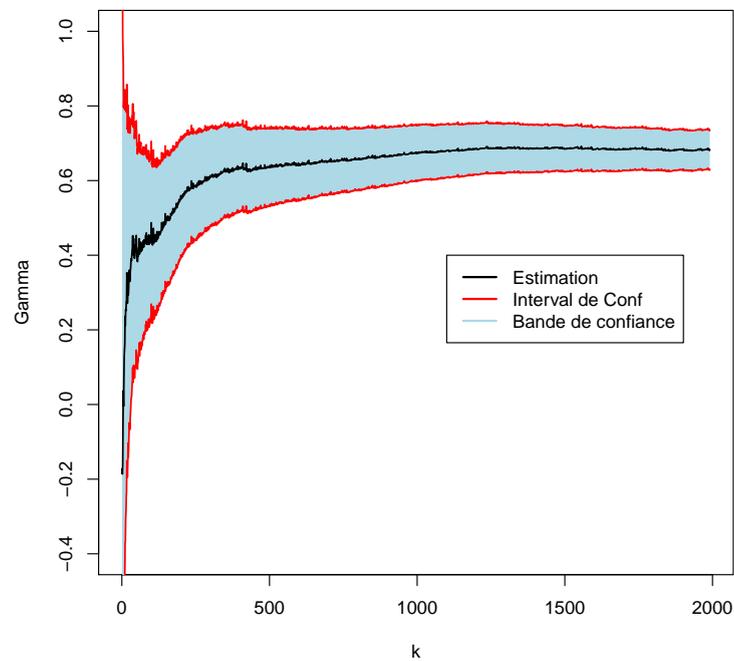


FIGURE 1.7 – Comportement asymptotique de l'estimateur UH avec des intervalles de confiance au niveau de confiance 0.95 sur les données d'assurance incendie Danoise pour la période allant de 01/01/1980 au 31/12/1990.

1.2.2.3 Domaine d'attraction de Weibull

Théorème 1.2.9 Une fonction de répartition $F(\cdot)$ appartient au domaine d'attraction de Weibull avec $\gamma < 0$ si et seulement si $y_F < +\infty$ et en plus $1 - F^*$ est une fonction à variations régulières d'indice $-1/\gamma$ c'est-à-dire

$$F^*(y) = 1 - (y_F - y)^{-1/\gamma} L\left((y_F - y)^{-1}\right)$$

avec

$$F^*(y) = \begin{cases} 0 & \text{si } y \leq 0, \\ F(y_F - y^{-1}) & \text{sinon.} \end{cases}$$

où $L(\cdot)$ est une fonction à variations lentes à l'infini.

Dans ce domaine d'attraction les suites de normalisation sont déterminées comme suit :

$$a_n = y_F - F^{\leftarrow}\left(1 - \frac{1}{n}\right) \text{ et } b_n = y_F.$$

1.2.2.4 Domaine d'attraction de Gumbel

Théorème 1.2.10 Une fonction de répartition $F(\cdot)$ appartient au domaine d'attraction de Gumbel si et seulement si

$$y_F \leq \infty \text{ et } \bar{F}(y) = c(y) \exp\left\{-\int_{y_0}^y \frac{g(t)}{a(t)} dt\right\}, \quad y_0 < y < y_F,$$

où $c(y) \rightarrow c > 0$, $g(y) \rightarrow 1$ et $a'(y) \rightarrow 0$ si $y \rightarrow y_F$. Et un choix possible pour la fonction $a(\cdot)$ est la fonction moyenne des excès définie par

$$a(y) = \frac{1}{\bar{F}(y)} \int_y^{y_F} F(t) dt, \quad y < y_F.$$

Dans ce cas les suites a_n et b_n sont ainsi définies :

$$b_n = F^{\leftarrow}\left(1 - \frac{1}{n}\right) \text{ et } a_n = a(b_n).$$

Rappelons maintenant quelques représentations graphiques de modélisation des valeurs extrêmes.

1.2.2.5 Représentations graphiques

En présence de données réelles, le modélisateur des valeurs extrêmes cherche à répondre à cette question :

Dans quel domaine d'attraction, parmi les trois domaines, est-il raisonnable de se placer?

Pour répondre à cette question plusieurs représentations graphiques s'offrent à lui selon chaque domaine d'attraction. Ainsi, il peut visualiser quelques graphes pour avoir une meilleure idée sur le comportement de la queue de distribution.

- « Le Pareto quantile plot » (D. A. de Fréchet)

Ce domaine d'attraction a été largement étudié dans la littérature car il renferme de nombreuses applications pratiques. Cette méthode est basée sur la fonction queue définie par

$$U(y) = \inf \left\{ x : F(x) \geq 1 - \frac{1}{y} \right\}.$$

On peut l'écrire aussi en utilisant la fonction à variations lentes :

$$U(y) = y^\gamma l_U(y),$$

où $l_U(\cdot)$ est une fonction à variations lentes à l'infini.

« Le Pareto quantile plot » correspond au graphe de

$$\left(\log \frac{n+1}{i}, \log Y_{(n-i+1)}, i = 1, \dots, n \right).$$

Cette représentation graphique vérifie l'appartenance à la famille de Fréchet à certaines observations réelles.

Nous pouvons écrire cette approximation en utilisant les propriétés sur les fonctions à variations lentes à l'infini :

$$\log U(y) \sim \gamma \log y \quad \text{quand } y \rightarrow \infty.$$

En utilisant la version de la fonction queue on obtient

$$\log Y_{(n-i+1)} \sim \gamma \log \left(\frac{n+1}{i} \right) \quad \text{quand } \left(\frac{n+1}{i} \right) \rightarrow \infty.$$

Cette dernière écriture révèle que le « Pareto quantile plot » sera linéaire de pente γ .

- « L'exponentiel quantile plot » (Domaine de Gumbel)

Cette méthode est similaire à la précédente. Elle est la représentation gra-

phique de

$$\left(\log \frac{n+1}{i}, Y_{(n-i+1)}, i = 1, \dots, n \right).$$

Si $\gamma = 0$ ce graphe semble être approximativement linéaire.

- « Quantile plot généralisé »

En dehors de ces cas la distribution F appartient au domaine d'attraction de Weibull ($\gamma < 0$). Le modélisateur peut aussi tracer le « quantile plot généralisé » proposé par Beirlant et al. (1996, [13]), défini par le graphe de

$$\left(\log \frac{n+1}{i}, \log UH_{i,n}, i = 1, \dots, n \right)$$

où

$$UH_{i,n} = Y_{(n-i+1)} \left(\frac{1}{i} \sum_{j=1}^i \log Y_{(n-j+1)} - \log Y_{(n-i+1)} \right).$$

1.2.3 Quelques rappels sur la modélisation des valeurs extrêmes conditionnelles

1.2.3.1 Introduction

DANS ce paragraphe, nous allons rappeler la modélisation des valeurs extrêmes en présence de covariables à design fixe ou aléatoire. Dans la littérature, on distingue trois approches de modélisation de quantiles extrêmes conditionnels: l'approche paramétrique, semi-paramétrique et non-paramétrique.

- **L'approche paramétrique**

Smith (1989, [83]) et **Davison et Smith (1990, [33])** ont proposé des familles de modèles paramétriques basées sur les excès au dessus d'un seuil. Ainsi, ils modélisent les excès par une loi de Pareto généralisée dont les paramètres (forme ou échelle) sont des fonctions de la covariable. De ce fait, l'estimation est basée sur la méthode du maximum de vraisemblance ou moindres carrés.

- **L'approche semi-paramétrique**

Beirlant et Goegebeur (2003, [7]) adoptent l'approche semi-paramétrique après avoir transformé tout d'abord les données, puis ils les utilisent dans un modèle de régression exponentiel où les paramètres de ce dernier sont estimés par la méthode du maximum de vraisemblance. Pour plus de détails, le lecteur pourra consulter le livre de **Beirlant et al. (2004, [8])**.

- **L'approche non-paramétrique**

Hall et al. (2000, [65]) et **Davison et al. (2000, [34])** modélisent les valeurs extrêmes conditionnelles où la covariable est le **temps** dont les paramètres

sont estimés par la méthode non-paramétrique. Par contre, **Beirlant et al.** (2001, [11]) puis (2003, [12]) utilisent les statistiques d'ordre pour estimer les quantiles extrêmes conditionnels avec la méthode des polynômes locaux. Enfin, **Gardes et Girard** (2008, [51]) puis Gardes et al. (2010, [54]), ont proposé des estimateurs de quantiles extrêmes conditionnels avec la méthode non-paramétrique dont un résultat est la thèse de **Lékina** (2010, [72]). Dans cette thèse, il propose des estimateurs de l'indice et de quantiles extrêmes conditionnels à design fixe et aléatoire avec l'approche non-paramétrique. Ainsi, ce que nous allons rappeler dans cette partie, est l'un des résultats de cette thèse.

1.2.3.2 Estimation de l'indice et de quantiles extrêmes conditionnels

Soit Y une variable aléatoire réelle mesurée conjointement avec une covariable non-aléatoire $X \in \mathcal{X}$, où \mathcal{X} est un espace métrique muni d'une distance d . Soit $\{(x_i, Y_i), i = 1, \dots, n\}$ un échantillon d'observations indépendantes et identiquement distribuées du couple $(X, Y) \in \mathcal{X} \times \mathbb{R}$. On suppose que la distribution conditionnelle de Y , F est à « queue lourde » c'est-à-dire

$$\bar{F}(y, x) = y^{-1/\gamma(x)}l(y, x)$$

où $l \in \mathcal{RV}_0$ et $\gamma(\cdot)$ est une fonction positive et inconnue de la covariable x que l'on appelle « indice de queue conditionnel » ou « indice des valeurs extrêmes ». En utilisant la méthode de sélection proposée au chapitre 2 de la thèse, on obtient l'échantillon retenu par la procédure :

$$\{Z_i^x, i = 1, \dots, m_{n,x}\}$$

et son échantillon ordonné associé

$$Z_{(1)}^x \leq \dots \leq Z_{(m_{n,x})}^x.$$

Le quantile quantile extrême conditionnel d'ordre $\alpha_{m_{n,x}}$ est ainsi défini :

$$\hat{q}_1(\alpha_{m_{n,x}}, x) = Z_{(m_{n,x} - \lfloor m_{n,x} \alpha_{m_{n,x}} \rfloor)}^x$$

si $\alpha_{m_{n,x}}$ converge lentement vers 0, ie $\alpha_{m_{n,x}} \rightarrow 0$ et $m_{n,x} \alpha_{m_{n,x}} \rightarrow \infty$ lorsque $m_{n,x} \rightarrow \infty$ où $\lfloor a \rfloor$ désigne la partie entière de a . Par contre si $\alpha_{m_{n,x}}$ converge rapidement vers 0, ie $\alpha_{m_{n,x}} \rightarrow 0$ et $m_{n,x} \alpha_{m_{n,x}} \rightarrow c \in [0, 1[$ lorsque $m_{n,x} \rightarrow \infty$, le quantile extrême conditionnel est ainsi défini :

$$\hat{q}_2(\alpha_{m_{n,x}}, x) = \hat{q}_1(\beta_{m_{n,x}}, x) \left(\frac{\beta_{m_{n,x}}}{\alpha_{m_{n,x}}} \right)^{\hat{\gamma}_n(x)}$$

où $\beta_{m_n, x}$ converge lentement vers 0 et $\hat{\gamma}_n(x)$ est un estimateur de l'indice des valeurs extrêmes conditionnel.

Proposition 1.2.6 (Théorème 4.3.1 dans [72])

Soit $\alpha_{m_n, x}$ une suite qui converge lentement vers zéro. Sous la condition (A.1) et w_n définies dans le chapitre 2, si pour n tendant vers l'infini, s'il existe $\delta > 0$ tel que

$$(m_{n,x}\alpha_{m_{n,x}})^2 w_n \left(m_{n,x}^{-(1+\delta)} \right) \longrightarrow 0$$

alors,

$$(m_{n,x}\alpha_{m_{n,x}})^{1/2} \left(\frac{\hat{q}_1(\alpha_{m_{n,x}}, x)}{q(\alpha_{m_{n,x}}, x)} - 1 \right) \xrightarrow{d} \mathcal{N}(0, \sigma^2(x))$$

où $\sigma^2(x)$ est la variance de $\hat{\gamma}_n(x)$.

1.3 Quelques généralités sur la censure

Ce paragraphe est dédié aux rappels des notions essentielles sur la censure.

1.3.1 Fonction de survie et de risque

Définition 2 La fonction de survie est,

$$\begin{aligned} S(y) &= 1 - F(y) \\ &= \mathbb{P}(Y > y) \end{aligned}$$

c'est-à-dire pour y fixé, la probabilité de survivre jusqu'à l'instant y .

Définition 3 La fonction de risque instantané λ (ou taux de hasard), pour y fixé représente la probabilité de mourir dans un petit intervalle de temps après y , conditionnellement au fait d'avoir survécu jusqu'au temps y (c'est-à-dire le risque de mort instantané pour ceux qui ont survécu) :

$$\lambda(y) = \lim_{h \rightarrow 0} \frac{\mathbb{P}(y \leq Y < y + h | Y > y)}{h} = \frac{f(y)}{S(y)}$$

où f est la densité de probabilité de Y .

Définition 4 La fonction de risque cumulé est définie par :

$$A(y) = \int_0^y \lambda(u) du$$

Ces fonctions permettent de caractériser la durée de survie à étudier.

1.3.2 La notion de censure

Une des caractéristiques des données de survie est l'existence d'observations incomplètes. En effet, les données sont souvent recueillies partiellement, notamment, à cause des processus de censure et de troncature. Les données censurées proviennent du fait qu'on n'a pas accès à toute l'information. Au lieu d'observer des réalisations indépendantes et identiquement distribuées de durée Y , on observe la réalisation de la variable Y soumise à diverses perturbations indépendantes ou non de l'événement étudié.

1.3.2.1 Définition

Définition 5 La *variable de censure* C est définie par la non-observation de l'événement étudié. Si au lieu d'observer Y , on observe C , et que l'on sait que $Y > C$ (respectivement $Y < C$, $C_1 < Y < C_2$), on dit qu'il y a **censure à droite** (respectivement **censure à gauche**, **censure par intervalle**).

1.3.2.2 Caractéristiques

La censure est le phénomène le plus couramment rencontré lors du recueil de données en statistique .

Pour un individu donné i , nous allons considérer

- son temps de survie Y_i , de fonction de répartition F
- sa variable de censure C_i , de fonction de répartition G
- sa variable réellement observée Z_i de fonction de répartition H .

Dans la littérature on distingue trois types de censure :

☆ Censure à droite

La variable d'intérêt est dite censurée à droite si l'individu concerné n'a aucune information sur sa dernière observation. Ainsi, en présence de censure à droite les variables d'intérêt ne sont pas toutes observées. Un exemple typique est celui où l'événement considéré est le décès d'un patient malade et la durée d'observation est une durée totale d'hospitalisation. On trouve aussi ce genre de phénomène dans les études de fiabilité quand la panne d'un appareil ou d'un composant électronique ne permet pas de continuer l'observation pour un autre appareil ou composant . Nous pouvons aussi trouver ces genres de phénomènes en hydrologie, en pluviométrie, ... L'expérimentateur peut fixer une date de fin d'expérience et les observations pour les individus pour lesquels on n'a pas observé l'événement d'intérêt avant cette date seront censurées à droite.

☆ Censure à gauche

Il y a censure à gauche lorsque l'individu a déjà subi l'événement avant qu'il soit observé. On sait uniquement que la variable d'intérêt est inférieure ou égale à une variable connue. Par exemple si nous voulons étudier en fiabilité un certain composant électronique qui est branché en parallèle avec un ou plusieurs autres composants: le système peut continuer à fonctionner, quoique de façon aberrante, jusqu'à ce que cette panne soit détectée (par exemple lors d'un contrôle ou en cas de l'arrêt du système). Ainsi donc, la durée observée pour ce composant est censurée à gauche. Dans la vie courante il y a plusieurs phénomènes qui présentent à la fois des données censurées à droite et à gauche.

☆ Censure double ou mixte

Il y a censure mixte ou double dans un échantillon de données s'il y a à la fois censure à gauche et à droite dans cet échantillon. Les données sont censurées à la fois à droite et à gauche. Plusieurs modèles non paramétriques ont été présentés pour l'étude de la double censure. Par exemple, le modèle de Turnbull (1974, [88]) est le plus utilisé, et plusieurs travaux sont basés sur ce modèle. Dans la littérature d'autres modèles ont été proposés notamment la censure par intervalle.

☆ Censure par intervalle

Dans ce cas, comme son nom l'indique, on observe à la fois une borne inférieure et une borne supérieure de la variable d'intérêt. On retrouve ce modèle en général dans des études de suivi médical où les patients sont contrôlés périodiquement, si un patient ne se présente pas à un ou plusieurs contrôles et se présente ensuite après que l'événement d'intérêt se soit produit. Nous avons aussi ce genre de données qui sont censurées à droite ou, plus rarement, à gauche. Un avantage de ce type est qu'il permet de présenter les données censurées à droite ou à gauche par des intervalles du type $[c, +\infty[$ et $[0, c]$ respectivement.

Ces quatre catégories de censure décrites ci-dessus peuvent se présenter en fonction du mode ou mécanisme de censure. Ainsi, dans la littérature on retrouve les types suivants :

❶ La censure de type I

L'expérimentateur fixe une valeur (une date par exemple non aléatoire de fin d'expérience). Par exemple en épidémiologie on fixe la durée maximale de participation et vaut, pour chaque observation, la différence entre la date de fin d'expérience et la date d'entrée du patient dans l'étude. Le nombre d'événements observés est, quant à lui, aléatoire.

Soit C une valeur fixée. Par exemple en censure à droite, au lieu d'observer

les variables Y_1, \dots, Y_n qui nous intéressent, on observe Y_i uniquement lorsque $Y_i \leq C$, sinon on sait que on sait uniquement que $Y_i > C$. Ainsi, la variable observée est $Z_i = Y_i \wedge C$.

② La censure de type II

L'expérimentateur fixe a priori le nombre d'événements à observer. La date de fin d'expérience devient alors aléatoire, le nombre d'événements étant quant à lui, non aléatoire. Ce modèle est souvent utilisé dans les études de fiabilité, d'épidémiologie. Par exemple en épidémiologie on décide d'observer les durées de survie des n patients jusqu'à ce que k d'entre eux soient décédés et d'arrêter l'étude à ce moment là. Soient $Y_{(i)}$ et $Z_{(i)}$ les statistiques d'ordre des variables Y_i et Z_i . La date de censure est donc $Y_{(k)}$ et on observe uniquement les variables suivantes :

$$\begin{aligned} Z_{(1)} &= Y_{(1)} \\ &\vdots \\ Z_{(k)} &= Y_{(k)} \\ Z_{(k+1)} &= Y_{(k)} \\ &\vdots \\ Z_{(n)} &= Y_{(k)}. \end{aligned}$$

③ La censure de type III

C'est la version aléatoire du **type I**. C'est ce modèle qui est utilisé pour les essais thérapeutiques, en épidémiologie, en fiabilité, en statistique spatiale. Par exemple en utilisant la catégorie **censure aléatoire** on observe

$$Z_i = Y_i \wedge C_i.$$

1.3.3 Estimation de la fonction de survie

Dans la littérature plusieurs auteurs se sont intéressés pour l'estimation de la fonction de survie dans le cas où les données sont censurées. Parmi ces derniers nous pouvons citer Kaplan et Meier (1958, [69]) ont proposé un estimateur de la fonction de survie. Cet estimateur a été généralisé par Beran (1981, [15]) dans le cas conditionnel. Quelques propriétés asymptotiques de cet estimateur de Kaplan-Meier généralisé ont été présentées par Gonzalez-Manteiga et Cadarso-Suarez (1994, [62]).

1.3.3.1 Estimateur de Kaplan-Meier

Soit $(Z_i, \delta_i)_{1 \leq i \leq n}$, l'échantillon réellement observé et soit $(Z_{(i)}, \delta_{(i)})_{1 \leq i \leq n}$ sa statistique d'ordre croissant. L'estimateur de Kaplan-Meier est défini :

$$\begin{aligned} \widehat{S}_n(t) = 1 - \widehat{F}_n(t) &= \prod_{i=1}^n \left(\frac{n-i}{n-i+1} \right)^{\delta_{(i)} \mathbb{1}_{\{Z_{(i)} \leq t\}}} \\ &= \prod_{i=1}^n \left[1 - \frac{\delta_{(i)} \mathbb{1}_{\{Z_{(i)} \leq t\}}}{n-i+1} \right]. \end{aligned}$$

Il est aussi appelé « **Produit limite** » car il s'obtient comme la limite d'un produit.

- Cet estimateur de Kaplan-Meier est une fonction étagée avec des sauts seulement aux observations non-censurées.
- La hauteur des sauts de cet estimateur est aléatoire.
- Quand toutes les observations sont non-censurées alors on obtient la fonction de répartition empirique.

L'estimateur de Kaplan-Meier est asymptotiquement gaussien; précisément on a le résultat suivant :

Théorème 1.3.1 (Droesbeke et Saporta (2011, [42]))

Si les fonctions de répartition de la survie et de la censure n'ont aucune discontinuité commune, alors :

$$\sup_{t \geq 1} |\widehat{S}_n(t) - S(t)| \xrightarrow{p.s.} 0$$

et pour tout $t \geq 0$,

$$\sqrt{n} (\widehat{S}_n(t) - S(t)) \xrightarrow{d} W_t$$

où $(W_t)_{t \geq 0}$ est un processus gaussien centré qui vérifie pour tous t et s strictement positifs

$$Cov(W_s, W_t) = S(t)S(s) \int_0^{t \wedge s} \frac{dF(u)}{(1-F(u))^2(1-G(u))}.$$

On peut trouver dans Gill (1980, [56]) une démonstration de la normalité asymptotique de \widehat{S}_n , fondée sur la théorie des processus ponctuels.

Cet estimateur de Kaplan-Meier a été généralisé dans le cas conditionnel.

1.3.3.2 Estimateur de Kaplan-Meier généralisé

Beran (1981, [15]) a ajouté un aspect local à l'estimateur de Kaplan-Meier en utilisant du lissage avec les poids de Nadaraya-Watson. Ainsi, il étudiait des problèmes

de régression avec des données censurées dans un contexte complètement non paramétrique. L'estimateur ainsi proposé se définit comme suit :

$$1 - \widehat{F}_n(z|x) = \widehat{S}_{GKM}(z|x) = \begin{cases} \prod_{i=1}^n \left\{ 1 - \frac{B_i(x)}{\sum_{j=1}^n \mathbb{1}_{\{Z_j \geq z_i\}} B_j(x)} \right\}^{\mathbb{1}_{\{Z_{(i)} \leq z, \delta_{(i)}=1\}}} & \text{si } z < Z_{(n)}, \\ 0 & \text{sinon} \end{cases}$$

où

$$B_i(x) = \frac{K\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)}$$

représentent les poids de Nadaraya-Watson, $h_n \rightarrow 0$, la fenêtre et K , le noyau. Cet estimateur est appelé estimateur de Kaplan-Meier conditionnel ou estimateur de Kaplan-Meier généralisé. Notons que si les poids $B_i(x)$ tous égaux à $1/n$, ce qui correspond à ce qu'on ferait sans conditionner sur la covariable X , on obtient en ce moment l'estimateur de Kaplan-Meier standard. Notons aussi que cet estimateur en absence de censure correspondrait à l'estimateur de la fonction de répartition empirique, étudiée par Stone (1977, [84]). Depuis son point de départ par Beran (1981, [15]), cet estimateur a été étudié par un grand nombre d'auteurs. Parmi ces auteurs nous avons d'abord, Dabrowska (1987, [24]), (1988, [25]), (1989, [26]), (1992a,b, [27], [28]), McKeague et Utikal (1990, [75]) et Akritas (1994, [1]) pour le cas où la covariable est aléatoire. Puis dans le cas d'un design fixe, Gonzalez-Manteiga et Cadarso-Suarez (1994, [62]). Puis récemment Van Keilegom et Veraverbeke (1997, [90]), (1998, [91]), (2001, [92]) ont beaucoup travaillé sur les propriétés asymptotiques de cet estimateur et son quantile correspondant.

Dans la suite nous allons considérer l'espace métrique (\mathcal{X}, d) . Soit $H^u(z|x) = H^1(z|x) = \mathbb{P}(Z \leq z, \delta = 1 | X = x)$ ¹ la sous-distribution conditionnelle des observations non-censurées. Ainsi, cet estimateur de Kaplan-Meier conditionnel ou généralisé a une structure compliquée car étant le produit d'éléments dépendants. Comme ont fait Lo et Singh (1986, [73]) pour l'estimateur de Kaplan-Meier standard, Dabrowska a décomposé cet estimateur en une somme pondérée de termes indépendants et un terme de reste d'ordre négligeable facile à manipuler. Pour les propriétés asymptotiques de cet estimateur nous allons supposer quelques hypothèses comme dans Van Keilegom et Veraverbeke (1998, [91]) et dans Gonzalez-Manteiga et Cadarso-Suarez (1994, [62]) :

1. « u » comme "uncensored" en anglais

- **(c1)** Les fonctions $H(\cdot|x_i)$ et $H^1(\cdot|x_i)$, pour $1 \leq i \leq n$, sont dans les familles $H^1(\cdot|x)_{x \in \mathcal{X}}$ et $H^1(\cdot|x)_{x \in \mathcal{X}}$, continues et dérivables par rapport à la première variable et deux fois dérivables par rapport à la seconde variable et que les dérivées sont continues.
- **(c2)** Le noyau K est une fonction de densité symétrique lipschitzienne d'ordre 1 avec support borné vérifiant :

$$\int K^2(u)du < \infty \text{ et } \int |x|^2 K(x)dx < \infty.$$
- **(c3)** $\max_i |s_i - s_{i-1}| \approx \max_i |x_i - x_{i-1}| = O(1/n)$

Théorème 1.3.2 (Dreesbeke et Saporta (2011, [42]))

Supposons les conditions **(c1)**, **(c2)** et **(c3)**, si $\frac{\log n}{nh_n} \rightarrow 0$, $\frac{nh_n^5}{\log n} = O(1)$ alors, pour tout $z < \tau(x)$ avec $\inf_{x \in \mathcal{X}} (1 - H(\tau(x)|x)) > 0$,

$$\begin{aligned} \hat{F}_n(z|x) - F(z|x) &= \sum_{i=1}^n B_i(x) \xi(Z_i, \delta_i, z|x) + r_n(x, z) \\ &= (nh_n)^{-1} f_X(x)^{-1} \sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right) \xi(Z_i, \delta_i, z|x) + r_n(x, z) \end{aligned}$$

où

$$\begin{aligned} \xi(Z_i, \delta_i, z|x) &= (1 - F(z|x)) \times \left[\int_0^z \frac{\mathbb{1}_{\{Z_i \leq s\}} - H(s|x)}{(1 - H(s|x))^2} dH^u(s|x) \right. \\ &\quad \left. + \frac{\mathbb{1}_{\{Z_i \leq z, \delta_i=1\}} - H^u(z|x)}{1 - H(z|x)} \right. \\ &\quad \left. - \int_0^z \frac{\mathbb{1}_{\{Z_i \leq s, \delta_i=1\}} - H^u(s|x)}{(1 - H(s|x))^2} dH^u(s|x) \right] \end{aligned}$$

et

$$\sup_{0 \leq z \leq (1 - H(\tau(x)|x))} |r_n(x, z)| = O\left((nh_n)^{-3/4} (\log n)^{3/4}\right) \quad p.s.$$

et f_X est la densité de la variable X .

Grâce à ce résultat, nous pouvons calculer la covariance et le biais asymptotique de l'estimateur de Kaplan-Meier conditionnel :

$$AsCov\left(\hat{F}_n(t|x), \hat{F}_n(s|x)\right) = (nh_n)^{-1} \Gamma(s, t|x)$$

et

$$AsB\left(\hat{F}_n(t|x)\right) = h_n^2 b(t|x)$$

où

$$\Gamma(s, t|x) = \int K(u)du (1 - F(s|x)) (1 - F(t|x)) \int_0^{s \wedge t} \frac{dH^u(y|x)}{(1 - H(y|x))^2}$$

et

$$b(t|x) = \frac{1}{2} \int u^2 K(u)du (1 - F(t|x)) \times \left[\int_0^t \left\{ \frac{\ddot{H}(s|x) dH^u(s|x)}{(1 - H(s|x))^2} + \frac{d\ddot{H}^u(s|x)}{1 - H(s|x)} \right\} \right. \\ \left. + 2f'_X(x)f_X^{-1}(x) \int_0^t \left\{ \frac{\dot{H}(s|x) dH^u(s|x)}{(1 - H(s|x))^2} + \frac{d\dot{H}^u(s|x)}{1 - H(s|x)} \right\} \right]$$

où $\dot{H}(s|x)$ et $\ddot{H}(s|x)$ représentent les dérivées d'ordre un et deux respectivement de $H(s|x)$ par rapport à x . Sous ces résultats nous obtenons les formules sur le paramètre h_n qui minimise l'erreur quadratique moyenne asymptotique (en anglais "asymptotic mean squared error" : AsMSE) définie par :

$$\begin{aligned} AsMSE(h_n) &= \mathbb{E}_\infty \left[\left(\widehat{F}_n(t|x) - F(t|x) \right)^2 \right] \\ &= AsVar \left(\widehat{F}_n(t|x) \right) + \left(AsB \left(\widehat{F}_n(t|x) \right) \right)^2 \\ &= (nh_n)^{-1} \Gamma(t, t|x) + h_n^4 b^2(t|x). \end{aligned}$$

où \mathbb{E}_∞ est l'espérance mathématique suivant la distribution asymptotique. La valeur de h_n qui minimise cette fonction, $AsMSE(h_n)$, est donnée par :

$$h_{n_{opt}} = \left(\frac{\Gamma(t, t|x)}{4b^2(t|x)} \right)^{1/5} n^{-1/5}.$$

Sous ces conditions nous avons la propriété asymptotique de l'estimateur de Beran définit dans ce théorème.

Théorème 1.3.3 (Droesbeke et Saporta (2011, [42]))

Supposons que les conditions **(c1)**-**(c3)** sont réalisées.

- Si $nh_n^5 \rightarrow 0$ et $\frac{(\log n)^3}{nh_n} \rightarrow 0$, alors, pour $n \rightarrow \infty$

$$(nh_n)^{1/2} \left(\widehat{F}_n(\cdot|x) - F(\cdot|x) \right) \rightarrow W(\cdot|x)$$

- Si $h_n = Cn^{-1/5}$ pour un certain $C > 0$, alors, pour $n \rightarrow \infty$

$$(nh_n)^{1/2} \left(\widehat{F}_n(\cdot|x) - F(\cdot|x) \right) \rightarrow \widetilde{W}(\cdot|x)$$

où $W(\cdot|x)$ et $\widetilde{W}(\cdot|x)$ sont des processus gaussiens avec comme fonction de covariance, $\Gamma(\cdot, \cdot|x)$ et pour $\widetilde{W}(\cdot|x)$, fonction de moyenne, $b(\cdot|x)C^{5/2}$.

Nous allons maintenant rappeler un résultat fondamental qui sera utile pour la suite, les quantiles conditionnels obtenus avec cet estimateur. Pour obtenir la propriété asymptotique des quantiles conditionnels d'ordre α ($0 < \alpha < 1$), nous allons rappeler cette relation énoncée sous forme de théorème.

Théorème 1.3.4 (Droesbeke et Saporta (2011, [42]))

Supposons les conditions **(c1)**-**(c3)**, si $\frac{\log n}{nh_n} \rightarrow 0$, $\frac{nh_n^5}{\log n} = O(1)$. Soit $0 < \xi_0 < \alpha_0 < 1$ telle que $F^{-1}(\alpha_0|x) < \inf_{x \in \mathcal{X}} (1 - H(\tau(x)|x))$ et $\inf_{\xi_0 \leq \alpha \leq \alpha_0} f(F^{-1}(\alpha|x)|x) = \lambda > 0$:

$$\begin{aligned} \widehat{F}_n^{-1}(\alpha|x) - F^{-1}(\alpha|x) &= - \left\{ \widehat{F}_n(F^{-1}(\alpha|x) - \alpha) \right\} / f(F^{-1}(\alpha|x)) + r_{n1}(\alpha, x) \\ &= (nh_n)^{-1} f_X(x)^{-1} \sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right) \\ &\quad \xi(Z_i, \delta_i, F^{-1}(\alpha|x)|x) / f(F^{-1}(\alpha|x)) + r_{n2}(\alpha, x) \end{aligned}$$

où la fonction ξ a été définie dans les théorèmes précédents et

$$\sup_{\xi_0 \leq \alpha \leq \alpha_0} |r_{n1}(\alpha, x)| = O\left((nh_n)^{-3/4} (\log n)^{3/4}\right) \quad p.s.$$

$$\sup_{\xi_0 \leq \alpha \leq \alpha_0} |r_{n2}(\alpha, x)| = O\left((nh_n)^{-3/4} (\log n)^{3/4}\right) \quad p.s.$$

si $n \rightarrow \infty$.

Cette décomposition nous permet de déterminer la propriété asymptotique des quantiles conditionnels.

Théorème 1.3.5 (Droesbeke et Saporta (2011, [42]))

Supposons les conditions **(c1)**, **(c2)** et **(c3)**. Soit $0 < \alpha < 1$ telle que $F^{-1}(\alpha|x) < \inf_{x \in \mathcal{X}} (1 - H(\tau(x)|x))$ et $f(F^{-1}(\alpha|x)|x) > 0$

- Si $nh_n^5 \rightarrow 0$ et $\frac{(\log(n))^3}{\log nh_n} \rightarrow 0$, alors, pour $n \rightarrow \infty$

$$(nh_n)^{1/2} \left(\widehat{F}_n^{-1}(\alpha|x) - F^{-1}(\alpha|x) \right) \xrightarrow{d} \mathcal{N} \left(0, \sigma^2(\alpha|x) \right)$$

- Si $h_n = Cn^{-1/5}$ pour un certain $C > 0$, alors, pour $n \rightarrow \infty$

$$(nh_n)^{1/2} \left(\widehat{F}_n^{-1}(\alpha|x) - F^{-1}(\alpha|x) \right) \xrightarrow{d} \mathcal{N} \left(\beta(\alpha|x), \sigma^2(\alpha|x) \right)$$

où

$$\sigma^2(\alpha|x) = \frac{\Gamma(F^{-1}(\alpha|x), F^{-1}(\alpha|x))}{f^2(F^{-1}(\alpha|x)|x)}$$

et

$$\beta(\alpha|x) = -\frac{b(F^{-1}(\alpha|x)|x)}{f(F^{-1}(\alpha|x)|x)}.$$

Pour plus de détails sur les trois derniers théorèmes, le lecteur est invité à consulter les papiers de Dabrowska ([24], [25], [26], [27], [28]).

1.3.4 Représentations graphiques en présence de données censurées

a) Avec l'estimateur de Kaplan-Meier

Le « Pareto quantile plot » peut être déterminé quand le cas où les données censurées aléatoirement à droite. De ce fait, il suffit de remplacer la fonction de répartition empirique par son équivalent qui est l'estimateur de Kaplan-Meier. Il devient alors la représentation des points :

$$\left(-\log \left(1 - \widehat{F}_n \left(Z_{(n-i+1)} \right) \right), \log Z_{(n-i+1)} \right), i = 1, \dots, n-1.$$

b) Avec l'estimateur de Kaplan-Meier généralisé

Ce « Pareto quantile plot » peut aussi être aussi tracé en utilisant des données censurées aléatoirement à droite avec une covariable x :

$$\left(-\log \left(1 - \widehat{F}_n \left(Z_{(n-i+1)} | x \right) \right), \log Z_{(n-i+1)} \right), i = 1, \dots, n-1$$

où \widehat{F}_n est l'estimateur de Kaplan-Meier conditionnel.

Estimation des quantiles extrêmes conditionnels en design fixe en présence de censure

Résumé

Dans ce chapitre, nous proposons dans le cas des fonctions à queues lourdes des estimateurs de l'indice et de quantiles extrêmes conditionnels à design fixe avec des données censurées aléatoirement à droite. Puis, nous étudions leurs propriétés asymptotiques. Enfin, nous illustrons leur comportement par des simulations et application sur des données réelles.

Sommaire

2.1 Introduction	36
2.2 Modèle et notations	36
2.3 Définitions des estimateurs	37
2.4 Propriétés asymptotiques	41
2.4.1 Propriété asymptotique de l'estimateur de l'indice	43
2.4.2 Propriété asymptotique de l'estimateur des quantiles	45
2.5 Simulation et illustration sur des données réelles	47
2.5.1 Expériences numériques par simulation	47
2.5.2 Exemple sur des données réelles	60
2.6 Discussion et perspectives	66
2.6.1 Taux de censure sur les estimations	67
2.6.2 Effet de l'estimateur standard	67
2.6.3 Propriétés asymptotiques	67
2.6.4 Perspectives	68
2.7 Preuves des théorèmes	68

2.1 Introduction

DANS ce chapitre, nous considérons le problème de l'estimation de l'indice des valeurs extrêmes conditionnels et des quantiles extrêmes conditionnels en présence de données censurées aléatoirement à droite. Les domaines privilégiés dans ce chapitre concernent ceux dont les fonctions de répartition sont à queues lourdes telles que l'hydrologie avec les inondations, la biologie avec la présence de données manquantes dans les bases de données, la finance avec les cracks boursiers, la télécommunication avec les trafics routiers, etc. Dans la littérature, il existe plusieurs méthodes d'estimation de quantiles extrêmes conditionnels. Il y a des méthodes basées sur les modèles paramétriques proposées par Smith et Davison (1989, [83]) et Smith (1990, [33]). L'approche non-paramétrique des quantiles extrêmes conditionnels a été proposée par Davison et Ramesh (2000, [34]) où ils présentent des estimateurs par ajustement polynomial. Par contre, Beirlant et Goegebeur (2003, [12]) ont présenté l'approche semi-paramétrique pour estimer des quantiles extrêmes conditionnels.

Dans le cas censuré sans covariable, Einmahl et al. (2008, [44]) ont proposé des estimateurs de l'indice des valeurs extrêmes et de quantiles extrêmes. Dans la même année Gardes et Girard (2008, [51]) ont quant eux présenté des estimateurs de quantiles extrêmes conditionnels avec l'utilisation de la méthode de la fenêtre mobile. Dans ce travail, nous proposons des adaptations de ces estimateurs dans le cas censuré aléatoirement à droite et en présence de covariable en design fixe. Puis une étude théorique (propriétés asymptotiques) de ces nouveaux estimateurs comme dans Lékina (2010, [72]) sera proposée. Enfin, leur comportement est illustré sur la base de simulations et sur des données médicales du S.I.D.A. (Syndrome d'Immuno Déficience Acquise).

2.2 Modèle et notations

Nous travaillons dans l'espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ et dans l'espace métrique (\mathcal{X}, d) . Soient $(x_i)_{1 \leq i \leq n}$, un échantillon déterministe de X , $(Y_i)_{1 \leq i \leq n}$ et $(C_i)_{1 \leq i \leq n}$ des réalisations respectives des variables aléatoires Y et C dans l'espace $(\Omega, \mathcal{A}, \mathbb{P})$.

Nous supposons que Y et C ont respectivement comme fonctions de répartition conditionnelles F et G . Soient $(Z_i)_{1 \leq i \leq n} = (Y_i \wedge C_i)_{1 \leq i \leq n}$ et $\delta_i = \mathbb{1}_{\{Y_i \leq C_i\}}$, où $\mathbb{1}_{\{\cdot\}}$ est une fonction indicatrice. Nous supposons aussi que Y et C sont indépendantes conditionnellement à X . Ainsi, dire que nous sommes en présence de données censurées aléatoirement à droite revient à dire que l'échantillon $(Y_i)_{1 \leq i \leq n}$ n'est pas observé mais est cen-

suré par $(C_i)_{1 \leq i \leq n}$. Nous observons, de ce fait, uniquement $(Z_1, \delta_1, x_1), \dots, (Z_n, \delta_n, x_n)$ des copies indépendantes de (Z, δ, X) . Dans la suite nous supposons que la variable Z a comme fonction de répartition conditionnelle H .

Dans tout ce chapitre, nous considérons que les fonctions de répartition conditionnelles de Y et C pour tout $x \in \mathcal{X}$ sont ainsi définies :

$$F(u|x) = 1 - u^{-1/\gamma_1(x)} L_1(u, x),$$

$$G(u|x) = 1 - u^{-1/\gamma_2(x)} L_2(u, x),$$

où $\gamma_1(\cdot)$ et $\gamma_2(\cdot)$ sont des paramètres fonctionnels positifs de la covariable x fixé appelés indices de queue conditionnels ou indices des valeurs extrêmes conditionnels, $L_1(\cdot, x)$ et $L_2(\cdot, x)$ sont des fonctions à variations lentes, i.e. pour tout $\lambda > 0$,

$$\lim_{u \rightarrow \infty} \frac{L_i(\lambda u, x)}{L_i(u, x)} = 1, i = 1, 2.$$

Si Y et C sont indépendantes, alors la fonction de répartition conditionnelle $H(\cdot|x)$ de Z pour $X = x$ fixé, est une fonction à queue lourde avec comme indice de queue conditionnel $\gamma(x) = \gamma_1(x)\gamma_2(x)/(\gamma_1(x) + \gamma_2(x))$. Ce qui donne pour tout u et x ,

$$\begin{aligned} 1 - H(u|x) &= (1 - F(u|x))(1 - G(u|x)) \\ &= u^{-1/\gamma_1(x)} L_1(u, x) u^{-1/\gamma_2(x)} L_2(u, x) \\ &= u^{-1/\gamma(x)} L(u, x), \end{aligned}$$

où $\gamma(x)$ est défini au dessus et $L(u, x) = L_1(u, x)L_2(u, x)$. Alors,

$$\lim_{u \rightarrow \infty} \frac{L(\lambda u, x)}{L(u, x)} = \lim_{u \rightarrow \infty} \frac{L_1(\lambda u, x)}{L_1(u, x)} \frac{L_2(\lambda u, x)}{L_2(u, x)} = 1.$$

Finalement,

$$H(u|x) = 1 - u^{-1/\gamma(x)} L(u, x)$$

donc H est une fonction de répartition conditionnelle appartenant au domaine d'attraction de Fréchet de paramètre fonctionnel, $\gamma(\cdot)$ (cf chap 1). Dans ces conditions nous allons proposer des estimateurs.

2.3 Définitions des estimateurs

Nous allons considérer $(Z_1, \delta_1, x_1), \dots, (Z_n, \delta_n, x_n)$, l'échantillon observé. L'objectif de cette partie est, pour un $x \in \mathcal{X}$ fixé, déterminer un estimateur de $\gamma_1(x)$ et son quan-

tile extrême correspondant. Pour cela nous n'allons pas utiliser toutes les données mais simplement les observations dont leurs covariables sont proches de x . Ainsi, l'approche que nous allons utiliser est la méthode de la **fenêtre mobile** comme dans Gardes et Girard (2008, [51]), définie par une boule, $B(x, h_{n,x})$ où x est le centre de la boule et $h_{n,x}$ est le rayon de la boule :

$$B(x, h_{n,x}) = \{t \in \mathcal{X}, d(x, t) \leq h_{n,x}\}.$$

Cette méthode renferme trois étapes :

- Etape 1 : On fixe une covariable x déterministe, centre de la boule,
- Etape 2 : Puis on fixe le rayon, $h_{n,x}$,
- Etape 3 : Enfin on sélectionne uniquement les observations (Z_i, δ_i) pour lesquelles $x_i \in B(x, h_{n,x})$.

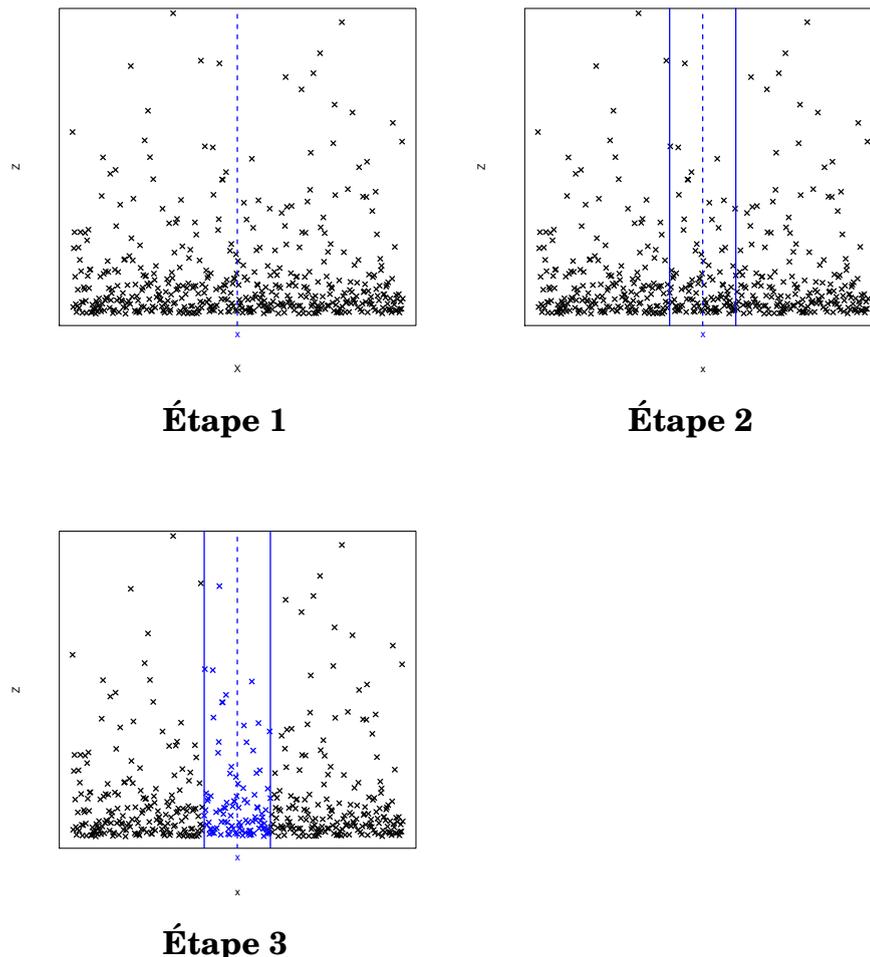


FIGURE 2.1 – Les différentes étapes de sélection des données. En ordonnée on a la variable Z et en abscisse la covariable X .

$m_{n,x} = n\phi(h_{n,x})$ est le nombre de données qui tombent dans $(0, \infty) \times B(x, h_{n,x})$, où

$$\phi(h_{n,x}) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{x_i \in B(x, h_{n,x})\}}.$$

Notons $\{Z_i^x, \delta_i^x, i = 1, \dots, m_{n,x}\}$ les observations retenues lors de la sélection c'est-à-dire

$$(Z_i^x, \delta_i^x) = \left\{ (Z_i, \delta_i) \mathbb{1}_{\{|x_i - x| \leq h_{n,x}\}} \right\}, i = 1, \dots, n; \quad h_{n,x} \rightarrow 0 \text{ quand } n \rightarrow \infty.$$

Puis nous notons aussi

$$\left(Z_{(1)}^x, \delta_{(1)}^x \right), \dots, \left(Z_{(m_{n,x})}^x, \delta_{(m_{n,x})}^x \right)$$

l'échantillon ordonné associé.

Dans le cas non censuré la fonction $\gamma(\cdot)$ peut être estimée par l'estimateur de Hill adapté comme dans Gardes et Girard (2008, [51]) :

$$\hat{\gamma}_{k_x, m_{n,x}}^{(H)}(x) = \frac{1}{k_x} \sum_{i=1}^{k_x} i \log \left(\frac{Z_{(m_{n,x}-i+1)}^x}{Z_{(m_{n,x}-i)}^x} \right).$$

Pour adapter cet estimateur dans le cas censuré nous allons diviser cet estimateur par la proportion de données non censurées des k_x plus grandes valeurs de Z^x pour obtenir :

$$\hat{\gamma}_{k_x, m_{n,x}}^{(c,H)}(x) = \frac{\hat{\gamma}_{k_x, m_{n,x}}^{(H)}(x)}{\hat{p}_x}$$

où $\hat{p}_x = \frac{1}{k_x} \sum_{i=1}^{k_x} \delta_{(m_{n,x}-i+1)}^x$ estime $p_x = \frac{\gamma_2(x)}{\gamma_1(x) + \gamma_2(x)}$, avec $\delta_{(1)}^x, \dots, \delta_{(m_{n,x})}^x$ les indicateurs de censure retenues respectivement avec l'échantillon $Z_{(1)}^x, \dots, Z_{(m_{n,x})}^x$, lors de la sélection.

Dans le cas général on obtient,

$$\hat{\gamma}_{k_x, m_{n,x}}^{(c,\cdot)}(x) = \frac{\hat{\gamma}_{k_x, m_{n,x}}^{(\cdot)}(x)}{\hat{p}_x} \tag{2.1}$$

où $\hat{\gamma}_{k_x, m_{n,x}}^{(\cdot)}(x)$ peut être :

☞ **Estimateur de Hill adapté (1975, [67])**

$$\hat{\gamma}_{k_x, m_{n,x}}^{(H)}(x) = M_{k_x, m_{n,x}}^{(1)} = \frac{1}{k_x} \sum_{i=1}^{k_x} i \log \left(\frac{Z_{(m_{n,x}-i+1)}^x}{Z_{(m_{n,x}-i)}^x} \right)$$

☞ **Estimateur de Dekkers-Einmahl-de Haan adapté (1989, [40])**

$$\hat{\gamma}_{k_x, m_{n,x}}^{(M)}(x) = M_{k_x, m_{n,x}}^{(1)} + 1 - \frac{1}{2} \left(1 - \frac{(M_{k_x, m_{n,x}}^{(1)})^2}{M_{k_x, m_{n,x}}^{(2)}} \right)^{-1}$$

où

$$M_{k_x, m_{n,x}}^{(2)} = \frac{1}{k_x} \sum_{i=1}^{k_x} \left(i \log \left(\frac{Z_{(m_{n,x}-i+1)}^x}{Z_{(m_{n,x}-i)}^x} \right) \right)^2$$

☞ **Estimateur UH adapté (1996, [13])**

$$\hat{\gamma}_{k_x, m_{n,x}}^{(UH)}(x) = \frac{1}{k_x} \sum_{i=1}^{k_x} \log \left(Z_{(m_{n,x}-i)}^x \hat{\gamma}_{i, m_{n,x}}^{(H)}(x) \right) - \log \left(Z_{(m_{n,x}-k_x)}^x \hat{\gamma}_{k_x, m_{n,x}}^{(H)}(x) \right)$$

☞ **etc.**

Pour plus de détails par rapport à ces estimateurs le lecteur pourra consulter le chapitre 1.

L'objectif essentiel de la théorie des valeurs extrêmes est d'estimer les quantiles extrêmes. Nous allons proposer un estimateur de quantiles extrêmes conditionnels en essayant de résoudre cette équation :

$$\mathbb{P}(Y > q(\alpha_{m_{n,x}}, x) | X = x) = \alpha_{m_{n,x}}$$

où l'ordre du quantile $\alpha_{m_{n,x}} \rightarrow 0$ quand $n \rightarrow +\infty$ voir **Lékina (2010, [72])**.

Pour résoudre ce problème avec l'hypothèse que F appartient au domaine d'attraction de Fréchet de paramètre fonctionnel $\gamma_1(\cdot)$. Or, d'après ce qui précède $F \in \text{D.A.}(\text{Fréchet})$ implique

$$\bar{F}(u|x) = u^{-1/\gamma_1(x)} L_1(u, x)$$

donc

$$\bar{F}\left(Z_{(m_{n,x}-k_x)}^x | x\right) = Z_{(m_{n,x}-k_x)}^x{}^{-1/\gamma_1(x)} L_1\left(Z_{(m_{n,x}-k_x)}^x, x\right).$$

Ainsi, on a

$$\frac{\bar{F}(u|x)}{\bar{F}\left(Z_{(m_{n,x}-k_x)}^x | x\right)} = \left(\frac{u}{Z_{(m_{n,x}-k_x)}^x} \right)^{-1/\gamma_1(x)} \frac{L_1(u, x)}{L_1\left(Z_{(m_{n,x}-k_x)}^x, x\right)}.$$

Pour k_x suffisamment petit et $u \geq Z_{(m_{n,x}-k_x)}^x$, on peut négliger, au premier ordre, le rapport des fonctions à variations lentes; il vient alors

$$\bar{F}(u|x) \approx \bar{F}(Z_{(m_{n,x}-k_x)}^x|x) \left(\frac{u}{Z_{(m_{n,x}-k_x)}^x} \right)^{-1/\gamma_1(x)}.$$

En inversant cette fonction à l'ordre $\alpha_{m_{n,x}}$, on obtient :

$$q(\alpha_{m_{n,x}}, x) = Z_{(m_{n,x}-k_x)}^x \left(\frac{\bar{F}(Z_{(m_{n,x}-k_x)}^x|x)}{\alpha_{m_{n,x}}} \right)^{\gamma_1(x)}. \quad (2.2)$$

Dans cette dernière équation (2.2) les fonctions $F(\cdot)$ et $\gamma_1(\cdot)$ sont inconnues. Or, nous sommes en présence de données censurées pour proposer un estimateur de quantile extrême nous allons remplacer $F(\cdot)$ par son estimateur de Kaplan-Meier (1958, [69]) :

$$1 - \hat{F}_{m_{n,x}}(t|x) = \prod_{i=1}^{m_{n,x}} \left(\frac{m_{n,x} - i}{m_{n,x} - i + 1} \right)^{\delta_{(i)}^x \mathbb{1}_{\{Z_{(i)}^x \leq t\}}}$$

et $\gamma_1(\cdot)$ par son estimateur défini au niveau de l'équation (2.1). Ce qui donne l'estimateur de Weissman adapté à notre cas :

$$\hat{q}^{(c,\cdot)}(\alpha_{m_{n,x}}, x) = Z_{(m_{n,x}-k_x)}^x \left(\frac{1 - \hat{F}_{m_{n,x}}(Z_{(m_{n,x}-k_x)}^x|x)}{\alpha_{m_{n,x}}} \right)^{\hat{\gamma}_{k_x, m_{n,x}}^{(c,\cdot)}(x)}. \quad (2.3)$$

Par exemple, si l'estimateur standard pris est celui de Hill, on obtient

$$\hat{q}^{(c,H)}(\alpha_{m_{n,x}}, x) = Z_{(m_{n,x}-k_x)}^x \left(\frac{1 - \hat{F}_{m_{n,x}}(Z_{(m_{n,x}-k_x)}^x|x)}{\alpha_{m_{n,x}}} \right)^{\hat{\gamma}_{k_x, m_{n,x}}^{(c,H)}(x)}.$$

2.4 Propriétés asymptotiques

Dans ce paragraphe, nous allons d'abord définir quelques concepts liés aux propriétés asymptotiques en théorie des valeurs extrêmes. Parmi ces concepts nous avons les fonctions liées à la propriété asymptotique. Tout d'abord nous avons la fonction quantile conditionnel ou fonction de queue conditionnelle définie par

$$U(y|x) = H^{\leftarrow} \left(1 - \frac{1}{y} | x \right) = \inf \left\{ u : H(u|x) \geq 1 - \frac{1}{y} \right\}.$$

Ainsi, en théorie des valeurs extrêmes cette fonction quantile est très importante car elle contrôle le biais des estimateurs via la condition du second ordre. S'il existe

une fonction a positive et une fonction auxiliaire a_2 positive telle $a_2(y|x) \rightarrow 0$ quand $y \rightarrow \infty$, alors elle vérifie

$$\lim_{y \rightarrow \infty} \frac{1}{a_2(y|x)} \left\{ \frac{U(uy|x) - U(y|x)}{a(y|x)} - h_{\gamma(x)}(u, x) \right\} = k(u, x)$$

où k est définie sur $]0, +\infty[$ et s'il existe une constante c et une fonction $\rho < 0$ alors,

$$\lim_{y \rightarrow \infty} \frac{1}{a_2(y|x)} \left\{ \frac{a(uy|x)}{a(y|x)} - u^{\gamma(x)} \right\} = cu^{\gamma(x)} h_{\rho(x)}(u, x)$$

avec $h_{\rho(x)}(y, x) = \int_1^y z^{\rho(x)-1} dz$.

La fonction k admet cette représentation

$$k(u, x) = c \int_1^u t^{\gamma(x)-1} h_{\rho(x)}(t, x) dt + Ah_{\gamma(x)+\rho(x)}(u, x),$$

où $A \in \mathbb{R}$. Si $0 < \gamma(x) < -\rho(x)$ alors nous avons une autre représentation de la fonction quantile

$$U(y, x) = l_+ y^{\gamma(x)} \left\{ \frac{1}{\gamma(x)} + Dy^{-\gamma(x)} + \frac{A}{\gamma(x) + \rho(x)} a_2(y|x) (1 + o(1)) \right\} \quad (2.4)$$

où l_+ est une constante positive.

Pour plus de détails le lecteur pourra consulter [44].

Sous toutes ces conditions la fonction biais sera définie ainsi,

$$b(y, x) = \begin{cases} A\rho(x)(\rho(x) + \gamma(x)(1 - \rho(x)))a_2(y|x) & \text{si } 0 < -\rho(x) < \gamma(x) \text{ ou} \\ & 0 < \gamma(x) < -\rho(x), \text{ avec } D = 0 \\ -\frac{\gamma(x)^3}{(1+\gamma(x))} x^{-\gamma(x)} L_2(y, x) & \text{si } \gamma(x) = -\rho(x) \\ -\frac{\gamma(x)^3 D}{(1+\gamma(x))} x^{-\gamma(x)} & \text{si } 0 < \gamma(x) < -\rho(x) \text{ avec } D \neq 0 \end{cases} \quad (2.5)$$

et

$$\tilde{\rho}(x) = \begin{cases} \rho(x) & \text{si } 0 < -\rho(x) \leq \gamma(x) \text{ ou } 0 < \gamma(x) < -\rho(x) \text{ avec } D = 0 \\ -\gamma(x) & \text{si } 0 < \gamma(x) < -\rho \text{ avec } D \neq 0. \end{cases} \quad (2.6)$$

L'ensemble de ces définitions nous a permis de déterminer les propriétés asymptotiques des estimateurs définis précédemment.

2.4.1 Propriété asymptotique de l'estimateur de l'indice

Pour déterminer les propriétés asymptotiques de l'estimateur de l'indice des valeurs extrêmes nous avons besoin de la fonction suivante comme définie dans [44] :

$$p_x(z) = \mathbb{P}(\delta = 1 | Z = z, x)$$

la sous-distribution conditionnelle des observations non-censurées. Nous pouvons l'écrire d'une autre manière

$$p_x(z) = \frac{(1 - G(z|x))f(z|x)}{(1 - G(z|x))f(z|x) + (1 - F(z|x))g(z|x)},$$

où f et g désignent respectivement les densités conditionnelles de Y et C , alors,

$$\lim_{z \rightarrow \infty} p_x(z) = \frac{\gamma_2(x)}{\gamma_1(x) + \gamma_2(x)} = p_x.$$

Nous proposons quelques exemples avec des lois à « queues lourdes ».

Exemple 2 Supposons Y et C sont respectivement de $\text{Pareto}(\gamma_1(\cdot))$ et $\text{Pareto}(\gamma_2(\cdot))$, c'est-à-dire pour tout $z \geq 1$, $F(z|x) = 1 - z^{-1/\gamma_1(x)}$ et $G(z|x) = 1 - z^{-1/\gamma_2(x)}$ où $\gamma_1(\cdot)$ et $\gamma_2(\cdot)$ sont des paramètres fonctionnels positifs. On obtient ,

$$H(z|x) = 1 - z^{-\frac{\gamma_1(x) + \gamma_2(x)}{\gamma_1(x)\gamma_2(x)}}$$

ce qui implique Z suit une $\text{Pareto}(\gamma_1(x)\gamma_2(x)/(\gamma_1(x) + \gamma_2(x)))$.

Nous pouvons à présent calculer la fonction $p_x(\cdot)$,

$$\begin{aligned} p_x(z) &= \frac{(1 - G(z|x))f(z|x)}{(1 - G(z|x))f(z|x) + (1 - F(z|x))g(z|x)} \\ &= \frac{z^{-1/\gamma_2(x)} \frac{1}{\gamma_1(x)} z^{-1/\gamma_1(x)}}{z^{-1/\gamma_2(x)} \frac{1}{\gamma_1(x)} z^{-1/\gamma_1(x)} + z^{-1/\gamma_1(x)} \frac{1}{\gamma_2(x)} z^{-1/\gamma_2(x)}} \\ &= \frac{\frac{1}{\gamma_1(x)} z^{-1/\gamma_1(x) - 1/\gamma_2(x)}}{(\frac{1}{\gamma_1(x)} + \frac{1}{\gamma_2(x)}) z^{-1/\gamma_1(x) - 1/\gamma_2(x)}} \\ &= \frac{\frac{1}{\gamma_1(x)}}{\frac{1}{\gamma_1(x)} + \frac{1}{\gamma_2(x)}} \\ &= \frac{\gamma_2(x)}{\gamma_1(x) + \gamma_2(x)} \end{aligned}$$

Exemple 3 Soient $F(z|x) = 1 - (1 + \gamma_1(x)z)^{-1/\gamma_1(x)}$ et $G(z|x) = 1 - (1 + \gamma_2(x)z)^{-1/\gamma_2(x)}$, pour tout $z \geq 0$, $\gamma_1(\cdot) > 0$, $\gamma_2(\cdot) > 0$.

Nous avons $H(z|x) = 1 - (1 + \gamma_1(x)z)^{-1/\gamma_1(x)}(1 + \gamma_2(x)z)^{-1/\gamma_2(x)}$

$$\begin{aligned}
 p(z|x) &= \frac{(1 - G(z|x))f(z|x)}{(1 - G(z|x))f(z|x) + (1 - F(z|x))g(z|x)} \\
 &= \frac{(1 + \gamma_1(x)z)^{-1/\gamma_1(x)-1}(1 + \gamma_2(x)z)^{-1/\gamma_2(x)}}{(1 + \gamma_1(x)z)^{\frac{-1}{\gamma_1(x)-1}}(1 + \gamma_2(x)z)^{\frac{-1}{\gamma_2(x)}} + (1 + \gamma_1(x)z)^{\frac{-1}{\gamma_1(x)}}(1 + \gamma_2(x)z)^{\frac{-1}{\gamma_2(x)-1}}} \\
 &= \frac{(1 + \gamma_1(x)z)^{-1/\gamma_1(x)}(1 + \gamma_2(x)z)^{-1/\gamma_2(x)}\{(1 + \gamma_1(x)z)^{-1}\}}{(1 + \gamma_1(x)z)^{\frac{-1}{\gamma_1(x)}}(1 + \gamma_2(x)z)^{\frac{-1}{\gamma_2(x)}}\{(1 + \gamma_1(x)z)^{-1} + (1 + \gamma_2(x)z)^{-1}\}} \\
 &= \frac{(1 + \gamma_1(x)z)^{-1}}{(1 + \gamma_1(x)z)^{-1} + (1 + \gamma_2(x)z)^{-1}} \\
 &= \frac{\frac{1}{z} + \gamma_2(x)}{\frac{2}{z} + (\gamma_2(x) + \gamma_1(x))} \longrightarrow \frac{\gamma_2(x)}{\gamma_1(x) + \gamma_2(x)}, \quad \text{quand } z \longrightarrow +\infty.
 \end{aligned}$$

Pour déterminer les propriétés asymptotiques de nos estimateurs nous allons besoin de quelques hypothèses de régularité (ces conditions sont des adaptations au cas conditionnel comme dans [51] et [19]). Pour tout $x \in \mathcal{X}$, nous supposons les assertions suivantes :

- C1.** $F(\cdot|x)$ et $G(\cdot|x)$ sont absolument continues,
- C2.** Il existe une fonction $\rho(x) < 0$ et une fonction à variations régulières $b(\cdot, x)$ d'indice $\rho(x)$ telles que pour tout $u > 0$,

$$\lim_{t \rightarrow \infty} \frac{H^{\leftarrow} \left(1 - \frac{1}{tu} | x\right) / H^{\leftarrow} \left(1 - \frac{1}{t} | x\right) - u^{\gamma(x)}}{b(t, x)} = u^{\gamma(x)} \frac{u^{\rho(x)} - 1}{\rho(x)}.$$

- C3.** Il existe des constantes positives c_U , z_U et $\alpha_U \leq 1$ telles que pour tout $x \in B(x', 1)$

$$\sup_{z \geq z_U} \left| \frac{\log U(z|x)}{\log U(z|x')} - 1 \right| \leq c_U d^{\alpha_U}(x, x')$$

$$\text{où } U(z|x) = \inf \left\{ s; H(s|x) \geq 1 - \frac{1}{z} \right\}.$$

Pour tout $x \in \mathcal{X}$, nous supposons les assertions suivantes (comme dans [44]) si $n \rightarrow \infty$, $k_x \rightarrow \infty$, $k_x/m_{n,x} \rightarrow 0$ et :

- C4.** $\sqrt{k_x} b\left(\frac{m_{n,x}}{k_x}, x\right) \rightarrow \lambda(x) < \infty$,
- C5.** $\frac{1}{\sqrt{k_x}} \sum_{i=1}^{k_x} \left[p_x \left(H^{\leftarrow} \left(1 - \frac{i}{m_{n,x}} | x\right) \right) - p_x \right] \rightarrow \varepsilon(x) < \infty$. Soient $C > 0$ et $\mathcal{A}(s, t) := \{1 - k_x/m_{n,x} \leq t < 1, |t - s| \leq C\sqrt{k_x}/m_{n,x}, s < 1\}$. Nous assumons que si $n \rightarrow \infty$,
- C6.** $\sqrt{k_x} \sup_{\mathcal{A}(s,t)} |p_x(H^{\leftarrow}(t|x)) - p_x(H^{\leftarrow}(s|x))| \rightarrow 0$.

Sous ces conditions, nous avons les résultats asymptotiques de nos estimateurs.

Théorème 2.4.1 Soit $x \in \mathcal{X}$. Sous les conditions **C5-C6** et s'il existe des fonctions $m(\cdot)$ et $\sigma(\cdot)$ telles que $\sqrt{k_x}(\hat{\gamma}_{k_x, m_{n,x}}^{(\cdot)}(x) - \gamma(x)) \xrightarrow{\mathcal{D}} \mathcal{N}(m(x)\lambda(x), \sigma^2(x))$, alors, nous avons :

$$\sqrt{k_x}(\hat{\gamma}_{k_x, m_{n,x}}^{(c,\cdot)}(x) - \gamma_1(x)) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\frac{\lambda(x)m(x) - \gamma_1(x)\varepsilon(x)}{p_x}, \frac{\sigma^2(x) + \gamma_1(x)^2 p_x(1 - p_x)}{p_x^2}\right).$$

À partir de ce théorème 2.4.1, nous en déduisons le corollaire suivant :

Corollaire 2.4.1 Sous les hypothèses **C1-C6** et $k_x^{1/2} h_{n,x}^{\alpha_U} \rightarrow 0$, nous avons,

$$\sqrt{k_x}(\hat{\gamma}_{k_x, m_{n,x}}^{(c,H)}(x) - \gamma_1(x)) \xrightarrow{d} \mathcal{N}\left(\frac{-\gamma_1(x)\varepsilon(x)}{p_x} + \frac{\lambda(x)}{p_x(1 - \rho(x))}, \frac{\gamma_1^3(x)}{\gamma(x)}\right)$$

$$\sqrt{k_x}(\hat{\gamma}_{k_x, m_{n,x}}^{(c,UH)}(x) - \gamma_1(x)) \xrightarrow{d} \mathcal{N}\left(\frac{-\gamma_1(x)\varepsilon(x)}{p_x} + \frac{\lambda(x)}{p_x(1 - \rho(x))}, \frac{\gamma_1^2(x)}{\gamma^2(x)}(1 + \gamma_1(x)\gamma(x))\right)$$

$$\sqrt{k_x}(\hat{\gamma}_{k_x, m_{n,x}}^{(c,M)}(x) - \gamma_1(x)) \xrightarrow{d} \mathcal{N}\left(\frac{-\gamma_1(x)\varepsilon(x)}{p_x} + \frac{\lambda(x)}{p_x(1 - \rho(x))}, \frac{\gamma_1^2(x)}{\gamma^2(x)}(1 + \gamma_1(x)\gamma(x))\right).$$

Ce corollaire se déduit directement du théorème précédent en notant $m(x) = (1 - \rho(x))^{-1}$ (pour tout $\gamma_{k_x, m_{n,x}}^{(H)}(x)$, $\gamma_{k_x, m_{n,x}}^{(M)}(x)$, $\gamma_{k_x, m_{n,x}}^{(UH)}(x)$) et

$$\sigma^2(x) = \begin{cases} \gamma^2(x) & \text{pour } \gamma_{k_x, m_{n,x}}^{(H)}(x) \\ 1 + \gamma^2(x) & \text{pour } \gamma_{k_x, m_{n,x}}^{(M)}(x) \text{ et } \gamma_{k_x, m_{n,x}}^{(UH)}(x) \end{cases}$$

pour plus de détails le lecteur pourra lire (2005, [6]) et (1989, [40]). Notons que ces trois estimateurs ont le même biais ce qui est cohérent d'après les résultats de Einmahl et al. (2008, [44]) (nous considérons seulement le cas où $\gamma_1(\cdot)$ et $\gamma_2(\cdot)$ sont des fonctions positives).

2.4.2 Propriété asymptotique de l'estimateur des quantiles

Dans ce paragraphe nous allons déterminer les propriétés asymptotiques des quantiles extrêmes conditionnels. Or, pour estimer ce quantile au point x nous n'avons pas utilisé toutes les observations mais seulement celles dont leurs cova-

riables sont proches de x . Cette méthode de sélection va engendrer un biais au niveau de la seconde variable du quantile. Nous allons aussi contrôler le biais dû par la première variable du quantile. Pour quantifier ces deux types de biais nous allons supposer deux hypothèses comme dans Lékina (2010, [72]).

C7. La fonction quantile extrême est ainsi définie :

$$\alpha \in]0, 1[\mapsto q(\alpha, x) \in]0, +\infty[$$

qui est dérivable et la fonction biais

$$\alpha \in]0, 1[\mapsto \Delta(\alpha, x) = \gamma_1(x) + \alpha \frac{\partial \log q(\alpha, x)}{\partial \alpha}$$

est continue et telle que

$$\lim_{\alpha \rightarrow 0} \Delta(\alpha, x) = 0$$

et nous notons aussi

$$\bar{\Delta}(a, x) = \sup_{\alpha \in]0, a[} |\Delta(\alpha, x)|.$$

Cette hypothèse permet de contrôler le biais dû à la première variable du quantile extrême conditionnel. La plus grande oscillation de la fonction log quantile est donnée par :

$$\omega_n(a) = \sup \left\{ \left| \log \frac{q(\alpha, t)}{q(\alpha, t')} \right| ; \alpha \in (a, 1 - a), (t, t') \in (B(x, h_{n,x}))^2 \right\}.$$

où $a \in]0, 1/2[$.

Cette fonction $\omega_n(\cdot)$ quant à elle permet de quantifier le biais dû du fait que nous travaillons dans le cas conditionnel.

Sous l'hypothèse **C7**, nous énonçons le théorème suivant pour les propriétés asymptotiques de l'estimateur des quantiles extrêmes conditionnels.

Théorème 2.4.2 *Supposons les conditions du Théorème 2.4.1 et C7 sont vérifiées. Soient deux suites $(\alpha_{m_{n,x}})_{n \geq 1}$ et $(\beta_{m_{n,x}})_{n \geq 1} := (1 - \widehat{F}_{m_{n,x}}(Z_{(m_{n,x}-k_x)}^x | x))_{n \geq 1}$ telles que $\alpha_{m_{n,x}} < \beta_{m_{n,x}}$ et $\xi_{m_{n,x}} = (m_{n,x} \beta_{m_{n,x}})^{1/2} \log\left(\frac{\beta_{m_{n,x}}}{\alpha_{m_{n,x}}}\right)$. Si $n \rightarrow \infty$, il existe $\delta > 0$ tel que $(m_{n,x} \beta_{m_{n,x}})^2 \omega_n(m_{n,x}^{-(1+\delta)}) \rightarrow 0$ et $k_x^{1/2} \max\{\xi_{m_{n,x}}^{-1}, \overline{\Delta}(\beta_{m_{n,x}}, x)\} \rightarrow 0$. Alors,*

$$\frac{\sqrt{k_x}}{\log\left(\frac{\beta_{m_{n,x}}}{\alpha_{m_{n,x}}}\right)} \log\left(\frac{\widehat{q}^{(c,\cdot)}(\alpha_{m_{n,x}}, x)}{q(\alpha_{m_{n,x}}, x)}\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\frac{\lambda(x)m(x) - \gamma_1(x)\varepsilon(x)}{p_x}, \frac{\sigma^2(x) + \gamma_1(x)^2 p_x(1 - p_x)}{p_x^2}\right).$$

Ce Théorème 2.4.2 révèle que la propriété asymptotique de l'estimateur du quantile extrême conditionnel provient de celle de l'estimateur de l'indice de queue conditionnel.

Nous allons à présent étudier les comportements de nos estimateurs via simulations et application sur des données réelles médicales.

2.5 Simulation et illustration sur des données réelles

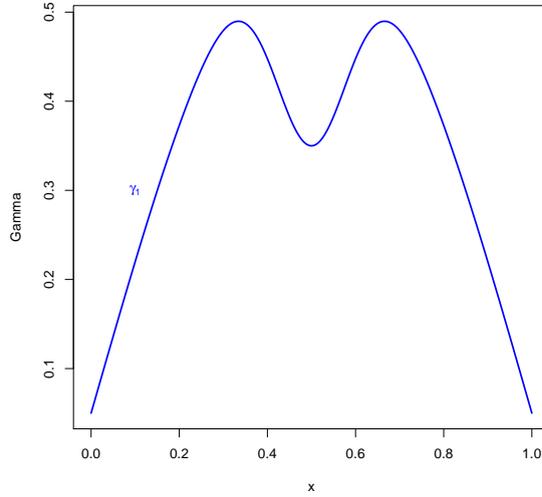
2.5.1 Expériences numériques par simulation

Dans cette partie, nous voulons vérifier trois choses :

- Effet du pourcentage de censure sur les estimations,
- Influence du type de l'estimateur standard pris sur les estimations,
- Consistance et normalités asymptotiques des estimateurs que nous avons proposés.

Pour cela on génère un échantillon $\{(x_i, Y_i, C_i), i = 1, \dots, n\}$ de taille n suivant la loi du triplet $(X, Y, C) \in \mathcal{X} \times \mathbb{R} \times \mathbb{R}$ où X est la covariable de loi uniforme standard, Y est la variable d'intérêt de loi conditionnelle à X , la loi de Paréto de paramètre fonctionnel $\gamma_1(\cdot)$ défini :

$$x \in [0, 1] \mapsto 0.5 \left(0.1 + \sin(\pi x) \times \left(1.1 - 0.5 \exp\left(-64(x - 0.5)^2\right) \right) \right)$$

FIGURE 2.2 – Comportement graphique de la fonction γ_1 sur $[0, 1]$.

et C est la variable de censure de loi Paréto de paramètre fonctionnel $\gamma_2(\cdot)$ défini selon le pourcentage de censure,

$$\gamma_2(x) = \begin{cases} 0.5, & \text{si \% de censure} \simeq 40\% \\ 1, & \text{si \% de censure} \simeq 25\% \\ 2.5, & \text{si \% de censure} \simeq 10\%. \end{cases} \quad (2.7)$$

Cette fonction $\gamma_2(\cdot)$ définie dans l'équation (2.7) nous permet de connaître le pourcentage de censure de l'échantillon considéré.

Pour toute réplication R , nous estimons $\gamma_1(\cdot)$ au point $x = 0.5$ ($\gamma_1(0.5) = 0.35$) en utilisant les estimateurs de Hill, Moments et UH. Le choix de k_x est largement discuté dans la littérature. Soit $\hat{\gamma}_{i,m_{n,x}}^{(c,\cdot),j}(x)$ l'estimateur de $\gamma_1(x)$ obtenu à la j -ième réplication ($j = 1, \dots, R$) avec $k_x = i$ ($i = 1, \dots, m_{n,x}$). Nous définissons l'estimateur de k_x comme suit :

$$k_x^{opt} := \arg \min_{1 \leq i \leq m_{n,x}} \left\{ MSE(\hat{\gamma}_{i,m_{n,x}}^{(c,\cdot)}(x)) \right\} = \arg \min_{1 \leq i \leq m_{n,x}} \left\{ \frac{1}{R} \sum_{j=1}^R (\hat{\gamma}_{i,m_{n,x}}^{(c,\cdot),j}(x) - \gamma_1(x))^2 \right\}.$$

C'est dans ce cadre que de Haan et Peng (1998, [36]) ont proposé le k_x^{opt} défini comme suit :

$$k_x^{opt} = \begin{cases} 1 + n^{\frac{2\gamma_1(x)}{2\gamma_1(x)+1}} \left(\frac{(1 + \gamma_1(x))^2}{2\gamma_1(x)} \right)^{1/(2\gamma_1(x)+1)} & \text{si } \gamma_1(x) \in]0, 1[, \\ 2n^{2/3} & \text{si } \gamma_1(x) > 1. \end{cases}$$

Nous avons utilisé la méthode de la fenêtre mobile avec une boule de centre $x = 0.5$ et de rayon $h = 0.1$. L'ordre du quantile est fixé à $\alpha_{m_{n,x}} = 1/5000$. Toutes ces considérations donnent une vraie valeur $\gamma_1(0.5) = 0.35$ et son quantile correspondant 19.70786.

Ainsi, nous obtenons les résultats suivants :

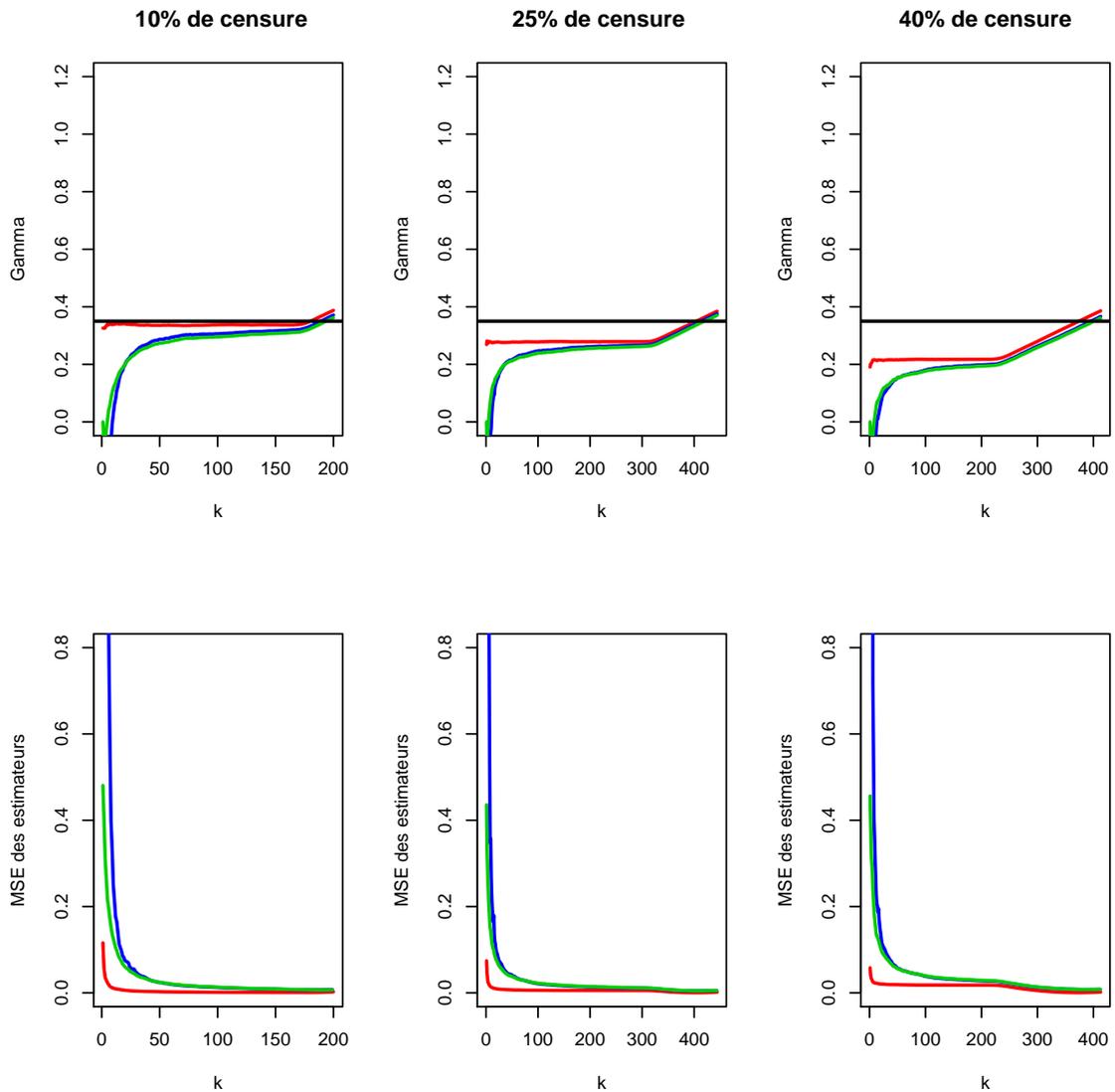


FIGURE 2.3 – Comportement graphique de l'estimateur de l'indice en fonction du pourcentage de censure et du type d'estimateur ($n = 1500$): *hill*, *moments* et *uh*.

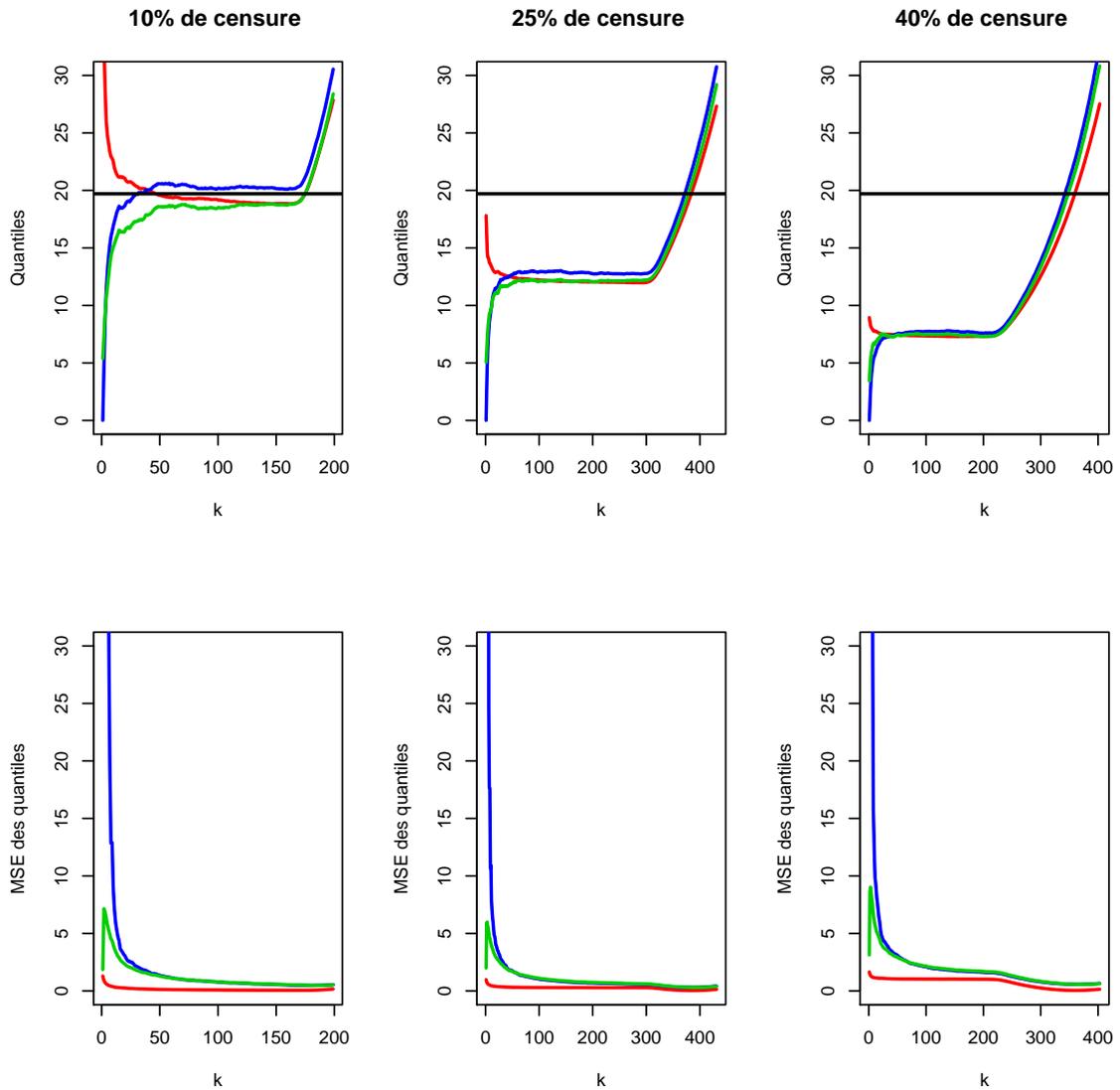


FIGURE 2.4 – Comportement graphique de l'estimateur des quantiles en fonction du pourcentage de censure et du type d'estimateur : ($n = 1500$) *hill*, *moments* et *uh*.

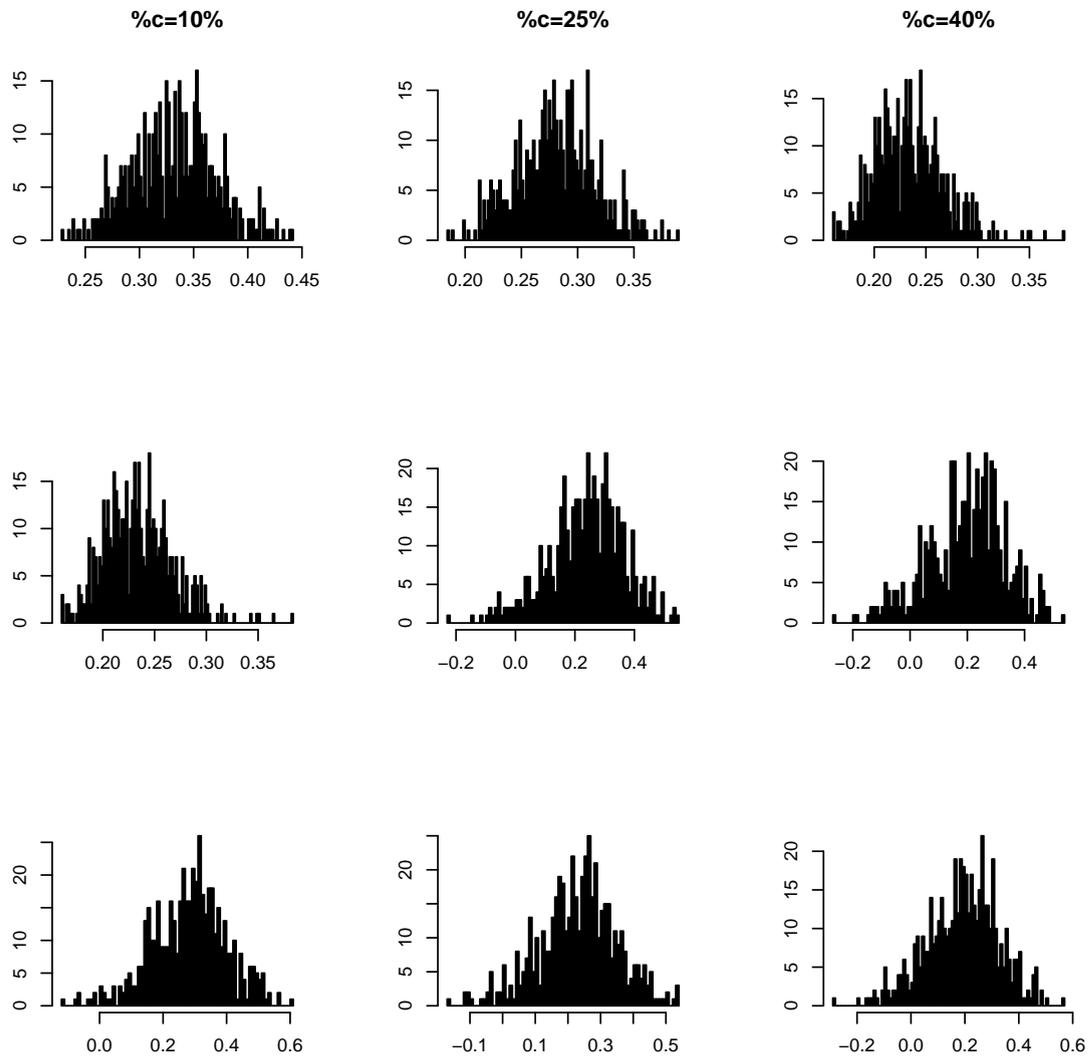


FIGURE 2.5 – Histogramme pour $R = 1000$ Hill (1^{re} ligne), Moments (2^e ligne) et UH (3^e ligne) estimation de l'indice des valeurs extrêmes au point $x = .5$ ($\gamma_1(.5) = .35$), pour $c = 10\%$ (colonne de gauche), $c = 25\%$ (milieu), $c = 40\%$ (droite). La taille de l'échantillon est 500.

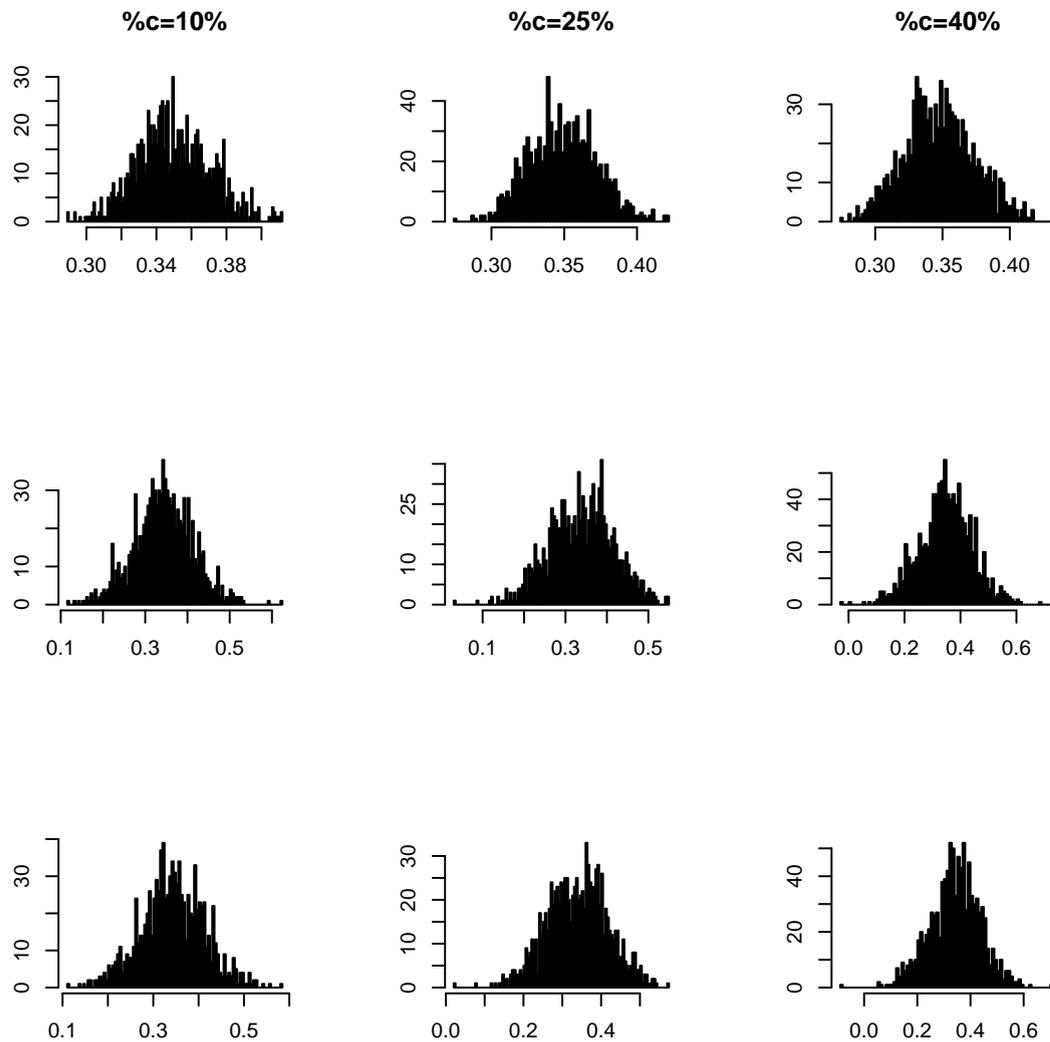


FIGURE 2.6 – Histogramme pour $R = 1000$ Hill (1^{re} ligne), Moments (2^e ligne) et UH (3^e ligne) estimation de l'indice des valeurs extrêmes au point $x = .5$ ($\gamma_1(.5) = .35$), pour $c = 10\%$ (colonne de gauche), $c = 25\%$ (milieu), $c = 40\%$ (droite). La taille de l'échantillon est 1500.

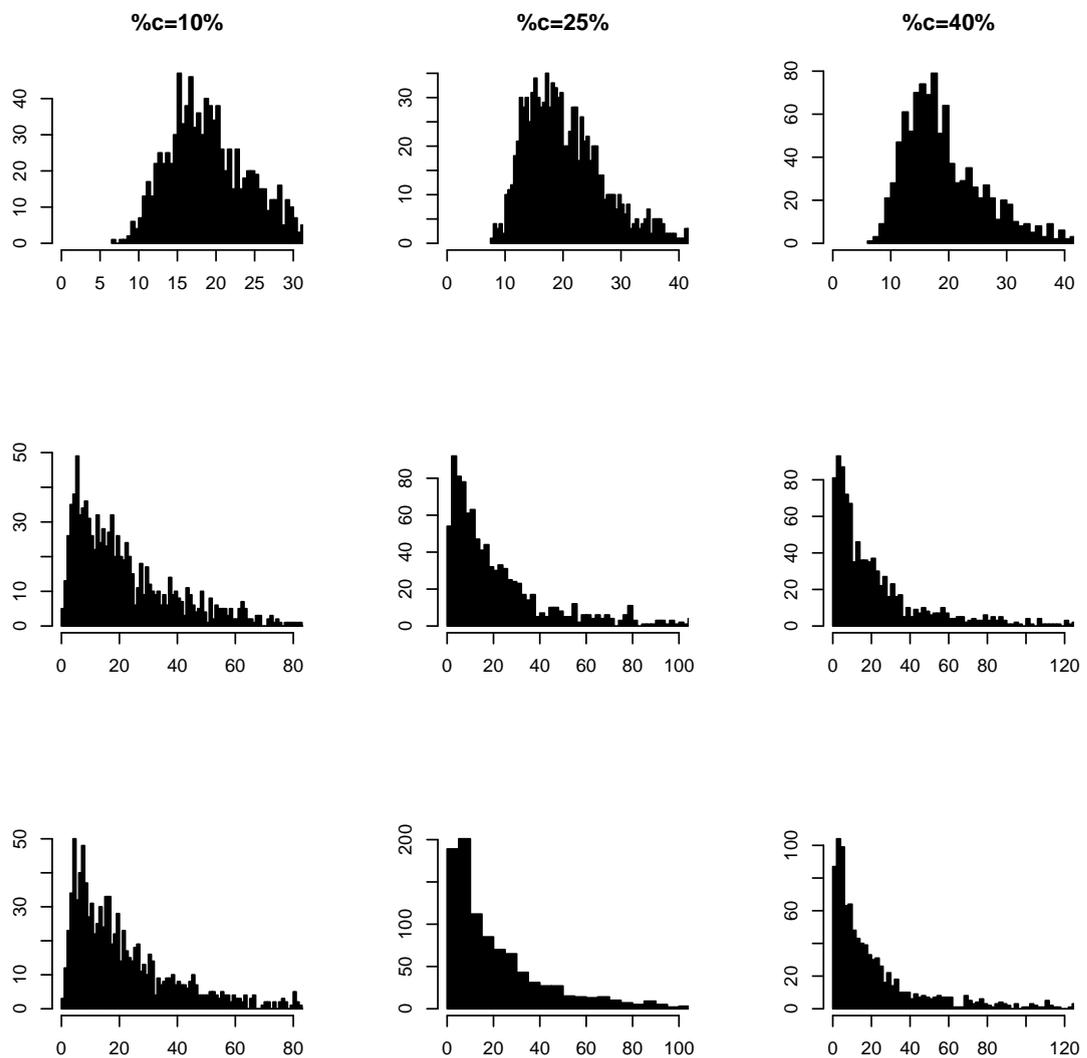


FIGURE 2.7 – Histogramme pour $R = 1000$ de l'estimateur de $q(1/5000, .5) \approx 19.70786$, basés sur les estimateurs de Hill (1^{re} ligne), des Moments (2^e ligne) et UH (3^e ligne) estimateurs de l'indice de queue conditionnel, pour $c = 10\%$ (colonne de gauche), $c = 25\%$ (milieu), $c = 40\%$ (droite). La taille de l'échantillon est 500.

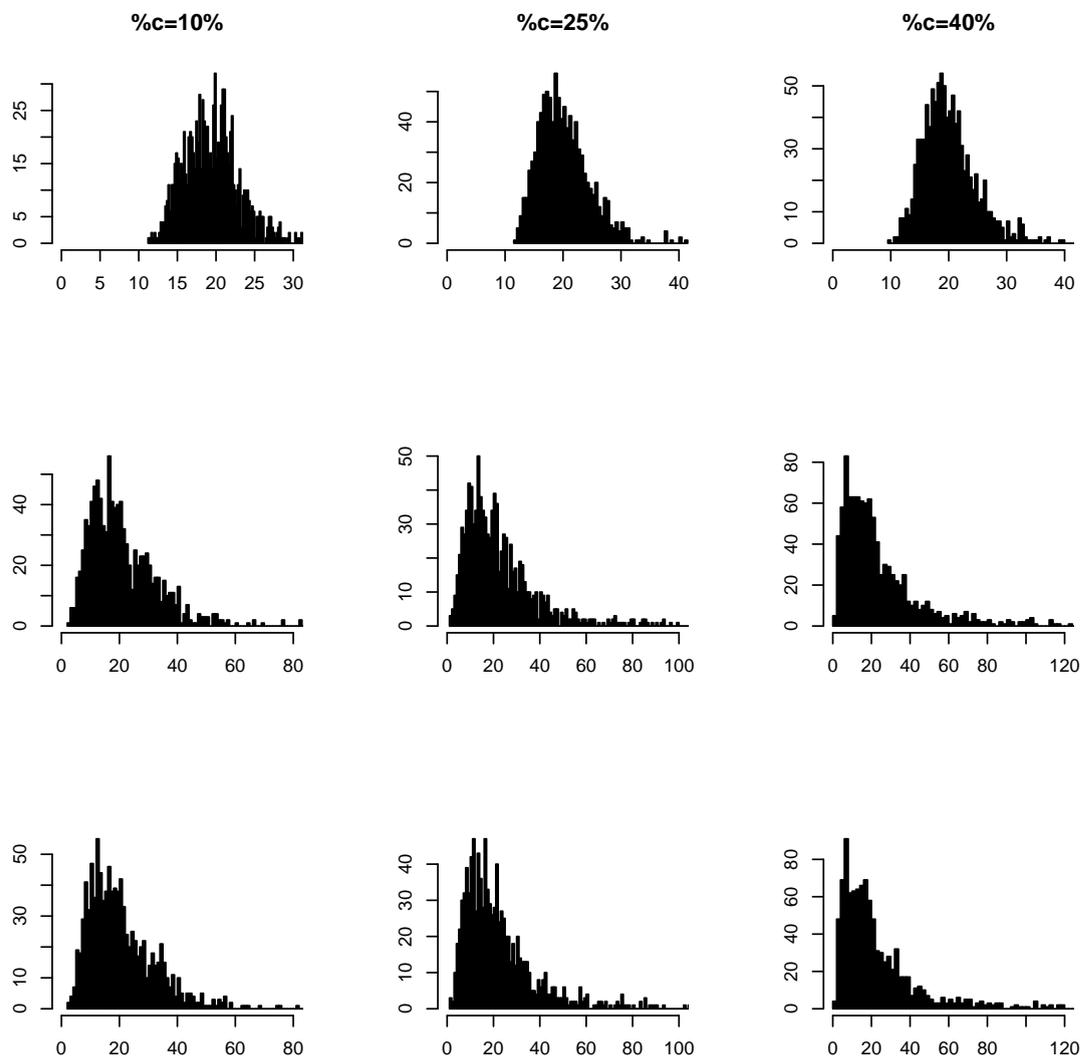


FIGURE 2.8 – Histogramme pour $R = 1000$ de l'estimateur de $q(1/5000, .5) \approx 19.70786$, basés sur les estimateurs de Hill (1^{re} ligne), des Moments (2^e ligne) et UH (3^e ligne) estimateurs de l'indice de queue conditionnel, pour $c = 10\%$ (colonne de gauche), $c = 25\%$ (milieu), $c = 40\%$ (droite). La taille de l'échantillon est 1500 .

n	Estimateur de Hill			Estimateur des Moments			Estimateur UH		
	10%	25%	40%	10%	25%	40%	10%	25%	40%
500	.349	.350	.345	.323	.315	.317	.322	.304	.323
	(.037)	(.043)	(.045)	(.116)	(.134)	(.172)	(.114)	(.137)	(.171)
	[.030]	[.034]	[.035]	[.090]	[.107]	[.136]	[.090]	[.109]	[.135]
	[.273,	[.270,	[.256,	[.081,	[.040,	[0*,	[.084,	[.029,	[0*,
	.425]	.428]	.434]	.566]	.591]	.652]	.561]	.580]	.652]
.758 [†]	.961 [†]	.963 [†]	.900 [†]	.908 [†]	.950 [†]	.945 [†]	.960 [†]	.964 [†]	
1000	.346	.349	.346	.337	.330	.334	.339	.336	.326
	(.027)	(.029)	(.032)	(.081)	(.097)	(.119)	(.081)	(.099)	(.119)
	[.022]	[.023]	[.026]	[.065]	[.076]	[.093]	[.065]	[.079]	[.093]
	[.293,	[.290,	[.283,	[.166,	[.124,	[.095,	[.171,	[.135,	[.088,
	.400]	.408]	.409]	.509]	.535]	.572]	.508]	.536]	.561]
.969 [†]	.986 [†]	.990 [†]	.959 [†]	.964 [†]	.970 [†]	.965 [†]	.968 [†]	.970 [†]	
1500	.347	.348	.345	.344	.339	.335	.342	.340	.339
	(.021)	(.025)	(.030)	(.067)	(.071)	(.101)	(.067)	(.072)	(.101)
	[.017]	[.020]	[.024]	[.053]	[.057]	[.080]	[.053]	[.058]	[.081]
	[.304,	[.300,	[.289,	[.208,	[.171,	[.122,	[.206,	[.183,	[.129,
	.389]	.396]	.401]	.480]	.506]	.549]	.477]	.511]	.551]
.973 [†]	.993 [†]	.995 [†]	.971 [†]	.977 [†]	.981 [†]	.975 [†]	.980 [†]	.984 [†]	
2000	.349	.349	.348	.337	.340	.335	.342	.345	.337
	(.019)	(.020)	(.022)	(.058)	(.067)	(.087)	(.058)	(.068)	(.087)
	[.015]	[.016]	[.017]	[.045]	[.052]	[.068]	[.046]	[.053]	[.068]
	[.311,	[.310,	[.303,	[.212,	[.202,	[.160,	[.219,	[.209,	[.164,
	.387]	.389]	.394]	.463]	.477]	.509]	.466]	.480]	.512]
.987 [†]	.995 [†]	.998 [†]	.980 [†]	.984 [†]	.986 [†]	.981 [†]	.984 [†]	.987 [†]	

Tableau 2.1 – Résultats des simulations pour l’estimateur (2.1) de $\gamma_1(x)$. Pour chaque scénario, la première ligne donne la moyenne pour $R = 1000$, estimation de $\gamma_1(x)$. (·): Racine carrée de l’erreur quadratique moyenne (RMSE). [·]: Erreur absolue moyenne (MAE). [·, ·]: intervalle de confiance au niveau de confiance 0.95 pour $\gamma_1(x)$ (* la borne inférieure de l’intervalle de confiance est négative et donc tronquée à 0). †: probabilité de couverture.

n	Estimateur de Hill			Estimateur des Moments			Estimateur UH		
	10%	25%	40%	10%	25%	40%	10%	25%	40%
500	19.777	20.225	20.072	27.490	27.343	27.279	25.313	29.046	37.537
	(.258)	(.265)	(.310)	(.807)	(.877)	(1.164)	(.806)	(.892)	(1.213)
	[.326]	[.333]	[.383]	[1.008]	[1.122]	[1.431]	[1.007]	[1.154]	[1.508]
	[16.00,	[15.90,	[15.56,	[15.81,	[14.76,	[14.25,	[14.56,	[15.82,	[20.05,
	25.88]	27.77]	28.25]	105.03]	184.95]	317.34]	96.71]	176.50]	292.11]
0.594†	0.936†	0.970†	0.680†	0.880†	0.882†	0.230†	0.921†	0.926†	
1000	19.381	19.960	20.086	24.065	22.762	27.813	23.189	23.465	27.695
	(.182)	(.206)	(.222)	(.565)	(.671)	(.820)	(.562)	(.676)	(.826)
	[.226]	[.259]	[.280]	[.712]	[.843]	[1.022]	[.709]	[.841]	[1.025]
	[16.54,	[16.71,	[16.55,	[15.60,	[14.00,	[16.81,	[15.03,	[14.53,	[16.82,
	23.39]	24.77]	25.53]	52.60]	60.69]	80.31]	50.68]	60.91]	78.30]
0.708†	0.971†	0.989†	0.863†	0.942†	0.948†	0.262†	0.958†	0.961†	
1500	19.841	19.981	19.905	22.506	21.621	22.762	22.584	23.106	26.684
	(.142)	(.161)	(.179)	(.468)	(.534)	(.639)	(.473)	(.544)	(.638)
	[0.177]	[0.199]	[0.223]	[0.597]	[0.666]	[0.818]	[0.600]	[0.674]	[0.684]
	[17.75,	[17.21,	[17.26,	[16.32,	[14.39,	[15.15,	[16.38,	[15.49,	[17.97,
	22.47]	23.59]	23.70]	36.19]	43.37]	45.72]	36.32]	45.42]	51.72]
0.910†	0.990†	0.992†	0.904†	0.957†	0.974†	0.972†	0.978†	0.981†	
2000	19.887	19.841	20.048	21.889	22.506	22.132	21.159	22.584	24.696
	(.131)	(.142)	(.160)	(.400)	(.468)	(.616)	(.397)	(.473)	(.624)
	[.164]	[.177]	[.202]	[.503]	[.597]	[.766]	[.500]	[.600]	[.771]
	[17.79,	[17.42,	[17.60,	[15.85,	[15.61,	[15.28,	[15.32,	[15.71,	[17.20,
	21.53]	23.04]	23.28]	35.36]	40.28]	40.07]	34.18]	40.13]	43.72]
0.922†	0.992†	0.994†	0.953†	0.963†	0.978†	0.277†	0.981†	0.983†	

Tableau 2.2 – Résultats des simulations pour l'estimateur (2.3) de $q(1/5000, .5)$. Pour chaque scénario, la première ligne donne la moyenne pour $R = 1000$, estimation de $q(1/5000, .5)$. (\cdot): Racine carrée de l'erreur quadratique moyenne (RMSE). [\cdot]: Erreur absolue moyenne (MAE). [\cdot, \cdot]: intervalle de confiance au niveau de confiance 0.95 pour $q(1/5000, .5)$ (* la borne inférieure de l'intervalle de confiance est négative et donc tronquée à 0). †: probabilité de couverture.

Pour évaluer la performance de l'estimateur proposé dans (2.1), nous allons fournir une comparaison avec deux autres stratégies d'estimation simples. La première est une procédure nommée cas-complet ("CC", pour abrégé) qui consiste à éliminer toutes les observations censurées des échantillons simulés et de faire le calcul en utilisant la méthode de la fenêtre mobile avec les estimateurs proposés dans Gardes et Girard (2008, [51]) de l'indice des valeurs extrêmes conditionnel avec l'ensemble des données qui en résultent. Dans la seconde variante, nous appliquons l'estimateur fenêtre mobile à l'ensemble des jeux de données simulées en traitant chaque observation Z_i comme s'il s'agissait non censurée. Nous nous référons à cette alternative comme "CI" (Cas-Ignoré). Nous considérons les mêmes scénarios de simulation que dans le Tableau 2.1. Pour cela, nous procédons comme suit.

- **Cas-Complet**

Nous avons simulé des données censurées aléatoirement à droite mesurées conjointement avec des covariables à design fixe. Puis nous avons tronqué l'échantillon en prenant seulement les observations qui ne sont pas censurées pour pouvoir appliquer quelques estimateurs de la théorie des valeurs extrêmes existants dans la littérature.

- **Cas-Ignoré**

Nous avons aussi considéré ces mêmes données en imposant l'indicateur de censure égal partout à un. C'est-à-dire nous avons négligé la censure. De ce fait la proportion des données non censurées est égale à un partout. Autrement dit nous utilisons les estimateurs de l'indice et des quantiles extrêmes avec des données non censurées aléatoirement à droite.

Nous déterminons également la racine carrée de l'erreur quadratique moyenne (RMSE) et l'erreur absolue moyenne (MAE). Les résultats de ces simulations sont rapportés dans les Tableaux 2.3 et 2.4.

Estimation pour le Cas-Complet									
n	Estimateur de Hill			Estimateur des Moments			Estimateur UH		
	10%	25%	40%	10%	25%	40%	10%	25%	40%
500	.335 (.045) [.036]	.277 (.081) [.073]	.216 (.138) [.133]	.294 (.141) [.107]	.215 (.202) [.159]	.138 (.277) [.222]	.283 (.139) [.107]	.203 (.201) [.161]	.145 (.263) [.218]
1000	.332 (.032) [.026]	.278 (.076) [.071]	.217 (.134) [.132]	.319 (.090) [.070]	.253 (.136) [.110]	.193 (.187) [.159]	.310 (.093) [.073]	.246 (.140) [.115]	.190 (.189) [.163]
1500	.339 (.025) [.026]	.282 (.071) [.067]	.218 (.028) [.131]	.333 (.070) [.056]	.265 (.115) [.094]	.195 (.177) [.159]	.324 (.072) [.057]	.258 (.119) [.099]	.191 (.180) [.160]
2000	.335 (.023) [.019]	.278 (.073) [.071]	.218 (.132) [.131]	.330 (.060) [.048]	.268 (.102) [.085]	.203 (.162) [.147]	.323 (.061) [.049]	.264 (.106) [.090]	.198 (.167) [.152]
Estimation pour le Cas-Ignoré									
n	Estimateur de Hill			Estimateur des Moments			Estimateur UH		
	10%	25%	40%	10%	25%	40%	10%	25%	40%
500	.336 (.041) [.032]	.277 (.077) [.072]	.219 (.132) [.130]	.314 (.118) [.093]	.252 (.140) [.113]	.187 (.193) [.166]	.303 (.115) [.092]	.245 (.143) [.118]	.182 (.195) [.171]
1000	.334 (.029) [.024]	.278 (.074) [.071]	.218 (.131) [.129]	.325 (.084) [.072]	.266 (.110) [.091]	.204 (.161) [.146]	.317 (.083) [.071]	.260 (.114) [.095]	.199 (.165) [.151]
1500	.332 (.026) [.021]	.277 (.074) [.069]	.217 (.130) [.126]	.330 (.069) [.055]	.268 (.102) [.086]	.208 (.153) [.141]	.323 (.069) [.054]	.263 (.105) [.090]	.204 (.157) [.145]
2000	.334 (.022) [.018]	.279 (.071) [.067]	.219 (.127) [.124]	.333 (.061) [.052]	.277 (.089) [.075]	.213 (.145) [.136]	.327 (.062) [.051]	.273 (.092) [.079]	.210 (.148) [.139]

Tableau 2.3 – Estimation de $\gamma_1(x)$: le Cas-Complet et Cas-Ignoré. Pour chaque situation de simulation, la première ligne donne l'estimation de $\gamma_1(x)$. (·): Racine carrée de l'erreur quadratique moyenne (RMSE). [·]: Erreur absolue moyenne (MAE).

Estimation pour le Cas-Complet									
n	Estimateur de Hill			Estimateur des Moments			Estimateur UH		
	10%	25%	40%	10%	25%	40%	10%	25%	40%
500	18.603 (.380) [.339]	11.258 (.693) [.625]	6.637 (1.173) [1.134]	20.392 (1.145) [1.122]	12.173 (1.256) [1.252]	7.787 (1.764) [1.528]	17.995 (1.165) [1.112]	11.099 (1.672) [1.342]	7.776 (2.180) [1.816]
1000	17.403 (.270) [.229]	10.958 (.648) [.611]	6.507 (1.144) [1.127]	19.379 (.762) [.716]	11.876 (1.137) [.928]	7.102 (1.536) [1.334]	18.031 (.789) [.719]	11.039 (1.180) [.973]	7.102 (1.593) [1.381]
1500	18.374 (.214) [.183]	11.233 (.604) [.577]	6.508 (1.132) [1.120]	20.138 (.614) [.605]	11.776 (.972) [.800]	6.773 (1.490) [1.323]	18.670 (.607) [.606]	11.125 (1.013) [.844]	6.498 (1.362) [1.362]
2000	17.704 (.197) [.168]	10.814 (.626) [.610]	6.471 (1.132) [1.123]	18.708 (.515) [.513]	11.222 (.870) [.728]	6.716 (1.378) [1.252]	18.569 (0.519) [.516]	10.668 (.908) [.766]	6.445 (1.414) [1.291]
Estimation pour le Cas-Ignoré									
n	Estimateur de Hill			Estimateur des Moments			Estimateur UH		
	10%	25%	40%	10%	25%	40%	10%	25%	40%
500	18.246 (.334) [.331]	10.975 (.658) [.614]	6.593 (1.131) [1.114]	21.410 (1.101) [1.090]	12.408 (1.193) [1.129]	7.024 (1.590) [1.447]	19.227 (1.113) [1.092]	11.521 (1.219) [1.161]	6.757 (1.668) [1.532]
1000	17.639 (.236) [.232]	10.907 (.633) [.607]	6.472 (1.130) [1.122]	19.174 (.726) [.723]	11.523 (.937) [.777]	6.711 (1.371) [1.245]	17.971 (.725) [.721]	10.914 (.972) [.898]	6.441 (1.405) [1.285]
1500	17.280 (.222) [.182]	10.750 (.635) [.617]	6.431 (1.139) [1.130]	19.226 (.647) [.645]	11.272 (.868) [.733]	6.767 (1.308) [1.205]	18.127 (.670) [.667]	10.710 (.901) [.771]	6.502 (1.342) [1.242]
2000	17.416 (.188) [.177]	10.949 (.615) [.598]	6.497 (1.123) [1.116]	18.716 (.560) [.554]	11.721 (.758) [.646]	6.755 (1.236) [1.160]	17.846 (.573) [.671]	11.326 (.786) [.676]	6.559 (1.263) [1.189]

Tableau 2.4 – Estimation de $q(1/5000, .5)$: le CC et CI. Pour chaque scénario de simulation, la première ligne donne l'estimation de $q(1/5000, .5)$. (·): Racine carrée de l'erreur quadratique moyenne (RMSE). [·]: Erreur absolue moyenne (MAE).

2.5.2 Exemple sur des données réelles

Le Syndrome d'Immuno Déficience Acquis, plus connu sous son acronyme SIDA, est un ensemble de symptômes consécutifs à la destruction de plusieurs cellules du système immunitaire par un rétrovirus. Le sida est le dernier stade de l'infection par ce virus et termine par affaiblir le système immunitaire qui sera vulnérable pour d'autres maladies. Ainsi, le VIH affecte le système immunitaire en infectant les cellules centrales de ce système : les lymphocytes T CD4+¹. Les lymphocytes T CD4+ sont des cellules du système immunitaire dont la prolifération permet de diriger et d'activer d'autres cellules de l'immunité, comme les lymphocytes B, pour éliminer un pathogène. Ils doivent leur nom aux marqueurs protéiques CD4 retrouvé sur la surface de leur membrane. Aussi nommés lymphocytes T auxiliaires, ils constituent les cibles privilégiées du VIH, et leur destruction conduit à l'arrêt du fonctionnement du système immunitaire, amenant alors au stade Sida.

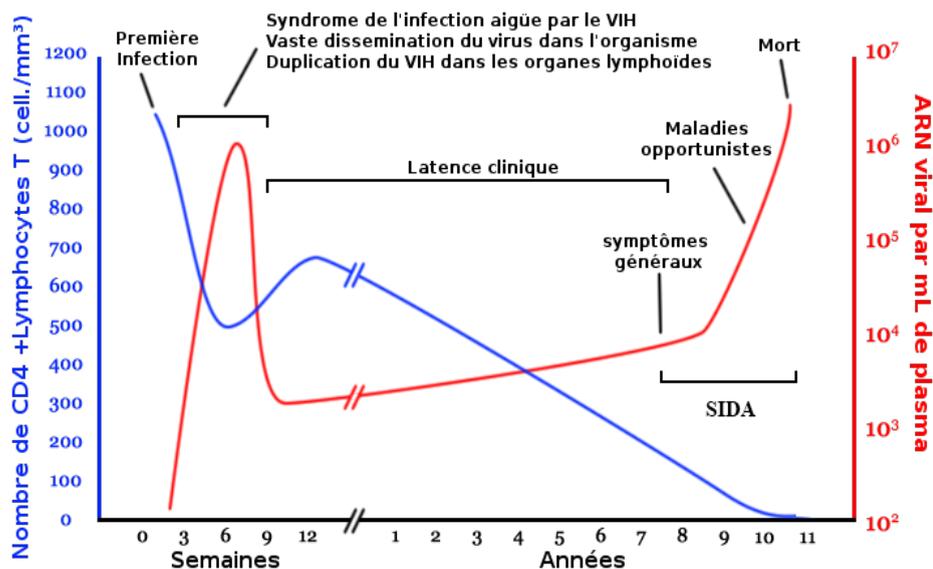


FIGURE 2.9 – Évolution de la charge virale et du système immunitaire à l'absence de traitement (Source : Wikipédia).

L'organisation mondiale de la santé a publié dans son site que plus de 36 millions de personnes sont mortes depuis le début de la maladie en 1981 et continue à faire ses ravages. Et qu'en 2012, il y avait environ 35.3 millions de personnes vivantes avec le virus de l'immunodéficience humaine (VIH). D'où l'intérêt d'illustrer notre méthodologie sur un ensemble de données de survie du SIDA. Cette base de données a été auparavant analysée par Einmahl et al. (2008, [44]). Nous comparons les résultats des deux analyses. La base de données contient $n = 2754$ patients masculins

1. Une molécule à la surface des cellules sur laquelle le VIH se fixe.

atteints de SIDA en Australie avant le 1^{er} Juillet 1991. La source des données est le Dr. PJ Salomon et le Centre national australien dans l'épidémiologie du VIH et de recherche clinique (voir Venables et Ripley, 2002, [93]). Ces données sont dans le package MASS sous R dans la base de données Aids2. Les informations sur chaque patient comprennent la date du diagnostic, l'âge au moment du diagnostic, la date du décès ou à la fin de l'observation et un indicateur qui vaut 1 si le patient est décédé et 0 sinon. 1708 patients sont décédés; les autres temps de survie sont censurées aléatoirement à droite. Einmahl et al. (2008, [44]) ont utilisé ces données pour estimer l'indice de queue γ_1 et son quantile extrême $F^{\leftarrow}(1 - \alpha)$ (avec $\alpha = 1/1000$) correspondant dans le cas non-conditionnel avec la fonction de répartition F de la variable d'intérêt Y et ils obtiennent $\hat{\gamma}_1 = 0.14$ et $\hat{F}^{\leftarrow}(1 - 1/1000) = 25$ ans. Nous réanalysons ces données pour calculer le temps de latence en considérant l'âge au diagnostic comme covariable \mathbf{x} .

Nous estimons le quantile $q(1/1000, \mathbf{x})$ d'ordre $1 - 1/1000$ de la distribution conditionnelle de Y pour \mathbf{x} donné, par exemple $\mathbf{x} = 37$ (médiane de x_1, \dots, x_n), $\mathbf{x} = 27$ (médiane - $\text{sd}(x_1, \dots, x_n)$) et $x = 47$ (médiane + $\text{sd}(x_1, \dots, x_n)$), où $\text{sd}(x_1, \dots, x_n)$ est la valeur du standard déviation de x_1, \dots, x_n . Pour calculer l'estimateur proposé dans l'équation (2.3), nous avons besoin les valeurs appropriées de $h_{n,x}$ et de k_x . Nous proposons la procédure suivante. Nous considérons un ensemble raisonnable pour $h_{n,x}$, noté $\mathcal{H} = \{3, 4, 5, 6, 7\}$ et pour tout $h_{n,x} \in \mathcal{H}$ et $k = 1, \dots, m_{n,x}$, nous calculons l'estimateur proposé dans (2.3) de $q(1/1000, x)$ basé sur les estimateurs de Hill et UH estimés en utilisant l'équation (2.1) de $\gamma_1(x)$. Ces estimateurs sont notés par $\hat{q}_{h,k}^{(c,H)}(1/1000, x)$ et $\hat{q}_{h,k}^{(c,UH)}(1/1000, x)$ respectivement. Alors, nous allons calculer la différence $d_{h,k}(x) := |\hat{q}_{h,k}^{(c,H)}(1/1000, x) - \hat{q}_{h,k}^{(c,UH)}(1/1000, x)|$ et nous retenons le couple (h_x^*, k_x^*) tel que $(h_x^*, k_x^*) = \underset{h \in \mathcal{H}, 1 \leq k \leq m_{n,x}}{\text{arg min}} d_{h,k}(x)$. Finalement, nous estimons $\gamma_1(x)$ et $q(1/1000, x)$ de (2.1) et (2.3) avec $(h_{n,x}, k_x) = (h_x^*, k_x^*)$. Les résultats numériques sont rapportés dans le Tableau 2.5 avec l'estimateur de Hill comme celui standard. Nous proposons aussi quelques graphes pour le choix de (h_x^*, k_x^*) . Pour x donné, nous traçons les fonctions $k \mapsto d_{h,k}(x)$ ($h_{n,x} \in \mathcal{H}$). Le couple (h_x^*, k_x^*) peut être déterminé en choisissant celui qui minimise la courbe $d_{h,k}(x)$. Les graphes pour le choix (h_x^*, k_x^*) dans l'équation (2.3) quand $x = 27, 37$ et 47 sont donnés dans la Figure 2.10.

age x	Estimateur de Hill			Cas-Complet			Cas-Ignoré		
	27	37	47	27	37	47	27	37	47
h_x^*	3	3	6	7	4	6	7	7	7
k_x^*	101	110	96	114	81	82	81	113	80
est. de $\gamma_1(x)$.304	.340	.323	.276	.255	.256	.237	.236	.271
est. de $q(1/1000, x)$	10.04	12.77	11.29	8.38	7.71	8.90	7.53	6.52	8.32

Tableau 2.5 – Application sur les données du SIDA. Nous utilisons notre méthode pour estimer $\gamma_1(x)$ et $q(1/1000, x)$ avec les estimateurs proposés dans (2.1) et (2.3).

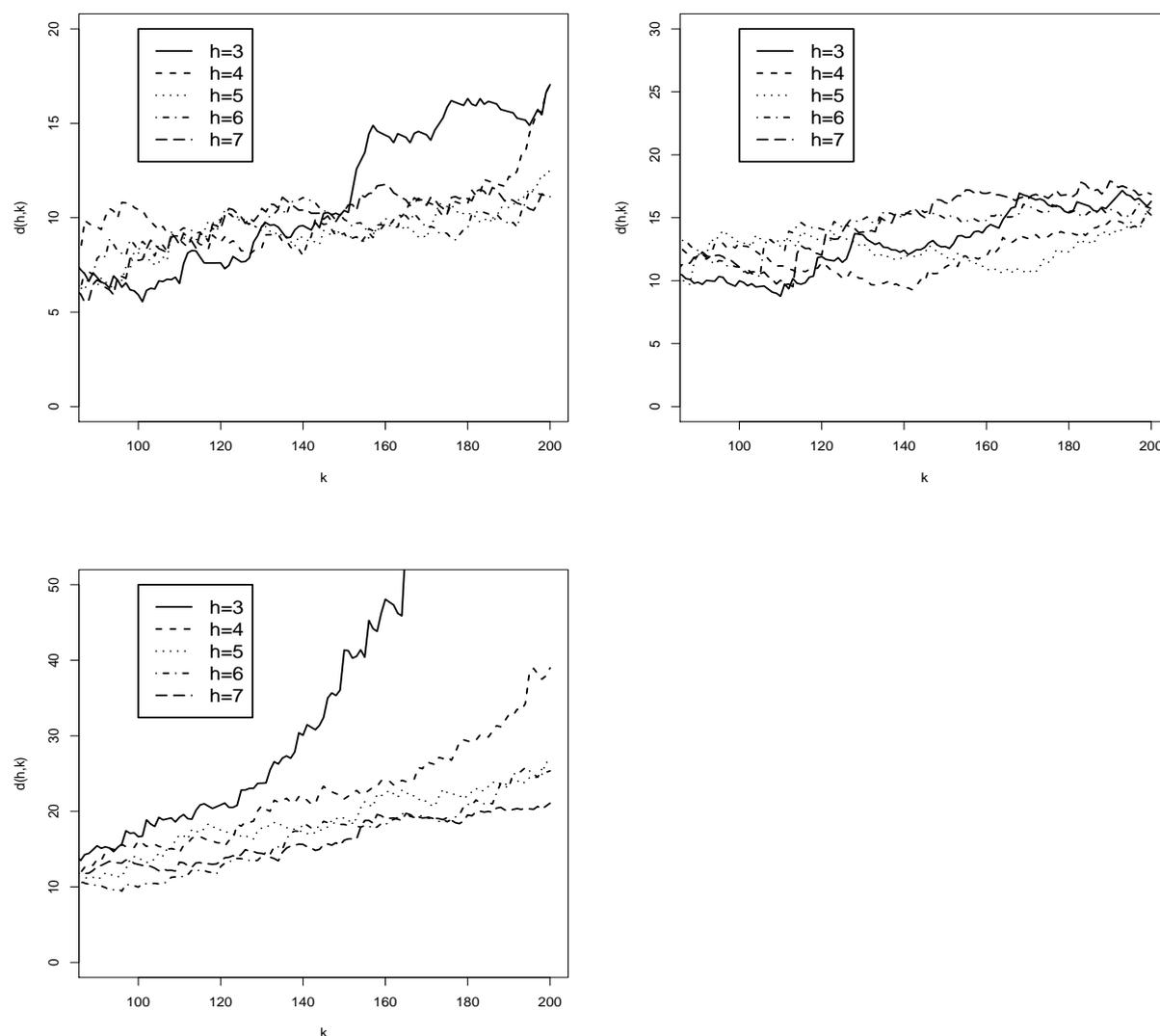


FIGURE 2.10 – Choix de (h_x^*, k_x^*) . Graphe de la fonction $k \mapsto |d_{h,k}(x)|$ (pour $h \in \mathcal{H}$) quand $x = 27$ (à gauche), $x = 37$ (à droite) et $x = 47$ (en bas) : notre cas.

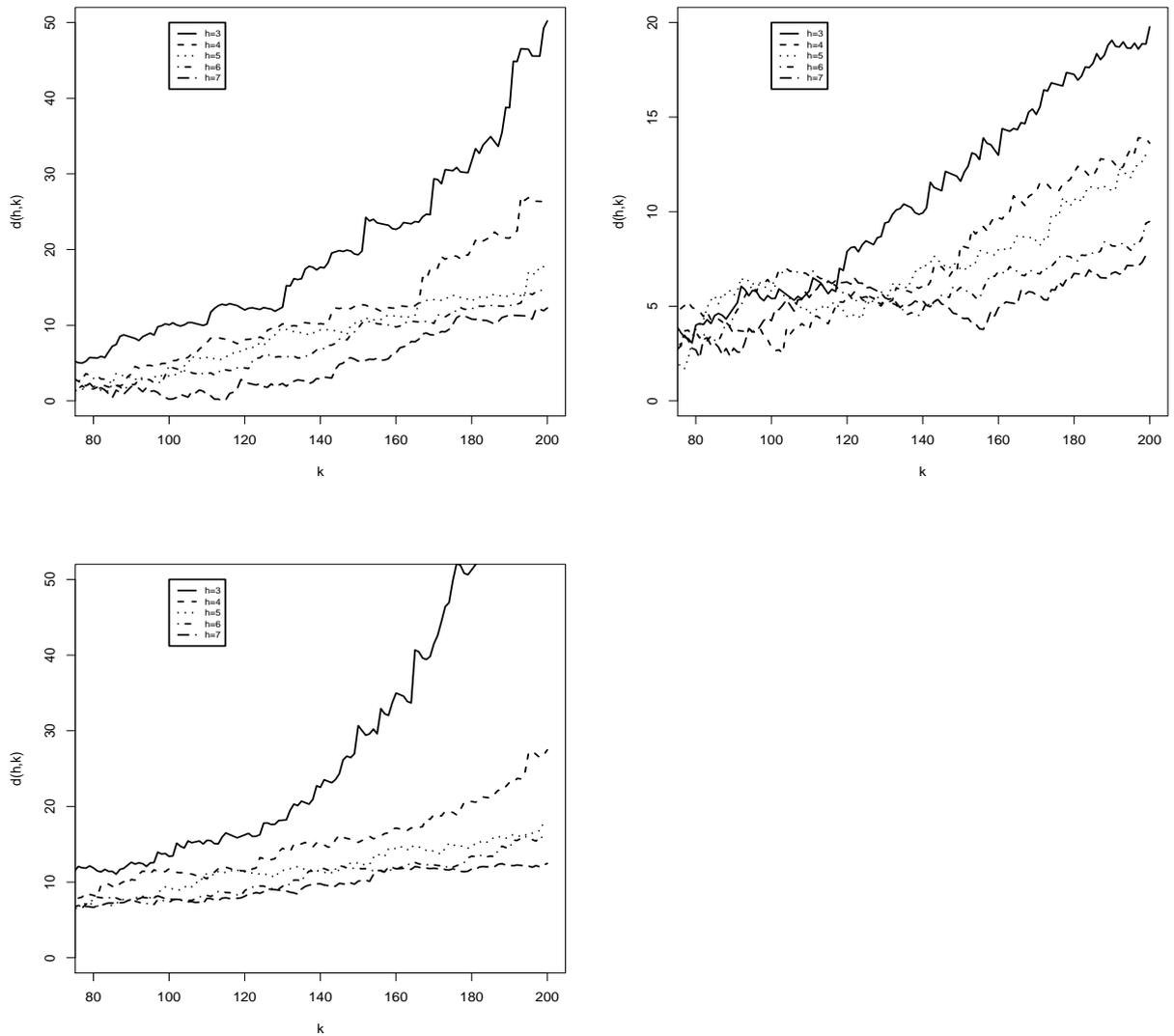


FIGURE 2.11 – Choix de (h_x^*, k_x^*) . Graphe de la fonction $k \mapsto |d_{h,k}(x)|$ (pour $h \in \mathcal{H}$) quand $x = 27$ (à gauche), $x = 37$ (à droite) et $x = 47$ (en bas) : Cas-Complet.

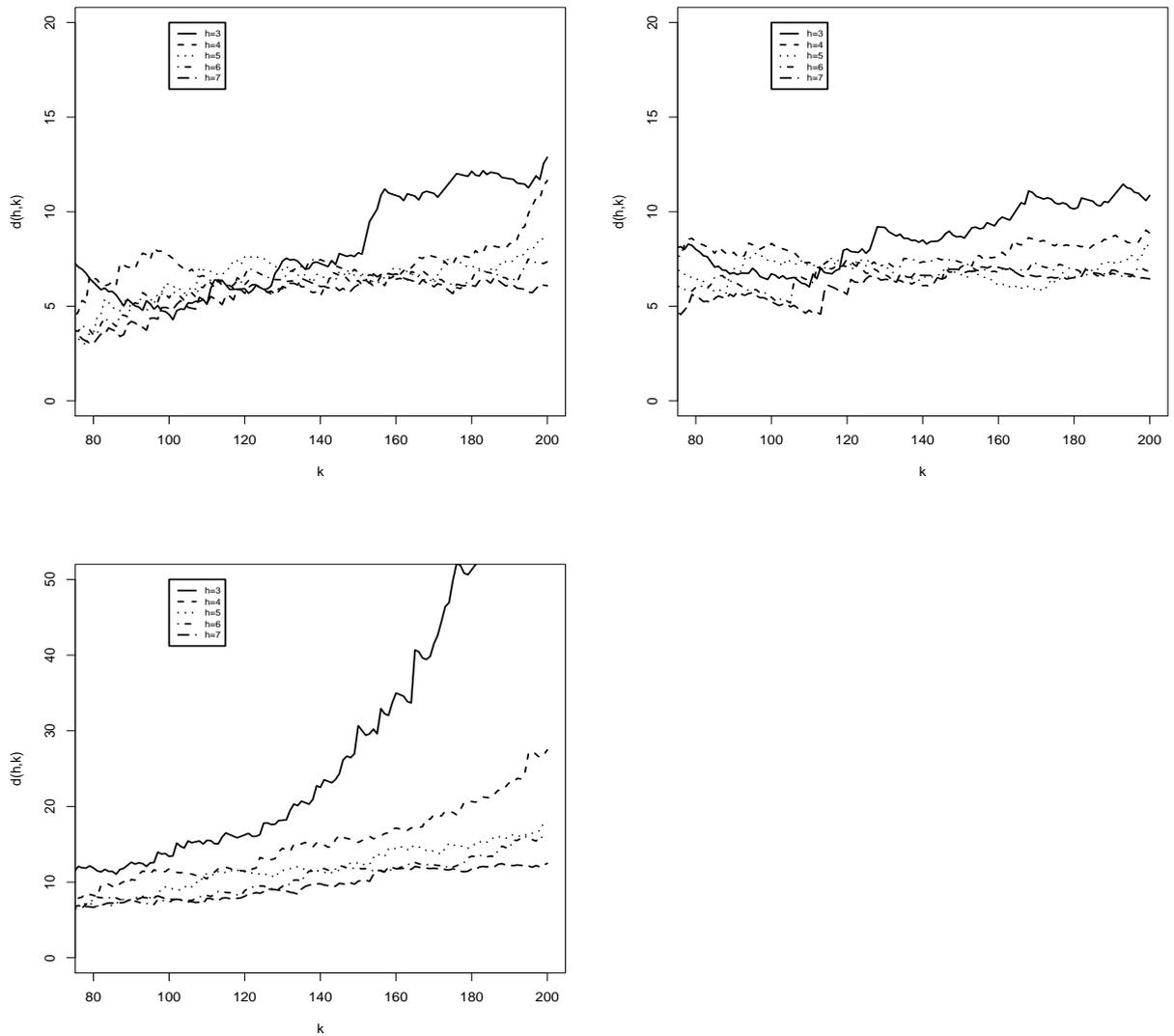


FIGURE 2.12 – Choix de (h_x^*, k_x^*) . Graphe de la fonction $k \mapsto |d_{h,k}(x)|$ (pour $h \in \mathcal{H}$) quand $x = 27$ (à gauche), $x = 37$ (à droite) et $x = 47$ (en bas) : Cas-Ignoré.

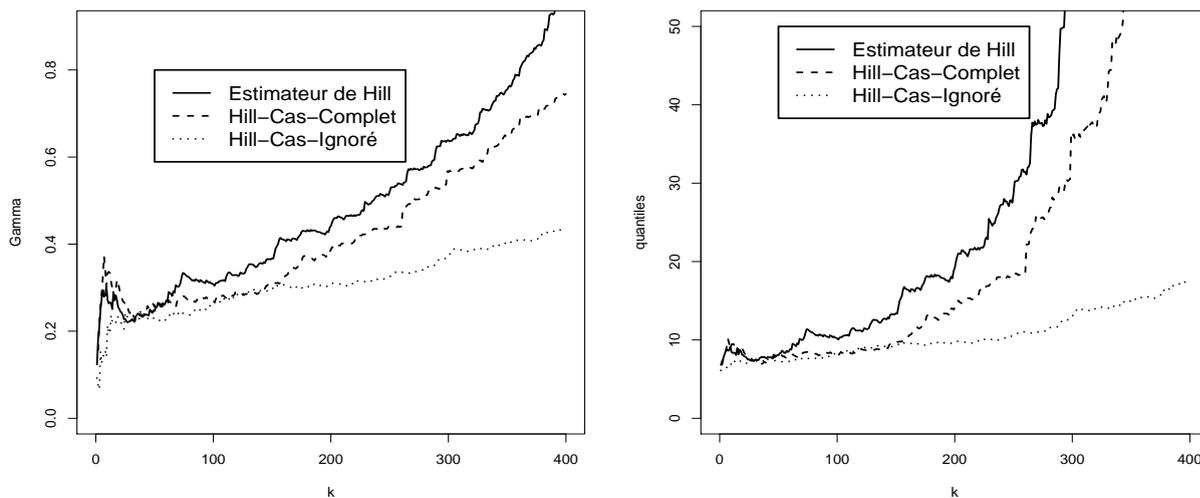


FIGURE 2.13 – Application sur les données du SIDA. Estimateur de Hill $\gamma_1(27)$ (estimateurs proposés dans (2.1), Cas-Complet et Cas-Ignoré) (à gauche) et $q(1/1000, 27)$ (estimateurs proposés dans (2.3), Cas-Complet et Cas-Ignoré) (à droite).

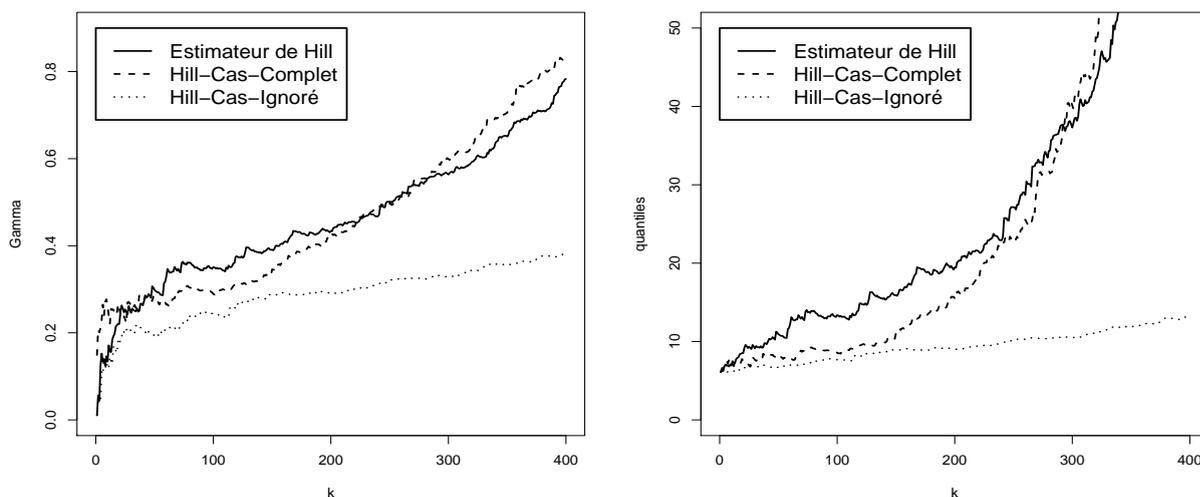


FIGURE 2.14 – Application sur les données du SIDA. Estimateur de Hill $\gamma_1(37)$ (estimateurs proposés dans (2.1), Cas-Complet et Cas-Ignoré) (à gauche) et $q(1/1000, 37)$ (estimateurs proposés dans (2.3), Cas-Complet et Cas-Ignoré) (à droite).

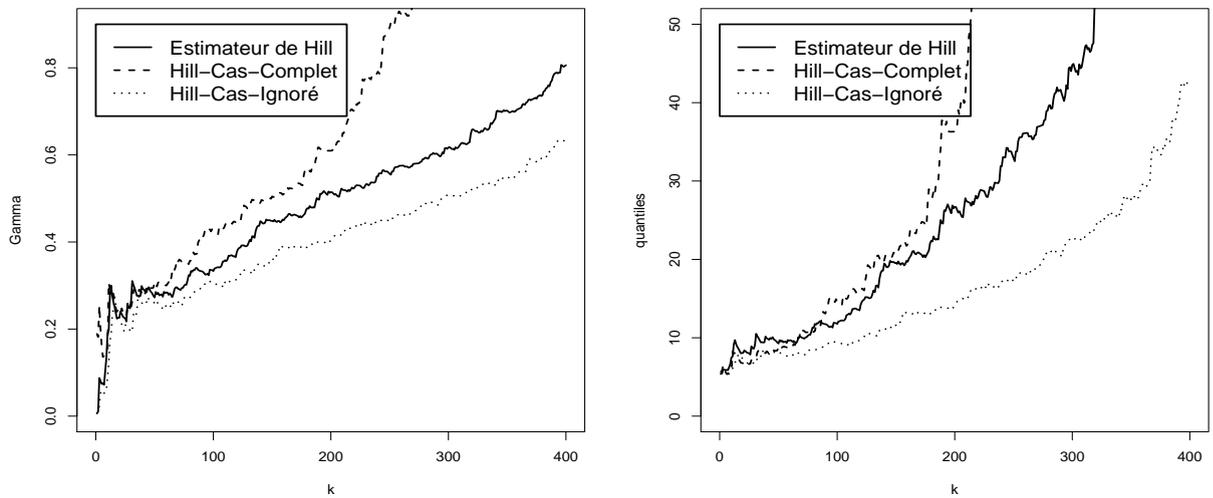


FIGURE 2.15 – Application sur les données du SIDA. Estimateur de Hill $\gamma_1(47)$ (estimateurs proposés dans (2.1), Cas-Complet et Cas-Ignoré) (à gauche) et $q(1/1000, 47)$ (estimateurs proposés dans (2.3), Cas-Complet et Cas-Ignoré) (à droite).

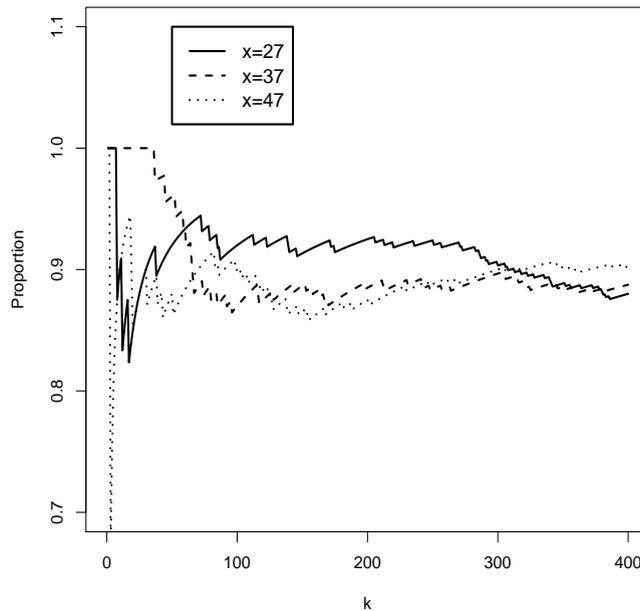


FIGURE 2.16 – Estimation de la proportion des données non censurées pour les k plus grandes valeurs de Z qui sont proches de x ($x = 27, 37, 47$).

2.6 Discussion et perspectives

Dans ce chapitre, nous avons proposé un estimateur de l'indice de queue conditionnel en présence de données censurées aléatoirement à droite et celui de son quan-

tile extrême conditionnel correspondant. Puis nous avons illustré leur comportement par des simulations. Dans l'illustration, nous avons voulu connaître l'effet du taux de censure sur les estimations, l'influence du type de l'estimateur standard pris enfin vérifier numériquement les propriétés asymptotiques des estimateurs proposés.

2.6.1 Taux de censure sur les estimations

Le taux de censure mesure le degré de données censurées ici aléatoirement à droite. Les estimations montrent qu'au fur et à mesure que l'on augmente le pourcentage de censure, les estimations se dégradent en terme d'erreur quadratique moyenne (MSE) et d'erreur absolue moyenne (MAE). Ainsi, si l'on considère les figures 2.3 et 2.4, les courbes en couleur s'éloignent de la vraie valeur si le taux de censure est augmenté. Ce qui témoigne que la censure a un effet considérable sur les estimations. Les erreurs quadratiques moyennes, erreurs absolues moyennes, longueurs des intervalles, les probabilités de couvertures des intervalles de confiance augmentent au fur et à mesure que le pourcentage de censure passe de 10%, de 25% et de 40% dans les tableaux 2.1, 2.2, 2.3 et 2.4.

2.6.2 Effet de l'estimateur standard

Les estimateurs que nous avons proposés dépendent d'un estimateur standard (hill, moments, uh, pickands, etc.). Ainsi, nous avons jugé important de comparer différents types d'estimateurs. Les estimateurs que nous avons voulu comparer dans les simulations sont : l'estimateur de Hill, l'estimateur des moments et l'estimateur uh. Les figures 2.3 et 2.4 révèlent que les estimations faites avec l'estimateur de Hill donnent de meilleures performances par rapport aux estimations faites avec les estimateurs moments et uh. D'ailleurs numériquement les tableaux 2.1, 2.2, 2.3 et 2.4 donnent une place importante en terme de MSE, MAE et probabilité de couverture des intervalles de confiance, l'estimateur de Hill par rapport aux autres estimateurs.

2.6.3 Propriétés asymptotiques

Après la détermination théorique des propriétés asymptotiques des estimateurs, nous avons illustré leur comportement en simulant. Au niveau des simulations nous avons considéré plusieurs tailles d'échantillon (500, 1000, 1500 et 2000) pour vérifier numériquement la consistance et la normalité asymptotique. Et les résultats reportés dans les tableaux 2.1, 2.2, 2.3 et 2.4 révèlent que plus nous augmentons la taille de l'échantillon, plus nous nous approchons de la vraie valeur des paramètres à estimer d'où la consistance. Les probabilités de couverture des intervalles de confiance

calculées avec les normalités asymptotiques diminuent lorsque la taille de l'échantillon augmente. Les erreurs quadratiques moyennes (MSE) et les erreurs absolues moyennes diminuent constamment lorsque la taille de l'échantillon augmente. Ce qui témoigne les résultats théoriques que nous avons obtenus dans les Théorèmes 2.4.1 et 2.4.2.

2.6.4 Perspectives

Dans ce chapitre, nous avons utilisé l'idée du plus proche voisin appelée méthode de la fenêtre mobile basée sur les boules de centre x et de rayon $h_{n,x}$. Nous avons choisi le rayon de la boule arbitrairement au niveau des simulations. Il serait intéressant de trouver des méthodes de calcul pour le choix de ce rayon $h_{n,x}$. Nous pouvons aussi souligner un autre problème d'ordre pratique; en pratique la vraie valeur n'est pas connue et que pour le choix du k_x nous avons choisi celui qui minimise les MSE qui suppose la connaissance de la vraie valeur. Dans le futur il serait aussi intéressant d'élaborer des méthodes de calculs sur le choix du k_x dans le cas conditionnel.

2.7 Preuves des théorèmes

Dans ce paragraphe, nous allons démontrer les théorèmes sur les résultats théoriques.

Preuve du théorème 2.4.1

Pour démontrer le théorème 2.4.1, nous allons considérer la décomposition suivante

$$\begin{aligned} \sqrt{k_x} \left(\hat{\gamma}_{k_x, m_{n,x}}^{(c, \cdot)}(x) - \gamma_1(x) \right) &= \frac{1}{\hat{p}_x} \sqrt{k_x} \left(\hat{\gamma}_{k_x, m_{n,x}}^{(\cdot)}(x) - \gamma(x) \right) + \frac{1}{\hat{p}_x} \sqrt{k_x} (\gamma(x) - \gamma_1(x) \hat{p}_x) \\ &= \frac{1}{\hat{p}_x} \sqrt{k_x} \left(\hat{\gamma}_{k_x, m_{n,x}}^{(\cdot)}(x) - \gamma(x) \right) - \frac{\gamma_1(x)}{\hat{p}_x} \sqrt{k_x} (\hat{p}_x - p_x). \end{aligned}$$

Or, par hypothèse,

$$\sqrt{k_x} \left(\hat{\gamma}_{k_x, m_{n,x}}^{(\cdot)}(x) - \gamma(x) \right) \xrightarrow{d} \mathcal{N} \left(m(x) \lambda(x), \sigma^2(x) \right)$$

que nous pouvons diviser par p_x pour obtenir la normalité asymptotique du premier terme de la décomposition

$$\frac{\sqrt{k_x}}{\hat{p}_x} \left(\hat{\gamma}_{k_x, m_{n,x}}^{(\cdot)}(x) - \gamma(x) \right) \xrightarrow{d} \mathcal{N} \left(\frac{m(x) \lambda(x)}{p_x}, \frac{\sigma^2(x)}{p_x^2} \right).$$

Prouvons maintenant la deuxième partie de la décomposition. Ainsi, cette deuxième partie sera montrée par construction parce que nous ne pouvons pas utiliser le théorème central limite car le \hat{p}_x est calculé à partir d'observations ordonnées. Nous procédons comme dans Einmahl et al. (2008, [44]) mais dans le cas conditionnel.

Soit $U[0, 1]$ une variable uniforme sur $[0, 1]$ indépendante de Z^x définie par

$$\delta^x = \begin{cases} 1, & \text{si } U \leq p_x(Z^x) \\ 0, & \text{si } U > p_x(Z^x) \end{cases} \quad (2.8)$$

et

$$\tilde{\delta}^x = \begin{cases} 1, & \text{si } U \leq p_x \\ 0, & \text{si } U > p_x. \end{cases} \quad (2.9)$$

On répète ce dispositif de construction jusqu'à la taille n . Nous obtenons l'échantillon $(Z_{(i)}^x, \delta_{(i)}^x)_{1 \leq i \leq n}$ qui a la même distribution que celle de l'échantillon de départ dans lequel nous appliquons la méthode de sélection, fenêtre mobile de taille $m_{n,x}$. Donc, Z^x et δ^x sont indépendantes et nous pouvons écrire

$$\mathbb{P}(|\delta^x - \tilde{\delta}^x| = 1 | Z^x = z, X = x) = |p_x - p_x(z)|.$$

Nous notons $U_{(1)}, \dots, U_{(m_{n,x})}$ l'échantillon ordonné associé à U et nous pouvons écrire

$$\hat{p}_x = \frac{1}{k_x} \sum_{i=1}^{k_x} \mathbb{1}_{\{U_{(m_{n,x}-i+1)} \leq p_x(Z_{(m_{n,x}-i+1)}^x)\}}$$

et

$$\tilde{p}_x = \frac{1}{k_x} \sum_{i=1}^{k_x} \tilde{\delta}_{(m_{n,x}-i+1)}^x = \frac{1}{k_x} \sum_{i=1}^{k_x} \mathbb{1}_{\{U_{(m_{n,x}-i+1)} \leq p_x\}}.$$

Nous décomposons à nouveau :

$$\sqrt{k_x}(\hat{p}_x - p_x) = \sqrt{k_x}(\hat{p}_x - \tilde{p}_x) + \sqrt{k_x}(\tilde{p}_x - p_x).$$

Or,

$$\tilde{p}_x \stackrel{d}{=} \sum_{i=1}^{k_x} \mathbb{1}_{\{U_i \leq p_x\}}$$

car les $U_{(1)}, \dots, U_{(m_{n,x})}$ sont indépendantes et identiquement distribuées et nous pouvons appliquer le théorème central limite pour obtenir :

$$\sqrt{k_x}(\tilde{p}_x - p_x) \xrightarrow{d} \mathcal{N}(0, p_x(1 - p_x)).$$

Nous calculons maintenant $\sqrt{k_x}(\hat{p}_x - \tilde{p}_x)$ en remplaçant chaque terme par sa valeur :

$$\sqrt{k_x}(\hat{p}_x - \tilde{p}_x) \stackrel{d}{=} \frac{1}{\sqrt{k_x}} \sum_{i=1}^{k_x} \left[\mathbb{1}_{\{U_i \leq p_x(Z_{(m_{n,x}^{x-i+1})}^x)\}} - \mathbb{1}_{\{U_i \leq p_x\}} \right]. \quad (2.10)$$

Nous incorporons le terme $\mathbb{1}_{\{U_i \leq p_x(H^{\leftarrow}\left(1 - \frac{i}{m_{n,x}}|x\right))\}}$ dans l'équation (2.10) :

$$\begin{aligned} \sqrt{k_x}(\hat{p}_x - \tilde{p}_x) &= \frac{1}{\sqrt{k_x}} \sum_{i=1}^{k_x} \left[\mathbb{1}_{\{U_i \leq p_x(Z_{(m_{n,x}^{x-i+1})}^x)\}} - \mathbb{1}_{\{U_i \leq p_x(H^{\leftarrow}\left(1 - \frac{i}{m_{n,x}}|x\right))\}} \right] \\ &\quad + \frac{1}{\sqrt{k_x}} \sum_{i=1}^{k_x} \left[\mathbb{1}_{\{U_i \leq p_x(H^{\leftarrow}\left(1 - \frac{i}{m_{n,x}}|x\right))\}} - \mathbb{1}_{\{U_i \leq p_x\}} \right]. \end{aligned}$$

Nous obtenons la nouvelle décomposition suivante :

$$\sqrt{k_x}(\hat{p}_x - \tilde{p}_x) =: T_{1,k_x} + T_{2,k_x}.$$

Sous les conditions **C5** et **C6**, l'expression T_{2,k_x} converge vers un réel $\varepsilon(x)$ en utilisant les résultats de Chow et Teicher (1997, page 356, [20]). Il reste maintenant à montrer $T_{1,k_x} \xrightarrow{\mathbb{P}} 0$. Pour terminer soit $V_i^x = H(Z_i^x|x)$ donc $Z_i^x = H^{\leftarrow}(V_i^x|x)$ où H^{\leftarrow} représente l'inverse généralisée de H . Soit maintenant $r_x(t) = p_x(H^{\leftarrow}(t|x))$. Et nous pouvons aussi écrire,

$$T_{1,k_x} = \frac{1}{\sqrt{k_x}} \sum_{i=1}^{k_x} \left[\mathbb{1}_{\{U_i \leq r_x(V_{(m_{n,x}^{x-i+1})}^x)\}} - \mathbb{1}_{\{U_i \leq r_x\left(1 - \frac{i}{m_{n,x}}\right)\}} \right].$$

Or, le résultat sur la convergence de la queue uniforme des processus de quantiles donne pour tout $1 \leq i \leq k_x$

$$V_{(m_{n,x}^{x-i+1})}^x - \left(1 - \frac{i}{m_{n,x}}\right) = O_{\mathbb{P}}\left(\frac{\sqrt{k_x}}{m_{n,x}}\right).$$

Soit $\eta > 0$. En utilisant la condition **C6**, nous obtenons pour $m_{n,x}$ assez grand,

$$\begin{aligned}
 |T_{1,k_x}| &\leq \frac{1}{\sqrt{k_x}} \sum_{i=1}^{k_x} \left| \mathbb{1}_{\left\{U_i \leq r_x \left(V_{(m_{n,x}-i+1)}^x\right)\right\}} - \mathbb{1}_{\left\{U_i \leq r_x \left(1 - \frac{i}{m_{n,x}}\right)\right\}} \right| \\
 &\stackrel{d}{=} \frac{1}{\sqrt{k_x}} \sum_{i=1}^{k_x} \mathbb{1}_{\left\{U_i \leq \left| r_x \left(V_{(m_{n,x}-i+1)}^x\right) - r_x \left(1 - \frac{i}{m_{n,x}}\right) \right|\right\}} \\
 &\leq \frac{1}{\sqrt{k_x}} \sum_{i=1}^{k_x} \mathbb{1}_{\left\{U_i \leq \frac{\eta}{\sqrt{k_x}}\right\}} \\
 &\leq \frac{1}{\sqrt{k_x}} \sum_{i=1}^{k_x} \mathbb{1}_{\left\{\frac{U_i \sqrt{k_x}}{\eta} \leq 1\right\}}.
 \end{aligned}$$

Soit $T_i = \frac{U_i \sqrt{k_x}}{\eta}$ pour tout $1 \leq i \leq k_x$ donc T_i est une variable uniforme sur $\left[0, \frac{\sqrt{k_x}}{\eta}\right]$. Donc, $\mathbb{E}(T_i) = \frac{\sqrt{k_x}}{2\eta}$ et

$$|T_{1,k_x}| \leq \frac{1}{\sqrt{k_x}} \sum_{i=1}^{k_x} \mathbb{E}(T_i) \mathbb{1}_{\{|T_i| \leq 1\}}.$$

Le résultat dans le livre de Chow et Teicher ([20], page 356) et le choix d'un η petit permettent d'avoir

$$T_{1,k_x} \xrightarrow{\mathbb{P}} 0.$$

Ainsi donc,

$$\begin{aligned}
 \sqrt{k_x} \left(\widehat{\gamma}_{k_x, m_{n,x}}^{(c, \cdot)}(x) - \gamma_1(x) \right) &= \frac{1}{\widehat{p}_x} \left(\sqrt{k_x} \left(\widehat{\gamma}_{k_x, m_{n,x}}^{(\cdot)}(x) - \gamma(x) \right) - \gamma_1(x) \sqrt{k_x} (\widehat{p}_x - p_x) \right) \\
 &\quad - \frac{\gamma_1(x) \varepsilon(x)}{\widehat{p}_x} + o_{\mathbb{P}}(1).
 \end{aligned}$$

Ce qui prouve la relation,

$$\sqrt{k_x} \left(\widehat{\gamma}_{k_x, m_{n,x}}^{(c, \cdot)}(x) - \gamma_1(x) \right) \xrightarrow{d} \mathcal{N} \left(\frac{1}{p_x} (\lambda(x) m(x) - \gamma_1(x) \varepsilon(x)), \frac{\sigma^2(x) + \gamma_1(x)^2 p_x (1 - p_x)}{p_x^2} \right).$$

□

Preuve du corollaire 2.4.1

Pour alléger les écritures, nous notons dans cette partie, h_x pour $h_{n,x}$, b_x pour $b\left(\frac{m_{n,x}}{k_x}, x\right)$, m_x pour $m_{n,x}$. La preuve du corollaire 2.4.1 sera basée sur le Théorème 2.4.1 et sur les travaux existants dans la littérature. Il s'agit de montrer

$$\sqrt{k_x} \left(\widehat{\gamma}_{k_x, m_x}^{(\cdot)}(x) - \gamma(x) \right) \xrightarrow{d} \mathcal{N} \left(m(x)\lambda(x), \sigma^2(x) \right)$$

pour les estimateurs de Hill, Moment et UH adaptés au cas conditionnel.

• Hill Conditionnel

Il s'agit de prouver

$$\sqrt{k_x} \left(\widehat{\gamma}_{k_x, m_x}^{(H)}(x) - \gamma(x) \right) \xrightarrow{d} \mathcal{N} \left(\frac{\lambda(x)}{1 - \rho(x)}, \gamma^2(x) \right).$$

En effet, sous la condition **C3**, le Théorème 1 dans Gardes et Girard (2008, [51]) donne

$$\begin{aligned} k_x \widehat{\gamma}_{k_x, m_x}^{(H)}(x) &\stackrel{d}{=} \gamma(x) (1 + O_p(h_x^{\alpha_U})) \sum_{i=1}^{k_x} F_i \\ &\quad + (1 + O_p(h_x^{\alpha_U})) b_x \sum_{i=1}^{k_x} \left(\frac{i}{k_x + 1} \right)^{-\rho(x)} F_i \\ &\quad + (1 + O_p(h_x^{\alpha_U})) \sum_{i=1}^{k_x} \beta_{i, m_x}(x) \\ &\quad + k_x O_p(b_x) \end{aligned}$$

où les F_i sont des v.a. exponentielles standards, les β_{i, m_x} sont des v.a. vérifiant $\frac{1}{k_x} \sum_{i=1}^{k_x} \beta_{i, m_x} = o_p(b_x)$. Notons $T_{1,n} := \sum_{i=1}^{k_x} (F_i - 1)$, $T_{2,n} := \sum_{i=1}^{k_x} \left(\frac{i}{k_x + 1} \right)^{-\rho(x)} (F_i - 1)$,

$T_{3,n} := \sum_{i=1}^{k_x} \beta_{i, m_x}(x)$, $T_{4,n} := b_x \sum_{i=1}^{k_x} \left(\frac{i}{k_x + 1} \right)^{-\rho(x)}$, $T_{5,n} := k_x$, $T_{6,n} := k_x$, $T_{7,n} := \sqrt{k_x}$.

Nous obtenons la relation suivante :

$$\begin{aligned} \frac{T_{5,n}}{T_{7,n}} \left(\widehat{\gamma}_{k_x, m_x}^{(H)}(x) - \gamma(x) - \frac{T_{4,n}}{T_{5,n}} \right) &\stackrel{d}{=} \left(\gamma(x) \frac{T_{1,n}}{T_{7,n}} + b_x \frac{T_{2,n}}{T_{7,n}} + \frac{T_{3,n}}{T_{7,n}} \right) (1 + O_p(h_x^{\alpha_U})) \\ &\quad + \left(\frac{T_{4,n}}{T_{7,n}} + \frac{T_{5,n}}{T_{7,n}} \right) O_p(h_x^{\alpha_U}) + \frac{T_{6,n}}{T_{7,n}} O_p(b_x). \end{aligned} \quad (2.11)$$

Or, pour $\delta > 0$, $\frac{k_x}{T_{7,n}^{2+\delta}} = \frac{1}{k_x^{\delta/2}} \rightarrow 0$ donc le théorème de Lindeberg nous donne

$$\frac{T_{1,n}}{T_{7,n}} = \frac{1}{\sqrt{k_x}} \sum_{i=1}^{k_x} (F_i - 1) \xrightarrow{d} \mathcal{N}(0, 1). \quad (2.12)$$

Et $\frac{T_{2,n}}{T_{7,n}} = \frac{1}{\sqrt{k_x}} \sum_{i=1}^{k_x} \left(\frac{i}{k_x+1}\right)^{-\rho(x)} (F_i - 1)$ est centré et admettant une variance finie donc

$$\frac{T_{2,n}}{T_{7,n}} = O_p(1). \quad (2.13)$$

Le Théorème 1 dans Gardes et Girard (2008, [51]) donne $\frac{1}{k_x} \sum_{i=1}^{k_x} \beta_{i,m_x}(x) = o_p(b_x)$ donc

$$\frac{T_{3,n}}{T_{7,n}} = \frac{1}{\sqrt{k_x}} \sum_{i=1}^{k_x} \beta_{i,m_x}(x) = o_p(\sqrt{k_x} b_x) = o_p(1). \quad (2.14)$$

Nous avons aussi sous la condition de l'intégrale de Riemann (valeur moyenne d'une fonction)

$$\lim_{k_x \rightarrow \infty} \frac{1}{k_x} \sum_{i=1}^{k_x} \left(\frac{1}{k_x+1}\right)^{-\rho(x)} = \int_0^1 s^{-\rho(x)} ds = \frac{1}{1-\rho(x)}$$

qui donne la relation suivante

$$\frac{T_{4,n}}{T_{5,n}} = b_x \frac{1}{k_x} \sum_{i=1}^{k_x} \left(\frac{i}{k_x+1}\right)^{-\rho(x)} = b_x \frac{1}{1-\rho(x)} (1 + o(1)) \quad (2.15)$$

et

$$\frac{T_{4,n}}{T_{7,n}} = \sqrt{k_x} b_x \times \frac{1}{k_x} \sum_{i=1}^{k_x} \left(\frac{i}{k_x+1}\right)^{-\rho(x)} = O(k_x^{1/2} b_x) = O(1), \quad (2.16)$$

$$\frac{T_{5,n}}{T_{7,n}} = \frac{T_{6,n}}{T_{7,n}} = k_x^{1/2} (1 + o(1)). \quad (2.17)$$

Remplaçons les équations (2.12)-(2.17) dans (2.11) :

$$k_x^{1/2} \left(\widehat{\gamma}_{k_x, m_x}^{(H)}(x) - \gamma(x) - b_x \frac{1}{1-\rho(x)} \right) \stackrel{d}{=} \gamma(x) \frac{T_{1,n}}{T_{7,n}} + O_p(k_x^{1/2} h_x^{\alpha_U}) + o_p(1) \quad (2.18)$$

d'où le résultat d'après (2.12).

• UH Conditionnel

L'estimateur UH s'écrit ainsi,

$$\begin{aligned} \widehat{\gamma}_{k_x, m_x}^{(UH)}(x) &= \frac{1}{k_x} \sum_{i=1}^{k_x} \log UH_{(i)}^x - \log UH_{(k+1)}^x \\ &= \sum_{i=1}^{k_x} i \log \frac{UH_{(i)}^x}{UH_{(i+1)}^x} \\ &= \frac{1}{k_x} \sum_{i=1}^{k_x} \log \left(Z_{(m_x-i)}^x \widehat{\gamma}_{i, m_x}^{(H)}(x) \right) - \log \left(Z_{(m_x-k_x)}^x \widehat{\gamma}_{k_x, m_x}^{(H)}(x) \right) \end{aligned}$$

où $UH_{(j)}^x = Z_{(j)}^x \left(\frac{1}{j} \sum_{i=1}^j \log Z_{(m_x-i+1)}^x - \log Z_{(m_x-i)}^x \right) = Z_{(j)}^x \left(\widehat{\gamma}_{j, m_x}^{(H)}(x) \right)$.

Dans Beirlant et al. (2005, [9], page 956) et le Théorème 1 dans Gardes et Girard

(2008, [51]), nous avons la relation adaptée au cas conditionnel suivant :

$$\log UH_{(j)}^x \stackrel{d}{=} \left(\gamma(x) \log U_{(k+1)}^{-1} + \log \left(l_+ \left(\frac{k_x}{j} \right)^{\gamma(x)} \right) + \left[\frac{1}{\sqrt{k_x}} \frac{N_{k_x, m_x} \left(\frac{j}{k_x} \right)}{\frac{j}{k_x}} + b_x \left(\frac{j}{k_x} \right)^{-\rho(x)} \right] (1 + o_p(1)) \right) (1 + O_p(h_x^{\alpha_U}))$$

où l_+ est une constante positive et $U_{(1)}, \dots, U_{(m_x)}$ est la statistique d'ordre des v.a. uniformes sur $]0, 1[$,

$$N_{k_x, m_x} \left(\frac{j}{k_x} \right) = P_{k_x, m_x}^{(2)} \left(\frac{j}{k_x} \right) + \frac{j}{k_x} \left[P_{m_x}^{(1)} \left(\frac{j}{m_x} \right) - P_{m_x}^{(1)} \left(\frac{k_x}{m_x} \right) \right]$$

avec

$$P_{m_x}^{(1)} \left(\frac{j}{m_x} \right) = \gamma(x) \sqrt{k_x} \left(\log U_{(j+1)}^{-1} - \log \left(\frac{m_x}{j} \right) \right)$$

et

$$P_{k_x, m_x}^{(2)} \left(\frac{j}{k_x} \right) = \sqrt{k_x} \left(\frac{j}{k_x} \right) \left\{ \frac{1}{j} \sum_{i=1}^j \log \left(\frac{U_{(j+1)}}{U_{(i)}} \right) - 1 \right\}, \quad 1 \leq j \leq k_x.$$

Dans l'Appendice B de Beirlant et al. (2005, [9]), nous avons

$$\gamma(x) \left[P_{m_x}^{(1)} \left(\frac{j}{m_x} \right) - P_{m_x}^{(1)} \left(\frac{k_x}{m_x} \right) \right] \approx \frac{W \left(\frac{j}{k_x} \right)}{\frac{j}{k_x}} - W(1)$$

et $P_{k_x, m_x}^{(2)} \left(\frac{j}{k_x} \right) \approx \int_0^{k_x} \frac{W(s)}{s} ds - W(t)$ où $(W(t))_{t \geq 1}$ est un processus de Wiener. Par conséquent $\{N_{k_x, m_x}(t); t \in]0, 1[\}$ est représenté asymptotiquement par

$$\int_0^t \frac{W(s)}{s} ds + (\gamma(x) - 1) W(t) - \gamma(x) t W(1) := W^0(t) - \gamma(x) t W(1).$$

Ce qui implique que

$$W^0(t) = \int_0^t \frac{W(s)}{s} ds + (\gamma(x) - 1) W(t).$$

Calculons la covariance suivante :

$$\text{Cov} \left(W^0(s), W^0(t) \right) = \mathbb{E} \left(W^0(t) W^0(s) \right) - \mathbb{E} \left(W^0(t) \right) \mathbb{E} \left(W^0(s) \right) \quad \text{avec } s \leq t.$$

Or,

$$\mathbb{E} \left(W^0(t) \right) = \int_0^t \frac{\mathbb{E} [W(u)]}{u} du - (\gamma(x) - 1) \mathbb{E} [W(t)] = 0$$

donc

$$\begin{aligned} \text{Cov} \left(W^0(s), W^0(t) \right) &= \mathbb{E} \left(W^0(t) W^0(s) \right) \\ &= \mathbb{E} \left[\left(\int_0^t \frac{W(u)}{u} du + (\gamma(x) - 1) W(t) \right) \left(\int_0^s \frac{W(v)}{v} dv + (\gamma(x) - 1) W(s) \right) \right] \\ &= \mathbb{E} \left[\int_0^t \frac{W(u)}{u} du \int_0^s \frac{W(v)}{v} dv \right] \\ &\quad + \mathbb{E} \left[(\gamma(x) - 1) \int_0^t \frac{W(s)W(u)}{u} du \right] \\ &\quad + \mathbb{E} \left[(\gamma(x) - 1) \int_0^s \frac{W(t)W(v)}{v} dv \right] \\ &\quad + \mathbb{E} \left[(\gamma(x) - 1)^2 W(t)W(s) \right] \\ &:= A + B + C + D. \end{aligned}$$

Ce qui donne en utilisant le fait que $(W(t))_{0 < t < 1}$ est un processus de Wiener,

$$D = \mathbb{E} \left[(\gamma(x) - 1)^2 W(t)W(s) \right] = (\gamma(x) - 1)^2 \mathbb{E} [W(t)W(s)] = (\gamma(x) - 1)^2 s,$$

$$\begin{aligned} C &= \mathbb{E} \left[(\gamma(x) - 1) \int_0^s \frac{W(t)W(v)}{v} dv \right] = (\gamma(x) - 1) \mathbb{E} \left[\int_0^s \frac{W(t)W(v)}{v} dv \right] \\ &= (\gamma(x) - 1) \mathbb{E} \left[\int_0^s \frac{W(t)W(v)}{v} dv \right] \\ &= (\gamma(x) - 1) \int_0^s \frac{\mathbb{E} [W(t)W(v)]}{v} dv \\ &= (\gamma(x) - 1) \int_0^s dv = (\gamma(x) - 1) s, \end{aligned}$$

$$\begin{aligned} B &= \mathbb{E} \left[(\gamma(x) - 1) \int_0^t \frac{W(s)W(u)}{u} du \right] \\ &= (\gamma(x) - 1) \mathbb{E} \left[\int_0^t \frac{W(s)W(u)}{u} du \right] \\ &= (\gamma(x) - 1) \int_0^t \frac{1}{u} \mathbb{E} [W(s)W(u)] du \\ &= (\gamma(x) - 1) \left[\int_0^s \frac{1}{u} \mathbb{E} [W(s)W(u)] du - \int_s^t \frac{1}{u} \mathbb{E} [W(s)W(u)] du \right] \\ &= (\gamma(x) - 1) \left[\int_0^s du + \int_s^t \frac{s}{u} du \right] \\ &= (\gamma(x) - 1) s + (\gamma(x) - 1) s \times \log \left(\frac{t}{s} \right) \end{aligned}$$

et

$$\begin{aligned}
A &= \mathbb{E} \left[\int_0^t \frac{W(u)}{u} du \int_0^s \frac{W(v)}{v} dv \right] \\
&= \mathbb{E} \left[\left(\int_0^s \frac{W(v)}{v} dv \right) \left(\int_0^v \frac{W(u)}{u} du + \int_v^t \frac{W(u)}{u} du \right) \right] \\
&= \mathbb{E} \left[\int_0^s \frac{W(v)}{v} dv \int_0^v \frac{W(u)}{u} du + \int_0^s \frac{W(v)}{v} dv \int_v^t \frac{W(u)}{u} du \right] \\
&= \mathbb{E} \left[\int_0^s \frac{W(v)}{v} dv \int_0^v \frac{W(u)}{u} du \right] + \mathbb{E} \left[\int_0^s \frac{W(v)}{v} dv \int_v^t \frac{W(u)}{u} du \right] \\
&= \int_0^s \left(\int_0^v \frac{1}{uv} \mathbb{E} [W(u)W(v)] du \right) dv + \int_0^s \left(\int_v^t \frac{1}{uv} \mathbb{E} [W(u)W(v)] du \right) dv \\
&= \int_0^s \left(\int_0^v \frac{1}{v} du \right) dv + \int_0^s \left(\int_v^t \frac{1}{u} du \right) dv \\
&= \int_0^s dv + \int_0^s \left([\log u]_v^t \right) dv \\
&= 2s + s \log \left(\frac{t}{s} \right).
\end{aligned}$$

D'où

$$Cov(W^0(s), W^0(t)) = s \left[1 + \gamma^2(x) + \gamma(x) \log \left(\frac{t}{s} \right) \right]$$

et par conséquent $Var(W^0(t)) = t(1 + \gamma^2(x))$ et la variance asymptotique de $\hat{\gamma}_{k_x, m_x}^{(UH)}(x)$ est égale à $1 + \gamma^2(x)$. Et nous avons aussi par application de la valeur moyenne d'une fonction

$$\lim_{k_x \rightarrow \infty} \frac{1}{k_x} \sum_{i=1}^{k_x} \left(\frac{1}{k_x} \right)^{-\rho(x)} = \int_0^1 s^{-\rho(x)} ds = \frac{1}{1 - \rho(x)}$$

donc

$$\sqrt{k_x} b_x \frac{1}{k_x} \sum_{i=1}^{k_x} \left(\frac{1}{k_x} \right)^{-\rho(x)} = \frac{\lambda(x)}{1 - \rho(x)}$$

d'où

$$\sqrt{k_x} \left(\hat{\gamma}_{k_x, m_x}^{(UH)}(x) - \gamma(x) \right) \xrightarrow{d} \mathcal{N} \left(\frac{\lambda(x)}{1 - \rho(x)}, 1 + \gamma^2(x) \right).$$

• Moment Conditionnel

Pour la normalité asymptotique de l'estimateur des Moments Conditionnel, le lecteur pourra se référer à Stupfler (2013, Théorème 2, page 2309, [85]).

□

Preuve du Théorème 2.4.2

L'estimateur de Weissman (1978, [95]) adapté au cas conditionnel s'écrit :

$$\hat{q}^{(c, \cdot)}(\alpha_{m_{n,x}}, x) = Z_{(m_{n,x} - k_x)}^x \left(\frac{\beta_{m_{n,x}}}{\alpha_{m_{n,x}}} \right) \hat{\gamma}_{k_x, m_{n,x}}^{(c, \cdot)}(x)$$

donc

$$\log \widehat{q}^{(c,\cdot)}(\alpha_{m_{n,x}}, x) = \log Z_{(m_{n,x}-k_x)}^x + \widehat{\gamma}_{k_x, m_{n,x}}^{(c,\cdot)}(x) \log \left(\frac{\beta_{m_{n,x}}}{\alpha_{m_{n,x}}} \right)$$

Puis nous décomposons comme suit :

$$\begin{aligned} \log \left(\frac{\widehat{q}^{(c,\cdot)}(\alpha_{m_{n,x}}, x)}{q(\alpha_{m_{n,x}}, x)} \right) &= \log \left(\frac{Z_{(m_{n,x}-k_x)}^x}{q(\beta_{m_{n,x}}, x)} \right) \\ &+ \log \left(\frac{\beta_{m_{n,x}}}{\alpha_{m_{n,x}}} \right) \left(\widehat{\gamma}_{k_x, m_{n,x}}^{(c,\cdot)}(x) - \gamma_1(x) \right) \\ &- \log \frac{q(\alpha_{m_{n,x}}, x)}{q(\beta_{m_{n,x}}, x)} + \gamma_1(x) \log \left(\frac{\beta_{m_{n,x}}}{\alpha_{m_{n,x}}} \right) \end{aligned}$$

Notons,

$$\xi_{1, m_{n,x}} = \log \left(\frac{Z_{(m_{n,x}-k_x)}^x}{q(\beta_{m_{n,x}}, x)} \right)$$

$$\xi_{2, m_{n,x}} = \log \left(\frac{\beta_{m_{n,x}}}{\alpha_{m_{n,x}}} \right) \left(\widehat{\gamma}_{k_x, m_{n,x}}^{(c,\cdot)}(x) - \gamma_1(x) \right)$$

et

$$\xi_{3, m_{n,x}} = \log \frac{q(\alpha_{m_{n,x}}, x)}{q(\beta_{m_{n,x}}, x)} - \gamma_1(x) \log \left(\frac{\beta_{m_{n,x}}}{\alpha_{m_{n,x}}} \right)$$

donc

$$\log \left(\frac{\widehat{q}^{(c,\cdot)}(\alpha_{m_{n,x}}, x)}{q(\alpha_{m_{n,x}}, x)} \right) = \xi_{1, m_{n,x}} + \xi_{2, m_{n,x}} - \xi_{3, m_{n,x}}.$$

Par un calcul simple nous obtenons

$$\xi_{3, m_{n,x}} = - \int_{\alpha_{m_{n,x}}}^{\beta_{m_{n,x}}} \frac{\Delta(u, x)}{u} du$$

donc

$$\frac{k_x^{1/2}}{\log \left(\frac{\beta_{m_{n,x}}}{\alpha_{m_{n,x}}} \right)} |\xi_{3, m_{n,x}}| \leq \overline{\Delta}(\beta_{m_{n,x}}, x) k_x^{1/2} \rightarrow 0$$

sous la condition additionnelle $k_x^{1/2} \max\{\xi_{m_{n,x}}^{-1}, \overline{\Delta}(\beta_{m_{n,x}}, x)\} \rightarrow 0$ et nous avons $m_{n,x} \beta_{m_{n,x}} \rightarrow \infty$ quand $m_{n,x} \rightarrow \infty$, $(m_{n,x} \beta_{m_{n,x}})^{1/2} \xi_{1, m_{n,x}}$ est donc gaussienne d'après (Proposition 1.2.6 dans la partie rappel), ce qui entraîne,

$$\frac{k_x^{1/2}}{\log(\beta_{m_{n,x}}/\alpha_{m_{n,x}})} \xi_{1, m_{n,x}} = k_x^{1/2} \xi_{m_{n,x}}^{-1} (m_{n,x} \beta_{m_{n,x}})^{1/2} \xi_{1, m_{n,x}} \xrightarrow{\mathbb{P}} 0. \quad (2.19)$$

Nous obtenons aussi,

$$\frac{k_x^{1/2}}{\log(\beta_{m_{n,x}}/\alpha_{m_{n,x}})} \xi_{2, m_{n,x}} = k_x^{1/2} \left(\widehat{\gamma}_{k_x, m_{n,x}}^{(c,\cdot)}(x) - \gamma_1(x) \right).$$

En utilisant le Théorème 2.4.1, on a :

$$\frac{k_x^{1/2}}{\log(\beta_{m_n,x}/\alpha_{m_n,x})} \xi_{2,m_n,x} \xrightarrow{d} \mathcal{N}\left(\frac{1}{p_x} (\lambda(x)m(x) - \gamma_1(x)\varepsilon(x)), \frac{\sigma^2(x) + \gamma_1^2(x)p_x(1-p_x)}{p_x^2}\right). \quad (2.20)$$

Enfin, nous utilisons à nouveau la condition additionnelle pour obtenir ,

$$\frac{k_x^{1/2}}{\log(\beta_{m_n,x}/\alpha_{m_n,x})} |\xi_{3,m_n,x}| \leq \overline{\Delta}(\beta_{m_n,x}, x) k_x^{1/2} \rightarrow 0. \quad (2.21)$$

Les équations (2.19), (2.20) et (2.21) terminent la démonstration du Théorème 2.4.2.

□

Estimation des quantiles extrêmes conditionnels en design aléatoire en présence de censure

Résumé

Dans ce chapitre, nous allons présenter des estimateurs de l'indice de queue conditionnel et de quantiles extrêmes conditionnels en présence de censure et de covariable aléatoire où les fonctions de répartition conditionnelles sont à queues lourdes. Ce problème d'estimation est très complexe car la fonction de survie conditionnelle n'est pas connue et nous utilisons l'estimateur de la fonction de survie conditionnelle proposé par Beran en 1981. Enfin nous étudierons les propriétés asymptotiques de l'indice avant d'illustrer son comportement par simulation.

Sommaire

3.1 Introduction	81
3.2 Construction de l'estimateur et propriétés asymptotiques	83
3.2.1 Définition de l'estimateur	83
3.2.2 Résultats asymptotiques	85
3.3 Estimation de quantiles extrêmes conditionnels	87
3.4 Expériences numériques	88
3.4.1 Génération des données	88
3.4.2 Résultats des simulations	89
3.4.3 Estimation des quantiles extrêmes conditionnels	90
3.4.4 Comparaison	91
3.5 Conclusion et Perspectives	91
3.6 Démonstrations	107
3.6.1 Résultats préliminaires	107
3.6.2 Démonstrations des principaux résultats	113

3.1 Introduction

DANS ce chapitre, nous considérons le problème de l'estimation de l'indice des valeurs extrêmes conditionnel et des quantiles extrêmes conditionnels en présence de censure aléatoire à droite à design aléatoire. L'estimation des quantiles extrêmes est devenue un enjeu crucial dans de nombreux domaines tels que l'hydrologie (les crues, des pluies diluviennes dévastatrices), la finance (les crises financières), la médecine (les épidémies), la fiabilité (durée de vie d'un appareil). Soit Y une variable aléatoire ayant comme fonction de répartition F et Y_1, \dots, Y_n des copies indépendantes de Y . Les quantiles extrêmes de F sont définies comme suit

$$F^{\leftarrow}(1 - \alpha) = \inf\{y : F(y) \geq 1 - \alpha\},$$

où α est si faible que le quantile correspondant tombe au-delà des observations Y_1, \dots, Y_n . Ce problème est étroitement lié à l'estimation de l'indice des valeurs extrêmes de Y . Cet indice de queue joue un rôle fondamental dans l'analyse des extrêmes car il donne la forme de la queue de distribution F à sa droite. Les théories récentes sur la statistique des valeurs extrêmes et en particulier sur l'estimation de l'indice des valeurs extrêmes et quantiles extrêmes sont développées dans Embrechts et al. (1997, [46]), Beirlant et al. (2004, [8]), Reiss et Thomas (2007, [80]) et Novak (2012, [77]).

Lorsqu'une covariable X est disponible et la distribution de Y dépend de X , le problème est d'estimer l'indice de queue conditionnel $\gamma_1(x)$ et son quantile extrême conditionnel correspondant, $F^{\leftarrow}(1 - \alpha|x) = \inf\{y : F(y|x) \geq 1 - \alpha\}$ de la fonction de répartition conditionnelle $F(\cdot|x)$ de Y sachant $X = x$. Motivés par les pluies extrêmes mesurées conjointement avec leurs positions géographiques (longitude, altitude, latitude) considérées comme covariables, Gardes et Girard (2010, [52]) ont proposé un estimateur de l'indice de queue conditionnel et son quantile extrême conditionnel. En suivant la même logique, Ferrez et al. (2011, [48]) ont fait l'analyse des températures extrêmes avec des paramètres topologiques et Pisarenko et Sornette (2003, [79]) ont étudié les séismes extrêmes avec comme covariable leurs emplacements. L'estimation de l'indice des valeurs extrêmes conditionnel et quantiles extrêmes conditionnels avec covariables fixes (ou non aléatoires) a été étudiée assez largement dans la littérature récente de la statistique des valeurs extrêmes.

Nous nous référons à Beirlant et al. (2004, [8]), à Gardes et Girard (2008, [51]), à Gardes et al. (2010, [54]), à Stupfler (2013, [85]) pour un aperçu de la méthodologie disponible, y compris le cas où la variable explicative est fonctionnelle (Gardes et Girard, 2012, [53]). A ce jour, beaucoup d'auteurs s'attardent sur le cas à cova-

riable aléatoire, en dépit de son intérêt pratique. Gardes et Stupfler (2014, [55]) et Goegebeur et al. (2014b, [59]) adaptent l'estimateur de Hill de l'indice de queue d'une distribution à queue lourde en présence d'une covariable aléatoire. L'estimateur des moments introduit par Dekkers et al. (1989, [40]) a été adapté aux covariables aléatoires par Goegebeur et al. (2014a, [58]). Daouia et al. (2011, [30]) ont proposé un estimateur basé sur le noyau de quantiles extrêmes conditionnels avec des covariables aléatoires.

Dans cette partie, nous abordons l'estimation de l'indice des valeurs extrêmes conditionnel et des quantiles extrêmes conditionnels de variables aléatoires lorsque les observations Y_1, \dots, Y_n sont aléatoirement censurées à droite. Par exemple, Y peut représenter la durée jusqu'à la survenue d'un événement d'intérêt (comme le décès d'un patient, la ruine d'une entreprise, ...). En présence de censure, les observations sont tirées d'un triplet $(Z_i, \delta_i, X_i), i = 1, \dots, n$ où $Z_i = \min(Y_i, C_i)$, $\delta_i = \mathbb{1}_{\{Y_i \leq C_i\}}$. Dans le cas sans covariable, Delafosse et Guillou (2002, [41]), Gomes et Oliveira (2003, [61]), Beirlant et al. (2007, [9]), Einmahl et al. (2008, [44]), Beirlant et al. (2010, [8]), Gomes et Neves (2011, [60]), Brahimi et al. (2013, [19]) et Worms et Worms (2014, [96]) ont proposé des estimateurs de l'indice des valeurs extrêmes conditionnel et parfois leurs quantiles extrêmes conditionnels. Par contre, Ndao et al. (2014, [76], voir le chapitre précédent) ont proposé un estimateur de l'indice de queue conditionnel et un estimateur de quantile extrême conditionnel basé sur une adaptation de l'estimateur de Weissman (1978, [95]) et de la méthode de la fenêtre mobile. À notre connaissance, l'estimation de l'indice des valeurs extrêmes conditionnel et de quantiles extrêmes conditionnels à design aléatoire et la censure n'a pas encore été abordée. C'est le sujet de la présente étude. Nous construisons d'abord un estimateur de l'indice des valeurs extrêmes $\gamma_1(x)$ et nous établissons la normalité asymptotique. Notre proposition combine une version du type noyau de l'estimateur de Hill de $\gamma(x)$ (comme développé dans Goegebeur et al. (2014b, [59]) dans le cas non censuré) avec le principe « inverse-probability-of-censoring-weighted kernel » (tels que ceux utilisés dans Einmahl et al. (2008, [44]) et Brahimi et al. (2013, [19]) pour estimer l'indice des valeurs extrêmes inconditionnel avec la censure). Ensuite, nous proposons une adaptation de l'estimateur de type-Weissman du quantile extrême conditionnel en présence de censure. Les performances des estimateurs proposés sont évaluées par des simulations. Le reste de ce chapitre est organisé comme suit. Dans le paragraphe 2, nous construisons notre estimateur de $\gamma_1(x)$ et établissons sa normalité asymptotique. Dans le paragraphe 3, nous proposons un estimateur des quantiles extrêmes conditionnels en présence de censure. Dans le paragraphe 4, nous évaluons par simulation le comportement de nos estimateurs. Quelques perspectives sont données dans le paragraphe 5.

3.2 Construction de l'estimateur et propriétés asymptotiques

3.2.1 Définition de l'estimateur

Nous travaillons dans l'espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ et dans l'espace métrique (\mathcal{X}, d) . Soient $(X_i, Y_i), i = 1, \dots, n$ des copies indépendantes du couple aléatoire (X, Y) où Y est une variable aléatoire non négative et $X \in \mathcal{X}$ (\mathcal{X} un certain ensemble borné dans \mathbb{R}^p) de dimension p avec une densité de probabilité g . Nous supposons que Y est censurée aléatoirement à droite par une variable aléatoire C . Ainsi, nous observons uniquement les triplets indépendants $(X_i, \delta_i, Z_i), i = 1, \dots, n$ où $Z_i = \min(Y_i, C_i)$, $\delta_i = \mathbb{1}_{\{Y_i \leq C_i\}}$ et $\mathbb{1}_{\{A\}}$ est la fonction indicatrice de l'événement A . Nous supposons aussi que C_1, \dots, C_n sont indépendantes et que Y et C sont indépendantes pour tout $X = x$ donné. Soient $F(\cdot|x)$ et $G(\cdot|x)$ les fonctions de répartition conditionnelles respectives de Y et C . Et soient aussi $\bar{F}(\cdot|x) = 1 - F(\cdot|x)$ et $\bar{G}(\cdot|x) = 1 - G(\cdot|x)$ les fonctions de survie conditionnelles respectives de Y et C pour $X = x$ donné. Nous nous intéressons dans ce chapitre aux fonctions à queue lourde c'est-à-dire aux distributions de Fréchet. Théoriquement cela signifie les fonctions de survie conditionnelles vérifient pour tout $X = x$:

(H1) $\bar{F}(u|x) = u^{-1/\gamma_1(x)} L_1(u|x)$ et $\bar{G}(u|x) = u^{-1/\gamma_2(x)} L_2(u|x)$,

où $\gamma_1(\cdot)$ et $\gamma_2(\cdot)$ sont des paramètres fonctionnels positifs de la covariable x et pour x fixé $L_1(\cdot|x)$ et $L_2(\cdot|x)$ sont des fonctions à variations lentes, i.e. pour tout $\lambda > 0$,

$$\lim_{u \rightarrow \infty} \frac{L_i(\lambda u|x)}{L_i(u|x)} = 1, i = 1, 2.$$

La condition **(H1)** est équivalente à dire que les fonctions de survie conditionnelles $\bar{F}(\cdot|x)$ et $\bar{G}(\cdot|x)$ sont respectivement des fonctions à variations régulières à l'infini d'indices $-1/\gamma_1(x)$ et $-1/\gamma_2(x)$.

Remarque 2

✪ *Par indépendance conditionnelle de Y et C , la fonction de répartition conditionnelle $H(\cdot|x)$ de Z pour $X = x$ donné, est à queue lourde d'indice conditionnel fonctionnel $\gamma(x) = \gamma_1(x)\gamma_2(x) / (\gamma_1(x) + \gamma_2(x))$. À ce point, notons que pour tout u et x*

$$\begin{aligned} \bar{H}(u|x) := 1 - H(u|x) &= (1 - F(u|x))(1 - G(u|x)) \\ &= u^{-1/\gamma_1(x)} L_1(u|x) u^{-1/\gamma_2(x)} L_2(u|x) \\ &= u^{-1/\gamma(x)} L(u|x) \end{aligned}$$

où $\gamma(x)$ est défini au dessus et $L(u|x) = L_1(u|x)L_2(u|x)$. De plus,

$$\lim_{u \rightarrow \infty} \frac{L(\lambda u|x)}{L(u|x)} = \lim_{u \rightarrow \infty} \frac{L_1(\lambda u|x)}{L_1(u|x)} \frac{L_2(\lambda u|x)}{L_2(u|x)} = 1.$$

□

✪ Pour les propriétés asymptotiques, nous supposons que les $L_i, i = 1, 2$ sont dans la classe de Hall.

Dans le cas où les Z_i ne sont pas censurées Goegebeur et al. (2014b, [59]) ont proposé un estimateur de l'indice des valeurs extrêmes conditionnel par la version du noyau de l'estimateur de Hill :

$$\gamma_{t_n}^H(x) = \frac{\sum_{i=1}^n K_h(x - X_i) (\log Z_i - \log t_n) \mathbb{1}_{\{Z_i > t_n\}}}{\sum_{i=1}^n K_h(x - X_i) \mathbb{1}_{\{Z_i > t_n\}}} \quad (3.1)$$

où $K_h(x) = h^{-p}K(x/h)$, K est une densité de probabilité dans \mathbb{R}^p , $h := h_n$ est une suite positive non aléatoire appelée fenêtre telle que $h \rightarrow 0$ quand $n \rightarrow \infty$ et t_n est une suite positive non aléatoire appelée seuil telle que $t_n \rightarrow \infty$ si $n \rightarrow \infty$.

L'estimateur décrit dans l'équation (3.1) n'est pas consistant pour $\gamma_1(x)$ car il est calculé directement avec l'échantillon censuré $(X_i, \delta_i, Z_i), i = 1, \dots, n$. En effet, sous certaines hypothèses de régularité appropriées, l'estimateur proposé dans (3.1) converge vers l'indice de queue conditionnel $\gamma(x)$ de la distribution conditionnelle de Z sachant $X = x$ (Goegebeur et al., 2014b, [59]). Pour l'adapter au cas censuré, nous proposons de diviser (3.1) par $\hat{p}_{t_n}(x)$ la proportion des observations non censurées dans $Z_i, i = 1, \dots, n$ qui sont plus grandes que t_n dans un voisinage de x :

$$\hat{p}_{t_n}(x) = \frac{\overline{H}_n^1(t_n|x)}{\overline{H}_n(t_n|x)}$$

où $\overline{H}_n(t_n|x) = \sum_{i=1}^n B_i(x) \mathbb{1}_{\{Z_i > t_n\}}$, $\overline{H}_n^1(t_n|x) = \sum_{i=1}^n B_i(x) \mathbb{1}_{\{Z_i > t_n, \delta_i = 1\}}$ et les poids de Nadarya-Watson définis par

$$B_i(x) = K\left(\frac{x - X_i}{h}\right) / \sum_{j=1}^n K\left(\frac{x - X_j}{h}\right).$$

Nous proposons un estimateur de $\gamma_1(x)$ par

$$\hat{\gamma}_{t_n}^{(c,H)}(x) = \frac{\hat{\gamma}_{t_n}^H(x)}{\hat{p}_{t_n}(x)} \quad (3.2)$$

appelé l'estimateur de l'indice de queue conditionnel à design aléatoire et en présence de censure à droite. Cet estimateur dépend de la fenêtre h et du seuil t_n dont

des méthodes de choix sont discutées dans la partie simulation. Pour la suite, nous déterminons les propriétés asymptotiques de $\hat{\gamma}_{t_n}^{(c,H)}(x)$ et de $\hat{p}_{t_n}(x)$.

3.2.2 Résultats asymptotiques

Dans ce paragraphe, nous étudions théoriquement les propriétés asymptotiques des estimateurs proposés. Pour cela, nous allons considérer quelques hypothèses de régularité comme dans Daouia et al. (2011, [30]) et Goegebeur et al. (2014b, [59]).

Hypothèses lipschitziennes

Tout d'abord, nous allons déterminer quelques conditions lipschitziennes. Pour tout $(x, x') \in \mathcal{X} \times \mathcal{X}$, nous supposons :

(H2) Il existe $c_\gamma > 0$ tel que $\left| \frac{1}{\gamma(x)} - \frac{1}{\gamma(x')} \right| \leq c_\gamma d(x, x')$

(H3) Il existe $c_L > 0$ et $u_0 > 1$ tels que

$$\sup_{u \geq u_0} \left| \frac{\log L(u|x)}{\log u} - \frac{\log L(u|x')}{\log u} \right| \leq c_L d(x, x')$$

(H4) Il existe $c_g > 0$ tel que $|g(x) - g(x')| \leq c_g d(x, x')$.

Hypothèse classique (condition du second ordre)

(H5) Pour tout $x \in \mathbb{R}^p$, il existe une fonction $\rho(x) < 0$ et une fonction à variations régulières à l'infini $b(\cdot|x)$ d'indice $\rho(x)$ telles que pour tout $u > 0$,

$$\lim_{t \rightarrow \infty} \frac{H^{\leftarrow} \left(1 - \frac{1}{tu} \mid x \right) / H^{\leftarrow} \left(1 - \frac{1}{t} \mid x \right) - u^{\gamma(x)}}{b(t|x)} = u^{\gamma(x)} \frac{u^{\rho(x)} - 1}{\rho(x)}.$$

Hypothèse sur le noyau

(H6) K est une fonction positive, bornée, intégrable et à support compact $S \subseteq \mathbb{R}^p$.

Nous pouvons maintenant énoncer nos résultats. Nous déterminons d'abord les propriétés asymptotiques de $\hat{p}_{t_n}(x)$. La preuve ce résultat est dans le paragraphe démonstrations.

Proposition 3.2.1 *Supposons les hypothèses (H1)-(H6) vérifiées. Soit t_n une suite non aléatoire telle que $t_n \rightarrow \infty$, $nh^p \bar{H}(t_n|x) \rightarrow \infty$, $nh^{p+2} \bar{H}(t_n|x)(\log t_n)^2 \rightarrow 0$ et $\sqrt{nh^p \bar{H}(t_n|x)} b\left(\frac{1}{\bar{H}(t_n|x)}|x\right) \rightarrow 0$ quand $n \rightarrow \infty$. Alors, pour tout $x \in \mathcal{X}$ tel que $g(x) > 0$,*

$$\sqrt{nh^p \bar{H}(t_n|x)} (\hat{p}_{t_n}(x) - p_x) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{p_x(1-p_x) \|K\|_2^2}{g(x)}\right) \quad \text{quand } n \rightarrow \infty,$$

où $p_x = \gamma_2(x)/(\gamma_1(x) + \gamma_2(x))$ et $\|K\|_2^2 = \int K^2(u) du$.

Nous allons à présent déterminer la normalité asymptotique de l'estimateur proposé dans (3.2) de $\gamma_1(x)$. La démonstration de ce théorème est dans le paragraphe démonstrations.

Théorème 3.2.1 *Supposons les conditions (H1)-(H6) vérifiées. Soit t_n une suite non aléatoire telle que $t_n \rightarrow \infty$, $nh^p \bar{H}(t_n|x) \rightarrow \infty$, $nh^{p+2} \bar{H}(t_n|x)(\log t_n)^2 \rightarrow 0$, $\sqrt{nh^p \bar{H}(t_n|x)} b\left(\frac{1}{\bar{H}(t_n|x)}|x\right) \rightarrow 0$ et $\sqrt{nh^p \bar{H}(t_n|x)} b(t_n|x) \rightarrow 0$ quand $n \rightarrow \infty$. Alors, pour tout $x \in \mathcal{X}$ tel que $g(x) > 0$,*

$$\sqrt{nh^p \bar{H}(t_n|x)} (\hat{\gamma}_{t_n}^{(c,H)}(x) - \gamma_1(x)) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{\gamma_1^3(x) \|K\|_2^2}{\gamma(x) g(x)}\right) \quad \text{quand } n \rightarrow \infty. \quad (3.3)$$

Remarque 3 (Sur les conditions)

- ❶ $\sqrt{nh^p \bar{H}(t_n|x)} b\left(\frac{1}{\bar{H}(t_n|x)}|x\right) \rightarrow 0$ nous permet de spécifier le biais dû par $\hat{p}_{t_n}(x)$.
- ❷ $\sqrt{nh^p \bar{H}(t_n|x)} b(t_n|x) \rightarrow 0$ permet de quantifier le biais de $\hat{\gamma}_{t_n}^H(x)$.
- ❸ $nh^{p+2} \bar{H}(t_n|x)(\log t_n)^2 \rightarrow 0$ est une condition pour que le carré du biais asymptotique, de l'ordre de $(h \log t_n)^2$, soit négligeable devant la variance asymptotique, de l'ordre de $\frac{1}{nh^p \bar{H}(t_n|x)}$.

Notons que la variance asymptotique dans (3.3) est la même (jusqu'à un facteur d'échelle $\|K\|_2^2/g(x)$) comme la variance asymptotique obtenue dans Beirlant et al. (2007, [9]) pour l'estimateur de Hill adapté selon le modèle de Hall avec des données censurées sans covariable.

3.3 Estimation de quantiles extrêmes conditionnels

Dans ce paragraphe, nous abordons le problème de l'estimation de quantiles extrêmes conditionnels $q(\alpha_n|x)$ d'ordre $1 - \alpha_n$ de la distribution conditionnelle $F(\cdot|x)$ de Y pour $X = x$ donné. Ces quantiles vérifient $\bar{F}(q(\alpha_n|x)|x) = \alpha_n$ où $\alpha_n \rightarrow 0$ quand $n \rightarrow +\infty$.

Tout d'abord, nous considérons l'estimateur de Kaplan-Meier généralisé (à noyau) de $\bar{F}(\cdot|x)$ défini comme dans Dabrowska (1992, [27]) pour tout $(x, t) \in \mathcal{X} \times [0, \infty)$:

$$\widehat{F}_n(t|x) = \prod_{i=1}^n \left[1 - \frac{B_i(x)}{\sum_{j=1}^n \mathbb{1}_{\{Z_j \geq Z_i\}} B_j(x)} \right]^{\mathbb{1}_{\{Z_i \leq t, \delta_i=1\}}} \quad \text{si } t \leq Z_{(n)} \quad (3.4)$$

et $\widehat{F}_n(t|x) = 0$ si $t > Z_{(n)}$ où $Z_{(1)} \leq \dots \leq Z_{(n)}$ étant la statistique d'ordre de Z_1, \dots, Z_n . Dans cette situation, un estimateur du quantile conditionnel $q(\alpha|x)$ (pour $\alpha \in (0, 1)$ fixé) peut être défini par l'inverse généralisée de $\widehat{F}_n(\cdot|x)$ comme suit :

$$\widehat{q}_n^c(\alpha|x) := \widehat{F}_n^{\leftarrow}(\alpha|x) = \inf\{u, \widehat{F}_n(u|x) \leq \alpha\}. \quad (3.5)$$

Lorsque $\alpha_n \rightarrow 0$ quand $n \rightarrow \infty$, nous proposons un estimateur des quantiles extrêmes conditionnels $q(\alpha_n|x)$ par une adaptation de l'estimateur de Weissman (1978, [95]) :

$$\widehat{q}_n^{(c,W)}(\alpha_n|x) := \widehat{q}_n^c\left(\widehat{F}_n(Z_{(n-k)}|x)|x\right) \left(\widehat{F}_n(Z_{(n-k)}|x)/\alpha_n\right)^{\widehat{\gamma}_{Z_{(n-k)}}^{(c,H)}(x)}. \quad (3.6)$$

Notons par $\widehat{q}_n^c(\widehat{F}_n(Z_{(n-k)}|x)|x)$ l'estimateur classique de (3.5), $(\widehat{F}_n(Z_{(n-k)}|x)/\alpha_n)^{\widehat{\gamma}_{Z_{(n-k)}}^{(c,H)}(x)}$ est la partie qui permet d'extrapoler au delà des observations et $\widehat{\gamma}_{Z_{(n-k)}}^{(c,H)}(x)$ est l'estimateur défini dans (3.2) avec t_n remplacé par $Z_{(n-k)}$. Le choix du seuil $Z_{(n-k)}$ (est équivalent au choix de k) est discuté dans ce paragraphe simulation.

Remarque 4 Quand il n'y a pas de censure, $\widehat{F}_n(t|x)$ se réduit à

$$\widetilde{F}_n(t|x) := \sum_{i=1}^n K_h(x - X_i) \mathbb{1}_{\{Z_i > t\}} / \sum_{i=1}^n K_h(x - X_i) \quad (3.7)$$

et $\widehat{q}_n^c(\alpha|x)$ devient

$$\widehat{q}_n(\alpha|x) := \inf\{u, \widetilde{F}_n(u|x) \leq \alpha\}. \quad (3.8)$$

Plusieurs auteurs ont considéré l'estimateur de (3.8) lorsque $\alpha \in (0, 1)$ est fixé. En particulier, la consistance de (3.8) été créée par Stone (1977, [84]) et Gannoun (1990,

[50]). Normalité asymptotique a été prouvée par Stute (1986, [86]), Samanta (1989, [82]) et Berlinet et al. (2001, [16]). La normalité asymptotique de (3.7) a été présentée par Daouia et al. (2011, [30]) quand $t := t_n \rightarrow \infty$ (c'est-à-dire lorsque (3.7) est utilisé pour estimer les probabilités de la queue avec les α petits). Daouia et al. (2011, [30]) ont également prouvé la normalité asymptotique de (3.8) lors de l'estimation des quantiles extrêmes (c'est-à-dire lorsque $\alpha := \alpha_n \rightarrow 0$ quand $n \rightarrow \infty$).

□

Dans ce paragraphe, nous évaluons les propriétés de l'estimateur proposé dans (3.8) via des simulations.

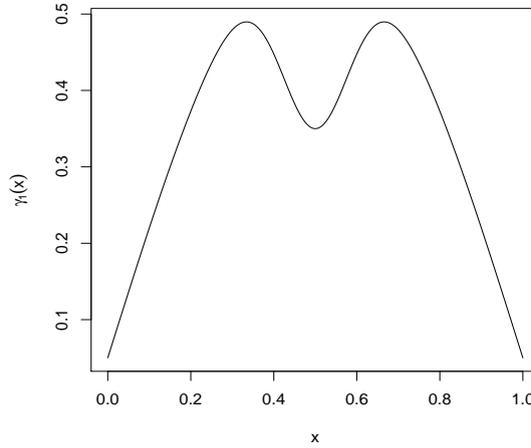
3.4 Expériences numériques

Dans ce paragraphe, nous évaluons la performance des estimateurs (3.2) et (3.6) proposés en échantillon fini.

3.4.1 Génération des données

Nous simulons $R = 100$ échantillons de taille n ($n = 200, 400, 600, 800$) de variables aléatoires indépendantes (Z_i, δ_i, X_i) où $Z_i = \min(Y_i, C_i)$ et $X_i \in \mathbb{R}$ uniforme sur $[0, 1]$. La fonction de répartition conditionnelle de Y_i pour $X_i = x$ est de Paréto simple c'est-à-dire $\bar{F}(y|x) = y^{-1/\gamma_1(x)}$ où $\gamma_1(x) = 0.5 \left(0.1 + \sin(\pi x) \times \left(1.1 - 0.5 \exp(-64(x - 0.5)^2) \right) \right)$. Par contre la fonction de répartition conditionnelle de C_i est aussi de Paréto simple où le paramètre fonctionnel est choisi pour donner le pourcentage désiré c ($c = 10\%, 25\%, 40\%$) correspondants à $\gamma_2(x) \in (2.5, 1, 0.5)$. Cette fonction $x \in [0, 1] \mapsto \gamma_1(x)$ est positive comme le montre la figure 3.1.

Pour toute réplication R simulée, nous estimons $\gamma_1(\cdot)$ au points $x = (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9)$ en utilisant l'estimateur proposé dans (3.2) et le noyau bi-quadratique $K(x) = \frac{15}{16}(1 - x^2)^2 \mathbf{1}_{\{|x| \leq 1\}}$. Pour calculer notre estimateur, nous devons choisir le paramètre de lissage h et le seuil t_n . Ainsi, nous avons choisi t_n comme la $(n - k)^{\text{ième}}$ statistique d'ordre correspondante à $Z_{(n-k)}$ qui est classique en statistique des extrêmes. Maintenant, nous proposons un algorithme pour choisir (h, k) . Cet algorithme est adapté à celui de Goegebeur et al. (2014b, [59]). Tout d'abord, nous sélectionnons le paramètre de lissage h en utilisant le critère suivant la validation croisée introduite par Yao (1999, [97]) et a également mis en

FIGURE 3.1 – La fonction de $\gamma_1(\cdot)$ sur $[0, 1]$.

oeuvre par Daouia et al. (2011, [30]) :

$$h^* := \arg \min_{h \in \mathbb{H}} \sum_{i=1}^n \sum_{j=1}^n \left(\mathbb{1}_{\{Z_i > Z_j\}} - \widehat{F}_{n,-i}(Z_j | x_i) \right)^2,$$

où $\widehat{F}_{n,-i}$ est l'estimateur conditionnel de Kaplan-Meier (3.4) (dépend de h) calculé avec l'échantillon $\{(X_j, \delta_j, Z_j), j = 1, \dots, n, j \neq i\}$ et \mathbb{H} est une grille de h (nous prenons $\mathbb{H} = \{h_1 < h_2 < \dots < h_M\}$ où les points h_2, \dots, h_{M-1} sont régulièrement répartis entre $h_1 = 1/(5 \log n)$ et $h_M = 0.5$, avec $M = 60$). Une fois que le paramètre de lissage est déterminé, on choisit k comme suit :

1. nous estimons $\widehat{\gamma}_{Z_{(n-k)}}^{(c,H)}(x)$ avec $k = 1, \dots, n-1$,
2. nous formons plusieurs « blocs » successifs d'estimations $\widehat{\gamma}_{Z_{(n-k)}}^{(c,H)}(x)$ (un bloc pour $k \in \{1, \dots, 15\}$, un deuxième bloc pour $k \in \{16, \dots, 30\}$ et ainsi de suite),
3. on calcule l'écart type de l'estimation dans chaque bloc,
4. on détermine la valeur de k à utiliser (par la suite désigné par k^*) à partir du bloc d'écart-type minimal. Précisément, nous prenons comme k^* le milieu de l'intervalle dont l'écart est minimal.

Finalement, nous estimons $\gamma_1(x)$ de (3.2) avec $(h, k) = (h^*, k^*)$.

3.4.2 Résultats des simulations

Pour chaque taille de l'échantillon de configuration \times pourcentage de censure \times valeur des paramètres à estimer, nous calculons l'estimateur de $\gamma_1(x)$, la racine carrée de l'erreur quadratique moyenne (RMSE) et l'erreur absolue moyenne (MAE). Nous avons obtenu également des intervalles de confiance pour

$\gamma_1(x)$ à 95% de niveau de confiance et les amplitudes des intervalles de confiance. Les résultats obtenus sont donnés dans le Tableau 3.1. Comme prévu, les estimations se détériorent lorsque le pourcentage de censure augmente de 10% à 40% et si la taille de l'échantillon diminue 800 à 200 en terme de la racine carrée de l'erreur quadratique moyenne et de l'erreur absolue moyenne. Nous obtenons le même résultat pour les intervalles de confiance. Cependant, les résultats numériques indiquent que l'estimateur se comporte toujours de façon raisonnable lorsque la censure est grande ou la taille de l'échantillon est modérée. En effet, dans les deux cas, le biais et la variabilité restent limités. Dans l'ensemble, les résultats sont assez robustes à la censure. La figure 3.3 montre les boîtes à moustaches des R réalisations de l'estimateur proposé dans (3.2) pour chaque x ($x = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$). De ce fait, il apparaît que l'estimateur proposé dans (3.2) se comporte très bien pour l'estimateur de l'indice des valeurs extrêmes conditionnel $\gamma_1(x)$.

3.4.3 Estimation des quantiles extrêmes conditionnels

Dans ce paragraphe, nous évaluons la performance de l'estimateur proposé dans (3.6) du quantile extrême conditionnel $q(1/1000|x)$ d'ordre $1 - 1/1000$ de la fonction de répartition conditionnelle de Y sachant $X = x$, pour $x = (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9)$. Les considérations de la simulation sont les mêmes que ci-dessus. Pour chaque configuration des paramètres de calcul de simulation, nous calculons les estimations de (3.6) avec (h^*, k^*) obtenu dans le paragraphe 3.4. En se basant sur les R réplifications simulées, nous obtenons l'estimation de $q(1/1000|x)$, la racine carrée de l'erreur quadratique moyenne et l'erreur absolue moyenne. Les résultats sont regroupés dans le Tableau 3.2.

De ce Tableau, on constate, comme prévu, que la performance de l'estimateur proposé se détériore avec l'augmentation de censure et les diminutions de taille de l'échantillon. Lorsque la taille de l'échantillon est modérée ($n = 200$, par exemple) et le pourcentage de censure est élevé ($c \geq 25\%$, par exemple), le biais peut être assez important. Pour la taille de l'échantillon modérée ($n = 200$), le biais reste toutefois limité quand la censure ne dépasse pas 10%. Lorsque la taille de l'échantillon est suffisamment grand ($n \geq 400$, par exemple), le biais de l'estimateur proposé reste limité dans la plupart des scénarios de simulation. Finalement, Figure 3.3 montre les boîtes à moustaches des R réalisations de l'estimateur proposé dans (3.6) pour tout x dans $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. De même que ci-dessus, nous observons que l'estimateur proposé dans (3.6) adapte très bien la forme de la fonction du quantile extrême conditionnel $x \mapsto q(1/1000|x)$.

3.4.4 Comparaison

Pour comparer nos estimateurs proposés dans (3.2) et (3.6) à ceux de la littérature, nous allons utiliser deux autres stratégies d'estimation simples. La première stratégie est une procédure dénommée Cas-Complet-Noyau ("CCN", pour abrégé) qui consiste à éliminer toutes les observations censurées des échantillons simulés et faire le calcul en utilisant les estimateurs proposés dans Goegebeur et al. (2014b, [59]) de l'indice des valeurs extrêmes conditionnel avec les données qui en résultent. Dans le second cas, nous estimons avec l'ensemble des jeux de données simulées en traitant chaque observation Z_i comme s'il s'agissait non censurée. Cette alternative est notée "CIN" (Cas-Ignoré-Noyau). Pour les faire, nous procédons comme suit :

- **Cas-Complet-Noyau**

Nous avons simulé des données censurées aléatoirement à droite mesurées conjointement avec des covariables à design aléatoire. Puis nous avons tronqué l'échantillon en prenant seulement les observations qui ne sont pas censurées pour pouvoir appliquer l'estimateur de l'indice de queue proposé par Goegebeur et al. (2014b, [59]).

- **Cas-Ignoré-Noyau**

Nous avons aussi considéré ces mêmes données en imposant l'indicateur de censure égal partout à un. Ce qui entraîne la proportion des données non censurées égale à un partout. Autrement dit nous utilisons les estimateurs de l'indice et des quantiles extrêmes avec des données non censurées aléatoirement à droite. Nous déterminons également l'erreur quadratique moyenne (RMSE) et l'erreur absolue moyenne (MAE) dans chaque stratégie. Les résultats de ces simulations sont rapportés dans les Tableaux 3.3 à 3.6.

3.5 Conclusion et Perspectives

Dans ce chapitre, nous avons abordé l'estimation de l'indice des valeurs extrêmes et de quantile extrême conditionnel d'une distribution à queue lourde quand une covariable aléatoire est disponible et que les données sont aléatoirement censurées à droite. Nous avons construit une version de l'estimateur de Hill de l'indice de queue conditionnel basé sur la méthode noyau et nous avons établi sa normalité asymptotique. Nous avons également proposé un estimateur de type-Weissman du quantile extrême conditionnel. Nous avons évalué leur performance via échantillon fini par simulations. À partir de ces simulations, les estimateurs proposés donnent de bons résultats à condition que la taille de l'échantillon est assez grande, même si le pourcentage de censure est élevé. Lorsque la taille de l'échantillon est modérée, les esti-

mateurs proposés donnent encore de bons résultats avec un pourcentage de censure élevé. En outre, les estimateurs proposés semblent adapter correctement la forme de la fonction de l'indice des valeurs extrêmes conditionnel inconnu et la fonction quantile extrême conditionnel.

Maintenant, plusieurs questions méritent une réflexion. En particulier, une adaptation rigoureuse des propriétés asymptotiques de l'estimateur proposé dans l'équation (3.6) de quantile extrême conditionnel est nécessaire. C'est l'un des sujets pour nos futures recherches. L'extension de nos résultats de convergence vers la convergence uniforme (au delà de x) est également intéressante. Ces résultats de convergence uniforme permettraient la construction de bandes de confiance simultanées pour la fonction de l'indice des valeurs extrêmes conditionnel et la fonction quantile extrême conditionnel.

n	$\gamma_1(0.1) = 0.2199$			$\gamma_1(0.2) = 0.3728$			$\gamma_1(0.3) = 0.4793$			$\gamma_1(0.4) = 0.4477$			$\gamma_1(0.5) = 0.35$		
	10%	25%	40%	10%	25%	40%	10%	25%	40%	10%	25%	40%	10%	25%	40%
200	0.2255	0.2258	0.2225	0.3595	0.3679	0.3623	0.4673	0.4724	0.4695	0.4250	0.4221	0.4295	0.3620	0.3652	0.3631
	(0.2082)	(0.2165)	(0.2191)	(0.2444)	(0.2690)	(0.2759)	(0.2764)	(0.2947)	(0.3227)	(0.2598)	(0.2623)	(0.2930)	(0.2412)	(0.2633)	(0.2682)
	[0.0032]	[0.0033]	[0.0039]	[0.0050]	[0.0081]	[0.0086]	[0.0090]	[0.0124]	[0.0200]	[0.0072]	[0.0074]	[0.0128]	[0.0056]	[0.0078]	[0.0082]
	[0.1394, 0.3115]	[0.1342, 0.3174]	[0.1222, 0.3228]	[0.2150, 0.5041]	[0.2078, 0.5280]	[0.1839, 0.5408]	[0.2800, 0.6547]	[0.2622, 0.6827]	[0.2261, 0.7130]	[0.2486, 0.6013]	[0.2256, 0.6185]	[0.2022, 0.6568]	[0.2244, 0.4996]	[0.2146, 0.5158]	[0.1964, 0.5299]
	0.1721†	0.1832†	0.2006†	0.2891†	0.3202†	0.3568†	0.3747†	0.4205†	0.4869†	0.3527†	0.3929†	0.4546†	0.2752†	0.3012†	0.3335†
400	0.2220	0.2195	0.2141	0.3666	0.3680	0.3664	0.4579	0.4621	0.4615	0.4368	0.4419	0.4409	0.3784	0.3830	0.3809
	(0.1534)	(0.1557)	(0.1609)	(0.1945)	(0.2053)	(0.2130)	(0.2433)	(0.2458)	(0.2551)	(0.2305)	(0.2560)	(0.2741)	(0.2118)	(0.2272)	(0.2298)
	[0.0011]	[0.0013]	[0.0017]	[0.0025]	[0.0027]	[0.0032]	[0.0051]	[0.0061]	[0.0069]	[0.0044]	[0.0063]	[0.0086]	[0.0034]	[0.0045]	[0.0048]
	[0.1624, 0.2815]	[0.1562, 0.2828]	[0.1454, 0.2827]	[0.2660, 0.4671]	[0.2581, 0.4780]	[0.2412, 0.4915]	[0.3257, 0.5901]	[0.3145, 0.6096]	[0.2926, 0.6304]	[0.3144, 0.5591]	[0.3047, 0.5792]	[0.2828, 0.5991]	[0.2833, 0.4735]	[0.2784, 0.4876]	[0.2620, 0.4999]
	0.1192†	0.1266†	0.1373†	0.2011†	0.2199†	0.2503†	0.2644†	0.2951†	0.3378†	0.2446†	0.2746†	0.3163†	0.1903†	0.2092†	0.2380†
600	0.2220	0.2192	0.2153	0.3718	0.3689	0.3638	0.4562	0.4532	0.4612	0.4368	0.4362	0.4396	0.3727	0.3770	0.376
	(0.1491)	(0.1523)	(0.1604)	(0.1906)	(0.1956)	(0.2121)	(0.2228)	(0.2323)	(0.2394)	(0.1996)	(0.2148)	(0.2184)	(0.1928)	(0.2006)	(0.2045)
	[0.0008]	[0.0010]	[0.0012]	[0.0022]	[0.0024]	[0.0030]	[0.0036]	[0.0043]	[0.0051]	[0.0022]	[0.0031]	[0.0039]	[0.0019]	[0.0026]	[0.0032]
	[0.1732, 0.2709]	[0.1680, 0.2704]	[0.1592, 0.2714]	[0.2892, 0.4544]	[0.2779, 0.4598]	[0.2618, 0.4658]	[0.3487, 0.5638]	[0.3339, 0.5725]	[0.3222, 0.6001]	[0.3371, 0.5366]	[0.3262, 0.5463]	[0.3122, 0.5671]	[0.3060, 0.4395]	[0.3041, 0.4500]	[0.2939, 0.4593]
	0.0977†	0.1024†	0.1121†	0.1651†	0.1819†	0.2040†	0.2150†	0.2385†	0.2779†	0.1995†	0.2201†	0.2549†	0.1549†	0.1681†	0.1898†
800	0.2238	0.2226	0.2205	0.3646	0.3627	0.3626	0.4619	0.4628	0.4637	0.4431	0.4385	0.4350	0.3734	0.3747	0.3767
	(0.1372)	(0.1420)	(0.1510)	(0.1713)	(0.1825)	(0.2000)	(0.1944)	(0.2179)	(0.2354)	(0.1874)	(0.1960)	(0.2097)	(0.1870)	(0.1956)	(0.2042)
	[0.0005]	[0.0006]	[0.0008]	[0.0014]	[0.0016]	[0.0024]	[0.0024]	[0.0036]	[0.0049]	[0.0021]	[0.0023]	[0.0031]	[0.0018]	[0.0024]	[0.0031]
	[0.1827, 0.2649]	[0.1787, 0.2665]	[0.1729, 0.2681]	[0.2932, 0.4359]	[0.2844, 0.4411]	[0.2751, 0.4502]	[0.3705, 0.5534]	[0.3600, 0.5656]	[0.3450, 0.5823]	[0.3580, 0.5283]	[0.3430, 0.5340]	[0.3250, 0.5449]	[0.2960, 0.4395]	[0.2906, 0.4500]	[0.2817, 0.4593]
	0.0822†	0.0878†	0.0952†	0.1427†	0.1567†	0.1751†	0.1829†	0.2056†	0.2373†	0.1703†	0.1910†	0.2199†	0.1335†	0.1459†	0.1654†

Tableau 3.1 – Résultats des simulations pour l’estimateur proposé dans (3.2) de $\gamma_1(x)$. Pour chaque scénario, la première ligne donne la moyenne de $\gamma_1(x)$ pour $R = 100$. (·) : Racine carrée de l’erreur quadratique moyenne (RMSE). [·] : Erreur absolue moyenne (MAE). [·, ·] : intervalle de confiance au niveau de confiance 0.95 pour $\gamma_1(x)$. † : Amplitude des intervalles.

n	$\gamma_1(0.6) = 0.4477$			$\gamma_1(0.7) = 0.4793$			$\gamma_1(0.8) = 0.3728$			$\gamma_1(0.9) = 0.2199$		
	10%	25%	40%	10%	25%	40%	10%	25%	40%	10%	25%	40%
200	0.4423 (0.2789) [0.0094] [0.2668, 0.6178] 0.3510†	0.4439 (0.2910) [0.0112] [0.2476, 0.6401] 0.3925†	0.4496 (0.3020) [0.0135] [0.2276, 0.6716] 0.4441†	0.4650 (0.2857) [0.0125] [0.2766, 0.6535] 0.3770†	0.4657 (0.2961) [0.0143] [0.2548, 0.6766] 0.4219†	0.4919 (0.3617) [0.0362] [0.2468, 0.7369] 0.4900†	0.3759 (0.2596) [0.0070] [0.2314, 0.5203] 0.2889†	0.3818 (0.2806) [0.0104] [0.2239, 0.5397] 0.3159†	0.3869 (0.2950) [0.0121] [0.2075, 0.5662] 0.3587†	0.2236 (0.1912) [0.0022] [0.1370, 0.3103] 0.1732†	0.2201 (0.1962) [0.0026] [0.1288, 0.3114] 0.1826†	0.2185 (0.2057) [0.0030] [0.1184, 0.3185] 0.2001†
400	0.4338 (0.2188) [0.0038] [0.3124, 0.5552] 0.2428†	0.4390 (0.2322) [0.0055] [0.3028, 0.5753] 0.2725†	0.4359 (0.2379) [0.0056] [0.2810, 0.5909] 0.3099†	0.4680 (0.2484) [0.0056] [0.3368, 0.5991] 0.2622†	0.4689 (0.2577) [0.0074] [0.3209, 0.6170] 0.2961†	0.4782 (0.2932) [0.0127] [0.3078, 0.6487] 0.3409†	0.3665 (0.2148) [0.0032] [0.2644, 0.4685] 0.2040†	0.3681 (0.2148) [0.0034] [0.2570, 0.4791] 0.2220†	0.3669 (0.2384) [0.0052] [0.2399, 0.4938] 0.2539†	0.2202 (0.1755) [0.0015] [0.1608, 0.2796] 0.1188†	0.2183 (0.1811) [0.0017] [0.1554, 0.2812] 0.1257†	0.2163 (0.1857) [0.0019] [0.1486, 0.2840] 0.1354†
600	0.4353 (0.2060) [0.0029] [0.3362, 0.5344] 0.1983†	0.4363 (0.2200) [0.0035] [0.3254, 0.5473] 0.2219†	0.4411 (0.2271) [0.0042] [0.3131, 0.5690] 0.2560†	0.4655 (0.2161) [0.0031] [0.3592, 0.5718] 0.2125†	0.4575 (0.2195) [0.0036] [0.3384, 0.5765] 0.2381†	0.4583 (0.2434) [0.0052] [0.3193, 0.5974] 0.2782†	0.3701 (0.1880) [0.0014] [0.2875, 0.4527] 0.1651†	0.3676 (0.1881) [0.0016] [0.2762, 0.4591] 0.1828†	0.3662 (0.1898) [0.0019] [0.2630, 0.4694] 0.2064†	0.2178 (0.1450) [0.0007] [0.1694, 0.2663] 0.0969†	0.2141 (0.1503) [0.0008] [0.1629, 0.2652] 0.1023†	0.2092 (0.1622) [0.0010] [0.1530, 0.2654] 0.1124†
800	0.4397 (0.1940) [0.0022] [0.3544, 0.5250] 0.1706†	0.4401 (0.1957) [0.0022] [0.3453, 0.5349] 0.1895†	0.4408 (0.2057) [0.0028] [0.3301, 0.5514] 0.2213†	0.4623 (0.2135) [0.0032] [0.3696, 0.5550] 0.1854†	0.4604 (0.2214) [0.0041] [0.3560, 0.5649] 0.2089†	0.4614 (0.2360) [0.0050] [0.3419, 0.5809] 0.2391†	0.3697 (0.1759) [0.0015] [0.2990, 0.4405] 0.1415†	0.3691 (0.1823) [0.0017] [0.2920, 0.4462] 0.1543†	0.3648 (0.1939) [0.0021] [0.2772, 0.4525] 0.1753†	0.2211 (0.1327) [0.0003] [0.1801, 0.2621] 0.0820†	0.2186 (0.1340) [0.0005] [0.1751, 0.2622] 0.0871†	0.2165 (0.1434) [0.0006] [0.1687, 0.2642] 0.0954†

Continuité du Tableau 3.1

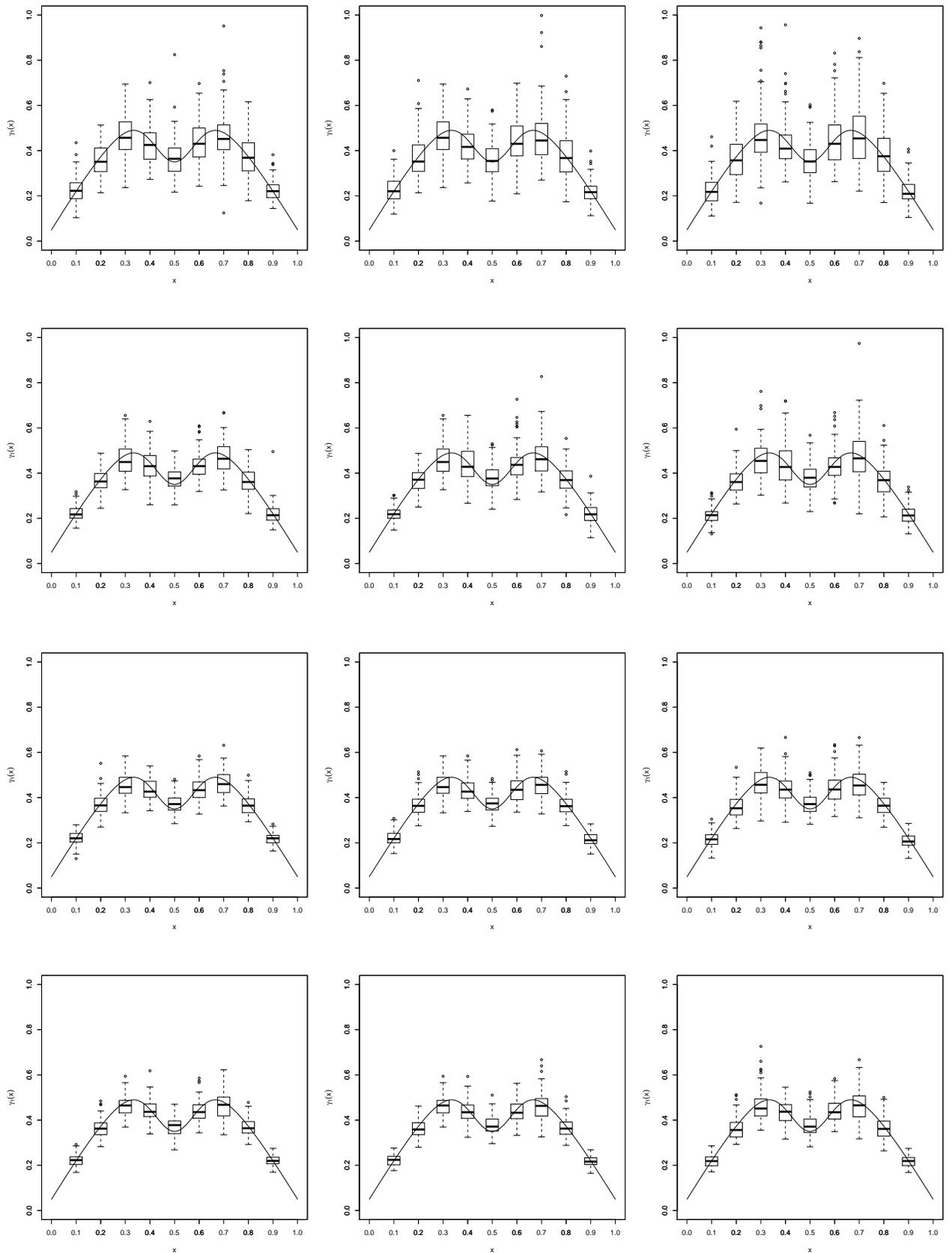


FIGURE 3.2 – Boxplots des R estimations de $\gamma_1(x)$ pour $n = 200$ (1^{ière} ligne), $n = 400$ (2^{ième} ligne), $n = 600$ (3^{ième} ligne), $n = 800$ (4^{ième} ligne). A gauche : $c = 10\%$, Au Centre : $c = 25\%$, A droite : $c = 40\%$.

n	$q(1/1000 0.1) = 4.5695$			$q(1/1000 0.2) = 13.1358$			$q(1/1000 0.3) = 27.4140$			$q(1/1000 0.4) = 22.0358$			$q(1/1000 0.5) = 11.2201$		
	10%	25%	40%	10%	25%	40%	10%	25%	40%	10%	25%	40%	10%	25%	40%
200	4.9051 (2.0984) [1.4987]	4.9217 (2.1909) [1.5564]	4.9775 (2.8898) [1.7181]	12.8627 (6.5390) [5.2294]	15.1259 (14.6945) [7.7024]	14.6754 (15.1103) [8.4554]	29.7765 (22.1035) [15.1600]	34.2158 (33.9952) [19.8775]	47.4512 (91.4621) [35.1628]	21.1591 (13.9045) [9.9591]	20.8619 (14.5388) [10.0064]	30.0911 (65.2830) [20.0666]	15.1480 (11.1821) [6.5352]	17.4580 (17.2417) [8.8535]	19.7782 (19.2146) [11.2048]
400	4.5509 (1.0149) [0.7783]	4.4909 (1.0231) [0.7846]	4.3758 (1.0838) [0.8026]	12.7265 (4.1748) [3.3625]	13.0571 (4.6262) [3.6549]	13.2056 (6.0845) [4.2554]	25.4270 (14.6121) [10.8706]	27.4535 (19.7593) [12.0761]	28.4287 (23.8713) [13.1128]	21.8468 (11.1816) [8.2532]	23.7294 (14.4708) [9.6850]	25.7412 (23.1045) [13.5354]	13.6515 (5.8417) [3.8271]	13.5535 (6.8835) [4.2458]	14.1533 (7.4824) [5.0220]
600	4.5291 (0.8798) [0.7116]	4.4845 (0.9537) [0.7504]	4.3823 (0.9731) [0.7849]	13.6544 (4.3521) [3.1794]	13.4414 (4.5436) [3.3703]	13.1955 (4.6639) [3.7097]	24.0469 (10.3738) [8.7212]	24.1875 (11.3229) [9.2147]	26.2753 (13.2170) [10.3688]	20.5668 (6.8417) [5.8623]	21.2501 (9.0705) [7.0867]	22.4150 (12.0648) [7.5981]	13.7379 (5.5173) [3.7719]	13.9236 (6.0139) [4.1005]	14.0653 (6.7546) [4.2414]
800	4.6094 (0.7210) [0.5762]	4.5768 (0.7395) [0.6098]	4.5582 (0.8746) [0.7028]	12.3614 (3.6101) [2.6401]	12.2954 (3.6328) [3.0161]	12.6640 (3.7755) [3.6113]	24.7095 (8.6029) [6.8036]	25.5801 (10.9137) [8.3678]	27.4010 (12.2523) [10.2965]	21.7130 (6.7012) [5.4887]	21.0711 (7.4548) [5.7443]	20.9701 (7.8365) [6.1833]	13.0676 (3.9094) [3.0581]	13.6430 (4.8973) [3.5205]	13.8643 (6.0034) [3.8947]

Tableau 3.2 – Résultats de simulation pour $q(1/1000|x)$. Pour toute configuration avec les différents paramètres (n, c, x) , première ligne l'estimation de (3.6) de $q(1/1000|x)$ pour $R = 100$. (\cdot) : RMSE. $[\cdot]$: MAE.

n	$q(1/1000 0.6) = 22.0358$			$q(1/1000 0.7) = 27.4140$			$q(1/1000 0.8) = 13.1358$			$q(1/1000 0.9) = 4.5695$		
	10%	25%	40%	10%	25%	40%	10%	25%	40%	10%	25%	40%
200	24.8973 (18.6830) [12.7777]	26.8478 (23.4396) [15.0691]	31.6122 (40.3262) [19.6097]	35.4726 (42.2866) [20.8515]	37.9542 (49.5125) [24.0656]	38.4741 (56.0340) [25.0722]	15.0307 (9.5039) [6.5695]	17.6032 (17.6637) [9.3951]	18.7812 (18.1308) [10.5987]	4.7138 (1.6713) [1.1836]	4.6911 (1.8586) [1.2528]	4.7025 (2.0443) [1.4079]
400	21.0084 (9.9573) [7.1595]	23.0927 (16.5794) [9.1380]	22.8012 (17.8647) [9.1503]	24.7805 (12.4526) [9.6803]	27.1309 (16.3368) [11.6037]	32.7301 (33.0097) [16.9023]	12.8802 (5.0678) [4.0083]	13.2370 (5.4929) [4.0705]	13.7795 (7.9506) [5.2544]	4.6228 (1.0399) [1.0226]	4.5499 (1.4602) [1.0585]	4.5391 (1.4660) [1.0976]
600	20.7236 (8.0589) [6.1070]	21.3903 (9.5612) [7.2747]	22.7184 (12.2271) [8.3499]	25.5869 (10.3777) [8.3564]	24.5530 (10.6816) [8.5976]	25.8020 (14.2859) [10.6484]	12.8633 (3.9845) [3.1846]	12.7666 (4.5512) [3.2996]	12.6675 (4.9016) [3.4189]	4.3906 (0.7658) [0.6193]	4.3147 (0.8289) [0.6974]	4.1995 (0.9169) [0.7860]
800	21.2026 (7.5896) [5.6381]	21.3579 (7.6956) [5.7519]	21.7652 (8.6432) [6.4576]	24.9908 (9.3165) [6.9324]	25.1628 (10.0739) [8.0160]	26.8321 (12.0559) [8.6939]	12.8467 (3.5802) [2.8018]	12.9886 (4.1186) [3.0938]	12.7717 (4.4042) [3.2062]	4.5190 (0.6725) [0.5373]	4.4598 (0.6914) [0.5651]	4.4162 (0.7651) [0.6219]

Continuité du Tableau 3.2

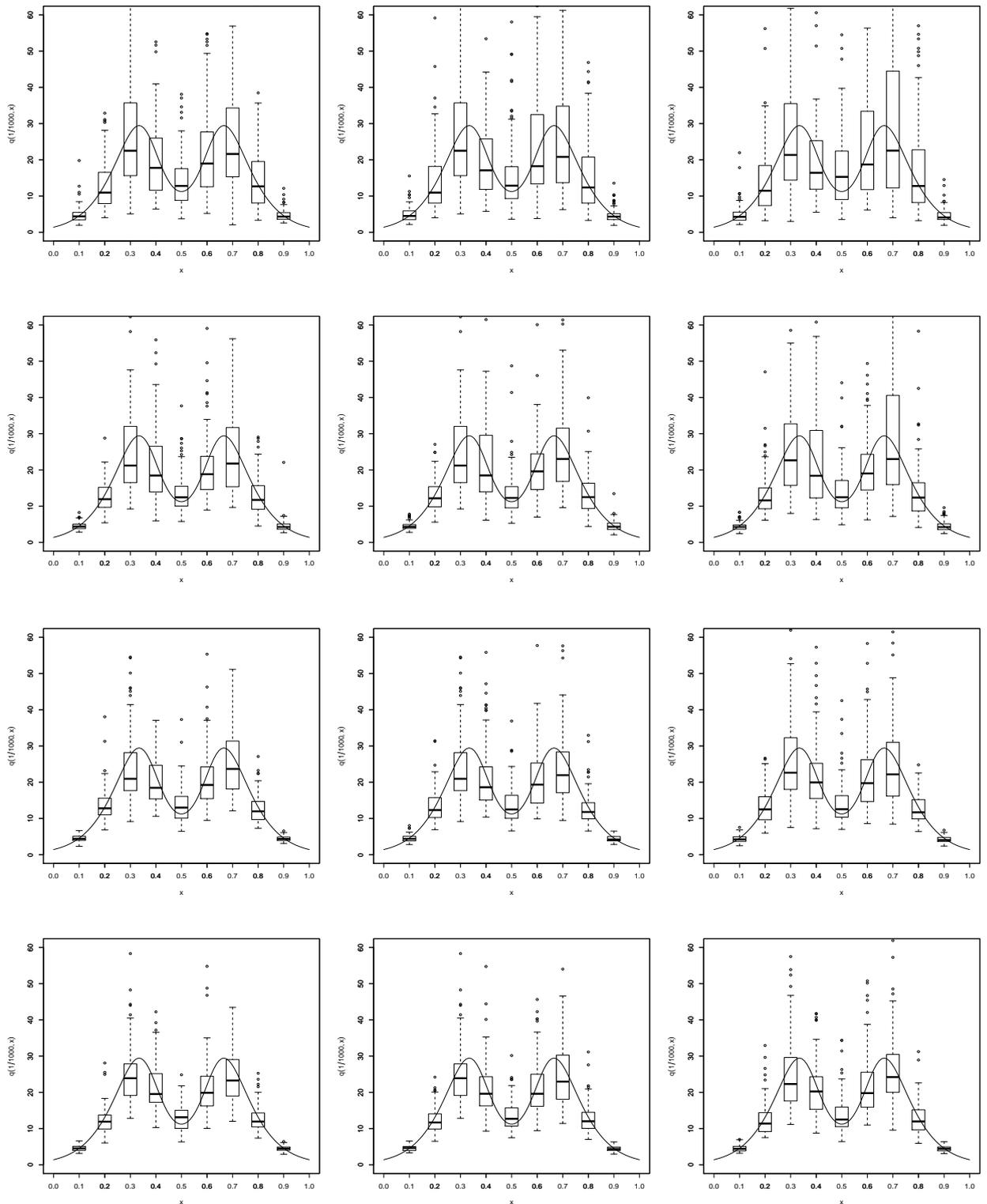


FIGURE 3.3 – Boxplots des R estimations de $q(1/1000|x)$ pour $n = 200$ (1^{ère} ligne), $n = 400$ (2^{ème} ligne), $n = 600$ (3^{ème} ligne), $n = 800$ (4^{ème} ligne). A gauche : $c = 10\%$, Au Centre : $c = 25\%$, A droite : $c = 40\%$.

n	$\gamma_1(.1) = 0.2199$			$\gamma_1(.2) = 0.3728$			$\gamma_1(.3) = 0.4793$			$\gamma_1(.4) = 0.4477$			$\gamma_1(.5) = 0.35$		
	10%	25%	40%	10%	25%	40%	10%	25%	40%	10%	25%	40%	10%	25%	40%
200	0.2014 (0.2137) [0.0034]	0.1719 (0.2344) [0.0041]	0.1556 (0.2993) [0.0069]	0.3181 (0.2939) [0.0104]	0.2720 (0.3251) [0.0141]	0.2099 (0.4053) [0.0309]	0.3799 (0.3262) [0.0153]	0.3107 (0.4122) [0.0390]	0.2453 (0.4830) [0.0613]	0.4112 (0.3130) [0.0134]	0.2953 (0.4716) [0.0447]	0.2229 (0.4741) [0.0548]	0.3689 (0.2556) [0.0060]	0.2828 (0.2756) [0.0080]	0.2208 (0.3625) [0.0203]
400	0.2060 (0.1648) [0.0012]	0.1817 (0.2107) [0.0025]	0.1434 (0.2767) [0.0064]	0.3220 (0.2518) [0.0054]	0.2680 (0.3248) [0.0131]	0.2149 (0.3974) [0.0265]	0.3859 (0.3085) [0.0118]	0.3107 (0.4106) [0.0309]	0.2507 (0.4782) [0.0543]	0.3717 (0.2848) [0.0088]	0.2996 (0.3850) [0.0243]	0.2314 (0.4651) [0.0487]	0.3234 (0.2275) [0.0026]	0.2672 (0.2905) [0.0078]	0.2146 (0.3579) [0.0199]
600	0.2008 (0.1623) [0.0011]	0.1836 (0.1973) [0.0020]	0.1477 (0.2689) [0.0057]	0.3209 (0.2340) [0.0038]	0.2661 (0.3237) [0.0126]	0.2041 (0.3907) [0.0253]	0.3981 (0.2884) [0.0089]	0.3082 (0.4037) [0.0280]	0.2485 (0.4605) [0.0510]	0.3817 (0.2656) [0.0065]	0.3044 (0.3786) [0.0226]	0.2342 (0.4621) [0.0465]	0.3231 (0.2013) [0.0023]	0.2702 (0.2842) [0.0075]	0.2124 (0.3510) [0.0198]
800	0.2039 (0.1564) [0.0008]	0.1753 (0.1870) [0.0018]	0.1446 (0.2545) [0.0052]	0.3186 (0.2338) [0.0037]	0.2690 (0.3223) [0.0116]	0.2086 (0.3752) [0.0228]	0.3863 (0.2759) [0.0073]	0.3218 (0.3969) [0.0261]	0.2379 (0.4514) [0.0491]	0.3735 (0.2639) [0.0062]	0.3091 (0.3723) [0.0205]	0.2338 (0.4525) [0.0447]	0.3267 (0.1967) [0.0021]	0.2697 (0.2839) [0.0072]	0.2099 (0.3443) [0.0202]

Tableau 3.3 – Résultats des simulations pour l'estimateur proposé dans (3.2)(Cas-Complet-Noyau) de $\gamma_1(x)$. Pour chaque scénario, la première ligne donne la moyenne de $\gamma_1(x)$ pour $R = 100$. (\cdot) : Racine carrée de l'erreur quadratique moyenne (RMSE). $[\cdot]$: Erreur absolue moyenne (MAE).

n	$\gamma_1(.6) = 0.4477$			$\gamma_1(.7) = 0.4793$			$\gamma_1(.8) = 3728$			$\gamma_1(.9) = 0.2199$		
	10%	25%	40%	10%	25%	40%	10%	25%	40%	10%	25%	40%
200	0.3698 (0.3029) [0.0121]	0.2973 (0.3900) [0.0284]	0.2272 (0.4707) [0.0534]	0.3902 (0.3265) [0.0169]	0.3245 (0.4066) [0.0309]	0.2370 (0.4922) [0.0636]	0.3126 (0.2681) [0.0072]	0.2594 (0.3394) [0.0161]	0.2144 (0.3994) [0.0289]	0.2091 (0.2111) [0.0031]	0.1806 (0.2194) [0.0034]	0.1492 (0.2735) [0.0071]
400	0.3685 (0.2936) [0.0092]	0.3127 (0.3688) [0.0259]	0.2289 (0.4678) [0.0497]	0.3801 (0.3189) [0.0129]	0.3153 (0.4059) [0.0304]	0.2408 (0.4884) [0.0594]	0.3129 (0.2593) [0.0061]	0.2569 (0.3305) [0.0149]	0.2072 (0.3969) [0.0282]	0.1988 (0.1897) [0.0023]	0.1754 (0.2124) [0.0029]	0.1507 (0.2631) [0.0064]
600	0.3734 (0.2753) [0.0074]	0.3045 (0.3584) [0.0220]	0.2260 (0.4608) [0.0491]	0.3916 (0.2982) [0.0096]	0.3127 (0.4042) [0.0292]	0.2417 (0.4875) [0.0579]	0.3153 (0.2446) [0.0046]	0.2701 (0.3206) [0.0116]	0.2084 (0.3954) [0.0277]	0.2024 (0.1547) [0.0018]	0.1745 (0.2051) [0.0027]	0.1464 (0.2613) [0.0059]
800	0.3749 (0.2723) [0.0070]	0.3010 (0.3430) [0.0186]	0.2313 (0.4552) [0.0483]	0.3894 (0.2905) [0.0095]	0.3167 (0.4033) [0.0280]	0.2412 (0.4863) [0.0576]	0.3231 (0.2358) [0.0035]	0.2684 (0.3132) [0.0112]	0.2128 (0.3814) [0.0263]	0.2023 (0.1453) [0.0008]	0.1742 (0.1958) [0.0025]	0.1487 (0.2579) [0.0055]

Continuité du Tableau 3.3.

n	$q(1/1000 .1) = 4.5695$			$q(1/1000 .2) = 13.1358$			$q(1/1000 .3) = 27.4140$			$q(1/1000 .4) = 22.0358$			$q(1/1000 .5) = 11.2201$		
	10%	25%	40%	10%	25%	40%	10%	25%	40%	10%	25%	40%	10%	25%	40%
200	4.1080 (3.0417) [2.2933]	3.2491 (3.6299) [2.4723]	2.8883 (3.9964) [2.8388]	9.8051 (8.9486) [6.8961]	9.6386 (15.2268) [7.8926]	13.4212 (15.5815) [10.0760]	13.6245 (25.4495) [16.3037]	8.6236 (39.4626) [21.8665]	5.5780 (93.1253) [37.8360]	13.5252 (14.3937) [11.1897]	7.7951 (15.8764) [15.2898]	4.5972 (67.5979) [27.4387]	11.0114 (13.5663) [7.8146]	6.9258 (21.2262) [11.1363]	4.3742 (22.9915) [13.8459]
400	4.0718 (2.0162) [0.8698]	3.4227 (2.3928) [1.2616]	2.5847 (3.0292) [1.9848]	9.2295 (5.2196) [4.7442]	6.2094 (7.2136) [6.9404]	4.2650 (8.9502) [8.8708]	13.9239 (15.4528) [15.5384]	8.3275 (20.2842) [18.0865]	5.4401 (25.0422) [20.9739]	12.6516 (10.5159) [9.7117]	7.5983 (14.7007) [14.4417]	4.7965 (27.3086) [17.2394]	9.0064 (6.5536) [5.0531]	6.1299 (8.3824) [7.1533]	4.2915 (9.6400) [9.5287]
600	3.9357 (0.9810) [0.8568]	3.4370 (1.2672) [1.1605]	2.7061 (1.9026) [1.8634]	8.7731 (4.8776) [4.5176]	6.0104 (7.1786) [6.1254]	4.9871 (8.1874) [8.1487]	15.2458 (13.2615) [12.3215]	8.1524 (19.4176) [17.2616]	5.4168 (20.0551) [19.9972]	13.1873 (9.7162) [8.9936]	7.9479 (14.3171) [14.0880]	4.8380 (17.2845) [17.1978]	9.1764 (6.4945) [4.9579]	6.1949 (8.2429) [7.0686]	4.1538 (9.1162) [9.0664]
800	3.9981 (0.8457) [0.7563]	3.2871 (1.2627) [1.0877]	2.6670 (1.8285) [1.5025]	8.6456 (4.7951) [4.5147]	6.1839 (7.0478) [5.9519]	4.0785 (8.0909) [8.0573]	14.0891 (12.9882) [11.3990]	8.8163 (18.7219) [16.5977]	4.9508 (19.4846) [19.4632]	12.6915 (8.9834) [8.3843]	8.1450 (14.0463) [13.8909]	4.8126 (17.2540) [16.2232]	9.2001 (4.8982) [4.4539]	6.2326 (5.1481) [6.9889]	4.1118 (7.1425) [8.1084]

Tableau 3.4 – Résultats de simulation pour $q(1/1000|x)$ (Cas-Complet-Noyau). Pour toute configuration avec les différents paramètres (n, c, x) , première ligne l'estimation de $q(1/1000|x)$ par l'estimateur proposé dans (3.6) pour $R = 100$. (\cdot) : RMSE. $[\cdot]$: MAE.

n	$q(1/1000 .6) = 22.0358$			$q(1/1000 .7) = 27.4140$			$q(1/1000 .8) = 13.1358$			$q(1/1000 .9) = 4.5695$		
	10%	25%	40%	10%	25%	40%	10%	25%	40%	10%	25%	40%
200	12.7008 (22.1353) [12.9255]	7.8237 (24.8535) [15.2122]	4.7413 (41.5405) [22.3945]	15.9442 (44.3813) [21.5389]	9.7660 (51.9383) [27.2877]	5.1182 (58.4641) [29.2957]	8.7584 (12.7927) [7.2900]	5.9670 (18.5359) [11.2382]	4.2613 (19.0613) [13.8745]	4.1983 (2.4890) [2.3626]	3.4545 (2.5222) [2.5279]	2.7492 (3.0214) [2.8785]
400	12.7121 (11.9283) [10.2847]	8.4974 (17.9500) [13.6622]	4.7249 (19.3630) [17.3110]	13.5699 (14.8161) [14.1996]	8.9238 (17.9546) [17.5721]	5.1427 (36.3483) [22.2713]	8.6218 (5.5306) [5.0160]	5.7283 (7.5235) [7.4075]	3.9875 (9.0031) [9.1483]	3.8840 (1.9958) [1.8479]	3.2568 (2.4203) [2.3574]	2.7641 (2.8869) [2.8754]
600	12.9843 (9.9377) [9.2349]	7.8050 (13.3955) [13.2308]	4.5757 (17.3035) [17.2601]	14.4691 (13.6632) [12.9716]	8.3937 (17.1534) [16.0203]	5.0858 (21.3662) [20.9282]	8.5311 (5.1311) [4.7467]	6.2330 (7.0562) [6.9028]	4.0638 (8.8028) [9.0720]	3.9726 (0.8761) [0.7527]	3.2510 (1.3547) [1.3308]	2.6818 (1.8280) [1.8577]
800	13.0067 (9.7905) [9.0793]	7.7419 (12.4117) [12.2940]	4.8322 (17.2368) [17.2036]	14.1145 (12.8020) [12.3011]	8.6326 (16.9134) [15.7814]	5.1386 (20.3035) [20.2754]	9.0117 (4.5578) [4.1976]	6.1506 (7.0047) [6.0852]	4.1788 (7.9912) [8.9570]	3.9813 (0.8668) [0.7406]	3.2770 (1.3321) [1.3160]	2.7368 (1.7768) [1.8327]

Continuité du Tableau 3.4.

n	$\gamma_1(.1) = 0.2199$			$\gamma_1(.2) = 0.3728$			$\gamma_1(.3) = 0.4793$			$\gamma_1(.4) = 0.4477$			$\gamma_1(.5) = 0.35$		
	10%	25%	40%	10%	25%	40%	10%	25%	40%	10%	25%	40%	10%	25%	40%
200	0.2096 (0.2118) [0.0041]	0.1794 (0.2192) [0.0050]	0.1510 (0.2644) [0.0058]	0.3094 (0.2686) [0.0071]	0.2632 (0.3311) [0.0141]	0.2106 (0.4128) [0.0288]	0.3800 (0.3433) [0.0151]	0.3097 (0.4321) [0.0421]	0.2372 (0.5021) [0.0640]	0.3765 (0.2863) [0.0093]	0.3063 (0.3760) [0.0240]	0.2355 (0.4647) [0.0476]	0.3236 (0.2455) [0.0046]	0.2664 (0.2978) [0.0098]	0.2054 (0.3803) [0.0227]
400	0.2071 (0.1563) [0.0015]	0.1828 (0.2092) [0.0026]	0.1527 (0.2594) [0.0050]	0.3199 (0.2431) [0.0046]	0.2665 (0.3261) [0.0125]	0.2078 (0.4063) [0.0279]	0.3727 (0.3278) [0.0140]	0.3000 (0.4235) [0.0339]	0.2302 (0.4991) [0.0633]	0.3718 (0.2845) [0.0088]	0.3032 (0.3702) [0.0231]	0.2330 (0.4634) [0.0473]	0.3269 (0.2182) [0.0042]	0.2724 (0.2807) [0.0073]	0.2175 (0.3640) [0.0183]
600	0.2038 (0.1517) [0.0010]	0.1788 (0.2067) [0.0023]	0.1480 (0.2483) [0.0045]	0.3181 (0.2377) [0.0040]	0.2642 (0.3196) [0.0122]	0.2057 (0.4048) [0.0274]	0.3988 (0.2873) [0.0082]	0.3201 (0.3990) [0.0265]	0.2425 (0.4866) [0.0568]	0.3623 (0.2740) [0.0078]	0.2974 (0.3678) [0.0227]	0.2287 (0.3686) [0.0386]	0.3217 (0.1965) [0.0026]	0.2705 (0.2719) [0.0070]	0.2107 (0.3533) [0.0179]
800	0.2019 (0.1431) [0.0008]	0.1790 (0.2033) [0.0021]	0.1504 (0.2137) [0.0031]	0.3192 (0.2356) [0.0039]	0.2676 (0.3144) [0.0118]	0.2098 (0.4037) [0.0270]	0.3911 (0.2783) [0.0065]	0.3184 (0.3012) [0.0240]	0.2430 (0.3861) [0.0564]	0.3734 (0.2728) [0.0066]	0.3033 (0.3400) [0.0217]	0.2342 (0.3621) [0.0360]	0.3245 (0.1963) [0.0021]	0.2699 (0.2630) [0.0063]	0.2127 (0.3406) [0.0162]

Tableau 3.5 – Résultats des simulations pour l'estimateur proposé dans (3.2)(Cas-Ignoré-Noyau) de $\gamma_1(x)$. Pour chaque scénario, la première ligne donne la moyenne de $\gamma_1(x)$ pour $R = 100$. (\cdot) : Racine carrée de l'erreur quadratique moyenne (RMSE). $[\cdot]$: Erreur absolue moyenne (MAE).

n	$\gamma_1(.6) = 0.4477$			$\gamma_1(.7) = 0.4793$			$\gamma_1(.8) = 3728$			$\gamma_1(.9) = 0.2199$		
	10%	25%	40%	10%	25%	40%	10%	25%	40%	10%	25%	40%
200	0.3689 (0.2958) [0.0105]	0.2971 (0.3881) [0.0259]	0.2291 (0.4676) [0.0497]	0.3946 (0.3180) [0.0141]	0.3165 (0.4070) [0.0316]	0.2432 (0.4959) [0.0589]	0.3253 (0.2616) [0.0083]	0.2763 (0.3198) [0.0128]	0.2133 (0.4094) [0.0293]	0.2031 (0.1920) [0.0024]	0.1784 (0.2157) [0.0028]	0.1529 (0.2590) [0.0053]
400	0.3691 (0.2868) [0.0089]	0.3038 (0.3793) [0.0228]	0.2352 (0.4609) [0.0465]	0.3882 (0.3105) [0.0121]	0.3149 (0.4055) [0.0293]	0.2370 (0.4923) [0.0598]	0.3184 (0.2423) [0.0049]	0.2647 (0.3288) [0.0132]	0.2068 (0.4074) [0.0285]	0.1994 (0.1776) [0.0017]	0.1760 (0.2129) [0.0026]	0.1470 (0.2202) [0.0057]
600	0.3702 (0.2808) [0.0079]	0.2998 (0.3746) [0.0211]	0.2298 (0.4568) [0.0443]	0.3908 (0.2980) [0.0096]	0.3116 (0.4035) [0.0291]	0.2366 (0.4827) [0.0597]	0.3187 (0.2387) [0.0042]	0.2649 (0.3285) [0.0124]	0.2089 (0.4049) [0.0274]	0.2028 (0.1590) [0.0009]	0.1781 (0.2045) [0.0021]	0.1499 (0.2147) [0.0052]
800	0.3742 (0.2712) [0.0067]	0.3057 (0.3668) [0.0210]	0.2322 (0.4542) [0.0412]	0.3888 (0.2204) [0.0092]	0.3164 (0.4017) [0.0273]	0.2405 (0.4187) [0.0575]	0.3213 (0.2285) [0.0034]	0.2669 (0.3255) [0.0117]	0.2104 (0.4030) [0.0267]	0.2053 (0.1406) [0.0006]	0.1804 (0.1997) [0.0018]	0.1503 (0.2040) [0.0051]

Continuité du Tableau 3.5.

n	$q(1/1000 .1) = 4.5695$			$q(1/1000 .2) = 13.1358$			$q(1/1000 .3) = 27.4140$			$q(1/1000 .4) = 22.0358$			$q(1/1000 .5) = 11.2201$		
	10%	25%	40%	10%	25%	40%	10%	25%	40%	10%	25%	40%	10%	25%	40%
200	4.2897 (2.3776) [2.1389]	3.4537 (2.4435) [2.4090]	2.8371 (2.8471) [2.7546]	8.7684 (8.5582) [6.9917]	6.2767 (15.1524) [9.8591]	4.3502 (17.8639) [10.7856]	15.0531 (24.5730) [17.2359]	9.0258 (38.8192) [22.4204]	5.3598 (95.7441) [42.0542]	14.3938 (15.3041) [11.2257]	8.6377 (18.8140) [15.8982]	5.2354 (68.8797) [26.8004]	9.7966 (13.2965) [8.0146]	6.5747 (15.4608) [9.0510]	4.2460 (27.1106) [13.9742]
400	4.0967 (1.8949) [1.7477]	3.4640 (2.2388) [2.1237]	2.8344 (2.7832) [2.7351]	9.1486 (4.8504) [4.4303]	6.2933 (7.0134) [6.8425]	4.1896 (8.5750) [8.7462]	13.5288 (15.7748) [13.9894]	8.1083 (20.4761) [19.3057]	4.9534 (25.4959) [22.4606]	13.5414 (12.0645) [9.1570]	8.3314 (16.9884) [13.7044]	5.0511 (26.0269) [15.9848]	9.5372 (6.2990) [4.8359]	6.5826 (6.9478) [6.6975]	4.4676 (9.9912) [7.7526]
600	4.0016 (0.8834) [0.7296]	3.3798 (1.2082) [1.2090]	2.7494 (1.7530) [1.7201]	8.8928 (4.7068) [4.3913]	6.1247 (7.0079) [6.0111]	4.1215 (8.0362) [8.0143]	15.8611 (12.5660) [11.9221]	9.1643 (18.3917) [18.2497]	5.3512 (21.0892) [21.0628]	12.2564 (9.9940) [8.9485]	7.8021 (14.3552) [13.2338]	4.8465 (15.2116) [15.1894]	9.1106 (5.8370) [4.4512]	6.3891 (6.9393) [5.8311]	4.2628 (6.9889) [5.9573]
800	3.9709 (0.8421) [0.6318]	3.3922 (1.1631) [1.1845]	2.7954 (1.7023) [1.6741]	9.0251 (4.6250) [4.2471]	6.3347 (6.9185) [5.8011]	4.2389 (7.9195) [7.8969]	15.0829 (11.1630) [10.4731]	9.0157 (18.1203) [17.3983]	5.3787 (20.0558) [20.0353]	13.1601 (9.3573) [8.8834]	8.1165 (14.0095) [12.9193]	5.0220 (15.0295) [15.0138]	9.3226 (4.7082) [3.2976]	6.4006 (5.7484) [4.8196]	4.3235 (5.9203) [5.8967]

Tableau 3.6 – Résultats de simulation pour $q(1/1000|x)$ (Cas-Ignoré-Noyau). Pour toute configuration avec les différents paramètres (n, c, x) , première ligne donne l'estimation de $q(1/1000|x)$ par l'estimateur proposé dans (3.6) pour $R = 100$. (\cdot) : RMSE. $[\cdot]$: MAE.

n	$q(1/1000 .6) = 22.0358$			$q(1/1000 .7) = 27.4140$			$q(1/1000 .8) = 13.1358$			$q(1/1000 .9) = 4.5695$		
	10%	25%	40%	10%	25%	40%	10%	25%	40%	10%	25%	40%
200	13.5598 (20.7934) [15.7558]	8.1817 (24.2654) [19.8598]	5.0081 (47.1019) [23.0277]	17.6563 (46.2637) [23.8800]	9.9514 (52.5574) [28.1125]	5.7560 (58.8686) [31.6580]	9.9470 (11.6445) [7.0836]	7.0556 (18.8980) [9.4752]	4.4594 (19.7960) [10.6764]	4.0331 (1.9243) [1.6999]	3.4000 (2.0226) [1.9085]	2.8543 (2.4991) [2.3158]
400	13.1138 (14.2812) [13.3751]	8.3195 (17.9705) [17.7164]	5.1619 (20.9315) [20.8739]	15.4384 (17.0349) [15.0190]	9.0572 (22.6133) [19.3568]	5.1732 (36.2739) [26.2408]	9.1203 (6.8832) [6.3280]	6.2692 (8.2716) [7.8666]	4.2023 (9.9819) [9.9335]	3.9124 (1.6807) [1.5269]	3.3383 (1.9761) [1.8638]	2.7317 (2.4784) [2.2378]
600	13.0675 (13.8580) [13.1027]	7.9925 (17.1865) [17.0433]	4.8972 (19.1670) [19.1387]	14.9991 (15.1700) [14.4417]	8.6069 (18.9371) [17.8071]	5.1462 (22.2930) [23.2178]	8.9504 (6.1652) [6.1672]	6.1782 (7.8561) [7.2576]	4.2053 (8.9552) [9.5305]	3.9653 (0.8865) [0.7655]	3.3457 (1.2941) [1.2238]	2.7773 (1.8213) [1.7922]
800	13.3315 (13.2773) [12.7043]	8.2681 (16.8695) [16.7677]	4.9618 (19.0896) [19.0740]	14.6017 (14.1858) [13.8123]	8.8815 (17.9119) [17.5325]	5.2610 (22.1671) [23.1530]	9.1107 (6.0907) [6.0647]	6.2694 (7.1299) [7.1664]	4.2371 (8.9132) [8.4987]	4.0549 (0.7301) [0.6367]	3.4273 (1.2121) [1.1474]	2.7940 (1.7983) [1.7755]

Continuité du Tableau 3.6.

3.6 Démonstrations

Nous proposons d'abord quelques notations supplémentaires et lemmes techniques intermédiaires. Les preuves de la proposition 3.2.1 et le théorème 3.2.1 sont indiquées au paragraphe 3.6.2.

3.6.1 Résultats préliminaires

Notons tout d'abord que les expressions $\bar{H}_n(t|x)$ et $\bar{H}_n^1(t|x)$ dans 3.2 peuvent être réécrites comme $\bar{H}_n(t|x) = \hat{\varphi}_n(t, x)/\hat{g}_n(x)$ et $\bar{H}_n^1(t|x) = \hat{\psi}_n(t, x)/\hat{g}_n(x)$ respectivement où $\hat{\varphi}_n(t, x) = \frac{1}{n} \sum_{i=1}^n K_h(x - X_i) \mathbb{1}_{\{Z_i > t\}}$, $\hat{\psi}_n(t, x) = \frac{1}{n} \sum_{i=1}^n K_h(x - X_i) \mathbb{1}_{\{Z_i > t, \delta_i = 1\}}$ et $\hat{g}_n(x) = \frac{1}{n} \sum_{i=1}^n K_h(x - X_i)$ est l'estimateur non-paramétrique classique de la fonction de densité g .

Rappelons que $\bar{H}(t|x) = \mathbb{P}(Z > t | X = x)$ est la fonction de survie conditionnelle de Z sachant $X = x$ et soit $\bar{H}^1(t|x) = \mathbb{P}(Z > t, \delta = 1 | X = x)$ la fonction de survie conditionnelle de la sous-distribution de Z . Nous définissons encore $\varphi(t, x) = \bar{H}(t|x)g(x)$, $\psi(t, x) = \bar{H}^1(t|x)g(x)$,

$$\mathbb{W}_{n,\varphi}(x) := \sqrt{nh^p \varphi(t_n, x)} \left(\frac{\hat{\varphi}_n(t_n, x) - \mathbb{E}[\hat{\varphi}_n(t_n, x)]}{\varphi(t_n, x)} \right)$$

et

$$\mathbb{W}_{n,\psi}(x) := \sqrt{nh^p \psi(t_n, x)} \left(\frac{\hat{\psi}_n(t_n, x) - \mathbb{E}[\hat{\psi}_n(t_n, x)]}{\psi(t_n, x)} \right).$$

Enfin, dans ce qui suit, \top désignera la transposée. Lemme 3.6.1 prône sur la limite de $\psi(t, x)/\varphi(t, x)$ quand $t \rightarrow \infty$. Un lemme similaire est obtenu dans Brahimi et al. (2013, [19]) dans le cas inconditionnel.

Lemme 3.6.1 *Sous l'hypothèse (H1) et soit $x \in \mathcal{X}$. Alors,*

$$\lim_{t \rightarrow \infty} \psi(t, x)/\varphi(t, x) = p_x.$$

Preuve

Notons tout d'abord que $\psi(t, x)/\varphi(t, x) = \bar{H}^1(t|x)/\bar{H}(t|x)$. Un simple calcul donne $\bar{H}^1(t|x) = \int_t^\infty \bar{G}(u|x) dF(u|x)$. Le changement de variable suivant $\bar{F}(u|x) = 1/v$ nous

donne

$$\begin{aligned}\bar{H}^1(t|x) &= \int_{1/\bar{F}(t|x)}^{\infty} v^{-2} \bar{G}(F^{\leftarrow}(1-1/v|x)|x) dv \\ &:= \int_{1/\bar{F}(t|x)}^{\infty} v^{-1} \bar{R}(v|x) dv,\end{aligned}$$

où $v \mapsto \bar{R}(v|x)$ est une fonction à variations régulières d'indice $-(1 + \gamma_1(x)/\gamma_2(x))$. En utilisant le Théorème 1.2.2 dans de Haan et Ferreira (2006, [35]) pour avoir,

$$\bar{H}^1(t|x) = \int_{1/\bar{F}(t|x)}^{\infty} v^{-1} \bar{R}(v|x) dv \sim \left(1 + \frac{\gamma_1(x)}{\gamma_2(x)}\right)^{-1} \bar{R}(1/\bar{F}(t|x)|x)$$

quand t est assez grand. Ce qui donne, $\bar{R}(1/\bar{F}(t|x)|x) = \bar{F}(t|x) \bar{G}(t|x) = \bar{H}(t|x)$ donc

$$\bar{H}^1(t|x) \sim \left(1 + \frac{\gamma_1(x)}{\gamma_2(x)}\right)^{-1} \bar{H}(t|x)$$

quand t est assez grand, ce qui implique $\lim_{t \rightarrow \infty} \bar{H}^1(t|x)/\bar{H}(t|x) = \gamma_2(x)/(\gamma_1(x) + \gamma_2(x))$, c'est-à-dire, $\lim_{t \rightarrow \infty} \psi(t, x)/\varphi(t, x) = p_x$. \square

Lemme 3.6.2 *Supposons (H1), (H2) et (H3) sont vérifiées. Si $t_n \rightarrow \infty$ et $h \log t_n \rightarrow 0$ quand $n \rightarrow \infty$, alors*

$$\sup_{d(x, x') \leq h} \left| \frac{\bar{H}(t_n|x)}{\bar{H}(t_n|x')} - 1 \right| = O(h \log t_n).$$

Preuve

La preuve de ce lemme est similaire à celle du Lemme 1 dans Daouia et al. (2011, [30]). La condition (H1) donne

$$\begin{aligned}\left| \log \left(\frac{\bar{H}(t_n|x)}{\bar{H}(t_n|x')} \right) \right| &\leq |\log t_n| \left(\left| \frac{1}{\gamma(x)} - \frac{1}{\gamma(x')} \right| + \left| \frac{\log L(t_n|x)}{\log t_n} - \frac{\log L(t_n|x')}{\log t_n} \right| \right) \\ &\leq (c_\gamma + c_L) \log t_n d(x, x'),\end{aligned}$$

Comme pour n assez grand, $t_n \geq 1$, alors

$$\sup_{d(x, x') \leq h} \left| \log \left(\frac{\bar{H}(t|x)}{\bar{H}(t|x')} \right) \right| = O(h \log t) \rightarrow 0$$

et il suffit de remarquer que $\log(x+1) \sim x$ quand $x \rightarrow 0$ pour avoir le résultat. \square

Lemme 3.6.3 *Supposons (H1), (H2) et (H3) sont vérifiées. Si $t_n \rightarrow \infty$ et $h \log t_n \rightarrow 0$*

quand $n \rightarrow \infty$, alors

$$\sup_{d(x,x') \leq h} \left| \frac{\bar{H}^1(t_n|x)}{\bar{H}^1(t_n|x')} - 1 \right| = O(h \log t_n).$$

Preuve

D'après le Lemme 3.6.1 nous avons

$$\bar{H}^1(t_n|x) \sim \left(1 + \frac{\gamma_1(x)}{\gamma_2(x)} \right)^{-1} \bar{H}(t_n|x) := a(x) \bar{H}(t_n|x).$$

Autrement dit

$$\begin{aligned} \frac{\bar{H}^1(t_n|x)}{\bar{H}^1(t_n|x')} - 1 &\sim \frac{a(x)}{a(x')} \frac{\bar{H}(t_n|x)}{\bar{H}(t_n|x')} - 1 \\ &\sim \frac{a(x)}{a(x')} \left[\frac{\bar{H}(t_n|x)}{\bar{H}(t_n|x')} \right] + \frac{a(x)}{a(x')} - 1 \end{aligned}$$

ce qui entraîne

$$\begin{aligned} \sup_{d(x,x') \leq h} \left| \frac{\bar{H}^1(t_n|x)}{\bar{H}^1(t_n|x')} - 1 \right| &\sim \sup_{d(x,x') \leq h} \left| \frac{a(x)}{a(x')} \left[\frac{\bar{H}(t_n|x)}{\bar{H}(t_n|x')} - 1 \right] + \frac{a(x)}{a(x')} - 1 \right| \\ &\leq \sup_{d(x,x') \leq h} \left| \frac{a(x)}{a(x')} \left[\frac{\bar{H}(t_n|x)}{\bar{H}(t_n|x')} - 1 \right] \right| + \sup_{d(x,x') \leq h} \left| \frac{a(x)}{a(x')} - 1 \right| \\ &\leq \underbrace{\sup_{d(x,x') \leq h} \left| \frac{a(x)}{a(x')} \right|}_{\mathbb{A}} \times \underbrace{\sup_{d(x,x') \leq h} \left| \frac{\bar{H}(t|x)}{\bar{H}(t|x')} - 1 \right|}_{\mathbb{B}} + \underbrace{\sup_{d(x,x') \leq h} \left| \frac{a(x)}{a(x')} - 1 \right|}_{\mathbb{C}}. \end{aligned}$$

D'après le Lemme 3.6.2, $\mathbb{B} = O(h \log t_n)$. Si $d(x, x') \leq h$ et que $h \rightarrow 0$ on a $d(x, x') \rightarrow 0$. Et les fonctions $\gamma_1(\cdot)$ et $\gamma_2(\cdot)$ sont continues alors $\frac{a(x)}{a(x')} \rightarrow 1$. Donc $\mathbb{A} \rightarrow 1$ et $\mathbb{C} \rightarrow 0$, d'où le résultat. \square

Le lemme suivant est dédié aux propriétés asymptotiques de $\hat{\varphi}_n(t, x)$ et $\hat{\psi}_n(t, x)$.

Lemme 3.6.4 *Supposons les hypothèses (H1)-(H6) satisfaites. Soit t_n une suite non aléatoire telle que $t_n \rightarrow \infty$, $h \log t_n \rightarrow 0$ et $nh^p \bar{H}(t_n|x) \rightarrow \infty$ quand $n \rightarrow \infty$. Alors, pour tout $x \in \mathcal{X}$ tel que $g(x) > 0$,*

1. $\mathbb{E}[\hat{\varphi}_n(t_n, x)] = \varphi(t_n, x)(1 + O(h \log t_n))$ et $\mathbb{E}[\hat{\psi}_n(t_n, x)] = \psi(t_n, x)(1 + O(h \log t_n))$.
2. Quand $n \rightarrow \infty$, $\mathbb{W}_{n,\varphi}(x) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \|K\|_2^2)$ et $\mathbb{W}_{n,\psi}(x) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \|K\|_2^2)$. Par ailleurs comme $n \rightarrow \infty$, $\mathbb{W}_n(x) := (\mathbb{W}_{n,\varphi}(x), \mathbb{W}_{n,\psi}(x))^\top$ converge en distribution vers un vecteur gaussien bivarié $\mathcal{N}(0, M)$ où

$$M := \begin{pmatrix} \|K\|_2^2 & \|K\|_2^2 \sqrt{p_x} \\ \|K\|_2^2 \sqrt{p_x} & \|K\|_2^2 \end{pmatrix}.$$

Preuve

1. Nous montrons que $\mathbb{E}[\hat{\varphi}_n(t_n, x)] = \varphi(t_n, x)(1 + O(h \log t_n))$. Les observations (Z_i, X_i, δ_i) , $i = 1, \dots, n$ sont identiquement distribuées donc

$$\begin{aligned}\mathbb{E}[\hat{\varphi}_n(t_n, x)] &= \int_{\mathbb{R}^p} K_h(x-s) \bar{H}(t_n|s) g(s) ds \\ &= \int_S K(u) \bar{H}(t_n|x-hu) g(x-hu) du\end{aligned}$$

sous **(H6)**. Il s'ensuit que

$$\begin{aligned}|\mathbb{E}[\hat{\varphi}_n(t_n, x)] - \varphi(t_n, x)| &\leq \bar{H}(t_n|x) \int_S K(u) |g(x-hu) - g(x)| du \\ &\quad + \bar{H}(t_n|x) \int_S K(u) \left| \frac{\bar{H}(t_n|x-hu)}{\bar{H}(t_n|x)} - 1 \right| g(x-hu) du \\ &:= A_{1,x} + A_{2,x}.\end{aligned}$$

Sous **(H4)** puisque $g(x) > 0$, nous avons

$$A_{1,x} \leq \bar{H}(t_n|x) c_g h \int_S d(u, 0) K(u) du = \varphi(t_n, x) O(h). \quad (3.9)$$

En utilisant le Lemme 3.6.2, nous avons

$$\sup_{u \in S} \left| \frac{\bar{H}(t_n|x-hu)}{\bar{H}(t_n|x)} - 1 \right| = O(h \log t_n)$$

donc

$$\begin{aligned}A_{2,x} &= \bar{H}(t_n|x) O(h \log t_n) \int_S K(u) g(x-hu) du \\ &= \varphi(t_n, x) O(h \log t_n).\end{aligned} \quad (3.10)$$

Combinons (3.9) et (3.10) pour conclure la preuve.

Remarque 5 Nous procédons de la même manière pour avoir

$$\mathbb{E}[\hat{\psi}_n(t_n, x)] = \psi(t_n, x)(1 + O(h \log t_n)).$$

□

2. Nous montrons d'abord la normalité asymptotique de $\mathbb{W}_{n,\varphi}(x)$. Soit

$$\mathbb{W}_{n,\varphi}(x) := \frac{\hat{\varphi}_n(t_n, x) - \mathbb{E}[\hat{\varphi}_n(t_n, x)]}{\varphi(t_n, x) \Lambda_n(x)}$$

où $\Lambda_n(x) = (nh^p \varphi(t_n, x))^{-1/2}$. Par des calculs simples nous pouvons réécrire

$\mathbb{W}_{n,\varphi}(x)$ ainsi :

$$\begin{aligned}\mathbb{W}_{n,\varphi}(x) &= \frac{1}{\varphi(t_n, x)\Lambda_n(x)} \frac{1}{n} \sum_{i=1}^n \left\{ K_h(x - X_i) \mathbb{1}_{\{Z_i > t_n\}} - \mathbb{E} \left[K_h(x - X_i) \mathbb{1}_{\{Z_i > t_n\}} \right] \right\} \\ &:= \frac{1}{\varphi(t_n, x)\Lambda_n(x)n} \sum_{i=1}^n T_{i,n},\end{aligned}$$

c'est-à-dire comme les sommes partielles dans la matrice triangulaire de variables aléatoires $(T_{i,n}, i = 1, \dots, n)$, $n \geq 1$ où pour n fixé, $T_{i,n}$, $i = 1, \dots, n$ sont des variables aléatoires centrées indépendantes et identiquement distribuées. Ainsi, pour établir la normalité asymptotique de $\mathbb{W}_{n,\varphi}(x)$, nous vérifions la condition de Lyapunov pour les tableaux triangulaires de variables aléatoires. Premièrement, nous calculons d'abord la variance de $\mathbb{W}_{n,\varphi}(x)$. Nous avons :

$$\begin{aligned}\text{var}(T_{i,n}) &= \text{var} \left(K_h(x - X_i) \mathbb{1}_{\{Z_i > t_n\}} \right) \\ &= \mathbb{E} \left[K_h^2(x - X_i) \mathbb{1}_{\{Z_i > t_n\}} \right] - \mathbb{E} \left[K_h(x - X_i) \mathbb{1}_{\{Z_i > t_n\}} \right]^2.\end{aligned}$$

Le Lemme 2 dans Goegebeur et al. (2014b, [59]) implique,

$$\mathbb{E} \left[K_h^2(x - X_i) \mathbb{1}_{\{Z_i > t_n\}} \right] = \varphi(t_n, x) h^{-p} \|K\|_2^2 (1 + O(h \log t_n))$$

et le résultat 1 dans le Lemme 3.6.4 donne,

$$\mathbb{E} \left[K_h(x - X_i) \mathbb{1}_{\{Z_i > t_n\}} \right] = \varphi(t_n, x) (1 + O(h \log t_n)).$$

Il s'ensuit que

$$\text{var}(T_{i,n}) = \varphi(t_n, x) h^{-p} \|K\|_2^2 (1 + O(h \log t_n)) - \varphi^2(t_n, x) (1 + O(h \log t_n))^2 \quad (3.11)$$

et donc nous avons,

$$\begin{aligned}\text{var}(\mathbb{W}_{n,\varphi}(x)) &= \frac{1}{\varphi^2(t_n, x)\Lambda_n^2(x)n^2} n \text{var}(T_{i,n}) \\ &= \|K\|_2^2 (1 + O(h \log t_n)) - h^p \varphi(t_n, x) (1 + O(h \log t_n))^2.\end{aligned}$$

De cela, on en déduit que $\text{var}(\mathbb{W}_{n,\varphi}(x)) \rightarrow \|K\|_2^2$ quand $n \rightarrow \infty$. Maintenant, nous prouvons que la condition de Lyapunov pour les tableaux triangulaires est satisfaite, donc nous montrons que $\sum_{i=1}^n \mathbb{E} |\tilde{T}_{i,n}|^3 \rightarrow 0$ et $n \rightarrow \infty$ où $\tilde{T}_{i,n} := T_{i,n} / \{\varphi(t_n, x)\Lambda_n(x)\}$. Et de ce fait, il est simple de démontrer que

$$|\tilde{T}_{i,n}| \leq \frac{2\|K\|_\infty}{\sqrt{nh^p\varphi(t_n, x)}}$$

où $\|\cdot\|_\infty$ désigne la norme sup et donc $|\tilde{T}_{i,n}|^3 \leq \frac{2\|K\|_\infty}{\sqrt{nh^p\varphi(t_n,x)}}|\tilde{T}_{i,n}|^2$. Concentrons nous sur les deux côtés de cette inégalité pour avoir

$$\mathbb{E}|\tilde{T}_{i,n}|^3 \leq \frac{2\|K\|_\infty h^{2p}}{(nh^p\varphi(t_n,x))^{3/2}} \text{var}(T_{i,n})$$

et en utilisant (3.11), nous obtenons

$$\sum_{i=1}^n \mathbb{E}|\tilde{T}_{i,n}|^3 \leq \frac{2\|K\|_\infty \|K\|_2^2}{\sqrt{nh^p\varphi(t_n,x)}}(1+o(1)) - \frac{2\|K\|_\infty h^p\varphi(t_n,x)}{\sqrt{nh^p\varphi(t_n,x)}}(1+o(1)) \rightarrow 0$$

quand $n \rightarrow \infty$. La condition de Lyapunov est vérifiée, ainsi $\mathbb{W}_{n,\varphi}(x)/\sqrt{\text{var}(\mathbb{W}_{n,\varphi}(x))} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$. Finalement, $\mathbb{W}_{n,\varphi}(x) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \|K\|_2^2)$.

Remarque 6 Nous procédons de la même manière pour avoir le résultat :

$$\mathbb{W}_{n,\psi}(x) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \|K\|_2^2).$$

□

Nous montrons à présent $\mathbb{W}_n(x) := (\mathbb{W}_{n,\varphi}(x), \mathbb{W}_{n,\psi}(x))^\top$ converge en distribution vers $\mathcal{N}(0, M)$. Selon la formule de Cramér-Wold (1998, [89]), il suffit de prouver que $\ell^\top \mathbb{W}_n(x) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \ell^\top M \ell)$ pour tout $\ell = (\ell_1, \ell_2)^\top \in \mathbb{R}^2$, $\ell \neq 0$. Un calcul simple implique :

$$\ell^\top \mathbb{W}_n(x) := \frac{1}{n\varphi(t_n,x)\Lambda_n(x)} \sum_{i=1}^n T_{i,n}^*$$

où

$$T_{i,n}^* = K_h(x - X_i) \left\{ \ell_1 \mathbb{1}_{\{Z_i > t_n\}} + \ell_2 \mathbb{1}_{\{Z_i > t_n, \delta_i = 1\}} \sqrt{\frac{\varphi(t_n, x)}{\psi(t_n, x)}} \right\} \\ - \mathbb{E} \left[K_h(x - X_i) \left\{ \ell_1 \mathbb{1}_{\{Z_i > t_n\}} + \ell_2 \mathbb{1}_{\{Z_i > t_n, \delta = 1\}} \sqrt{\frac{\varphi(t_n, x)}{\psi(t_n, x)}} \right\} \right].$$

Ces calculs sont similaires à ceux de $\text{var}(T_{i,n})$

$$\text{var}(T_{i,n}^*) = l_1^2 \left(\varphi(t_n, x) h^{-p} \|K\|_2^2 (1 + O(h \log t_n)) - \varphi^2(t_n, x) (1 + O(h \log t_n))^2 \right) \\ + l_2^2 \left(p_x^{-1} + o(1) \right) \left(\psi(t_n, x) h^{-p} \|K\|_2^2 (1 + O(h \log t_n)) \right. \\ \left. - \psi^2(t_n, x) (1 + O(h \log t_n))^2 \right) \\ + 2l_1 l_2 \left(p_x^{-1/2} + o(1) \right) \left(\psi(t_n, x) h^{-p} \|K\|_2^2 (1 + O(h \log t_n)) \right. \\ \left. - \varphi(t_n, x) \psi(t_n, x) (1 + O(h \log t_n))^2 \right),$$

ce qui donne donc

$$\begin{aligned} \lim_{n \rightarrow \infty} \text{var}(\ell^\top \mathbb{W}_n(x)) &= \lim_{n \rightarrow \infty} \frac{1}{h^{-p} \varphi(t_n, x)} \text{var}(T_{i,n}^*) \\ &= \ell_1^2 \|K\|_2^2 + \ell_2^2 \|K\|_2^2 + 2\ell_1 \ell_2 \|K\|_2^2 \sqrt{p_x} \\ &= \ell^\top M \ell. \end{aligned}$$

Déterminons la normalité asymptotique de $\ell^\top \mathbb{W}_n(x)$ en utilisant la condition de Lyapunov : $\lim_{n \rightarrow \infty} \sum_{i=1}^n \mathbb{E}|T_{i,n}^* / \{\varphi(t_n, x) \Lambda_n(x) n\}|^3 = 0$. Alors, pour tout $\ell \in \mathbb{R}^2$, $\ell \neq 0$, $\ell^\top \mathbb{W}_n(x)$ converge en distribution vers la distribution normale univariée $\mathcal{N}(0, \ell^\top M \ell)$. La formule de Cramér-Wold implique finalement que $\mathbb{W}_n(x) \xrightarrow{\mathcal{D}} \mathcal{N}(0, M)$. \square

3.6.2 Démonstrations des principaux résultats

Preuve de la Proposition 3.2.1

Nous décomposons comme suit :

$$\begin{aligned} \sqrt{nh^p \bar{H}(t_n|x)} (\hat{p}_{t_n}(x) - p_x) &= \sqrt{nh^p \bar{H}(t_n|x)} \left(\hat{p}_{t_n}(x) - \frac{\psi(t_n, x)}{\varphi(t_n, x)} \right) \\ &\quad + \sqrt{nh^p \bar{H}(t_n|x)} \left(\frac{\psi(t_n, x)}{\varphi(t_n, x)} - p_x \right). \end{aligned}$$

Nous montrons tout d'abord que le premier terme de cette somme est asymptotiquement normal. Cela résulte de l'application de la delta méthode. Nous avons :

$$\begin{aligned} \sqrt{nh^p \bar{H}(t_n|x)} \begin{pmatrix} \frac{\hat{\psi}_n(t_n, x)}{\varphi(t_n, x)} - \frac{\psi(t_n, x)}{\varphi(t_n, x)} \\ \frac{\hat{\varphi}_n(t_n, x)}{\varphi(t_n, x)} - 1 \end{pmatrix} &= \begin{pmatrix} \frac{1}{\sqrt{g(x)}} \sqrt{\frac{\psi(t_n, x)}{\varphi(t_n, x)}} \mathbb{W}_{n, \psi}(x) \\ \frac{1}{\sqrt{g(x)}} \mathbb{W}_{n, \varphi}(x) \end{pmatrix} \\ &\quad + \begin{pmatrix} \frac{1}{\sqrt{g(x)}} \sqrt{\frac{\psi(t_n, x)}{\varphi(t_n, x)}} \sqrt{nh^p \psi(t_n, x)} \left(\frac{\mathbb{E}[\hat{\psi}_n(t_n, x)]}{\psi(t_n, x)} - 1 \right) \\ \frac{1}{\sqrt{g(x)}} \sqrt{nh^p \varphi(t_n, x)} \left(\frac{\mathbb{E}[\hat{\varphi}_n(t_n, x)]}{\varphi(t_n, x)} - 1 \right) \end{pmatrix}. \end{aligned}$$

Quand $n \rightarrow \infty$, $\psi(t_n, x)/\varphi(t_n, x) \rightarrow p_x$ en utilisant Lemme 3.6.1. Par ailleurs, l'utilisation de l'assertion 1 dans le Lemme 3.6.4 donne, $\mathbb{E}[\hat{\psi}_n(t_n, x)]/\psi(t_n, x) - 1 = O(h \log t_n)$ et $\mathbb{E}[\hat{\varphi}_n(t_n, x)]/\varphi(t_n, x) - 1 = O(h \log t_n)$ (notons que, sous les conditions de la Proposition 3.2.1, la condition $h \log t_n \rightarrow 0$ du Lemme 3.6.4 est satisfaite). Finalement, sous l'hypothèse, $nh^{p+2} \bar{H}(t_n|x) (\log t_n)^2 \rightarrow 0$ quand $n \rightarrow \infty$, nous avons

$$\sqrt{nh^p \bar{H}(t_n|x)} \begin{pmatrix} \frac{\hat{\psi}_n(t_n, x)}{\varphi(t_n, x)} - \frac{\psi(t_n, x)}{\varphi(t_n, x)} \\ \frac{\hat{\varphi}_n(t_n, x)}{\varphi(t_n, x)} - 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{g(x)}} \sqrt{\frac{\psi(t_n, x)}{\varphi(t_n, x)}} \mathbb{W}_{n, \psi}(x) \\ \frac{1}{\sqrt{g(x)}} \mathbb{W}_{n, \varphi}(x) \end{pmatrix} + o(1).$$

L'utilisation de l'assertion 2 dans le Lemme 3.6.4 implique la convergence en loi vers un vecteur gaussien bivarié $\mathcal{N}(0, \Sigma)$ où

$$\Sigma := \begin{pmatrix} \frac{\|K\|_2^2 p_x}{g(x)} & \frac{\|K\|_2^2 p_x}{g(x)} \\ \frac{\|K\|_2^2 p_x}{g(x)} & \frac{\|K\|_2^2}{g(x)} \end{pmatrix}.$$

Nous considérons la fonction $\phi(x, y) = \frac{x}{y}$. La fonction ϕ est différentiable aux points (x, y) (si $y \neq 0$) avec comme dérivée $\phi'(x, y) = (\frac{1}{y}, -\frac{x}{y^2})$. Il résulte de ce delta méthode décrit dans van der Vaart (1998, [89]) que

$$\begin{aligned} & \sqrt{nh^p \overline{H}(t_n|x)} \left(\widehat{p}_{t_n}(x) - \frac{\psi(t_n, x)}{\varphi(t_n, x)} \right) \\ &= \sqrt{nh^p \overline{H}(t_n|x)} \left(\phi \left(\frac{\widehat{\psi}_n(t_n, x)}{\varphi(t_n, x)}, \frac{\widehat{\varphi}_n(t_n, x)}{\varphi(t_n, x)} \right) - \phi \left(\frac{\psi(t_n, x)}{\varphi(t_n, x)}, 1 \right) \right) \end{aligned}$$

converge en distribution vers $\mathcal{N}(0, \phi'(p_x, 1) \Sigma (\phi'(p_x, 1))^\top)$ c'est-à-dire à la distribution normale $\mathcal{N}(0, p_x(1-p_x) \|K\|_2^2 / g(x))$. En utilisant des arguments similaires que dans Beirlant et al. (2007, [9]), on montre que $\sqrt{nh^p \overline{H}(t_n|x)} \left(\frac{\psi(t_n, x)}{\varphi(t_n, x)} - p_x \right) = \sqrt{nh^p \overline{H}(t_n|x)} \left(\frac{\overline{H}^1(t_n|x)}{\overline{H}(t_n|x)} - p_x \right) = o(1)$.

En effet, nous avons pour $t_n \in [1, \infty[$,

$$\begin{aligned} \frac{\int_{t_n}^{\infty} z^{-1/\gamma_1(x)-1/\gamma_2(x)-1} L(z|x) dz}{t_n^{-1/\gamma_1(x)-1/\gamma_2(x)} L(t_n|x)} - \gamma(x) &= \frac{\int_{t_n}^{\infty} \frac{\overline{H}(z|x)}{z} dz}{\overline{H}(t_n|x)} - \gamma(x) \\ &=: \Delta(t_n|x). \end{aligned}$$

Cherchons

$$\lim_{t_n \rightarrow \infty} \int_{t_n}^{\infty} \frac{\overline{H}(u|x)}{\overline{H}(t_n|x)} \frac{du}{u} = \lim_{t_n \rightarrow \infty} \int_1^{\infty} \frac{\overline{H}(t_n s|x)}{\overline{H}(t_n|x)} \frac{ds}{s}$$

or

$$\frac{\overline{H}(t_n s|x)}{\overline{H}(t_n|x)} = \frac{(t_n s)^{-1/\gamma(x)} L(t_n s|x)}{(t_n)^{-1/\gamma(x)} L(t_n|x)} = s^{-1/\gamma(x)} \frac{L(st_n|x)}{L(t_n|x)} \rightarrow s^{-1/\gamma(x)} \text{ si } t_n \rightarrow \infty$$

donc

$$\lim_{t_n \rightarrow \infty} \int_{t_n}^{\infty} \frac{\overline{H}(u|x)}{\overline{H}(t_n|x)} \frac{du}{u} = \int_1^{\infty} s^{-1/\gamma(x)-1} ds = \gamma(x).$$

Ce qui entraîne donc

$$\int_{t_n}^{\infty} \frac{\overline{H}(u|x)}{u} du = \gamma(x) \overline{H}(t_n|x) (1 + O(\Delta(t_n|x))).$$

De plus,

$$\begin{aligned}\bar{H}^1(t_n|x) &= \mathbb{P}(\min(Y, C) > t_n, \delta = 1 | X = x) \\ &= \mathbb{P}(Y > t_n, Y - C < 0 | X = x).\end{aligned}$$

Nous déterminons d'abord la densité $f_{(Y, Y-C)}$ du couple $(Y, Y - C)$ sachant $X = x$ en fonction des densités marginales f_Y et f_C de Y et C respectivement par la formule du changement de variables et nous obtenons

$$f_{(Y, Y-C)}((u, v)|x) = f_Y(u|x)f_C(u - v|x).$$

De ce fait,

$$\begin{aligned}\bar{H}^1(t_n|x) &= \int_{t_n}^{\infty} \left(\int_{-\infty}^0 f_Y(u|x)f_C(u - v|x)dv \right) du \\ &= \int_{t_n}^{\infty} f_Y(u|x) \left(\int_u^{\infty} f_C(w|x)dw \right) du \\ &= \int_{t_n}^{\infty} f_Y(u|x)\bar{G}(u|x)du.\end{aligned}$$

Par conséquent,

$$\begin{aligned}\bar{H}^1(t_n|x) &= \int_{t_n}^{\infty} (1 - G(u|x)) dF(u|x) \\ &= \frac{1}{\gamma_1(x)} \int_{t_n}^{\infty} u^{-1/\gamma(x)-1} L(u|x)du - \int_{t_n}^{\infty} u^{-1/\gamma(x)} L'_1(u|x)du\end{aligned}$$

posons $\varepsilon_1(t_n|x) = t_n L'_1(t_n|x)/L_1(t_n|x)$, alors $\varepsilon_1(t_n|x) \rightarrow 0$ quand $t_n \rightarrow \infty$ pour tout $x \in \mathbb{R}^p$ (voir Bingham et al., 1987, page 15) et sous la classe de Hall $\varepsilon_1(t_n|x) = O(\Delta(t_n|x))$ et $\Delta(t_n|x) = O\left(b\left(\frac{1}{\bar{H}(t_n|x)}|x\right)\right)$.

Ce qui nous permet d'écrire

$$\begin{aligned}\bar{H}^1(t_n|x) &= \frac{1}{\gamma_1(x)} \int_{t_n}^{\infty} \frac{\bar{H}(u|x)}{u} du - \int_{t_n}^{\infty} \frac{\bar{H}(u|x)}{u} \varepsilon_1(u|x) du \\ &= \frac{1}{\gamma_1(x)} \gamma(x) \bar{H}(t_n|x) (1 + O(\Delta(t_n|x))) \\ &= p_x \bar{H}(t_n|x) (1 + O(\Delta(t_n|x)))\end{aligned}$$

donc

$$\frac{\bar{H}^1(t_n|x)}{\bar{H}(t_n|x)} - p_x = O(\Delta(t_n|x)).$$

Nous obtenons enfin,

$$\begin{aligned}
\sqrt{nh^p \bar{H}(t_n|x)} \left(\frac{\bar{H}^1(t_n|x)}{\bar{H}(t_n|x)} - p_x \right) &= O\left(\sqrt{nh^p \bar{H}(t_n|x)} \Delta(t_n|x)\right) \\
&= O\left(\sqrt{nh^p \bar{H}(t_n|x)} O\left(b\left(\frac{1}{\bar{H}(t_n|x)}|x\right)\right)\right) \\
&= O\left(O\left(\sqrt{nh^p \bar{H}(t_n|x)} b\left(\frac{1}{\bar{H}(t_n|x)}|x\right)\right)\right) \\
&= O(O(o(1))) \\
&= O(o(1)) \\
&= o(1)
\end{aligned}$$

ce qui conclut la preuve. \square

Preuve du Théorème 3.2.1

Premièrement, nous décomposons $\sqrt{nh^p \bar{H}(t_n|x)}(\hat{\gamma}_{t_n}^{(c,H)}(x) - \gamma_1(x))$ comme suit :

$$\begin{aligned}
\sqrt{nh^p \bar{H}(t_n|x)} \left(\hat{\gamma}_{t_n}^{(c,H)}(x) - \gamma_1(x) \right) &= \frac{1}{\hat{p}_{t_n}(x)} \sqrt{nh^p \bar{H}(t_n|x)} \left(\hat{\gamma}_{t_n}^H(x) - \gamma(x) \right) \\
&\quad - \frac{\gamma_1(x)}{\hat{p}_{t_n}(x)} \sqrt{nh^p \bar{H}(t_n|x)} (\hat{p}_{t_n}(x) - p_x) \\
&= \frac{1}{p_x} \sqrt{nh^p \bar{H}(t_n|x)} \left(\hat{\gamma}_{t_n}^H(x) - \gamma(x) \right) \\
&\quad - \frac{\gamma_1(x)}{p_x} \sqrt{nh^p \bar{H}(t_n|x)} (\hat{p}_{t_n}(x) - p_x) + o_{\mathbb{P}}(1) \\
&:= \frac{1}{p_x} P_{n,x} - \frac{\gamma_1(x)}{p_x} R_{n,x} + o_{\mathbb{P}}(1). \tag{3.12}
\end{aligned}$$

Alors, quelques calculs donnent

$$\begin{aligned}
P_{n,x} &= \sqrt{\frac{nh^p}{g(x)\varphi(t_n,x)}} \left(\frac{1}{n} \sum_{i=1}^n K_h(x - X_i) \log\left(\frac{Z_i}{t_n}\right) \mathbf{1}_{\{Z_i > t_n\}} - \int_{t_n}^{\infty} \frac{\varphi(z,x)}{z} dz \right) \\
&\quad - \gamma(x) \sqrt{\frac{nh^p}{g(x)\varphi(t_n,x)}} (\hat{\varphi}_n(t_n,x) - \varphi(t_n,x)) + o_{\mathbb{P}}(1)
\end{aligned}$$

$$\begin{aligned}
R_{n,x} &= \sqrt{\frac{nh^p}{g(x)\varphi(t_n,x)}} \left(\hat{\psi}_n(t_n,x) - \psi(t_n,x) \right) \\
&\quad - p_x \sqrt{\frac{nh^p}{g(x)\varphi(t_n,x)}} (\hat{\varphi}_n(t_n,x) - \varphi(t_n,x)) + o_{\mathbb{P}}(1).
\end{aligned}$$

Par conséquent, la normalité asymptotique de $\hat{\gamma}_{t_n}^{(c,H)}(x)$ sera prouvée si nous pouvons

établir la normalité asymptotique du vecteur aléatoire

$$\mathbb{X}_n(x) = \sqrt{\frac{nh^p}{g(x)\varphi(t_n, x)}} \begin{pmatrix} \hat{\varphi}_n(t_n, x) - \varphi(t_n, x) \\ \hat{\psi}_n(t_n, x) - \psi(t_n, x) \\ \frac{1}{n} \sum_{i=1}^n K_h(x - X_i) \log\left(\frac{Z_i}{t_n}\right) \mathbb{1}_{\{Z_i > t_n\}} - \int_{t_n}^{\infty} \frac{\varphi(z, x)}{z} dz \end{pmatrix},$$

après quoi le résultat est venu par une application directe de la delta méthode. La normalité asymptotique de $\mathbb{X}_n(x)$ est prouvée en utilisant la formule de Cramér-Wold.

Soit $\ell = (\ell_1, \ell_2, \ell_3)^\top \in \mathbb{R}^3$, $\ell \neq 0$. Alors

$$\ell^\top \mathbb{X}_n(x) := \sqrt{\frac{nh^p}{g(x)\varphi(t_n, x)}} \frac{1}{n} \sum_{i=1}^n T_{i,n}^\dagger,$$

où

$$\begin{aligned} T_{i,n}^\dagger &= l_1 K_h(x - X_i) \mathbb{1}_{\{Z_i > t_n\}} + l_2 K_h(x - X_i) \mathbb{1}_{\{Z_i > t_n, \delta_i = 1\}} \\ &\quad + l_3 K_h(x - X_i) \log\left(\frac{Z_i}{t_n}\right) \mathbb{1}_{\{Z_i > t_n\}} - \left(l_1 \varphi(t_n, x) + l_2 \psi(t_n, x) + l_3 \int_{t_n}^{\infty} \frac{\varphi(z, x)}{z} dz \right). \end{aligned}$$

Le Lemme 2 de Goegebeur et al. (2014b, [59]) donne :

$$\begin{aligned} \mathbb{E} \left[K_h(x - X_i) \log(Z_i/t_n) \mathbb{1}_{\{Z_i > t_n\}} \right] &= \gamma(x) \varphi(t_n, x) (1 + O(h \log t_n)), \\ \mathbb{E} \left[K_h^2(x - X_i) \log(Z_i/t_n) \mathbb{1}_{\{Z_i > t_n\}} \right] &= \gamma(x) \varphi(t_n, x) h^{-p} \|K\|_2^2 (1 + O(h \log t_n)) \end{aligned}$$

et

$$\mathbb{E} \left[K_h^2(x - X_i) \log^2(Z_i/t_n) \mathbb{1}_{\{Z_i > t_n\}} \right] = 2\gamma^2(x) \varphi(t_n, x) h^{-p} \|K\|_2^2 (1 + O(h \log t_n)).$$

Avec cela, les calculs simples impliquent

$$\begin{aligned} \text{var}(\ell^\top \mathbb{X}_n(x)) &= \frac{h^p}{g(x)\varphi(t_n, x)} \text{var}(T_{i,n}^\dagger) \\ &= \ell_1^2 \frac{\|K\|_2^2}{g(x)} + \ell_2^2 \frac{p_x \|K\|_2^2}{g(x)} + \ell_3^2 \frac{2\gamma^2(x) \|K\|_2^2}{g(x)} + 2\ell_1 \ell_2 \frac{p_x \|K\|_2^2}{g(x)} \\ &\quad + 2\ell_1 \ell_3 \frac{\gamma(x) \|K\|_2^2}{g(x)} + 2\ell_2 \ell_3 \frac{\gamma(x) p_x \|K\|_2^2}{g(x)} + o(1) \\ &= \ell^\top \Gamma \ell + o(1), \end{aligned}$$

où

$$\Gamma := \begin{pmatrix} \frac{\|K\|_2^2}{g(x)} & \frac{\|K\|_2^2 p_x}{g(x)} & \frac{\|K\|_2^2 \gamma(x)}{g(x)} \\ \frac{\|K\|_2^2 p_x}{g(x)} & \frac{\|K\|_2^2 p_x}{g(x)} & \frac{\|K\|_2^2 p_x \gamma(x)}{g(x)} \\ \frac{\|K\|_2^2 \gamma(x)}{g(x)} & \frac{\|K\|_2^2 p_x \gamma(x)}{g(x)} & \frac{2\|K\|_2^2 \gamma^2(x)}{g(x)} \end{pmatrix}.$$

Or, pour établir la normalité asymptotique de $\ell^\top \mathbb{X}_n(x)$, nous vérifions le critère de Lyapounov pour les tableaux triangulaires de variables aléatoires. Dans le contexte actuel, cela consiste à prouver que $\lim_{n \rightarrow \infty} \sum_{i=1}^n \mathbb{E} |T_{i,n}^\dagger / \{ng(x)h^{-p}\varphi(t_n, x)\}^{1/2}|^3 = 0$ (des calculs sont similaires comme dans la preuve du Lemme 3.6.4 et sont donc omis). Ce qui implique pour tout $\ell = (\ell_1, \ell_2, \ell_3)^\top \in \mathbb{R}^3$, $\ell \neq 0$, $\ell^\top \mathbb{X}_n(x)$ converge en distribution vers $\mathcal{N}(0, \ell^\top \Gamma \ell)$. En utilisant la formule de Cramér-Wold, nous avons, $\mathbb{X}_n(x) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Gamma)$. Finalement, sur la base de la décomposition (3.12), sur les expressions de $P_{n,x}$ et $R_{n,x}$ et sur la normalité asymptotique de $\mathbb{X}_n(x)$, une application directe de la delta méthode complète la preuve du Théorème 3.2.1. \square

Loi de Paréto généralisée conditionnelle en présence de censure

Résumé

*Dans ce chapitre, les données sont censurées aléatoirement à droite conditionnellement à une covariable à design fixe dans le cas des lois à queues lourdes. Nous proposons pour une covariable x donnée un estimateur du couple $(\gamma_1(x), \sigma_1(x))$ basé sur les méthodes de la **fenêtre mobile** et du maximum de vraisemblance sur une seule étape de l'algorithme de **Newton-Raphson**. Puis nous déterminerons leurs propriétés asymptotiques enfin nous illustrerons leur comportement par des simulations.*

Sommaire

4.1 Introduction	121
4.2 Modèle d'étude	121
4.3 Estimation des paramètres	122
4.3.1 Estimateurs de Beirlant et al. (2007) adaptés	123
4.3.2 Estimation par l'approche P.O.T.	124
4.3.3 Estimation du quantile extrême conditionnel	124
4.4 Résultats théoriques des estimateurs	125
4.5 Simulations	126
4.5.1 Données pour les simulations	127
4.5.2 Résultats	128
4.6 Discussion et perspectives	128
4.7 Preuve du Théorème	129

4.1 Introduction

LE « Generalized Extreme Value » est basé sur les maxima de chaque bloc; ce qui consiste à rejeter beaucoup de données. Cette perte de données constitue une perte d'information alors que l'information compte beaucoup en statistique. Pour palier ce genre de problème, une approche serait d'utiliser la méthode P.O.T. (« **Peaks-Over-Threshold** ») qui repose sur les résultats de Balkema et de Haan (1974, [5]) et aussi ceux de Pickands (1975, [78]) selon lesquels les excès absolus $E_i = Z_i - u > 0$ au delà d'un seuil u donné, peuvent être approchés par une loi de Pareto généralisée. C'est dans ce cadre que Beirlant et al. (2010, [10]) ont présenté des estimateurs du couple (γ_1, σ_1) dans le cas où les données sont censurées aléatoirement à droite. Pour le moment, le problème de l'estimation de quantiles extrêmes conditionnels en présence de données censurées aléatoirement à droite par la méthode P.O.T. reste non abordé dans la littérature.

Dans ce chapitre, nous allons nous intéresser au problème de l'estimation de l'indice des valeurs extrêmes conditionnel, du paramètre d'échelle conditionnel et leurs quantiles extrêmes conditionnels en présence des données censurées aléatoirement à droite en utilisant l'approche **P.O.T.** La méthode d'estimation des paramètres est celle du maximum de vraisemblance avec l'algorithme de **Newton-Raphson**.

4.2 Modèle d'étude

Nous travaillons dans l'espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ et dans l'espace métrique (\mathcal{X}, d) . Soient $(x_i)_{1 \leq i \leq n}$, un échantillon déterministe et $(Y_i)_{1 \leq i \leq n}$ et $(C_i)_{1 \leq i \leq n}$ des réalisations respectives des variables aléatoires X, Y et C dans l'espace $(\Omega, \mathcal{A}, \mathbb{P})$. Nous supposons que Y et C ont respectivement pour fonctions de répartition F et G . Soient $(Z_i)_{1 \leq i \leq n} = (Y_i \wedge C_i)_{1 \leq i \leq n}$ et $\delta_i = \mathbb{1}_{\{Y_i \leq C_i\}}$, où $\mathbb{1}_{\{\cdot\}}$ est une fonction indicatrice. Nous supposons aussi Y et C sont indépendantes conditionnellement à X . Ainsi, dire que nous sommes en présence de données censurées aléatoirement à droite revient à dire que l'échantillon $(Y_i)_{1 \leq i \leq n}$ n'est pas observé mais est censuré par $(C_i)_{1 \leq i \leq n}$. Nous observons uniquement $(Z_1, \delta_1), \dots, (Z_n, \delta_n)$ des copies indépendantes de (Z, δ) . Dans la suite nous supposons que la variable Z a pour fonction de répartition H . Dans tout ce chapitre nous considérons que les fonctions de répartition conditionnelles de Y et C pour tout $x \in \mathcal{X}$ sont ainsi définies:

$$F(u|x) = 1 - u^{-1/\gamma_1(x)} L_1(u, x),$$

$$G(u|x) = 1 - u^{-1/\gamma_2(x)} L_2(u, x),$$

où $\gamma_1(\cdot)$ et $\gamma_2(\cdot)$ sont des paramètres fonctionnels positifs de la covariable x fixé, $L_1(\cdot, x)$ et $L_2(\cdot, x)$ sont des fonctions à variations lentes, i.e. pour tout $\lambda > 0$,

$$\lim_{u \rightarrow \infty} \frac{L_i(\lambda u, x)}{L_i(u, x)} = 1, i = 1, 2.$$

Si Y et C sont indépendantes, alors pour tout x , nous avons,

$$\overline{H}(\cdot|x) = \overline{F}(\cdot|x)\overline{G}(\cdot|x).$$

Alors,

$$\overline{H}(u|x) = u^{-1/\gamma_1(x)} L_1(u, x) u^{-1/\gamma_2(x)} L_2(u, x) = u^{-\left(\frac{\gamma_1(x)+\gamma_2(x)}{\gamma_1(x)\gamma_2(x)}\right)} L_1(u, x) L_2(u, x).$$

Nous allons considérer

$$\gamma(x) = \frac{\gamma_1(x)\gamma_2(x)}{\gamma_1(x) + \gamma_2(x)} > 0 \text{ et } L(u, x) = L_1(u, x)L_2(u, x)$$

et

$$\lim_{u \rightarrow \infty} \frac{L(\lambda u, x)}{L(u, x)} = 1, \forall \lambda > 0$$

car le produit de fonctions à variations lentes donne une fonction à variations lentes. Finalement,

$$H(u|x) = 1 - u^{-1/\gamma(x)} L(u, x)$$

donc H est une fonction de répartition conditionnelle appartenant au domaine d'attraction de Fréchet (cf chapitre 1).

4.3 Estimation des paramètres

Nous allons considérer $(Z_1, \delta_1, x_1), \dots, (Z_n, \delta_n, x_n)$, l'échantillon observé. L'objectif de cette partie est, pour un $x \in \mathcal{X}$ fixé, déterminer des estimateurs de $\gamma_1(x)$, de $\sigma_1(x)$ et de leur quantile extrême conditionnel correspondant. Pour cela nous n'allons pas utiliser toutes les données mais simplement les observations dont leurs covariables sont proches de x . Ainsi, l'approche que nous allons utiliser est la méthode de la **fenêtre mobile** (2008, [51]), définie par une boule, $B(x, h_{n,x})$ où x est le centre de la boule et $h_{n,x}$ est le rayon de la boule:

$$B(x, h_{n,x}) = \{t \in \mathcal{X}, d(x, t) \leq h_{n,x}\}$$

comme dans le chapitre 2. Notons $\{Z_i^x, \delta_i^x, i = 1, \dots, m\}$ les observations retenues lors de la sélection. Soient

$$Z_{(1)}^x \leq \dots \leq Z_{(m)}^x$$

et $\delta_{(1)}^x, \dots, \delta_{(m)}^x$ la statistique d'ordre associée à cet échantillon.

4.3.1 Estimateurs de Beirlant et al. (2007) adaptés

Dans ce paragraphe, nous allons adapter les estimateurs de Beirlant et al. (2007, [9]) dans le cas conditionnel à design fixe. Pour cela, soit u un seuil donné. Dans le cas non censuré les estimateurs sont notés, $\hat{\gamma}_{Z^x, u}^{(H)}(x)$ pour l'estimateur de Hill (1975, [67]), $\hat{\gamma}_{Z^x, u}^{(M)}(x)$ pour l'estimateur des Moments dans Dekkers et al. (1989, [40]), l'estimateur UH par $\hat{\gamma}_{Z^x, u}^{(UH)}(x)$ dans Beirlant (1996, [9], [13]) et dans le cas général, $\hat{\gamma}_{Z^x, u}^{(\cdot)}(x)$ ((\cdot) désigne l'estimateur standard comme dans le chapitre 2) de la fonction $\gamma(\cdot)$. Pour estimer l'indice de queue, $\gamma_1(x)$, on divise les estimateurs précédents par la proportion de données non censurées au delà du seuil u , ce qui implique,

$$\hat{\gamma}_{Z^x, u}^{(c, \cdot)}(x) = \frac{\hat{\gamma}_{Z^x, u}^{(\cdot)}(x)}{\frac{1}{N_u} \sum_{i=1}^m \delta_i^x \mathbb{1}_{\{Z_i^x > u\}}}$$

où N_u est le nombre de données qui dépassent le seuil u et le paramètre d'échelle estimé par

$$\hat{\sigma}_{Z^x, u}^{(c, M)}(x) = \frac{\bar{H}_m(u|x)}{\bar{H}_m^1(u|x)} \sqrt{3 \left(M_{Z^x, u}^{(1)} \right)^2 - M_{Z^x, u}^{(2)}}$$

où

$$M_{Z^x, u}^{(1)} = \frac{\sum_{i=1}^m \log \left(\frac{Z_i^x}{u} \right) \mathbb{1}_{\{Z_i^x > u\}}}{m \bar{H}_m(u|x)}$$

$$M_{Z^x, u}^{(2)} = \frac{\sum_{i=1}^m \log \left(\frac{Z_i^x}{u} \right)^2 \mathbb{1}_{\{Z_i^x > u\}}}{m \bar{H}_m(u|x)}$$

et les distributions H^1 et H admettent respectivement comme estimateurs empiriques

$$H_m^1(z|x) = \frac{1}{m} \sum_{i=1}^m \delta_i^x \mathbb{1}_{\{Z_i^x \leq z\}}$$

et

$$H_m(z|x) = \frac{1}{m} \sum_{i=1}^m \mathbb{1}_{\{Z_i^x \leq z\}}.$$

4.3.2 Estimation par l'approche P.O.T.

Avec ces données, nous envisageons d'utiliser la méthode **P.O.T.** pour estimer le couple $(\gamma_1(x), \sigma_1(x))$. Pour cela, il est alors nécessaire d'adapter la fonction de vraisemblance à la censure en utilisant les résultats de Balkema et de Haan (1974, [5]). Ce qui nous permet de proposer un estimateur de l'indice des valeurs extrêmes conditionnel mais aussi celui du paramètre d'échelle conditionnel de la G.P.D., distribution des excès $E_i^x = Z_i^x - u$ sachant $Z_i^x > u$, au-delà d'un seuil $u \rightarrow z_H = \sup \{z : H(z|x) < 1\}$. Dans ce cas, la fonction de vraisemblance est la suivante:

$$L(\gamma_1(x), \sigma_1(x)) = \prod_{i=1}^{N_u} \left[g_{\gamma_1(x), \sigma_1(x)}(E_i^x | x) \right]^{\delta_i^x} \left[1 - G_{\gamma_1(x), \sigma_1(x)}(E_i^x | x) \right]^{1 - \delta_i^x} \quad (4.1)$$

où $1 - G_{\gamma_1(x), \sigma_1(x)}(z|x) = \left(1 + \frac{\gamma_1(x)}{\sigma_1(x)}z\right)^{-\frac{1}{\gamma_1(x)}}$, $g_{\gamma_1(x), \sigma_1(x)}$ est la densité associée. Pour plus de détails le lecteur peut consulter la thèse de **Toulemonde** (2008, [87]) ou Beirlant et al. (2010, [10]). Les estimateurs obtenus par la méthode du maximum de vraisemblance de (4.1) sont notés $(\hat{\gamma}_{Z^x, u}^{(c, os)}(x), \hat{\sigma}_{Z^x, u}^{(c, os)}(x))$.

4.3.3 Estimation du quantile extrême conditionnel

Avec ce résultat nous pouvons déterminer l'estimateur du quantile extrême conditionnel correspondant d'ordre α_m :

$$\hat{q}_u^{(c, os)}(x, \alpha_m) = u + \hat{\sigma}_{Z^x, u}^{(c, os)}(x) \frac{\left(\frac{1 - \hat{F}_m(u|x)}{\alpha_m}\right)^{\hat{\gamma}_{Z^x, u}^{(c, os)}(x)} - 1}{\hat{\gamma}_{Z^x, u}^{(c, os)}(x)}$$

où \hat{F}_m est l'estimateur de Kaplan-Meier défini

$$1 - \hat{F}_m(z|x) = \prod_{i=1}^m \left[1 - \frac{\delta_{(i)}^x \mathbb{1}_{\{Z_{(i)}^x \leq z\}}}{m - i + 1} \right].$$

En utilisant les résultats de de Haan et Ferreira (2006, Théorème 1.2.5, p.21, [35]) nous pouvons estimer le paramètre de dispersion $\sigma_1(x)$, par

$$\hat{\sigma}_{Z^x, u}^{(c, os)} = u \hat{\gamma}_{Z^x, u}^{(c, os)}.$$

Dans ce cas, l'estimateur du quantile extrême conditionnel rejoint celui de Weissman adapté,

$$\hat{q}_u^{(c, os)}(x, \alpha_m) = u \left(\frac{1 - \hat{F}_m(u|x)}{\alpha_m} \right)^{\hat{\gamma}_{Z^x, u}^{(c, os)}(x)}. \quad (4.2)$$

Nous allons maintenant déterminer les propriétés asymptotiques de ces estimateurs.

4.4 Résultats théoriques des estimateurs

Avant d'énoncer les propriétés asymptotiques de ces estimateurs, nous définissons et notons quelques concepts qui nous seront utiles pour la suite comme Beirlant et al. (2010, [10]). Notons $U_F(\cdot|x)$ (respectivement $U_G(\cdot|x)$) la fonction quantile conditionnelle de F (respectivement de G) définie par $U_F(z|x) = F^{\leftarrow}\left(1 - \frac{1}{z}|x\right) = \inf\left\{y : F(y|x) \geq 1 - \frac{1}{z}\right\}$ (respectivement $U_G(z|x) = G^{\leftarrow}\left(1 - \frac{1}{z}|x\right) = \inf\left\{s : G(s|x) \geq 1 - \frac{1}{z}\right\}$). Supposons qu'il existe deux fonctions conditionnelles positives a_F et a_G telles que

$$\lim_{z \rightarrow \infty} \frac{U_F(zu|x) - U_F(z|x)}{a_F(z|x)} = \int_1^u v^{\gamma_1(x)-1} dv = h_{\gamma_1(x)}(u|x) \quad \text{pour tout } u > 0 \quad (4.3)$$

$$\lim_{z \rightarrow \infty} \frac{U_G(zu|x) - U_G(z|x)}{a_G(z|x)} = \int_1^u v^{\gamma_2(x)-1} dv = h_{\gamma_2(x)}(u|x) \quad \text{pour tout } u > 0. \quad (4.4)$$

Ces conditions sont appelées condition du premier ordre conditionnelles. Les conditions du second ordre conditionnelles sont décrites comme suit:

$$\frac{U_F(zu|x) - U_F(z|x)}{a_F(z|x)} - h_{\gamma_1(x)}(u|x) \sim a_{2,F}(z|x)k_F(u|x), z \rightarrow \infty \quad (4.5)$$

$$\frac{U_G(zu|x) - U_G(z|x)}{a_G(z|x)} - h_{\gamma_2(x)}(u|x) \sim a_{2,G}(z|x)k_G(u|x), z \rightarrow \infty \quad (4.6)$$

où $a_{2,F}$, $a_{2,G} \rightarrow 0$ sont des fonctions à variations régulières d'indices respectifs $\rho_1 \leq 0$ et $\rho_2 \leq 0$ et

$$k_F(u|x) = A_F h_{\gamma_1(x)+\rho_1(x)}(u|x) + c_F \int_1^u t^{\gamma_1(x)-1} h_{\rho_1(x)}(t|x) dt$$

$$k_G(u|x) = A_G h_{\gamma_2(x)+\rho_2(x)}(u|x) + c_G \int_1^u t^{\gamma_2(x)-1} h_{\rho_2(x)}(t|x) dt$$

où A_F , A_G , c_F et c_G sont des constantes liées à ces fonctions. Avec ces conditions voici les différents termes de biais:

$$\begin{aligned} B_{1,u}(\gamma_1(x), \sigma_1(x)) &= \frac{1}{\gamma_1(x)\sigma_1(x)} \frac{1}{\overline{H}(z|x)} \int_u^\infty \frac{\overline{H}(z|x)}{1 + \frac{\gamma_1(x)}{\sigma_1(x)}(z-u)} dz \\ &\quad - \frac{1}{\gamma_1(x)\sigma_1(x)} \frac{1}{\overline{H}(z|x)} \int_u^\infty \frac{\overline{H}(z|x)}{\left(1 + \frac{\gamma_1(x)}{\sigma_1(x)}(z-u)\right)^2} dz \\ &\quad - \frac{1}{\sigma_1(x)} \frac{1}{\overline{H}(z|x)} \int_u^\infty \frac{\overline{H}^1(z|x)}{\left(1 + \frac{\gamma_1(x)}{\sigma_1(x)}(z-u)\right)^2} dz \end{aligned}$$

$$\begin{aligned}
B_{2,u}(\gamma_1(x), \sigma_1(x)) &= -\frac{\overline{H}^1(z|x)}{\overline{H}(z|x)} \frac{1}{\sigma_1(x)} \frac{1}{\overline{H}(z|x)} \int_u^\infty \frac{\overline{H}(z|x)}{(1 + \frac{\gamma_1(x)}{\sigma_1(x)}(z-u))^2} dz \\
&\quad + \frac{1}{\sigma_1(x)} \frac{1}{\overline{H}(z|x)} \int_u^\infty \frac{\overline{H}^1(z|x)}{(1 + \frac{\gamma_1(x)}{\sigma_1(x)}(z-u))^2} dz
\end{aligned}$$

avec $H^1(u|x) = \mathbb{P}(Z^x > u, \delta^x = 1 | X = x)$ appelée la sous-distribution conditionnelle des observations non-censurées comme dans le chapitre 2.

Théorème 4.4.1 *Supposons que U_F et U_G sont continues et satisfassent les conditions (4.3) à (4.6) et $u \rightarrow \infty$,*

$$\sqrt{N_u} B_{1,u}(\gamma_1(x), \sigma_1(x)) \xrightarrow{\mathbb{P}} 0 \quad (4.7)$$

$$\sqrt{N_u} B_{2,u}(\gamma_1(x), \sigma_1(x)) \xrightarrow{\mathbb{P}} 0 \quad (4.8)$$

alors, nous avons,

$$\sqrt{N_u} \begin{pmatrix} \widehat{\gamma}_{Z^x,u}^{(c,os)}(x) - \gamma_1(x) \\ \widehat{\sigma}_{Z^x,u}^{(c,os)}(x) - \sigma_1(x) \end{pmatrix} \xrightarrow{d} \mathcal{N} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Delta \right)$$

où

$$\Delta = \begin{pmatrix} \left(\frac{\gamma_1(x)}{\gamma(x)}\right)^3 (1 + \gamma(x))^2 & -\left(\frac{\gamma_1(x)}{\gamma(x)}\right)^2 (1 + \gamma(x)) \sigma_1(x) \\ -\left(\frac{\gamma_1(x)}{\gamma(x)}\right)^2 (1 + \gamma(x)) \sigma_1(x) & 2 \left(\frac{\gamma_1(x)}{\gamma(x)}\right) (1 + \gamma(x)) \sigma_1^2(x) \end{pmatrix}$$

pour tout $x \in \mathcal{X}$.

4.5 Simulations

Dans ce paragraphe, nous menons une étude de simulation complète pour évaluer la performance des estimateurs de vraisemblance proposés de l'indice des valeurs extrêmes conditionnel et de quantiles conditionnels proposé dans l'équation (4.2). Nous étudions à la fois la précision de ces estimateurs et la qualité de l'approximation gaussienne de leurs distributions asymptotiques. Nous identifions la portée de chacun de ces estimateurs de l'application (en fonction de la taille de l'échantillon et le pourcentage de censure).

4.5.1 Données pour les simulations

Nous procédons comme suit pour les simulations. Nous simulons $R = 1000$ répliques des échantillons de tailles n ($n = 1000, 1500, 2000, 2500$) des données indépendantes, (Z_i, δ_i, x_i) , où $Z_i = \min(Y_i, C_i)$ et $x_i \in [0, 1]$. La fonction de répartition conditionnelle de Y_i sachant $X = x_i$ est de Pareto de paramètre

$$\gamma_1(x) = 0.5 \left(0.1 + \sin(\pi x) \times \left(1.1 - 0.5 \exp \left(-64 (x - 0.5)^2 \right) \right) \right)$$

et celle de C_i est aussi de Pareto avec comme paramètre γ_2 qui nous permet de déterminer le pourcentage de censure c (10%, 25%, 40%). Pour chaque échantillon, nous estimons $\gamma_1(\cdot)$ au point $x = 0.5$ ($\gamma_1(0.5) = 0.35$). Nous utilisons aussi ici la méthode de la fenêtre mobile pour le choix des plus proches voisins dans la boule $B(0.5, 0.1)$ comme dans le chapitre 2. Soit $\hat{\gamma}_{Z^x, u}^{(c, os), j}(x)$ l'estimateur de $\gamma_1(x)$ obtenu à la j -ième réplique ($j = 1, \dots, R$) avec $k_x = i$ ($i = 1, \dots, m$). Nous définissons l'estimateur de k_x comme suit:

$$k_x^{opt} := \arg \min_{1 \leq i \leq m} \left\{ \frac{1}{R} \sum_{j=1}^R \left(\hat{\gamma}_{Z^x, u}^{(c, os), j}(x) - \gamma_1(x) \right)^2 \right\}.$$

Ensuite, nous déterminons l'estimation du quantile extrême $q(0.5, 1/1000) = 11.22018$ d'ordre $1 - 1/1000$ de la fonction de répartition conditionnelle de Y sachant $x = 0.5$. Pour chaque choix de paramètres de simulation, nous calculons l'estimation conditionnelle de l'estimateur de l'indice de queue conditionnel et du quantile extrême conditionnel.

4.5.2 Résultats

Après simulations, les résultats numériques sont présentés dans le Tableau 4.1.

n	Gamma			Quantile		
	10%	25%	40%	10%	25%	40%
1000	0.3976 (0.1086) [0.0868]	0.3770 (0.1559) [0.1213]	0.3892 (0.2069) [0.1632]	19.5698 (17.4202) [10.5830]	25.1856 (46.8966) [18.4180]	47.8178 (167.5705) [41.9614]
1500	0.3974 (0.1066) [0.0862]	0.3832 (0.1314) [0.1051]	0.3841 (0.1720) [0.1366]	19.5698 (17.0360) [10.5830]	20.9517 (25.1356) [13.4081]	29.3067 (55.3671) [22.7669]
2000	0.3899 (0.0902) [0.0727]	0.3904 (0.1140) [0.0894]	0.3863 (0.1591) [0.1248]	17.2913 (12.1376) [8.0196]	19.7244 (20.4324) [11.2041]	26.8328 (47.3183) [19.8159]
2500	0.3932 (0.0839) [0.0668]	0.3906 (0.1024) [0.0804]	0.3863 (0.1295) [0.1024]	17.1461 (11.0037) [7.3878]	18.4922 (16.7167) [9.5450]	21.2265 (25.7675) [13.4501]

Tableau 4.1 – Résultats des simulations pour les estimateurs de $\gamma_1(0.5)$ et du quantile $q(.5, 1/1000)$. Pour chaque scénario, la première ligne donne la moyenne avec $R = 1000$, estimations de $\gamma_1(0.5)$ et $q(.5, 1/1000)$. (·): La racine carrée de l'erreur quadratique moyenne (RMSE). [·]: Erreur absolue moyenne (MAE).

4.6 Discussion et perspectives

Dans ce travail, nous avons considéré l'estimation de l'indice de queue conditionnel et quantiles extrêmes conditionnels par la méthode du maximum de vraisemblance et de Pareto généralisée lorsque des covariables fixes sont disponibles et les données sont censurées aléatoirement à droite. On utilise l'algorithme de Newton-Raphson pour optimiser la fonction du maximum de vraisemblance. D'après le Tableau 4.1, les estimations des estimateurs de $\gamma_1(x)$ et de $q(0.5, 1/1000)$ considérés se dégradent au fur et à mesure que le pourcentage de censure augmente de 10% à 40% ou bien la taille de l'échantillon diminue de 2500 à 1000. Ce Tableau révèle que l'approximation gaussienne de la distribution asymptotique de l'estimateur du maximum de vraisemblance est raisonnable.

Dans nos futures recherches il serait intéressant de combiner l'approche P.O.T. et les covariables aléatoires dans le cas où les données sont censurées aléatoirement à droite en utilisant les résultats de Balkema et de Haan (1974, [5]) et de Pickands (1975, [78]). Pour cela, il faudrait utiliser la méthode du noyau pour proposer des estimateurs de l'indice de queue conditionnel et du quantile pour aborder la nature

aléatoire de la covariable. Il serait aussi intéressant de proposer des méthodes pour le choix du rayon de la boule.

4.7 Preuve du Théorème

Preuve du Théorème 4.4.1

• Nous nous referons au Théorème 4.2 plus précisément l'assertion (25) de Lehmann (1991, [71]) qui nous permet d'écrire le système suivant si l'on considère les estimateurs initiaux comme ceux des Moments (voir Stupfler (2013, [85]) pour les propriétés asymptotiques):

$$\left\{ \begin{array}{l} \left(\hat{\gamma}_{Z^x,u}^{(c,os)}(x) - \hat{\gamma}_{Z^x,u}^{(c,M)}(x) \right) L''_{11} \left(\hat{\gamma}_{Z^x,u}^{(c,M)}(x), \hat{\sigma}_{Z^x,u}^{(c,M)}(x) \right) + \\ \left(\hat{\sigma}_{Z^x,u}^{(c,os)}(x) - \hat{\sigma}_{Z^x,u}^{(c,M)}(x) \right) L''_{12} \left(\hat{\gamma}_{Z^x,u}^{(c,M)}(x), \hat{\sigma}_{Z^x,u}^{(c,M)}(x) \right) = -L'_1 \left(\hat{\gamma}_{Z^x,u}^{(c,M)}(x), \hat{\sigma}_{Z^x,u}^{(c,M)}(x) \right) \\ \\ \left(\hat{\gamma}_{Z^x,u}^{(c,os)}(x) - \hat{\gamma}_{Z^x,u}^{(c,M)}(x) \right) L''_{12} \left(\hat{\gamma}_{Z^x,u}^{(c,M)}(x), \hat{\sigma}_{Z^x,u}^{(c,M)}(x) \right) + \\ \left(\hat{\sigma}_{Z^x,u}^{(c,os)}(x) - \hat{\sigma}_{Z^x,u}^{(c,M)}(x) \right) L''_{22} \left(\hat{\gamma}_{Z^x,u}^{(c,M)}(x), \hat{\sigma}_{Z^x,u}^{(c,M)}(x) \right) = -L'_2 \left(\hat{\gamma}_{Z^x,u}^{(c,M)}(x), \hat{\sigma}_{Z^x,u}^{(c,M)}(x) \right) \end{array} \right. \quad (4.9)$$

où L'_i et L''_{ij} , $i = 1, 2, j = 1, 2$ désignent les premières, respectivement les secondes, dérivées évaluées au point $\left(\hat{\gamma}_{Z^x,u}^{(c,M)}(x), \hat{\sigma}_{Z^x,u}^{(c,M)}(x) \right)$ de la fonction log-vraisemblance, $\log L(\gamma_1(x), \sigma_1(x))$.

Ce système d'équations, (4.9), peut être écrit sous forme:

$$\begin{pmatrix} \hat{\gamma}_{Z^x,u}^{(c,os)}(x) \\ \frac{\hat{\sigma}_{Z^x,u}^{(c,os)}(x)}{\sigma_1(x)} \end{pmatrix} = \begin{pmatrix} \hat{\gamma}_{Z^x,u}^{(c,M)}(x) \\ \frac{\hat{\sigma}_{Z^x,u}^{(c,M)}(x)}{\sigma_1(x)} \end{pmatrix} - \begin{pmatrix} L''_{11} & \sigma_1(x)L''_{12} \\ \sigma_1(x)L''_{12} & \sigma_1^2(x)L''_{22} \end{pmatrix}^{-1} \begin{pmatrix} L'_1 \\ \sigma_1(x)L'_2 \end{pmatrix}. \quad (4.10)$$

• Ce système (4.10) peut être écrit en multipliant par $\sqrt{N_u}$ comme:

$$\begin{aligned} & \sqrt{N_u} \begin{pmatrix} \hat{\gamma}_{Z^x,u}^{(c,os)}(x) - \hat{\gamma}_{Z^x,u}^{(c,M)}(x) \\ \frac{\hat{\sigma}_{Z^x,u}^{(c,os)}(x)}{\sigma_1(x)} - \frac{\hat{\sigma}_{Z^x,u}^{(c,M)}(x)}{\sigma_1(x)} \end{pmatrix} \\ &= \frac{1}{\left(\frac{1}{N_u} L''_{11} \right) \left(\frac{\sigma_1^2(x)}{N_u} L''_{22} \right) - \left(\frac{\sigma_1(x)}{N_u} L''_{12} \right)^2} \begin{pmatrix} \frac{\sigma_1^2(x)}{N_u} L''_{22} & -\frac{\sigma_1(x)}{N_u} L''_{12} \\ -\frac{\sigma_1(x)}{N_u} L''_{12} & \frac{1}{N_u} L''_{11} \end{pmatrix} \begin{pmatrix} -\frac{1}{\sqrt{N_u}} L'_1 \\ -\frac{\sigma_1(x)}{\sqrt{N_u}} L'_2 \end{pmatrix} \end{aligned} \quad (4.11)$$

• En utilisant ce système (4.11), nous avons la convergence en probabilité des dérivées secondes de la fonction de log-vraisemblance:

$$-\frac{1}{N_u} L''_{11}(\gamma_1(x), \sigma_1(x)) \xrightarrow{\mathbb{P}} 2 \left(\frac{\gamma(x)}{\gamma_1(x)} \right)^3 \frac{1}{(1 + \gamma(x))(1 + 2\gamma(x))},$$

$$-\frac{\sigma_1(x)^2}{N_u} L''_{22}(\gamma_1(x), \sigma_1(x)) \xrightarrow{\mathbb{P}} \frac{\gamma(x)}{\gamma_1(x)} \frac{1}{1 + 2\gamma(x)}$$

et

$$-\frac{\sigma_1(x)}{N_u} L''_{12}(\gamma_1(x), \sigma_1(x)) \xrightarrow{\mathbb{P}} \left(\frac{\gamma(x)}{\gamma_1(x)} \right)^2 \frac{1}{(1 + \gamma(x))(1 + 2\gamma(x))}$$

où

$$\left\{ \begin{array}{l} L'_1(\gamma_1(x), \sigma_1(x)) := \frac{\partial \log L(\gamma_1(x), \sigma_1(x))}{\partial \gamma_1(x)} \\ = \frac{1}{\gamma_1(x)^2} \sum_{i=1}^{N_u} \log \left(1 + \frac{\gamma_1(x) E_i^x}{\sigma_1(x)} \right) - \frac{1}{\gamma_1(x)} \sum_{i=1}^{N_u} (1 + \delta_i^x) \frac{\gamma_1(x) E_i^x / \sigma_1(x)}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} E_i^x} \\ L'_2(\gamma_1(x), \sigma_1(x)) := \frac{\partial \log L(\gamma_1(x), \sigma_1(x))}{\partial \sigma_1(x)} \\ = -\frac{1}{\sigma_1(x)} \sum_{i=1}^{N_u} \delta_i^x + \frac{1}{\sigma_1(x)} \sum_{i=1}^{N_u} (1 + \delta_i^x) \frac{\gamma_1(x) E_i^x / \sigma_1(x)}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} E_i^x} \\ L''_{11}(\gamma_1(x), \sigma_1(x)) := \frac{\partial^2 \log L(\gamma_1(x), \sigma_1(x))}{\partial^2 \gamma_1(x)} \\ = -\frac{2}{\gamma_1(x)^3} \sum_{i=1}^{N_u} \log \left(1 + \frac{\gamma_1(x) E_i^x}{\sigma_1(x)} \right) \\ + \frac{2}{\gamma_1(x)^3} \sum_{i=1}^{N_u} \frac{\gamma_1(x) E_i^x / \sigma_1(x)}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} E_i^x} \\ + \frac{1}{\gamma_1(x)^2} \sum_{i=1}^{N_u} (1 + \delta_i^x) \frac{(\gamma_1(x) E_i^x / \sigma_1(x))^2}{\left(1 + \frac{\gamma_1(x)}{\sigma_1(x)} E_i^x \right)^2} \\ L''_{22}(\gamma_1(x), \sigma_1(x)) := \frac{\partial^2 \log L(\gamma_1(x), \sigma_1(x))}{\partial^2 \sigma_1(x)} \\ = \frac{1}{\sigma_1(x)^2} \sum_{i=1}^{N_u} \delta_i^x - \frac{2}{\sigma_1(x)^2} \sum_{i=1}^{N_u} (1 + \delta_i^x) \frac{\gamma_1(x) E_i^x / \sigma_1(x)}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} E_i^x} \\ + \frac{1}{\sigma_1(x)^2} \sum_{i=1}^{N_u} (1 + \delta_i^x) \frac{(\gamma_1(x) E_i^x / \sigma_1(x))^2}{\left(1 + \frac{\gamma_1(x)}{\sigma_1(x)} E_i^x \right)^2} \\ L''_{12}(\gamma_1(x), \sigma_1(x)) := \frac{\partial^2 \log L(\gamma_1(x), \sigma_1(x))}{\partial \sigma_1(x) \partial \gamma_1(x)} \\ = -\frac{1}{\gamma_1(x) \sigma_1(x)} \sum_{i=1}^{N_u} (1 + \delta_i^x) \frac{(\gamma_1(x) E_i^x / \sigma_1(x))^2}{\left(1 + \frac{\gamma_1(x)}{\sigma_1(x)} E_i^x \right)^2} \\ + \frac{1}{\gamma_1(x) \sigma_1(x)} \sum_{i=1}^{N_u} \delta_i^x \frac{\gamma_1(x) E_i^x / \sigma_1(x)}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} E_i^x} \end{array} \right.$$

• En se basant sur les dérivées premières de la fonction log-vraisemblance avec le système (4.10) au point $(\gamma_1(x), \sigma_1(x))$ (comme dans la preuve du Théorème 4.2 de Lehmann (1991, [71])) et en tenant compte des conditions (4.7) et (4.8), alors

$$V = \left(\begin{array}{c} -\sqrt{N_u} \left[\frac{1}{N_u} L'_1(\gamma_1, \sigma_1(x)) - B_{1,u}(\gamma_1, \sigma_1(x)) \right] \\ -\sqrt{N_u} \left[\frac{\sigma_1(x)}{N_u} L'_2(\gamma_1, \sigma_1(x)) - B_{2,u}(\gamma_1, \sigma_1(x)) \right] \end{array} \right).$$

Déterminons les termes relatifs à la normalité asymptotique:

$$\begin{aligned} & -\sqrt{N_u} \left[\frac{1}{N_u} L'_1(\gamma_1(x), \sigma_1(x)) - B_{1,u}(\gamma_1(x), \sigma_1(x)) \right] = -\sqrt{\frac{\bar{H}(u|x)}{\bar{H}_m(u|x)}} \sqrt{\frac{m}{\bar{H}(u|x)}} \\ & \left\{ \frac{1}{\gamma_1(x)^2} \left(\frac{1}{m} \sum_{i=1}^m \log \left(1 + \frac{\gamma_1(x)}{\sigma_1(x)} E_i^x \right) \mathbb{1}_{\{Z_i^x > u\}} - \frac{\gamma_1(x)}{\sigma_1(x)} \int_u^\infty \frac{\bar{H}(z|x)}{1 + \frac{\gamma_1(x)}{\sigma_1(x)}(z-u)} dz \right) \right. \\ & - \frac{1}{\gamma_1(x)^2} \left(\frac{1}{m} \sum_{i=1}^m \left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} E_i^x} \right) \mathbb{1}_{\{Z_i^x > u\}} - \frac{\gamma_1(x)}{\sigma_1(x)} \int_u^\infty \frac{\bar{H}(z|x)}{\left(1 + \frac{\gamma_1(x)}{\sigma_1(x)}(z-u) \right)^2} dz \right) \\ & - \frac{1}{\gamma_1(x)} \left(\frac{1}{m} \sum_{i=1}^m \delta_i^x \left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} E_i^x} \right) \mathbb{1}_{\{Z_i^x > u\}} - \frac{\gamma_1(x)}{\sigma_1(x)} \int_u^\infty \frac{\bar{H}(z|x)}{\left(1 + \frac{\gamma_1(x)}{\sigma_1(x)}(z-u) \right)^2} dz \right) \\ & + [\bar{H}_m(u|x) - \bar{H}(u|x)] \left[-\frac{1}{\gamma_1(x)\sigma_1(x)} \frac{1}{\bar{H}(u|x)} \int_u^\infty \frac{\bar{H}(u|x)}{1 + \frac{\gamma_1(x)}{\sigma_1(x)}(z-u)} dz \right] \\ & + [\bar{H}_m(u|x) - \bar{H}(u|x)] \left[\frac{1}{\gamma_1(x)\sigma_1(x)} \frac{1}{\bar{H}(u|x)} \int_u^\infty \frac{\bar{H}(u|x)}{\left(1 + \frac{\gamma_1(x)}{\sigma_1(x)}(z-u) \right)^2} dz \right] \\ & \left. + [\bar{H}_m(u|x) - \bar{H}(u|x)] \left[\frac{1}{\gamma_1(x)\sigma_1(x)} \frac{1}{\bar{H}(u|x)} \int_u^\infty \frac{\bar{H}^1(u|x)}{\left(1 + \frac{\gamma_1(x)}{\sigma_1(x)}(z-u) \right)^2} dz \right] \right\} \end{aligned}$$

et

$$\begin{aligned} & -\sqrt{N_u} \left[\frac{\sigma_1(x)}{N_u} L'_2(\gamma_1(x), \sigma_1(x)) - B_{2,u}(\gamma_1(x), \sigma_1(x)) \right] \\ & = -\sqrt{\frac{\bar{H}(u|x)}{\bar{H}_m(u|x)}} \sqrt{\frac{m}{\bar{H}(u|x)}} \left\{ -(\bar{H}_m^1(u|x) - \bar{H}^1(u|x)) \right. \\ & + \frac{1}{\gamma_1(x)} \left(\frac{1}{m} \sum_{i=1}^m \left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} E_i^x} \right) \mathbb{1}_{\{Z_i^x > u\}} - \frac{\gamma_1(x)}{\sigma_1(x)} \int_u^\infty \frac{\bar{H}(z|x)}{\left(1 + \frac{\gamma_1(x)}{\sigma_1(x)}(z-u) \right)^2} dz \right) \\ & + \left(\frac{1}{m} \sum_{i=1}^m \delta_i^x \left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} E_i^x} \right) \mathbb{1}_{\{Z_i^x > u\}} - \frac{\gamma_1(x)}{\sigma_1(x)} \int_u^\infty \frac{\bar{H}^1(z|x)}{\left(1 + \frac{\gamma_1(x)}{\sigma_1(x)}(z-u) \right)^2} dz \right) \\ & + [\bar{H}_m(u|x) - \bar{H}(u|x)] \left[\frac{\bar{H}^1(u|x)}{\bar{H}(u|x)} - \frac{1}{\sigma_1(x)} \frac{1}{\bar{H}(u|x)} \int_u^\infty \frac{\bar{H}(u|x)}{\left(1 + \frac{\gamma_1(x)}{\sigma_1(x)}(z-u) \right)^2} dz \right] \\ & \left. + [\bar{H}_m(u|x) - \bar{H}(u|x)] \left[-\frac{1}{\sigma_1(x)} \frac{1}{\bar{H}(u|x)} \int_u^\infty \frac{\bar{H}^1(u|x)}{\left(1 + \frac{\gamma_1(x)}{\sigma_1(x)}(z-u) \right)^2} dz \right] \right\} \end{aligned}$$

- La normalité asymptotique provient de cette matrice:

$$W = \sqrt{\frac{m}{\overline{H}(u|x)}} \begin{pmatrix} \overline{H}_m(u|x) - \overline{H}(u|x) \\ \overline{H}_m^1(u|x) - \overline{H}^1(u|x) \\ \frac{1}{m} \sum_{i=1}^m \log \left(1 + \frac{\gamma_1(x)}{\sigma_1(x)} E_i^x \right) \mathbb{1}_{\{Z_i^x > u\}} - \frac{\gamma_1(x)}{\sigma_1(x)} \int_u^\infty \frac{\overline{H}(z|x)}{1 + \frac{\gamma_1(x)}{\sigma_1(x)}(z-u)} dz \\ \frac{1}{m} \sum_{i=1}^m \left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} E_i^x} \right) \mathbb{1}_{\{Z_i^x > u\}} - \frac{\gamma_1(x)}{\sigma_1(x)} \int_u^\infty \frac{\overline{H}(z|x)}{\left(1 + \frac{\gamma_1(x)}{\sigma_1(x)}(z-u) \right)^2} dz \\ \frac{1}{m} \sum_{i=1}^m \delta_i^x \left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} E_i^x} \right) \mathbb{1}_{\{Z_i^x > u\}} - \frac{\gamma_1(x)}{\sigma_1(x)} \int_u^\infty \frac{\overline{H}^1(z|x)}{\left(1 + \frac{\gamma_1(x)}{\sigma_1(x)}(z-u) \right)^2} dz \end{pmatrix}.$$

- Pour les calculs des covariances le lecteur pourra consulter l'annexe.

• Nous appliquons le théorème central multivarié sur les couples $(Z_i^x, \delta_i^x)_{1 \leq i \leq m}$ qui sont indépendants et nous avons la matrice de variance-covariance de W définie par,

$$\Sigma = \begin{pmatrix} 1 & \frac{\gamma(x)}{\gamma_1(x)} & \gamma(x) & \frac{\gamma_1(x)}{\gamma(x)} a & a \\ \frac{\gamma(x)}{\gamma_1(x)} & \frac{\gamma(x)}{\gamma_1(x)} & \frac{\gamma(x)^2}{\gamma_1(x)} & a & a \\ \gamma(x) & \frac{\gamma(x)^2}{\gamma_1(x)} & 2\gamma(x)^2 & b & \frac{\gamma(x)}{\gamma_1(x)} b \\ \frac{\gamma_1(x)}{\gamma(x)} a & a & b & c & \frac{\gamma(x)}{\gamma_1(x)} c \\ a & a & \frac{\gamma(x)}{\gamma_1(x)} b & \frac{\gamma(x)}{\gamma_1(x)} c & \frac{\gamma(x)}{\gamma_1(x)} c \end{pmatrix}$$

où $a = \frac{\gamma(x)}{\gamma_1(x)} \left(1 - \frac{1}{1+\gamma(x)} \right)$, $b = \gamma(x) \left(1 - \frac{1}{(1+\gamma(x))^2} \right)$ et $c = 1 - \frac{2}{1+\gamma(x)} + \frac{1}{1+2\gamma(x)}$.

Nous en déduisons la matrice de variance-covariance de V :

$$\begin{pmatrix} 2 \left(\frac{\gamma(x)}{\gamma_1(x)} \right)^3 \frac{1}{(1+\gamma(x))(1+2\gamma(x))} & \left(\frac{\gamma(x)}{\gamma_1(x)} \right)^2 \frac{1}{(1+\gamma(x))(1+2\gamma(x))} \\ \left(\frac{\gamma(x)}{\gamma_1(x)} \right)^2 \frac{1}{(1+\gamma(x))(1+2\gamma(x))} & \frac{\gamma(x)}{\gamma_1(x)} \frac{1}{1+2\gamma(x)} \end{pmatrix}$$

ce qui implique

$$\sqrt{n_u} \begin{pmatrix} \hat{\gamma}_{Z^x, u}^{(c, os)}(x) - \gamma_1(x) \\ \hat{\sigma}_{Z^x, u}^{(c, os)}(x) - \sigma_1(x) \end{pmatrix} \xrightarrow{d} \mathcal{N} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Delta \right).$$

Conclusion et perspectives

L'OBJECTIF de cette thèse était de proposer de nouveaux estimateurs de l'indice de queue et de quantiles extrêmes conditionnels dans le cas où la variable d'intérêt Y est censurée aléatoirement à droite par une variable aléatoire C qui sont mesurées simultanément avec une covariable X fixe ou aléatoire. Pour cela, nous avons scindé le problème suivant la nature de la covariable (fixe ou aléatoire) mais aussi selon le modèle choisi (GEV, GPD) mais uniquement dans le domaine de Fréchet.

Pour faciliter la lecture du document, nous avons rappelé dans le chapitre 1 quelques concepts fondamentaux sur la statistique des valeurs extrêmes. Dans cette même lancée, nous avons aussi rappelé dans ce chapitre quelques notions fondamentales sur la statistique des données censurées pour mieux appréhender le domaine.

Dans le chapitre 2, nous avons présenté de nouveaux estimateurs de l'indice des valeurs extrêmes et de quantiles extrêmes conditionnels où la covariable est à design fixe en se basant sur la méthode de la fenêtre mobile comme dans Gardes et al. (2010, [54]). Puis, nous avons déterminé leurs propriétés asymptotiques et enfin nous avons appliqué ces théories sur les données simulées et réelles.

Le chapitre 3 est dédié aux estimateurs de l'indice de queue et de quantiles extrêmes conditionnels en présence de covariable à design aléatoire. Pour cela, nous avons utilisé la méthode non-paramétrique comme dans Goegebeur et al. (2014b, [59]) pour proposer un estimateur de l'indice de queue conditionnel basé sur la méthode noyau puis présenter un estimateur à noyau adapté de type Weissman des quantiles. Nous avons déterminé les propriétés asymptotiques de l'estimateur

de l'indice des valeurs extrêmes conditionnel et illustrer les comportements de ces estimateurs via simulations.

Enfin, dans le chapitre 4, nous nous sommes placés dans le cadre de données censurées aléatoirement à droite avec une covariable à design fixe. Dans ce cas, nous avons utilisé le modèle à seuil pour proposer des estimateurs de l'indice de queue et de quantiles extrêmes conditionnels. Pour le faire, nous avons adapté la vraisemblance à la censure avec les résultats de Balkema et de Haan (1974, [5]) et de Pickands (1975, [78]), qui ont approximé la distribution des excès à la loi de Pareto généralisée dans le cas inconditionnel comme dans Beirlant et al. (2010, [10]). La méthode d'optimisation de cette vraisemblance que nous avons préconisée est celle de Newton-Raphson avec « one-step ». Là aussi, nous avons proposé les propriétés de l'estimateur de l'indice de queue et terminé par étudier son comportement par simulations.

Au terme de ces travaux satisfaisants, nous préconiserons à court terme de déterminer les propriétés asymptotiques de l'estimateur de quantiles extrêmes conditionnels dans le cas où la covariable est aléatoire. Dans nos futures recherches nous modéliserons les valeurs extrêmes avec le modèle des excès conditionnel où la covariable est aléatoire. Il serait aussi intéressant à court terme de déterminer des méthodes appropriées pour choisir le rayon de la boule h_n dans le cas où la covariable est à design fixe.

Aussi, à long terme, il serait intéressant d'étendre cette théorie des valeurs extrêmes en présence des données censurées aléatoirement à droite dans tous les domaines (Weibull, Gumbel) et dans les autres types de censure (censuré à gauche, par intervalle). Dans l'avenir, nous nous attarderons aussi dans ces domaines où les covariables sont à design mixte (fixe et aléatoire en même temps). Enfin, nous nous combinerons toutes nos forces pour proposer des estimateurs de l'indice de queue conditionnel et des quantiles extrêmes conditionnels à design aléatoire avec les modèles (GEV, GPD) où le seuil t_n est une suite aléatoire.



Quelques outils

Rappel sur l'algorithme de Newton-Raphson

La méthode de Newton-Raphson permet une résolution numérique des équations du score du type $S(\gamma_1(x), \sigma_1(x)) =: S(\Delta(x)) = \frac{\partial L(\Delta(x))}{\partial \Delta(x)} = L'(\Delta(x)) = 0$. On part tout d'abord d'une valeur initiale arbitraire de $\Delta(x)$, notée $\Delta(x)^0$. On note $\Delta(x)^1 = \Delta(x)^0 + pas$ une valeur candidate pour être solution de $S(\Delta(x)) = 0$, c'est-à-dire $S(\Delta(x)^0 + pas) = 0$. Par une application de la formule de Taylor à l'ordre un de la fonction S , on obtient l'approximation suivante

$$S(\Delta(x)^0 + pas) = S(\Delta(x)^0) + pas S'(\Delta(x)^0).$$

On sait que $S(\Delta(x)^0 + pas) = 0$, on obtient donc la valeur suivante

$$pas = -[S'(\Delta(x)^0)]^{-1} S(\Delta(x)^0).$$

Il vient donc

$$\Delta(x)^1 = \Delta(x)^0 - [L''(\Delta(x)^0)]^{-1} L'(\Delta(x)^0)$$

où $L''(\Delta(x)) = \frac{\partial^2 L(\Delta(x))}{\partial \gamma_1(x) \partial \sigma_1(x)}$.

On itère le processus. La stratégie est la suivante

1. Choix d'un point de départ $\Delta(x)^0$;
2. Calculer $\Delta(x)^{k+1} = \Delta(x)^k - [L''(\Delta(x)^k)]^{-1} L'(\Delta(x)^k)$.

Algorithme 1 Choisir $\Delta(x)^0$

$k = 1$

Répéter

$\Delta(x)^{k+1} = \Delta(x)^k - [L''(\Delta(x)^k)]^{-1} L'(\Delta(x)^k)$

$k = k + 1$

Jusqu'à

$\Delta(x)^{k+1} \approx \Delta(x)^k$ et / ou $L(\Delta(x)^{k+1}) \approx L(\Delta(x)^k)$.

Rappel sur le Théorème de Cramér-Wold

Le résultat suivant qui permet de ramener la convergence en loi en dimension d à de la convergence en loi en dimension 1. Dans la littérature anglo-américaine, ce résultat est appelé sous le nom de « Cramér-Wold device ».

Théorème de Cramér-Wold

La suite $(Y_n)_{n \geq 1}$ de vecteurs aléatoires dans \mathbb{R}^d converge en loi dans \mathbb{R}^d vers la variable aléatoire Y si et seulement si pour toute forme linéaire $\phi : \mathbb{R}^d \mapsto \mathbb{R}$, la suite de variables aléatoires réelles $\phi(Y_n)_{n \geq 1}$ converge en loi dans \mathbb{R} vers $\phi(Y)$.

Condition de Lyapunov et Théorème Central Limite**Théorème Central Limite-Lyapunov**

Soient $(X_i)_{1 \leq i \leq n}$ des variables aléatoires indépendantes centrées (pas forcément de même loi) telles que $\mathbb{E}|X_i|^2 = \sigma_i^2$ et $s_n = \sum_{i=1}^n \sigma_i^2$. S'il existe $\varepsilon > 0$ tel que la **condition de Lyapunov** soit vérifiée:

$$\lim_{n \rightarrow \infty} \frac{1}{s_n^{2+\varepsilon}} \sum_{i=1}^n \mathbb{E}|X_i|^{2+\varepsilon} = 0$$

alors,

$$\frac{1}{\sqrt{s_n}} \sum_{i=1}^n X_i \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1).$$



Calculs de covariances

Dans cette annexe nous allons calculer quelques covariances asymptotiques.

$$\begin{aligned} & \text{Var} \left(\sqrt{\frac{m}{\bar{H}(u|x)}} \left(\bar{H}_m(u|x) - \bar{H}(u|x) \right) \right) \\ &= \frac{m}{\bar{H}(u|x)} \times \frac{1}{m} \text{Var} \left(\mathbf{1}_{\{Z^x > u\}} \right) \\ &= \frac{1}{\bar{H}(u|x)} \left[\mathbb{E}[\mathbf{1}_{\{Z^x > u\}}] - \left(\mathbb{E}[\mathbf{1}_{\{Z^x > u\}}] \right)^2 \right] \\ &= \frac{1}{\bar{H}(u|x)} \left(\bar{H}(u|x) - \left(\bar{H}(u|x) \right)^2 \right) \\ &= 1 - \bar{H}(u|x) \rightarrow 1 \quad \text{si } u \rightarrow \infty. \end{aligned}$$

$$\begin{aligned} & \text{Var} \left(\sqrt{\frac{m}{\bar{H}(u|x)}} \left(\bar{H}_m^1(u|x) - \bar{H}^1(u|x) \right) \right) \\ &= \frac{m}{\bar{H}(u|x)} \times \frac{1}{m} \text{Var} \left(\mathbf{1}_{\{Z^x > u, \delta^x = 1\}} \right) \\ &= \frac{1}{\bar{H}(u|x)} \left[\mathbb{E}[\mathbf{1}_{\{Z^x > u, \delta^x = 1\}}] - \left(\mathbb{E}[\mathbf{1}_{\{Z^x > u, \delta^x = 1\}}] \right)^2 \right] \\ &= \frac{1}{\bar{H}(u|x)} \left(\bar{H}^1(u|x) - \left(\bar{H}^1(u|x) \right)^2 \right) \\ &= \frac{\bar{H}^1(u|x)}{\bar{H}(u|x)} - \bar{H}^1(u|x) \rightarrow p_x = \frac{\gamma(x)}{\gamma_1(x)} \quad \text{si } u \rightarrow \infty \text{ (d'après lemme 3.6.1 du chapitre 3)}. \end{aligned}$$

$$\begin{aligned}
I &= \text{Var} \left(\sqrt{\frac{m}{\bar{H}(u|x)}} \left(\frac{1}{m} \sum_{i=1}^m \log \left(1 + \frac{\gamma_1(x)}{\sigma_1(x)} (Z_i^x - u) \right) \mathbb{1}_{\{Z_i^x > u\}} - \frac{\gamma_1(x)}{\sigma_1(x)} \int_u^\infty \frac{\bar{H}(z|x)}{1 + \frac{\gamma_1(x)}{\sigma_1(x)}(z-u)} dz \right) \right) \\
&= \text{Var} \left(\sqrt{\frac{m}{\bar{H}(u|x)}} \left(\frac{1}{m} \sum_{i=1}^m \log \left(1 + \frac{\gamma_1(x)}{\sigma_1(x)} (Z_i^x - u) \right) \mathbb{1}_{\{Z_i^x > u\}} \right) \right) \\
&= \frac{1}{\bar{H}(u|x)} \text{Var} \left(\log \left(1 + \frac{\gamma_1(x)}{\sigma_1(x)} (Z^x - u) \right) \mathbb{1}_{\{Z^x > u\}} \right)
\end{aligned}$$

posons $Y = \log \left(1 + \frac{\gamma_1(x)}{\sigma_1(x)} (Z^x - u) \right) \mathbb{1}_{\{Z^x > u\}}$ et par des calculs simples, sa loi est donnée comme suit $F_Y(y|x) = \left(1 - e^{-y/\gamma(x)} \right) \bar{H}(u|x)$ donc nous en déduisons $\mathbb{E}[Y] = \gamma(x) \bar{H}(u|x)$ et $\mathbb{E}[Y^2] = 2\gamma(x)^2 \bar{H}(u|x)$ d'où

$$\begin{aligned}
I &= \frac{1}{\bar{H}(u|x)} \left(2\gamma(x)^2 \bar{H}(u|x) - \gamma(x)^2 (\bar{H}(u|x))^2 \right) \\
&= 2\gamma(x)^2 - \gamma(x)^2 \bar{H}(u|x) \longrightarrow 2\gamma(x)^2 \quad \text{quand } u \rightarrow \infty.
\end{aligned}$$

$$\begin{aligned}
J &= \text{Var} \left(\sqrt{\frac{m}{\bar{H}(u|x)}} \left(\frac{1}{m} \sum_{i=1}^m \left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} E_i^x} \right) \mathbb{1}_{\{Z_i^x > u\}} - \frac{\gamma_1(x)}{\sigma_1(x)} \int_u^\infty \frac{\bar{H}(z|x)}{\left(1 + \frac{\gamma_1(x)}{\sigma_1(x)}(z-u) \right)^2} dz \right) \right) \\
&= \text{Var} \left(\sqrt{\frac{m}{\bar{H}(u|x)}} \left(\frac{1}{m} \sum_{i=1}^m \left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} E_i^x} \right) \mathbb{1}_{\{Z_i^x > u\}} \right) \right) \\
&= \frac{1}{\bar{H}(u|x)} \text{Var} \left(\left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} (Z^x - u)} \right) \mathbb{1}_{\{Z^x > u\}} \right)
\end{aligned}$$

posons $T = \left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} (Z^x - u)} \right) \mathbb{1}_{\{Z^x > u\}}$ et nous en déduisons par des calculs que $F_T(t|x) = \left(1 - (1-t)^{1/\gamma(x)} \right) \bar{H}(u|x) \mathbb{1}_{[0,1]}(t)$ et par conséquent

$$\mathbb{E}[T] = \frac{1}{\frac{1}{\gamma(x)} + 1} \bar{H}(u|x) \quad \text{et} \quad \mathbb{E}[T^2] = \frac{2\gamma(x)^2}{(1 + \gamma(x))(1 + 2\gamma(x))} \bar{H}(u|x).$$

D'où

$$\begin{aligned}
J &= \frac{1}{\bar{H}(u|x)} \left(\frac{2\gamma(x)^2}{(1 + \gamma(x))(1 + 2\gamma(x))} \bar{H}(u|x) - \frac{1}{\left(\frac{1}{\gamma(x)} + 1 \right)^2} (\bar{H}(u|x))^2 \right) \\
&= \frac{2\gamma(x)^2}{(1 + \gamma(x))(1 + 2\gamma(x))} - \frac{\bar{H}(u|x)}{\left(\frac{1}{\gamma(x)} + 1 \right)^2} \longrightarrow \frac{2\gamma(x)^2}{(1 + \gamma(x))(1 + 2\gamma(x))} \quad \text{quand } u \rightarrow \infty.
\end{aligned}$$

$$\begin{aligned}
R &= \text{Var} \left(\sqrt{\frac{m}{\overline{H}(u|x)}} \left(\frac{1}{m} \sum_{i=1}^m \delta_i^x \left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} E_i^x} \right) \mathbb{1}_{\{Z_i^x > u\}} - \frac{\gamma_1(x)}{\sigma_1(x)} \int_u^\infty \frac{\overline{H}^1(z|x)}{\left(1 + \frac{\gamma_1(x)}{\sigma_1(x)}(z-u)\right)^2} dz \right) \right) \\
&= \text{Var} \left(\sqrt{\frac{m}{\overline{H}(u|x)}} \left(\frac{1}{m} \sum_{i=1}^m \left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} E_i^x} \right) \mathbb{1}_{\{Z_i^x > u, \delta_i^x = 1\}} \right) \right) \\
&= \frac{1}{\overline{H}(u|x)} \text{Var} \left(\left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} (Z^x - u)} \right) \mathbb{1}_{\{Z^x > u, \delta^x = 1\}} \right)
\end{aligned}$$

posons $M = \left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} (Z^x - u)} \right) \mathbb{1}_{\{Z^x > u, \delta^x = 1\}}$ et nous en déduisons par des calculs que $F_M(t|x) = \left(1 - (1-t)^{1/\gamma(x)} \right) \overline{H}^1(u|x) \mathbb{1}_{[0,1]}(t)$ et par conséquent

$$\mathbb{E}[M] = \frac{1}{\frac{1}{\gamma(x)} + 1} \overline{H}^1(u|x) \quad \text{et} \quad \mathbb{E}[M^2] = \frac{2\gamma(x)^2}{(1 + \gamma(x))(1 + 2\gamma(x))} \overline{H}^1(u|x).$$

D'où

$$\begin{aligned}
R &= \frac{1}{\overline{H}(u|x)} \left(\frac{2\gamma(x)^2}{(1 + \gamma(x))(1 + 2\gamma(x))} \overline{H}^1(u|x) - \frac{1}{\left(\frac{1}{\gamma(x)} + 1\right)^2} (\overline{H}^1(u|x))^2 \right) \\
&= \frac{\overline{H}^1(u|x)}{\overline{H}(u|x)} \frac{2\gamma(x)^2}{(1 + \gamma(x))(1 + 2\gamma(x))} - \frac{\overline{H}^1(u|x)}{\overline{H}(u|x)} \frac{\overline{H}^1(u|x)}{\left(\frac{1}{\gamma(x)} + 1\right)^2} \\
&\rightarrow \frac{\gamma(x)}{\gamma_1(x)} \frac{2\gamma(x)^2}{(1 + \gamma(x))(1 + 2\gamma(x))} \quad \text{si } u \rightarrow \infty \quad (\text{d'après lemme 3.6.1 du chapitre 3}).
\end{aligned}$$

$$\begin{aligned}
&\text{Cov} \left(\sqrt{\frac{m}{\overline{H}(u|x)}} (\overline{H}_m(u|x) - \overline{H}(u|x)), \sqrt{\frac{m}{\overline{H}(u|x)}} (\overline{H}_m^1(u|x) - \overline{H}^1(u|x)) \right) \\
&= \frac{m}{\overline{H}(u|x)} \text{Cov} \left(\overline{H}_m(u|x) - \overline{H}(u|x), \overline{H}_m^1(u|x) - \overline{H}^1(u|x) \right) \\
&= \frac{m}{\overline{H}(u|x)} \text{Cov} \left(\overline{H}_m(u|x), \overline{H}_m^1(u|x) \right) \\
&= \frac{m}{\overline{H}(u|x)} \text{Cov} \left(\frac{1}{m} \sum_{i=1}^m \mathbb{1}_{\{Z_i^x > u\}}, \frac{1}{m} \sum_{i=1}^m \mathbb{1}_{\{Z_i^x > u, \delta_i^x = 1\}} \right) \\
&= \frac{1}{\overline{H}(u|x)} \text{Cov} \left(\mathbb{1}_{\{Z^x > u\}}, \mathbb{1}_{\{Z^x > u, \delta^x = 1\}} \right) \\
&= \frac{1}{\overline{H}(u|x)} \left[\mathbb{E}[\mathbb{1}_{\{Z^x > u\}} \mathbb{1}_{\{Z^x > u, \delta^x = 1\}}] - \mathbb{E}[\mathbb{1}_{\{Z^x > u\}}] \mathbb{E}[\mathbb{1}_{\{Z^x > u, \delta^x = 1\}}] \right] \\
&= \frac{1}{\overline{H}(u|x)} \left[\overline{H}^1(u|x) - \overline{H}(u|x) \overline{H}^1(u|x) \right] \\
&= \frac{\overline{H}^1(u|x)}{\overline{H}(u|x)} - \overline{H}^1(u|x) \rightarrow p_x = \frac{\gamma(x)}{\gamma_1(x)} \quad \text{si } u \rightarrow \infty \quad (\text{d'après lemme 3.6.1 du chapitre 3}).
\end{aligned}$$

$$\begin{aligned}
& Cov\left(\sqrt{\frac{m}{\overline{H}(u|x)}}\left(\overline{H}_m(u|x) - \overline{H}(u|x)\right),\right. \\
& \left.\sqrt{\frac{m}{\overline{H}(u|x)}}\left(\frac{1}{m}\sum_{i=1}^m \log\left(1 + \frac{\gamma_1(x)}{\sigma_1(x)}(Z_i^x - u)\right)\mathbf{1}_{\{Z_i^x > u\}} - \frac{\gamma_1(x)}{\sigma_1(x)}\int_u^\infty \frac{\overline{H}(z|x)}{1 + \frac{\gamma_1(x)}{\sigma_1(x)}(z - u)}dz\right)\right) \\
& = Cov\left(\sqrt{\frac{m}{\overline{H}(u|x)}}\left(\overline{H}_m(u|x) - \overline{H}(u|x)\right),\right. \\
& \left.\sqrt{\frac{m}{\overline{H}(u|x)}}\left(\frac{1}{m}\sum_{i=1}^m \log\left(1 + \frac{\gamma_1(x)}{\sigma_1(x)}(Z_i^x - u)\right)\mathbf{1}_{\{Z_i^x > u\}}\right)\right) \\
& = \frac{1}{\overline{H}(u|x)}Cov\left(\mathbf{1}_{\{Z^x > u\}}, \log\left(1 + \frac{\gamma_1(x)}{\sigma_1(x)}(Z^x - u)\right)\mathbf{1}_{\{Z^x > u\}}\right)
\end{aligned}$$

$$\begin{aligned}
& = \frac{1}{\overline{H}(u|x)}\left(\mathbb{E}\left[\log\left(1 + \frac{\gamma_1(x)}{\sigma_1(x)}(Z^x - u)\right)\mathbf{1}_{\{Z^x > u\}}\right] - \right. \\
& \left.\mathbb{E}[\mathbf{1}_{\{Z^x > u\}}]\mathbb{E}\left[\log\left(1 + \frac{\gamma_1(x)}{\sigma_1(x)}(Z^x - u)\right)\mathbf{1}_{\{Z^x > u\}}\right]\right) \\
& = \frac{1}{\overline{H}(u|x)}\left[\gamma(x)\overline{H}(u|x) - \overline{H}(u|x) \times \gamma(x)\overline{H}(u|x)\right] \\
& = \gamma(x) - \gamma(x)\overline{H}(u|x) \rightarrow \gamma(x) \quad \text{si } u \rightarrow \infty.
\end{aligned}$$

$$\begin{aligned}
& Cov\left(\sqrt{\frac{m}{\overline{H}(u|x)}}\left(\overline{H}_m(u|x) - \overline{H}(u|x)\right),\right. \\
& \left.\sqrt{\frac{m}{\overline{H}(u|x)}}\left(\frac{1}{m}\sum_{i=1}^m\left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)}E_i^x}\right)\mathbb{1}_{\{Z_i^x > u\}} - \frac{\gamma_1(x)}{\sigma_1(x)}\int_u^\infty \frac{\overline{H}(z|x)}{\left(1 + \frac{\gamma_1(x)}{\sigma_1(x)}(z - u)\right)^2}dz\right)\right) \\
&= Cov\left(\sqrt{\frac{m}{\overline{H}(u|x)}}\left(\overline{H}_m(u|x) - \overline{H}(u|x)\right),\right. \\
& \left.\sqrt{\frac{m}{\overline{H}(u|x)}}\left(\frac{1}{m}\sum_{i=1}^m\left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)}E_i^x}\right)\mathbb{1}_{\{Z_i^x > u\}}\right)\right) \\
&= \frac{1}{\overline{H}(u|x)}Cov\left(\mathbb{1}_{\{Z^x > u\}}, \left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)}(Z^x - u)}\right)\mathbb{1}_{\{Z^x > u\}}\right) \\
&= \frac{1}{\overline{H}(u|x)}\left(\mathbb{E}\left[\left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)}(Z^x - u)}\right)\mathbb{1}_{\{Z^x > u\}}\right] - \right. \\
& \left.\mathbb{E}[\mathbb{1}_{\{Z^x > u\}}]\mathbb{E}\left[\left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)}(Z^x - u)}\right)\mathbb{1}_{\{Z^x > u\}}\right]\right) \\
&= \frac{1}{\overline{H}(u|x)}\left[\frac{1}{\frac{1}{\gamma(x)} + 1}\overline{H}(u|x) - \overline{H}(u|x) \times \frac{1}{\frac{1}{\gamma(x)} + 1}\overline{H}(u|x)\right] \\
&= \frac{1}{\frac{1}{\gamma(x)} + 1} - \overline{H}(u|x) \times \frac{1}{\frac{1}{\gamma(x)} + 1} \rightarrow \frac{1}{\frac{1}{\gamma(x)} + 1} \quad \text{si } u \rightarrow \infty.
\end{aligned}$$

$$\begin{aligned}
& Cov\left(\sqrt{\frac{m}{\overline{H}(u|x)}}\left(\overline{H}_m(u|x) - \overline{H}(u|x)\right),\right. \\
& \left.\sqrt{\frac{m}{\overline{H}(u|x)}}\left(\frac{1}{m}\sum_{i=1}^m\delta_i^x\left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)}E_i^x}\right)\mathbb{1}_{\{Z_i^x > u\}} - \frac{\gamma_1(x)}{\sigma_1(x)}\int_u^\infty\frac{\overline{H}^1(z|x)}{\left(1 + \frac{\gamma_1(x)}{\sigma_1(x)}(z-u)\right)^2}dz\right)\right) \\
& = Cov\left(\sqrt{\frac{m}{\overline{H}(u|x)}}\left(\overline{H}_m(u|x) - \overline{H}(u|x)\right),\right. \\
& \left.\sqrt{\frac{m}{\overline{H}(u|x)}}\left(\frac{1}{m}\sum_{i=1}^m\delta_i^x\left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)}E_i^x}\right)\mathbb{1}_{\{Z_i^x > u\}}\right)\right) \\
& = \frac{1}{\overline{H}(u|x)}Cov\left(\mathbb{1}_{\{Z^x > u\}},\left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)}(Z^x - u)}\right)\mathbb{1}_{\{Z^x > u, \delta^x = 1\}}\right) \\
& = \frac{1}{\overline{H}(u|x)}\left(\mathbb{E}\left[\left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)}(Z^x - u)}\right)\mathbb{1}_{\{Z^x > u, \delta^x = 1\}}\right] - \right. \\
& \left.\mathbb{E}[\mathbb{1}_{\{Z^x > u\}}]\mathbb{E}\left[\left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)}(Z^x - u)}\right)\mathbb{1}_{\{Z^x > u, \delta^x = 1\}}\right]\right) \\
& = \frac{1}{\overline{H}(u|x)}\left[\frac{1}{\frac{1}{\gamma(x)} + 1}\overline{H}^1(u|x) - \overline{H}(u|x) \times \frac{1}{\frac{1}{\gamma(x)} + 1}\overline{H}^1(u|x)\right] \\
& = \frac{1}{\frac{1}{\gamma(x)} + 1}\frac{\overline{H}^1(u|x)}{\overline{H}(u|x)} - \overline{H}^1(u|x) \times \frac{1}{\frac{1}{\gamma(x)} + 1} \rightarrow \frac{\gamma(x)}{\gamma_1(x)}\frac{1}{\frac{1}{\gamma(x)} + 1} \quad \text{si } u \rightarrow \infty.
\end{aligned}$$

$$\begin{aligned}
& Cov\left(\sqrt{\frac{m}{\overline{H}(u|x)}}\left(\overline{H}_m^1(u|x) - \overline{H}^1(u|x)\right),\right. \\
& \left.\sqrt{\frac{m}{\overline{H}(u|x)}}\left(\frac{1}{m}\sum_{i=1}^m\log\left(1 + \frac{\gamma_1(x)}{\sigma_1(x)}(Z_i^x - u)\right)\mathbb{1}_{\{Z_i^x > u\}} - \frac{\gamma_1(x)}{\sigma_1(x)}\int_u^\infty\frac{\overline{H}(z|x)}{1 + \frac{\gamma_1(x)}{\sigma_1(x)}(z-u)}dz\right)\right) \\
& = Cov\left(\sqrt{\frac{m}{\overline{H}(u|x)}}\left(\overline{H}_m^1(u|x) - \overline{H}^1(u|x)\right),\right. \\
& \left.\sqrt{\frac{m}{\overline{H}(u|x)}}\left(\frac{1}{m}\sum_{i=1}^m\log\left(1 + \frac{\gamma_1(x)}{\sigma_1(x)}(Z_i^x - u)\right)\mathbb{1}_{\{Z_i^x > u\}}\right)\right) \\
& = \frac{1}{\overline{H}(u|x)}Cov\left(\mathbb{1}_{\{Z^x > u\}},\log\left(1 + \frac{\gamma_1(x)}{\sigma_1(x)}(Z^x - u)\right)\mathbb{1}_{\{Z^x > u\}}\right)
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{\overline{H}(u|x)} \left(\mathbb{E} \left[\log \left(1 + \frac{\gamma_1(x)}{\sigma_1(x)} (Z^x - u) \right) \mathbf{1}_{\{Z^x > u, \delta^x = 1\}} \right] - \right. \\
&\quad \left. \mathbb{E}[\mathbf{1}_{\{Z^x > u, \delta^x = 1\}}] \mathbb{E} \left[\log \left(1 + \frac{\gamma_1(x)}{\sigma_1(x)} (Z^x - u) \right) \mathbf{1}_{\{Z^x > u\}} \right] \right) \\
&= \frac{1}{\overline{H}(u|x)} \left[\gamma(x) \overline{H}^1(u|x) - \overline{H}^1(u|x) \times \gamma(x) \overline{H}(u|x) \right] \\
&= \gamma(x) \frac{\overline{H}^1(u|x)}{\overline{H}(u|x)} - \frac{\overline{H}^1(u|x)}{\overline{H}(u|x)} \times \gamma(x) \overline{H}(u|x) \rightarrow \frac{\gamma(x)^2}{\gamma_1(x)} \quad \text{si } u \rightarrow \infty.
\end{aligned}$$

$$\begin{aligned}
&Cov \left(\sqrt{\frac{m}{\overline{H}(u|x)}} \left(\overline{H}_m^1(u|x) - \overline{H}^1(u|x) \right), \right. \\
&\quad \left. \sqrt{\frac{m}{\overline{H}(u|x)}} \left(\frac{1}{m} \sum_{i=1}^m \left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} E_i^x} \right) \mathbf{1}_{\{Z_i^x > u\}} - \frac{\gamma_1(x)}{\sigma_1(x)} \int_u^\infty \frac{\overline{H}(z|x)}{\left(1 + \frac{\gamma_1(x)}{\sigma_1(x)} (z - u) \right)^2} dz \right) \right) \\
&= Cov \left(\sqrt{\frac{m}{\overline{H}(u|x)}} \left(\overline{H}_m^1(u|x) - \overline{H}^1(u|x) \right), \right. \\
&\quad \left. \sqrt{\frac{m}{\overline{H}(u|x)}} \left(\frac{1}{m} \sum_{i=1}^m \left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} E_i^x} \right) \mathbf{1}_{\{Z_i^x > u\}} \right) \right) \\
&= \frac{1}{\overline{H}(u|x)} Cov \left(\mathbf{1}_{\{Z^x > u, \delta^x = 1\}}, \left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} (Z^x - u)} \right) \mathbf{1}_{\{Z^x > u\}} \right) \\
&= \frac{1}{\overline{H}(u|x)} \left(\mathbb{E} \left[\left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} (Z^x - u)} \right) \mathbf{1}_{\{Z^x > u, \delta^x = 1\}} \right] - \right. \\
&\quad \left. \mathbb{E}[\mathbf{1}_{\{Z^x > u, \delta^x = 1\}}] \mathbb{E} \left[\left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} (Z^x - u)} \right) \mathbf{1}_{\{Z^x > u\}} \right] \right) \\
&= \frac{1}{\overline{H}(u|x)} \left[\frac{1}{\frac{1}{\gamma(x)} + 1} \overline{H}^1(u|x) - \overline{H}^1(u|x) \times \frac{1}{\frac{1}{\gamma(x)} + 1} \overline{H}(u|x) \right] \\
&= \frac{1}{\frac{1}{\gamma(x)} + 1} \frac{\overline{H}^1(u|x)}{\overline{H}(u|x)} - \overline{H}(u|x) \times \frac{1}{\frac{1}{\gamma(x)} + 1} \frac{\overline{H}^1(u|x)}{\overline{H}(u|x)} \rightarrow \frac{\gamma(x)}{\gamma_1(x)} \frac{1}{\frac{1}{\gamma(x)} + 1} \quad \text{si } u \rightarrow \infty.
\end{aligned}$$

$$\begin{aligned}
& Cov\left(\sqrt{\frac{m}{\overline{H}(u|x)}}\left(\overline{H}_m^1(u|x) - \overline{H}^1(u|x)\right)\right), \\
& \sqrt{\frac{m}{\overline{H}(u|x)}}\left(\sqrt{\frac{m}{\overline{H}(u|x)}}\left(\frac{1}{m}\sum_{i=1}^m\delta_i^x\left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)}E_i^x}\right)\mathbb{1}_{\{Z_i^x > u\}}\right.\right. \\
& \left.\left. - \frac{\gamma_1(x)}{\sigma_1(x)}\int_u^\infty\frac{\overline{H}^1(z|x)}{\left(1 + \frac{\gamma_1(x)}{\sigma_1(x)}(z-u)\right)^2}dz\right)\right) \\
& = Cov\left(\sqrt{\frac{m}{\overline{H}(u|x)}}\left(\overline{H}_m^1(u|x) - \overline{H}^1(u|x)\right)\right), \\
& \sqrt{\frac{m}{\overline{H}(u|x)}}\left(\frac{1}{m}\sum_{i=1}^m\delta_i^x\left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)}E_i^x}\right)\mathbb{1}_{\{Z_i^x > u\}}\right) \\
& = \frac{1}{\overline{H}(u|x)}Cov\left(\mathbb{1}_{\{Z^x > u, \delta^x=1\}},\left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)}(Z^x - u)}\right)\mathbb{1}_{\{Z^x > u, \delta^x=1\}}\right) \\
& = \frac{1}{\overline{H}(u|x)}\left(\mathbb{E}\left[\left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)}(Z^x - u)}\right)\mathbb{1}_{\{Z^x > u, \delta^x=1\}}\right] - \right. \\
& \left. \mathbb{E}[\mathbb{1}_{\{Z^x > u, \delta^x=1\}}]\mathbb{E}\left[\left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)}(Z^x - u)}\right)\mathbb{1}_{\{Z^x > u, \delta^x=1\}}\right]\right) \\
& = \frac{1}{\overline{H}(u|x)}\left[\frac{1}{\frac{1}{\gamma(x)} + 1}\overline{H}^1(u|x) - \overline{H}^1(u|x) \times \frac{1}{\frac{1}{\gamma(x)} + 1}\overline{H}^1(u|x)\right] \\
& = \frac{1}{\frac{1}{\gamma(x)} + 1}\frac{\overline{H}^1(u|x)}{\overline{H}(u|x)} - \overline{H}^1(u|x) \times \frac{1}{\frac{1}{\gamma(x)} + 1}\frac{\overline{H}^1(u|x)}{\overline{H}(u|x)} \rightarrow \frac{\gamma(x)}{\gamma_1(x)}\frac{1}{\frac{1}{\gamma(x)} + 1} \quad \text{si } u \rightarrow \infty.
\end{aligned}$$

$$\begin{aligned}
& Cov\left(\sqrt{\frac{m}{\overline{H}(u|x)}}\left(\frac{1}{m}\sum_{i=1}^m\log\left(1 + \frac{\gamma_1(x)}{\sigma_1(x)}E_i^x\right)\mathbb{1}_{\{Z_i^x > u\}} - \frac{\gamma_1(x)}{\sigma_1(x)}\int_u^\infty\frac{\overline{H}(z|x)}{1 + \frac{\gamma_1(x)}{\sigma_1(x)}(z-u)}dz\right)\right), \\
& \sqrt{\frac{m}{\overline{H}(u|x)}}\left(\frac{1}{m}\sum_{i=1}^m\left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)}E_i^x}\right)\mathbb{1}_{\{Z_i^x > u\}} - \frac{\gamma_1(x)}{\sigma_1(x)}\int_u^\infty\frac{\overline{H}(z|x)}{\left(1 + \frac{\gamma_1(x)}{\sigma_1(x)}(z-u)\right)^2}dz\right) \\
& = Cov\left(\sqrt{\frac{m}{\overline{H}(u|x)}}\left(\frac{1}{m}\sum_{i=1}^m\log\left(1 + \frac{\gamma_1(x)}{\sigma_1(x)}E_i^x\right)\mathbb{1}_{\{Z_i^x > u\}}\right)\right), \\
& \sqrt{\frac{m}{\overline{H}(u|x)}}\left(\frac{1}{m}\sum_{i=1}^m\left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)}E_i^x}\right)\mathbb{1}_{\{Z_i^x > u\}}\right)
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{\overline{H}(u|x)} \text{Cov} \left(\log \left(1 + \frac{\gamma_1(x)}{\sigma_1(x)} (Z^x - u) \right) \mathbb{1}_{\{Z^x > u\}}, \left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} (Z^x - u)} \right) \mathbb{1}_{\{Z^x > u\}} \right) \\
&= \frac{1}{\overline{H}(u|x)} \left(\mathbb{E} \left[\log \left(1 + \frac{\gamma_1(x)}{\sigma_1(x)} (Z^x - u) \right) \mathbb{1}_{\{Z^x > u\}} \left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} (Z^x - u)} \right) \mathbb{1}_{\{Z^x > u\}} \right] \right. \\
&\quad \left. \mathbb{E} \left[\log \left(1 + \frac{\gamma_1(x)}{\sigma_1(x)} (Z^x - u) \right) \mathbb{1}_{\{Z^x > u\}} \right] \mathbb{E} \left[\left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} (Z^x - u)} \right) \mathbb{1}_{\{Z^x > u\}} \right] \right) \\
&= \frac{1}{\overline{H}(u|x)} \left[\gamma(x) \left(1 - \frac{1}{(1 + \gamma(x))^2} \right) \overline{H}(u|x) - \gamma(x) \overline{H}(u|x) \times \frac{1}{1 + \frac{1}{\gamma(x)}} \overline{H}(u|x) \right] \\
&= \gamma(x) \left(1 - \frac{1}{(1 + \gamma(x))^2} \right) - \frac{\gamma(x)}{1 + \frac{1}{\gamma(x)}} \overline{H}(u|x) \rightarrow \gamma(x) \left(1 - \frac{1}{(1 + \gamma(x))^2} \right) \text{ quand } u \rightarrow \infty.
\end{aligned}$$

$$\begin{aligned}
&\text{Cov} \left(\sqrt{\frac{m}{\overline{H}(u|x)}} \left(\frac{1}{m} \sum_{i=1}^m \log \left(1 + \frac{\gamma_1(x)}{\sigma_1(x)} E_i^x \right) \mathbb{1}_{\{Z_i^x > u\}} - \frac{\gamma_1(x)}{\sigma_1(x)} \int_u^\infty \frac{\overline{H}(z|x)}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} (z - u)} dz \right), \right. \\
&\quad \left. \sqrt{\frac{m}{\overline{H}(u|x)}} \left(\frac{1}{m} \sum_{i=1}^m \delta_i^x \left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} E_i^x} \right) \mathbb{1}_{\{Z_i^x > u\}} - \frac{\gamma_1(x)}{\sigma_1(x)} \int_u^\infty \frac{\overline{H}(z|x)}{\left(1 + \frac{\gamma_1(x)}{\sigma_1(x)} (z - u) \right)^2} dz \right) \right) \\
&= \text{Cov} \left(\sqrt{\frac{m}{\overline{H}(u|x)}} \left(\frac{1}{m} \sum_{i=1}^m \log \left(1 + \frac{\gamma_1(x)}{\sigma_1(x)} E_i^x \right) \mathbb{1}_{\{Z_i^x > u\}} \right), \right. \\
&\quad \left. \sqrt{\frac{m}{\overline{H}(u|x)}} \left(\frac{1}{m} \sum_{i=1}^m \delta_i^x \left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} E_i^x} \right) \mathbb{1}_{\{Z_i^x > u\}} \right) \right) \\
&= \frac{1}{\overline{H}(u|x)} \text{Cov} \left(\log \left(1 + \frac{\gamma_1(x)}{\sigma_1(x)} (Z^x - u) \right) \mathbb{1}_{\{Z^x > u\}}, \left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} (Z^x - u)} \right) \mathbb{1}_{\{Z^x > u, \delta^x = 1\}} \right) \\
&= \frac{1}{\overline{H}(u|x)} \left(\mathbb{E} \left[\log \left(1 + \frac{\gamma_1(x)}{\sigma_1(x)} (Z^x - u) \right) \mathbb{1}_{\{Z^x > u\}} \left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} (Z^x - u)} \right) \mathbb{1}_{\{Z^x > u, \delta^x = 1\}} \right] \right. \\
&\quad \left. \mathbb{E} \left[\log \left(1 + \frac{\gamma_1(x)}{\sigma_1(x)} (Z^x - u) \right) \mathbb{1}_{\{Z^x > u\}} \right] \mathbb{E} \left[\left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} (Z^x - u)} \right) \mathbb{1}_{\{Z^x > u, \delta^x = 1\}} \right] \right) \\
&= \frac{1}{\overline{H}(u|x)} \left[\gamma(x) \left(1 - \frac{1}{(1 + \gamma(x))^2} \right) \overline{H}^1(u|x) - \gamma(x) \overline{H}(u|x) \times \frac{1}{1 + \frac{1}{\gamma(x)}} \overline{H}^1(u|x) \right] \\
&= \gamma(x) \left(1 - \frac{1}{(1 + \gamma(x))^2} \right) \frac{\overline{H}^1(u|x)}{\overline{H}(u|x)} - \frac{\gamma(x)}{1 + \frac{1}{\gamma(x)}} \overline{H}(u|x) \frac{\overline{H}^1(u|x)}{\overline{H}(u|x)} \\
&\rightarrow \gamma(x) \left(1 - \frac{1}{(1 + \gamma(x))^2} \right) \frac{\gamma(x)}{\gamma_1(x)} \text{ quand } u \rightarrow \infty \text{ (d'après lemme 3.6.1 du chapitre 3)}.
\end{aligned}$$

$$\begin{aligned}
& \text{Cov} \left(\sqrt{\frac{m}{\overline{H}(u|x)}} \left(\frac{1}{m} \sum_{i=1}^m \left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} E_i^x} \right) \mathbb{1}_{\{Z_i^x > u\}} - \frac{\gamma_1(x)}{\sigma_1(x)} \int_u^\infty \frac{\overline{H}(z|x)}{\left(1 + \frac{\gamma_1(x)}{\sigma_1(x)}(z-u)\right)^2} dz \right), \right. \\
& \left. \sqrt{\frac{m}{\overline{H}(u|x)}} \left(\frac{1}{m} \sum_{i=1}^m \delta_i^x \left(\mathbb{1} - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} E_i^x} \right) \mathbb{1}_{\{Z_i^x > u\}} - \frac{\gamma_1(x)}{\sigma_1(x)} \int_u^\infty \frac{\overline{H}^1(z|x)}{\left(1 + \frac{\gamma_1(x)}{\sigma_1(x)}(z-u)\right)^2} dz \right) \right) \\
&= \text{Cov} \left(\sqrt{\frac{m}{\overline{H}(u|x)}} \left(\frac{1}{m} \sum_{i=1}^m \left(1 - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} E_i^x} \right) \mathbb{1}_{\{Z_i^x > u\}} \right), \right. \\
& \left. \sqrt{\frac{m}{\overline{H}(u|x)}} \left(\frac{1}{m} \sum_{i=1}^m \delta_i^x \left(\mathbb{1} - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)} E_i^x} \right) \mathbb{1}_{\{Z_i^x > u\}} \right) \right) \\
&= \frac{1}{\overline{H}(u|x)} \text{Cov} \left(\left(\mathbb{1} - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)}(Z^x - u)} \right) \mathbb{1}_{\{Z^x > u\}}, \left(\mathbb{1} - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)}(Z^x - u)} \right) \mathbb{1}_{\{Z^x > u, \delta^x = 1\}} \right) \\
&= \frac{1}{\overline{H}(u|x)} \left(\mathbb{E} \left[\left(\mathbb{1} - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)}(Z^x - u)} \right) \mathbb{1}_{\{Z^x > u\}} \left(\mathbb{1} - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)}(Z^x - u)} \right) \mathbb{1}_{\{Z^x > u, \delta^x = 1\}} \right] \right. \\
& \left. - \mathbb{E} \left[\left(\mathbb{1} - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)}(Z^x - u)} \right) \mathbb{1}_{\{Z^x > u\}} \right] \mathbb{E} \left[\left(\mathbb{1} - \frac{1}{1 + \frac{\gamma_1(x)}{\sigma_1(x)}(Z^x - u)} \right) \mathbb{1}_{\{Z^x > u, \delta^x = 1\}} \right] \right) \\
&= \frac{1}{\overline{H}(u|x)} \left(\frac{2\gamma(x)^3}{\gamma_1(x)(1 + \gamma(x))(1 + 2\gamma(x))} \overline{H}(u|x) - \frac{\gamma(x)}{1 + \gamma(x)} \overline{H}(u|x) \times \frac{\gamma(x)}{1 + \gamma(x)} \overline{H}^1(u|x) \right) \\
&= \frac{2\gamma(x)^3}{\gamma_1(x)(1 + \gamma(x))(1 + 2\gamma(x))} - \frac{\gamma(x)}{1 + \gamma(x)} \times \frac{\gamma(x)}{1 + \gamma(x)} \overline{H}^1(u|x) \\
&\rightarrow \frac{2\gamma(x)^3}{\gamma_1(x)(1 + \gamma(x))(1 + 2\gamma(x))} \quad \text{si } u \rightarrow \infty.
\end{aligned}$$

Bibliographie

- [1] Akritas, M. G. Nearest neighbor estimation of a bivariate distribution under random censoring. *Ann. Statist.*, 22:1299–1327, 1994. (Cité en page 29.)
- [2] Al-Abbasi, J. N., Fahmi, K. J. Estimating maximum magnitude earthquakes in iraq using extreme value statistics. *Geophysical journal of the Royal Astronomical Society*, 82:535–548, 1985. (Cité en page xviii.)
- [3] Aldama, A. A., Ramirez A. I. Dam design flood estimation based on bivariate extreme-value distributions. in the extremes of the extremes : extraordinary floods. *Wallingford, IAHS Press. 1*, pages 257–262, 2002. (Cité en page xviii.)
- [4] Alvarado, E., Sandberg, D. V., Pickford, S. G. Modeling large forest fires as extreme events. *Northwest Science*, 1:66–75, 1998. (Cité en page xviii.)
- [5] Balkema, A., de Haan, L. Residual life at great age. *Ann. Probab.*, 2:792–804, 1974. (Cité en pages xxi, 121, 124, 128 et 135.)
- [6] Beirlant, J., Dierckx, G., Guillou, A. Estimation of the extreme-value index and generalized quantile plots. *Bernoulli*, 6:949–970, 2005. (Cité en pages 19 et 45.)
- [7] Beirlant, J., Goegebeur, Y. Regression with response distributions of pareto-type. *Computational Statistics and Data Analysis*, 42:595–619, 2003. (Cité en page 22.)
- [8] Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J. *Statistics of extremes, theory and applications*. Wiley, New York, 2004. (Cité en pages xix, 22, 81 et 82.)
- [9] Beirlant, J., Guillou, A. , Dierckx, G., Fils-Villetard, A. . Estimation of the extreme value index and extreme quantiles under random censoring. *Springer Science+Business Media, LLC 2007*, 10:151–174, 2007. (Cité en pages xx, 73, 74, 82, 86, 114 et 123.)

- [10] Beirlant, J., Guillou, A., Toulemonde, G. Peaks-over-threshold modeling under random censoring. *Communication in Statistics- Theory and Methods*, 39:1158–1179, 2010. (Cité en pages 121, 124, 125 et 135.)
- [11] Beirlant, J., Matthys, G. Extreme quantile estimation for heavy-tailed distributions. *Rapport technique, Department of Mathematics, K. U. Leuven*, 2001. (Cité en page 23.)
- [12] Beirlant, J., Matthys, G. Estimating the extreme value index and high quantiles with exponential regression models. *Statistica Sinica*, 13:853–880, 2003. (Cité en pages 23 et 36.)
- [13] Beirlant, J., Vjnckier, P., Teugels, J.L. Tail index estimation, pareto quantile plots, and regression diagnostics. *J. Amer Statist. Assoc.*, 91:1659–1667, 1996. (Cité en pages xix, 14, 22, 40 et 123.)
- [14] Bengtsson, A., Nilsson, C. Extreme value modelling of storm damage in swedish forests. *Nat. Hazards Earth Syst. Sci.*, 7:515–521, 2007. (Cité en page xviii.)
- [15] Beran, R. Nonparametric regression with randomly censored data. *Technical report. Univ. California, Berkeley*, 1981. (Cité en pages xx, 3, 27, 28 et 29.)
- [16] Berline, A., Gannoun, A., Matzner-Løber, E. Asymptotic normality of convergent estimates of conditional quantiles. *Statistics*, 35:139–169, 2001. (Cité en page 88.)
- [17] Bingham, N., Goldie, C., Teugels, J. *Regular variation*. Cambridge University Press, 1987. (Cité en page 12.)
- [18] Bouchaud, J.-P., Potters, M. *Theory of financial risk and derivative pricing : from statistical physics to risk management*. Cambridge University Press, Cambridge, 2nd edition, 2003. (Cité en page xviii.)
- [19] Brahimi, B., Meraghni, D., Necir, A. On the asymptotic normality of hill's estimator of the tail index under random censoring. *Preprint: arXiv-1302.1666*, 2013. (Cité en pages 44, 82 et 107.)
- [20] Chow, Y.S., Teicher, H. *Probability Theory. Independence, Interchangeability, Martingales, 3rd ed.* New York, 1997. (Cité en pages 70 et 71.)
- [21] Coles, S. *An introduction to statistical modeling of extreme values*. Springer Series in Statistics, Springer-Verlag, London, 2001. (Cité en page 12.)
- [22] Csörgő, S., Deheuvels, P., Mason, D.M. Kernel estimates of the tail index of a distribution. *Ann. Statist.*, 13:1050–1077, 1985. (Cité en page 13.)
- [23] Csörgő, S., Mason, D. Central limit theorems for sums of extreme values. *Mathematical Proceedings of the Cambridge Philosophical Society*, 98:547–558, 1985. (Cité en page 13.)

- [24] Dabrowska, D. M. Nonparametric regression with censored survival time data. *Scandina journal of statistics*, 14:181–197, 1987. (Cité en pages 29 et 33.)
- [25] Dabrowska, D. M. Kaplan-meier estimate on the plane. *Ann. Statist.*, 16:1475–1489, 1988. (Cité en pages 29 et 33.)
- [26] Dabrowska, D. M. Uniform consistency of the kernel kaplan-meier estimate. *Ann. Statist.*, 17:1157–1167, 1989. (Cité en pages 29 et 33.)
- [27] Dabrowska, D. M. Variable bandwidth conditional kaplan-meier estimate. *Scandina journal of statistics*, 19:351–361, 1992a. (Cité en pages 29, 33 et 87.)
- [28] Dabrowska, D. M. Nonparametric quantile regression with censored data. *Sankhya: The Indian journal of statistics*, 54:252–259, 1992b. (Cité en pages 29 et 33.)
- [29] Daniels, H.E. The statistical theory of the strength of bundles and threads. *Proc. Royal Soc.*, pages 405–435, 1945. (Cité en page xviii.)
- [30] Daouia, A., Gardes, L., Girard, S., Lekina, A. Kernel estimators of extreme level curves. *Test*, 20:311–333, 2011. (Cité en pages xix, 82, 85, 88, 89 et 108.)
- [31] Davis, R., Resnick, S. Tail estimates motivated by extreme value theory. *Annals of Statistics*, 12:1467–1487, 1984. (Cité en page 13.)
- [32] Davis, R.A., Resnick, S.I. Basic properties and prediction of max-arma processes. *Adv. in Appl. Probab.*, 21:781–803, 1989. (Cité en page 13.)
- [33] Davison, A. C. , Smith, R. L. Models for exceedances over high thresholds. *Journal of the Royal Statistical Society*, 52:393–442, 1990. (Cité en pages 22 et 36.)
- [34] Davison, A., Ramesh, N. I. Local likelihood smoothing of sample extremes. *Journal of the Royal Statistical Society*, 62:191–208, 2000. (Cité en pages 22 et 36.)
- [35] de Haan, L., Ferreira, A. *Extreme value theory: An introduction*. New York, Springer, 2006. (Cité en pages 11, 14, 15, 16, 108 et 124.)
- [36] de Haan, L., Peng, L. Comparison of tail index estimators. *Statistica Neerlandica*, 52 (1):60–70, 1998. (Cité en page 48.)
- [37] de Haan, L., Resnick, S.I. . A simple asymptotic estimate for the index of a stable distribution. *Journal of the Royal Stat. Soc. B*, 42:83–87, 1980. (Cité en page 13.)
- [38] de Haan, L., Resnick, S.I. On asymptotic normality of the hill estimator. *Stochastic Models*, 4:849–867, 1998. (Cité en page 13.)
- [39] Dekkers, A.L.M., de Haan, L. On the estimation of the extreme value index and large quantile estimation. *The Annals of Statistics*, 17:1795–1832, 1989. (Cité en page 17.)

- [40] Dekkers, A.L.M., Einmahl, J.H.J., de Haan, L. A moment estimator for the index of an extreme value index. *Annals of statistics*, 17:1833–1855, 1989. (Cité en pages [xix](#), [13](#), [14](#), [18](#), [40](#), [45](#), [82](#) et [123](#).)
- [41] Delafosse, E., Guillou, A. Almost sure convergence of a tail index estimator in the presence of censoring. *comptes rendus mathématiques. Académie des Sciences. Paris*, 335:375–380, 2002. (Cité en page [82](#).)
- [42] Dreesbeke, J.J., Saporta, G. *Approches non paramétriques en régression*. Editions Technip, 2011. (Cité en pages [28](#), [30](#), [32](#) et [33](#).)
- [43] Eichner, J. F, Kantelhardt, J. W., Bunde, A., Havlin, S. Extreme value statistics in records with long-term persistence. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 73:016130, 2006. (Cité en page [xviii](#).)
- [44] Einmahl, J. H. J., Fils-Villetard, A., Guillou, A. Statistics of extremes under random censoring. *Bernoulli*, 14:207–227, 2008. (Cité en pages [xx](#), [36](#), [42](#), [43](#), [44](#), [45](#), [60](#), [61](#), [69](#) et [82](#).)
- [45] Einmahl, J. H. J., Magnus, J. R. Records in athletics through extreme-value theory. *Journal of the American Statistical Association*, 103:1382–1391, 2008. (Cité en page [xviii](#).)
- [46] Embrechts, P., Kluppelberg, C., Mikosch, T. *Modelling Extremal Events for Insurance and Finance*. New York, 1997. (Cité en pages [6](#), [9](#), [12](#) et [81](#).)
- [47] Ferraty, F., Vieu, P. *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Series in Statistics, 2006. (Cité en page [xix](#).)
- [48] Ferrez, J., Davison, A.C., Rebetez, M. Extreme temperature analysis under forest cover compared to an open field. *Agricultural and Forest Meteorology*, 151:992–1001, 2011. (Cité en page [81](#).)
- [49] Fisher, R., Tippett, L. Limiting forms the frequency distribution of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society*, 24:180–190, 1928. (Cité en page [xviii](#).)
- [50] Gannoun, A. Estimation non paramétrique de la médiane conditionnelle, médiane anogramme et méthode du noyau. *Publications de l'institut de statistique de l'université de Paris 26*, pages 11–22, 1990. (Cité en page [88](#).)
- [51] Gardes, L., Girard, S. A moving window approach for nonparametric estimation of the conditional tail index. *Journal of Multivariate Analysis*, 99:2368–2388, 2008. (Cité en pages [xix](#), [23](#), [36](#), [38](#), [39](#), [44](#), [57](#), [72](#), [73](#), [74](#), [81](#) et [122](#).)
- [52] Gardes, L., Girard, S. Conditional extremes from heavy-tailed distributions: an application to the estimation of extreme rainfall return levels. *Extremes*, 13:177–204, 2010. (Cité en page [81](#).)

- [53] Gardes, L., Girard, S. Functional kernel estimators of large conditional quantiles. *Electronic Journal of Statistics*, 6:1715–1744, 2012. (Cité en page 81.)
- [54] Gardes, L., Girard, S., Lekina, A. Functional nonparametric estimation of conditional extreme quantiles. *Journal of Multivariate Analysis*, 101:419–433, 2010. (Cité en pages xix, 23, 81 et 134.)
- [55] Gardes, L., Stupfler, G. Estimation of the conditional tail index using a smoothed local hill estimator. *Extremes*, 17:45–75, 2014. (Cité en page 82.)
- [56] Gill, R.D. *Censoring and stochastic integrals*. Amsterdam: Mathematische Centrum, 1980. (Cité en page 28.)
- [57] Gnedenko, B. Sur la distribution limite du terme maximum d’une série aléatoire. *Annals of Mathematics*, 44:423–453, 1943. (Cité en page xviii.)
- [58] Goegebeur, Y., Guillou, A., Osmann, M. A local moment type estimator for the extreme value index in regression with random covariates. *The Canadian Journal of Statistics*, 42:487–507, 2014a. (Cité en page 82.)
- [59] Goegebeur, Y., Guillou, A., Schorgen, A. Nonparametric regression estimation of conditional tails: the random covariate case. *Statistics*, 48:732–755, 2014b. (Cité en pages 82, 84, 85, 88, 91, 111, 117 et 134.)
- [60] Gomes, M.I., Neves, M.M. Estimation of the extreme value index for randomly censored data. *Biometrical Letters*, 48:1–22, 2011. (Cité en page 82.)
- [61] Gomes, M.I., Oliveira, O. Censoring estimators of a positive tail index. *Statistics & Probability Letter*, 65:147–159, 2003. (Cité en page 82.)
- [62] Gonzalez-Manteiga, W., Cadarso-Suarez, C. Asymptotic properties estimation of a generalized kaplan-meier estimator with some applications. *Nonparametric Statistics*, 4:65–78, 1994. (Cité en pages 27 et 29.)
- [63] Grubel, R., Wolf, P. *Estimation of the tail index of a distribution*. Delft University of Technology, Faculty of Technical Mathematics and Informatics, 1994. (Cité en page 13.)
- [64] Hall, P. On some simple estimates of an exponent of regular variation. *Stat. Scoc., B*, 44:37–42, 1982. (Cité en page 13.)
- [65] Hall, P., Tajvidi, N. Nonparametric analysis of temporal trend when fitting parametric models to extreme-value data. *Statistical Science*, 15:153–167, 2000. (Cité en page 22.)
- [66] Hall, P., Welsh, A.H. Adaptive estimates of parameters of regular variation. *Ann. Statist.*, 13:331–341, 1985. (Cité en page 13.)
- [67] Hill, B. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3:1163–1174, 1975. (Cité en pages xix, 13, 39 et 123.)

- [68] Jenkinson, A. F. The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *The Quarterly Journal of the Royal Meteorological Society*, 81:158–171, 1955. (Cité en page [xix](#).)
- [69] Kaplan, E.L., Meier, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481, 1958. (Cité en pages [xix](#), [3](#), [27](#) et [41](#).)
- [70] Lavenda, B. H., Cipollone, E. Extreme value statistics and thermodynamics of earthquakes : large earthquakes. *Annali di Geofisica*, 43:469–484, 2001. (Cité en page [xviii](#).)
- [71] Lehmann, E. L. *Theory of Point Estimation*. Pacific Grove, CA Wadsworth & Brooks/Cole Advanced Books & Software, 1991. (Cité en pages [129](#) et [130](#).)
- [72] Lekina, A. Estimation non-paramétrique des quantiles extrêmes conditionnels. *Thèse de doctorat, Université de Grenoble*, 2010. (Cité en pages [xix](#), [23](#), [24](#), [36](#), [40](#) et [46](#).)
- [73] Lo, S.-H., Singh, K. The product-limit estimator and the bootstrap: some asymptotic representations. *Probab.Th. Rel.Fields*, 71:455–465, 1986. (Cité en page [29](#).)
- [74] Mason, D.M. Laws of large numbers for sums of extreme values. *Ann. Probab.*, 10:754–764, 1982. (Cité en page [13](#).)
- [75] McKeague, I. W., Utikal, K.J. Inference for a nonlinear counting process regression model. *Ann. Statist.*, 18:1172–1187, 1990. (Cité en page [29](#).)
- [76] Ndao, P., Diop, A., Dupuy, J.-F. Nonparametric estimation of the conditional tail index and extreme quantiles under random censoring. *Computational Statistics & Data*, 79:63–79, 2014. (Cité en page [82](#).)
- [77] Novak, S.Y. *Extreme value methods with applications to finance*. CRC Press, Boca Raton, FL, 2012. (Cité en page [81](#).)
- [78] Pickands, J. Statistical inference using extreme order statistics. *Annals of statistics*, 3:119–131, 1975. (Cité en pages [xix](#), [13](#), [14](#), [17](#), [121](#), [128](#) et [135](#).)
- [79] Pisarenko, V.F., Sornette, D. Characterization of the frequency of extreme earthquake events by the generalized pareto distribution. *Pure and Applied Geophysics*, 160:2343–2364, 2003. (Cité en page [81](#).)
- [80] Reiss, R. D., Thomas, M. S. *Statistical analysis of extreme values. From insurance, finance, hydrology and other fields*. Birkhauser Verlag, Basel, 2007. (Cité en pages [xx](#) et [81](#).)
- [81] Resnick, S.I. *Extreme Values, Regular Variation and Point Processes*. Springer-Verlag, 1987. (Cité en page [11](#).)

- [82] Samanta, M. Nonparametric estimation of conditional quantiles. *Statistics & Probability Letters*, 7:407–412, 1989. (Cité en page 88.)
- [83] Smith, R. L. Extreme value analysis of environmental time series : An application to trend detection in ground-level ozone. *Statistical Science*, 5:367–377, 1989. (Cité en pages 22 et 36.)
- [84] Stone, C.J. Consistent nonparametric regression (with discussion). *The Annals of Statistics*, 5:595–645, 1977. (Cité en pages 29 et 87.)
- [85] Stupfler, G. A moment estimator for the conditional extreme-value index. *Electronic Journal of Statistics*, 7:2298–2343, 2013. (Cité en pages 76, 81 et 129.)
- [86] Stute, W. Conditional empirical processes. *The Annals of Statistics*, 14:638–647, 1986. (Cité en page 88.)
- [87] Toulemonde, G. Estimation et tests en théorie des valeurs extrêmes. *Thèse de doctorat, Université Paris VI - Pierre et Marie Curie*, 2008. (Cité en page 124.)
- [88] Turnbull, B.W. Nonparametric estimation of a survivorship function with doubly censored data. *J. Amer. Statist. Assoc.*, 69:169–173, 1974. (Cité en page 26.)
- [89] van der Vaart, A. W. *Asymptotic Statistics*. Cambridge University Press, New York, 1998. (Cité en pages 112 et 114.)
- [90] Van Keilecom, I., Veraverbeke, N. Estimation and bootstrap with censored data in fixed design nonparametric regression. *Ann. Inst. Statist. Math.*, 49:467–491, 1997. (Cité en page 29.)
- [91] Van Keilegom, I., Veraverbeke, N. Bootstrapping quantiles in a fixed design regression model with censored data. *Journal of Statistical Planning and Inference*, 69:115–131, 1998. (Cité en page 29.)
- [92] Van Keilegom, I., Veraverbeke, N. Hazard rate estimation in nonparametric regression with censored data. *Ann. Inst. Statist. Math.*, 53:730–745, 2001. (Cité en page 29.)
- [93] Venables, W.N., Ripley, B.D. *Modern Applied Statistics with S*. Springer, New York, 2002. (Cité en page 61.)
- [94] Von Mises, R. La distribution de la plus grande de n valeurs. *Amer. Math. Soc.*, 2:271–294, 1954. (Cité en page xviii.)
- [95] Weissman, I. Estimation of parameters and large quantiles based on the k-largest observations. *Journal of the American Statistical Association*, 73:812–815, 1978. (Cité en pages xxi, 76, 82 et 87.)
- [96] Worms, J., Worms, R. New estimators of the extreme value index under random right censoring, for heavy-tailed distributions. *Extremes*, 17:337–358, 2014. (Cité en page 82.)

- [97] Yao, Q. Conditional predictive regions for stochastic processes. technical report, university of kent at canterbury. 1999. (Cité en page 88.)

Résumé:

CETTE thèse entre dans le cadre de la modélisation statistique des valeurs extrêmes conditionnelles. Notre objectif est de proposer des estimateurs de l'indice des valeurs extrêmes et des quantiles extrêmes conditionnels en présence de données censurées. Pour cela, nous avons supposé que la variable d'intérêt Y , est censurée aléatoirement à droite par la variable aléatoire C , qui sont mesurées simultanément avec une covariable X . Nous avons considéré que Y et C sont des variables indépendantes conditionnellement à X avec les lois conditionnelles appartenant à « queues lourdes ». Deux modèles GEV: « Generalized Extreme Value » et GPD: « Generalized Pareto Distribution » sont modélisés.

Avec le modèle GEV nous avons traité deux parties selon la nature de la covariable à design fixe ou aléatoire et la méthode de sélection de variables: la fenêtre mobile ou la méthode d'estimation à noyau. Nous avons proposé dans chaque partie des estimateurs de l'indice de queue conditionnel et des quantiles extrêmes conditionnels générés à partir de l'estimateur de Weissman adapté et de l'estimateur de Kaplan-Meier simple ou généralisé. Puis nous avons déterminé les propriétés asymptotiques de nos estimateurs à design fixe. À design aléatoire, nous avons aussi déterminé les propriétés asymptotiques de l'estimateur de l'indice des valeurs extrêmes. Enfin nous avons illustré les comportements de ces estimateurs par des simulations en vérifiant l'effet de la censure avec trois pourcentages de censure (10%, 25%, 40%), l'effet de l'estimateur standard (hill, moment, uh, etc.) sur les estimations mais aussi la consistance et la normalité associées.

Quant au modèle GPD, nous avons présenté des estimateurs de l'indice de queue et de quantiles extrêmes conditionnels à design fixe avec la méthode du maximum de vraisemblance adaptée à la censure. Nous avons aussi proposé les propriétés asymptotiques de l'estimateur de l'indice de queue obtenues à partir de cette méthode et terminer par vérifier leurs comportements via simulation.

Mots clés: Théorie des valeurs extrêmes, queues lourdes, censure aléatoire à droite, indice de queue conditionnel, estimateur de Kaplan-Meier, quantiles extrêmes conditionnels, fenêtre mobile.

Titre: Modélisation de valeurs extrêmes conditionnelles en présence de censure.

Spécialité: Probabilités-Statistique.

Lieux: Thèse préparée à l'Université Gaston Berger de Saint-Louis et à l'Institut National des Sciences Appliquées de Rennes.

Abstract:

THIS thesis is part of the statistical modeling of extreme values conditional. Our goal is to propose the estimators of extreme value index and the conditional extreme quantiles in the presence of censored data. For this, we assumed that the variable of interest Y is randomly censored on the right by the random variable C , which are measured simultaneously with a covariate X . We considered that Y and C are conditionally independent samples with X as conditionals owned « heavy tails ». Two models GEV (Generalized Extreme Value) and GPD (Generalized Pareto Distribution) are modeled.

With the GEV model we treated two parts depending on the nature of the covariate with fixed or random design as a method of selection of variables: the moving window or the kernel estimation method. We have proposed in each of the estimators of the conditional tail index and conditional extreme quantiles generated from the Weissman estimator adapted and Kaplan-Meier estimator or conditional Kaplan-Meier estimator. Then we determined the asymptotic properties of our estimators on the fixed design. At random design, we also determined the asymptotic properties of the estimator of the extreme value index. Finally we illustrate their behavior by simulations for checking the effect of censorship with three percentages of censorship (10%, 25%, 40%), the effect of the standard estimator (Hill, moment, uh, etc.) on estimates but also the normality and the associated consistency.

As for the GPD model we presented estimators of the extreme value index and the conditional extreme quantile in the fixed design with the maximum likelihood method suitable censorship. We took this opportunity to propose the asymptotic properties of the estimator of the extreme value index obtained from this method and end with their check behavior via simulation.

Keywords: Extreme value theory, tail heavy, random censorship, index tail, Kaplan-Meier estimator, conditional extreme quantiles, mobile window method.

Title: Modeling of the extreme value conditional with censored data.

Speciality: Probability-Statistic

Thesis prepared at the Gaston Berger University of Saint-Louis and the National Institute of Applied Sciences of Rennes.
