HAL open science

Estimation non-paramétrique des quantiles extrêmes conditionnels

Alexandre Lekina

To cite this version:

Alexandre Lekina. Estimation non-paramétrique des quantiles extrêmes conditionnels. Mathématiques [math]. Université Joseph-Fourier - Grenoble I; Université de Grenoble, 2010. Français. NNT : . tel-00529476v2

HAL Id: tel-00529476

https://theses.hal.science/tel-00529476v2

Submitted on 26 Feb 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité MATHÉMATIQUES APPLIQUÉES

Arrêté ministériel : 7 août 2006

Présentée et soutenue publiquement par

LEKINA Alexandre

le 13 octobre 2010

ESTIMATION NON-PARAMÉTRIQUE DES QUANTILES EXTRÊMES CONDITIONNELS

Thèse dirigée par GIRARD Stéphane et codirigée par GARDES Laurent

JURY

Civilité/Nom/Prénom	Fonction	Lieu de la fonction	Rôle
Mme PRIEUR Clémentine	Professeur	Université Grenoble 1	Président
Mme GUILLOU Armelle	Professeur	Université Strasbourg	Rapporteur
M. FERRATY Frédéric	Maître de Conférences	Université Toulouse 3	Rapporteur
M. BACRO Jean-Noël	Professeur	Université Montpellier 2	Examinateur
M. GIRARD Stéphane	Chargé de Recherche	INRIA Rhône-Alpes	Examinateur
M. GARDES Laurent	Maître de Conférences	Université Grenoble 2	Examinateur

Thèse préparée au sein du Laboratoire JEAN KUNTZMANN dans l'ECOLE DOCTORALE MATHÉMATIQUES, SCIENCES ET TECHNOLOGIE DE L'INFORMATION, INFORMATIQUE

À MOLO MBESSE *Christine épouse* LEKINA MANGA *Alexandre,*
Au peuple EKANG *et plus particulièrement aux* ETON BËTI.

Remerciements

LE moment est venu pour moi d'exprimer ma plus grande gratitude envers tous celles et ceux qui m'ont aidé et encouragé dans l'accomplissement de cette tâche dont le prix était l'obtention du grade de Docteur de l'Université de Grenoble.

Ainsi qu'il est d'usage, je tiens tout d'abord à exprimer toute ma reconnaissance à Stéphane GIRARD, mon directeur de thèse, pour sa très grande disponibilité, son écoute, ses conseils éclairés et ses encouragements. Il a toujours été présent pour m'inculquer sa grande rigueur. Inestimable a été pour moi le privilège de l'avoir pour guide. Stéphane, merci aussi pour toutes les blagues et la simplicité !

Je tiens également à remercier chaleureusement Laurent GARDES, mon co-directeur de thèse, pour ses hautes qualités humaines et statistiques. Je le remercie pour sa disponibilité et la confiance dont il a fait preuve. Dans les périodes difficiles, il m'a toujours offert du temps et de précieux conseils pour m'aider à avancer. Merci notamment pour l'exigence.

Stéphane et Laurent, merci de m'avoir laissé une grande liberté dans mon travail, merci pour vos relectures. Je garderai toujours avec moi le souvenir de nos discussions, de vos conseils et surtout de votre gentillesse. J'espère que nos relations et nos collaborations continueront longtemps après cette thèse.

Je remercie Armelle GUILLOU et Frédéric FERRATY d'avoir accepté la tâche de rapporteur et pour leur relecture très attentive malgré un emploi du temps chargé. Je remercie aussi Clémentine PRIEUR et Jean-Noël BACRO de m'avoir fait l'honneur de participer gentiment au jury de ma thèse respectivement comme présidente et examinateur. Je leur suis sincèrement reconnaissant de s'être rendus disponibles pour cette soutenance. Merci pour toutes vos critiques constructives sur les problèmes d'estimation non-paramétrique et des valeurs extrêmes. Vos remarques m'ouvrent de nouveaux horizons et de nouvelles perspectives.

Merci à ceux qui ont participé à la réalisation de ce mémoire et en particulier à Abdelaati DAOUIA pour son astuce très ingénieuse lors d'une preuve.

Aussi, je tiens à remercier Florence FORBES de m'avoir accueilli au sein de l'équipe projet Mistis de l'INRIA de Grenoble, Rhône-Alpes. Je dis merci à l'EPST INRIA pour les 37 mois de bourse Cordi sur Subvention qu'il m'a octroyé. Couplée au matériel pédagogique, celle-ci m'a permis de mener mes travaux de recherche doctorale dans de très bonnes conditions.

Je profite aussi de l'occasion pour remercier tous les membres de l'équipe, anciens et nouveaux, que j'ai côtoyés. Une attention particulière à :

- Caroline BERNARD-MICHEL pour avoir agrémenter mon séjour dans la région,

surtout à ses bons petits plats et son caractère taquin ;

- Lamiae AZIZI pour son esprit de partage et ses gâteries dont raffole Senan quelque soit l'espace et le temps ;
- Mathieu FAUVEL et Vassil KHALIDOV pour leur disponibilité et leurs friandises ;
- Senan DOYLE pour son aide précieuse sous Fédora. Par ailleurs, il a su éclairer certains points d'ombres que j'avais sur l'histoire de l'Irlande, son pays ;
- Juliette BLANCHET, Julie CARREAU et Eugen URSU pour leur présence et leurs conseils sur l'après thèse.

Je remercie mes amis, mes proches et ma famille (au sens large) qui m'ont soutenu durant cette aventure, et spécialement à celles ou ceux, elles ou ils se reconnaîtront, qui m'ont aidé lors des préparatifs. Remerciements bien merités à la petite sénégalaise Seynabou NDEYE NDIAYE, au kamite intransigeant « de nom » Gaëtan NDO, au geek des hauts plateaux Rodrigue CHAKODE et au seigneur de la forêt équatoriale Brice EKOBO AKOA.

Enfin, je ne saurais terminer cette partie sans exprimer ma gratitude à mes parents, mes grands parents, mes frères et mes sœurs qui m'ont toujours soutenu, encouragé et stimulé pendant mes études. Je n'aurais jamais pu arriver ici, d'autant plus que je vis seul depuis 2001, sans l'équilibre, la chaleur, le soutien et le bonheur dans lequel j'ai vécu. Merci !

Si une personne juge qu'elle a été oubliée alors qu'elle m'adresse un courriel argumenté et je lui payerais un resto afin de m'excuser :-).

Ici, s'achève ou débute, tout dépendant du lecteur, mon apport scientifique ./...

Table des matières

Introduction générale **1**

1 Quelques éléments de la théorie des valeurs extrêmes **9**
1.1 Introduction 10
1.2 Lois des valeurs extrêmes 10
1.3 Caractérisation des domaines d'attractions 13
1.3.1 Résultats sur les fonctions à variations régulières 14
1.3.2 Domaine d'attraction de Fréchet 15
1.3.3 Domaine d'attraction de Weibull 16
1.3.4 Domaine d'attraction de Gumbel 16
1.4 Estimation des quantiles extrêmes 16
1.4.1 Quelques résultats sur les statistiques d'ordres 18
1.4.2 Approche des quantiles extrêmes par la loi des valeurs extrêmes . . 19
1.4.3 Approche des quantiles extrêmes par la méthode des excès 20
1.4.4 Approche des quantiles extrêmes par l'approche semi-paramétrique 22
1.5 Estimation du paramètre de la loi des valeurs extrêmes 23
1.5.1 Estimateur de Hill 25
1.5.2 Estimateur de Pickands 26
1.5.3 Estimateur de Zipf 28

2 Un nouvel estimateur des quantiles extrêmes non conditionnels **31**
2.1 Introduction 31
2.2 Résultats asymptotiques 32
2.3 Comparaison graphique des estimateurs 34
2.4 Démonstrations 39

3 Quantiles conditionnels **45**
3.1 Introduction 45
3.2 Estimation paramétrique des quantiles conditionnels 46

3.3 Estimation non paramétrique des quantiles conditionnels 46
3.3.1 Méthode d'estimation indirecte 46
3.3.2 Méthode d'estimation directe 51

4 Estimation des quantiles extrêmes conditionnels en design fixe 53
4.1 Introduction 54
4.2 Contexte d'étude et définitions des estimateurs 55
4.2.1 Contexte d'étude et méthode d'estimation 55
4.2.2 Estimateurs des quantiles extrêmes conditionnels 58
4.3 Étude théorique des estimateurs 59
4.3.1 Hypothèses 59
4.3.2 Étude du comportement asymptotique des estimateurs 60
4.4 Exemples et discussion 62
4.4.1 Quelques exemples d'estimateurs de $\gamma(x)$ 63
4.4.2 Applications et discussion 64
4.5 Simulations et illustration sur données réelles 67
4.5.1 Simulations 67
4.5.2 Illustration sur données réelles 74
4.6 Démonstrations 77
4.6.1 Résultats préliminaires 77
4.6.2 Preuve des résultats théoriques 79

5 Estimation de courbes de niveaux extrêmes en design aléatoire 87
5.1 Introduction 88
5.2 Cadre de l'étude et définitions des estimateurs 89
5.2.1 Cadre de l'étude 89
5.2.2 Méthode d'estimation et définitions des estimateurs 89
5.3 Étude théorique des estimateurs 91
5.3.1 Hypothèses 91
5.3.2 Étude du comportement asymptotique des estimateurs 93
5.4 Application à l'estimation de l'indice de queue conditionnel 98
5.5 Expériences numériques et illustration sur données réelles 100
5.5.1 Expériences numériques 100
5.5.2 Illustration sur données réelles 113
5.6 Démonstrations 116
5.6.1 Preuve des résultats préliminaires 116
5.6.2 Preuve des lois asymptotiques des estimateurs 118

Conclusions et perspectives 125

A Loi limite d'une combinaison linéaire d'espacements entre les logarithmes des plus grandes statistiques d'ordre. 129

Bibliographie 131

Introduction générale

Généralités

Pour un profane, la statistique est associée à la notion de moyenne ou d'écart-type. En effet, dans de nombreuses applications, notamment dans les sciences sociales ou sciences physiques, les statistiques se résument parfois au calcul de moyennes ou à l'évaluation de la dispersion d'une série de valeurs autour de leur moyenne.

Par définition, les *événements rares* sont des événements ayant une faible probabilité d'apparition. Lorsque le comportement de ces événements est dû au hasard on peut étudier leur loi. Ils sont dits *extrêmes* quand il s'agit de valeurs beaucoup plus grandes ou plus petites que celles observées habituellement.

Les événements extrêmes et catastrophiques (tremblements de terre, inondations, accidents nucléaires, crises monétaires ou financières, krachs boursiers, émergence d'un nouveau phénomène endémique, etc.) dominent l'actualité quotidienne par leur caractère imprévisible. Compte tenu de l'importance des enjeux sociaux et scientifiques, aucun débat sérieux sur le hasard ne saurait être mené sans une réflexion sur les événements rares et extrêmes. Voilà qui justifie probablement ces propos de Cont (2009) :*« la loi des grands nombres et la distribution gaussienne, fondements de l'étude statistique des grandeurs moyennes, échouent à rendre compte des événements rares ou extrêmes. Pour ce faire, des outils statistiques plus adaptés existent... mais ne sont pas toujours utilisés ! »*.

Dès lors, la question que l'on pourrait se poser est de savoir ce que peuvent les statistiques face aux événements extrêmes ? Autrement dit, peut-on réellement *prévoir ou quantifier le risque des événements extrêmes ?* (dixit l'auteur).

Problématique et contribution de la thèse

Exemples illustratifs

La problématique de ce sujet de thèse peut être illustrée par les trois exemples suivants.

Exemple 1 (de Haan (1990))**.** *Le premier février 1953, lors d'une forte tempête, la mer passe par-dessus plusieurs digues aux Pays-Bas, les détruit et inonde la région. Il s'agit d'un accident majeur. Un comité est mis en place pour étudier le phénomène et proposer des recommandations sur les hauteurs de digues. Il doit tenir compte des facteurs économiques (coût de construction, coût des inondations, etc.), des facteurs physiques (rôle du vent sur la marée, etc.), et aussi des données enregistrées sur les hauteurs de marées. En fait il est plus judicieux de considérer les surcotes, c'est-à-dire la différence entre la hauteur réelle et la hauteur prévue de la marée, que les hauteurs des marées. En effet, on peut supposer, dans une première approximation, que les surcotes des marées lors des tempêtes sont des réalisations de variables aléatoires de même loi. Si on regarde les surcotes pour des marées de tempêtes séparées par quelques jours d'accalmie, on peut même supposer que les variables aléatoires sont indépendantes. L'étude statistique sur des surcotes a pour but de répondre aux questions suivantes :*

- *Trouver la hauteur de digues dont la probabilité d'être dépassée par la surcote est* $\alpha \in]0,1[$.
- *Étant donné une hauteur des digues, trouver la probabilité qu'elle soit dépassée par la plus haute surcote annuelle.*

La réponse aux questions soulevées dans cet exemple peut fournir des éléments indispensables pour construire aux endroits critiques des digues d'une hauteur appropriée, déterminer les zones inconstructibles, définir la périodicité des opérations de nettoyage des grands cours d'eaux et des estuaires afin de protéger efficacement la population et pourquoi pas les biens (voir Garrido (2002) pour des exemples de fiabilité des structures à **EDF** [1]).

Dans de nombreuses applications, la quantité d'intérêt (la surcote dans l'exemple 1) est parfois mesurée simultanément avec une covariable. Cette covariable peut être unidimensionnelle, multidimensionnelle ou fonctionelle, aléatoire ou déterministe. Elle est de nature diverse en fonction de l'expérience physique ou du phénomène observé. L'exemple suivant illustre une covariable unidimensionnelle.

Exemple 2 (**Modélisation de données de fiabilité de réacteurs nucléaires**)**.** *Afin d'optimiser le changement des cuves de réacteurs nucléaires, le CEA*[2] *dispose de centaines de mesures de la ténacité de l'acier en fonction de la température (voir Figure 1). Il souhaite, soit évaluer la probabilité (a priori très faible) d'apparition de micro fissures dans ses cuves, soit évaluer des bas-fractiles de la ténacité en fonction de la température.*

1. Électricité de France
2. Commissariat à l'Énergie Atomique

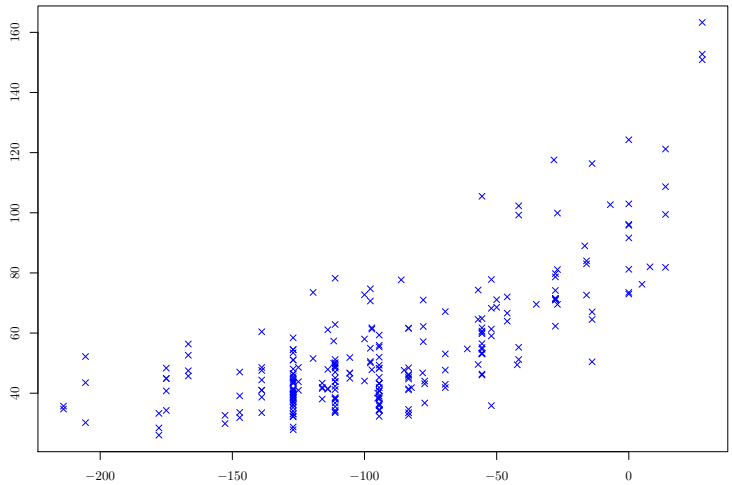


FIGURE 1 – La ténacité de la cuve (verticalement) et la température (horizontalement).

Dans l'exemple 2, la quantité d'intérêt est la ténacité et la covariable est la température. À la Figure 1, nous avons représenté la ténacité de la cuve en fonction de la température. Les données ayant servi à cette illustration ont été fournies par le Laboratoire de Conduite et Fiabilité des Réacteurs (LCFR) du CEA Cadarache. Présentons maintenant un exemple illustrant une covariable multidimensionnelle.

Exemple 3 (**Estimation de niveaux de retour de pluie dans une région).** *Le* **LTHE**[3] *de Grenoble a mesuré les hauteurs de pluies horaires (en mm) entre* 1972 *et* 2000 *sur environ* 300 *stations situées dans la région des Cévennes-Vivarais (France) (voir Figures 2 et 3). Les hydrologues aimeraient évaluer des hauteurs de pluies apparaissant en moyenne toutes les N années lorsque N est supérieur au nombre d'années de mesures. Ils souhaitent également calculer la probabilité d'une hauteur de pluies supérieure au maximum de toutes celles mesurées.*

Dans l'exemple 3, la covariable qui est la position géographique peut être unidimensionnelle (altitude), bidimensionnelle (longitude et latitude) ou tridimensionnelle (latitude, longitude et altitude). La variable d'intérêt est la hauteur de pluies. À la Figure 2, on a donné la localisation des Cévennes-Vivarais en Europe et plus précisément en France et à la Figure 3, on a mis en exergue le relief de cette région, ses rivières, quelques unes de ces importantes villes et les stations d'observation du LTHE.

3. Laboratoire d'Etude des Transferts en Hydrologie et Environnement.

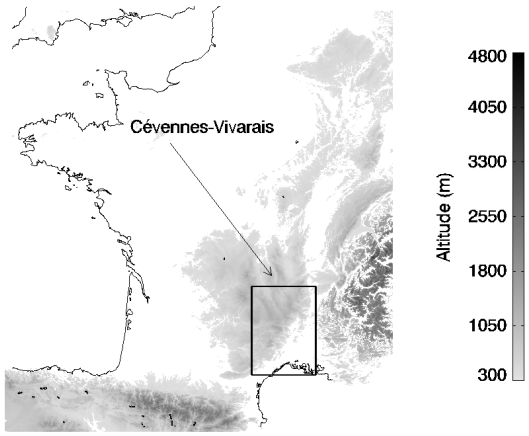


FIGURE 2 – Localisation de la région des Cévennes-Vivarais (France)

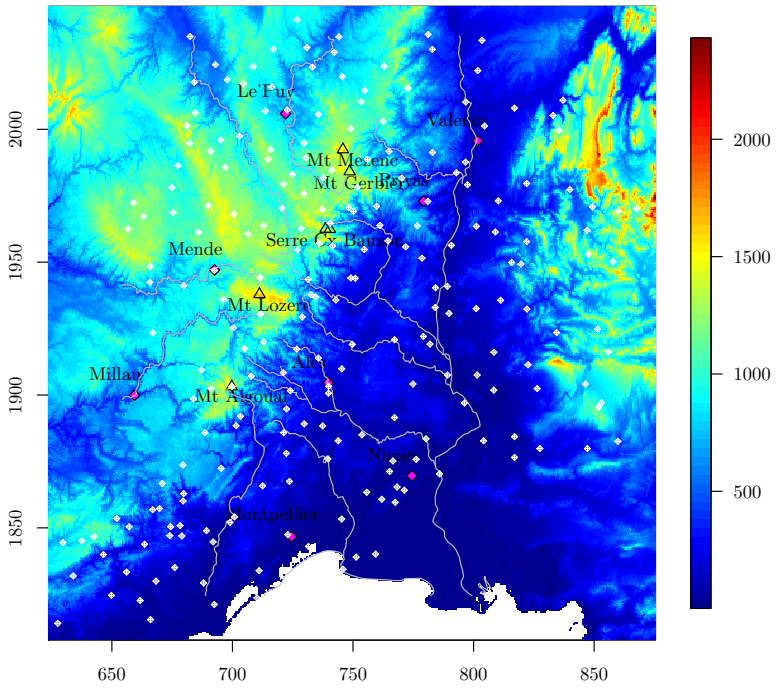


FIGURE 3 – La région des Cévennes-Vivarais et ses alentours représentés en fonction de leurs coordonnées géographiques. Horizontalement : la longitude (en kilomètres), verticalement : la latitude (en kilomètres), l'échelle des couleurs : altitude (en mètres). Sur la carte : les villes (losanges roses), les montagnes (triangles), les cours d'eaux (lignes grises) et 225 stations du LTHE (losanges blancs).

Formalisation du problème

Nous pouvons formaliser les questions soulevées dans les exemples précédents en quatre problèmes dont la description est donnée ci-dessous.

Partant d'un échantillon d'observations indépendantes $\{X_i,\ i = 1,\dots,n\}$, la statistique des valeurs extrêmes se propose d'estimer les quantités extrêmes q_{α_n} associées à une variable aléatoire $X \in \mathbb{R}$ définies par

$$\mathbb{P}\left(X > q_{\alpha_n}\right) = \alpha_n \text{ quand } \alpha_n \to 0 \text{ lorsque } n \to \infty. \tag{1}$$

De façon complémentaire, elle se propose aussi de calculer la probabilité α_n d'observer une quantité extrême y_n définie par

$$\alpha_n \overset{def}{=} \mathbb{P}\left(X > y_n\right) \text{ quand } y_n \to \infty \text{ lorsque } n \to \infty. \tag{2}$$

Les problèmes (1) et (2) ont une difficulté commune. Celle-ci se résume au fait que la probabilité $\mathbb{P}\left(X > y_n\right)$ est inconnue et difficile à évaluer au-delà de l'observation maximale puisque $\mathbb{P}\left(y_n > \max(X_i), \forall i = 1,\dots,n\right) \to 1$ si $n\alpha_n \to 0$ (se référer à la partie 1.1).

Ces deux problèmes s'adressent au cas sans covariable et en particulier à l'exemple 1. Ceux qui suivent prennent en compte la covariable.

Partant d'un échantillon d'observations indépendantes $\{(X_i, Y_i),\ i = 1,\dots,n\}$ du couple (X, Y) où X est une covariable aléatoire ou déterministe et Y une variable d'intérêt réelle et aléatoire, la statistique des valeurs extrêmes conditionnelles s'intéresse, en un point x de la covariable, à l'estimation des quantités extrêmes $q(\alpha_n|x)$ définies par

$$\mathbb{P}\left(Y > q(\alpha_n|x)|X = x\right) = \alpha_n \text{ quand } \alpha_n \to 0 \text{ lorsque } n \to \infty. \tag{3}$$

Par dualité, elle souhaite également évaluer la probabilité conditionnelle définie par

$$\alpha_n \overset{def}{=} \mathbb{P}\left(Y > y_n|X = x\right) \text{ quand } y_n \to \infty \text{ lorsque } n \to \infty. \tag{4}$$

De même, la difficulté des problèmes duaux (3) et (4) est commune. D'une part, la probabilité conditionnelle $\mathbb{P}\left(Y > y_n|X = x\right)$ est inconnue et difficile à estimer au delà du point maximal du sous-échantillon des observations prises dans un voisinage de x et d'autre part, la quantité $q(\alpha_n|x)$ est une fonction de la covariable.

Contributions

Cadre non conditionnel

Bien que les problèmes (1) et (2) ne soient pas le thème central de cette thèse, nous avons tout de même proposé, au **chapitre 2**, une nouvelle méthode d'estimation permettant d'améliorer les solutions existantes. Pour ce faire, nous avons utilisé l'une des modélisations usuelles de l'analyse des valeurs extrêmes. En supposant les observations

de l'échantillon aléatoire $\{X_i,\ i = 1,\dots,n\}$ indépendantes et identiquement distribuées de fonction de répartition F non nécessairement continue, nous avons fait l'hypothèse que la fonction $1-F$, appelée *fonction de survie*, est à décroissance polynomiale d'exposant $-1/\gamma$, avec $\gamma > 0$. Cette hypothèse permet aux statisticiens de ramener le problème (1) à l'estimation de quelques paramètres (voir chapitre 1). Le paramètre γ est d'une très grande importance puisque c'est lui qui contrôle la lourdeur de la queue de la fonction de survie et donc de la loi. Dans la contribution présentée au chapitre 2, nous avons proposé un nouvel estimateur de q_{α_n}. Lorsque la probabilité $\alpha_n < 1/n$, nous avons montré que la loi asymptotique de notre estimateur était gaussienne. En comparaison à quelques méthodes d'estimation courantes de q_{α_n}, l'un de nos principaux résultats montre que notre estimateur améliore, en l'occurrence, *le biais* par rapport à l'estimateur de Weissman (1978).

Cadre conditionnel

Le but principal de cette thèse est de proposer des solutions aux problèmes (3) et (4). Compte tenu de la nature de la covariable, l'étude a été scindée en deux volets :

- Dans le chapitre 4, nous considérons que la covariable X est déterministe et de dimension non nécessairement finie.
- Au chapitre 5, nous nous intéressons aux covariables aléatoires de dimension $p \in \mathbb{N}^*$.

Dans les deux cas, nous supposons que la probabilité conditionnelle que $Y > y_n$ connaissant la valeur x de la covariable X est à décroissance polynomiale d'exposant $-1/\gamma(x)$, où $\gamma(x)$ est le paramètre décrivant la lourdeur de la queue de distribution conditionnelle. Dans un tel contexte, $\gamma(.)$ est une fonction inconnue de la covariable.

Contribution du chapitre 4 : Pour construire des estimateurs de $q(\alpha_n|x)$, on se propose d'utiliser une méthode qui consiste à ne sélectionner que les variables d'intérêts Y_i pour lesquelles les covariables X_i sont suffisamment proches du point x. Des estimateurs de $q(\alpha_n|x)$ sont alors construits en utilisant uniquement les variables d'intérêts Y_i retenues par la méthode de sélection. En fonction de la vitesse de convergence de α_n vers zéro, on distingue trois situations. L'étude théorique des estimateurs ainsi construits montre que dans l'une des situations, la loi asymptotique de l'estimateur de $q(\alpha_n|x)$ n'est pas forcément gaussienne. Cette contribution a donné lieu à la publication Gardes *et al.* (2010). Toutefois, signalons que la contribution présentée dans ce chapitre comporte un petit supplément (cf. Corollaire 4.4.2) par rapport à l'article citée précedemment. *In fine*, nous nous sommes proposés d'adapter l'estimateur du chapitre 2 au cas conditionnel.

Contribution du chapitre 5 : Pour estimer les petites probabilités conditionnelles du problème (4), on se propose d'utiliser les variables aléatoires $\{(X_i, Y_i),\ i = 1,\dots,n\}$ dont les points d'observations X_i sont distants au plus de $h_n > 0$ du point x. Les variables d'intérêts Y_i ainsi retenues se voient attribuer des poids qui tiennent compte de

la distance de leur covariable X_i au point x ce qui n'était pas le cas dans le chapitre 4. Mathématiquement, ces probabilités sont estimées comme une moyenne pondérée de la variable réponse $\mathbb{1}_{\{Y>y_n\}}$, i.e par le rapport :

$$\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right)\mathbb{1}_{\{Y_i>y_n\}} \Bigg/ \sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right),$$

où $\mathbb{1}_{\{Y_i>y_n\}}$ est la fonction qui vaut 1 si $Y_i > y_n$ et 0 sinon et K une densité de probabilité (à support compact). On donne une condition nécessaire et suffisante pour que l'estimateur de la probabilité conditionnelle $\mathbb{P}\left(Y > y_n | X = x\right)$ soit asymptotiquement gaussien.

De cette méthode d'estimation des petites probabilités conditionnelles, on déduit des solutions au problème (3). En fonction de la vitesse de convergence de α_n vers zéro, on a envisagé deux situations. Pour chacune d'elles, on propose un estimateur de $q(\alpha_n|x)$. L'étude théorique des estimateurs ainsi construits nous permet d'introduire des estimateurs de la fonction $\gamma(x)$. Nous avons donné la loi limite de ces estimateurs. Ces travaux ont donné suite à la publication Daouia *et al.* (2010).

Organisation de la thèse

Cette thèse s'organise en cinq chapitres. Ceux-ci ont été rédigés de façon à être lus indépendamment. Ainsi, le chapitre 1 présente l'état de l'art en théorie des valeurs extrêmes. On se limite au cas univarié réel. Au cours de ce chapitre, on présente tout d'abord le résultat principal de la théorie des valeurs extrêmes. Celui-ci montre que, à l'exception de certaines lois pathologiques, on peut regrouper les lois usuelles en des groupes appelés *domaines d'attractions*. Nous exposons les critères pour qu'une loi appartienne à l'un de ces groupes. Nous présentons ensuite les méthodes classiques permettant de résoudre les problèmes (1) et (2) et donc d'estimer q_{α_n} et γ.

Nous proposons, au chapitre 2, un nouvel estimateur de q_{α_n} puis du paramètre γ dont nous établissons les lois asymptotiques. Afin d'apprécier le comportement de ces nouveaux outils statistiques, des comparaisons graphiques sont proposées.

Le chapitre 3 sera consacré à l'état de l'art des problèmes (3) et (4) quand $\alpha_n = \alpha \in]0,1[$ est fixé. Nous y présentons les différentes approches et méthodes d'estimation et donnons quelques résultats clés dont certains nous servirons de critères de comparaison dans la suite.

Dans le chapitre 4, après avoir présenté les contributions existantes dans le cas $\alpha_n \to 0$, on expose clairement le cadre d'étude dans lequel on va investiguer. Ensuite, on définit nos estimateurs. En dehors de l'étude théorique qui y sera menée, on illustrera notre approche d'estimation d'abord sur des données fonctionnelles et ensuite sur un

exemple d'hydrologie.

À quelques exceptions, le plan du chapitre 5 est similaire à celui du chapitre 4. Après avoir mené l'étude théorique de nos estimateurs, on s'attardera sur des simulations dont le but est d'apprécier la qualité de nos estimateurs. Comme illustration, nous proposons une application de notre méthodologie à la télédétection spatiale.

Finalement, on conclura ce mémoire par des perspectives.

1

Quelques éléments de la théorie des valeurs extrêmes

Résumé

Dans ce chapitre, on regroupe des définitions et des résultats sur la théorie des valeurs extrêmes dans le cas univarié réel.

Sommaire

1.1 Introduction **10**
1.2 Lois des valeurs extrêmes **10**
1.3 Caractérisation des domaines d'attractions **13**
1.3.1 Résultats sur les fonctions à variations régulières 14
1.3.2 Domaine d'attraction de Fréchet 15
1.3.3 Domaine d'attraction de Weibull 16
1.3.4 Domaine d'attraction de Gumbel 16
1.4 Estimation des quantiles extrêmes **16**
1.4.1 Quelques résultats sur les statistiques d'ordres 18
1.4.2 Approche des quantiles extrêmes par la loi des valeurs extrêmes . . . 19
1.4.3 Approche des quantiles extrêmes par la méthode des excès 20
1.4.4 Approche des quantiles extrêmes par l'approche semi-paramétrique 22
1.5 Estimation du paramètre de la loi des valeurs extrêmes **23**
1.5.1 Estimateur de Hill 25
1.5.2 Estimateur de Pickands 26
1.5.3 Estimateur de Zipf 28

1.1 Introduction

ON souhaite 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* et on parle de *quantile extrême* lorsque l'ordre du quantile (probabilité d'observation) converge vers zéro quand la taille de l'échantillon tend vers l'infini.

Plus précisément, on considère n variables aléatoires réelles $\{X_i,\ i = 1,\dots,n\}$ indépendantes et identiquement distribuées de fonction de répartition F non nécessairement continue. A partir des observations de ces variables aléatoires, on souhaite estimer le quantile extrême d'ordre $\alpha_n \to 0$ quand $n \to \infty$ défini par

$$q_{\alpha_n} \overset{def}{=} \bar{F}^{\leftarrow}(\alpha_n) = \inf\{x : \bar{F}(x) \leq \alpha_n\}, \tag{1.1}$$

avec la convention $\inf\{\emptyset\} = \infty$, où $\bar{F}^{\leftarrow}$ est l'inverse généralisé de $\bar{F} = 1 - F$. En particulier, pour n tendant vers l'infini, on a

$$\begin{aligned}
\mathbb{P}\big(\max(X_1,\dots,X_n) < q_{\alpha_n}\big) &= \mathbb{P}\big(X_i < q_{\alpha_n},\ \forall i = 1,\dots,n\big) \\
&= (1-\alpha_n)^n \\
&= \exp\big(n\log(1-\alpha_n)\big) \\
&= \exp(-n\alpha_n(1+o(1)) \ \text{ quand } \alpha_n \to 0.
\end{aligned}$$

Par conséquent, comme $\alpha_n \to 0$, supposer que $n\alpha_n \to 0$ quand $n \to \infty$ implique que $\mathbb{P}\big(\max(X_1,\dots,X_n) < q_{\alpha_n}\big) \to 1$. On ne peut donc pas estimer q_{α_n} en inversant simplement la fonction de répartition empirique. Plusieurs méthodes d'estimation du quantile extrême q_{α_n} ont été proposées dans la littérature. Mais avant d'exposer celles-ci, il convient d'exposer la théorie sous-jacente à l'étude du maximum d'un échantillon.

Ainsi, dans ce chapitre, on fera une brève introduction sur l'étude du comportement asymptotique du maximum d'un échantillon. On commencera par introduire, dans la partie 1.2, un résultat décrivant les limites possibles d'un tel maximum. Dans la partie 1.3, nous donnons des critères pour que la limite en loi du maximum suive telle ou telle loi des valeurs extrêmes. Ces critères faisant appel à la notion de fonctions à variations régulières, on rappelle au préalable la définition de telles fonctions et on en donne quelques propriétés. Puis, dans la partie 1.4, nous présentons quelques méthodes d'estimation des quantiles extrêmes. Enfin, dans la partie 1.5, nous exposons deux méthodes statistiques qui permettent d'identifier la loi limite.

1.2 Lois des valeurs extrêmes

Le principal résultat de la théorie des valeurs extrêmes repose sur le Théorème de Fisher et Tippet (1928) dont la première preuve rigoureuse est due à Gnedenko (1943). Celui-ci fait appel à la notion de statistique d'ordre associée aux variables aléatoires

$\{X_i, i = 1, \ldots, n\}$. De ce fait, nous ne saurons aborder ce paragraphe sans en donner au préalable une définition.

Définition 1.2.1. *Notons par $X_{1,n} \leq \ldots \leq X_{n,n}$ le réarrangement croissant de l'échantillon $\{X_i,\ i = 1, \ldots, n\}$. Pour tout $i = 1, \ldots, n$, la variable aléatoire $X_{i,n}$ s'appelle la ième statistique d'ordre de l'échantillon.*

Nous pouvons maintenant exposer le résultat principal de la théorie des valeurs extrêmes. Celui-ci montre qu'il existe des suites $(a_n)_{n\geq 1} > 0$ et $(b_n)_{n\geq 1} \in \mathbb{R}$ telles que la suite de variables aléatoires $\left(a_n^{-1}\left(X_{n,n} - b_n\right)\right)_{n\geq 1}$ converge en loi vers une limite non dégénérée. Ce résultat est la base de la théorie des valeurs extrêmes. C'est l'équivalent du Théorème central limite **(TCL)** en ce qui concerne les maxima.

Théorème 1.2.1 (Fisher et Tippet (1928); Gnedenko (1943))**.** *Sous certaines conditions de régularité sur la fonction de répartition F, il existe un paramètre réel γ et deux suites $(a_n)_{n\geq 1} > 0$ et $(b_n)_{n\geq 1} \in \mathbb{R}$ tels que pour tout $x \in \mathbb{R}$,*

$$\lim_{n\to\infty} \mathbb{P}\left[\frac{X_{n,n} - b_n}{a_n} \leq x\right] = \mathscr{H}_\gamma(x),$$

avec

$$\begin{aligned}
&\textit{loi de Fréchet} \quad \mathscr{H}_\gamma(x) = \begin{cases} 0 & \textit{si } x < 0 \\ \exp\left[-(x)^{-1/\gamma}\right] & \textit{si } x \geq 0 \end{cases} \quad \textit{et } \gamma > 0,\\
&\textit{loi de Weibull} \quad \mathscr{H}_\gamma(x) = \begin{cases} \exp\left[-(-x)^{-1/\gamma}\right] & \textit{si } x < 0 \\ 1 & \textit{si } x \geq 0 \end{cases} \quad \textit{et } \gamma < 0,\\
&\textit{loi de Gumbel} \quad \mathscr{H}_0(x) = \exp\left[-\exp(-x)\right] \ \textit{si } x \in \mathbb{R} \quad \textit{et } \gamma = 0,
\end{aligned}$$

et $\mathscr{H}_\gamma$ la fonction de répartition de la **loi des valeurs extrêmes (EVD)**.

Le Théorème 1.2.1 est vrai pour la majorité des lois usuelles. Si l'on fait un parallèle avec le TCL, la suite a_n joue le rôle de $n^{-1/2}\sigma(X)$ où $\sigma(X)$ désigne l'écart type de X et la suite b_n joue le rôle de l'espérance. La suite a_n (resp. b_n) s'interprète comme un paramètre d'échelle (resp. un paramètre de position ou de centrage). De plus, les suites a_n et b_n ne sont pas uniques.

La Figure 1.1 illustre dans le cas d'une loi normale centrée réduite, la convergence de la suite de variables aléatoires $\left(a_n^{-1}\left(X_{n,n} - b_n\right)\right)_{n\geq 1}$ en loi vers une limite non dégénérée $\mathscr{H}_0$. Dans cet exemple, nous avons utilisé les suites de renormalisation théoriques associées à la loi normale standard (Embrechts *et al.*, 1997, voir page 145), à savoir :

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

Grâce aux travaux de von Mises (1936) et de Jenkinson (1955), on a une forme unifiée de la fonction de répartition de la loi EVD à un facteur d'échelle et de position près.

Définition 1.2.2. *La représentation de Jenkinson-von Mises de la loi EVD que l'on appelle* **loi des valeurs extrêmes généralisée** *notée* **(GEVD)** *ou* **(GEV)** *a pour fonction de répartition*

$$\Lambda_\gamma(x) = \begin{cases} \exp\left[-(1+\gamma x)^{-1/\gamma}\right] & si\ \gamma \neq 0 \\ \exp\left[-\exp(-x)\right] & si\ \gamma = 0 \end{cases} \quad et\ pour\ tout\ x\ tel\ que\ 1+\gamma x > 0.$$

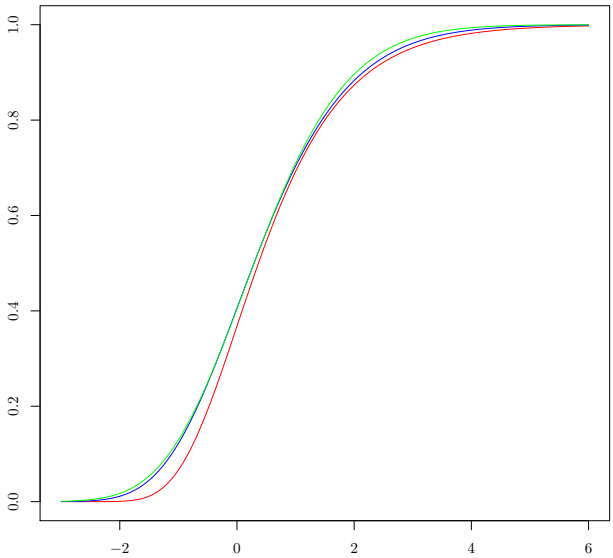


FIGURE 1.1 – Illustration de la théorie des valeurs extrêmes sur une loi normale centrée réduite. Comparaison entre $\mathscr{H}_0(x)$ (rouge), $\mathbb{P}\left[\frac{X_{n,n}-b_n}{a_n} \leq x\right]$ avec $n = 100$ (bleu) et $\mathbb{P}\left[\frac{X_{n,n}-b_n}{a_n} \leq x\right]$ avec $n = 10$ (vert).

Définition 1.2.3. *Le paramètre γ du Théorème 1.2.1 ou de la Définition 1.2.2 est un paramètre de forme que l'on appelle « indice des valeurs extrêmes » ou « indice de queue ».*

Si F vérifie le Théorème 1.2.1, alors on dit alors que F appartient au « *domaine d'attraction* » de $\mathscr{H}_\gamma$ ou Λ_γ et selon le signe de γ, on distingue trois domaines d'attraction dont quelques densités ont été représentées pour γ fixé à la Figure 1.2.

1. Si $\gamma > 0$, on dit que F appartient au domaine d'attraction de **Fréchet** (voir Fréchet, 1927) et on notera $F \in \mathscr{D}(Fréchet)$. Ce domaine d'attraction est celui des distributions à queues lourdes, i.e qui ont une fonction de survie à décroissance polynomiale. Comme exemple de lois appartenant à ce domaine d'attraction on a les lois de Cauchy, de Pareto, du Chi-deux, de Student, de Burr, de Fréchet, etc.

2. Si $\gamma < 0$, on dit que F appartient au domaine d'attraction de **Weibull** et on notera $F \in \mathscr{D}(Weibull)$. Ce domaine d'attraction est celui des fonctions de survie dont le support est borné supérieurement. Pour le domaine d'attraction de Weibull on trouve les lois Uniforme, Beta, de Weibull, etc.
3. Si $\gamma = 0$, on dit que F appartient au domaine d'attraction de **Gumbel** et on notera $F \in \mathscr{D}(Gumbel)$. Ce domaine d'attraction est celui des distributions à queues légères, i.e qui ont une fonction de survie à décroissance exponentielle. Dans ce domaine d'attraction, on regroupe les lois Normale, Exponentielle, Log-normale, Gamma, Weibull$_m$, etc. Il apparait important de signaler ici que les lois de Weibull et Weibull$_m$ sont deux lois différentes. La loi de Weibull$_m$ de paramètres $\tau > 0$ et $c > 0$ a pour fonction de répartition

$$F(x) = \begin{cases} 1 - \exp(-cx^{\tau}) & \text{si } x \geq 0 \\ 0 & \text{sinon.} \end{cases}$$

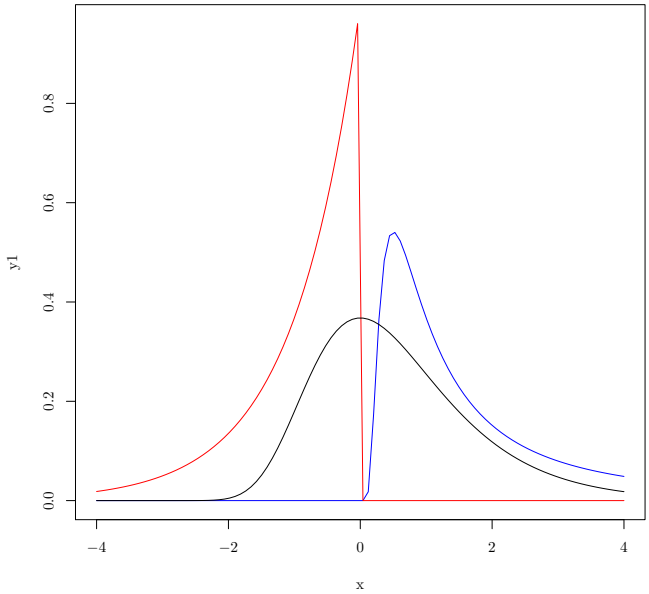


FIGURE 1.2 – Exemple de densités associées à la loi EVD avec $\gamma = -1$ (rouge), $\gamma = 0$ (noir) et $\gamma = 1$ (bleu).

1.3 Caractérisation des domaines d'attractions

Le lecteur pourra se référer aux ouvrages de Resnick (1987) et Embrechts *et al.* (1997) pour une présentation plus détaillée. Dans cette partie, on se contentera de donner des conditions sur la fonction de répartiton F pour qu'elle appartienne à l'un des domaines

d'attraction définis précédemment. Ces conditions faisant appel à la notion de fonctions à variations régulières, on rappelle au préalable la définition de telles fonctions et on en donne quelques propriétés.

1.3.1 Résultats sur les fonctions à variations régulières

Pour décrire plus en détail les fonctions à variations régulières, il est nécessaire de commencer par une définition des fonctions à variations lentes.

Définition 1.3.1. *On dit qu'une fonction ℓ est à variations lentes à l'infini si $\ell(x) > 0$ pour x assez grand et si pour tout $\lambda > 0$, on a*

$$\lim_{x\to\infty} \frac{\ell(\lambda x)}{\ell(x)} = 1.$$

Par exemple, la fonction constante, les fonctions qui tendent vers une constante et $\log(x)$ sont des fonctions à variations lentes à l'infini. Les fonctions à variations lentes jouent un rôle prépondérant dans l'étude des lois de valeurs extrêmes.

1.3.1.1 Représentation de Karamata

Théorème 1.3.1 (Resnick (1987), **Corollaire du Théorème 0.6**). *Toute fonction à variations lentes ℓ s'écrit sous la forme*

$$\ell(x) = c(x) \exp \int_1^x \frac{\varepsilon(t)}{t} dt. \tag{1.2}$$

où $c > 0$ et ε sont deux fonctions mesurables telles que

$$\lim_{x\to\infty} c(x) = c_0 \in]0,\infty[\text{ et } \lim_{x\to\infty} \varepsilon(x) = 0.$$

Si la fonction c est constante, alors on dit que ℓ est *normalisée*. L'équation (1.2) implique que si ℓ est normalisée alors ℓ est dérivable de dérivée ℓ' avec pour tout $x > 0$,

$$\ell'(x) = \frac{\varepsilon(x)\ell(x)}{x}.$$

En particulier, on a

$$\lim_{x\to\infty} x\frac{\ell'(x)}{\ell(x)} = 0.$$

1.3.1.2 Définition d'une fonction à variations régulières et quelques résultats

Définition 1.3.2. *On dit qu'une fonction G est à variations régulières d'indice $\rho \in \mathbb{R}$ à l'infini si G est positive à l'infini (i.e s'il existe A tel que pour tout $x \geq A$, $G(x) > 0$) et si pour tout $\lambda > 0$*

$$\lim_{x\to\infty} \frac{G(\lambda x)}{G(x)} = \lambda^\rho.$$

Dans le cas particulier où $\rho = 0$, G est une fonction à variations lentes à l'infini. Une fonction à variations régulières d'indice ρ peut toujours s'écrire sous la forme $x^\rho \ell(x)$ où ℓ est une fonction à variations lentes à l'infini.

Le résultat du Lemme suivant montre que la convergence du rapport de deux fonctions à variations régulières d'indice ρ de la Définition 1.3.2 est localement uniforme lorsque x tend vers l'infini.

Lemme 1.3.1 (Resnick (1987), **Proposition 0.5**). *Si G est une fonction à variations régulières d'indice ρ, alors pour tout $0 < a < b$*

$$\lim_{x\to\infty} \sup_{\lambda\in[a,b]} \left| \frac{G(\lambda x)}{G(x)} - \lambda^\rho \right| = 0.$$

Lemme 1.3.2 (**Inverse d'une fonction à variation régulières**).

- *Si G est à variations régulières d'indice $\rho > 0$, alors $G^{\leftarrow}(x)$ est à variations régulières d'indice $1/\rho$.*
- *Si G est à variations régulières d'indice $\rho < 0$, alors $G^{\leftarrow}(1/x)$ est à variations régulières d'indice $-1/\rho$.*

Pour une preuve du Lemme 1.3.2, le lecteur pourra se référer au Théorème 1.5.12 de l'ouvrage de Bingham *et al.* (1987) ou à la Proposition 2.6 du livre de Resnick (1987).

Lemme 1.3.3. *Soit ℓ une fonction à variations lentes et soient u_n et v_n deux suites positives telles que $u_n \to +\infty$ et $v_n \to +\infty$. Si $u_n \sim v_n$ (ie $u_n/v_n \to 1$ quand $n \to +\infty$), alors $\ell(u_n) \sim \ell(v_n)$.*

Le Lemme 1.3.3, qui est une conséquence du Lemme 1.3.1, montre que les fonctions à variations lentes conservent les équivalents. Par conséquent, ce résultat implique que si G est à variations régulières d'indice ρ et si $u_n \sim v_n$, alors $G(u_n) \sim G(v_n)$ et on dit que les fonctions à variations régulières conservent aussi les équivalents.

1.3.2 Domaine d'attraction de Fréchet

Avant de donner les expressions de la fonction de répartition dans chaque domaine, il convient d'introduire une définition utile.

Définition 1.3.3. *On appelle point terminal (en anglais « upper endpoint ») de la fonction F, le réel x_F défini par*

$$x_F = \sup\{x : F(x) < 1\},$$

avec la convention $\sup\{\emptyset\} = \infty$.

Théorème 1.3.2. *$F \in \mathcal{D}(Fréchet)$ avec un indice des valeurs extrêmes $\gamma > 0$ si et seulement si $x_F = +\infty$ et la fonction $\bar{F}$ est à variations régulières d'indice $-1/\gamma$. Dans ce cas, un choix possible pour les suites a_n et b_n est*

$$a_n = q_{1/n} \text{ et } b_n = 0.$$

De ce Théorème 1.3.2, on en déduit que $F \in \mathscr{D}(Fréchet)$ si et seulement si le point terminal x_F est infini et $\bar{F}(x) = x^{-1/\gamma}\ell(x)$, où ℓ est une fonction à variations lentes à l'infini et γ un réel strictement positif.

1.3.3 Domaine d'attraction de Weibull

Théorème 1.3.3. *$F \in \mathscr{D}(Weibull)$ avec un indice des valeurs extrêmes $\gamma < 0$ si et seulement $x_F < +\infty$ et la fonction $\left(1 - \tilde{F}\right)$ est à variations régulières d'indice $1/\gamma$ avec*

$$\tilde{F}(x) = \begin{cases} 0 & si\ x \leq 0, \\ F(x_F - x^{-1}) & si\ x > 0. \end{cases}$$

Dans ce cas, un choix possible pour les suites a_n et b_n est

$$a_n = x_F - q_{1/\mathrm{n}} \text{ et } b_n = x_F.$$

Du Théorème 1.3.3, on en déduit que $F \in \mathscr{D}(Weibull)$ si et seulement si le point terminal x_F est fini et $\bar{F}(x) = (x_F - x)^{-1/\gamma}\ell\left[(x_F - x)^{-1}\right]$ avec ℓ est une fonction à variations lentes à l'infini et γ un réel strictement négatif.

1.3.4 Domaine d'attraction de Gumbel

Les résultats concernant le domaine d'attraction de la loi de Gumbel sont plus délicats. Se référer aux livres de Beirlant *et al.* (2004b) ou de Haan et Ferreira (2006) pour une présentation exhaustive.

Théorème 1.3.4. *$F \in \mathscr{D}(Gumbel)$ si et seulement si*

$$\bar{F}(x) = c(x)\exp\left\{-\int_z^x \frac{1}{a(t)}dt\right\},$$

où $\lim_{x \to x_F} c(x) = c$. Dans ce cas, un choix possible pour les suites a_n et b_n est

$$a_n = q_{1/\mathrm{n}} \text{ et } b_n = \frac{1}{\bar{F}(a_n)}\int_{a_n}^{x_F} \bar{F}(y)dy = q_{1/(\mathrm{ne})} - q_{1/\mathrm{n}}.$$

1.4 Estimation des quantiles extrêmes

Dans tout ce qui suit, on fait l'hypothèse que F appartient à l'un des domaines d'attractions définis précédemment. Afin de résumer le problème d'estimation investigué dans ces travaux, on introduit le résultat suivant appelé *approximation de Poisson.*

Lemme 1.4.1 (Embrechts *et al.* (1997), **Proposition 3.1.1**). *Si $\alpha_n \to 0$ et $n\alpha_n \to c$ (non nécessairement fini) quand $n \to \infty$, alors*

$$\mathbb{P}\left(X_{n,n} < q_{\alpha_n}\right) \to e^{-c}.$$

Ainsi, d'après le Lemme 1.4.1, deux situations peuvent alors être distinguées en fonction de c lorsque l'on désire estimer les quantiles d'ordre $\alpha_n \to 0$ quand $n \to \infty$.

Primo, si $c = \infty$ alors, $\mathbb{P}(X_{n,n} < q_{\alpha_n}) = 0$. Dans un tel contexte, un estimateur naturel de q_{α_n} est le quantile empirique qui n'est rien d'autre que la $\lfloor n\alpha_n \rfloor$ ième plus grande observation de l'échantillon $\{X_i,\ i = 1,\dots,n\}$ c'est-à-dire la statistique d'ordre $X_{n-\lfloor n\alpha_n \rfloor+1,n}$. Le résultat principal de cet estimateur de quantile extrême est donné au paragraphe 1.4.1 par le Théorème 1.4.1.

Secundo, si $c = 0$ alors, $\mathbb{P}(X_{n,n} < q_{\alpha_n}) = 1$. Par conséquent, on ne peut pas estimer le quantile empiriquement. Pour résoudre ce problème, on a répertorié trois catégories principales de méthodes :

- La théorie des valeurs extrêmes présentée par Guida et Longo (1988) et dont les premiers éléments bibliographiques remontent à Fisher et Tippet (1928) et Gnedenko (1943) consiste à diviser l'échantillon en m_0 sous-groupes disjoints de taille $n_0 = n/m_0$ desquels on détermine les maxima. La loi de ces maxima est alors approchée, pour n_0 assez grand, par une loi des valeurs extrêmes. En utilisant la relation $\mathbb{P}\left(\max(X_1,\dots,X_n) < q_{\alpha_n}\right) = F^n(q_{\alpha_n})$ on peut alors estimer le quantile extrême q_{α_n}. Cette méthode d'estimation est présentée au paragraphe 1.4.2.
- La méthode des excès initialement présentée par Pickands (1975). Elle préconise de ne retenir que les observations dépassant un seuil fixé u. La loi des k_n observations ainsi retenues que l'on note par $\{Y_i,\ i = 1,\dots,k_n\}$ peut-être approchée, si u est assez grand par *une loi de Pareto généralisée* **(GPD)** (voir Définition 1.4.2). Pour estimer le quantile extrême q_{α_n}, il suffit alors d'utiliser le résultat de (Balkema et de Haan, 1974; Pickands, 1975) qui établit l'équivalence entre la convergence en loi du maximum vers une loi des valeurs extrêmes et la convergence en loi d'un excès vers une GPD. Nous exposons cette approche d'estimation des quantiles extrêmes au paragraphe 1.4.3.
- Les méthodes semi-paramétriques où l'on suppose que pour tout $\gamma > 0$ on a $\mathbb{P}(X > x) \sim x^{-1/\gamma}$ lorsque x tend vers l'infini, c'est-à-dire que la fonction de survie $\bar{F}(x)$ décroit en $x^{-1/\gamma}$. Cette hypothèse permet de construire des estimateurs non paramétriques du paramètre γ dont le plus célèbre est l'estimateur de Hill (1975). Partant de ce résultat, Weissman (1978) proposera trois ans plus tard un estimateur du quantile extrême q_{α_n}. En effet comme nous venons de voir (c.f Lemme 1.3.2), supposer que $\mathbb{P}(X > x) \sim x^{-1/\gamma}$ revient à supposer que le quantile q_{α_n} décroit en $\alpha_n^{-\gamma}$. Une présentation détaillée de cette méthode d'estimation est donnée au paragraphe 1.4.4.

1.4.1 Quelques résultats sur les statistiques d'ordres

Pour des raisons de commodités, commençons par rappeler un résultat sur la transformation des quantiles.

Lemme 1.4.2 (**Transformation des quantiles**). *Soient* $\{X_i,\ i = 1,\dots,n\}$ *des variables aléatoires indépendantes et de fonction de répartition* F. *Soient* $\{V_{1,n},\dots,V_{n,n}\}$ *les statistiques d'ordres associées aux variables aléatoires indépendantes de loi uniforme standard* $\{V_i,\ i = 1,\dots,n\}$. *Alors,*

- $F^{\leftarrow}(V_1) \overset{\mathscr{L}}{=} X_1$.
- $\left\{F^{\leftarrow}(V_{1,n}),\dots,F^{\leftarrow}(V_{n,n})\right\} \overset{\mathscr{L}}{=} \{X_{1,n},\dots,X_{n,n}\}$, *pour tout* $n \in \mathbb{N}$.
- $F(X_1)$ *suit une loi uniforme standard si et seulement si* F *est continue.*

Ce Lemme nous assure que l'étude des quantiles empiriques associés à une loi quelconque peut se déduire de l'étude des statistiques d'ordres associées à la loi uniforme sur $[0,1]$. Donnons à présent un résultat sur les statistiques d'ordres de lois uniforme et exponentielle. Soit $\{E_i,\ i = 1,\dots,n\}$ une suite de variables aléatoires de loi exponentielle standard. On note $T_n = \sum_{i=1}^{n} E_i$. Soit $\{V_i,\ i = 1,\dots,n\}$ une suite de variables aléatoires indépendantes de loi uniforme standard.

Lemme 1.4.3 (**Représentation de Rényi (1953)**). *La suite de variables aléatoires* $\{V_{1,n},\dots,V_{n,n}\}$ *a même loi que la suite de variables variables aléatoires* $\left\{\frac{T_1}{T_{n+1}},\dots,\frac{T_n}{T_{n+1}}\right\}$, *ie*

$$\left\{V_{j,n} \overset{\mathscr{L}}{=} \frac{T_j}{T_{n+1}}\right\}_{\{j=1,\dots,n\}}.$$

Ce Lemme nous permet d'établir la convergence faible d'une suite de statistiques d'ordres de loi exponentielle.

Proposition 1.4.1. *Soit* $(k_n)_{n\geq 1}$ *une suite d'entiers telle que* $1 \leq k_n \leq n$, $k_n \to \infty$ *et* $k_n/n \to 0$ *quand* $n \to \infty$. *Alors*

$$\sqrt{k_n}\left(E_{n-k_n+1,n} - \log(n/k_n)\right) \overset{\mathscr{L}}{\to} \mathscr{N}(0,1).$$

Définition 1.4.1. *Pour tout* $t > 0$, *la fonction queue (en anglais « tail quantile function ») est donnée par*

$$U(t) \overset{def}{=} \inf\{x : F(x) \geq 1 - 1/t\} = q_{1/t}.$$

L'utilité de cette définition est essentiellement d'ordre pratique. En effet, il peut être plus commode de travailler non pas sur la fonction de survie $\bar{F}$ elle-même, mais plutôt sur la fonction queue U.

Théorème 1.4.1 (de Haan et Ferreira (2006), **Théorème 2.2.1**). *Soit* $(k_n)_{n\geq 1}$ *une suite d'entiers telle que* $1 \leq k_n \leq n$, $k_n \to \infty$ *et* $k_n/n \to 0$ *quand* $n \to \infty$. *Si* $\lim_{x\to x_F^-} \frac{\bar{F}(x)F''(x)}{(F'(x))^2} = -\gamma - 1$, *alors*

$$\sqrt{k_n}\left(\frac{X_{n-k_n+1,n} - U(n/k_n)}{(n/k_n)U'(n/k_n)}\right) \xrightarrow{\mathscr{L}} \mathscr{N}(0,1).$$

Le Théorème 1.4.1 montre que l'estimateur de quantile extrême $X_{n-k_n+1,n}$ avec $k_n \overset{def}{=} \lfloor n\alpha \rfloor$ est asymptotiquement gaussien. Aussi, remarquons que la Proposition 1.4.1 est un cas particulier du Théorème 1.4.1 avec $U = \log$.

Proposition 1.4.2. *Dans le cas particulier où* $\gamma > 0$, *si la condition* $\lim_{x\to\infty} \frac{xU'(x)}{U(x)} = \gamma$ *est vérifiée alors, on a*

$$\sqrt{k_n}\left(\frac{X_{n-k_n+1,n}}{q_{k_n/n}} - 1\right) \xrightarrow{\mathscr{L}} \mathscr{N}\left(0,\gamma^2\right).$$

La Proposition 1.4.2 est encore un cas particulier du Théorème 1.4.1. Il apparait que pour des lois dont la fonction de répartition $F \in \mathscr{D}(Fréchet)$, l'estimateur de quantile extrême $X_{n-k_n+1,n}$ est asymptotiquement normal avec une variance asymptotiquement proportionnelle à γ^2/k_n.

1.4.2 Approche des quantiles extrêmes par la loi des valeurs extrêmes

Pour estimer le quantile q_{α_n}, Guida et Longo (1988) utilisent l'approximation $\mathbb{P}\left(X_{n,n} \leq a_n x + b_n\right) = F^n(a_n x + b_n) \simeq \mathscr{H}_\gamma(x)$. En effet d'après le Théorème 1.2.1, on peut écrire

$$\lim_{n\to\infty} n \log F(a_n x + b_n) = \log \mathscr{H}_\gamma(x),$$

soit encore

$$\lim_{n\to\infty} n \log\left[1 - \bar{F}(a_n x + b_n)\right] = \log \mathscr{H}_\gamma(x).$$

Comme $n \to \infty$, on peut montrer que $a_n x + b_n \to x_F$ et par conséquent $\bar{F}(a_n x + b_n)$ converge vers 0. Un développement limité au premier ordre de $\log(1+u)$ donne donc

$$\bar{F}(a_n x + b_n) \sim -\frac{1}{n} \log \mathscr{H}_\gamma(x).$$

Pour tout γ, on peut alors approcher le quantile q_{α_n} par :

$$q_{\alpha_n} \simeq a_n x_{\alpha_n} + b_n \text{ où } x_{\alpha_n} \text{ vérifie } -\log \mathscr{H}_\gamma(x_{\alpha_n}) = n\alpha_n.$$

On a alors un estimateur de quantile extrême de type

$$\begin{aligned} \hat{q}_{\alpha_n} &= \hat{a}_n x_{\alpha_n} + \hat{b}_n \\ &= \begin{cases} \hat{a}_n (n\alpha_n)^{-\hat{\gamma}} + \hat{b}_n & \text{si } F \in \mathscr{D}(Fréchet) \\ -\hat{a}_n (n\alpha_n)^{-\hat{\gamma}} + \hat{b}_n & \text{si } F \in \mathscr{D}(Weibull) \\ -\hat{a}_n \log(n\alpha_n) + \hat{b}_n & \text{si } F \in \mathscr{D}(Gumbel) \end{cases} \end{aligned} \tag{1.3}$$

où $(\hat{a}_n, \hat{b}_n)$ et $\hat{\gamma}$ sont respectivement des estimateurs des suites (a_n, b_n) et de l'indice de queue γ.

Dans le cas particulier où $\gamma = 0$, les auteurs proposent d'utiliser l'approche basée sur la loi GEV dont le résultat s'énonce ainsi.

Théorème 1.4.2 (Weinstein (1973)). *Soit $F \in \mathscr{D}(Gumbel)$, il existe deux suites $(a_n)_{n\geq 1} > 0$ et $(b_n)_{n\geq 1} \in \mathbb{R}$ telles que pour tout $x \in \mathbb{R}$ et $v > 0$,*

$$\lim_{n\to\infty} n\bar{F}\left[\left(b_n^v + c_n x\right)^{1/v}\right] = \exp(-x), \tag{1.4}$$

où $c_n = a_n v b_n^{v-1}$.

Dans une telle situation, on approche le quantile par

$$q_{\alpha_n} \simeq \left(b_n^v + c_n x_{\alpha_n}\right)^{1/v} \text{ où } x_{\alpha_n} \text{ vérifie } \exp(-x_{\alpha_n}) = n\alpha_n,$$

et un estimateur du quantile extrême est obtenu en remplaçant les suites b_n et c_n respectivement par leurs estimateurs $\hat{b}_n$ et $\hat{c}_n$, i.e

$$\hat{q}_{\alpha_n} = \left(\hat{b}_n^v - \hat{c}_n \log(n\alpha_n)\right)^{1/v}.$$

L'avantage d'utiliser le résultat (1.4) provient du fait qu'il existe des valeurs de v pour lesquelles la convergence dans (1.4) est plus rapide que dans le cas $v = 1$. Dans ce cas, sur simulation, l'approximation du quantile q_{α_n} est de meilleure qualité que l'approximation basée sur l'approche EVD, i.e avec $v = 1$. Les auteurs fournissent la valeur optimale du paramètre v.

Les paramètres γ, a_n, b_n et c_n de ces lois peuvent être estimés par la méthode du maximum de vraisemblance (Prescott et Walden, 1980, 1983) ou la méthode des moments pondérés (Hosking *et al.*, 1985). Dans le cas de l'approche EVD, Smith (1985) fait une étude détaillée du comportement asymptotique des estimateurs des paramètres γ, a_n, b_n obtenus par la méthode du maximum de vraisemblance. Toutefois, il est conseillé d'utiliser les estimateurs des moments pondérés car ceux-ci sont non seulement explicites et faciles à calculer mais aussi parce qu'ils donnent de meilleurs résultats que les estimateurs du maximum de vraisemblance quand on a des échantillons de petite ou de moyenne taille. La principale difficulté de l'estimation des paramètres γ, a_n, b_n et c_n est due au fait qu'il faut un échantillon de maxima, lequel est parfois difficile à extraire des données initiales.

1.4.3 Approche des quantiles extrêmes par la méthode des excès

Avant de présenter cette approche, il convient de commencer par une définition.

Définition 1.4.2 (**Loi de Pareto Généralisée (GPD)**). *La loi de Pareto Généralisée est la loi dont la fonction de répartition est donnée par*

$$\mathscr{G}_{\gamma,\beta}(y) = \begin{cases} 1-\left(1+\gamma\dfrac{y}{\beta}\right)^{-1/\gamma} & si\ \gamma \neq 0\ et\ \beta > 0 \\ 1-\exp\left(-\dfrac{y}{\beta}\right) & si\ \gamma = 0\ et\ \beta > 0 \end{cases}$$

avec $y \in \mathbb{R}_+$ *si* $\gamma \geq 0$ *ou* $[0, -\beta/\gamma[$ *si* $\gamma < 0$.

Dans l'expression précédente, β représente le paramètre d'échelle et γ le paramètre de forme : il s'agit du même paramètre de forme introduit dans la partie 1.2 et que l'on appelle indice des valeurs extrêmes.

La loi GPD présente quelques particularités. En voici une liste non exhaustive :

- Si $\beta = 1$, on parle la loi GPD standard.
- Si $\gamma = 0$, la GPD correspond à une loi exponentielle d'espérance β.
- Si $\gamma = -1$, elle correspond à une loi uniforme sur $[0, \beta]$.
- Si $\gamma > 0$, on retrouve la loi de Pareto décentrée.

Dans cette approche d'estimation des quantiles extrêmes, on ne retient que les observations dépassant un seuil fixé $u < x_F$. On définit alors l'excès Y de la variable X au dessus du seuil u par $X - u$ sachant $X > u$. Si l'on note par F_u la fonction de répartition d'un excès au dessus du seuil u, on a pour tout $y > 0$

$$\begin{aligned} 1 - F_u(y) &= \mathbb{P}(Y > y) \\ &= \mathbb{P}(X - u > y | X > u) \\ &= \frac{\mathbb{P}(X > u + y, X > u)}{\mathbb{P}(X > u)} \\ &= \frac{1 - F(u+y)}{1 - F(u)}. \end{aligned}$$

Lorsque le seuil u est grand, on peut approcher cette quantité par la fonction de survie d'une loi GPD. Afin d'approcher le quantile, il suffit alors d'utiliser le résultat de Balkema et de Haan (1974) et Pickands (1975) qui établit l'équivalence entre la convergence en loi du maximum vers une loi des valeurs extrêmes $\mathscr{H}_\gamma$ et la convergence en loi d'un excès vers une GPD. Ce résultat s'énonce comme suit.

Théorème 1.4.3 (Balkema et de Haan (1974); Pickands (1975)). *Si F appartient au domaine d'attraction de $\mathscr{H}_\gamma$, alors*

$$\lim_{u \to x_F} \sup_{y \in]0, x_F - u[} \left| F_u(y) - \mathscr{G}_{\gamma,\beta}(y) \right| = 0.$$

D'après ce résultat, si pour une fonction de répartition F inconnue, l'échantillon des maxima normalisés converge en loi vers une distribution non dégénérée, alors il s'en

déduit que la distribution des excès au-dessus d'un seuil élevé converge vers une GPD lorsque le seuil tend vers la limite supérieure du support de F. Cette caractérisation est à la base des méthodes d'estimation de type *Peaks Over Threshold* **(POT)**.

Comme $1-F(u+y) = [1-F(u)]\left[1-F_u(y)\right]$, si pour tout $y \geq 0$ on pose $q_{\alpha_n} = u+y$, alors

$$\begin{aligned} \alpha_n = 1-F(q_{\alpha_n}) &= [1-F(u)]\left[1-F_u(q_{\alpha_n}-u)\right] \\ &\simeq [1-F(u)]\left(1-\mathscr{G}_{\gamma,\beta}(q_{\alpha_n}-u)\right). \end{aligned}$$

Pour k_n excès au-dessus du seuil u, l'approximation $1-F(u) \simeq k_n/n$ conduit à

$$\frac{k_n}{n}\left(1-\mathscr{G}_{\gamma,\beta}(q_{\alpha_n}-u)\right) \simeq \alpha_n,$$

et si $\gamma \neq 0$, alors on approche le quantile par

$$q_{\alpha_n} \simeq u + \frac{\beta}{\gamma}\left(\left(\frac{k_n}{n\alpha_n}\right)^{\gamma} - 1\right).$$

On a alors un estimateur de type

$$\hat{q}_{\alpha_n} = \frac{\left(\frac{k_n}{n\alpha}\right)^{\hat{\gamma}} - 1}{\hat{\gamma}}\hat{\beta} + u, \tag{1.5}$$

où $\hat{\gamma}$ et $\hat{\beta}$ sont respectivement des estimateurs des paramètres de forme et d'échelle. On peut noter la similitude entre l'estimateur de quantile (1.5) et l'expression du quantile (1.3) avec $\hat{\beta} = \hat{a}_n$ et $u = \hat{b}_n$.

Les paramètres γ et β de la GPD peuvent être estimés par la méthode des moments, la méthode des moments pondérées (Hosking et Wallis, 1987) ou la méthode du maximum de vraisemblance (Smith, 1987; Davison et Smith, 1990).

Cette méthode présente un avantage par rapport à la précédente en ce sens qu'il est plus facile d'avoir un échantillon d'excès que de maxima. Dans la pratique, on remplace u par $X_{n-k_n+1,n}$ c'est-à-dire la k_n plus grande observation de l'échantillon $\{X_i, i = 1,\ldots,n\}$.

Deux variantes de cette méthode ont été présentées par Breiman *et al.* (1990) sous les appellations *Exponential tail* **(ET)** et *Quadratique tail* **(QT)**.

1.4.4 Approche des quantiles extrêmes par l'approche semi-paramétrique

On se restreint aux fonctions $F \in \mathscr{D}(Fréchet)$ pour lesquelles on a la caractérisation suivante

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

avec ℓ une fonction à variations lentes à l'infini et $\gamma > 0$. Conformément au Lemme 1.3.2, cette caractérisation implique que

$$q_{\alpha_n} := \bar{F}^{\leftarrow}(\alpha_n) = \alpha_n^{-\gamma} L(1/\alpha_n) \text{ avec } \alpha_n \leq 1/n,$$
$$q_{\beta_n} := \bar{F}^{\leftarrow}(\beta_n) = \beta_n^{-\gamma} L(1/\beta_n) \text{ avec } \beta_n \geq 1/n,$$

où L est une fonction à variations lentes à l'infini. En ce qui concerne les fonctions L et ℓ, il apparait important de signaler ici qu'il ne s'agit pas de la même fonction à variations lentes.

Vu la définition d'une fonction à variations lentes (se référer à la Définition 1.3.1), pour β_n suffisamment petit, on a

$$\bar{F}^{\leftarrow}(\alpha_n) \simeq \bar{F}^{\leftarrow}(\beta_n) \left(\frac{\beta_n}{\alpha_n}\right)^{\gamma}.$$

En remplaçant $\bar{F}^{\leftarrow}(\beta_n)$ et γ par des estimateurs, on obtient l'estimateur de Weissman (1978) défini par

$$\hat{q}^{\mathrm{W}}_{\alpha_n} = X_{n-\lfloor n\beta_n \rfloor+1,n} \left(\frac{\beta_n}{\alpha_n}\right)^{\hat{\gamma}}.$$

Pour les propriétés de l'estimateur de Weissman, on peut se référer à l'ouvrage de Embrechts *et al.* (1997).

Comme autre estimateur des quantiles extrêmes, on peut citer celui obtenu par l'approximation

$$\bar{F}^{\leftarrow}(\alpha_n) \simeq \frac{(\beta_n/\alpha_n)^{\gamma} - 1}{1 - 2^{-\gamma}} \left(\bar{F}^{\leftarrow}(\beta_n) - \bar{F}^{\leftarrow}(2\beta_n)\right) + \bar{F}^{\leftarrow}(\beta_n),$$

et valable quel que soit le domaine d'attraction de la fonction F. La normalité asymptotique de l'estimateur de quantile extrême qui en découle, i.e

$$\hat{q}^{\mathrm{DH}}_{\alpha_n} = \frac{(\beta_n/\alpha_n)^{\hat{\gamma}} - 1}{1 - 2^{-\hat{\gamma}}} \left(X_{n-\lfloor n\beta \rfloor+1,n} - X_{n-\lfloor 2n\beta_n \rfloor+1,n}\right) + X_{n-\lfloor n\beta_n \rfloor+1,n},$$

a été établie par Dekkers et de Haan (1989). Il apparait clairement que cet estimateur de quantile extrême peut se mettre sous la forme (1.5) et donc (1.3) avec $\hat{a}_n = \dfrac{\hat{\gamma}}{1 - 2^{-\hat{\gamma}}} \left(X_{n-\lfloor n\beta_n \rfloor+1,n} - X_{n-\lfloor 2n\beta_n \rfloor+1,n}\right)$ et $\hat{b}_n = X_{n-\lfloor n\beta_n \rfloor+1,n}$.

1.5 Estimation du paramètre de la loi des valeurs extrêmes

Dans la littérature de la théorie des valeurs extrêmes on trouve plusieurs techniques semi-paramétriques pour l'estimation de l'indice de queue. On peut citer l'estimateur de Hill (1975) valable pour $\gamma > 0$. Il est considéré comme le plus simple des estimateurs de l'indice de queue. Pour pallier les limitations de l'estimateur Hill, mais aussi pour

l'étendre aux deux autres domaines d'attractions Dekkers *et al.* (1989) ont proposé un estimateur des moments valable quel que soit $\gamma \in \mathbb{R}$. Dans la littérature, certains auteurs l'appellent plutôt l'estimateur de Dekkers-Einmahl-de Haan. Les auteurs ont établi la consistance forte, faible et la normalité asymptotique de leur estimateur. Soulignons que la méthode des moments fut initialement utilisée par Hosking et Wallis (1987) qui proposa un estimateur des moments pondérés définit pour tout $\gamma < 1$. Cet estimateur est consistant si $\gamma < 1$ et asymptotiquement gaussien pour $\gamma < 1/2$. Toutefois, Pickands (1975) fût le premier à proposer un estimateur de l'indice de queue plus général que l'estimateur de Hill, i.e valable quel que soit le signe de γ. Smith (1985) s'est quant à lui intéressé au comportement des estimateurs du maximum de vraisemblance [1] dans le cas d'une loi GEV [2]. Il a été démontré que l'estimateur du maximum de vraisemblance est consistant, asymptotiquement efficace et asymptotiquement normal pour tout $\gamma > -1/2$. Falk (1995) a proposé un complément à l'estimateur du maximum de vraisemblance. Son estimateur, que l'on appelle aussi l'estimateur négatif de Hill (en anglais *The Negative Hill Estimator*), est consistant si $\gamma < -1/2$ et asymptotiquement normal si $-1 < \gamma < -1/2$.

Comme autres estimateurs, on peut citer l'estimateur du rapport des moments (Danielsson *et al.*, 1996), l'estimateur de Peng (1998), l'estimateur basé sur le QQ-plot, l'estimateur basé sur le graphique de la moyenne des excès (Beirlant *et al.*, 1996), l'estimateur construit par des méthodes de régression (Beirlant *et al.*, 2002), d'optimisation avec contrainte (Csörgö *et al.*, 1985; Bacro et Brito, 1994) ou sans contrainte (Schultze et Steinebach, 1996; Kratz et Resnick, 1996) et d'autres (voir Embrechts *et al.*, 1997; Beirlant *et al.*, 2004b; de Haan et Ferreira, 2006).

À côté de cette énumération, on pourrait aussi ajouter les méthodes de correction du biais basées sur des méthodes de régression (Beirlant *et al.*, 1999, 2005; Diebolt *et al.*, 2008) ou sur des méthodes de bootstrap (Gomes et Oliveira, 2001) ou jackknife (Gomes, 1999; Gomes *et al.*, 2005b) et les estimateurs à poids ou combinaisons linéaires (Viharos, 1993, 1995; Gomes *et al.*, 2005a) qui englobent d'une part les estimateurs construits par des méthodes d'optimisation avec et sans contrainte et d'autres part les estimateurs construits par des méthodes de régression.

D'un point de vue théorique, toutes ces méthodes partagent les mêmes propriétés de consistance et de normalité asymptotique. Cependant, les simulations montrent qu'il y a de grandes différences entre ces différents estimateurs. En général, il n'y a pas une meilleure méthode dans toutes les situations. Les méthodes les plus utilisées sont celles de Hill, Pickands et des moments. Certaines études de comparaison (théorique et par simulation) entre les différentes méthodes peuvent être trouvées dans Rosen et Weissman (1996), Peng (1998), de Haan et Peng (1998), Groeneboom *et al.* (2003) et Tsourti et Panaretos (2001, 2003).

1. D'après la théorie du Maximum de Vraisemblance, le support de la loi ne dépend pas des paramètres.
2. Dans le cas des lois EVD et GPD, le support dépend des paramètres sauf dans le cas particulier ou $\gamma = 0$.

Tsourti et Panaretos (2001, 2003) pensent que la performance d'une méthode dépend de la distribution de la série étudiée. En d'autres termes, elle dépend de la vraie valeur de l'indice de queue. Ils recommandent l'utilisation de techniques pour déterminer le domaine d'attraction de la loi des valeurs extrêmes, et donc l'intervalle le plus probable pour l'indice des valeurs extrêmes. Les méthodes les plus utilisées à cette fin sont graphiques : le graphique loglog, la moyenne empirique des excès, le graphique des rapports du maximum et de la somme, le graphique du rapport de Hill, la statistique de Jackson. Pour de plus amples explications sur ces méthodes voir El-Adlouni *et al.* (2007).

Dans ce paragraphe, nous exposerons uniquement trois estimateurs de l'indice de queue. Notre choix s'explique par notre volonté d'introduire une adaptation des estimateurs ainsi proposés au cas conditionnel (voir chapitres 4 et 5).

1.5.1 Estimateur de Hill

Cet estimateur a été introduit par Hill (1975) pour estimer d'une manière non-paramétrique le paramètre de queue des lois appartenant au $\mathscr{D}(Fréchet)$. Pour construire son estimateur, Hill utilise la méthode du maximum de vraisemblance sur l'ensemble des k_n plus grandes observations d'un échantillon. Un grand nombre de travaux théoriques ont été consacrés à l'étude des propriétés de l'estimateur de Hill. Mason (1982) a démontré la consistence faible et Deheuvels, Haeusler et Mason ont établi la consistance forte dans Deheuvels *et al.* (1988). La normalité asymptotique est due entre autres à Davis et Resnick (1984), Csörgö et Mason (1985), Haeusler et Teugels (1985) et Smith (1987).

Définition 1.5.1. *Soit $(k_n)_{n\geq 1}$ une suite d'entiers avec $1 < k_n \leq n$, l'estimateur de Hill est défini par*

$$\hat{\gamma}^{\mathrm{H}}_{k_n} = \frac{1}{k_n - 1} \sum_{i=1}^{k_n-1} \log X_{n-i+1,n} - \log X_{n-k_n+1,n}.$$

Théorème 1.5.1 (Propriétés de l'estimateur de Hill). *Soit $(k_n)_{n\geq 1}$ une suite d'entiers telle que $1 < k_n \leq n$, $k_n \to \infty$ et $k_n/n \to 0$ quand $n \to \infty$.*

- *Alors, $\hat{\gamma}^{\mathrm{H}}_{k_n}$ converge en probabilité vers γ.*
- *Si de plus $k_n/\log\log n \to \infty$ quand $n \to \infty$, alors $\hat{\gamma}^{\mathrm{H}}_{k_n}$ converge presque sûrement vers γ.*

Pour établir la normalité asymptotique de l'estimateur $\hat{\gamma}^{\mathrm{H}}_{k_n}$, on a besoin d'une hypothèse sur la fonction à variations lentes ℓ. Il est en effet nécessaire d'imposer une condition qui spécifie la vitesse de convergence du rapport des fonctions à variations lentes vers 1 telle que défini au paragraphe 1.3.1.

(C.1) : *Il existe une constante réelle* $\rho < 0$ *et une fonction* $\varepsilon(x) \to 0$ *quand* $x \to \infty$, *telles que pour tout* $\lambda > 1$

$$\log \frac{\ell(\lambda x)}{\ell(x)} \sim \varepsilon(x) \frac{\lambda^{\rho} - 1}{\rho} \text{ quand } x \to \infty.$$

Cette condition appelée « *condition du second ordre* » est satisfaite pour la plupart des lois appartenant au $\mathscr{D}(Fréchet)$. Plus la constante $\rho < 0$ de la condition **(C.1)** est proche de zéro, plus difficile est l'estimation de l'indice de queue γ.

Remarque 1.5.1. *La condition* **(C.1)** *implique que* $\forall\ \varepsilon > 0, \exists\ x_0$ *tel que* $\forall\ x \geq x_0, \forall\ \lambda > 1$

$$\frac{(1+\varepsilon)\,\lambda^{\rho+\varepsilon} - 1}{\rho} \leq \frac{1}{\varepsilon(x_n)} \log\left(\frac{\ell(\lambda x)}{\ell(x)}\right) \leq \frac{(1-\varepsilon)\,\lambda^{\rho-\varepsilon} - 1}{\rho}.$$

Théorème 1.5.2 (**Normalité asymptotique de l'estimateur de Hill).** *s Soit* $(k_n)_{n\geq 1}$ *une suite d'entiers telle que* $1 < k_n \leq n$, $k_n \to \infty$ *et* $k_n/n \to 0$. *Si la condition* **(C.1)** *est satisfaite avec* $\sqrt{k_n}\varepsilon(n/k_n) \to 0$ *quand* $n \to \infty$, *alors*

$$\sqrt{k_n}\left(\hat{\gamma}^{\mathrm{H}}_{k_n} - \gamma\right) \xrightarrow{\mathscr{L}} \mathscr{N}\left(0, \gamma^2\right).$$

En pratique, le choix du paramètre k_n pose des problèmes. Si l'on trace le diagramme de Hill (voir Figure 1.3), i.e la fonction $k_n \mapsto \hat{\gamma}^{\mathrm{H}}_{k_n}$, on observe une extrême volatilité qui rend difficile l'utilisation de cet estimateur en pratique si l'on n'a aucune indication sur le choix de k_n. De plus, cet estimateur est biaisé (voir Figure 1.4). Ce biais est de l'ordre de $\varepsilon(n/k_n)$. La condition $\sqrt{k_n}\varepsilon(n/k_n) \to 0$ impose au bais d'être négligeable devant l'écart type de l'estimateur qui est quand à lui égal à $\sqrt{k_n}$ .Une minimisation de l'erreur en moyenne quadratique peut-être utilisée comme critère. Cette méthode reste néanmoins inutilisable en pratique puisque l'erreur en moyenne quadratique reste inconnue ! Se référer à Beirlant *et al.* (1996) et Drees et Kaufmann (1998) pour des exemples de sélection du paramètre k_n.

Le résultat sur la normalité asymptotique de l'estimation de Hill permet de donner un intervalle de confiance pour l'estimation. En pratique on se contentera de remplacer γ par sa valeur estimée. Par conséquent, si k_n est petit, on aura *a fortiori*, compte tenu des remarques faites précédemment, une estimation avec un intervalle de confiance large et *a contrario*, si k_n est grand, on aura un intervalle de confiance plus étroit mais pas centrée sur la vraie valeur.

1.5.2 Estimateur de Pickands

L'estimateur de Pickands est construit en utilisant trois statistique d'ordres. Cet estimateur a l'avantage d'être valable quel que soit le domaine d'attraction de la distribution et par conséquent, du domaine de définition de l'indice des valeurs extrêmes. Pickands (1975) démontre la consistance faible de son estimateur. La convergence forte ainsi que la normalité asymptotique ont été démontrées par Dekkers et de Haan (1989).

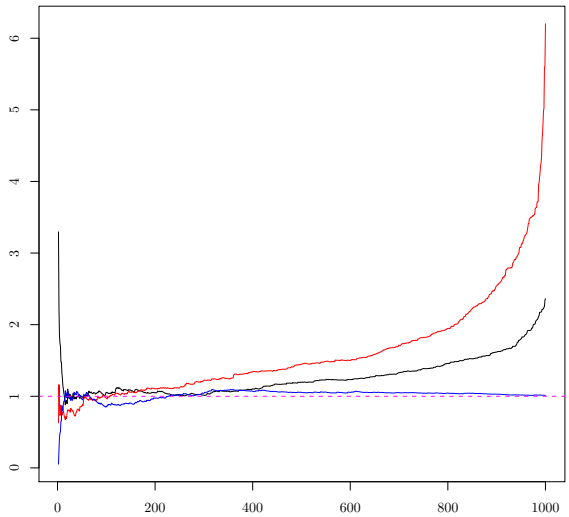


FIGURE 1.3 – Graphique de Hill à $n = 1000$ et trois lois appartenant au domaine d'attraction de Fréchet. La loi de Pareto (bleu), Burr (rouge), Fréchet (noir) et la vraie valeur de $\gamma = 1$ en trait interrompu. En ordonnée on a l'indice de queue estimé et en abscisse le seuil k_n. Pour une loi de Pareto, comme la fonction à variations lentes ℓ est une constante alors, on peut prendre k_n aussi grand que l'on veut.

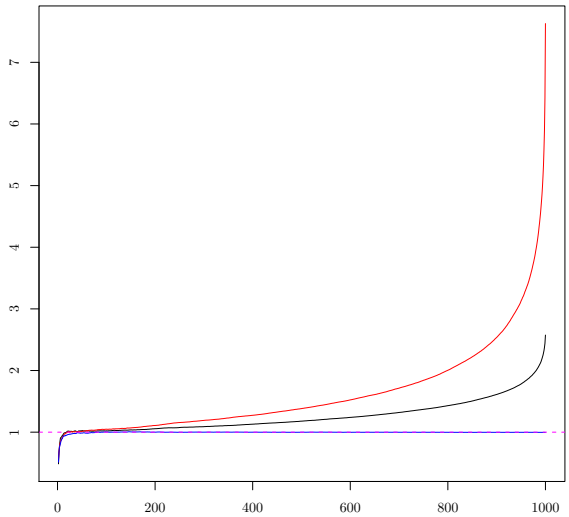


FIGURE 1.4 – Graphique de Hill en moyenne sur 100 réalisations à $n = 1000$ et trois lois appartenant au domaine d'attraction de Fréchet. La loi de Pareto (bleu), Burr (rouge), Fréchet (noir) et la vraie valeur de $\gamma = 1$ en trait interrompu. En ordonnée on a l'indice de queue estimé et en abscisse le seuil k_n. Pour une loi de Pareto comme $\varepsilon = 0$ alors, il n'y a pas de biais asymptotique.

Définition 1.5.2. *On suppose que* $\{X_i,\ i = 1,\dots,n\}$ *est une suite de variables aléatoires indépendantes de loi* F *appartenant à l'un des domaines d'attractions. Soit* $(k_n)_{n\geq 1}$ *une suite d'entiers avec* $1 \leq k_n < n$*, l'estimateur de Pickands est défini par*

$$\hat{\gamma}^{\mathrm{P}}_{k_n} = \frac{1}{\log 2}\log\left(\frac{X_{n-k_n+1,n} - X_{n-2k_n+1,n}}{X_{n-2k_n+1,n} - X_{n-4k_n+1,n}}\right).$$

Théorème 1.5.3 (Propriétés de l'estimateur de Pickands). *Soit* $(k_n)_{n\geq 1}$ *une suite d'entiers telle que* $1 \leq k_n < n$*,* $k_n \to \infty$ *et* $k_n/n \to 0$ *quand* $n \to \infty$*.*

- *Alors,* $\hat{\gamma}^{\mathrm{P}}_{k_n}$ *converge en probabilité vers* γ*.*
- *Si de plus* $k_n/\log\log n \to \infty$ *quand* $n \to \infty$*, alors* $\hat{\gamma}^{\mathrm{P}}_{k_n}$ *converge presque sûrement vers* γ*.*
- *Sous des conditions additionnelles sur la suite* k_n *et la fonction de répartition* F *que l'on pourra consulter dans Dekkers et de Haan (1989),*

$$\sqrt{k_n}\left(\hat{\gamma}^{\mathrm{P}}_{k_n} - \gamma\right) \xrightarrow{\mathscr{L}} \mathcal{N}\left(0, \frac{\gamma^2(2^{2\gamma+1}+1)}{4(\log 2)^2(2^{\gamma}-1)^2}\right).$$

Comme l'estimateur de Hill, cet estimateur est biaisé et le résultat sur sa normalité asymptotique permet de donner un intervalle de confiance pour l'estimation.

Compte tenu de la variance asymptotique de $\hat{\gamma}^{\mathrm{P}}_{k_n}$ qui est assez importante comparativement à $\hat{\gamma}^{\mathrm{H}}_{k_n}$ (voir Fig 1.5), certains auteurs ont proposés des estimateurs à « *variance minimale* » construits à partir de combinaisons linéaires des logarithmes des accroissements des statistiques d'ordres. Par exemple Drees (1995) propose de faire la moyenne de plusieurs estimateurs de Pickands utilisant un nombre de plus grandes observations différents dans le but d'obtenir un estimateur moins sensible au choix de k_n .

1.5.3 Estimateur de Zipf

Dans le but d'améliorer le biais asymptotique des estimateurs précédents, Kratz et Resnick (1996) et Schultze et Steinebach (1996) ont indépendamment proposé d'estimer l'indice de queue par la méthode des moindres carrés classique. Leur estimateur connu sous le nom de *Zipf* est asymptotiquement gaussien. Il est défini par

$$\hat{\gamma}^{\mathrm{Z}}_{k_n} = \frac{\frac{1}{k_n}\sum_{j=1}^{k_n}\log\frac{k_n+1}{j}\log X_{n-j+1,n} - \frac{1}{k_n}\sum_{j=1}^{k_n}\log\frac{k_n+1}{j}\left(\frac{1}{k_n}\sum_{j=1}^{k_n}\log X_{n-i+1,n}\right)}{\frac{1}{k_n}\sum_{j=1}^{k_n}\left(\log\frac{k_n+1}{j}\right)^2 - \left(\frac{1}{k_n}\sum_{j=1}^{k_n}\log\frac{k_n+1}{j}\right)^2}. \tag{1.6}$$

Toutefois, sa variance asymptotique deux fois supérieure à celle de l'estimateur de Hill.

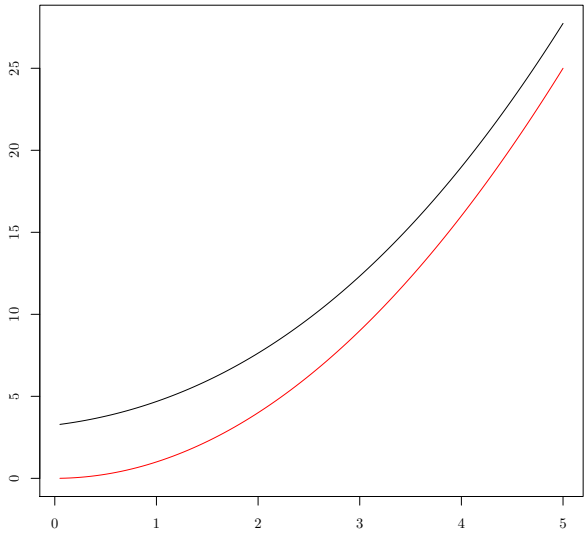


FIGURE 1.5 – La variance asymptotique de l'estimateur $\hat{\gamma}^{\mathrm{P}}_{k_n}$ (noir) et de $\hat{\gamma}^{\mathrm{H}}_{k_n}$ (rouge) en fonction de γ.

2

Un nouvel estimateur des quantiles extrêmes non conditionnels

Résumé

Dans ce chapitre, nous nous proposons d'étudier un nouvel estimateur des quantiles extrêmes dans le cas non conditionnel. Nous établissons sa loi limite et nous le comparons à quelques estimateurs de quantiles existants.

Sommaire

2.1 Introduction 31
2.2 Résultats asymptotiques 32
2.3 Comparaison graphique des estimateurs 34
2.4 Démonstrations 39

2.1 Introduction

Soit $\{X_i,\ i = 1,\dots,n\}$ un échantillon de n variables aléatoires réelles indépendantes et identiquement distribuées de fonction de répartition F. On dénote par $X_{1,n} \leq \dots \leq X_{n,n}$ les statistiques ordonnées associées aux observations $\{X_i,\ i = 1,\dots,n\}$ et on se propose d'estimer le quantile d'ordre α_n lorsque $F \in \mathscr{D}(Fréchet)$.

Pour ce faire, on se propose de faire la moyenne géométrique des estimateurs de Weissman (1978). Ainsi que nous l'avons souligné au dernier paragraphe du chapitre 1, l'idée de faire la moyenne de plusieurs estimateurs a déjà été appliquée par Drees (1995) qui proposait alors de faire la moyenne d'estimateurs de Pickands (1975) pour plusieurs valeurs de k_n dans le but d'obtenir un estimateur moins sensible au choix du seuil.

Définition 2.1.1. *Soit $(k_n)_{n\geq 1}$ une suite d'entiers telle que $1 \leq k_n < n$, on définit un estimateur des quantiles extrêmes par*

$$\hat{q}_{\alpha_n}^{\mathrm{WG}} = \left[\prod_{i=1}^{k_n} X_{n-i+1,n} \left(\frac{i g_{k_n}}{n\alpha_n} \right)^{\hat{\gamma}_i^{\mathrm{H}}} \right]^{1/k_n}, \tag{2.1}$$

où $g_{k_n} = \exp\left[\log(k_n+1) - 1 - \log(k_n!)/k_n\right]$ et $\hat{\gamma}_i^{\mathrm{H}}$ est l'estimateur de Hill (1975) donné par

$$\hat{\gamma}_i^{\mathrm{H}} = \frac{1}{i} \sum_{j=1}^{i} j \left\{ \log X_{n-j+1,n} - \log X_{n-j,n} \right\}. \tag{2.2}$$

En ce qui concerne l'estimateur de Hill (1975), il n'y aucune différence entre (2.2) et la Définition 1.5.1. La forme de l'estimateur (2.2) introduite par Beirlant *et al.* (2002) repose sur le modèle de régression exponentiel pour des écarts de logarithmes entre les i plus grandes statistiques d'ordre d'un échantillon de variables aléatoires indépendantes et de loi $F \in \mathscr{D}(Fréchet)$. Ainsi, on montre facilement que

$$\hat{\gamma}_i^{\mathrm{H}} = \frac{1}{i} \sum_{j=1}^{i} j \left\{ \log X_{n-j+1,n} - \log X_{n-j,n} \right\} = \frac{1}{i} \sum_{j=1}^{i} \log X_{n-j+1,n} - \log X_{n-i,n}.$$

2.2 Résultats asymptotiques

Dans cette partie, nous étudions le comportement asymptotique de notre nouvel estimateur de quantile extrême. Nous commençons par présenter les résultats auxiliaires permettant d'en déduire sa normalité asymptotique. Ainsi, notre premier résultat auxiliaire est dédié à la représentation en loi de l'estimateur $\hat{q}_{\alpha_n}^{\mathrm{WG}}$. Dans tout ce qui suit, on pose $\Delta_j = j\left\{\log X_{n-j+1,n} - \log X_{n-j,n}\right\}$.

Proposition 2.2.1. *On a la décomposition suivante :*

$$\log \hat{q}_{\alpha_n}^{\mathrm{WG}} \stackrel{\mathscr{L}}{=} \hat{\gamma}_{k_n}^{\mathrm{H}} - \gamma \log V_{k_n+1,n} + \log L\left(1/V_{k_n+1,n}\right) + \log\left(\frac{1}{\mathrm{e}} \frac{(k_n+1)}{n\alpha_n} \right) \hat{\gamma}_{k_n}^{\pi}, \tag{2.3}$$

avec $\hat{\gamma}_{k_n}^{\mathrm{H}}$ l'estimateur de Hill défini précédemment, L une fonction à variations lente à l'infini, $V_{k_n+1,n}$ la $(n-k_n)$-ième plus grande statistique d'ordre d'un échantillon de variables aléatoires indépendantes de loi uniforme standard $\{V_i,\ i = 1,\dots,n\}$ et

$$\hat{\gamma}_{k_n}^{\pi} = \sum_{j=1}^{k_n} \Delta_j \pi_j \Bigg/ \sum_{j=1}^{k_n} \pi_j \quad \textit{avec}\ \pi_j = \sum_{i=j}^{k_n} \frac{1}{i} \log\left(\frac{i g_{k_n}}{n\alpha_n} \right).$$

En utilisant le schéma de la preuve de la Proposition 2.2.1 (voir partie 2.4), on montre que la représentation en loi de l'estimateur de Weissman (1978) défini comme $\hat{q}_{\alpha_n}^{\mathrm{W}} = X_{n-k_n+1,n} \left(\frac{k_n}{n\alpha_n} \right)^{\hat{\gamma}_{k_n}^{\mathrm{H}}}$ est donnée par

$$\log \hat{q}_{\alpha_n}^{\mathrm{W}} \stackrel{\mathscr{L}}{=} -\gamma \log V_{k_n,n} + \log L\left(1/V_{k_n,n}\right) + \log\left(\frac{k_n}{n\alpha_n} \right) \hat{\gamma}_{k_n}^{\mathrm{H}}, \tag{2.4}$$

où $V_{k_n,n}$ est la $(n-k_n+1)$-ième plus grande statistique d'ordre d'un échantillon de variables aléatoires indépendantes de loi uniforme standard $\{V_i,\ i=1,\dots,n\}$.

En comparant (2.3) et (2.4), on remarque que la représentation en loi de $\hat{q}^{\mathrm{WG}}_{\alpha_n}$ fait intervenir non seulement un terme supplémentaire mais aussi un nouvel estimateur de queue noté $\hat{\gamma}^{\pi}_{k_n}$. Cet estimateur de queue est une somme pondérée des écarts de logarithmes entre les k_n plus grandes observations de l'échantillon $\{X_i,\ i=1,\dots,n\}$. Grâce aux travaux de Beirlant *et al.* (2002), on sait étudier un tel estimateur. D'après les auteurs, si l'on montre que les poids $\{\pi_j, j=1,\dots,k_n\}$ satisfont certaines conditions énoncées dans leur contribution scientifique (voir annexe A) alors, $\hat{\gamma}^{\pi}_{k_n}$ est asymptotiquement gaussien. Ainsi, les deux lemmes suivants sont des outils dont le but est de montrer que nos poids satisfont les conditions requises.

Lemme 2.2.1. *Soit $(\alpha_n)_{n\geq 1}$ une suite telle que $n\alpha_n \to 0$. Si $k_n \to \infty$ et $\log(k_n)/\log(n\alpha_n) \to 0$ quand $n\to\infty$ alors, pour tout $j=1,\dots,k_n$,*

$$\pi_j = \tau_j \log\left(\frac{1}{n\alpha_n}\right)(1+o(1)) \text{ avec } \tau_j = \sum_{i=j}^{k_n} 1/i,$$

où le $o(1)$ est uniforme en j.

D'après le Lemme 2.2.1, la loi asymptotique de $\hat{\gamma}^{\pi}_{k_n}$ ne dépend plus que du comportement des poids $\{\tau_j, j=1,\dots,k_n\}$. Le résultat suivant est dédié à l'étude de τ_j.

Lemme 2.2.2. *Pour tout $j=1,\dots,k_n$,*

$$\tau_j = W^{\tau}(j/(k_n+1))(1+o(1)) \text{ avec } W^{\tau}(s) = -\log s \text{ et } s>0,$$

uniformément en $j=1,\dots,k_n$.

L'estimateur de Zipf noté $\hat{\gamma}^{\mathrm{Z}}_{k_n}$ (voir paragraphe 1.5.3) s'interprétant comme un estimateur de la pente dans un graphe de coordonnées

$$\left(\log\left(\frac{n+1}{j}\right), \log X_{n-j+1,n}\right)_{\{j=1,\dots,n\}},$$

alors il peut se réécrire $\hat{\gamma}^{\mathrm{Z}}_{k_n} = \sum_{j=1}^{k_n}\Delta_j\mu_j \Big/ \sum_{j=1}^{k_n}\mu_j$ avec $\mu_j = -\log\left(\frac{j}{k_n+1}\right)(1+o(1))$ uniformément en $j=1,\dots,k_n$. Par conséquent, $\hat{\gamma}^{\mathrm{Z}}_{k_n}$ et $\hat{\gamma}^{\pi}_{k_n}$ sont très proches. En particulier, ils ont même loi limite. La proposition suivante établit la normalité asymptotique de $\hat{\gamma}^{\pi}_{k_n}$.

Proposition 2.2.2. *Si l'hypothèse* **(C.1)** *est vérifiée et si de plus la suite $(k_n)_{n\geq 1}$ satisfait $k_n\to\infty$, $\log(k_n)/\log(n\alpha_n)\to 0$ et $k_n^{1/2}\varepsilon(n/k_n)\to\lambda\in\mathbb{R}$ quand $n\to\infty$, alors*

$$k_n^{1/2}\left(\hat{\gamma}^{\pi}_{k_n} - \gamma - \frac{\varepsilon(n/k_n)}{(1-\rho)^2}\right) \xrightarrow{\mathscr{L}} \mathcal{N}\left(0, 2\gamma^2\right).$$

Si $k_n \to \infty$, $k_n/n \to 0$ et $k_n^{1/2}\varepsilon(n/k_n) \to \lambda \in \mathbb{R}$ quand $n \to \infty$, alors on peut montrer que (se référer à la démonstration de la Proposition 2.2.2)

$$k_n^{1/2}\left(\hat{\gamma}_{k_n}^{\mathrm{H}} - \gamma - \frac{\varepsilon(n/k_n)}{(1-\rho)}\right) \xrightarrow{\mathscr{L}} \mathscr{N}\left(0, \gamma^2\right).$$

Comparativement, il apparaît donc que le biais asymptotique de l'estimateur $\hat{\gamma}_{k_n}^{\pi}$ est de meilleure qualité que celui de l'estimateur de $\hat{\gamma}_{k_n}^{\mathrm{H}}$ et *a contrario*, sa variance asymptotique est de moins bonne qualité. Ce problème est récurrent en modélisation des données ou en estimation. On observe fréquemment qu'une réduction significative de la variance peut conduire à une augmentation du biais et *vice-versa*, une réduction du biais initial peut mener à une augmentation nette de la variance : on parle alors *du dilemme biais / variance*.

Théorème 2.2.1. *Supposons la condition* **(C.1)** *satisfaite. Soit* $(\alpha_n)_{n\geq 1}$ *une suite telle que* $n\alpha_n \to 0$. *Si* $k_n \to \infty$, $\log(k_n)/\log(n\alpha_n) \to 0$ *et* $k_n^{1/2}\varepsilon(n/k_n) \to \lambda \in \mathbb{R}$ *quand* $n \to \infty$ *alors,*

$$\frac{k_n^{1/2}}{\log\left(\frac{k_n}{n\alpha_n}\right)}\left\{\log\left(\frac{\hat{q}_{\alpha_n}^{\mathrm{WG}}}{q_{\alpha_n}}\right) - \log\left(\frac{k_n}{n\alpha_n}\right)\frac{\varepsilon(n/k_n)}{(1-\rho)^2}\right\} \xrightarrow{\mathscr{L}} \mathscr{N}\left(0, 2\gamma^2\right).$$

Si $k_n \to \infty$, $k_n/n \to 0$ et $k_n^{1/2}\varepsilon(n/k_n) \to \lambda \in \mathbb{R}$ quand $n \to \infty$ alors, l'estimateur de Weissman (1978), dont la décomposition en loi est donnée en (2.4), est asymptotiquement gaussien (se référer à la technique de preuve utilisée pour le Théorème 2.2.1), i.e

$$\frac{k_n^{1/2}}{\log\left(\frac{k_n}{n\alpha_n}\right)}\left\{\log\left(\frac{\hat{q}_{\alpha_n}^{\mathrm{W}}}{q_{\alpha_n}}\right) - \log\left(\frac{k_n}{n\alpha_n}\right)\frac{\varepsilon(n/k_n)}{(1-\rho)}\right\} \xrightarrow{\mathscr{L}} \mathscr{N}\left(0, \gamma^2\right).$$

Les estimateurs de quantile extrême $\hat{q}_{\alpha_n}^{\mathrm{WG}}$ et $\hat{q}_{\alpha_n}^{\mathrm{W}}$ convergent à la même vitesse. Le biais asymptotique de l'estimateur $\hat{q}_{\alpha_n}^{\mathrm{WG}}$ est donnée par $\varepsilon(n/k_n)/(1-\rho)^2$. Il est, à un facteur d'échelle $1/(1-\rho)$ près, de meilleure qualité que celui de l'estimateur $\hat{q}_{\alpha_n}^{\mathrm{W}}$. Mais la variance asymptotique de $\hat{q}_{\alpha_n}^{\mathrm{W}}$ est, à un facteur d'échelle 2, de meilleure qualité que celle de $\hat{q}_{\alpha_n}^{\mathrm{WG}}$.

Notons que Fils et Guillou (2004) ont étudié un estimateur de quantile extrême basé sur l'estimateur des moindres carrés (1.6) et défini par

$$\hat{q}_{\alpha_n}^{\mathrm{FG}} = \alpha_n^{-\hat{\gamma}_{k_n}^{\mathrm{Z}}} \exp\left(\frac{1}{k_n}\sum_{j=1}^{k_n}\log X_{n-j+1,n} - \frac{\hat{\gamma}_{k_n}^{\mathrm{Z}}}{k_n}\sum_{j=1}^{k_n}\log\left(\frac{n+1}{j}\right)\right).$$

Sous les hypothèses de convergence énoncées précédemment, le résultat établi par les auteurs montrent que si $\lambda = 0$ alors $\hat{q}_{\alpha_n}^{\mathrm{WG}}$ et $\hat{q}_{\alpha_n}^{\mathrm{FG}}$ ont même loi asymptotique.

2.3 Comparaison graphique des estimateurs

Nous allons ici comparer les performances des estimateurs $\hat{q}_{\alpha_n}^{\mathrm{WG}}$, $\hat{q}_{\alpha_n}^{\mathrm{FG}}$ et $\hat{q}_{\alpha_n}^{\mathrm{W}}$ sur quatre lois. Pour une loi donnée, on simule $N = 1000$ échantillons de taille $n = 1000$. On

s'intéresse à l'estimation du quantile d'ordre $\alpha_n = \frac{1}{10n}$ et les lois considérées sont celles de Student à 10 degrés de liberté, Pareto, Burr et Fréchet.

Pour chaque loi, on compare la moyenne empirique des estimateurs du quantile extrême d'ordre α_n (voir graphes **(a)** et **(c)** aux Figures 2.1 et 2.2) et leur erreur relative quadratique moyenne[1] en fonction de k_n (voir Figure 2.3). Ensuite on représente les estimateurs médians[2] (voir graphes **(b)** et **(d)** aux Figures 2.1 et 2.2). Pour une bonne visualisation des graphiques, nous avons opté pour l'échelle logarithmique.

Ces simulations tentent de confirmer que, pour des lois dont $\varepsilon \neq 0$ (exception faite de la loi Pareto où $\varepsilon = 0$), notre nouvel estimateur de quantile extrême $\hat{q}^{\mathrm{WG}}_{\alpha_n}$ est moins biaisé. De plus, il est aussi lisse que $\hat{q}^{\mathrm{FG}}_{\alpha_n}$. Enfin, pour un échantillon de taille finie et du point de vue l'erreur relative quadratique, les simulations semblent montrer qu'à partir d'un certain seuil k_n pas trop grand, l'estimateur $\hat{q}^{\mathrm{WG}}_{\alpha_n}$ semble moins mauvais que $\hat{q}^{\mathrm{W}}_{\alpha_n}$ (voir Figure 2.3).

1. Elle est donnée par $\left(\frac{1}{N}\sum_{j=1}^{N}\frac{(q_{\alpha_n}-\hat{q}_{\alpha_n,j})^2}{q_{\alpha_n}}\right)^{1/2}$
2. Ce sont les estimateurs correspondant à la médiane des erreurs relatives quadratiques moyennes.

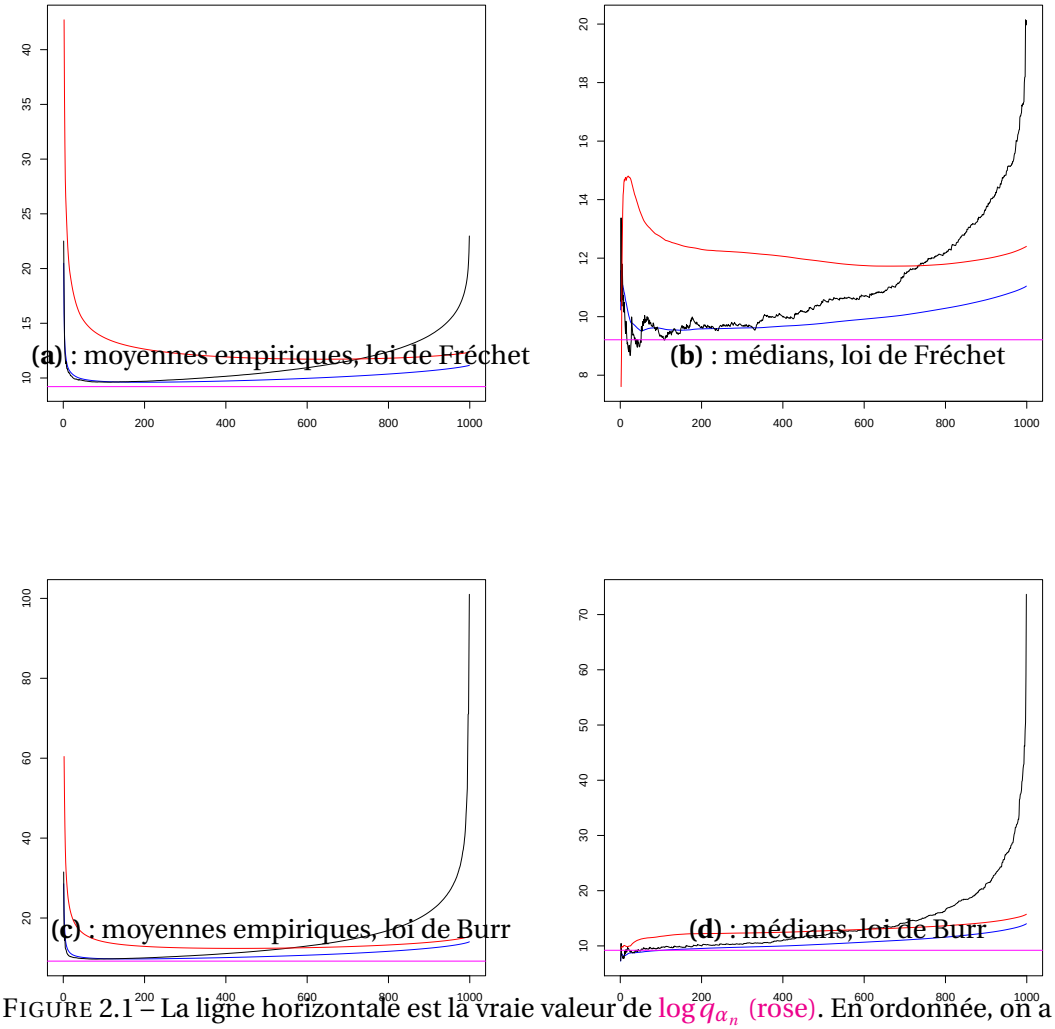


FIGURE 2.1 – La ligne horizontale est la vraie valeur de $\log q_{\alpha_n}$ (rose). En ordonnée, on a le quantile extrême estimé et en abscisse le seuil k_n.
(a) & **(c)** : Comparaison des moyennes empiriques des estimateurs $\log \hat{q}^{\mathrm{FG}}_{\alpha_n}$ (rouge), $\log \hat{q}^{\mathrm{W}}_{\alpha_n}$ (noir) et $\log \hat{q}^{\mathrm{WG}}_{\alpha_n}$ (bleu).
(b) & **(d)** : Comparaison des estimateurs médians de $\log \hat{q}^{\mathrm{FG}}_{\alpha_n}$ (rouge), $\log \hat{q}^{\mathrm{W}}_{\alpha_n}$ (noir) et $\log \hat{q}^{\mathrm{WG}}_{\alpha_n}$ (bleu).

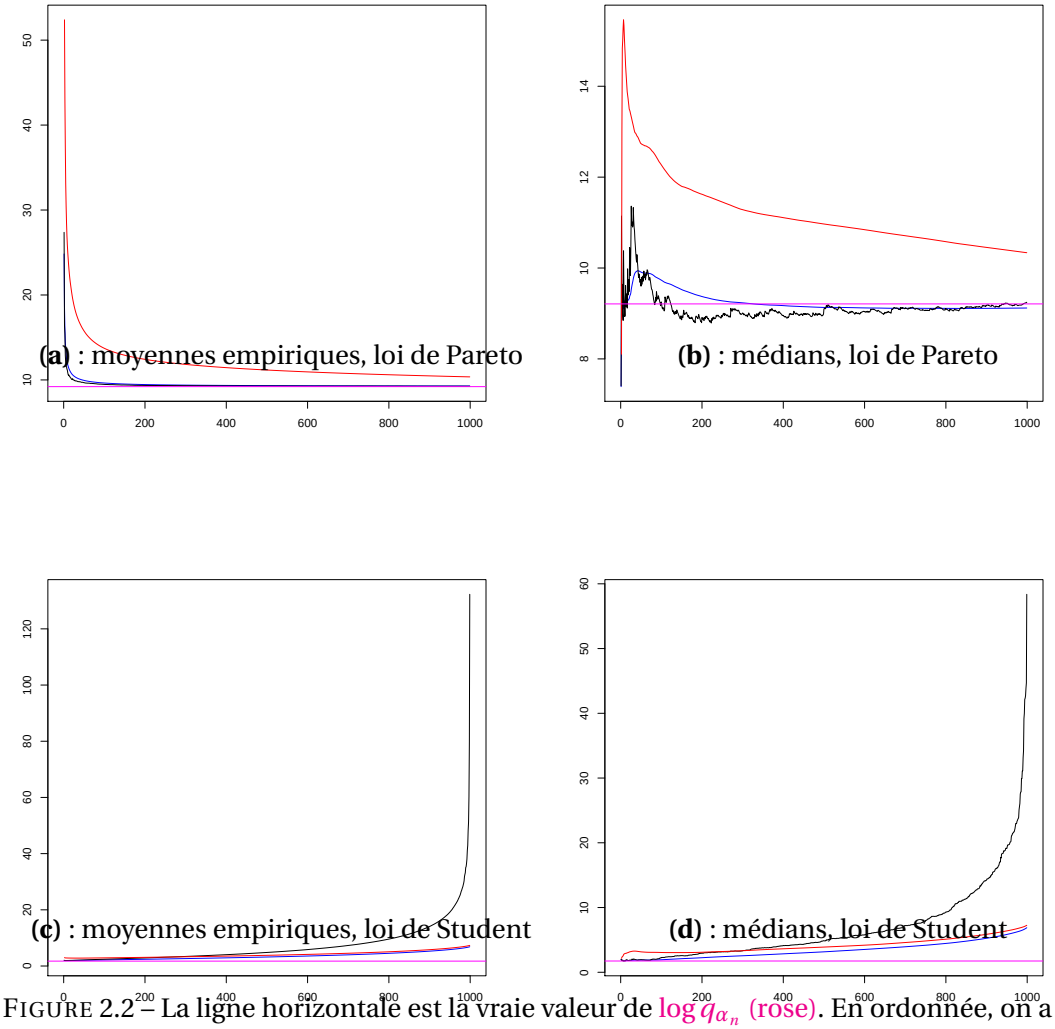


FIGURE 2.2 – La ligne horizontale est la vraie valeur de $\log q_{\alpha_n}$ (rose). En ordonnée, on a le quantile extrême estimé et en abscisse le seuil k_n.
(a) & **(c)** : Comparaison des moyennes empiriques des estimateurs $\log \hat{q}^{\mathrm{FG}}_{\alpha_n}$ (rouge), $\log \hat{q}^{\mathrm{W}}_{\alpha_n}$ (noir) et $\log \hat{q}^{\mathrm{WG}}_{\alpha_n}$ (bleu).
(b) & **(d)** : Comparaison des estimateurs médians de $\log \hat{q}^{\mathrm{FG}}_{\alpha_n}$ (rouge), $\log \hat{q}^{\mathrm{W}}_{\alpha_n}$ (noir) et $\log \hat{q}^{\mathrm{WG}}_{\alpha_n}$ (bleu).

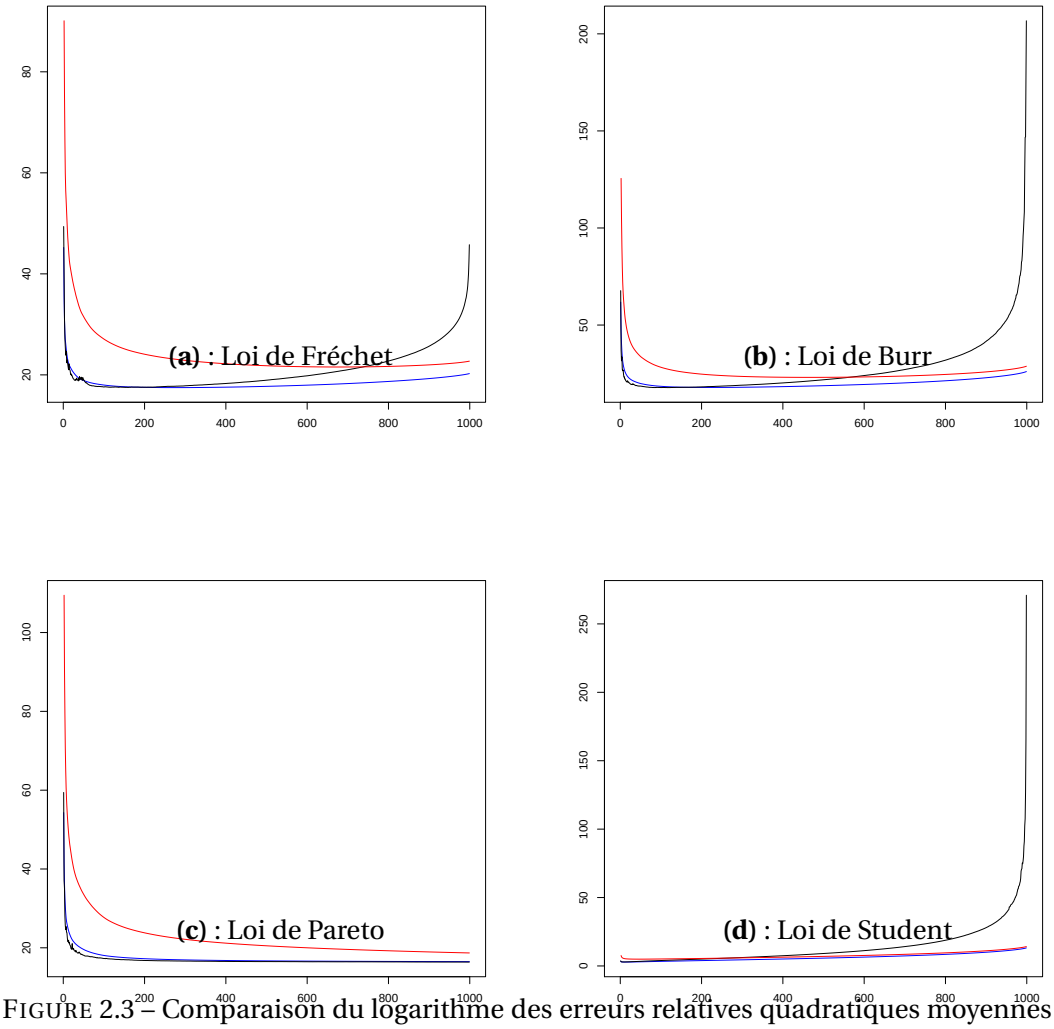


FIGURE 2.3 – Comparaison du logarithme des erreurs relatives quadratiques moyennes de $\hat{q}^{\mathrm{FG}}_{\alpha_n}$ (rouge), $\hat{q}^{\mathrm{W}}_{\alpha_n}$ (noir) et $\hat{q}^{\mathrm{WG}}_{\alpha_n}$ (bleu). En ordonnée, on a l'erreur quadratique moyenne et en abscisse le seuil k_n.

2.4 Démonstrations

Démonstration de la Proposition 2.3. En passant au logarithme, l'estimateur (2.1) se réécrit

$$\begin{aligned}
\log \hat{q}_{\alpha_n}^{\mathrm{WG}} &= \left(\frac{1}{k_n}\sum_{i=1}^{k_n}\log X_{n-i+1,n} - \log X_{n-k_n,n}\right) + \log X_{n-k_n,n} + \frac{1}{k_n}\sum_{i=1}^{k_n}\frac{1}{i}\log\left(\frac{i g_{k_n}}{n\alpha_n}\right)\sum_{j=1}^{i}\Delta_j \\
&= \frac{1}{k_n}\sum_{i=1}^{k_n} i\left(\log X_{n-i+1,n} - \log X_{n-i,n}\right) + \log X_{n-k_n,n} + \frac{1}{k_n}\sum_{i=1}^{k_n}\frac{1}{i}\log\left(\frac{i g_{k_n}}{n\alpha_n}\right)\sum_{j=1}^{i}\Delta_j \\
&= \hat{\gamma}_{k_n}^{\mathrm{H}} + \log X_{n-k_n,n} + \frac{1}{k_n}\sum_{i=1}^{k_n}\frac{1}{i}\log\left(\frac{i g_{k_n}}{n\alpha_n}\right)\sum_{j=1}^{i}\Delta_j.
\end{aligned}$$

D'après le Lemme 1.4.2 sur la transformation des quantiles, nous avons

$$X_{n-k_n,n} \overset{\mathscr{L}}{=} F^{\leftarrow}(V_{n-k_n,n}) = F^{\leftarrow}(1-(1-V_{n-(k_n+1)+1,n})) = \bar{F}^{\leftarrow}(V_{k_n+1,n}) = V_{k_n+1,n}^{-\gamma} L\left(V_{k_n+1,n}^{-1}\right),$$

où $V_{k_n+1,n}$ est la $(n-k_n)$-ième plus grande statistique d'ordre d'un échantillon de variables aléatoires indépendantes de loi uniforme standard $\{V_i,\ i=1,\dots,n\}$. Il en découle alors que

$$\begin{aligned}
\log \hat{q}_{\alpha_n}^{\mathrm{WG}} &\overset{\mathscr{L}}{=} \hat{\gamma}_{k_n}^{\mathrm{H}} - \gamma\log V_{k_n+1,n} + \log L\left(1/V_{k_n+1,n}\right) + \frac{1}{k_n}\sum_{i=1}^{k_n}\frac{1}{i}\log\left(\frac{i g_{k_n}}{n\alpha_n}\right)\sum_{j=1}^{i}\Delta_j \\
&\overset{\mathscr{L}}{=} \hat{\gamma}_{k_n}^{\mathrm{H}} - \gamma\log V_{k_n+1,n} + \log L\left(1/V_{k_n+1,n}\right) + \frac{1}{k_n}\sum_{j=1}^{k_n}\sum_{i=j}^{k_n}\frac{1}{i}\log\left(\frac{i g_{k_n}}{n\alpha_n}\right)\Delta_j \\
&\overset{\mathscr{L}}{=} \hat{\gamma}_{k_n}^{\mathrm{H}} - \gamma\log V_{k_n+1,n} + \log L\left(1/V_{k_n+1,n}\right) + \frac{1}{k_n}\sum_{j=1}^{k_n}\pi_j\Delta_j.
\end{aligned}$$

En introduisant la variable aléatoire définie par $\hat{\gamma}_{k_n}^{\pi} = \sum_{j=1}^{k_n}\Delta_j\pi_j \Big/ \sum_{j=1}^{k_n}\pi_j$ où

$$\begin{aligned}
\sum_{j=1}^{k_n}\pi_j = \sum_{j=1}^{k_n}\sum_{i=j}^{k_n}\frac{1}{i}\log\left(\frac{i g_{k_n}}{n\alpha_n}\right) &= \sum_{j=1}^{k_n}\sum_{j=1}^{i}\frac{1}{i}\log\left(\frac{i g_{k_n}}{n\alpha_n}\right) \\
&= \sum_{j=1}^{k_n}\sum_{j=1}^{i}\frac{\log i}{i} + \sum_{j=1}^{k_n}\sum_{j=1}^{i}\frac{1}{i}\log\left(\frac{g_{k_n}}{n\alpha_n}\right) \\
&= \log(k_n!) + k_n\log\left(\frac{g_{k_n}}{n\alpha_n}\right),
\end{aligned}$$

on aboutit à

$$\begin{aligned}
\log \hat{q}_{\alpha_n}^{\mathrm{WG}} &\overset{\mathscr{L}}{=} \hat{\gamma}_{k_n}^{\mathrm{H}} - \gamma\log V_{k_n+1,n} + \log L\left(1/V_{k_n+1,n}\right) + \frac{1}{k_n}\left(\log(k_n!) + k_n\log\left(\frac{g_{k_n}}{n\alpha_n}\right)\right)\hat{\gamma}_{k_n}^{\pi} \\
&\overset{\mathscr{L}}{=} \hat{\gamma}_{k_n}^{\mathrm{H}} - \gamma\log V_{k_n+1,n} + \log L\left(1/V_{k_n+1,n}\right) + \log\left(\frac{1}{\mathrm{e}}\frac{(k_n+1)}{n\alpha_n}\right)\hat{\gamma}_{k_n}^{\pi}
\end{aligned}$$

puisque $g_{k_n} = \exp\left[\log(k_n+1) - 1 - \log(k_n!)/k_n\right]$. Ce qui achève la démonstration. □

Démonstration du Lemme 2.2.1. Comme la formule de Stirling donnée par

$$k_n! = (2\pi k_n)^{1/2}\left(\frac{k_n}{\mathrm{e}}\right)^{k_n}(1+o(1)),$$

montre que $g_{k_n} = \exp\left\{\frac{\log(k_n+1)}{k_n} - \frac{\log(2\pi k_n)}{k_n} + o(1)\right\} \to 1$ quand $k_n \to \infty$ alors, $\forall\, \varepsilon \in]0,1[, \exists\, N \in \mathbb{N}$ tel que $\forall\ n \geq N$, $1-\varepsilon \leq g_{k_n} \leq 1+\varepsilon$. Ceci implique que pour tout $i = 1,\dots,k_n$, nous avons

$$0 < \frac{i(1-\varepsilon)}{n\alpha_n} \leq \frac{ig_{k_n}}{n\alpha_n} \leq \frac{i(1+\varepsilon)}{n\alpha_n}.$$

En passant au logarithme de part et d'autre des inégalités, nous avons pour $n \geq N$

$$\log\left(\frac{i(1-\varepsilon)}{n\alpha_n}\right) \leq \log\left(\frac{ig_{k_n}}{n\alpha_n}\right) \leq \log\left(\frac{i(1+\varepsilon)}{n\alpha_n}\right) \Rightarrow \log\left(\frac{(1-\varepsilon)}{n\alpha_n}\right) \leq \log\left(\frac{ig_{k_n}}{n\alpha_n}\right) \leq \log\left(\frac{k_n(1+\varepsilon)}{n\alpha_n}\right).$$

Nous pouvons réécrire

$$1 - \frac{\log(1-\varepsilon)}{\log(n\alpha_n)} \leq \log\left(\frac{ig_{k_n}}{n\alpha_n}\right)\Big/\log\left(\frac{1}{n\alpha_n}\right) \leq 1 - \frac{\log(1+\varepsilon)}{\log(n\alpha_n)} - \frac{\log k_n}{\log(n\alpha_n)}.$$

La condition $\log(k_n)/\log(n\alpha_n) \to 0$ entraine que

$$1 - \frac{\log(1-\varepsilon)}{\log(n\alpha_n)} \to 1 \text{ et } 1 - \frac{\log(1+\varepsilon)}{\log(n\alpha_n)} - \frac{\log k_n}{\log(n\alpha_n)} \to 1 \text{ si } n\alpha_n \to 0.$$

Donc, uniformément pour tout $i = 1,\dots,k_n$, nous avons

$$\log\left(\frac{ig_{k_n}}{n\alpha_n}\right) = \log\left(\frac{1}{n\alpha_n}\right)(1+o(1)).$$

Par conséquent,

$$\sum_{j=1}^{i}\frac{1}{i}\log\left(\frac{ig_{k_n}}{n\alpha_n}\right) = \log\left(\frac{1}{n\alpha_n}\right)(1+o(1))\sum_{i=j}^{k_n} 1/i,$$

permet de conclure la preuve. □

Démonstration du Lemme 2.2.2. Pour $i > 1$, la méthode des rectangles nous assure que

$$\int_i^{i+1}\frac{1}{t}dt \leq \frac{1}{i} \leq \int_{i-1}^{i}\frac{1}{t}dt.$$

En passant à la somme pour $j > 1$,

$$\sum_{i=j}^{k_n}\int_i^{i+1}\frac{1}{t}dt \leq \tau_j \leq \sum_{i=j}^{k_n}\int_{i-1}^{i}\frac{1}{t}dt,$$

on obtient,

$$\log\frac{k_n+1}{j} \leq \tau_j \leq \log\frac{k_n}{j-1}.$$

On peut donc réécrire

$$0 \le \tau_j / \log\big((k_n+1)/j\big) - 1 \le \frac{\log k_n - \log(j-1)}{\log(k_n+1) - \log j} - 1.$$

En introduisant la fonction $h_n(j) = \dfrac{\log k_n - \log(j-1)}{\log(k_n+1) - \log j} - 1$, on a

$$h_n'(j) = \frac{\varphi_n(j)}{j(j-1)(\log(k_n+1) - \log(j))^2},$$

où $\varphi_n(j) = -j\log(k_n+1) + j\log(j) + j\log(k_n) - j\log(j-1) - \log(k_n) + \log(j-1)$. Comme pour $j > 1$, $\varphi_n''(j) = \frac{1}{j} - \frac{1}{j-1} < 0$, $\varphi_n'(k_n+1) = 0$ et $\varphi_n(k_n+1) = 0$, il en découle que pour tout $j < k_n+1$, $\varphi_n(j) < 0$. Ainsi, puisque $h_n(k_n) > 0$, on a finalement que la fonction h_n est décroissante et positive sur $[2, k_n]$ et par conséquent on a pour $j > 1$,

$$0 \le h_n(j) \le h_n(2) = \frac{\log k_n}{\log(k_n+1) - \log 2} - 1 \to 0 \text{ quand } k_n \to \infty.$$

Donc, uniformément pour tout j tel que $1 < j \le k_n$, on a

$$\tau_j = -\log(j/(k_n+1))(1+o(1)).$$

D'autre part, si $j = 1$, on a

$$\sum_{i=1}^{k_n} 1/i - \log(k_n+1) = C_{Euler} + o(1),$$

où C_{Euler} est la constante de Euler. Ceci implique que

$$\frac{\tau_1}{\log(k_n+1)} - 1 = o(1) \Rightarrow \tau_1 \sim \log(k_n+1).$$

En conclusion, uniformément pour tout $j = 1, \dots, k_n$, nous avons

$$\tau_j = -\log(j/(k_n+1))(1+o(1)) \overset{def}{=} W^\tau(j/(k_n+1))(1+o(1)).$$

Ceci achève la preuve. □

Démonstration de la Proposition 2.2.2. Commençons par remarquer que

$$\begin{aligned}
\hat{\gamma}^{\pi}_{k_n} = \frac{\sum_{j=1}^{k_n} \Delta_j \pi_j}{k_n \log\left(\frac{1}{e}\frac{(k_n+1)}{n\alpha_n}\right)} &= \frac{\sum_{j=1}^{k_n} \Delta_j W^\tau\left(\frac{j}{k_n+1}\right)(1+o(1))}{k_n \log\left(\frac{1}{e}\frac{(k_n+1)}{n\alpha_n}\right) \Big/ \log\left(\frac{1}{n\alpha_n}\right)} \\
&= \frac{1}{k_n}\sum_{j=1}^{k_n} \Delta_j W^\tau\left(\frac{j}{k_n+1}\right)(1+o(1)) \qquad (2.5) \\
&\quad \text{puisque } \log\left(\frac{1}{e}\frac{(k_n+1)}{n\alpha_n}\right) \Big/ \log\left(\frac{1}{n\alpha_n}\right) \to 1.
\end{aligned}$$

Pour tout $s \in]0,1[$, la fonction $W^\tau(s) \overset{def}{=} -\log(s)$ satisfait les conditions de Beirlant *et al.* (2002) (voir Annexe A, Théorème A.0.1) avec $u(t) = -\log(t) - 1$ et $g(s) = 1 - \log(s)$. D'une part, puisque la fonction W^τ est monotone et décroissante, si $j > 1$ alors, on a

$$\left| k_n \int_{(j-1)/k_n}^{j/k_n} u(x)dx \right| \leq W^\tau\left(\frac{j-1}{k_n}\right) \leq 1 - \log\left(\frac{j-1}{k_n}\right) \leq 1 - \log\left(\frac{1}{k_n+1}\right) = g\left(\frac{1}{k_n+1}\right),$$

et dans le cas particulier où $j = 1$, on a

$$\left| k_n \int_{(j-1)/k_n}^{j/k_n} u(x)dx \right| = W^\tau\left(\frac{1}{k_n}\right) \leq 1 - \log\left(\frac{1}{k_n}\right) \leq 1 - \log\left(\frac{1}{k_n+1}\right) = g\left(\frac{1}{k_n+1}\right).$$

D'autre part, on a

$$\int_0^1 \left(\log(1/s) \vee 1\right) g(s)ds < \infty \text{ et } \int_0^1 \left|W^\tau\right|^{2+\delta}(s)ds < \infty \text{ pour tout } \delta > 0.$$

Par conséquent la variable aléatoire

$$k_n^{1/2}\left[\frac{1}{k_n}\sum_{j=1}^{k_n} \Delta_j W^\tau\left(\frac{j}{k_n+1}\right) - \frac{1}{k_n}\sum_{j=1}^{k_n} W^\tau\left(\frac{j}{k_n+1}\right)\left\{\gamma + \varepsilon(n/k_n)\left(\frac{j}{k_n+1}\right)^{-\rho}\right\}\right] \tag{2.6}$$

converge vers une loi normale centrée de variance $\gamma^2 \int_0^1 (W^\tau)^2(s)ds$. La fonction $f(s) = \log(s)s^{-\rho}$ est continue sur $[0,1]$, donc intégrable au sens de Reimann et on a

$$\begin{aligned} \frac{1}{k_n+1}\sum_{j=1}^{k_n+1} W^\tau\left(\frac{j}{k_n+1}\right)\left(\frac{j}{k_n+1}\right)^{-\rho} &= \frac{1}{k_n+1}\sum_{j=1}^{k_n+1} \log\left(\frac{j}{k_n+1}\right)\left(\frac{j}{k_n+1}\right)^{-\rho} \\ &\to \int_0^1 \log(s)s^{-\rho}ds = -\frac{1}{(1-\rho)^2}. \end{aligned} \tag{2.7}$$

Comme le $o(1)$ de la variable aléatoire $\hat{\gamma}_{k_n}^\pi$ est uniforme en $j = 1, \ldots, k_n$ alors,

$$(2.5) \Rightarrow (1 + o(1))\hat{\gamma}_{k_n}^\pi = \frac{1}{k_n}\sum_{j=1}^{k_n} \Delta_j W^\tau\left(\frac{j}{k_n+1}\right). \tag{2.8}$$

Aussi, puisque $\tau_j = W^\tau\left(\frac{j}{k_n+1}\right)(1 + o(1))$ uniformément en $j = 1, \ldots, k_n$ il s'en suit que

$$\gamma\frac{1}{k_n}\sum_{j=1}^{k_n} \Delta_j W^\tau\left(\frac{j}{k_n+1}\right) = \gamma\frac{(1+o(1))}{k_n}\sum_{j=1}^{k_n} \tau_j = \gamma\frac{(1+o(1))}{k_n}\sum_{j=1}^{k_n}\sum_{i=j}^{k_n}\frac{1}{i} = \gamma(1+o(1)). \tag{2.9}$$

En combinant (2.9), (2.8), (2.7) et (2.6) on obtient

$$k_n^{1/2}(1+o(1))\left[\hat{\gamma}_{k_n}^\pi - \gamma - \frac{\varepsilon(n/k_n)(1+o(1))}{(1-\rho)^2}\right] \overset{\mathscr{L}}{\to} \mathscr{N}\left(0, \gamma^2\int_0^1 \log^2(1/s)ds\right),$$

qui entraine que

$$k_n^{1/2}\left(\hat{\gamma}_{k_n}^\pi - \gamma - \frac{\varepsilon(n/k_n)(1+o(1))}{(1-\rho)^2}\right) \overset{\mathscr{L}}{\to} \mathscr{N}\left(0, 2\gamma^2\right).$$

Ainsi, si $k_n^{1/2}\varepsilon(n/k_n) \to \lambda \in \mathbb{R}$ alors,

$$k_n^{1/2}\left(\hat{\gamma}_{k_n}^{\pi} - \gamma - \frac{\varepsilon(n/k_n)}{(1-\rho)^2}\right) \xrightarrow{\mathscr{L}} \mathcal{N}\left(0, 2\gamma^2\right).$$

Ce qui conclut la preuve de la Proposition. □

Démontration du Théorème 2.2.1. D'après la décomposition de la Proposition 2.2.2, nous avons

$$\log \hat{q}_{\alpha_n}^{\mathrm{WG}} \overset{\mathscr{L}}{=} \hat{\gamma}_{k_n}^{\mathrm{H}} - \gamma \log V_{k_n+1,n} + \log L\left(1/V_{k_n+1,n}\right) + \log\left(\frac{1}{\mathrm{e}}\frac{(k_n+1)}{n\alpha_n}\right)\hat{\gamma}_{k_n}^{\pi},$$

et comme par définition $\log q_{\alpha_n} = -\gamma \log \alpha_n + \log L(1/\alpha_n)$, alors

$$\begin{aligned}
\log \hat{q}_{\alpha_n}^{\mathrm{WG}} - \log q_{\alpha_n} &= \hat{\gamma}_{k_n}^{\mathrm{H}} - \gamma \log V_{k_n+1,n} + \gamma \log \alpha_n + \log\left(\frac{L\left(V_{k_n+1,n}^{-1}\right)}{L(1/\alpha_n)}\right) + \log\left(\frac{1}{\mathrm{e}}\frac{(k_n+1)}{n\alpha_n}\right)\hat{\gamma}_{k_n}^{\pi} \\
&= (\hat{\gamma}_{k_n}^{\mathrm{H}} - \gamma) + \gamma\left(\log V_{k_n+1,n}^{-1} - \log(n/(k_n+1)\right) + \log\left(\frac{L\left(V_{k_n+1,n}^{-1}\right)}{L(1/\alpha_n)}\right) \\
&+ \log\left(\frac{1}{\mathrm{e}}\frac{(k_n+1)}{n\alpha_n}\right)\left(\hat{\gamma}_{k_n}^{\pi} - \gamma\right). \qquad (2.10)
\end{aligned}$$

D'après le Lemme 1.4.3 sur la représentation de Rényi (1953) et la conséquence de la Proposition 1.4.1 on a

$$x_n \overset{def}{=} V_{k_n+1,n}^{-1} \overset{\mathbb{PS}}{\sim} n/(k_n+1) \to \infty \text{ et } \lambda_n \overset{def}{=} \frac{V_{k_n+1,n}}{\alpha_n} \overset{\mathbb{PS}}{\sim} \frac{k_n+1}{n\alpha_n} \to \infty \text{ si } \alpha_n < 1/n.$$

D'où

$$\log\left(L(1/\alpha_n)/L\left(V_{k_n+1,n}^{-1}\right)\right) = \log\left(L(\lambda_n x_n)/L(x_n)\right).$$

La conséquence de la condition **(C.1)** (voir Remarque 1.5.1) implique que pour n assez grand on a

$$\frac{(1+\varepsilon)\,\lambda_n^{\rho+\varepsilon} - 1}{\rho} \le \frac{1}{\varepsilon(x_n)}\log\left(\frac{L(1/\alpha_n)}{L\left(V_{k_{n+1}}^{-1}\right)}\right) \le \frac{(1-\varepsilon)\,\lambda_n^{\rho-\varepsilon} - 1}{\rho}.$$

Si l'on prend $0 < \varepsilon < -\rho$, les fonctions de λ_n que sont $\dfrac{(1+\varepsilon)\,\lambda_n^{\rho+\varepsilon} - 1}{\rho}$ et $\dfrac{(1-\varepsilon)\,\lambda_n^{\rho-\varepsilon} - 1}{\rho}$ sont croissantes et comme $\lambda_n \to \infty$ alors, on a

$$\log\left(L(1/\alpha_n)/L\left(V_{k_n+1,n}^{-1}\right)\right) = -\frac{1}{\rho}\varepsilon(x_n)(1 + o_{\mathrm{P}}(1)).$$

De même, comme $x_n \overset{\mathbb{PS}}{\sim} n/(k_n+1) \sim n/k_n$ et d'après la Proposition 1.4.1, on a

$$\log\left(L(1/\alpha_n)/L\left(V_{k_n+1,n}^{-1}\right)\right) = -\frac{1}{\rho}\varepsilon(n/k_n)(1 + o_{\mathrm{P}}(1)). \qquad (2.11)$$

Puisque $E_{n-k_n,n} \overset{\mathscr{L}}{=} \log V_{k_n+1,n}^{-1}$, la conséquence de la Proposition 1.4.1 implique que

$$\log V_{k_n+1,n}^{-1} - \log(n/(k_n+1) = \frac{\xi^{\mathrm{E}}}{k_n^{1/2}}, \tag{2.12}$$

avec ξ^{E} une variable aléatoire qui converge vers une loi normale centrée réduite. La démonstration de la Proposition 2.2.2, montre que pour tout $s \in [0,1]$, si l'on pose $W^{\tau}(s) = 1$ alors,

$$\hat{\gamma}_{k_n}^{\mathrm{H}} - \gamma = \frac{\xi^{\mathrm{H}}}{k_n^{1/2}} + \frac{\varepsilon(n/k_n)}{1-\rho}, \tag{2.13}$$

avec ξ^{H} une variable aléatoire qui converge vers une loi normale centrée de variance γ^2. Aussi, la Proposition 2.2.2 entraine que

$$\hat{\gamma}_{k_n}^{\pi} - \gamma = \frac{\xi^{\pi}}{k_n^{1/2}} + \frac{\varepsilon(n/k_n)}{(1-\rho)^2}, \tag{2.14}$$

avec ξ^{π} une variable aléatoire qui converge vers une loi normale centrée de variance $2\gamma^2$. Finalement, en combinant (2.10), (2.11), (2.12), (2.13) et (2.14) on a donc que

$$\log \frac{\hat{q}_{\alpha_n}^{\mathrm{WG}}}{q_{\alpha_n}} = \frac{\xi^{\mathrm{H}}}{k_n^{1/2}} + \frac{\varepsilon(n/k_n)}{(1-\rho)} + \frac{\xi^{\mathrm{E}}}{k_n^{1/2}} - \frac{1}{\rho}(1+o_{\mathrm{P}}(1)) + \log\left(\frac{1}{\mathrm{e}}\frac{(k_n+1)}{n\alpha_n}\right)\left(\frac{\xi^{\pi}}{k_n^{1/2}} + \frac{\varepsilon(n/k_n)}{(1-\rho)^2}\right).$$

L'hypothèse $\alpha_n < 1/n$ implique que $\log\left(\frac{1}{\mathrm{e}}\frac{(k_n+1)}{n\alpha_n}\right) \to \infty$ et nous avons

$$\begin{aligned}\frac{k_n^{1/2}}{\log\left(\frac{1}{\mathrm{e}}\frac{(k_n+1)}{n\alpha_n}\right)}\log\left(\frac{\hat{q}_{\alpha_n}^{\mathrm{WG}}}{q_{\alpha_n}}\right) - \frac{k_n^{1/2}\varepsilon(n/k_n)}{(1-\rho)^2} &= O_{\mathrm{P}}\left(\frac{1}{\log\left(\frac{1}{\mathrm{e}}\frac{(k_n+1)}{n\alpha_n}\right)}\right) + \frac{k_n^{1/2}\varepsilon(n/k_n)}{(1-\rho)\log\left(\frac{1}{\mathrm{e}}\frac{(k_n+1)}{n\alpha_n}\right)} \\ &- \frac{k_n^{1/2}\varepsilon(n/k_n)}{\rho\log\left(\frac{1}{\mathrm{e}}\frac{(k_n+1)}{n\alpha_n}\right)}(1+o_{\mathrm{P}}(1)) + \xi^{\pi}.\end{aligned}$$

Si de plus, $k_n^{1/2}\varepsilon(n/k_n) \to \lambda \in \mathbb{R}$, alors

$$\frac{k_n^{1/2}}{\log\left(\frac{1}{\mathrm{e}}\frac{(k_n+1)}{n\alpha_n}\right)}\log\left(\frac{\hat{q}_{\alpha_n}^{\mathrm{WG}}}{\log q_{\alpha_n}}\right) - \frac{k_n^{1/2}\varepsilon(n/k_n)}{(1-\rho)^2} = O_{\mathrm{P}}\left(\frac{1}{\log\left(\frac{1}{\mathrm{e}}\frac{(k_n+1)}{\mathrm{e}n\alpha_n}\right)}\right) + O\left(\frac{1}{\log\left(\frac{1}{\mathrm{e}}\frac{(k_n+1)}{n\alpha_n}\right)}\right) + \xi^{\pi},$$

et par conséquent

$$\frac{k_n^{1/2}}{\log\left(\frac{1}{\mathrm{e}}\frac{(k_n+1)}{n\alpha_n}\right)}\left(\log\hat{q}_{\alpha_n}^{\mathrm{WG}} - \log q_{\alpha_n}\right) - \frac{k_n^{1/2}\varepsilon(n/k_n)}{(1-\rho)^2} \xrightarrow{\mathscr{L}} \mathscr{N}(0, 2\gamma^2).$$

Ceci permet de conclure la preuve du Théorème. □

3

Quantiles conditionnels

Résumé

Dans ce chapitre bibliographique, on présente les différentes approches d'estimation des quantiles conditionnels. Quelques définitions et des résultats utiles y sont donnés.

Sommaire

3.1 Introduction **45**
3.2 Estimation paramétrique des quantiles conditionnels **46**
3.3 Estimation non paramétrique des quantiles conditionnels **46**
3.3.1 Méthode d'estimation indirecte 46
3.3.2 Méthode d'estimation directe 51

3.1 Introduction

Nous recensons dans la littérature deux types d'approches pour l'estimation des quantiles conditionnels que sont :

- l'approche paramétrique, brièvement présentée dans la partie 3.2 ;
- l'approche non-paramétrique, présentée de façon plus détaillée dans la partie 3.3.

Dans le cadre de ces approches, on distingue selon la nature de la variable explicative :

- Le modèle dit « *à plan fixe* » ou « *design fixe* » dont les données sont des couples $\{(x_i, Y_i),\ i = 1, \ldots, n\}$ où les observations Y_i sont des variables aléatoires réelles non nécessairement indépendantes et les x_i sont des points d'observations non aléatoires.
- Le modèle dit « *à plan aléatoire* » ou « *design aléatoire* » pour lequel les données sont des couples $\{(X_i, Y_i),\ i = 1, \ldots, n\}$ de variables aléatoires réelles non nécessairement indépendantes et identiquement distribuées de même loi.

3.2 Estimation paramétrique des quantiles conditionnels

Elle est généralement utilisée quand l'on dispose d'un échantillon de petite taille. Afin d'estimer le quantile conditionnel, une façon de procéder consiste à supposer la fonction de répartition conditionnelle gaussienne. L'estimateur correspondant du quantile conditionnel $q(\alpha|x)$ est défini par :

$$\hat{q}_n(\alpha|x) = \hat{m}_n(x) + z_\alpha \hat{\sigma}_n(x),$$

où $\hat{m}_n(x)$ (resp. $\hat{\sigma}_n(x)$) désigne l'estimateur de l'espérance conditionnelle (resp. l'écart type conditionnel) de Y sachant $X = x$ et z_α le quantile d'ordre α de la loi normal centrée réduite. Afin d'estimer $m(x)$ et $\sigma^2(x)$, Royston et Wright (1998) ont utilisé un modèle polynomial qu'ils ont associé à la méthode des moindres carrés. Cependant, dans la pratique, lorsque l'on dispose des données connues pour leurs valeurs aberrantes comme en biologie, il est parfois nécessaire de les transformer dans l'espoir d'obtenir des résidus normalement distribués (voir Cole (1988) et Tango (1998)). Hormis la restrictivité des hypothèses paramétriques, il semblerait même que l'existence d'une telle transformation ne soit pas toujours possible (Cole, 1988).

Pour pallier ces problèmes de modélisation et ou d'hypothèses, une nouvelle approche dite non-paramétrique a été mise en oeuvre.

3.3 Estimation non paramétrique des quantiles conditionnels

Dans le cadre de l'approche non-paramétrique, on distingue deux méthodes d'estimation. La première consiste à estimer au préalable la fonction de répartition conditionnelle puis à l'inverser pour en obtenir un estimateur du quantile conditionnel. Elle sera présentée au paragraphe 3.3.1. La seconde consiste quant à elle en une estimation directe basée sur le principe des moindres carrés. Elle sera exposée au paragraphe 3.3.2. De nombreuses études ont été menées dans le cas d'un processus Markovien (Roussas, 1969), lorsque les données sont indépendantes et identiquement distribuées (Stute, 1986; Samanta, 1989) ou α-mélangeantes (Berlinet *et al.*, 1998).

3.3.1 Méthode d'estimation indirecte

On dénombre deux techniques d'estimation indirectes de la fonction de répartition conditionnelle :

- Les estimateurs à noyaux qui peuvent être construits suivant :
  - (a) la méthode d'estimation du simple noyau ;
  - (b) la méthode d'estimation par noyau produit ou de Roussas (1969) ;
  - (c) la méthode du médianogramme ou de la fenêtre mobile.
- Les estimateurs au sens des plus proches voisins.

3.3.1.1 Estimateurs à noyaux

L'expression de l'estimateur du quantile conditionnel construit en inversant la fonction de répartition conditionnelle est donnée par :

$$\hat{q}_n(\alpha|x) = \inf\left\{y \mid \hat{F}_n\left(y|x\right) \geq \alpha\right\}. \tag{3.1}$$

(a)- Les estimateurs par simple noyau de la fonction de répartition conditionnelle

On estime la fonction de répartition conditionnelle par :

$$\hat{F}_n\left(y|x\right) = \sum_{i=1}^{n} w_{ni}(x)\mathbb{1}_{\{Y_i \leq y\}} \text{ avec } \sum_{i=1}^{n} w_{ni}(x) = 1. \tag{3.2}$$

Notons que si l'on ne conditionne pas, i.e $w_{ni}(x) = 1/n$, on retrouve l'expression classique de la fonction de répartition empirique. En posant $Y^* = \mathbb{1}_{\{Y \leq y\}}$, on a $F(y|x) = \mathbb{E}[Y^*|X = x]$ et $\hat{q}_n(\alpha|x)$ est appelé *l'estimateur du quantile de régression.*

Cas du design aléatoire : La convergence ponctuelle en probabilité de l'estimateur (3.2) a été établie par Stone (1977) [1] lorsque les variables aléatoires $(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}$ sont indépendantes et identiquement distribuées. Collomb (1980) après avoir étudié les propriétes asymptotiques d'un estimateur à noyau de probabilité conditionnelle d'un couple de variables aléatoires indépendantes et identiquement distribuées à valeurs dans $\mathbb{R}^p \times \mathbb{R}$ a proposé d'estimer la fonction de répartition conditionnelle en posant

$$w_{ni}(x) = K\left(\frac{x - X_i}{h_n}\right) \Bigg/ \sum_{j=1}^{n} K\left(\frac{x - X_j}{h_n}\right),$$

où K est une densité de probabilité appelée « *noyau* » et h_n un paramètre qui converge vers zéro lorsque n tend vers l'infini. Il a par ailleurs démontré la convergence ponctuelle et uniforme en x de son estimateur. Il n'a pas démontré la convergence uniforme en y. Il a également donné l'erreur quadratique moyenne optimale ainsi que la normalité asymptotique de ses estimateurs.

En utilisant un noyau de probabilité continu et borné sur $[-1, 1]$ Stute (1986) posa :

$$w_{ni}(x) = K\left(\frac{G_n(x) - G_n(X_i)}{h_n}\right) \Bigg/ \sum_{j=1}^{n} K\left(\frac{G_n(x) - G_n(X_j)}{h_n}\right) \text{ avec } G_n(x) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{\{X_j \leq x\}}.$$

Il a démontré la convergence uniforme de son estimateur de la fonction de répartition conditionnelle. Il a établi un résultat sur la normalité asymptotique de l'estimateur du quantile conditionnel associé.

1. Plus précisément, il donne des conditions sur les poids w_{ni} pour que l'estimateur converge ponctuellement en probabilité.

Horváth et Yandell (1988) donnent des conditions permettant d'approcher la suite $\phi_n(y|x) = (nh_n)^{1/2}\left(\hat{F}_n\left(y|x\right) - F(y|x)\right)$ par un processus gaussien. Ils énoncent le résultat sur la convergence uniforme (en probabilité) en y de la suite $\phi_n(y|x)$ des estimateurs de Collomb (1980) et Stute (1986). Toutefois, notons que Stute (1986) a établi un résultat sur la convergence faible de la suite $\phi_n(y|x)$ vers un processus gaussien.

Gannoun (1989) étudie les propriétés de l'estimateur de Collomb (1980) dans le cas de données indépendantes et identiquement distribuées puis α-mélangeantes. Hart (1991) propose quant à lui d'utiliser un noyau de probabilité de support $[0,1]$ et de modifier l'estimateur de Stute (1986) en prenant une famille de poids définie par :

$$w_{ni}(x) = \frac{1}{h_n}\int_{\frac{(i-1)}{n}}^{i/n} K\left(\frac{G_n(x)-u}{h_n}\right)du.$$

Lorsque la covariable X est de nature fonctionnelle ou de dimension infinie, Ferraty et Vieu (2000) proposent de modifier la famille de poids introduite par Collomb (1980) en posant

$$w_{ni}(x) = K\left(\frac{d(x,X_i)}{h_n}\right)\Bigg/\sum_{j=1}^{n} K\left(\frac{d(x,X_j)}{h_n}\right),$$

où d est une distance semi-métrique qui mesure la proximité entre deux objets fonctionnels et K un noyau qui est positif et décroissant sur $[0,1]$. Les vitesses de convergence presque complète ont été obtenues dans le cas d'un échantillon indépendant et dans le cas dépendant. On pourra se référer à l'ouvrage (Ferraty et Vieu, 2006, chapitre 6).

Cas du design fixe : Antoch et Janssen (1989) étendent l'étude d'un modèle de régression initialement introduit par Gasser et Hüller (1984) à l'estimation des quantiles conditionnels. Ils proposent alors d'estimer la fonction de répartition conditionnelle en posant :

$$\begin{cases} w_{ni}(x) = \dfrac{1}{h_n}\displaystyle\int_{x_{i-1}}^{x_i} K\left(\frac{x-u}{h_n}\right)du \quad \text{pour } 2 \le i \le n-1 \\ w_{n1}(x) = \dfrac{1}{h_n}\displaystyle\int_{-\infty}^{x_1} K\left(\frac{x-u}{h_n}\right)du \\ w_{nn}(x) = \dfrac{1}{h_n}\displaystyle\int_{x_{n-1}}^{\infty} K\left(\frac{x-u}{h_n}\right)du \end{cases}$$

où K est un noyau de probabilité. Les auteurs donnent une représentation de type Bahadur de l'estimateur de quantile qui en découle.

(b)- L'estimateur de Roussas (1969)

Une version plus lisse et régulière des estimateurs simple noyau de la fonction de répartition conditionnelle (3.2) avait été introduite par Roussas (1969) dans le cas d'un processus (X_i, Y_i) à valeurs dans $\mathbb{R}\times\mathbb{R}$, supposé Markovien. Elle consiste

tout d'abord à remplacer la fonction indicatrice de l'estimateur (3.2) par une densité symétrique, puis à effectuer préalablement à l'inversion un lissage par noyau de type Parzen (1979) de la fonction de densité marginale $g(x)$ de X et de la fonction de densité conjointe $f(x,y)$ de (X,Y). Cela revient à supposer que la variable aléatoire (X,Y) admet une densité de probabilité $f(.,.)$ et que la densité conditionnelle de Y sachant $X=x$ admet une version régulière $f(.|x)$.

L'estimateur de la densité conditionnelle étant défini par le rapport entre les estimateurs de la densité du couple $f(x,y)$ et de la densité marginale $g(x)$, il en découle que l'estimateur à noyau de la fonction de répartition conditionnelle est [2] :

$$\hat{F}_n\left(y|x\right) = \int_{-\infty}^{y} \hat{f}_n\left(u|x\right)du = \frac{\int_{-\infty}^{y}\hat{f}_n\left(x,u\right)du}{\hat{g}_n\left(x\right)} \overset{def}{=} \frac{\hat{\psi}_n(x,y)}{\hat{g}_n\left(x\right)}, \tag{3.3}$$

où
$$\hat{f}_n\left(x,y\right) = \frac{1}{nh_n^{d+1}}\sum_{i=1}^{n} K_0\left(\frac{x-X_i}{h_n}\right)K_1\left(\frac{y-Y_i}{h_n}\right),$$
$$\hat{g}_n(x) = \int_{\mathbb{R}^p}\hat{f}_n\left(x,y\right)dy = \frac{1}{nh_n^{d}}\sum_{i=1}^{n} K_0\left(\frac{x-X_i}{h_n}\right),$$
$$\hat{\psi}_n(x,y) = \frac{1}{nh_n^{d+1}}\sum_{i=1}^{n} K_0\left(\frac{x-X_i}{h_n}\right)\int_{-\infty}^{y} K_1\left(\frac{z-Y_i}{h_n}\right)dz,$$

avec K_0 et K_1 des noyaux de probabilité. Ainsi que l'on peut le remarquer, cet estimateur a été construit pour le *design aléatoire*. À notre connaissance, il en n'existe pas de version pour le *design fixe*.

De nombreuses approches ont été explorées suivant la structure de dépendance de données. Samanta (1989) énonce des résultats sur la consistance forte et de normalité asymptotique dans le cas des observations indépendantes et identiquement distribuées [3]. Berlinet *et al.* (1998) étendent ces résultats aux données non indépendantes. Ils considèrent le processus (X_i,Y_i) stationnaire et α-mélangeant. Ils énoncent des théorèmes qui établissent qu'un estimateur convergent du quantile $q(.|x)$ construit à partir d'un estimateur convenable de $F(.|x)$ est asymptotiquement normal, quelle que soit la structure de dépendance des données.

Théorème 3.3.1 (Berlinet *et al.* (1998))**.** *Soit $x\in\mathbb{R}^p$ et $(\mathcal{U}_n)_{n\geq 1}$ une suite de réels. On suppose que $F(.|x)$ admet un quantile conditionnel unique $q(.|x)$ d'ordre $\alpha\in(0,1)$. On suppose que $f(.|x)$, $\hat{f}_n(.|x)$ et $\hat{F}_n(.|x)$ existent, que $f(.|x)$ est continue et que $f(q(\alpha|x)|x)\neq 0$. Si, pour n tendant vers l'infini*

1. *$\hat{q}_n(\alpha|x)$ converge presque sûrement vers $q(\alpha|x)$,*
2. *$\mathcal{U}_n\left(\hat{F}_n\left(q(\alpha|x)|x\right)-F(q(\alpha|x)|x)\right)$ converge en loi vers $\mathcal{N}\left(m_q(x),\sigma_q^2(x)\right)$,*
3. *$\hat{f}_n(.|x)$ converge uniformément en probabilité vers $f(.|x)$ sur un voisinage de $\mathcal{V}(q(.|x))$,*

2. Il s'agit de la relation classique entre la fonction de répartition conditionnelle et ses deux densités.
3. L'auteur travaille sur des données univariées

alors

$$\mathscr{U}_n\left(\hat{q}_n(\alpha|x) - q(\alpha|x)\right) \xrightarrow{\mathscr{L}} \mathscr{N}\left(\frac{m_q(x)}{f(q(\alpha|x)|x)}, \frac{\sigma_q^2(x)}{f^2(q(\alpha|x)|x)}\right).$$

Si l'on suppose que $\{(X_i, Y_i),\ i = 1, \ldots, n\}$ sont des copies indépendantes du couple aléatoire $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$, alors le Théorème 3.3.1 est verifié avec $\mathscr{U}_n = \left(nh_n^p\right)^{1/2}$, $m_q(x) = 0$ et $\sigma_q^2(x) = \dfrac{\alpha(1-\alpha)}{g(x)}\|K\|_2^2$ (voir Berlinet *et al.*, 2001, Théorème 6.3). Signalons que les auteurs ont aussi établi la convergence presque complète de leurs estimateurs.

Lorsque X est de nature fonctionnelle, Ferraty *et al.* (2005) proposent de modifier l'estimateur (3.3) dans le cas des données non nécessairement dépendantes. L'estimateur ainsi défini est donné par

$$\hat{F}_n(y|x) = \frac{\displaystyle\sum_{i=1}^n K_0\left(\frac{d(x, X_i)}{h_{n,0}}\right)\int_{-\infty}^y K_1\left(\frac{z - Y_i}{h_{n,1}}\right)dz}{\displaystyle\sum_{i=1}^n K_0\left(\frac{d(x, X_i}{h_{n,0}}\right)}$$

où d est une distance semi-métrique, K_1 un noyau positif et décroissant sur $[0, 1]$, K_0 un noyau symétrique, $h_{n,0}$ et $h_{n,1}$ deux suites positives. Les auteurs établissent la vitesse de convergence presque complète de l'estimateur de quantile conditionnel qui en découle. Pour plus de détails, on pourra aussi consulter (Ferraty et Vieu, 2006, chapitres 5 et 6).

(c)- L'estimateur de la fenêtre mobile

Encore appelé médianogramme, son ancêtre le régressogramme fut introduit par Tukey (1961). Cet estimateur est valable quel que soit le design. Pour un réel fixé $h_n > 0$, on se place en un point fixé x et on sélectionne les seuls Y_i pour lesquels les points d'observations X_i (ou x_i) appartiennent à la boule centrée en x et de rayon h_n. Ceci revient alors à estimer la fonction de répartition conditionnelle par (3.2) en posant :

$$w_{ni}(x) = \mathbb{1}_{\{X_i \in B(x,h_n)\}} \Big/ \sum_{i=1}^n \mathbb{1}_{\{X_i \in B(x,h_n)\}} \text{ (design aléatoire)},$$

$$w_{ni}(x) = \mathbb{1}_{\{x_i \in B(x,h_n)\}} \Big/ \sum_{i=1}^n \mathbb{1}_{\{x_i \in B(x,h_n)\}} \text{ (design fixe)},$$

où $B(x, h_n)$ est une boule centrée en x et de rayon $h_n \to 0$ quand $n \to \infty$. L'estimateur du quantile conditionnel est alors construit en inversant l'estimateur de la fonction de répartition empirique du sous-échantillon des observations dont les covariables sont dans la boule. On peut alors remarquer que l'estimateur de la fenêtre mobile est un estimateur à noyau particulier correspondant au cas où le noyau est $K(x) = \mathbb{1}_{\{x \in [-1;1]\}}$.

L'estimateur de la médiane conditionnelle est construit suivant cette méthode. Truong (1989) a donné la vitesse optimale de la convergence d'un estimateur empirique de la médiane conditionnelle. Gannoun (1989) a établi la convergence uniforme presque complète de cet estimateur.

3.3.1.2 Estimateurs au sens des plus proches voisins

La méthode de construction de ces estimateurs est analogue à ceux de la fenêtre mobile. Bhattacharyya et Gangopadhyay (1990) définissent l'estimateur de la fonction de répartition conditionnelle de Y sachant $X = x$ au sens des k *plus proches voisins* comme étant la fonction de répartition empirique des observations aux k points X_i (ou x_i) les plus proches du point fixe x, i.e

$$\hat{F}_k(y|x) = \frac{1}{k}\sum_{i=1}^{k} \mathbb{1}_{\{Y_{\lfloor ni \rfloor} \leq y\}},$$

où k est un entier inférieur ou égal à n et les $Y_{\lfloor ni \rfloor}$ sont les concomitants des statistiques d'ordres $U_{n,1} < \ldots < U_{n,n}$ associés aux $U_i = |X_i - x|$ (ou $U_i = |x_i - x|$). Le quantile conditionnel d'ordre α au sens des k *plus proches voisins* est alors défini par :

$$\hat{q}_k(\alpha|x) = \inf\left\{ y \mid \hat{F}_k(y|x) \geq \frac{\lfloor k\alpha \rfloor}{k} \right\}.$$

Les auteurs montrent que leur estimateur de quantile conditionnel est asymptotiquement gaussien pour $k = \lfloor n^{4/5} s \rfloor$ où $s \in [a, b]$ avec $0 < a < b$.

3.3.2 Méthode d'estimation directe

Elle consiste à ramener le problème d'estimation des quantiles conditionnels à un problème d'optimisation

$$(\mathscr{P}) : \qquad q(\alpha|x) = \underset{\theta \in \mathbb{R}}{\arg\min}\, \mathbb{E}\left[\rho_\alpha(Y - \theta)|X = x\right],$$

où la fonction ρ_α est définie par $\rho_\alpha(y) = \alpha y \mathbb{1}_{[0,\infty[} - (1-\alpha) y \mathbb{1}_{]-\infty,0[}$. La théorie sous-jacente de cette technique d'estimation est décrite dans Bassett et Koenker (1982).

Contrairement aux méthodes d'estimation indirectes, on n'a pas besoin de déduire le comportement asymptotique du quantile conditionnel de la convergence de l'estimateur de la fonction de répartition. Dans la littérature, on a répertorié trois méthodes d'estimations directes à savoir :

- La méthode d'estimation par la constante locale qui est un cas particulier de la méthode d'estimation par les polynômes locaux.
- La méthode d'estimation par des fonctions splines de régression développée par He et Shi (1994). Ils proposent d'approcher la fonction quantile par une fonction spline. Cardot *et al.* (2004, 2005) généraliseront cet estimateur au cas où la variable explicative X est de type fonctionnel.

- La méthode d'estimation par des fonctions splines de lissage proposée par Koenker *et al.* (1994). Elle avait été spécialement developpée pour *le design fixe*.

Dans les lignes qui suivent, on ne détaillera pas les deux dernières méthodes. Le lecteur curieux pourra se référer au travail de synthèse de Poiraud-Casanova et Thomas-Agnan (1998).

3.3.2.1 Méthode de la constante locale

La technique consiste à approcher le quantile par une fonction linéaire

$$q(\alpha|z) \approx q(\alpha|x) + q'(\alpha|x)(z-x) \overset{def}{=} a + b(z-x),$$

pour z dans un voisinage de x. De façon naturelle, ceci revient à utiliser le principe des moindre carrés, i.e

$$\min_{(a,b)} \sum_{i=1}^{n} \rho_\alpha (Y_i - a + b(X_i - x)) K\left(\frac{x - X_i}{h_n}\right).$$

Si $b = 0$, on parle de la méthode dite de la constante locale introduite par Tsybakov (1986) et on estime le quantile conditionnel par :

$$\hat{q}_n(\alpha|x) = \arg\min_{a\in\mathbb{R}} \sum_{i=1}^{n} \rho_\alpha (Y_i - a) K\left(\frac{x - X_i}{h_n}\right).$$

Les résultats sur la consistance faible et la convergence en moyenne quadratique de cet estimateur sont dus à Stone (1977) et Yu et Jones (1998). Fan *et al.* (1994) établissent la convergence en loi. La convergence uniforme sur un support compact peut être obtenue en adaptant le résultat de Berlinet *et al.* (2001) (se référer à Gannoun *et al.* (2002)). Par ailleurs, cette méthode se comporte bien face aux effets de bord (Fan *et al.*, 1994; Koenker *et al.*, 1992; Mint El Mouvit, 2000).

3.3.2.2 Estimation par des polynômes locaux

Plus générale que la méthode de la constante locale, elle consiste à adapter la méthode des polynômes locaux classique pour la moyenne conditionnelle aux quantiles conditionnels. Ainsi, si l'on approche le quantile conditionnel par un polynôme de degré k, cela revient à résoudre le problème d'optimisation

$$\min_{(b_0,\ldots,b_k)} \sum_{i=1}^{n} \rho_\alpha \left(Y_i - \sum_{j=0}^{k} b_j (X_i - x)^j\right) K\left(\frac{x - X_i}{h_n}\right).$$

L'estimateur de quantile conditionnel est alors défini par :

$$\hat{q}_n(\alpha|x) = \hat{b}_0.$$

Chauduri (1991) a démontré la consistance forte ponctuelle dans le cas d'un noyau de type fonction indicatrice d'intervalles. Pour $k = 1$, on retrouve l'estimateur de la constante locale présenté précédemment.

4

Estimation des quantiles extrêmes conditionnels en design fixe

Résumé

Dans ce chapitre, nous proposons dans le cas des lois à queues lourdes une méthode d'estimation des quantiles extrêmes en présence d'une covariable de dimension finie ou infinie (i.e covariable fonctionnelle). La loi limite des estimateurs ainsi construits est ensuite donnée en fonction de la vitesse de convergence de l'ordre du quantile vers un. Pour conclure, des simulations et des illustrations seront présentées sur données réelles.

Sommaire

4.1 Introduction **54**
4.2 Contexte d'étude et définitions des estimateurs **55**
4.2.1 Contexte d'étude et méthode d'estimation 55
4.2.2 Estimateurs des quantiles extrêmes conditionnels 58
4.3 Étude théorique des estimateurs **59**
4.3.1 Hypothèses 59
4.3.2 Étude du comportement asymptotique des estimateurs 60
4.4 Exemples et discussion **62**
4.4.1 Quelques exemples d'estimateurs de $\gamma(x)$ 63
4.4.2 Applications et discussion 64
4.5 Simulations et illustration sur données réelles **67**
4.5.1 Simulations 67
4.5.2 Illustration sur données réelles 74
4.6 Démonstrations **77**
4.6.1 Résultats préliminaires 77
4.6.2 Preuve des résultats théoriques 79

4.1 Introduction

DAns la littérature, on dénombre plusieurs méthodes d'estimation des quantiles extrêmes conditionnels. Smith (1989) et Davison et Smith (1990) proposent des familles de modèles paramétriques basées sur les excès au dessus du seuil. Beirlant et Goegebeur (2003) en adoptant une approche semi-paramétrique proposent de transformer tout d'abord les données, puis de les utiliser dans un modèle de régression exponentiel où les paramètres dudit modèle sont estimés par la méthode du maximum de vraisemblance. Hall et Tajvidi (2000) proposent dans le cas d'une série temporelle $\{(Y_i, t_i), i = 1, .., n\}$ où t_i est le temps, de combiner l'estimation non-paramétrique de la tendance temporelle avec une hypothèse sur la distribution conditionnelle de Y_i sachant t_i. L'estimation non-paramétrique des quantiles extrêmes conditionnels a été introduite à notre connaissance dans Davison et Ramesh (2000) où les auteurs proposent des estimateurs par ajustement polynomial. Beirlant et Goegebeur (2004) étendent ces résultats aux covariables multidimensionnelles. En outre, ils donnent les propriétés asymptotiques des estimateurs qui en découlent. Dans le cas d'une covariable unidimensionnelle, Beirlant *et al.* (2004a) se proposent d'adapter les estimateurs de quantiles proposés dans Beirlant et Matthys (2001) et Beirlant et Matthys (2003) au cas conditionnel en remplaçant tout simplement les statistiques d'ordres par les quantiles estimés par la méthode des polynômes locaux (voir Koenker et Bassett, 1978). Plus récemment, Chavez-Demoulin et Davison (2005) en utilisant la méthode du maximum de vraisemblance pénalisé proposent des estimateurs splines de quantiles extrêmes conditionnels dans le cas d'une covariable unidimensionnelle. D'autres méthodes d'estimation ont été proposées par Chernozhukov (1998, 2001) et Hall et Ronde (2000).

Il existe de nombreux exemples en chimiométrie ou en astrophysique où les covariables sont des courbes. On parle alors de *covariables fonctionnelles*. À notre connaissance, aucun auteur ne s'est encore attardé sur l'estimation des quantiles extrêmes conditionnels pour des jeux de données présentant cette particularité. Notons que dans un tel contexte, l'estimation du quantile requiert des techniques de lissage non-paramétrique adaptées aux données fonctionnelles afin de mieux prendre en compte la covariable (Ramsay et Silverman, 1997; Bosq, 2000; Ferraty et Vieu, 2006; Ramsay et Silverman, 2002).

Dans la suite du chapitre, nous nous attarderons sur cinq points. Ainsi, nous commencerons par présenter, dans la partie 4.2, le contexte d'étude puis notre méthode d'estimation et enfin nos estimateurs de quantiles extrêmes. Dans la partie 4.3, nous établirons la loi limite de ces estimateurs. En particulier, c'est dans celle-ci que seront énumérées les hypothèses servant à l'étude théorique. Dans la partie 4.4, nous présenterons d'abord quelques exemples d'estimateurs de l'indice de queue conditionnel. Ensuite, nous les appliquerons à nos estimateurs de quantiles extrêmes conditionnels. Enfin, nous illustrerons le biais de quelques lois usuelles à queues lourdes. Dans la partie 4.5, nous nous focaliserons sur les simulations puis nous présenterons une application concrète de nos résultats sur un exemple d'hydrologie. Enfin, la partie 4.6 sera

dédiée aux preuves.

4.2 Contexte d'étude et définitions des estimateurs

Cette partie se subdivise en deux paragraphes. Le paragraphe 4.2.1 est dédié au contexte d'étude et à la présentation de notre méthode d'estimation des quantiles extrêmes conditionnels. Le paragraphe 4.2.2 est consacré à la présentation *stricto sensu* des estimateurs.

4.2.1 Contexte d'étude et méthode d'estimation

Soit $Y \in \mathbb{R}$ une variable aléatoire associée à une covariable non-aléatoire $x \in E$, où E désigne un espace métrique, non nécessairement de dimension finie, muni d'une distance d. D'une manière précise, le problème est le suivant. Soient $\{(x_i, Y_i),\ i = 1, \dots, n\}$ des observations indépendantes du couple $(x, Y) \in E \times \mathbb{R}$, on veut estimer le quantile d'ordre $\alpha_n \to 0$ quand $n \to \infty$ dans le cas particulier où la fonction de répartition conditionnelle de Y sachant x et notée $F(y, x)$ est à queue lourde. Ceci signifie que pour tout $y > 0$,

$$F(y, x) = 1 - y^{-1/\gamma(x)} \ell(y, x), \tag{4.1}$$

où $\gamma(.)$ est une fonction positive et inconnue de la covariable x que l'on appelle « *indice de queue conditionnel* » ou « *indice des valeurs extrêmes conditionnel* » et pour tout x fixé, $\ell(., x)$ est une fonction à variations lentes à l'infini (c.f Définition 1.3.1), i.e pour tout $\lambda > 0$,

$$\lim_{y \to \infty} \frac{\ell(\lambda y, x)}{\ell(y, x)} = 1.$$

Ceci revient donc à supposer que pour tout x fixé, la fonction de répartition conditionnelle $F(., x) \in \mathscr{D}(Fréchet)$. Dans ce cas, pour tout $x \in E$, le quantile conditionnel d'ordre $(1 - \alpha_n)$ et noté $q(\alpha_n, x)$ est à décroissance polynomiale d'indice $-\gamma(x)$ (Bingham *et al.*, 1987), i.e pour tout $\lambda > 0$,

$$\lim_{\alpha_n \to 0} \frac{q(\lambda \alpha_n, x)}{q(\alpha_n, x)} = \lambda^{-\gamma(x)}.$$

On dit aussi que le quantile conditionnel $q(., x)$ est à variations régulières d'indice $-\gamma(x)$.

Dans l'optique de combiner des techniques de lissage non-paramétrique avec des méthodes d'analyse de valeurs extrêmes dont le but est d'obtenir des estimateurs ayant de bonnes propriétés asymptotiques, nous utilisons une méthode d'estimation dite de la *fenêtre mobile* pour construire nos estimateurs. Pour cela, on introduit une boule centrée en x, de rayon $r > 0$, notée $B(x, r)$ et définie par :

$$B(x, r) = \{t \in E : d(x, t) \leq r\}.$$

Étant donné $(r_{n,x})_{n\geq 1}$ une suite positive convergeant vers zéro quand n tend vers l'infini, on se propose de ne sélectionner que les observations Y_i pour lesquelles les covariables x_i sont dans la boule $B(x, r_{n,x})$. La proportion de tels points est ainsi donnée par

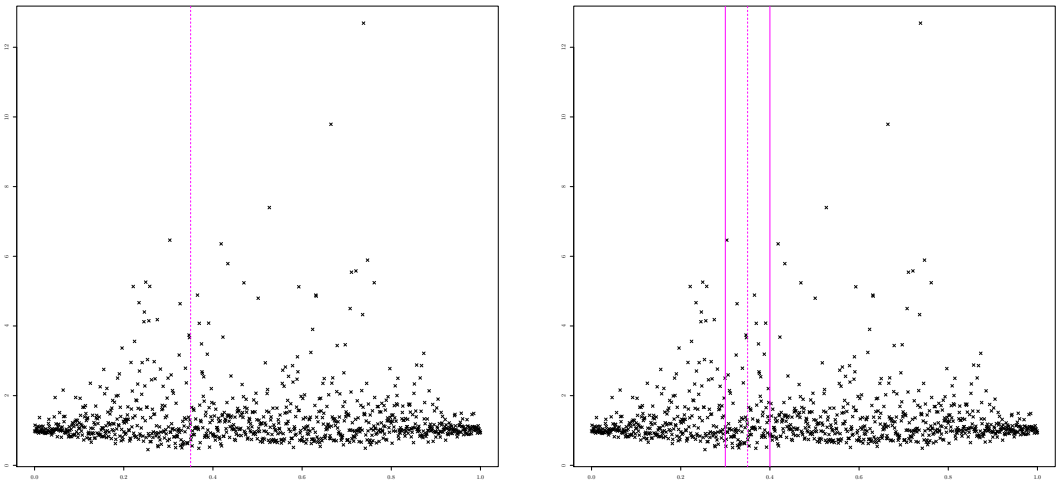
$$\varphi(r_{n,x}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{x_i \in B(x, r_{n,x})\}},$$

et joue un rôle central dans cette étude. De façon similaire à la notion de probabilité de petite boule utilisée en analyse fonctionnelle dans Ferraty et Vieu (2006), $\varphi(r_{n,x})$ décrit comment cet ensemble de points se concentre dans un voisinage de x lorsque $r_{n,x} \to 0$. Dans la suite, $m_{n,x}$ désignera le nombre d'observations dans la boule $B(x, r_{n,x})$. Nous noterons par $\{Z_i(x), i = 1, \ldots, m_{n,x}\}$, les observations de Y retenues par la procédure de sélection et nous désignerons par $Z_{1,m_{n,x}}(x) \leq \ldots \leq Z_{m_{n,x},m_{n,x}}(x)$ les statistiques ordonnées correspondantes. Dans un tel contexte, il apparait que dans la région $]0,\infty[\times B(x, r_{n,x})$, le nombre d'observations retenues par la procédure de sélection est lié à l'ensemble $\varphi(r_{n,x})$ par la relation $m_{n,x} = n\varphi(r_{n,x})$. Les graphiques de la Figure 4.1 illustrent notre méthode d'estimation dans le cas d'une covariable unidimensionnelle de $E = [0, 1]$.

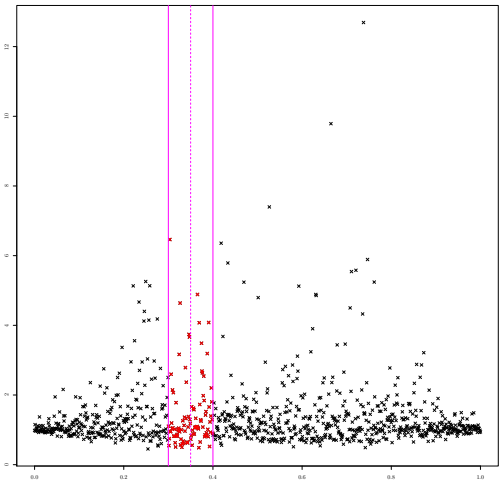
L'avantage de notre approche par rapport à celles présentées à la partie introductive du chapitre est que très peu d'hypothèses sont faites sur la régularité de la fonction $q(., x)$ ou $\gamma(x)$ [1] et sur la nature de la covariable x. Des résultats sur le comportement asymptotique des nos estimateurs de quantiles extrêmes conditionnels sont établis sans faire d'hypothèse sur la dimension de E. D'un point de vue pratique, nos estimateurs sont faciles à calculer car ils ne requièrent aucune technique d'optimisation.

Une approche similaire à notre méthode d'estimation et basée sur *les plus proches voisins* a été développée par Gardes et Girard (2010).

1. Par exemple, Beirlant et Goegebeur (2004) suppose que la fonction indice de queue conditionnel doit être au moins deux fois continûment dérivable.



Étape 1 **Étape 2**



Étape 3

FIGURE 4.1 – Les différentes étapes de notre procédure de sélection. En ordonnée on a la variable d'intérêt Y et en abscisse la covariable x . Pour construire des estimateurs en un point x on procède comme suit.
Étape 1 : On se place au point x (dans cet exemple $x = 0.37$) qui sera le centre de la boule.
Étape 2 : On fixe le rayon $r_{n,x}$ de boule de centre x : $B(x, r_{n,x})$.
Étape 3 : On ne sélectionne que les observations Y_i (★★★) pour lesquelles $x_i \in B(x, r_{n,x})$.

4.2.2 Estimateurs des quantiles extrêmes conditionnels

Dans ce chapitre, on veut construire en tout point $x \in E$, un estimateur du quantile extrême conditionnel d'ordre $1 - \alpha_{m_{n,x}}$. Ici, on parle de quantile extrême si on a $\alpha_{m_{n,x}}$ qui converge vers zéro quand $m_{n,x}$ tend vers l'infini. En fonction de la vitesse de convergence de $\alpha_{m_{n,x}}$ vers 0, trois situations sont envisagées :

(S.1) $\alpha_{m_{n,x}}$ converge *lentement* vers 0, i.e $\alpha_{m_{n,x}} \to 0$ et $m_{n,x}\alpha_{m_{n,x}} \to \infty$ lorsque $m_{n,x} \to \infty$.

(S.2) $\alpha_{m_{n,x}}$ converge *rapidement* vers 0, i.e $\alpha_{m_{n,x}} \to 0$, $m_{n,x}\alpha_{m_{n,x}} \to c \in [1,\infty[$ et $\lfloor m_{n,x}\alpha_{m_{n,x}} \rfloor \to \lfloor c \rfloor$ lorsque $m_{n,x} \to \infty$.

(S.3) $\alpha_{m_{n,x}}$ converge *très rapidement* vers 0, i.e $\alpha_{m_{n,x}} \to 0$ et $m_{n,x}\alpha_{m_{n,x}} \to c \in [0,1[$ lorsque $m_{n,x} \to \infty$.

Dans la situation **(S.1)**, $\alpha_{m_{n,x}}$ converge moins vite vers zéro que le rapport $1/m_{n,x}$. Par conséquent, l'estimation du quantile extrême conditionnel requiert d'interpoler à l'intérieur de l'échantillon car $q(\alpha_{m_{n,x}}, x)$ est presque sûrement inférieur à l'observation maximale (voir Proposition 4.3.2). On propose alors d'estimer $q(\alpha_{m_{n,x}}, x)$ par

$$\hat{q}_1(\alpha_{m_{n,x}}, x) = Z_{m_{n,x} - \lfloor m_{n,x}\alpha_{m_{n,x}} \rfloor + 1, m_{n,x}}(x). \tag{4.2}$$

Dans la situation intermédiaire **(S.2)**, pour n assez grand on a $\lfloor m_{n,x}\alpha_{m_{n,x}} \rfloor = \lfloor c \rfloor > 0$. Ainsi, l'estimation du quantile extrême conditionnel repose sur les plus grandes observations situées au voisinage la frontière de l'échantillon, mais toujours dans l'ensemble des données. Par conséquent, on peut réutiliser l'estimateur défini en (4.2).

Dans la situation **(S.3)**, $\alpha_{m_{n,x}}$ converge au moins aussi vite vers zéro que le rapport $1/m_{n,x}$. Estimer le quantile extrême conditionnel nécessite d'extrapoler au-delà des observations puisque $q(\alpha_{m_{n,x}}, x)$ est supérieur à l'observation maximale avec probabilité $e^{-c} \geq e^{-1}$ (voir Proposition 4.3.2). Dans une telle situation, on propose d'adapter l'estimateur de Weissman (1978) au cas conditionnel. On estime alors $q(\alpha_{m_{n,x}}, x)$ par

$$\hat{q}_2(\alpha_{m_{n,x}}, x) = \hat{q}_1(\beta_{m_{n,x}}, x) \left(\beta_{m_{n,x}} / \alpha_{m_{n,x}}\right)^{\hat{\gamma}_n(x)}, \tag{4.3}$$

où $\beta_{m_{n,x}}$ satisfait la situation **(S.1)** et $\hat{\gamma}_n(x)$ est un estimateur de l'indice des valeurs extrêmes conditionnel.

Les situations **(S.1)**, **(S.2)** et **(S.3)** ont déjà été étudiées dans le cas non conditionnel. de Haan (1984) en énonce le premier résultat dans la situation **(S.3)** en posant $c = 0$. Dekkers et de Haan (1989) étudient les situations **(S.1)** et **(S.3)** avec $c \neq 0$. Leurs résultats peuvent être consultés dans (Embrechts *et al.*, 1997, Théorèmes 6.4.14 et 6.4.15). Dans la situation **(S.2)**, si la constante c n'est pas entière alors, $m_{n,x}\alpha_{m_{n,x}} \to c$ implique que $\lfloor m_{n,x}\alpha_{m_{n,x}} \rfloor \to \lfloor c \rfloor$ quand $m_{n,x} \to \infty$. Sinon, si c est un entier, alors la condition $\lfloor m_{n,x}\alpha_{m_{n,x}} \rfloor \to \lfloor c \rfloor$ est nécessaire car elle empêche à la suite $\left(\lfloor m_{n,x}\alpha_{m_{n,x}} \rfloor\right)_{m_{n,x}>0}$ d'avoir deux valeurs d'adhérence.

4.3 Étude théorique des estimateurs

Cette partie est subdivisée en deux paragraphes. Dans le premier on donne nos hypothèses de travail et dans le second on établit les lois asymptotiques. Dans tout ce qui suit, on suppose que $x \in E$.

4.3.1 Hypothèses

Dans ce paragraphe, on énumère les hypothèses nous permettant d'étudier le comportement asymptotique de nos estimateurs de quantiles extrêmes conditionnels.

Hypothèse sur le quantile conditionnel

(A.1) La fonction quantile conditionnel

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

est dérivable et la fonction biais

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

est continue et telle que : $\lim_{\alpha \to 0} \Delta(\alpha, x) = 0$.

L'hypothèse **(A.1)** a pour but de contrôler le comportement de la fonction log-quantile quant à sa première variable. C'est une condition suffisante pour que la fonction de répartition conditionelle $F(., x)$ soit à queue lourde (Bingham *et al.*, 1987). Afin de simplifier les notations, pour tout $a \in]0,1[$ on notera :

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

Il parait important de préciser que dans **(A.1)**, on suppose implicitement que pour tout $x \in E$ ($dim\ E \leq \infty$) la fonction à variation lentes $\ell(., x)$ est dérivable (se référer au sous-paragraphe 1.3.1.1) .

Hypothèse sur le paramètre de lissage

(A.2) Soit $(k_{n,x})_{n \geq 1}$ une suite d'entiers telle que $1 \leq k_{n,x} < m_{n,x}$, on suppose que

$$n\varphi(r_{n,x})/k_{n,x} \to \infty \text{ et } k_{n,x} \to \infty \text{ quand } n \to \infty.$$

Cette hypothèse implique que $n\varphi(r_{n,x}) \to \infty$, c'est-à-dire que le nombre d'observations retenues par notre méthode de fenêtres mobiles tend vers l'infini lorsque la taille de l'échantillon tend vers l'infini.

4.3.2 Étude du comportement asymptotique des estimateurs

Cette partie est consacrée à l'étude théorique de nos estimateurs. Pour cela, il convient de donner quelques résultats auxiliaires utiles pour établir leur loi limite. Cependant, il apparait nécessaire de commencer par une définition de l'oscillation de la fonction log-quantile.

Définition 4.3.1. *Pour tout $a \in (0,1/2)$, la plus grande oscillation de la fonction log-quantile par rapport à sa seconde variable est donnée par :*

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

Notre premier résultat est dédié à l'étude de la représentation en loi des plus grandes variables aléatoires de l'échantillon $\{Z_i(x),\ i=1,\dots,m_{n,x}\}$ retenues par notre procédure de sélection.

Proposition 4.3.1. *Soit $J_{k_{n,x}} = \{1,\dots,k_{n,x}\}$. Sous* **(A.1)** *et* **(A.2)**, *si $k_{n,x}^2\omega_n\left(m_{n,x}^{-(1+\delta)}\right) \to 0$ pour certain $\delta > 0$, alors, il existe un événement $\mathscr{A}_n$ avec une probabilité convergeant vers un (i.e $\mathbb{P}(\mathscr{A}_n) \to 1$) quand n tend vers l'infini tel que :*

$$\left\{\left(\log Z_{m_{n,x}-i+1,m_{n,x}}(x), i \in J_{k_{n,x}}\right)|\mathscr{A}_n\right\} \overset{\mathscr{L}}{=} \left\{\left(\log q(V_{i,m_{n,x}}, T_i), i \in J_{k_{n,x}}\right)|\mathscr{A}_n\right\},$$

où $V_{1,m_{n,x}} \le \dots \le V_{m_{n,x},m_{n,x}}$ sont les statistiques ordonnées associées à la suite de variables aléatoires de loi uniforme standard $\{V_i,\ i=1,\dots,m_{n,x}\}$ et $\{T_i,\ i=1,\dots,k_{n,x}\}$ sont des variables aléatoires appartenant à la boule $B(x,r_{n,x})$.

La condition $k_{n,x}^2\omega_n\left(m_{n,x}^{-(1+\delta)}\right) \to 0$ montre qu'il est plus facile de contrôler les $k_{n,x}$ plus grandes observations dans la région $]0,\infty[\times B(x,r_{n,x})$ lorsque l'oscillation de la fonction log-quantile par rapport à sa seconde variable est petite. Un rapprochement peut être fait entre notre Proposition et l'approximation utilisée par (Falk *et al.*, 2004, Théorème 3.5.2) dans l'étude de la loi des k-plus proches voisins en utilisant la distance de Hellinger.

Donnons à présent un résultat sur la position du quantile extrême conditionnel $q(\alpha_{m_{n,x}},x)$ dans l'ensemble des données.

Proposition 4.3.2. *Sous* **(A.1)**, *si $\omega_n\left(m_{n,x}^{-(1+\delta)}\right) \to 0$ pour un certain $\delta > 0$, alors*

- *dans la situation* **(S.1)**, $\mathbb{P}\left(Z_{m_{n,x},m_{n,x}} < q(\alpha_{m_{n,x}},x)\right) \to 0$,
- *dans la situation* **(S.2)** *ou* **(S.3)**, $\mathbb{P}\left(Z_{m_{n,x},m_{n,x}} < q(\alpha_{m_{n,x}},x)\right) \to e^{-c}$.

Les résultats suivants établissent la loi limite d'un estimateur du quantile extême conditionnel construit à partir de notre procédure d'estimation. Focalisons nous tout d'abord à l'estimation du quantile conditionnel dans la situation **(S.1)**.

Théorème 4.3.1. *Soit* $(\alpha_{m_{n,x}})_{n\geq 1}$ *une suite satisfaisant la situation* **(S.1)**. *Sous* **(A.1)**, *si pour n tendant vers l'infini, il existe* $\delta > 0$ *tel que* $\left(m_{n,x}\alpha_{m_{n,x}}\right)^2 \omega_n \left(m_{n,x}^{-(1+\delta)}\right) \to 0$ *alors,*

$$\left(m_{n,x}\alpha_{m_{n,x}}\right)^{1/2}\left(\frac{\hat{q}_1(\alpha_{m_{n,x}},x)}{q(\alpha_{m_{n,x}},x)}-1\right) \xrightarrow{\mathscr{L}} \mathscr{N}\left(0,\gamma^2(x)\right).$$

Dans cette situation, il apparaît que l'estimateur est asymptotiquement gaussien avec une variance asymptotique proportionnelle à $\gamma^2(x)/(m_{n,x}\alpha_{m_{n,x}})$. Ainsi, plus grande est la valeur de l'indice de queue conditionnel, c'est-à-dire plus lourde est la queue de la loi, et plus grande est la variance de notre estimateur de quantile extrême conditionnel. De plus, la variance asymptotique étant inversement proportionnelle à $\alpha_{m_{n,x}}$, l'estimation de quantile extrême est d'autant plus stable que l'on s'éloigne de la frontière de l'échantillon.

Considérons maintenant la situation intermédiaire **(S.2)** dont l'estimation du quantile repose sur les plus grandes observations situées au voisinage de la frontière de l'échantillon.

Théorème 4.3.2. *Soit* $(\alpha_{m_{n,x}})_{n\geq 1}$ *une suite satisfaisant la situation* **(S.2)**. *Sous* **(A.1)**, *si pour n tendant vers l'infini, il existe* $\delta > 0$ *tel que* $\left(m_{n,x}\alpha_{m_{n,x}}\right)^2 \omega_n \left(m_{n,x}^{-(1+\delta)}\right) \to 0$ *alors,*

$$\left(\frac{\hat{q}_1(\alpha_{m_{n,x}},x)}{q(\alpha_{m_{n,x}},x)}-1\right) \xrightarrow{\mathscr{L}} \mathscr{E}\left(c,\gamma(x)\right),$$

où $\mathscr{E}(c,\gamma(x))$ *est une loi non dégénérée.*

Dans la situation **(S.2)**, la loi asymptotique du quantile extrême conditionnel n'est pas gaussienne et son expression est assez compliquée. En outre, l'estimateur $\hat{q}_1(.,x)$ n'est pas consistant.

La dernière situation **(S.3)** est beaucoup plus intéressante mais aussi complexe puisqu'elle fait intervenir d'une part, l'estimateur de quantile $\hat{q}_1(.,x)$ de la situation **(S.1)** et d'autre part, l'estimateur de l'indice de queue conditionnel $\hat{\gamma}_n(x)$. Ainsi, sa loi asymptotique peut provenir de l'un ou de l'autre.

Théorème 4.3.3. *Soit* $(\beta_{m_{n,x}})_{n\geq 1}$ *une suite satisfaisant la situation* **(S.1)** *et soit* $(\alpha_{m_{n,x}})_{n\geq 1}$ *une suite telle que* $\alpha_{m_{n,x}} < \beta_{m_{n,x}}$. *On pose* $\zeta_{m_{n,x}} = \left(m_{n,x}\beta_{m_{n,x}}\right)^{1/2}\log(\beta_{m_{n,x}}/\alpha_{m_{n,x}})$. *Sous* **(A.1)**, *si pour n tendant vers l'infini, il existe* $\delta > 0$ *tel que* $\left(m_{n,x}\beta_{m_{n,x}}\right)^2 \omega_n \left(m_{n,x}^{-(1+\delta)}\right) \to 0$ *et s'il existe une suite positive* $\mathscr{V}_n(x)$ *tendant vers l'infini et une loi* $\mathscr{D}$ *telle que :*

$$\mathscr{V}_n(x)\left(\hat{\gamma}_n(x)-\gamma(x)\right) \xrightarrow{\mathscr{L}} \mathscr{D}, \tag{4.4}$$

alors, deux situations se présentent :

(i) Sous la condition additionnelle

$$\zeta_{m_{n,x}} \max\left\{\mathscr{V}_n^{-1}(x), \bar{\Delta}\left(\beta_{m_{n,x}},x\right)\right\} \to 0, \tag{4.5}$$

nous avons

$$\left(m_{n,x}\beta_{m_{n,x}}\right)^{1/2}\left(\frac{\hat{q}_2(\alpha_{m_{n,x}},x)}{q(\alpha_{m_{n,x}},x)}-1\right)\xrightarrow{\mathscr{L}}\mathscr{N}\left(0,\gamma^2(x)\right).$$

(ii) Sous la condition additionnelle

$$\mathcal{V}_n(x)\max\left\{\zeta^{-1}_{m_{n,x}},\bar{\Delta}\left(\beta_{m_{n,x}},x\right)\right\}\to 0, \tag{4.6}$$

nous avons

$$\frac{\mathcal{V}_n(x)}{\log(\beta_{m_{n,x}}/\alpha_{m_{n,x}})}\left(\frac{\hat{q}_2(\alpha_{m_{n,x}},x)}{q(\alpha_{m_{n,x}},x)}-1\right)\xrightarrow{\mathscr{L}}\mathscr{D}.$$

De ce Théorème, on remarque que la loi asymptotique de $\hat{q}_2(.,x)$ dépend d'une part du comportement de $\hat{q}_1(.,x)$ et d'autre part du comportement de $\hat{\gamma}_n(x)$. Dans la situation (i), la normalité asymptotique de $\hat{q}_2(.,x)$ est donnée par la statistique d'ordre $\hat{q}_1(.,x)$ qui n'est rien d'autre que la $m_{n,x}\alpha_{m_{n,x}}$ ième plus grande observation parmi les $\{Z_i(x), i=1,\dots,m_{n,x}\}$ observations contenues dans la boule $B(x,r_{n,x})$. À l'opposé, dans la situation (ii), $\hat{q}_2(.,x)$ hérite de la loi limite de l'estimateur de l'indice de queue conditionnel.

La loi asymptotique de l'estimateur de quantile extrême conditionnel fait apparaitre un biais asymptotique dont l'ordre (en valeur absolue) est donné par[2] :

$$\left|\int_{\alpha_{m_{n,x}}}^{\beta_{m_{n,x}}}\frac{\Delta(u,x)}{u}du\right|\leq\log\left(\frac{\beta_{m_{n,x}}}{\alpha_{m_{n,x}}}\right)\bar{\Delta}\left(\beta_{m_{n,x}},x\right).$$

Ainsi, dans la situation (i), la condition $\zeta_{m_{n,x}}\bar{\Delta}\left(\beta_{m_{n,x}},x\right)\to 0$ impose au biais d'être négligeable devant l'écart type de $\hat{q}_2(\alpha_{m_{n,x}},x)$ qui vaut quant à lui $\left(m_{n,x}\beta_{m_{n,x}}\right)^{1/2}$. Dans la situation (ii), la condition $\mathcal{V}_n(x)\bar{\Delta}\left(\beta_{m_{n,x}},x\right)\to 0$, impose au biais d'être négligeable devant l'écart type de $\hat{q}_2(\alpha_{m_{n,x}},x)$ qui est ici égal à $\mathcal{V}_n(x)\big/\log(\beta_{m_{n,x}}/\alpha_{m_{n,x}})$.

Notons toutefois que, même si l'intérêt principal du Théorème 4.3.3 est d'établir le comportement asymptotique du quantile extrême conditionnel dans la situation où $\alpha_{m_{n,x}}$ converge très rapidement vers zéro, il peut également être appliqué dans toutes les situations moins restrictives que la situation **(S.3)**, où $\alpha_{m_{n,x}}$ est plus petit que $\beta_{m_{n,x}}$. Par exemple, il apparaît que, dans la situation **(S.2)**, $\hat{q}_2(\alpha_{m_{n,x}},x)$ est un estimateur faiblement consistant, i.e

$$\frac{\hat{q}_2(\alpha_{m_{n,x}},x)}{q_2(\alpha_{m_{n,x}},x)}\xrightarrow{\mathbb{P}}1.$$

4.4 Exemples et discussion

La partie suivante comprend deux paragraphes. Tandis que le premier expose quelques méthodes d'estimation de l'indice de queue conditionnel, le deuxième s'articule quant-à lui autour de deux points, à savoir des applications (voir sous-paragraphe 4.4.2.1) et des illustrations (voir sous-paragraphe 4.4.2.2).

2. On se pourra se référer à la démonstration du Théorème 4.3.3

4.4.1 Quelques exemples d'estimateurs de $\gamma(x)$

Dans la littérature, on dénombre plusieurs approches d'estimation de l'indice de queue conditionnel. Celles-ci sont pour la majorité inhérentes aux méthodes d'estimations présentées dans l'introduction (voir partie 4.1). Ainsi, on pourra se référer aux travaux de Davison et Smith (1990), Davison et Ramesh (2000), Hall et Tajvidi (2000), Beirlant *et al.* (2002), Beirlant et Goegebeur (2004) et Chavez-Demoulin et Davison (2005) pour de plus amples explications.

En ce qui nous concerne, nous allons présenter la famille d'estimateurs proposée par Gardes et Girard (2008) qui est une extension des estimateurs proposés par Beirlant *et al.* (2002) dans le cas univarié. Les raisons de ce choix sont essentiellement dûs à l'utilisation de la méthode des fenêtres mobiles.

Les estimateurs de l'indice de queue conditionnel que l'on présente ici sont une somme pondérée des écarts de logarithmes entre les plus grandes observations retenues par la méthode de fenêtres mobiles décrite au paragraphe 4.2.1.

Définition 4.4.1. *Soient $Z_{1,m_{n,x}}(x) \leq \ldots \leq Z_{m_{n,x},m_{n,x}}(x)$ les statistiques ordonnées correspondantes aux observations dans la boule $B(x, r_{n,x})$. La famille d'estimateurs définie dans Gardes et Girard (2008) est donnée par*

$$\hat{\gamma}_n(x,W) = \sum_{i=1}^{k_{n,x}} i \log\left(\frac{Z_{m_{n,x}-i+1,m_{n,x}}(x)}{Z_{m_{n,x}-i,m_{n,x}}(x)}\right) W(i/k_{n,x},x) \Bigg/ \sum_{i=1}^{k_{n,x}} W(i/k_{n,x},x)\ ,$$

où $W(.,x)$ est une fonction de poids définie sur $]0,1[$ telle que $\int_0^1 W(s,x)ds \neq 0$.

L'étude de cet estimateur repose sur le modèle de régression exponentielle pour des écarts pondérés de logarithmes entre les $k_{n,x}$ plus grandes observations dans la boule $B(x, r_{n,x})$. Sous certaines hypothèses de régularité de la distribution conditionnelle de Y sachant x et sous certaines conditions sur la fonction de poids, les auteurs montrent que leur estimateur est asymptotiquement gaussien avec un biais dépendant à la fois de la distribution théorique et de la fonction de poids $W(.,x)$:

$$k_{n,x}^{1/2}\left(\hat{\gamma}_n(x,W) - \gamma(x) - \Delta\left(k_{n,x}/m_{n,x},x\right)\mathscr{AB}(x,W)\right) \xrightarrow{\mathscr{L}} \mathscr{N}\left(0,\gamma^2(x)\mathscr{AV}(x,W)\right), \qquad (4.7)$$

où $\mathscr{AB}(x,W) = \int_0^1 W(s,x)s^{-\rho(x)}ds$ avec $\rho(x)$ une fonction négative appelée *paramètre du second ordre* et $\mathscr{AV}(x,W) = \int_0^1 W^2(s,x)ds$.

La variance asymptotique de leur estimateur est proportionnelle, à un facteur d'échelle près, au carré de l'indice de queue conditionnel. Ce facteur d'échelle ou terme multiplicatif donné par $\mathscr{AV}(x,W)$ fait intervenir la fonction de poids. Ceci montre qu'un mauvais choix de la fonction de poids $W(.,x)$ a forcément des répercutions sur la qualité des estimateurs et en particulier sur la construction des intervalles de confiances.

Pour cette famille d'estimateurs, $\mathcal{V}_n = k_{n,x}^{1/2}$ et on a

$$\frac{\zeta_{m_{n,x}}}{\mathcal{V}_n} = \log\left(\frac{k_{n,x}}{m_{n,x}\alpha_{m_{n,x}}}\right) \to \infty.$$

Par conséquent, la situation (i) du Théorème 4.3.3 n'est pas viable car la condition (4.5) ne peut être satisfaite.

4.4.2 Applications et discussion

Dans ce paragraphe, on illustre tout d'abord le Théorème 4.3.3 sur les quantiles extrêmes conditionnels avec l'estimateur de l'indice de queue présenté précédemment. Ensuite, nous adapterons l'estimateur de la Définition 2.1.1 au cas conditionnel. Enfin, nous donnerons des exemples de quelques lois à queues lourdes.

4.4.2.1 Application du résultat de Gardes & Girard à nos estimateurs de quantile extrême

Pour tout $\beta_{m_{n,x}} = k_{n,x}/m_{n,x} \to 0$, l'estimateur de quantile extrême conditionnel de la situation **(S.3)** peut se réécrire :

$$\hat{q}_2(\alpha_{m_{n,x}}, x, W) = Z_{m_{n,x}-k_{n,x}+1, m_{n,x}}(x)\left(\frac{k_{n,x}}{m_{n,x}\alpha_{m_{n,x}}}\right)^{\hat{\gamma}_n(x,W)}.$$

Compte tenu de la remarque faite précédemment, seule la situation (ii) du Théorème 4.3.3 présente un intérêt certain au vu de la vitesse de convergence des estimateurs de l'indice de queue conditionnel introduits par Gardes et Girard (2008). Cette situation (ii) peut être réécrite de façon à mieux intégrer le comportement asymptotique de ces estimateurs.

Corollaire 4.4.1. *Supposons les hypohèses de (Gardes et Girard, 2008, Théorème 2) satisfaites. Sous* **(A.1)** – **(A.2)**, *si la suite* $(k_{n,x})_{n\geq 1}$ *est telle que :*

$$k_{n,x}^{1/2}\bar{\Delta}\left(k_{n,x}/m_{n,x}, x\right) \to 0 \text{ et} \tag{4.8}$$

$$k_{n,x}^{1/2}\omega_n\left(m_{n,x}^{-(1+\delta)}\right) \to 0 \text{ pour un certain } \delta > 0, \tag{4.9}$$

et si $(\alpha_{m_{n,x}})_{n\geq 1}$ *est une suite satisfaisant la situation* **(S.2)** *ou* **(S.3)**, *alors*

$$\frac{k_{n,x}^{1/2}}{\log\left(\frac{k_{n,x}}{m_{n,x}\alpha_{m_{n,x}}}\right)}\left(\frac{\hat{q}_2(\alpha_{m_{n,x}}, x, W)}{q(\alpha_{m_{n,x}}, x)} - 1\right) \to \mathcal{N}\left(0, \gamma^2(x)\mathcal{AV}(x, W)\right).$$

Comme exemple de famille de poids, nous pouvons citer :

- *La famille de poids constante* qui consiste à poser pour tout $s \in [0,1]$, $W^{\mathrm{H}}(s,x) = 1$. Alors, l'estimateur de queue obtenu est une adaptation de l'estimateur de Hill (1975) au cas conditionnel. Dans ce cas, l'estimateur de quantile extrême conditionnel $\hat{q}_2(\alpha_{m_{n,x}}, x, W^{\mathrm{H}})$ est asymptotiquement gaussien avec $\mathcal{AB}(x, W^{\mathrm{H}}) = 1/(1-\rho(x))$ et $\mathcal{AV}(x, W^{\mathrm{H}}) = 1$.

- *La Famille de poids logarithmique* définie pour tout $s \in]0,1]$ par $W^{\mathrm{Z}}(s,x) = -\log(s)$. L'estimateur de queue qui en découle est une adaptation de l'estimateur de Zipf (Schultze et Steinebach, 1996; Kratz et Resnick, 1996) au cas conditionnel. Dans ce cas, $\hat{q}_2(\alpha_{m_{n,x}}, x, W^{\mathrm{Z}})$ est asymptotiquement gaussien avec $\mathscr{AB}(x, W^{\mathrm{Z}}) = 1/(1-\rho(x))^2$ et $\mathscr{AV}(x, W^{\mathrm{Z}}) = 2$.

Pour d'autres exemples de familles de poids, le lecteur pourra se référer à l'article de Gardes et Girard (2008).

On peut estimer $q(\alpha_{m_{n,x}}, x)$ en adaptant l'estimateur de quantile extême introduit au chapitre 2 (cf. Définition 2.1.1) au cas conditionnel. Ainsi, on a donc

$$\hat{q}_3(\alpha_{m_{n,x}}, x) = \left[\prod_{i=1}^{k_{n,x}} Z_{m_{n,x}-i+1,m_{n,x}}(x)\left(\frac{i\, g_{k_{n,x}}}{m_{n,x}\alpha_{m_{n,x}}}\right)^{\hat{\gamma}_i(x,W^{\mathrm{H}})}\right]^{1/k_{n,x}},$$

avec
$$g_{k_{n,x}} = \exp\left[\log(k_{n,x}+1) - 1 - \log(k_{n,x}!)/k_{n,x}\right],$$
$$\hat{\gamma}_i(x, W^{\mathrm{H}}) \overset{def}{=} \frac{1}{i}\sum_{j=1}^{i} j \log\left(\frac{Z_{m_{n,x}-j+1,m_{n,x}}(x)}{Z_{m_{n,x}-j,m_{n,x}}(x)}\right).$$

Corollaire 4.4.2. *Supposons les hypohèses de (Gardes et Girard, 2008, Théorème 2) satisfaites. Sous* **(A.1)** – **(A.2)**, *si la suite* $(k_{n,x})_{n\geq 1}$ *vérifie les conditions (4.8), (4.9) et* $\log(k_{n,x})/\log(m_{n,x}\alpha_{m_{n,x}}) \to 0$ *et si* $(\alpha_{m_{n,x}})_{n\geq 1}$ *est une suite satisfaisant la situation* **(S.3)**, *alors*

$$\frac{k_{n,x}^{1/2}}{\log\left(\frac{k_{n,x}}{m_{n,x}\alpha_{m_{n,x}}}\right)}\left(\frac{\hat{q}_3(\alpha_{m_{n,x}}, x)}{q(\alpha_{m_{n,x}}, x)} - 1\right) \xrightarrow{\mathscr{L}} \mathscr{N}\left(0, 2\gamma^2(x)\right).$$

Il apparait que les estimateurs $\hat{q}_2(\alpha_{m_{n,x}}, x, W^{\mathrm{Z}})$ et $\hat{q}_3(\alpha_{m_{n,x}}, x)$ ont même loi asymptotiquement. Pour une analogie, on pourra se référer aux commentaires du Théorème 2.2.1 (voir chapitre 2).

4.4.2.2 Illustration sur quelques lois à queue lourde

La première illustration que nous proposons est la loi de Pareto. Cette loi est l'exemple le plus simple des lois à queues lourdes car sa fonction à variations lentes vaut un, ie $\ell(y,x) = 1$. En effet, son quantile conditionnel d'ordre $1-\alpha_{m_{n,x}}$ a pour expression $q(\alpha_{m_{n,x}}, x) = \alpha_{m_{n,x}}^{-\gamma(x)}$. Par conséquent, sa fonction biais $\Delta(\alpha_{m_{n,x}}, x) = 0$ et la condition (4.8) du Corollaire 4.4.1 est trivialement satisfaite.

La deuxième illustration est la loi de Fréchet dont le quantile conditionnel est défini par

$$q(\alpha_{m_{n,x}}, x) = \alpha_{m_{n,x}}^{-\gamma(x)}\left\{\frac{1}{\alpha_{m_{n,x}}}\log\left(\frac{1}{1-\alpha_{m_{n,x}}}\right)\right\}^{-\gamma(x)}.$$

Ce quantile conditionnel décroit aussi comme un polynôme en $\alpha_{m_{n,x}}$ de puissance $-\gamma(x)$, i.e $q(\alpha_{m_{n,x}}, x) \sim \alpha_{m_{n,x}}^{-\gamma(x)}$ et la qualité de cette approximation est contrôlée par

$$\Delta(\alpha_{m_{n,x}}, x) = -\frac{\gamma(x)}{2}\alpha_{m_{n,x}}\left(1 + O(\alpha_{m_{n,x}})\right).$$

Une autre illustration peut être donnée par la loi de Burr dont le quantile conditionnel et son erreur d'approximation sont respectivement donnée pour tout $\rho(x) < 0$ par :

$$q(\alpha_{m_{n,x}}, x) = \alpha_{m_{n,x}}^{-\gamma(x)}\left(1 - \alpha_{m_{n,x}}^{-\rho(x)}\right)^{-\gamma(x)/\rho(x)}$$

et

$$\Delta(\alpha_{m_{n,x}}, x) = -\gamma(x)\alpha_{m_{n,x}}^{-\rho(x)}\left(1 + O(\alpha_{m_{n,x}}^{-\rho(x)})\right).$$

On remarque que pour les lois de Fréchet ou de Burr, la fonction de biais $\Delta(\alpha_{m_{n,x}}, x)$ est asymptotiquement proportionnelle à $\alpha_{m_{n,x}}^{-\rho(x)}$ quand $\alpha_{m_{n,x}} \to 0$ avec la convention $\rho(x) = -1$ pour la loi de Fréchet. Le paramètre du second ordre $\rho(x)$ contrôle la qualité d'approximation du quantile conditionnel $q(\alpha_{m_{n,x}}, x)$ par le polynôme $\alpha_{m_{n,x}}^{-\gamma(x)}$. En outre, comme pour ces deux lois, la valeur absolue de la fonction biais, i.e $\left|\Delta(\alpha_{m_{n,x}}, x)\right|$ est croissante alors, la condition (4.8) du Corollaire 4.4.1 peut se réécrire

$$m_{n,x}^{2\rho(x)} k_{n,x}^{1-\rho(x)} \to 0.$$

Ce qui montre que l'on peut prendre $k_{n,x}$ grand lorsque la fonction $\rho(x)$ est éloignée de zéro, c'est-à-dire petite. Enfin, si on suppose que la fonction indice de queue conditionnelle $\gamma(.)$ et le paramètre du second ordre $\rho(.)$ sont lipschitziens, i.e s'il existe des constantes $c_\gamma > 0$ et $c_\rho > 0$ telles que

$$\left|\gamma(t) - \gamma(t')\right| \leq c_\gamma d(t, t') \text{ et } \left|\rho(t) - \rho(t')\right| \leq c_\rho d(t, t'),$$

pour tout $(t, t') \in B(x, r_{n,x})^2$, alors la plus grande oscillation de la fonction log-quantile par rapport à sa seconde variable peut être bornée et on a

$$\omega_n(a) = O\left(r_{n,x}\log(1/a)\right) \text{ quand } a \to 0.$$

On peut alors simplifier la condition (4.9) du Corollaire 4.4.1 et on obtient

$$k_{n,x}^2 r_{n,x} \log m_{n,x} \to 0.$$

Ci-joint le tableau récapitulatif de nos illustrations.

	$q(\alpha_{m_{n,x}}, x)$	$\Delta(\alpha_{m_{n,x}}, x)$
Pareto	${\alpha_{m_{n,x}}}^{-\gamma(x)}$	0
Fréchet	${\alpha_{m_{n,x}}}^{-\gamma(x)} \left\{ \frac{1}{\alpha_{m_{n,x}}} \log\left(\frac{1}{1-\alpha_{m_{n,x}}} \right) \right\}^{-\gamma(x)}$	$-\frac{\gamma(x)}{2} \alpha_{m_{n,x}} (1 + O(\alpha_{m_{n,x}}))$
Burr	$\alpha_{m_{n,x}}^{-\gamma(x)} \left(1 - \alpha_{m_{n,x}}^{-\rho(x)}\right)^{-\gamma(x)/\rho(x)}$	$-\gamma(x) \alpha_{m_{n,x}}^{-\rho(x)} (1 + O(\alpha_{m_{n,x}}^{-\rho(x)}))$

TABLE 4.1 – Quelques exemples de lois à queues lourdes. $\gamma(x) > 0$ est la fonction indice de queue conditionnel et $\rho(x) < 0$ est le paramètre du second-ordre

4.5 Simulations et illustration sur données réelles

4.5.1 Simulations

Dans cette partie, on se propose d'illustrer notre approche d'estimation des quantiles extrêmes conditionnels sur des données fonctionnelles issues de la spectrométrie. En physique, la spectrométrie désigne l'ensemble des méthodes d'analyse spectrale permettant d'accéder à la composition et à la structure de la matière. La spectrométrie est fondée sur l'étude (qualitative et quantitative) des spectres fournis par l'interaction de la matière avec divers rayonnements comme la lumière, les rayons X et les électrons. La spectrométrie peut aider les planétologues à comprendre l'histoire géologique des planètes. Ainsi, étant donné un spectre recueilli par l'instrument OMEGA à bord de la sonde européenne Mars Express en orbite autour de Mars, ils aimeraient pouvoir estimer les propriétés physiques associées au sol martien (taille des grains de CO_2, les proportions d'eau, de poussière, de CO_2, etc.).

Pour ce faire, le **LPG** [3] dispose d'un outil de simulation, appelé *radiative transfer model*, permettant à partir de valeurs d'un paramètre physique, de construire les spectres correspondants. Ainsi, pour répondre à la question posée précédemment, on peut construire une base de données d'apprentissage permettant de mettre en œuvre notre méthode d'estimation. Ici, nous nous intéresserons uniquement à la proportion de CO_2 (gaz carbonique). Étant donné des proportions $\{y_i,\ i = 1, \ldots, 16\}$ en entrée de CO_2, le *radiative transfer model* nous fournit en sortie une suite de spectres correspondants $\{x_i,\ i = 1, \ldots, 16\}$ (voir Figure 4.2). Naturellement, les spectres obtenus ici ne sont pas aléatoires mais déterministes. Ils sont des fonctions de la longueur d'onde et dans

3. Laboratoire de Planétologie de Grenoble

cette illustration, nous considérons leur version discrétisée $\{x_{i,l},\ l = 1,\dots,256\}$ [4].

Partant de cette méthode de construction des données d'apprentissage, plusieurs approches d'estimation de la proportion de CO_2 associée à un spectre observé ont été proposées dans la littérature. Dans cette optique, on peut citer la méthode des plus proches voisins, la méthode SIR [5] et la méthode des SVM [6]. On pourra consulter Bernard-Michel *et al.* (2009a) pour un aperçu de ces différentes approches. Pour toutes ces méthodes, l'estimation de la proportion de CO_2 est perturbée par un terme d'erreur aléatoire. Dans cet exemple, nous proposons de modéliser cette perturbation par :

$$Y_{i,j} = \log(1/y_i) + \sigma(\varepsilon_j(x_i) - \Gamma(1-\gamma(x_i))),\ j = 1,\dots,n_i,\ i = 1,\dots,16,$$

où

$$\gamma(x_i) = 0.3\frac{\|x_i\|_2^2 - \min_l \|x_l\|_2^2}{\max_l \|x_l\|_2^2 - \min_l \|x_l\|_2^2} + 0.2,\ \sigma = \min_i \frac{\log(1/y_i)}{\Gamma(1-\gamma(x_i))},$$

et $\varepsilon_j(x_i),\ j = 1,\dots,n_i$ sont des variables aléatoires indépendantes et identiquement distribuées suivant une loi de Fréchet d'indice de queue $\gamma(x_i)$ (cf. Tableau 4.1). Notons que $\|x_i\|_2^2$ est une approximation de l'énergie totale du spectre x_i. Les définitions ci-dessus nous assurent que $\gamma(x_i) \in [0.2, 0.5]$ et que $Y_{i,j} > 0$ pour tout $i = 1,\dots,16$ et $j = 1,\dots,n_i$. En outre, puisque l'espérance de $\varepsilon_j(x_i)$ est donnée par $\Gamma(1-\gamma(x_i))$, les variables aléatoires $Y_{i,j}$ sont centrées sur $\log(1/y_i)$. Notre objectif est d'estimer le quantile conditionnel

$$q(\alpha, x_i) = \bar{F}^{\leftarrow}(\alpha, x_i),\ \text{pour } i = 1,\dots,16,$$

où $\bar{F}(., x_i)$ est la fonction de survie de $Y_{i,1}$. Ainsi, puisque la covariable est de nature fonctionnelle, on peut utiliser l'estimateur $\hat{q}_2(\alpha, x_i, W^Z)$ défini au sous-paragraphe 4.4.2.1 à condition de se munir d'une distance appropriée. Dans un tel contexte, (Ferraty et Vieu, 2006, voir chapitre 9) recommandent d'utiliser la distance semi-métrique basée sur la dérivée seconde :

$$d^2(x_i, x_j) = \int \left(x_i^{(2)}(t) - x_j^{(2)}(t)\right)^2 dt,$$

où $x^{(2)}$ désigne la dérivée seconde de x. Pour calculer cette distance semi-métrique, on peut utiliser une approximation des fonctions x_i et x_j basée sur les B-splines tel que proposé dans (Ferraty et Vieu, 2006, voir chapitre 3). Ici, nous nous limiterons à une version discrétisée $\tilde{d}$ de d :

$$\tilde{d}^2(x_i, x_j) = \sum_{l=2}^{255} \left\{(x_{i,l+1} - x_{j,l+1}) + (x_{i,l-1} - x_{j,l-1}) - 2(x_{i,l} - x_{j,l})\right\}^2.$$

Nous avons évalué la performance de notre estimateur sur $N = 100$ échantillons $\{(x_i, Y_{i,j}),\ i = 1,\dots,16,\ j = 1,\dots,n_i\}$ avec $n_1 = \dots = n_{16} = 100$. Nous nous sommes intéressés à l'estimation des quantiles d'ordre $\alpha \in \{1/300, 1/500\}$. Dans la suite de nos propos,

4. On suppose ici que chaque spectre est un vecteur réponse associé à 256 longueurs d'ondes.
5. Aussi appelée la régression inverse par tranches
6. Connue aussi sous le nom de Machine à vecteurs de support

nous supposons que les hyperparamètres r (la largeur de la fenêtre mobile) et k (les plus grandes observations dans la boule) ne dépendent pas du spectre. Pour sélectionner ces paramètres, on se propose de minimiser la distance entre deux estimateurs différents de quantiles extrêmes conditionnels :

$$(\hat{r}_{\text{select}}, \hat{k}_{\text{select}}) = \underset{r,k}{\arg\min}\, \mathbb{D}\left(\hat{q}_2(\alpha, ., W^{\text{H}}), \hat{q}_2(\alpha, ., W^{\text{Z}})\right),$$

où pour deux fonctions f et g,

$$\mathbb{D}(f, g) = \left\{ \sum_{i=1}^{16} \left(f(x_i) - g(x_i)\right)^2 \right\}^{1/2}.$$

L'estimateur de quantile associé à ces paramètres sera noté $\hat{q}_{\text{select}}$. Nous allons le comparer à celui dont les paramètres notés $\hat{r}_{\text{oracle}}$ et $\hat{k}_{\text{oracle}}$ sont définies par :

$$(\hat{r}_{\text{oracle}}, \hat{k}_{\text{oracle}}) = \underset{r,k}{\arg\min}\, \mathbb{D}(\hat{q}_2(\alpha, ., W^{\text{H}}), q(\alpha, .)).$$

L'estimateur de quantile associé à ces deux derniers paramètres sera noté $\hat{q}_{\text{oracle}}$. Notons que $\hat{r}_{\text{select}}$, $\hat{k}_{\text{select}}$, $\hat{r}_{\text{oracle}}$ et $\hat{k}_{\text{oracle}}$ dépendent de α. Remarquons que l'estimateur construit à l'aide des paramètres $\hat{r}_{\text{oracle}}$ et $\hat{k}_{\text{oracle}}$ n'est pas utilisable en pratique puisque $q(\alpha, .)$ est inconnu. Cependant, il nous donne la borne inférieure de la distance $\mathbb{D}$ (comprendre l'erreur) qui peut être atteinte avec notre méthode d'estimation. Afin de valider notre choix de $\hat{r}_{\text{select}}$ et $\hat{k}_{\text{select}}$, les histogrammes de $\mathbb{D}(\hat{q}_{\text{select}}(\alpha, ., W^{\text{Z}}), q(\alpha, .))$ et $\mathbb{D}(\hat{q}_{\text{oracle}}(\alpha, ., W^{\text{Z}}), q(\alpha, .))$, calculés pour les $N = 100$ échantillons, ont été superposés à la Figure 4.3. Il apparait que les erreurs moyennes sont approximativement égales. Remarquons aussi que les erreurs obtenues avec notre méthode de sélection semblent avoir une queue droite plus lourde que celles obtenues avec la méthode de référence proposée. Pour chaque spectre x_i, l'intervalle de confiance empirique à 90% de $\hat{q}_{\text{oracle}}(\alpha, x_i, W^{\text{Z}})$ est représenté à la Figure 4.4 pour $\alpha = 1/300$ et à la Figure 4.5 pour $\alpha = 1/500$. Les intervalles de confiance sont classés par ordre croissant de l'indice de queue. On observe que plus grand est l'indice de queue et moins étroit sont les intervalles de confiance. Ceci est en adéquation avec les résultats présentés dans le Corollaire 4.4.1 et les commentaires du Théorème 4.3.1. Enfin, sur les Figures 4.6 ($\alpha = 1/300$) et 4.7 ($\alpha = 1/500$), nous avons représenté les estimateurs $\hat{q}_{\text{select}}(\alpha, x_i, W^{\text{Z}})$ et $\hat{q}_{\text{oracle}}(\alpha, x_i, W^{\text{Z}})$ correspondants à la situation de l'erreur médiane $\mathbb{D}(\hat{q}_{\text{select}}(\alpha, ., W^{\text{Z}}), q(\alpha, .))$ en fonction de $\|x_i\|_2^2$. Il apparait que l'estimateur de la stratégie oracle est à peine meilleur que celui obtenu avec le critère de sélection.

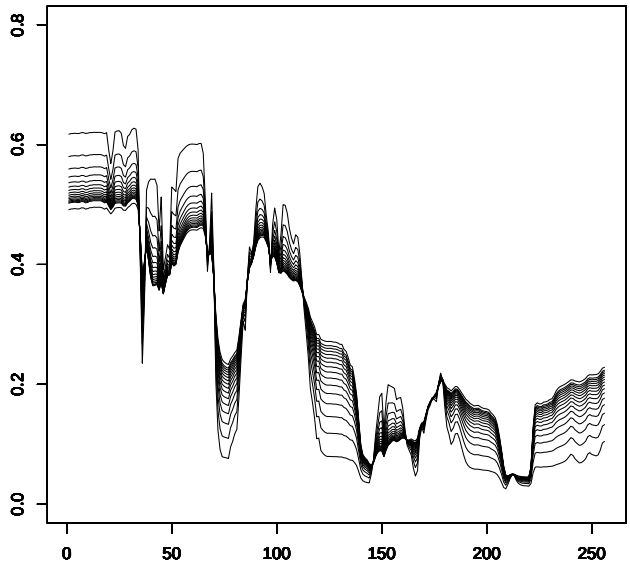


FIGURE 4.2 – Représentation des 16 spectres en fonction des longueurs d'ondes.

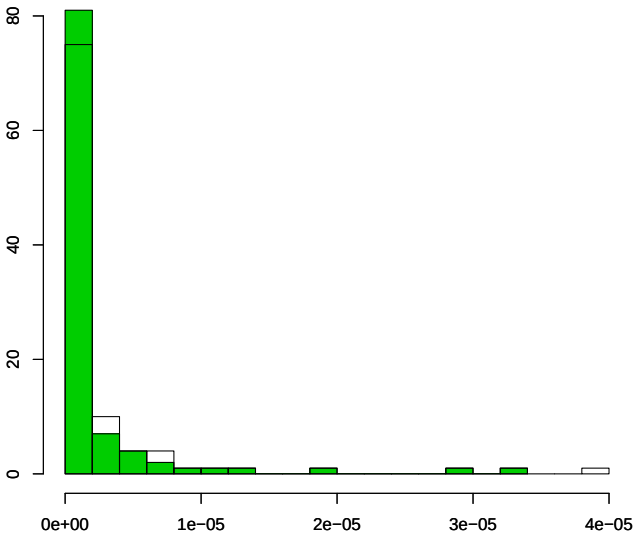


FIGURE 4.3 – Comparaison des histogrammes des erreurs calculées sur $N = 100$ échantillons avec la stratégie de référence oracle (vert) et le critère de sélection (transparent).

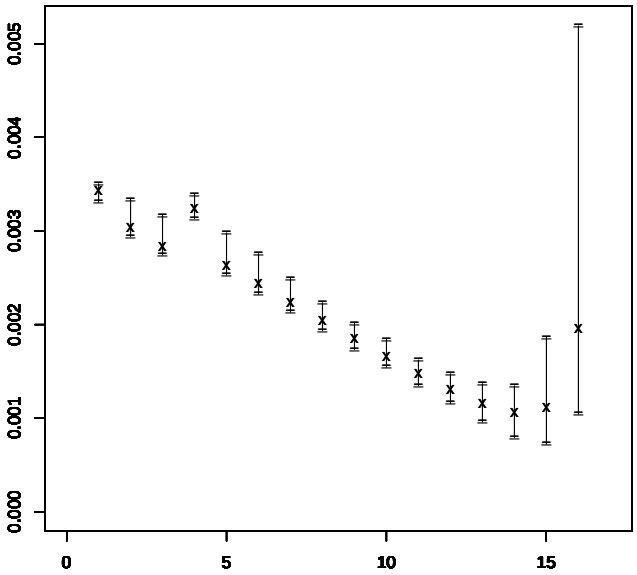


FIGURE 4.4 – Intervalles de confiance empirique à 90% de $\hat{q}_{\text{oracle}}(1/300, ., W^Z)$ classés par ordre croissant de l'indice de queue.

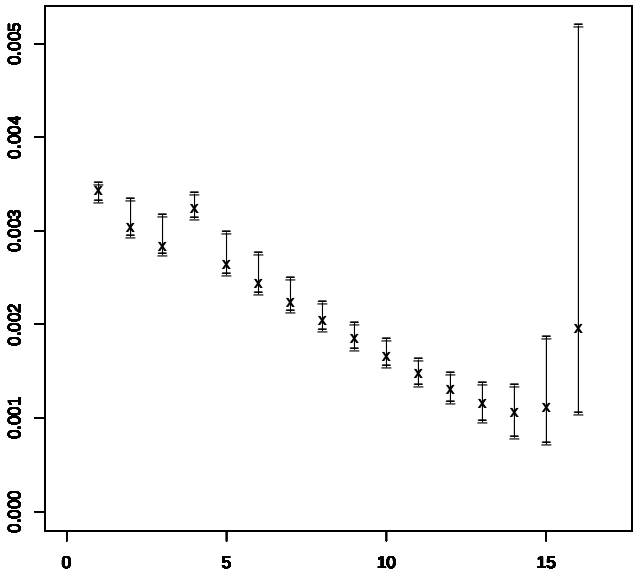


FIGURE 4.5 – Intervalles de confiance empirique à 90% de $\hat{q}_{\text{oracle}}(1/500, ., W^Z)$ classés par ordre croissant de l'indice de queue.

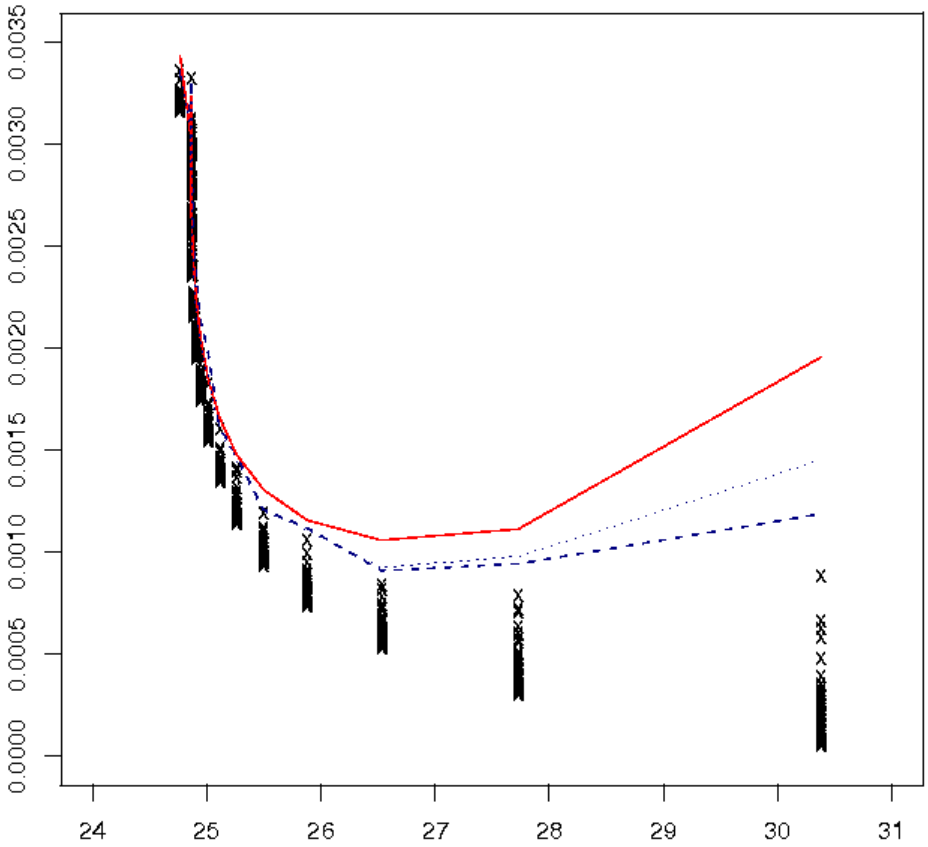


FIGURE 4.6 – Comparaison du vrai quantile $q(1/300,.)$ (rouge) avec les estimateurs obtenus par le critère de sélection $\hat{q}_{\text{select}}(1/300,.,W^Z)$ (traits interrompus) et la stratégie oracle $\hat{q}_{\text{oracle}}(1/300,.,W^Z)$ (traits pointillés). Les estimateurs représentés ici sont ceux correspondants à la médiane de l'erreur $\mathbb{D}(\hat{q}_{\text{select}}(1/300,.,W^Z), q(1/300,.))$. L'échantillon associé est representé par les points ("×"). En abscisse on a $\|x_i\|_2^2$.

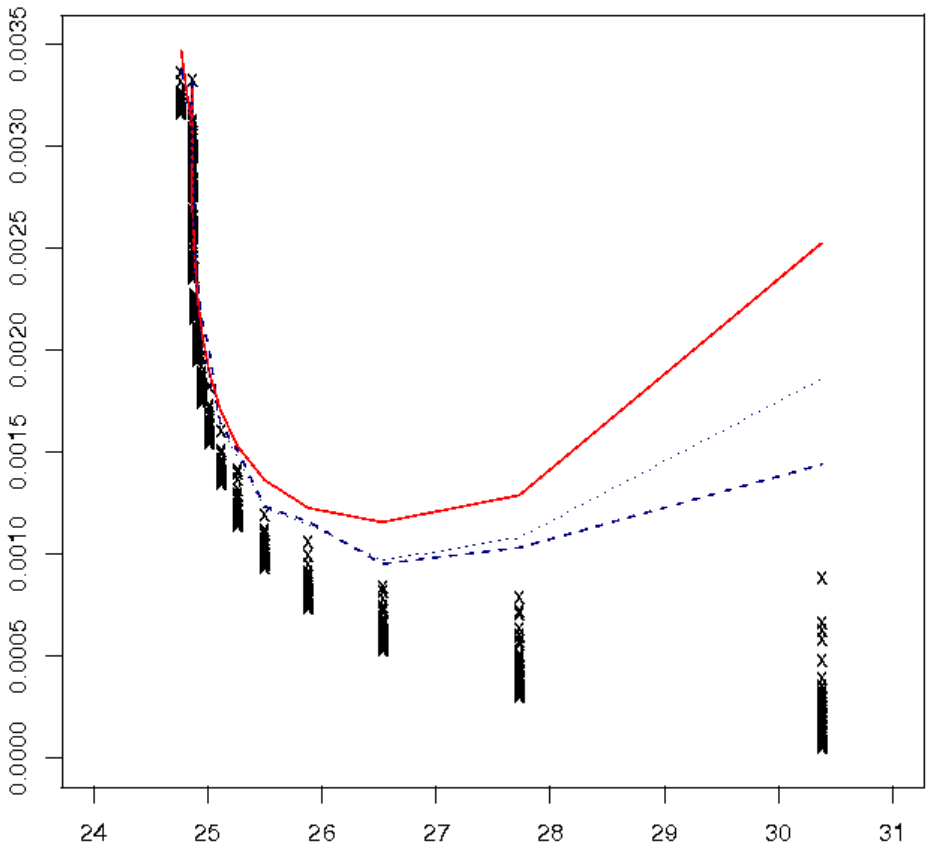


FIGURE 4.7 – Comparaison du vrai quantile $q(1/500,.)$ (rouge) avec les estimateurs obtenus par le critère de sélection $\hat{q}_{\text{select}}(1/500,.,W^Z)$ (traits interrompus) et la stratégie oracle $\hat{q}_{\text{oracle}}(1/500,.,W^Z)$ (traits pointillés). Les estimateurs représentés ici sont ceux correspondants à la médiane de l'erreur $\mathbb{D}(\hat{q}_{\text{select}}(1/500,.,W^Z), q(1/500,.))$. L'échantillon associé est representé par les points ("×"). En abscisse on a $\|x_i\|_2^2$.

4.5.2 Illustration sur données réelles

Le LTHE de Grenoble a mesuré les hauteurs de pluies journalières (en mm) entre les années 1958 et 2000 sur 225 stations situées dans la région des Cévennes-Vivarais (sud de la France). L'étude des pluies et débits extrêmes revêt un intérêt certain en termes d'aménagement du territoire et de pré-détermination du risque hydrologique. Les hydrologues du LTHE aimeraient estimer la hauteur de pluies journalières pouvant être dépassée une fois toutes les N années. La réponse à cette question peut fournir des éléments indispensables pour construire aux endroits critiques des digues d'une hauteur appropriée ou définir la périodicité des opérations de nettoyage des fleuves et des estuaires afin de protéger efficacement la population. D'une manière formelle, le problème posé dans cette application est celui d'estimer le quantile des précipitations journalières d'ordre $1/(365 \times N)$ que nous appellerons ici le *niveau de retour* sur N ans.

Diverses approches d'estimation des quantiles des précipitations ont été considérées dans la littérature. Certains auteurs modélisent les précipitations par un processus max-stable (Buishand *et al.*, 2008; S. A. Padoan *et al.*, 2010) et d'autres utilisent une approche d'estimation bayésienne où les excès sont approchés par une GPD avec quelques informations *a priori* sur ses paramètres (Coles et Tawn, 1996).

Dans notre contexte, les variables d'intérêts sont les précipitations journalières et les covariables sont les coordonnées géographiques des stations. Les coordonnées géographiques sont définies par la longitude, la latitude et l'altitude. La Figure 4.8 montre la disposition des 225 stations dans le plan. Nous disposons au total de $n = 821925$ observations. Soulignons que Bois *et al.* (1997) et Gardes et Girard (2010) ont déjà étudié les précipitations extrêmes dans la région des Cévennes-Vivarais. Les premiers contributeurs ont utilisé un jeu de données de pluies horaires mesurées entre les années 1948 et 1991 sur 48 stations. Ils se sont intéressés aux quantiles décennaux [7] de précipitations qu'ils ont estimés par une loi de Gumbel et la méthode de krigeage (Krige, 1951) [8]. Les seconds contributeurs ont plutôt supposé que les pluies extrêmes pouvaient être modélisées par une loi à queue lourde. Partant d'un jeu de données de pluies horaires mesurées entre les années 1993 et 2000 sur 142 stations, ils ont estimé le quantile décennal des précipitations par la méthode des plus proches voisins. Les résultats obtenus emmènent ces auteurs à remettre en cause l'hypothèse selon laquelle la loi des précipitations dans la région des Cévennes-Vivarais appartiendrait au $\mathscr{D}(Gumbel)$. Ils affirment : « *The large estimated values* ($\hat{\gamma}_n \in [0.15, 0.28]$) *are consistent with the credibility intervals found in Coles et Tawn (1996) but contradict the Gumbel assumption of Bois* et al. *(1997). (...) The Gumbel assumption seems therefore to be unrealistic.* »

Compte tenu de toutes ces remarques, nous avons supposé que la loi des pluies journalières dans la région Cévennes-Vivarais était à queue lourde. Nous nous sommes in-

7. niveau de retour sur 10-ans

8. Le krigeage est une méthode d'interpolation spatiale. Cette méthode porte le nom de son précurseur, l'ingénieur minier sud-africain Daniel Gerhardus Krige.

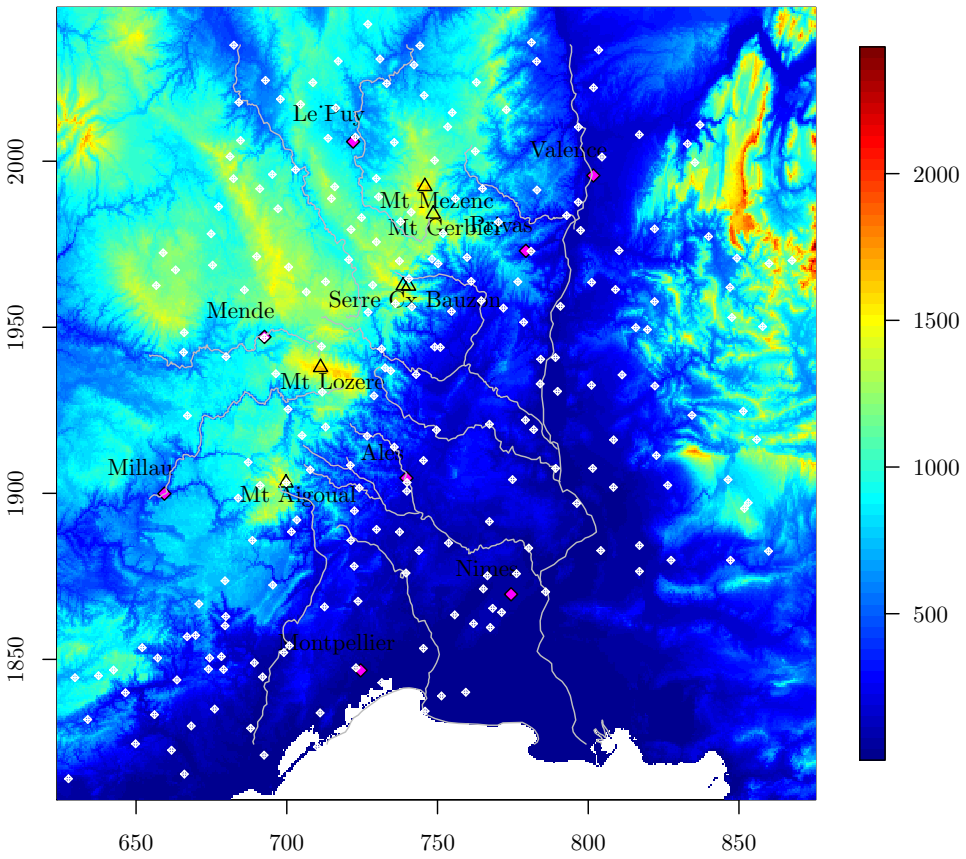


FIGURE 4.8 – 225 stations d'observations (losanges blancs) représentées en fonction de leurs cordonnées géographiques. Horizontalement : la longitude (en kilomètres), verticalement : la latitude (en kilomètres), l'échelle des couleurs : l'altitude (en mètres). Sur la carte : les montagnes (triangles), les cours d'eaux (lignes grises) et les villes (losanges roses)

téressés à l'estimation du quantile conditionnel centenal. Pour ce faire, on a discrétisé le plan (longitude, latitude) selon une grille de taille 240×228. On se propose d'utiliser les estimateurs $\hat{q}_2(\alpha_{m_{n,x}}, ., W^{\mathrm{H}})$, $\hat{q}_2(\alpha_{m_{n,x}}, ., W^{\mathrm{Z}})$ et $\hat{q}_3(\alpha_{m_{n,x}}, .)$. Ces estimateurs dépendent des paramètres $(r_{n,x}, k_{n,x})$ que nous nous proposons de choisir en chaque point x de la covariable en minimisant la distance entre ces trois estimateurs, i.e

$$\left(\hat{r}_{n,x}, \hat{k}_{n,x}\right) = \arg\min_{r_{n,x}, k_{n,x}} \mathbb{D}\left(\hat{q}_2\left(\alpha_{m_{n,x}}, x, W^{\mathrm{H}}\right), \hat{q}_2\left(\alpha_{m_{n,x}}, x, W^{\mathrm{Z}}\right), \hat{q}_3\left(\alpha_{m_{n,x}}, x\right)\right), \tag{4.10}$$

avec $\mathbb{D}(v_1, v_2, v_3) = (v_1 - v_2)^2 + (v_1 - v_3)^2 + (v_2 - v_3)^2$. D'après cette heuristique, si le couple

$(\hat{r}_{n,x}, \hat{k}_{n,x})$ est correctement choisi, alors la valeur des trois estimateurs doit être approximativement la même. Dans (4.10), on recherche le rayon $r_{n,x} \in \{5, \ldots, 130\}$ (en km) et le nombre des plus grandes statistiques d'ordres $k_{n,x} \in \{2, \ldots, 0.1 m_{n,x}\}$.

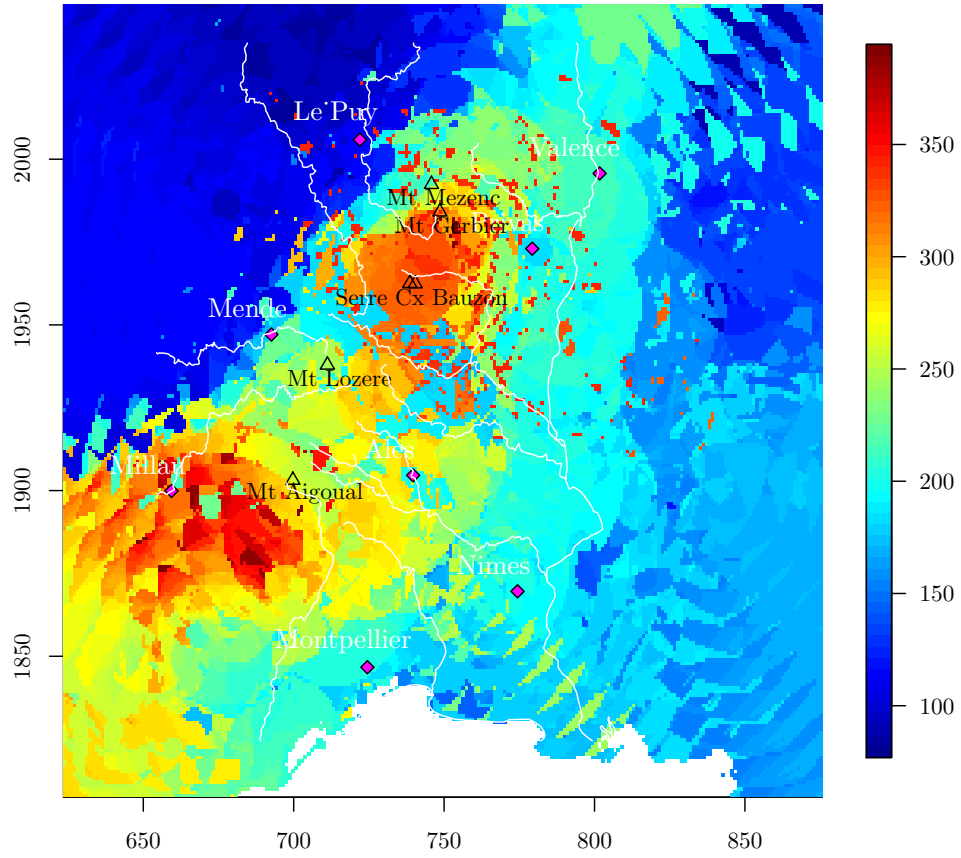


FIGURE 4.9 – Carte des niveaux de retour sur 100 ans en fonction de la longitude et de la latitude, estimée avec $\hat{q}_2(\alpha_{m_{n,x}}, ., W^Z)$.

Sur la Figure 4.9, il apparait que le niveau de retour décroit globalement avec l'altitude. La dérive du taux de précipitations en fonction de l'altitude est en adéquation avec les statistiques descriptives des précipitations dans la région des Cévennes-Vivarais (Molinié *et al.*, 2008). Étant donné que dans cette région des zones de basse altitude sont des terrains plats et proches de la mer, le processus physique de déconvolution impliqué dans le rapport entre le taux de précipitations observé et l'altitude est complexe. Par conséquent, deux phénomèmes pourraient expliquer la chute de pluies

extrêmes :

- *un phénomène régional* : les vents en provenance de la mer Méditerranée alimentent en air chaud et humide le massif cevenol,
- *un phénomène universel* : les terrains plats capturent efficacement l'énergie solaire qui est nécessaire a la formation des nuages convectifs.

4.6 Démonstrations

4.6.1 Résultats préliminaires

Notre premier résultat préliminaire se résume au Lemme 4.6.1 qui un outil permettant de déterminer la loi asymptotique d'une variable aléatoire conditionnellement à un évenement de probabilité tendant vers un.

Lemme 4.6.1. *Soient* $(X_n)_{n\geq 1}$ *et* $(Y_n)_{n\geq 1}$ *deux suites de variables aléatoires. S'il existe un évenement* $\mathscr{A}_n$ *tel que* $(X_n|\mathscr{A}_n) \overset{\mathscr{L}}{=} (Y_n|\mathscr{A}_n)$ *avec* $\mathbb{P}(\mathscr{A}_n) \to 1$, *alors*

$$Y_n \overset{\mathscr{L}}{\to} Y \Longrightarrow X_n \overset{\mathscr{L}}{\to} Y.$$

Démonstration du lemme 4.6.1. Pour tout $x \in \mathbb{R}$ et pour tout système complet $\{\mathscr{A}_n, \mathscr{A}_n^C\}$, où $\mathscr{A}_n^C$ est le complémentaire de l'événement $\mathscr{A}_n$, on a la relation suivante :

$$\mathbb{P}(X_n \leq x) = \mathbb{P}(\{X_n \leq x\}|\mathscr{A}_n)\mathbb{P}(\mathscr{A}_n) + \mathbb{P}(\{X_n \leq x\}|\mathscr{A}_n^C)\mathbb{P}(\mathscr{A}_n^C),$$

et donc l'inégalité :

$$\mathbb{P}(\{X_n \leq x\}|\mathscr{A}_n)\mathbb{P}(\mathscr{A}_n) \leq \mathbb{P}(X_n \leq x) \leq \mathbb{P}(\{X_n \leq x\}|\mathscr{A}_n) + \mathbb{P}(\mathscr{A}_n^C).$$

Comme $(X_n|\mathscr{A}_n) \overset{\mathscr{L}}{\to} (Y_n|\mathscr{A}_n)$, il s'en suit que :

$$\mathbb{P}(\{X_n \leq x\} \cap \mathscr{A}_n)\mathbb{P}(\mathscr{A}_n) \leq \mathbb{P}(X_n \leq x) \leq \mathbb{P}(\{X_n \leq x\} \cap \mathscr{A}_n) + \mathbb{P}(\mathscr{A}_n^C).$$

En remarquant que

$$\mathbb{P}(Y_n \leq x) - \mathbb{P}(\mathscr{A}_n^C) \leq \mathbb{P}(\{Y_n \leq x\} \cap \mathscr{A}_n) \leq \mathbb{P}(Y_n \leq x)$$

on aboutit à

$$\mathbb{P}(Y_n \leq x) - \mathbb{P}(\mathscr{A}_n^C) \leq \mathbb{P}(X_n \leq x) \leq \mathbb{P}(Y_n \leq x) + \mathbb{P}(\mathscr{A}_n^C).$$

$\mathbb{P}(Y_n \leq x) \to \mathbb{P}(Y \leq x)$ et $\mathbb{P}(\mathscr{A}_n^C) \to 0$ permettent de conclure la preuve. □

Notre second résultat préliminaire établit la loi asymptotique sur la statistique d'ordre d'une suite de variables aléatoires indépendantes et de loi uniforme standard dans une situation analogue à **(S.1)** dans le cas univarée.

Lemme 4.6.2. *Soit $\{V_1,\dots,V_M\}$ une suite de variables aléatoires indépendantes et de loi uniforme standard. Pour toute suite $(\theta_M)_{M\geq 1}\subset]0,1[$ telle que $\theta_M\to 0$ et $M\theta_M\to\infty$, nous avons*

$$\left(\frac{M}{\theta_M}\right)^{1/2}(V_{\lfloor M\theta_M\rfloor,M}-\theta_M)\xrightarrow{\mathscr{L}}\mathscr{N}(0,1).$$

Démonstration du lemme 4.6.2. Afin de simplifier les notations, on pose $k_M=\lfloor M\theta_M\rfloor$. En se référant au Lemme 1.4.3 sur la représentation de Rényi, on peut écrire

$$V_{k_M,M}\overset{\mathscr{L}}{=}\sum_{i=1}^{k_M}E_i\Big/\sum_{i=1}^{M+1}E_i\,,$$

où $\{E_1,\dots,E_{M+1}\}$ est une suite de variables aléatoires indépendantes et de loi exponentielle standard. Ainsi, en posant

$$\begin{aligned}\xi_M &\overset{\mathscr{L}}{=}\left(\frac{M}{\theta_M}\right)^{1/2}(V_{k_M,M}-\theta_M)\\ &\overset{\mathscr{L}}{=}\left(\frac{1}{M}\sum_{i=1}^{M+1}E_i\right)^{-1}\left(\frac{M}{\theta_M}\right)^{1/2}\times\left[\frac{1}{k_M}\sum_{i=1}^{k_M}E_i\left(\frac{k_M}{M}-\theta_M\right)+\theta_M\left(\frac{1}{k_M}\sum_{i=1}^{k_M}E_i-1\right)\right.\\ &\quad-\left.\theta_M\left(\frac{1}{M}\sum_{i=1}^{M+1}E_i-1\right)\right],\end{aligned}$$

on a, d'après la loi des grands nombres,

$$\begin{aligned}\xi_M &\overset{\mathbb{P}}{\sim}\left(\frac{M}{\theta_M}\right)^{1/2}\left(\frac{k_M}{M}-\theta_M\right)(1+o_P(1))+(M\theta_M)^{1/2}\left(\frac{1}{k_M}\sum_{i=1}^{k_M}E_i-1\right)\\ &\quad-(M\theta_M)^{1/2}\left(\frac{1}{M}\sum_{i=1}^{M+1}E_i-1\right)\\ &\overset{\mathscr{L}}{=}\xi_{1,M}+\xi_{2,M}-\xi_{3,M}.\end{aligned}$$

Intéressons nous maintenant aux trois termes séparément.

Étude du terme $\xi_{1,M}$: Comme $k_M=\lfloor M\theta_M\rfloor$, on peut écrire $k_M=M\theta_M-\tau_M$ avec $\tau_M\in[0,1[$ et, nous avons alors

$$\xi_{1,M}\overset{\mathbb{P}}{\sim}\left(\frac{M}{\theta_M}\right)^{1/2}\frac{\tau_M}{M}=\frac{\tau_M}{(M\theta_M)^{1/2}}\to 0, \tag{4.11}$$

puisque par hypothèse $M\theta_M\to\infty$.

Étude du terme $\xi_{2,M}$: Puisque $k_M\sim M\theta_M$, une application du TCL nous assure la convergence en loi du terme $\xi_{2,M}$. Plus précisément, on a :

$$\xi_{2,M}\sim k_M^{1/2}\left(\frac{1}{k_M}\sum_{i=1}^{k_M}E_i-1\right)\xrightarrow{\mathscr{L}}\mathscr{N}(0,1). \tag{4.12}$$

Étude du terme $\xi_{3,M}$: Comme précédemment, une simple application du TCL couplée à la condition $\theta_M \to 0$ conduit à :

$$\xi_{3,M} = O_{\mathrm{P}}(\theta_M^{1/2}) = o_{\mathrm{P}}(1). \tag{4.13}$$

Les relations (4.11), (4.12) et (4.13) nous permettent de conclure la preuve. □

4.6.2 Preuve des résultats théoriques

Démonstration de la Proposition 4.3.1. Sous l'hypothèse **(A.1)**, l'indépendance des variables aléatoires $\{Z_i(x),\ i = 1,\dots,m_{n,x}\}$ nous assure que

$$\{\log Z_i(x),\ i = 1,\dots,m_{n,x}\} \overset{\mathscr{L}}{=} \{\log q(V_i, x_i)\ i = 1,\dots,m_{n,x}\},$$

où x_i est la covariable associée à l'observations $Z_i(x)$. En notant par $\psi(i)$ l'indice aléatoire de la covariable associée à la statistique ordonnée $Z_{m_{n,x}-i+1,m_{n,x}}(x)$, on obtient

$$\{\log Z_{m_{n,x}-i+1,m_{n,x}}(x),\ i = 1,\dots,m_{n,x}\} \overset{\mathscr{L}}{=} \{\log q(V_{\psi(i)}, x_{\psi(i)})\ i = 1,\dots,m_{n,x}\}.$$

Considérons maintenant l'événement $\mathscr{A}_n = \mathscr{A}_{1,n} \cap \mathscr{A}_{2,n}$, où

$$\begin{aligned}
\mathscr{A}_{1,n} &= \left\{ \min_{i=1,\dots,k_{n,x}-1} \log \frac{q(V_{i,m_{n,x}}, u_i)}{q(V_{i+1,m_{n,x}}, u_{i+1})} > 0, \forall (u_1,\dots,u_{k_{n,x}}) \subset B(x, r_{n,x}) \right\} \text{ et} \\
\mathscr{A}_{2,n} &= \left\{ \min_{i=k_{n,x}+1,\dots,m_{n,x}} \log \frac{q(V_{k_{n,x},m_{n,x}}, u_{k_{n,x}})}{q(V_{i,m_{n,x}}, u_i)} > 0, \forall (u_{k_{n,x}+1},\dots,u_{m_{n,x}}) \subset B(x, r_{n,x}) \right\}.
\end{aligned}$$

Conditionnellement à l'événement $\mathscr{A}_{1,n}$, les variables aléatoires $\{q(V_{i,m_{n,x}}, u_i), i = 1,\dots,k_{n,x}\}$ sont ordonnées ainsi

$$q(V_{k_{n,x},m_{n,x}}, u_{k_{n,x}}) \le q(V_{k_{n,x}-1,m_{n,x}}, u_{k_{n,x}-1}) \le \dots \le q(V_{1,m_{n,x}}, u_1),$$

et conditionnellement à $\mathscr{A}_{2,n}$, les variables aléatoires $\{q(V_{i,m_{n,x}}, u_i),\ i = k_{n,x}+1,\dots,m_{n,x}\}$ restantes leur sont toutes plus petites puisque

$$\max_{i=k_{n,x}+1,\dots,m_{n,x}} q(V_{i,m_{n,x}}, u_i) \le q(V_{k_{n,x},m_{n,x}}, u_{k_{n,x}}).$$

Par conséquent, conditionnellement à $\mathscr{A}_n$, les $k_{n,x}$ plus grandes observations de l'ensemble $\{\log q(V_{\psi(i)}, x_{\psi(i)}),\ i = 1,\dots,m_{n,x}\}$ sont $\{\log q(V_{i,m_{n,x}}, x_{\psi(i)}),\ i = 1,\dots,k_{n,x}\}$. En conséquence, pour $J_{k_{n,x}} = \{1,\dots,k_{n,x}\}$ et $T_i \overset{def}{=} x_{\psi(i)}$, nous avons :

$$\left\{\log Z_{m_{n,x}-i+1,m_{n,x}}(x), i \in J_{k_{n,x}} | \mathscr{A}_n\right\} \overset{\mathscr{L}}{=} \left\{\log q(V_{i,m_{n,x}}, T_i), i \in J_{k_{n,x}} | \mathscr{A}_n\right\}.$$

Afin de conclure la preuve, il ne nous reste plus qu'à montrer que $\mathbb{P}(\mathscr{A}_n) \to 1$ quand $n \to \infty$. Pour ce faire, définissons $\delta_{m_{n,x}} = m_{n,x}^{-(1+\delta)}$ et considérons les événements

$$\begin{aligned}
\mathscr{A}_{3,n} &= \{V_{1,m_{n,x}} > \delta_{m_{n,x}}\} \cap \{V_{m_{n,x},m_{n,x}} < 1 - \delta_{m_{n,x}}\} \text{ et} \\
\mathscr{A}_{2,n} &= \left\{ \min_{i=1,\dots,k_{n,x}} \log \frac{q(V_{i,m_{n,x}}, x)}{q(V_{i+1,m_{n,x}}, x)} > 2\omega_n(\delta_{m_{n,x}}) \right\}.
\end{aligned}$$

Sous $\mathscr{A}_{3,n}$, nous avons $\delta_{m_{n,x}} < V_{i,m_{n,x}} < 1-\delta_{m_{n,x}}$ pour tout $i = 1,\dots,m_{n,x}$. Ainsi, en se référant à la Définition 4.3.1, pour tout $(u_i,u_j) \in B(x,r_{n,x})^2$, il s'en suit d'une part que

$$\begin{aligned}\log\frac{q(V_{j,m_{n,x}},u_j)}{q(V_{i,m_{n,x}},u_i)} &= \log\frac{q(V_{j,m_{n,x}},x)}{q(V_{i,m_{n,x}},x)} + \log\frac{q(V_{j,m_{n,x}},u_j)}{q(V_{j,m_{n,x}},x)} + \log\frac{q(V_{i,m_{n,x}},x)}{q(V_{i,m_{n,x}},u_i)}\\ &\geq \log\frac{q(V_{j,m_{n,x}},x)}{q(V_{i,m_{n,x}},x)} - 2\omega_n(\delta_{m_{n,x}}),\end{aligned}$$

et d'autre part que,

$$\begin{aligned}\min_{i=k_{n,x}+1,\dots,m_{n,x}} \log\frac{q(V_{k_{n,x},m_{n,x}},u_{k_{n,x}})}{q(V_{i,m_{n,x}},u_i)} &\geq \min_{i=k_{n,x}+1,\dots,m_{n,x}} \log\frac{q(V_{k_{n,x},m_{n,x}},x)}{q(V_{i,m_{n,x}},x)} - 2\omega_n(\delta_{m_{n,x}})\\ &\geq \log\frac{q(V_{k_{n,x},m_{n,x}},x)}{q(V_{k_{n,x}+1,m_{n,x}},x)} - 2\omega_n(\delta_{m_{n,x}}).\end{aligned}$$

Par conséquent $\mathscr{A}_{3,n}\cap\mathscr{A}_{2,n}\subset\mathscr{A}_n$. En remarquant que

$$\mathbb{P}(\mathscr{A}_{3,n}) \geq \mathbb{P}(V_{1,m_{n,x}} > \delta_{m_{n,x}}) + \mathbb{P}(V_{m_{n,x},m_{n,x}} < 1-\delta_{m_{n,x}}) - 1 = 2\mathbb{P}(V_{1,m_{n,x}} > \delta_{m_{n,x}}) - 1 \to 1,$$

puisque $V_{m_{n,x},m_{n,x}} \overset{\mathscr{L}}{=} 1 - V_{1,m_{n,x}}$ et $\mathbb{P}(V_{1,m_{n,x}} > \delta_{m_{n,x}}) = \left(1-\delta_{m_{n,x}}\right)^{m_{n,x}} \to 1$, il reste donc à prouver que $\mathbb{P}(\mathscr{A}_{2,n}) \to 1$. De (Bingham *et al.*, 1987, paragraphe 1.3.1), la condition **(A.1)** implique qu'il existe une fonction $x \mapsto c(x) > 0$ telle que pour tout $\alpha \in (0,1)$,

$$q(\alpha,x) = c(x)\exp\left\{\int_\alpha^1 \frac{\gamma(x)+\Delta(u,x)}{u}\,du\right\}.$$

D'où, pour tout $i \in J_{k_{n,x}}$,

$$\log\frac{q(V_{i,m_{n,x}},x)}{q(V_{i+1,m_{n,x}},x)} = \int_{V_{i,m_{n,x}}}^{V_{i+1,m_{n,x}}} \frac{\gamma(x)+\Delta(u,x)}{u}\,du,$$

et il s'en suit que

$$\log\frac{q(V_{i,m_{n,x}},x)}{q(V_{i+1,m_{n,x}},x)} \geq (\gamma(x) - \bar{\Delta}(V_{k_{n,x}+1,m_{n,x}},x))\log\frac{V_{i+1,m_{n,x}}}{V_{i,m_{n,x}}},$$

ce qui nous conduit à

$$\begin{aligned}\mathbb{P}(\mathscr{A}_{2,n}) &\geq \mathbb{P}\left((\gamma(x) - \bar{\Delta}(V_{k_{n,x}+1,m_{n,x}},x)) \min_{i=1,\dots,k_{n,x}} \log\frac{V_{i+1,m_{n,x}}}{V_{i,m_{n,x}}} > 2\omega_n(\delta_{m_{n,x}})\right)\\ &\geq \mathbb{P}\left(\left\{\min_{i=1,\dots,k_{n,x}} \log\frac{V_{i+1,m_{n,x}}}{V_{i,m_{n,x}}} \geq \frac{4\omega_n(\delta_{m_{n,x}})}{\gamma(x)}\right\} \cap \left\{\bar{\Delta}(V_{k_{n,x}+1,m_{n,x}},x) < \gamma(x)/2\right\}\right)\\ &\geq \mathbb{P}\left(\min_{i=1,\dots,k_{n,x}} \log\frac{V_{i+1,m_{n,x}}}{V_{i,m_{n,x}}} \geq \frac{4\omega_n(\delta_{m_{n,x}})}{\gamma(x)}\right) + \mathbb{P}\left(\bar{\Delta}(V_{k_{n,x}+1,m_{n,x}},x) < \gamma(x)/2\right) - 1\\ &\overset{def}{=} P_{1,m_{n,x}} + P_{2,m_{n,x}} - 1.\end{aligned}$$

Compte tenu du Lemme 1.4.3 sur la représentation de Rényi (1953) qui nous assure que,

$$\{i\log(V_{i,m_{n,x}}^{-1}/V_{i+1,m_{n,x}}^{-1}),\ i \in J_{k_{n,x}}\} \overset{\mathscr{L}}{=} \{F_i,\ i \in J_{k_{n,x}}\},$$

où $F_1, \ldots, F_{k_{n,x}}$ sont des variables aléatoires de loi exponentielle standard, nous avons

$$\begin{aligned} P_{1,m_{n,x}} &= \mathbb{P}\left(\min_{i=1,\ldots,k_{n,x}} \frac{F_i}{i} \geq \frac{4\omega_n(\delta_{m_{n,x}})}{\gamma(x)}\right) = \prod_{i=1}^{k_{n,x}} \exp\left(-\frac{4i\omega_n(\delta_{m_{n,x}})}{\gamma(x)}\right) \\ &= \exp\left(-\frac{2}{\gamma(x)} k_{n,x}(k_{n,x}+1)\omega_n(\delta_{m_{n,x}})\right) \to 1, \end{aligned}$$

puisque $k_{n,x}^2 \omega_n(\delta_{m_{n,x}}) \to 0$. En outre, comme sous **(A.2)**

$$V_{k_{n,x}+1,m_{n,x}} = (k_{n,x}/m_{n,x})(1+o_{\mathrm{P}}(1)) \xrightarrow{\mathbb{P}} 0,$$

alors sous **(A.1)**, la condition $\Delta(\alpha, x) \to 0$ quand $\alpha \to 0$ entraine que $P_{2,m_{n,x}} \to 1$. Ce qui achève la démonstration de la Proposition. □

Démonstration de la Proposition 4.3.2. D'après la Proposition 4.3.1, il existe un événement $\mathscr{A}_n$ avec $\mathbb{P}(\mathscr{A}_n) \to 1$ tel que $(Z_{m_{n,x},m_{n,x}}(x)|\mathscr{A}_n) \overset{\mathscr{L}}{=} (q(V_{1,m_{n,x}}, T_1)|\mathscr{A}_n)$ et donc,

$$\begin{aligned} \mathbb{P}(Z_{m_{n,x},m_{n,x}}(x) < q(\alpha_{m_{n,x}}, x)) &= \mathbb{P}\left(\left\{\log \frac{q(V_{1,m_{n,x}}, T_1)}{q(\alpha_{m_{n,x}}, x)} < 0\right\} \cap \mathscr{A}_n\right) \\ &+ \mathbb{P}\left(\left\{\log \frac{Z_{m_{n,x},m_{n,x}}(x)}{q(\alpha_{m_{n,x}}, x)} < 0\right\} \cap \mathscr{A}_n^C\right) \\ &\overset{def}{=} P_{3,m_{n,x}} + P_{4,m_{n,x}}. \end{aligned} \tag{4.14}$$

Comme $P_{4,m_{n,x}} \leq \mathbb{P}(\mathscr{A}_n^C) \to 0$, il ne nous reste plus qu'à traiter le terme $P_{3,m_{n,x}}$. En introduisant $\delta_{m_{n,x}} = m_{n,x}^{-(1+\delta)}$ et un évènement $\mathscr{A}_{3,n} = \{V_{1,m_{n,x}} \in [\delta_{m_{n,x}}, 1-\delta_{m_{n,x}}]\}$, nous avons

$$\begin{aligned} P_{3,m_{n,x}} &= \mathbb{P}\left(\left\{\log \frac{q(V_{1,m_{n,x}}, T_1)}{q(\alpha_{m_{n,x}}, x)} < 0\right\} \cap \mathscr{A}_n \cap \mathscr{A}_{3,n}\right) \\ &+ \mathbb{P}\left(\left\{\log \frac{q(V_{1,m_{n,x}}, T_1)}{q(\alpha_{m_{n,x}}, x)} < 0\right\} \cap \mathscr{A}_n \cap \mathscr{A}_{3,n}^C\right), \end{aligned}$$

et la formule de l'union binaire nous conduit à

$$\begin{aligned} &\mathbb{P}\left(\left\{\log \frac{q(V_{1,m_{n,x}}, T_1)}{q(\alpha_{m_{n,x}}, x)} < 0\right\} \cap \mathscr{A}_{3,n}\right) + \mathbb{P}(\mathscr{A}_n) - 1 \leq P_{3,m_{n,x}} \\ \leq\ &\mathbb{P}\left(\left\{\log \frac{q(V_{1,m_{n,x}}, T_1)}{q(\alpha_{m_{n,x}}, x)} < 0\right\} \cap \mathscr{A}_{3,n}\right) + \mathbb{P}(\mathscr{A}_{3,n}^C). \end{aligned}$$

De plus, en se référant à la Définition 4.3.1, sous $\mathscr{A}_{3,n}$, on a

$$\left|\log \frac{q(V_{1,m_{n,x}}, T_1)}{q(V_{1,m_{n,x}}, x)}\right| \leq \omega_n(\delta_{m_{n,x}}),$$

et donc

$$\begin{aligned} &\mathbb{P}\left(\left\{\log \frac{q(V_{1,m_{n,x}}, x)}{q(\alpha_{m_{n,x}}, x)} < -\omega_n(\delta_{m_{n,x}})\right\} \cap \mathscr{A}_{3,n}\right) + \mathbb{P}(\mathscr{A}_n) - 1 \leq P_{3,m_{n,x}} \\ \leq\ &\mathbb{P}\left(\left\{\log \frac{q(V_{1,m_{n,x}}, x)}{q(\alpha_{m_{n,x}}, x)} < \omega_n(\delta_{m_{n,x}})\right\} \cap \mathscr{A}_{3,n}\right) + \mathbb{P}(\mathscr{A}_{3,n}^C), \end{aligned}$$

qui entraine

$$\begin{aligned} & \mathbb{P}\left(\log\frac{q(V_{1,m_{n,x}},x)}{q(\alpha_{m_{n,x}},x)} < -\omega_n(\delta_{m_{n,x}})\right) + \mathbb{P}(\mathscr{A}_{3,n}) + \mathbb{P}(\mathscr{A}_n) - 2 \leq P_{1,m_{n,x}} \\ \leq \quad & \mathbb{P}\left(\log\frac{q(V_{1,m_{n,x}},x)}{q(\alpha_{m_{n,x}},x)} < \omega_n(\delta_{m_{n,x}})\right) + \mathbb{P}(\mathscr{A}_{3,n}^C). \end{aligned} \tag{4.15}$$

Maintenant, focalisons nous sur l'équalité suivante

$$\begin{aligned} P_{5,m_{n,x}} &\overset{def}{=} \mathbb{P}\left(\log\frac{q(V_{1,m_{n,x}},x)}{q(\alpha_{m_{n,x}},x)} < \pm\omega_n(\delta_{m_{n,x}})\right) \\ &= \left[\mathbb{P}\left(\log\frac{q(V_1,x)}{q(\alpha_{m_{n,x}},x)} < \pm\omega_n(\delta_{m_{n,x}})\right)\right]^{m_{n,x}} \\ &= \left[\mathbb{P}\left(q(V_1,x) < \mathrm{e}^{\pm\omega_n(\delta_{m_{n,x}})}q(\alpha_{m_{n,x}},x)\right)\right]^{m_{n,x}} \\ &= \left[\mathbb{P}\left(1-V_1 < F\left(\mathrm{e}^{\pm\omega_n(\delta_{m_{n,x}})}q(\alpha_{m_{n,x}},x),x\right)\right)\right]^{m_{n,x}} \\ &= \exp\left[m_{n,x}\log F\left(\mathrm{e}^{\pm\omega_n(\delta_{m_{n,x}})}q(\alpha_{m_{n,x}},x),x\right)\right]. \end{aligned}$$

Puisque $\mathrm{e}^{\pm\omega_n(\delta_{m_{n,x}})}q(\alpha_{m_{n,x}},x) \to \infty$, si l'on introduit la fonction de survie conditionnelle $\bar{F}(.,x) = 1 - F(.,x)$, après un développement limité l'on obtient

$$\begin{aligned} m_{n,x}\log F\left(\mathrm{e}^{\pm\omega_n(\delta_{m_{n,x}})}q(\alpha_{m_{n,x}},x),x\right) &= -m_{n,x}\bar{F}\left(\mathrm{e}^{\pm\omega_n(\delta_{m_{n,x}})}q(\alpha_{m_{n,x}},x),x\right)(1+o(1)) \\ &= -m_{n,x}\alpha_{m_{n,x}}\frac{\bar{F}\left(\mathrm{e}^{\pm\omega_n(\delta_{m_{n,x}})}q(\alpha_{m_{n,x}},x),x\right)}{\bar{F}\left(q(\alpha_{m_{n,x}},x),x\right)}(1+o(1)). \end{aligned}$$

L'hypothèse **(A.1)** nous assure que $\bar{F}(.,x)$ est une fonction à variations régulières d'indice $-1/\gamma(x)$ à l'infini. Ainsi, puisque $\mathrm{e}^{\pm\omega_n(\delta_{m_{n,x}})} \to 1$, nous avons donc que (voir Bingham *et al.*, 1987, Théorème 1.5.2)

$$\frac{\bar{F}\left(\mathrm{e}^{\pm\omega_n(\delta_{m_{n,x}})}q(\alpha_{m_{n,x}},x),x\right)}{\bar{F}\left(q(\alpha_{m_{n,x}},x),x\right)} \to 1.$$

Finalement, on a que

$$P_{5,m_{n,x}} = \left[1-\alpha_{m_{n,x}}(1+o(1))\right]^{m_{n,x}}, \tag{4.16}$$

et en injectant (4.16) dans (4.15) on a

$$\left[1-\alpha_{m_{n,x}}(1+o(1))\right]^{m_{n,x}} + \mathbb{P}(\mathscr{A}_{3,n}) + \mathbb{P}(\mathscr{A}_n) - 2 \leq P_{3,m_{n,x}} \leq \left[1-\alpha_{m_{n,x}}(1+o(1))\right]^{m_{n,x}} + \mathbb{P}(\mathscr{A}_{3,n}^C).$$

Puisque $\mathbb{P}(\mathscr{A}_{3,n}) \to 1$ et $\mathbb{P}(\mathscr{A}_n) \to 1$, il s'en suit que $P_{3,m_{n,x}} \to 0$ sous **(S.1)** et $P_{3,m_{n,x}} \to \mathrm{e}^{-c}$ sous **(S.1)** ou **(S.3)**. L'équation (4.14) conclut la démonstration. □

Démonstration du Théorème 4.3.1. Dans un soucis de simplification, on pose $k_{n,x} = \lfloor m_{n,x}\alpha_{m_{n,x}}\rfloor$. D'après la Proposition 4.3.1, il existe un événement $\mathscr{A}_n$ tel que :

$$\left((m_{n,x}\alpha_{m_{n,x}})^{1/2}\log\frac{\hat{q}_1(\alpha_{m_{n,x}},x)}{q(\alpha_{m_{n,x}},x)}\middle|\mathscr{A}_n\right) \xrightarrow{\mathscr{L}} \left((m_{n,x}\alpha_{m_{n,x}})^{1/2}\log\frac{q(V_{k_{n,x},m_{n,x}},T_{k_{n,x}})}{q(\alpha_{m_{n,x}},x)}\middle|\mathscr{A}_n\right),$$

où $\mathbb{P}(\mathscr{A}_n) \to 1$. En vertu du Lemme 4.6.1, la convergence en loi

$$(m_{n,x}\alpha_{m_{n,x}})^{1/2} \log \frac{q(V_{k_{n,x},m_{n,x}}, T_{k_{n,x}})}{q(\alpha_{m_{n,x}}, x)} \xrightarrow{\mathscr{L}} \mathscr{N}(0, \gamma^2(x)), \tag{4.17}$$

est une condition suffisante pour obtenir

$$(m_{n,x}\alpha_{m_{n,x}})^{1/2} \log \frac{\hat{q}_1(\alpha_{m_{n,x}}, x)}{q(\alpha_{m_{n,x}}, x)} \xrightarrow{\mathscr{L}} \mathscr{N}(0, \gamma^2(x)).$$

Une simple application de la δ-méthode permet de conclure la preuve. Il ne nous reste plus donc qu'à prouver la convergence en loi (4.17). Pour cela, considérons une suite

$$R_n = \left| \log \frac{q(V_{k_{n,x},m_{n,x}}, T_{k_{n,x}})}{q(V_{k_{n,x},m_{n,x}}, x)} \right|$$

et posons $\delta_{m_{n,x}} = m_{n,x}^{-(1+\delta)}$. Remarquons que sous **(S.1)**,

$$\mathbb{P}(R_n \leq \omega_n(\delta_{m_{n,x}})) \geq \mathbb{P}(V_{k_{n,x},m_{n,x}} \in [\delta_{m_{n,x}}, 1 - \delta_{m_{n,x}}]) \to 1.$$

Donc, $R_n = O_{\mathrm{P}}(\omega_n(\delta_{m_{n,x}}))$ et nous avons

$$\log \frac{q(V_{k_{n,x},m_{n,x}}, T_{k_{n,x}})}{q(\alpha_{m_{n,x}}, x)} = \log \frac{q(V_{k_{n,x},m_{n,x}}, x)}{q(\alpha_{m_{n,x}}, x)} + O_{\mathrm{P}}(\omega_n(\delta_{m_{n,x}})). \tag{4.18}$$

En introduisant la fonction log-quantile $g(.) = \log q(., x)$, on a pour tout $\alpha \in (0,1)$,

$$g'(\alpha) = \frac{\Delta(\alpha, x) - \gamma(x)}{\alpha}$$

et un développement limité à l'ordre un entraine que

$$\begin{aligned}(m_{n,x}\alpha_{m_{n,x}})^{1/2} \log \frac{q(V_{k_{n,x},m_{n,x}}, x)}{q(\alpha_{m_{n,x}}, x)} &= (m_{n,x}\alpha_{m_{n,x}})^{1/2} g'(\theta_{m_{n,x}})(V_{k_{n,x},m_{n,x}} - \alpha_{m_{n,x}}) \\ &= \alpha_{m_{n,x}} g'(\theta_{m_{n,x}}) \left(\frac{m_{n,x}}{\alpha_{m_{n,x}}}\right)^{1/2} (V_{k_{n,x},m_{n,x}} - \alpha_{m_{n,x}}),\end{aligned}$$

où $\theta_{m_{n,x}} \in [\min(\alpha_{m_{n,x}}, V_{k_{n,x},m_{n,x}}), \max(\alpha_{m_{n,x}}, V_{k_{n,x},m_{n,x}})]$. Comme $V_{k_{n,x},m_{n,x}} \overset{\mathbb{P}}{\sim} \alpha_{m_{n,x}}$ entraine que $\theta_{m_{n,x}} \overset{\mathbb{P}}{\sim} \alpha_{m_{n,x}} \to 0$, d'après l'hypothèse **(A.1)**, on a

$$\alpha_{m_{n,x}} g'(\theta_{m_{n,x}}) \overset{\mathbb{P}}{\sim} \theta_{m_{n,x}} g'(\theta_{m_{n,x}}) = \Delta(\theta_{m_{n,x}}, x) - \gamma(x) \xrightarrow{\mathbb{P}} -\gamma(x).$$

Le Lemme 4.6.2 implique alors que

$$(m_{n,x}\alpha_{m_{n,x}})^{1/2} \log \frac{q(V_{k_{n,x},m_{n,x}}, x)}{q(\alpha_{m_{n,x}}, x)} \xrightarrow{\mathscr{L}} \mathscr{N}(0, \gamma(x)^2). \tag{4.19}$$

Les relations (4.18) et (4.19) couplées à la remarque suivante

$$(m_{n,x}\alpha_{m_{n,x}})^2 \omega_n(\delta_{m_{n,x}}) \to 0 \Rightarrow (m_{n,x}\alpha_{m_{n,x}})^{1/2} \omega_n(\delta_{m_{n,x}}) \to 0$$

permettent de conclure la preuve. □

Démonstration du Théorème 4.3.2. Puisque $q(.,x)$ est une fonction à variations régulières d'indice $-\gamma(x)$ en zéro, sous **(S.2)** nous avons $q(1/m_{n,x},x)/q(\alpha_{m_{n,x}},x) \sim (m_{n,x}\alpha_{m_{n,x}})^{\gamma(x)} \to c^{\gamma(x)}$ et il s'en suit la décomposition asymptotique suivante

$$\begin{aligned}\log\frac{\hat{q}_1(\alpha_{m_{n,x}},x)}{q(\alpha_{m_{n,x}},x)} &= \log\frac{\hat{q}_1(\alpha_{m_{n,x}},x)}{q(1/m_{n,x},x)} + \frac{q(1/m_{n,x},x)}{q(\alpha_{m_{n,x}},x)} \\ &= \log\frac{\hat{q}_1(\alpha_{m_{n,x}},x)}{q(1/m_{n,x},x)} + \gamma(x)\log(c) + o(1).\end{aligned}$$

Dans la situation **(S.2)**, pour n assez grand, on a $\lfloor m_{n,x}\alpha_{m_{n,x}}\rfloor = \lfloor c\rfloor$. Ainsi, d'après la Proposition 4.3.2, il existe un événement $\mathscr{A}_n$ tel que $\mathbb{P}(\mathscr{A}_n) \to 1$ et

$$\left(\log\frac{\hat{q}_1(\alpha_{m_{n,x}},x)}{q(1/m_{n,x},x)}\middle|\mathscr{A}_n\right) \overset{\mathscr{L}}{=} \left(\log\frac{q(V_{\lfloor c\rfloor,m_{n,x}},T_{\lfloor c\rfloor})}{q(1/m_{n,x},x)}\middle|\mathscr{A}_n\right).$$

En suivant le même raisonnement qu'à la démonstration du Théorème 4.3.1, on obtient

$$\log\frac{q(V_{\lfloor c\rfloor,m_{n,x}},T_{\lfloor c\rfloor})}{q(1/m_{n,x},x)} = \log\frac{q(V_{\lfloor c\rfloor,m_{n,x}},x)}{q(1/m_{n,x},x)} + O_{\mathrm{P}}(\omega_n(\delta_{m_{n,x}})).$$

Pour conclure, il suffit de remarquer que $q(V_{\lfloor c\rfloor,m_{n,x}},x)$ est la $\lfloor c\rfloor$ième plus grande statistique d'ordre associée à une loi à queue lourde. Dans une telle situation, (voir Embrechts *et al.*, 1997, Corollaire 4.2.4) $q(V_{\lfloor c\rfloor,m_{n,x}},x)/q(1/m_{n,x},x)$ converge vers une loi non dégénérée. □

Démonstration du Théorème 4.3.3. On remarquera que

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

Ce qui conduit à la décomposition suivante

$$\begin{aligned}\log\frac{\hat{q}_2(\alpha_{m_{n,x}},x)}{q(\alpha_{m_{n,x}},x)} &= \log\frac{\hat{q}_1(\beta_{m_{n,x}},x)}{q(\beta_{m_{n,x}},x)} \\ &+ \log\left(\frac{\beta_{m_{n,x}}}{\alpha_{m_{n,x}}}\right)(\hat{\gamma}_n(x)-\gamma(x)) \\ &- \log\frac{q(\alpha_{m_{n,x}},x)}{q(\beta_{m_{n,x}},x)} + \gamma(x)\log\left(\frac{\beta_{m_{n,x}}}{\alpha_{m_{n,x}}}\right) \\ &\overset{def}{=} \xi_{4,m_{n,x}} + \xi_{5,m_{n,x}} - \xi_{6,m_{n,x}}.\end{aligned}$$

Premièrement, comme nous l'avons déjà exposé au cours de la démonstration de la Proposition 4.3.2, sous l'hypothèse **(A.1)**, nous avons

$$\log\frac{q(\alpha_{m_{n,x}},x)}{q(\beta_{m_{n,x}},x)} = \int_{\alpha_{m_{n,x}}}^{\beta_{m_{n,x}}}\frac{\gamma(x)+\Delta(u,x)}{u}du,$$

et par conséquent, on peut simplifier $\xi_{6,m_{n,x}}$ et réécrire

$$\xi_{6,m_{n,x}} = \int_{\alpha_{m_{n,x}}}^{\beta_{m_{n,x}}}\frac{\Delta(u,x)}{u}du. \tag{4.20}$$

En bornant le terme (4.20), on a alors

$$|\xi_{6,m_{n,x}}| \leq \bar{\Delta}(\beta_{m_{n,x}}, x) \log\left(\frac{\beta_{m_{n,x}}}{\alpha_{m_{n,x}}}\right).$$

(i) Sous la condition additionnelle (4.5), la loi asymptotique est imposée par $\xi_{4,m_{n,x}}$ et le Théorème 4.3.2 entraine que

$$(m_{n,x}\beta_{m_{n,x}})^{1/2}\xi_{4,m_{n,x}} \xrightarrow{\mathscr{L}} \mathscr{N}(0, \gamma^2(x)) \tag{4.21}$$

et d'après (4.4) et (4.5),

$$(m_{n,x}\beta_{m_{n,x}})^{1/2}\xi_{5,m_{n,x}} = \zeta_{m_{n,x}}\mathscr{V}_n^{-1}(x)\mathscr{V}_n(x)(\hat{\gamma}_n(x) - \gamma(x)) \xrightarrow{\mathbb{P}} 0. \tag{4.22}$$

Aussi, d'après (4.4),

$$(m_{n,x}\beta_{m_{n,x}})^{1/2}|\xi_{6,m_{n,x}}| \leq \zeta_{m_{n,x}}\bar{\Delta}(\beta_{m_{n,x}}, x) \to 0. \tag{4.23}$$

Les relations (4.21), (4.22) et (4.23) permettent de conclure la preuve dans la situation (i).

(ii) Sous la condition additionnelle (4.6), le Théorème 4.3.2 implique que

$$\frac{\mathscr{V}_n(x)}{\log(\beta_{m_{n,x}}/\alpha_{m_{n,x}})}\xi_{4,m_{n,x}} = \mathscr{V}_n(x)\zeta_{m_{n,x}}^{-1}(m_{n,x}\beta_{m_{n,x}})^{1/2}\xi_{4,m_{n,x}} \xrightarrow{\mathbb{P}} 0. \tag{4.24}$$

De plus, comme d'après (4.4),

$$\frac{\mathscr{V}_n(x)}{\log(\beta_{m_{n,x}}/\alpha_{m_{n,x}})}\xi_{5,m_{n,x}} = \mathscr{V}_n(x)(\hat{\gamma}_n(x) - \gamma(x)) \xrightarrow{\mathscr{L}} \mathscr{D}. \tag{4.25}$$

en se référant à (4.6), on a finalement que

$$\frac{v_n(t)}{\log(\beta_{m_{n,x}}/\alpha_{m_{n,x}})}|\xi_{6,m_{n,x}}| \leq \bar{\Delta}(\beta_{m_{n,x}}, t)v_n(t) \to 0. \tag{4.26}$$

Les relations (4.24), (4.25) et (4.26) concluent la preuve dans la situation (ii). □

Démonstration du Corollaire 4.4.1. Elle est immédiate et découle du Théorème 4.3.3. □

Démonstration du Corollaire 4.4.2. En se référant à la démonstration de la Proposition 2.2, on remarque que

$$\log \hat{q}_3(\alpha_{m_{n,x}}, x) = \hat{\gamma}_n(x, W^{\mathrm{H}}) + \log Z_{m_{n,x}-k_{n,x},m_{n,x}}(x) + \hat{\gamma}_n(x, W^{\pi}) \log\left(\frac{1}{\mathrm{e}}\frac{k_{n,x}+1}{m_{n,x}\alpha_{m_{n,x}}}\right),$$

où $\hat{\gamma}_n(x, W^{\mathrm{H}})$ et $\hat{\gamma}_n(x, W^{\pi})$ sont des estimateurs de l'indice de queue conditionnel de la Définition 4.4.1 avec $\mathscr{AB}(x, W^{\pi}) = \mathscr{AB}(x, W^{\mathrm{Z}})$ et $\mathscr{AV}(x, W^{\pi}) = \mathscr{AV}(x, W^{\mathrm{Z}})$ (voir preuve

de la Proposition 2.2.2). Ce qui conduit à la décomposition suivante

$$\begin{aligned}
\log\frac{\hat{q}_3(\alpha_{m_{n,x}},x)}{q(\alpha_{m_{n,x}},x)} &= -\int_{\alpha_{m_{n,x}}}^{\frac{k_{n,x}+1}{m_{n,x}}}\frac{\Delta(u,x)}{u}du+\left(\hat{\gamma}_n(x,W^{\mathrm{H}})-\gamma(x)-\Delta\left(\frac{k_{n,x}}{m_{n,x}}\right)\mathscr{AB}(x,W^{\mathrm{H}})\right)\\
&+ \log\left(\frac{1}{\mathrm{e}}\frac{k_{n,x}+1}{m_{n,x}\alpha_{m_{n,x}}}\right)\left(\hat{\gamma}_n(x,W^{\pi})-\gamma(x)-\Delta\left(\frac{k_{n,x}}{m_{n,x}}\right)\mathscr{AB}(x,W^{\pi})\right)\\
&+ \log\left(\frac{1}{\mathrm{e}}\frac{k_{n,x}+1}{m_{n,x}\alpha_{m_{n,x}}}\right)\Delta\left(\frac{k_{n,x}}{m_{n,x}}\right)\mathscr{AB}(x,W^{\pi})+\mathscr{AB}(x,W^{\mathrm{H}})\Delta\left(\frac{k_{n,x}}{m_{n,x}}\right)\\
&\overset{def}{=} -\xi_{7,m_{n,x}}+\xi_{8,m_{n,x}}+\xi_{9,m_{n,x}}+\xi_{10,m_{n,x}}+\xi_{11,m_{n,x}}.
\end{aligned}$$

La condition $\log(k_{n,x})/\log(m_{n,x}\alpha_{m_{n,x}})\to 0$ entraine que

$$\log\left(\frac{1}{\mathrm{e}}\frac{k_{n,x}+1}{m_{n,x}\alpha_{m_{n,x}}}\right)\sim\log\left(\frac{1}{m_{n,x}\alpha_{m_{n,x}}}\right)\to\infty, \tag{4.27}$$

et d'après (4.7), comme $k_{n,x}^{1/2}\xi_{8,m_{n,x}}\overset{\mathscr{L}}{\to}\mathscr{N}(0,\gamma^2(x))$ alors,

$$\frac{k_{n,x}^{1/2}}{\log\left(1/(m_{n,x}\alpha_{m_{n,x}})\right)}\xi_{8,m_{n,x}}\overset{\mathbb{P}}{\to}0. \tag{4.28}$$

Aussi, d'après (4.7),

$$\begin{aligned}
\frac{k_{n,x}^{1/2}}{\log\left(\frac{1}{\mathrm{e}}\frac{k_{n,x}+1}{m_{n,x}\alpha_{m_{n,x}}}\right)}\xi_{9,m_{n,x}} &= k_{n,x}^{1/2}\left(\hat{\gamma}_n(x,W^{\pi})-\gamma(x)-\Delta\left(\frac{k_{n,x}}{m_{n,x}}\right)\mathscr{AB}(x,W^{\pi})\right)\\
&\overset{\mathscr{L}}{\to} \mathscr{N}\left(0,\gamma^2(x)\mathscr{AV}(x,W^{\pi})\right).
\end{aligned} \tag{4.29}$$

La condition (4.8) montre que

$$\frac{k_{n,x}^{1/2}}{\log\left(\frac{1}{\mathrm{e}}\frac{k_{n,x}+1}{m_{n,x}\alpha_{m_{n,x}}}\right)}\left|\xi_{7,m_{n,x}}\right|\le k_{n,x}^{1/2}(1+o(1))\bar{\Delta}\left(\frac{k_{n,x}+1}{m_{n,x}}\right)\to 0, \tag{4.30}$$

$$\frac{k_{n,x}^{1/2}}{\log\left(\frac{1}{\mathrm{e}}\frac{k_{n,x}+1}{m_{n,x}\alpha_{m_{n,x}}}\right)}\xi_{10,m_{n,x}}=k_{n,x}^{1/2}\Delta\left(\frac{k_{n,x}}{m_{n,x}}\right)\to 0, \tag{4.31}$$

$$\frac{k_{n,x}^{1/2}}{\log\left(\frac{1}{\mathrm{e}}\frac{k_{n,x}+1}{m_{n,x}\alpha_{m_{n,x}}}\right)}\xi_{11,m_{n,x}}\to 0. \tag{4.32}$$

Les relations (4.28), (4.29), (4.30), (4.31) et (4.31) permettent de conclure la preuve. □

5

Estimation de courbes de niveaux extrêmes en design aléatoire

Résumé

Dans ce chapitre, nous nous intéressons à l'estimation des courbes de niveaux extrêmes dans le cas des lois à queues lourdes. Ce problème d'estimation est équivalent à l'étude des quantiles conditionnels quand l'ordre du quantile converge vers un. Nous montrons que sous certaines conditions, il est possible d'estimer de telles courbes au moyen d'un estimateur à noyau de la fonction de survie conditionnelle. En conséquence, ce résultat nous permet d'introduire deux versions lisses de l'estimateur de l'indice de queue conditionnel indispensable lorsque l'on veut extrapoler. Nous établissons la loi limite des estimateurs ainsi construits. Pour conclure, des expériences numériques ainsi que des illustrations seront présentées.

Sommaire

5.1 Introduction **88**
5.2 Cadre de l'étude et définitions des estimateurs **89**
5.2.1 Cadre de l'étude 89
5.2.2 Méthode d'estimation et définitions des estimateurs 89
5.3 Étude théorique des estimateurs **91**
5.3.1 Hypothèses 91
5.3.2 Étude du comportement asymptotique des estimateurs 93
5.4 Application à l'estimation de l'indice de queue conditionnel **98**
5.5 Expériences numériques et illustration sur données réelles **100**
5.5.1 Expériences numériques 100
5.5.2 Illustration sur données réelles 113
5.6 Démonstrations **116**
5.6.1 Preuve des résultats préliminaires 116
5.6.2 Preuve des lois asymptotiques des estimateurs 118

5.1 Introduction

L'objectif premier de ce chapitre est d'introduire puis d'étudier un modèle de régression non-paramétrique pour les quantiles extrêmes conditionnels. Contrairement au problème d'estimation non-paramétrique des quantiles de régression classique qui a été largement considéré (voir la paragraphe 3.3.1 pour une présentation plus détaillée), peu d'attention a été accordée, d'une façon certaine, aux quantiles extrêmes de régression. Toutefois, signalons que dans la littérature, des modèles paramétriques, des méthodes semi-paramétriques et des estimateurs non-paramétriques ont été proposés respectivement dans (Smith, 1989; Davison et Smith, 1990), (Beirlant et Goegebeur, 2003; Hall et Ronde, 2000) et (Davison et Ramesh, 2000; Chavez-Demoulin et Davison, 2005; Chernozhukov, 1998, 2001), consulter la partie 4.1 pour plus de détails. Dans les deux dernières références bibliographiques, les auteurs se focalisent d'une part sur des covariables univariées et d'autre part sur les propriétés de leurs estimateurs pour un échantillon de taille finie. Nombreux sont les auteurs qui se consacrèrent au cas particulier où la loi conditionnelle de la variable d'intérêt en tout point x de la covariable est bornée supérieurement par une fonction $\varphi(x)$ (se référer à la Définition 1.3.3). Dans une telle situation, la fonction $\varphi(.)$ est appelée *frontière* et peut-être estimée au moyen d'une courbe de niveau extrême. On pourra consulter Girard et Jacob (2008) où les auteurs proposent un estimateur à noyau de la fonction $\varphi(.)$ avec un ordre de quantile égal à $(1-1/n)$ où n désigne la taille de l'échantillon. En outre, ils établissent la normalité asymptotique de leur estimateur dans la situation où la variable d'intérêt en tout point x de la covariable est uniformément distribuée sur $[0,\varphi(x)]$. En ce qui concerne l'estimation des frontières, on considère généralement que ses pères fondateurs sont Rényi et Sulanke (1963, 1964) et Geffroy (1964). Le dernier auteur proposait d'estimer la frontière du support avec un simple histogramme basé sur les plus grandes observations. Ainsi, il considérait que la densité conditionnelle de la variable d'intérêt en tout point de la covariable était bornée inférieurement. Des extensions de son modèle ont été proposées pour des densités conditionnelles à décroissance polynomiale dans (Gijbels et Peng, 2000; Hall *et al.*, 1997; Härdle *et al.*, 1995). Rényi et Sulanke (1963, 1964) considéraient quant à eux des supports bidimensionnels convexes qu'ils estimaient en prenant simplement l'enveloppe convexe des observations. Plus récemment, des modèles de régression utilisant des valeurs extrêmes d'un processus ponctuel de Poisson d'intensité inconnue ont été introduits dans (Gardes, 2002; Girard et Jacob, 2004; Girard et Menneteau, 2005; Menneteau, 2008).

Dans ce chapitre, nous nous intéressons au cas où la densité conditionnelle de la variable d'intérêt en tout point de la covariable n'est plus à support borné. Dans un tel contexte, la fonction frontière $\varphi(.)$ n'existe pas et le quantile conditionnel tend vers l'infini quand l'ordre du quantile converge vers un. La partie 5.2 de ce chapitre sera donc dédiée à la formulation mathématique du problème investigué. C'est notamment dans celle-ci que nous introduirons les outils permettant de traiter le sujet. Puis, dans la partie 5.3, nous nous attarderons sur l'étude théorique des outils ainsi introduits. La partie 5.4 de ce chapitre sera consacrée à l'application de nos résultats à l'estimation de

l'indice de queue conditionnel indispensable lorsque l'on veut estimer des quantités dont la probabilité d'observation est quasi nulle. Des expériences numériques suivies d'une illustration en télédétection hyperspectrale seront présentées dans la partie 5.5. Enfin, dans la partie 5.6, nous conclurons le chapitre par la preuve de nos différents résultats.

5.2 Cadre de l'étude et définitions des estimateurs

Cette partie se subdivise en deux paragraphes. Dans le premier, nous exposons le contexte de l'étude et dans le second nous présentons notre méthode d'estimation ainsi que les estimateurs associés.

5.2.1 Cadre de l'étude

Soient $\{(X_i, Y_i),\ i = 1, \dots, n\}$ des copies indépendantes du couple aléatoire $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$ où Y est une variable d'intérêt associée à une covariable X. Pour tout $x \in \mathbb{R}^p$ et pour toute suite réelle $\alpha_n \to 0$, on se propose d'estimer les courbes de niveaux extrêmes définies comme les graphes de fonctions $x \in \mathbb{R}^p \mapsto q(\alpha_n|x) \in \mathbb{R}$ vérifiant

$$\mathbb{P}\left(Y > q(\alpha_n|x)|X = x\right) = \alpha_n,$$

lorsque la fonction de survie conditionnelle de Y sachant $X = x$ est à variations régulières d'indice $-1/\gamma(x)$ à l'infini. De façon plus précise, compte tenu de la remarque faite au sous-paragraphe 1.3.1.2, ceci signifie que pour tout $y > 0$,

$$\bar{F}\left(y|x\right) \overset{def}{=} 1 - F(y|x) = y^{-1/\gamma(x)}\ell(y|x), \tag{5.1}$$

avec $\gamma(.)$ une fonction inconnue et positive de la covariable x et $\ell(.|x)$ une fonction à variations lentes à l'infini. Tel que nous l'avons exposé au paragraphe 4.2.1, cette hypothèse revient à supposer que la loi conditionnelle de Y sachant $X = x$ est à queue lourde, i.e $F(.|x) \in \mathscr{D}(Fréchet)$. Dans un tel contexte, la fonction $\gamma(x)$ est l'indice de queue conditionnel.

5.2.2 Méthode d'estimation et définitions des estimateurs

Un estimateur naturel de la fonction $x \in \mathbb{R}^d \mapsto q(\alpha_n|x)$ appelée *quantile conditionnel* est donné par l'inverse généralisé de l'estimateur de la fonction de survie conditionnelle défini par :

$$\hat{q}_n(\alpha_n|x) \overset{def}{=} \hat{\bar{F}}_n^{\leftarrow}(\alpha_n|x) = \inf\left\{t,\ \hat{\bar{F}}_n(t|x) \leq \alpha_n\right\}. \tag{5.2}$$

Ainsi, l'estimation d'une courbe de niveau $\alpha_n \to 0$ nécessite d'estimer la probabilité

$$\bar{F}\left(y|x\right) \to 0 \text{ lorsque } y \to \infty.$$

Ici, nous parlons d'extrême lorsque α_n converge vers zéro quand n tend vers l'infini. Par ailleurs, notons que la condition $\alpha_n \to 0$ quand $n \to \infty$ en ce qui concerne le quantile

est équivalente à la condition $y \to \infty$ quand $n \to \infty$ en ce qui concerne la fonction de survie conditionnelle.

Afin d'estimer la fonction de survie conditionnelle de Y sachant $X = x$, on peut utiliser la méthode de la *fenêtre mobile* introduite au chapitre 4 (on pourra aussi se référer au paragraphe 3.3.1.1) et pour tout $(x, y) \in \mathbb{R}^p \times \mathbb{R}$ on a :

$$\hat{\bar{F}}_n(y|x) = \sum_{i=1}^{n} \mathbb{1}_{\{Y_i > y\}} \mathbb{1}_{\{X_i \in B(x,h_n)\}} \Big/ \sum_{i=1}^{n} \mathbb{1}_{\{X_i \in B(x,h_n)\}},$$

où $B(x, h_n)$ est une boule centrée en x de rayon h_n strictement positif. Malheureusement cet estimateur présente le désavantage d'être discontinu par nature. Sa généralisation naturelle est l'estimateur à noyau dont la simplicité de construction et la facilité d'utilisation vont de pair avec ses bonnes propriétés asymptotiques.

Ainsi, pour estimer la fonction de survie conditionnelle de Y sachant $X = x$, on se propose d'utiliser un estimateur à noyau introduit par Collomb (1980) [1]. Il est défini pour $(x, y) \in \mathbb{R}^p \times \mathbb{R}$ par :

$$\hat{\bar{F}}_n(y|x) = \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right) \mathbb{1}_{\{Y_i > y\}} \Big/ \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right), \tag{5.3}$$

où la fonction $K(.)$ appelée *noyau* est une densité de probabilité sur $\mathbb{R}^p$ et h_n est une suite non aléatoire telle que $h_n \to 0$ quand $n \to \infty$ appelée *paramètre de lissage*. Sans rentrer dans tous les détails, à l'exception du *noyau uniforme*, il s'avère que le choix du noyau n'a pas d'influence majeure sur l'estimation. L'estimateur (5.3) peut se mettre sous la forme

$$\hat{\bar{F}}_n(y|x) = \frac{\hat{\psi}_n(x, y)}{\hat{g}_n(x)},$$

où le dénominateur de la fonction de survie conditionnelle de Y sachant $X = x$ est l'estimateur à noyau classique de la densité $g(x)$ de X et la fonction $\hat{\psi}_n(x, y)$ est un estimateur de $\psi(y, x) = \bar{F}(y|x)g(x)$.

Dans ce chapitre, nous intéressons à l'estimation du réel $q(\alpha_n|x)$ en fonction de la vitesse de convergence α_n vers zéro. Nous avons envisagé les deux situations suivantes :

(D.1) Dans la première situation, la suite α_n *converge lentement vers zéro* en ce sens que $nh_n^p\alpha_n \to \infty$ quand $n \to \infty$.

(D.2) Dans la deuxième situation, on autorise la suite α_n à converger vers zéro plus vite que dans la situation **(D.1)**, c'est-à-dire que l'on ne suppose plus que $nh_n^p\alpha_n \to \infty$ quand $n \to \infty$. On dit alors que α_n *converge rapidement vers zéro*.

1. Dans les faits, Collomb (1980) introduisit juste un estimateur la fonction de répartition conditionnelle.

Dans la situation **(D.1)**, la condition $nh_n^p\alpha_n \to \infty$ quand $n \to \infty$ revient à supposer que le quantile $q(\alpha_n|x)$ ne tend pas trop vite vers l'infini quand n tend vers l'infini. Dans une telle situation, l'estimation du quantile extrême conditionnel requiert d'interpoler à l'intérieur de l'ensemble des données car il y a presque sûrement un point de l'échantillon dans la région $B(x,h_n) \times \big(q(\alpha_n|x),\infty\big)$ de $\mathbb{R}^{p+1}$ où $B(x,h_n)$ est une boule centrée en x de rayon h_n (voir Lemme 5.3.3). On se propose alors d'estimer le quantile extrême conditionnel par (5.2).

De façon équivalente, dans la situation **(D.2)**, il est supposé que le quantile $q(\alpha_n|x)$ tend vite vers l'infini quand n tend vers l'infini. Dans un tel contexte, l'estimation du quantile conditionnel peut nécessiter d'extrapoler au-delà des observations. On se propose donc d'adapter l'estimateur de Weissman (1978) au cas conditionnel. On estime alors $q(\alpha_n|x)$ par

$$\hat{q}_n^{\mathrm{W}}(\alpha_n|x) = \hat{q}_n\big(\beta_n|x\big)\big(\alpha_n/\beta_n\big)^{-\hat{\gamma}_n(x)}, \tag{5.4}$$

où β_n est telle que $nh_n^p\beta_n \to \infty$ quand $n \to \infty$. $\hat{\gamma}_n(x)$ est un estimateur de l'indice de queue conditionnel dont nous donnons deux exemples à la partie 5.4. On notera des similitudes entre les situations **(D.1)**, **(D.2)** et les situations **(S.1)**, **(S.2)**, **(S.3)** de la partie 4.2.2.

5.3 Étude théorique des estimateurs

Cette partie est composée de deux paragraphes. Le premier nous permet d'énumerer les hypothèses qui nous servirons à étudier le comportement asymptotiques de nos estimateurs et le second est consacrée à l'étude théorique proprement dite.

5.3.1 Hypothèses

Dans ce paragraphe, on présente toutes les conditions utiles pour établir la loi asymptotique de tous nos estimateurs. Dans tout ce qui suit, on désigne par $d(x,x')$ la distance entre deux points $(x,x') \in \mathbb{R}^p \times \mathbb{R}^p$.

Hypothèses sur la fonction à variations lentes

(F.1) $\ell(.|x)$ est normalisée.

D'après L'hypothèse **(F.1)**, la fonction à variations lentes peut se réécrire sous la forme (se référer au sous-paragraphe 1.3.1.1)

$$\ell(y|x) = c(x)\exp\left(\int_1^y \frac{\varepsilon(u|x)}{u}du\right), \tag{5.5}$$

avec $c(.)$ une fonction positive et $\varepsilon(y|x) \to 0$ quand $y \to \infty$. Ceci implique donc que la fonction $\ell(.|x)$ est dérivable et pour tout $x \in \mathbb{R}^p$, la fonction auxiliaire $\varepsilon(y|x)$ est définie par

$$\varepsilon(y|x) = y\frac{\ell'(y|x)}{\ell(y|x)}.$$

Poser par hypothèse (comme on le fait dans ce chapitre) que la fonction $\bar{F}\big(y|x\big)$ est à variations régulières d'indice $-1/\gamma(x)$ à l'infini revient à supposer implicitement que pour tout $x \in \mathbb{R}^p$ et pour tout $\lambda > 0$,

$$\lim_{y\to\infty} \frac{\ell(\lambda y|x)}{\ell(y|x)} = 1. \tag{5.6}$$

Ainsi, la fonction $\varepsilon(y|x)$ qui représente le terme de *biais* occupera une place prépondérante dans cette étude puisqu'elle contrôle la vitesse de convergence dans (5.6). En effet, plus la convergence de $\ell(\lambda y|x)/\ell(y|x)$ vers un quand y tend vers l'infini est rapide, et plus facile est l'estimation des courbes de niveaux extrême ou de l'indice de queue conditionnel. Le contrôle du terme de biais est d'une importance capitale quand on cherche à établir des résultats sur la normalité asymptotique des estimateurs de l'indice de queue et de quantile extrême. Il parait donc intéressant de préciser au moins une condition permettant de contrôler sa vitesse de convergence asymptotique vers zéro. Certains auteurs supposent parfois que le biais est une fonction à variations régulières d'indice $\rho < 0$ à l'infini et des auteurs tels que Alves *et al.* (2003a,b) et Gomes *et al.* (2002) en ont proposés des méthodes d'estimation. En ce qui nous concerne, nous introduisons l'hypothèse **(F.2)** connue sous le nom de *condition du second ordre* et dont le but est de contrôler le comportement de la fonction de survie conditionnelle de Y sachant $X = x$ par rapport à sa première variable.

(F.2) $|\varepsilon(.|x)|$ est continue asymptotiquement décroissante.

Hypothèses de régularité lipschitzienne

(L.1) Il existe $c_\gamma > 0$ tel que $\left|\dfrac{1}{\gamma(x)} - \dfrac{1}{\gamma(x')}\right| \le c_\gamma d(x, x')$.

(L.2) Il existe $c_\ell > 0$ et $y_0 > 1$ tels que $\sup\limits_{y\ge y_0}\left|\dfrac{\log \ell(y|x)}{\log y} - \dfrac{\log \ell(y|x')}{\log y}\right| \le c_\ell d(x, x')$.

(L.3) Il existe $c_g > 0$ tel que $\big|g(x) - g(x')\big| \le c_g d(x, x')$.

Hypothèse sur le noyau

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

Les hypothèses **(F.1)**, **(L.1)**et **(L.2)** nous assurent la régularité de la loi conditionnelle de Y sachant $X = x$. La supposition **(L.2)** sert à contrôler localement les variations de la fonction de survie conditionnelle de Y sachant $X = x$ par rapport à sa seconde variable dans un voisinage de taille h_n quand la quantité d'intérêt y tend vers l'infini. Contrairement au chapitre 4 (cf. Définition 4.3.1), nous ne nous servirons pas de la définition sur la plus grande oscillation de la fonction log-quantile. Ici, nous procédons autrement [2]

2. De façon plus précise, la différence avec le chapitre 4 est qu'ici on fait des hypothèses sur $\bar{F}(.|x)$ plutôt que sur son inverse $q(.|x)$.

car notre objectif est de produire des résultats de convergence asymptotique de nos estimateurs sous des conditions de type lipschitz. Les conditions **(L.3)** et **(K)** sont des hypothèses classiques sur l'estimation non-paramétrique par la méthode du noyau.

5.3.2 Étude du comportement asymptotique des estimateurs

5.3.2.1 Résultats préliminaires

Nous ne saurions présenter la loi asymptotique de nos estimateurs sans en introduire les outils nous permettant de mener à bien l'étude théorique de telles quantités. Il semble donc inéluctable de commencer par un résultat dédié au contrôle des variations de la fonction de survie conditionnelle par rapport à la covariable x dans un voisinage de taille h_n quand la quantité d'intérêt $y \to \infty$.

Lemme 5.3.1. *Supposons les conditions* **(L.1)** *et* **(L.2)** *satisfaites. Si* $y_n \to \infty$ *et* $h_n \log y_n \to 0$ *quand* $n \to \infty$*, alors*

$$\sup_{d(x,x')\leq h_n} \left| \frac{\bar{F}\left(y_n|x\right)}{\bar{F}\left(y_n|x'\right)} - 1 \right| = O(h_n \log y_n).$$

Le Lemme 5.3.1 montre que sous condition de Lipschitz, le rapport de deux fonctions de survies conditionnelles évaluées aux points $(x, x') \in \mathbb{R}^p \times \mathbb{R}^p$ distants au plus de h_n converge uniformément vers un lorsque n tend vers l'infini.

À présent, donnons une conséquence de la condition du second ordre sur la fonction quantile extrême conditionnel.

Lemme 5.3.2. *Supposons les conditions* **(F.1)** *et* **(F.2)** *satisfaites.*

(i) Soient $(\alpha_n)_{n\geq 1}$ *et* $\left(\beta_n\right)_{n\geq 1}$ *deux suites positives telles que* $0 < \beta_n < \alpha_n$ *avec* $\alpha_n \to 0$ *quand* $n \to \infty$*. Alors*

$$\left|\log q\left(\beta_n|x\right) - \log q\left(\alpha_n|x\right) + \gamma(x)\log(\beta_n/\alpha_n)\right| = O(\log(\beta_n/\alpha_n)\varepsilon\left(q\left(\alpha_n|x\right)\right)|x\right).$$

(ii) Si de plus $\liminf \beta_n/\alpha_n > 0$*, alors*

$$\frac{\beta_n^{\gamma(x)} q\left(\beta_n|x\right)}{\alpha_n^{\gamma(x)} q\left(\alpha_n|x\right)} = 1 + O(\varepsilon\left(q\left(\alpha_n|x\right)\right)|x\right).$$

Ce résultat stipule que le quantile extrême conditionnel est à décroissance polynomiale d'indice $-\gamma(x)$ puisque $\varepsilon\left(q\left(\alpha_n|x\right)\right)|x\right) \to 0$ quand $n \to \infty$. On pourra se référer au paragraphe 4.2.1 pour une analogie.

On peut maintenant donner une interprétation géométrique de la condition $nh_n^p\alpha_n \to \infty$ introduite précédemment lors de la présentation des situations que l'on envisage d'examiner dans ce chapitre (voir paragraphe 5.2.2).

Lemme 5.3.3. *Supposons les conditions* **(L.1)**, **(L.2)** *et* **(L.3)** *satisfaites. Considérons la région de* $\mathbb{R}^{p+1}$ *définie par* $R_n(x) = B(x,h_n) \times (q(\alpha_n|x),\infty)$ *où* $x \in \mathbb{R}^p$ *est tel que* $g(x) > 0$. *Si* $h_n \log q(\alpha_n|x) \to 0$ *quand* $n \to \infty$, *alors* $\mathbb{P}(\exists i \in \{1,\dots,n\}, (X_i,Y_i) \in R_n(x)) \to 1$ *quand* $n \to \infty$ *si et seulement si,* $nh_n^p\alpha_n \to \infty$.

Le Lemme 5.3.3 montre qu'il n'est pas judicieux d'utiliser (5.2) pour estimer une courbe de niveau extrême dans la situation **(D.2)** où $nh_n^p\alpha_n$ peut converger vers une constante. Une similitude peut être faite avec les situations **(S.2)** et **(S.3)** du paragraphe 4.2.2 (voir chapitre 4).

Remarque 5.3.1. *Comme* α_n *converge vers zéro, alors la condition* $nh_n^p\alpha_n \to \infty$ *est équivalente à la condition* $nh_n^p\bar{F}\left(y_n|x\right) \to \infty$ *où la quantité d'intérêt* y_n *tend vers l'infini. Ainsi, le Lemme 5.3.3 peut se réécrire en remplaçant* $R_n(x) = B(x,h_n) \times (q(\alpha_n|x),\infty)$ *par* $R_n^*(x) = B(x,h_n) \times (y_n,\infty)$ *et* $nh_n^p\alpha_n \to \infty$ *par* $nh_n^p\bar{F}\left(y_n|x\right) \to \infty$.

5.3.2.2 Loi asymptotique des estimateurs

Comme nous l'avons fait remarquer, l'estimateur (5.3) peut se mettre sous la forme $\hat{\bar{F}}_n\left(y|x\right) = \hat{\psi}_n(x,y)/\hat{g}_n(x)$ où

$$\hat{\psi}_n(x,y) = \frac{1}{n}\sum_{i=1}^{n} K_h(x-X_i)\mathbb{1}_{\{Y_i>y\}} \text{ et} \tag{5.7}$$

$$\hat{g}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x-X_i), \tag{5.8}$$

avec $K_h(t) \overset{def}{=} K(t/h_n)/h_n^p$. Le comportement asymptotique de l'estimateur $\hat{g}_n(x)$ a été étudié par Collomb (1980). En particulier, sous les conditions **(L.3)** et et **(K)**, l'auteur montre que

$$\hat{g}_n(x) \xrightarrow{\mathbb{P}} g(x).$$

On pourra consulter le paragraphe 5.6.1 où deux résultats de l'auteur sont clairement exposés au Lemme 5.6.1. Ainsi, le premier résultat de ce paragraphe sera dédié aux propriétés asymptotiques de la variable aléatoire $\hat{\psi}_n(x,y)$ quand la variable d'intérêt y tend vers l'infini.

Lemme 5.3.4. *Supposons que les conditions* **(L.1)**, **(L.2)**, **(L.3)** *et* **(K)** *soient réalisées. Soit* $(y_n)_{n\geq 1}$ *une suite positive satisfaisant les conditions* $y_n \to \infty$, $h_n \log y_n \to 0$ *et* $nh_n^p\bar{F}\left(y_n|x\right) \to \infty$ *quand* $n \to \infty$ *et soit* $\{a_j, j = 1,\dots,J\}$ *une suite strictement positive et croissante. Alors, pour tout* $x \in \mathbb{R}^p$ *tel que* $g(x) > 0$,

(i) $\left\{\mathbb{E}(\hat{\psi}_n(a_j y_n,x))\right\}_{\{j=1,\dots,J\}} = \left\{\psi(a_j y_n,x)(1+O(h_n \log y_n))\right\}_{\{j=1,\dots,J\}}$.

(ii) $\left\{\sqrt{nh_n^p\psi(y_n,x)}\left(\frac{\hat{\psi}_n(a_j y_n,x)-\mathbb{E}(\hat{\psi}_n(a_j y_n,x))}{\psi(a_j y_n,x)}\right)\right\}_{\{j=1,\dots,J\}} \xrightarrow{\mathscr{L}} \mathcal{N}\left(0_{\mathbb{R}^J}, \|K\|_2^2 C(x)\right)$,
où $C_{j,j'}(x) = a_{j\wedge j'}^{1/\gamma(x)}$ *pour* $(j,j') \in \{1,\dots,J\}^2$.

D'après ce Lemme, pour tout $x \in \mathbb{R}^p$, l'estimateur $\hat{\psi}_n\left(y_n, x\right)$ est asymptotiquement gaussien et sans biais lorsque la suite y_n ne tend pas trop vite vers l'infini. Par conséquent, sous les conditions de Lipschitz énoncées précédemment au paragraphe 5.3.1, on a

$$\frac{\hat{\psi}_n\left(y_n, x\right)}{\psi\left(y_n, x\right)} \xrightarrow{\mathbb{P}} 1.$$

Le Théorème suivant établit la normalité asymptotique de l'estimateur $\hat{\bar{F}}_n(.|x)$.

Théorème 5.3.1. *Supposons que les conditions* **(L.1)**, **(L.2)**, **(L.3)** *et* **(K)** *soient réalisées. Soit* $(y_n)_{n\geq 1}$ *une suite positive satisfaisant les conditions* $y_n \to \infty$, $nh_n^p \bar{F}\left(y_n|x\right) \to \infty$ *et* $nh_n^{p+2}\log^2(y_n)\bar{F}(y_n|x) \to 0$ *quand* $n \to \infty$ *et soit* $\{a_j, j = 1, \ldots, J\}$ *une suite strictement positive et croissante. Alors, pour tout* $x \in \mathbb{R}^p$ *tel que* $g(x) > 0$,

$$\left\{\sqrt{nh_n^p \bar{F}\left(y_n|x\right)}\left(\frac{\hat{\bar{F}}_n\left(a_j y_n|x\right)}{\bar{F}\left(a_j y_n|x\right)} - 1\right)\right\}_{\{j=1,\ldots,J\}} \xrightarrow{\mathscr{L}} \mathscr{N}\left(0_{\mathbb{R}^J}, \frac{\|K\|_2^2}{g(x)} C(x)\right),$$

où $C_{j,j'}(x) = a_{j\wedge j'}^{1/\gamma(x)}$ *pour* $(j, j') \in \{1, \ldots, J\}^2$.

Il apparaît que l'estimateur de la fonction de survie conditionnelle de Y sachant $X = x$ est asymptotiquement gaussienne avec une variance asymptotique inversement proportionnelle aux nombre de points dans la région $B(x, h_n) \times (y_n, \infty) \in \mathbb{R}^{p+1}$.

Ainsi que l'illustre le graphique de la Figure 5.1, $nh_n^p \bar{F}\left(y_n|x\right) \to \infty$ est une condition nécessaire et suffisante pour que l'on ait presque sûrement au moins un point de l'échantillon $\{(X_i, Y_i),\ i = 1, \ldots, n\}$ dans la région $B(x, h_n) \times (y_n, \infty) \in \mathbb{R}^{p+1}$. On remarquera que si y_n est borné alors, on retrouve la condition de normalité asymptotique classique $nh_n^p \to \infty$.

La condition $nh_n^{p+2}\log^2(y_n)\bar{F}(y_n|x) \to 0$ est une condition pour que le carré du bais asymptotique, de l'ordre de

$$\left(h_n \log y_n\right)^2,$$

soit négligeable devant la variance asymptotique, de l'ordre de

$$\frac{1}{nh_n^p \bar{F}\left(y_n|x\right)}.$$

Le résultat du Théorème 5.3.1 peut être comparé à celui de Einmahl (1990) qui a étudié le comportement asymptotique de la fonction de survie empirique dans le cas non conditionnel sans énoncer une hypothèse sur la loi de l'échantillon.

Un autre parallèle peut être fait avec le résultat de (Berlinet *et al.*, 2001, Théorème 6.3) [3] où les auteurs trouvent pour un $y_n = y$ fixé et fini, une variance asymptotique proportionnelle à

$$\frac{\bar{F}\left(y|x\right)\left(1 - \bar{F}\left(y|x\right)\right)}{nh_n^p} \frac{\|K\|_2^2}{g(x)}.$$

3. On pourra se référer aux commentaires du Théorème 3.3.1 de cette thèse.

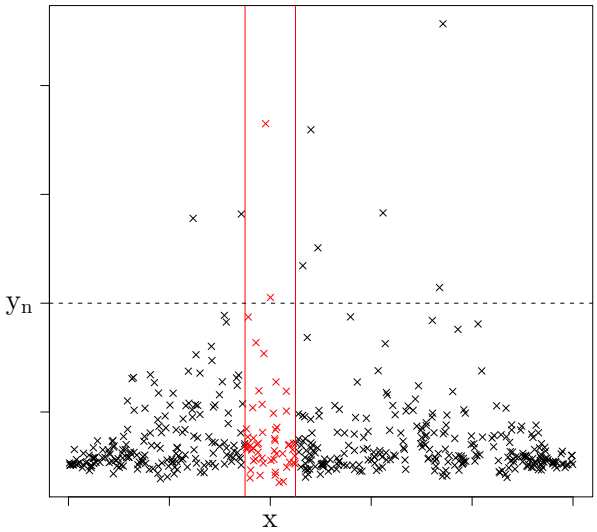


FIGURE 5.1 – Illustration géométrique de la condition $nh_n^p \bar{F}\left(y_n|x\right) \to \infty$ avec $p = 1$. Nuage de points $\{(X_i, Y_i), i = 1, \ldots, 1000\}$ ($\times\times\times\times$) et points dans la boule $B(x, h_n)$ (rouge). Dans cet exemple on a deux points de l'échantillon (ici les points en rouge au-dessus de y_n) dans la région $B(x, h_n) \times (y_n, \infty) \in \mathbb{R}^2$.

Ce qui montre bien que la variance asymptotique trouvée au Théorème 5.3.1 est asymptotiquement plus grande puisque qu'elle fait intervenir le terme additionnel $1/\bar{F}\left(y_n|x\right) \to \infty$ quand $n \to \infty$. Intéressons nous maintenant à l'estimation du quantile extrême conditionnel d'ordre $\alpha_n \to 0$ quand $n \to \infty$ dans la situation **(D.1)**.

Théorème 5.3.2. *Supposons que les conditions* **(F.1)**, **(L.1)**, **(L.2)**, **(L.3)** *et* **(K)** *soient réalisées. Soit* $(\alpha_n)_{n\geq 1}$ *une suite positive satisfaisant les conditions* $\alpha_n \to 0$, $nh_n^p \alpha_n \to \infty$ *et* $nh_n^{p+2} \alpha_n \log^2(\alpha_n) \to 0$ *quand* $n \to \infty$ *et soit* $\{\tau_j, j = 1, \ldots, J\}$ *une suite strictement positive et décroissante. Alors, pour tout* $x \in \mathbb{R}^p$ *tel que* $g(x) > 0$,

$$\left\{\sqrt{nh_n^p \alpha_n}\left(\frac{\hat{q}_n\left(\tau_j \alpha_n|x\right)}{q\left(\tau_j \alpha_n|x\right)} - 1\right)\right\}_{\{j=1,\ldots,J\}} \xrightarrow{\mathscr{L}} \mathscr{N}\left(0_{\mathbb{R}^J}, \gamma^2(x)\frac{\|K\|_2^2}{g(x)}\Sigma\right),$$

où $\Sigma_{j,j'} = 1/\tau_{j\wedge j'}$ *pour* $(j, j') \in \{1, \ldots, J\}^2$.

Dans la situation **(D.1)**, la variance asymptotique étant inversement proportionnelle à $nh_n^p \alpha_n$, l'estimation des courbes de niveaux extrêmes est d'autant plus stable que l'on s'éloigne de la frontière de l'ensemble des données. Aussi, comme cette variance est proportionnelle à $\gamma^2(x)$, ceci implique que l'estimation de $q(\alpha_n|x)$ est plus difficile pour des grandes valeurs de l'indice de queue conditionnel.

Aussi, du point de vue de la variance asymptotique, notre modèle est équivalent à une fonction de survie conditionnelle avec un indice de queue constant $\tilde{\gamma}(x) = 1$ et une densité de probabilité proportionnelle à $\tilde{g}(x) = g(x)/\gamma^2(x)$. Ainsi, le nombre de points dans la boule $B(x, h_n)$ est asymptotiquement inversement proportionnel à $\gamma^2(x)$. Ce qui justifie une fois de plus qu'une valeur élevée de l'indice de queue conditionnel en un point $x \in \mathbb{R}^p$ implique donc une estimation difficile de $q(\alpha_n|x)$.

Enfin, le résultat du Théorème 5.3.2 peut être comparé à celui énoncé dans (Berlinet *et al.*, 2001, Théorème 6.4) pour $\alpha_n = \alpha \in]0,1[$ fixé où les auteurs trouvent une variance asymptotique de l'ordre de

$$\frac{\alpha(1-\alpha)}{nh_n^p}\frac{\|K\|_2^2}{g(x)}.$$

Ainsi, notre variance asymptotique est aussi plus grande car elle fait intervenir le terme additionnel $1/\alpha_n \to \infty$ quand $n \to \infty$.

Les conditions $nh_n^p\alpha_n \to \infty$ et $nh_n^{p+2}\alpha_n \log^2(\alpha_n) \to 0$ entrainent

$$\frac{n\alpha_n}{\log^p(1/\alpha_n)} \to \infty,$$

qui implique

$$\alpha_n > \frac{\log^p(n)}{n}.$$

Ce qui montre bien que l'on ne peut pas extrapoler en utilisant (5.2).

Théorème 5.3.3. *Supposons que les conditions* **(F.1)**, **(L.1)**, **(L.2)**, **(L.3)** *et* **(K)** *soient réalisées. Soit* $(\beta_n)_{n\geq 1}$ *une suite positive satisfaisant les conditions* $\beta_n \to 0$, $nh_n^p\beta_n \to \infty$ *et* $nh_n^{p+2}\beta_n \log^2(\beta_n) \to 0$ *quand* $n \to \infty$. *Soit* $\hat{\gamma}_n(x)$ *un estimateur de l'indice de queue tel que*

$$\sqrt{nh_n^p\beta_n}\left(\hat{\gamma}_n(x) - \gamma(x)\right) \xrightarrow{\mathscr{L}} \mathcal{N}\left(0, v^2(x)\right) \text{ avec } v(x) > 0. \tag{5.9}$$

Si $(\alpha_n)_{n\geq 1}$ *est une suite positive tendant vers zéro et telle que* $\alpha_n/\beta_n \to 0$ *alors, pour tout* $x \in \mathbb{R}^p$, *on a*

$$\frac{\sqrt{nh_n^p\beta_n}}{\log(\beta_n/\alpha_n)}\left(\frac{\hat{q}_n^{\mathrm{W}}(\alpha_n|x)}{q(\alpha_n|x)} - 1\right) \xrightarrow{\mathscr{L}} \mathcal{N}\left(0, v^2(x)\right).$$

La loi asymptotique de $\hat{q}_n^{\mathrm{W}}(.|x)$ dépend du comportement de $\hat{\gamma}_n(x)$. Si l'estimateur $\hat{\gamma}_n(x)$ converge moins vite que dans (5.9), alors la loi limite de $\hat{q}_n^{\mathrm{W}}(.|x)$ peut dépendre du comportement de $\hat{q}_n(.|x)$. Pour un exemple d'une telle situation, on pourra se référer au Théorème 4.3.3. Aussi, remarquons que l'estimateur $\hat{q}_n^{\mathrm{W}}(.|x)$ peut être utilisé dans les deux situations **(D.1)** et **(D.2)**.

Afin de pouvoir utiliser convenablement ce résultat, il semble adéquat de disposer au moins d'un estimateur à noyau de l'indice de queue conditionnel. C'est à quoi l'on s'attelle dans la partie suivante.

5.4 Application à l'estimation de l'indice de queue conditionnel

Du Théorème 5.3.2, on déduit des estimateurs de l'indice de queue conditionnel. L'intérêt de construire de tels estimateurs est double. D'une part il nous permet de construire des intervalles de confiance du quantile extrême conditionnel $q(\alpha_n|x)$ et d'autre part de pouvoir extrapoler au-delà des observations, c'est-à-dire de pouvoir utiliser en pratique l'estimateur $\hat{q}_n^{\mathrm{W}}(\alpha_n|x)$. Le premier estimateur de $\gamma(x)$ que nous proposons est basé sur trois statistiques d'ordre.

Définition 5.4.1. *Soit $(\beta_n)_{n\geq 1}$ une suite positive telle que $\beta_n \to 0$ et $nh_n^p\beta_n \to \infty$ quand $n \to \infty$. Un estimateur à noyau de type Pickands (1975) est donné par*

$$\hat{\gamma}_n^{\mathrm{P}}(x) = \frac{1}{\log 2}\log\left(\frac{\hat{q}_n(\beta_n|x) - \hat{q}_n(2\beta_n|x)}{\hat{q}_n(2\beta_n|x) - \hat{q}_n(4\beta_n|x)}\right).$$

Corollaire 5.4.1. *Supposons la condition* **(F.2)** *vérifiée. Sous les hypothèses du Théorème 5.3.2, si $\sqrt{nh_n^p\beta_n}\,\varepsilon\left(q(2\beta_n|x)|x\right) \to 0$ quand $n \to \infty$, alors*

$$\sqrt{nh_n^p\beta_n}\left(\hat{\gamma}_n^{\mathrm{P}}(x) - \gamma(x)\right) \xrightarrow{\mathscr{L}} \mathscr{N}\left(0, \frac{\|K\|_2^2}{g(x)}\frac{\gamma^2(x)(2^{2\gamma(x)+1}+1)}{4(\log 2)^2(2^{\gamma(x)}-1)^2}\right).$$

Par comparaison, la variance asymptotique de l'estimateur $\hat{\gamma}_n^{\mathrm{P}}(x)$ est, à un facteur d'échelle $\|K\|_2^2/g(x)$ près, identique à celle de l'estimateur classique de Pickands (1975) dans le cas non conditionnel (voir Théorème 1.5.3). Puisque cette variance est assez importante pour des grandes valeurs de l'indice de queue conditionnel (se référer à la Figure 1.5), il nous parait préférable de proposer un estimateur à noyau de variance inférieure.

Définition 5.4.2. *Soit $(\beta_n)_{n\geq 1}$ une suite positive telle que $\beta_n \to 0$ et $nh_n^p\beta_n \to \infty$ quand $n \to \infty$. Un estimateur à noyau de type Hill (1975) est donné pour tout $J > 1$ par*

$$\hat{\gamma}_n^{\mathrm{H}}(x) = \sum_{j=1}^{J}\left[\log\hat{q}_n(\tau_j\beta_n|x) - \log\hat{q}_n(\tau_1\beta_n|x)\right] \Big/ \sum_{j=1}^{J}\log(\tau_1/\tau_j),$$

où $(\tau_j)_{j\geq 1}$ est une suite de poids strictement positive et décroissante.

Corollaire 5.4.2. *Supposons la condition* **(F.2)** *vérifiée. Sous les hypothèses du Théorème 5.3.2, si $\sqrt{nh_n^p\beta_n}\,\varepsilon\left(q(\tau_1\beta_n|x)|x\right) \to 0$ quand $n \to \infty$, alors*

$$\sqrt{nh_n^p\beta_n}\left(\hat{\gamma}_n^{\mathrm{H}}(x) - \gamma(x)\right) \xrightarrow{\mathscr{L}} \mathscr{N}\left(0, \gamma^2(x)\frac{\|K\|_2^2}{g(x)}V_J\right),$$

où $V_J = \left(\sum_{j=1}^{J}\frac{2(J-j)+1}{\tau_j} - J^2\right) \Big/ \left(\sum_{j=1}^{J}\log(\tau_1/\tau_j)\right)^2.$

Le Corollaire 5.4.2 montre aussi que la variance asymptotique de l'estimateur $\hat{\gamma}_n^{\mathrm{H}}(x)$ est, à un facteur d'échelle $V_J\|K\|_2^2/g(x)$ près, identique à celle de l'estimateur classique de Hill (1975) dans le cas non conditionnel (voir Théorème 1.5.2). Ce facteur d'échelle fait intervenir les poids $(\tau_j)_{j\geq 1}$. Ainsi, un mauvais choix de ces poids aura forcément une influence sur la volatilité de notre estimateur à noyau de l'indice de queue conditionnel. Comme exemple de poids nous proposons de manière non exhaustive (voir la Figure 5.2 pour un aperçu graphique) :

- *La suite de poids harmonique* définie pour tout $j = 1,\dots,J$ par $\tau_j^{Ha} = 1/j$. Dans ce cas, $V_J^{Ha} = J(J-1)(2J-1)/(6\log^2(J!))$ est une fonction convexe de J. Son minimum qui est atteint en $J_{opt}^{Ha} = 9$ vaut $V_9^{Ha} \simeq 1.245$.
- *La suite de poids géométrique* consiste à poser pour tout $j = 1,\dots,J$, $\tau_j^G = \left(\frac{1}{j}\right)^{j/J}$. La fonction V_J^G est convexe et minimum pour $J_{opt}^G = 15$ et $V_{15}^G \simeq 1.117$.
- *La suite de la balle rebondissante* qui est une application de la collision physique entre deux corps [4]. Pour tout $j = 1,\dots,J$ elle est définie par $\tau_j^{BR} = b^{2j}$ où $0 < b < 1$. Pour $b = \{8/9, 15/16, 19/20\}$ [5], les minima sont obtenus respectivement pour $J_{opt}^{BR} = \{11, 20, 26\}$ et $V_{J_{opt}}^{BR} \simeq \{1.141, 1.316, 1.135\}$.

Pour $j = 1,\dots,J$, il est possible d'obtenir $V_{J_{opt}} \simeq 1.00$ avec *une suite de poids affine* de la forme $\tau_j^A = \frac{1-1/J}{1-J}j + \left(1 - \frac{1-1/J}{1-J}\right)$. Malheureusement, la fonction V_J^A est strictement décroissante et son minimum est obtenu en $J_{opt}^A = +\infty$. Par exemple, pour $J = \{9, 11, 15, 20, 26, 100, 130\}$ on a respectivement $V_{J_{opt}}^A \simeq \{1.142, 1.115, 1.083, 1.062, 1.047, 1.012, 1.009\}$.

Pour construire un intervalle de confiance du quantile $\hat{q}_n^{\mathrm{W}}(\alpha_n|x)$ ou $\hat{q}_n(\alpha_n|x)$, il suffit de remplacer $\gamma(x)$ et $g(x)$ par leur estimateur respectif.

4. D'après Issac Newton, si l'on lâche un corps verticalement, il va rebondir et l'on peut quantifier les grandeurs physiques intervenant dans les rebonds grâce au coefficient de restitution mis en jeu. La hauteur maximal h_n après n rebonds est donnée par $h_n = C_e^{2n} h_0$ où C_e est le coefficient de restitution de la collision et h_0 la hauteur initiale avant de lâcher le corps. Il s'agit donc d'une suite géométrique.

5. $8/9, 15/16$ et $19/20$ sont respectivement les coefficients de restitutions obtenus par percussion de deux billes d'ivoire, de verre et d'acier.

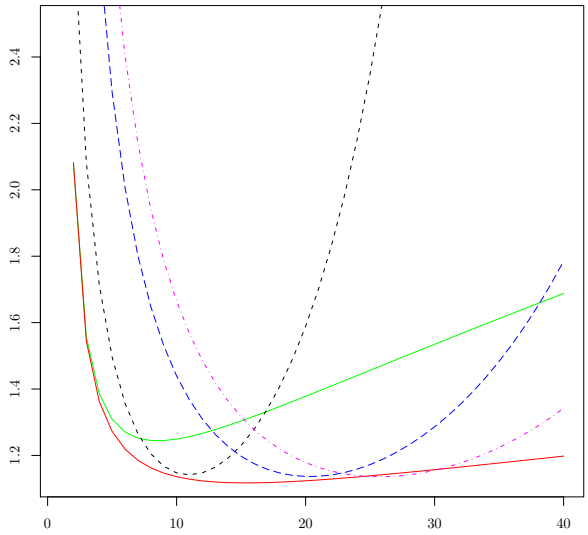


FIGURE 5.2 – V_J^{Ha} (vert), V_J^G (rouge) et V_J^{BR} (avec $b = 8/9$ (noir), $b = 15/16$ (bleu) et $b = 19/20$ (rose)) en trait interrompu en fonction de J. En ordonnée on a la fonction V_J et en abscisse J. On remarquera que V_J^G croit moins vite que V_J^{Ha} et V_J^{BR} pour $J > J_{opt}$.

5.5 Expériences numériques et illustration sur données réelles

5.5.1 Expériences numériques

On considère deux lois usuelles à queue lourde où la fonction indice de queue conditionnel est définie par

$$x \in [0,1] \mapsto \gamma(x) = \frac{1}{2}\left(\frac{1}{10} + \sin(\pi x)\right)\left(\frac{11}{10} - \frac{1}{2}\exp\left(-64\,(x-1/2)^2\right)\right).$$

Pour chacune de ces lois, on génère $N = 100$ réplications d'un échantillon $\{(X_i, Y_i), i = 1, \ldots, n\}$ de taille $n \in \{500, 1000\}$ suivant la loi du couple $(X, Y) \in \mathbb{R} \times \mathbb{R}$ où X est une covariable de loi uniforme standard et dont le quantile conditionnel de Y sachant $X = x$ est donné par la

- loi de Pareto, i.e $q(\alpha_n|x) = \alpha_n^{-\gamma(x)}$,
- loi de Fréchet, i.e $q(\alpha_n|x) = \left(-\log(1-\alpha_n)\right)^{-\gamma(x)}$.

On se focalise d'une part sur l'estimation du quantile extrême conditionnel d'ordre $\alpha_n \in \{5\log(n)/n, 1/2n\}$ et d'autre part sur l'estimation de l'indice de queue conditionnel. La construction des estimateurs de ce chapitre dépend du noyau K et du paramètre de lissage h_n. Pour nos expérimentations, on se contentera d'un noyau bi-quadratique d'expression

$$K(x) = \frac{15}{16}\left(1 - x^2\right)^2 \mathbb{1}_{\{|x| \le 1\}}.$$

En ce qui concerne le choix du noyau, le lecteur curieux pourra se référer à l'article précurseur de Gasser et Müller (1979), aux travaux de Berlinet (1993) qui propose un choix automatique du noyau ou de Vieu (1999) pour des éléments bibliographiques. Le choix automatique du paramètre de lissage a été largement discuté dans la littérature. On pourra se reporter aux travaux de Marron (1988), Jones *et al.* (1996), Vieu (1994) et Herrmann (2000). Notons toutefois que les principales méthodes de choix de la suite h_n sont basées sur des techniques de validation croisé (Yao, 1999), de réechantillonnage (Cao-Abad, 1991) ou de plug-in (Herrmann, 1997). La méthode que nous allons présenter[6] dans ce chapitre s'inspire des idées de validation croisée pour la sélection de modèles. On choisira donc le paramètre de lissage par validation croisée suivant le critère proposé par Yao (1999)

$$\hat{h}_{cv} = \arg\min_{h_n \in \mathscr{H}} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\{ \mathbb{1}_{\{Y_i \geq Y_j\}} - \hat{\bar{F}}_{n,-i}\left(Y_j | X_i\right) \right\}^2, \tag{5.10}$$

avec $\hat{\bar{F}}_{n,-i}$ l'estimateur de $\bar{F}$ calculé à partir de l'échantillon $\{(X_\ell, Y_\ell), \ell = 1, \ldots, n\}$ privé de sa ième observation (X_i, Y_i) et $\mathscr{H} = \{h_1 \leq h_2 \leq \cdots \leq h_M\}$ où $h_1 = 1/(5\log(n))$ et $h_M = 1/4$. La valeur minimal h_1 est choisie de façon à obtenir approximativement $2nh_1\alpha_n = 2$ observations dans la région $[x - h_1, x + h_1] \times [q(\alpha_n|x), \infty[$ quand $\alpha_n = 5\log(n)/n$.

Dans ce paragraphe, nous nous intéressons aux estimateurs $\hat{q}_n(\alpha_n|x)$, $\hat{\gamma}_n^{\mathrm{H}}(x)$ et $\hat{q}_n^{\mathrm{W}}(\alpha_n|x)$. Ici, l'estimateur à noyau de type Weissman sera construit en se servant de l'estimateur à noyau de type Hill. Afin de simplifier les commentaires lors de l'interprétation des graphiques, on note $\hat{\gamma}_{n,1}^{\mathrm{H}}(x)$, $\hat{\gamma}_{n,2}^{\mathrm{H}}(x)$ et $\hat{\gamma}_{n,3}^{\mathrm{H}}(x)$ (resp. $\hat{q}_{n,1}^{\mathrm{W}}(\alpha_n|x)$, $\hat{q}_{n,2}^{\mathrm{W}}(\alpha_n|x)$ et $\hat{q}_{n,3}^{\mathrm{W}}(\alpha_n|x)$) les estimateurs de l'indice de queue conditionnel (resp. du quantile extrême conditionnel) construits en utilisant respectivement la suite de poids harmonique, géométrique et de la balle rebondissante avec $b = 8/9$.

En plus du noyau K et du paramètre h_n, la construction des estimateurs à noyau de Hill et de Weissman dépend du paramètre β_n appelé *seuil*. On se propose de le choisir

- soit en regardant la plage de stabilité de l'estimateur $\hat{\gamma}_n^{\mathrm{H}}(.)$ (resp. $\hat{q}_n^{\mathrm{W}}(\alpha_n|.)$) dans un graphique de Hill (resp. graphique de Weissman),
- soit de minimiser la distance entre deux estimateurs $\hat{\gamma}_n^{\mathrm{H}}(.)$ (resp. $\hat{q}_n^{\mathrm{W}}(\alpha_n|.)$) construits en utilisant deux suites de poids différentes, i.e pour $(i, j) \in \{1, \ldots, 3\}^2$ tels que $i \neq j$

$$\hat{\beta}_{i,j} = \arg\min_{\beta_n \in]0,1]} \mathbb{D}\left(\hat{\gamma}_{n,i}^{\mathrm{H}}(.), \hat{\gamma}_{n,j}^{\mathrm{H}}(.)\right), \tag{5.11}$$

$$\left(\text{resp. } \hat{\beta}_{i,j} = \arg\min_{\beta_n \in]0,1]} \mathbb{D}\left(\hat{q}_{n,i}^{\mathrm{W}}(\alpha_n|.), \hat{q}_{n,j}^{\mathrm{W}}(\alpha_n|.)\right)\right), \tag{5.12}$$

6. Le choix du paramètre de lissage basé sur les techniques de validation croisée est connu pour être populaire tant sur le plan pratique que du point de vue des résultats asymptotiques.

où pour deux fonctions u et v,

$$\mathbb{D}(u,v) = \left\{ \sum_{\ell=1}^{L} (u(t_\ell) - v(t_\ell))^2 \right\}^{1/2}, \tag{5.13}$$

et où $t_1, \ldots, t_L$ sont des points d'une grille régulière sur $[0,1]$.

Nous nous proposons de comparer les estimateurs de la statégie décrite précédemment aux estimateurs de la stratégie de référence appelée *oracle*.

Primo, en ce qui concerne l'estimateur des courbes de niveaux proposé dans la situation **(D.1)**, elle consiste à sélectionner le paramètre h_n en minimisant la distance entre une courbe de niveau extrême estimée et la vraie courbe de niveau extrême, i.e

$$\hat{h}_{oracle} = \arg\min_{h_n \in \mathscr{H}} \mathbb{D}\left(\hat{q}_n(\alpha_n|.), q(\alpha_n|.)\right). \tag{5.14}$$

Secundo, en ce qui concerne l'estimateur des courbes de niveaux extrêmes proposé dans la situation **(D.2)**, elle consiste à sélectionner les paramètres (h_n, β_n) en minimisant la distance entre une courbe de niveau de type Weissman et la vraie courbe de niveau extrême, i.e

$$\left(\hat{h}_{i,oracle}, \hat{\beta}_{i,oracle}\right) = \arg\min_{h_n \in \mathscr{H}, \beta_n \in]0,1]} \mathbb{D}\left(\hat{q}^{\mathrm{W}}_{n,i}(\alpha_n|.), q(\alpha_n|.)\right). \tag{5.15}$$

Notons que le couple $(\hat{h}_{i,oracle}, \hat{\beta}_{i,oracle})$ dépend de la suite de poids i avec $i = 1$ pour la suite de poids harmonique, $i = 2$ pour la suite de poids géométrique et $i = 3$ pour la suite de poids de la balle rebondissante (cf. page précédente).

Tiertio, en ce qui concerne l'estimateur de l'indice de queue, elle consiste à sélectionner les paramètres (h_n, β_n) en minimisant la distance entre un estimateur à noyau de type Hill et la vraie fonction indice de queue, i.e

$$\left(\hat{h}_{i,oracle}, \hat{\beta}_{i,oracle}\right) = \arg\min_{h_n \in \mathscr{H}, \beta_n \in]0,1]} \mathbb{D}\left(\hat{\gamma}^{\mathrm{H}}_{n,i}(.) - \gamma(.)\right). \tag{5.16}$$

Notons qu'en pratique, l'on ne fera jamais mieux que la stratégie oracle. Celle-ci nous sert à apprécier la qualité des critères de sélection introduits dans ce chapitre.

Pour chaque réplication $r \in \{1, \ldots, N\}$, on sélectionne les paramètres $\hat{h}^{(r)}_{cv}$, $\hat{h}^{(r)}_{oracle}$, $\hat{\beta}^{(r)}_{i,j}$ et $\left(\hat{h}^{(r)}_{i,oracle}, \hat{\beta}^{(r)}_{i,oracle}\right)$. Les erreurs de type (5.13) associées au choix de $\hat{h}^{(r)}_{cv}$, $\left(\hat{h}^{(r)}_{cv}, \hat{\beta}^{(r)}_{i,j}\right)$, $\hat{h}^{(r)}_{oracle}$ et $\left(\hat{h}^{(r)}_{i,oracle}, \hat{\beta}^{(r)}_{i,oracle}\right)$ seront respectivement notées $\mathbb{D}^{(r)}_{cv}$, $\mathbb{D}^{(r)}_{i,j}$, $\mathbb{D}^{(r)}_{oracle}$ et $\mathbb{D}^{(r)}_{i,oracle}$.

Une illustration des distributions empiriques de $\mathbb{D}^{(r)}_{cv}$ et $\mathbb{D}^{(r)}_{oracle}$, obtenues pour des échantillons de taille $n \in \{500, 1000\}$ ont été superposées à la Figure 5.11. Il apparait que les moyennes des erreurs sont approximativement égales. On remarque que la loi des erreurs obtenues par validation croisée a une queue droite plus lourde que la loi des

erreurs obtenues par la stratégie oracle. Ceci pourrait s'expliquer par l'influence des points de valeurs plus fortes sur le choix du paramètre de lissage.

Aux Figures 5.3, 5.5, 5.7 et 5.9, on a superposé la vraie courbe de niveau extrême à ses estimateurs obtenus par le critère de validation croisée (5.10) et les stratégies oracle (5.14) et (5.15). On remarque que le meilleur de ces estimateurs est sans aucun doute celui obtenu par le critère (5.15). Il est plus lisse que ces concurrents. C'est certainement le mieux que l'on puisse avoir dans le cas d'espèce. Le premier décile de l'erreur $\mathbb{D}_{cv}^{(r)}$ montre que l'estimateur obtenu par validation croisée fait aussi bien que les estimateurs obtenus par la stratégie oracle. On peut aussi noter que les estimateurs médians obtenus par validation croisée sont relativement stables et d'aussi bonne qualité que ceux obtenus par la même stratégie de référence. Par contre, le neuvième décile de l'erreur $\mathbb{D}_{cv}^{(r)}$ montre que les estimateurs obtenus par validation croisée sont moins précis que les estimateurs oracle. Ainsi que nous l'avons souligné plus haut, cette imprécision est certainement due au fait que le critère de validation croisée est plus sensible aux points extrêmes que la stratégie oracle. Enfin, on peut être satisfait des résultats obtenus par la stratégie oracle.

Aux Figures 5.4, 5.6, 5.8 et 5.10, on a plutôt superposé la vraie courbe de niveau extrême à ses estimateurs obtenus en associant le critère de validation croisée (5.10) au critère (5.12) puis en utilisant la stratégie oracle (5.15). On note des similitudes avec l'expérience précédente. Les estimateurs obtenus avec la loi de Pareto semblent de meilleures qualité que ceux obtenus avec la loi de Fréchet puisque pour une telle loi, la fonction à variations lentes $\ell(.|x)$ est constante (cf. Figure 1.3). L'estimation semble s'améliorer avec la taille de l'échantillon malgré que le fait que $\alpha_{1000} < \alpha_{500}$. Enfin, les simulations tendent à confirmer la difficulté d'estimation des quantiles extrêmes conditionnels pour les grandes valeurs de l'indice de queue conditionnel (cf. Théorème 5.3.3).

L'estimateur à noyau $\hat{q}_n^{\mathrm{W}}(\alpha_n|x)$ étant défini comme une fonction de $\hat{\gamma}_n(x)$, pour des raisons de lisibilité, nous avons choisi de ne pas illustrer le critère de sélection de β_n sur l'estimateur à noyau de type Hill.

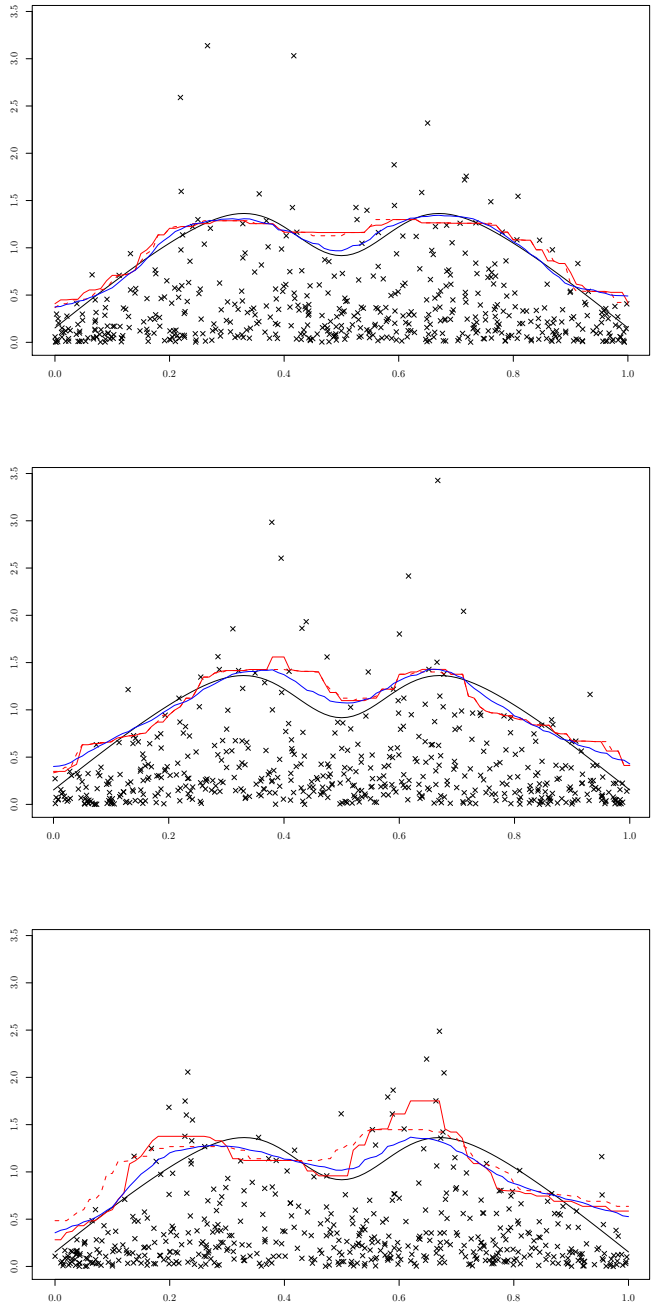


FIGURE 5.3 – Comparaison du vrai quantile $q\big(5\log(n)/n|.\big)$ (noir) avec les quantiles $\hat{q}_n\big(5\log(n)/n|.\big)$ (rouge) obtenus par le critère de Yao (trait continu), la stratégie oracle (5.14) (trait interrompu) et le quantile $\hat{q}^{\mathrm{W}}_{n,2}\big(5\log(n)/n|.\big)$ (bleu) obtenu par la stratégie oracle (5.15). La taille de l'échantillon est $n = 500$, la loi conditionnelle de Y sachant $X = x$ est distribuée selon la **loi Pareto**. L'axe des ordonnées est en échelle logarithmique. En bas : la situation correspondante au neuvième décile de l'erreur $\mathbb{D}_{cv}$, au milieu : la situation correspondante à la médiane de l'erreur $\mathbb{D}_{cv}$, en haut : la situation correspondante au premier décile de l'erreur $\mathbb{D}_{cv}$.

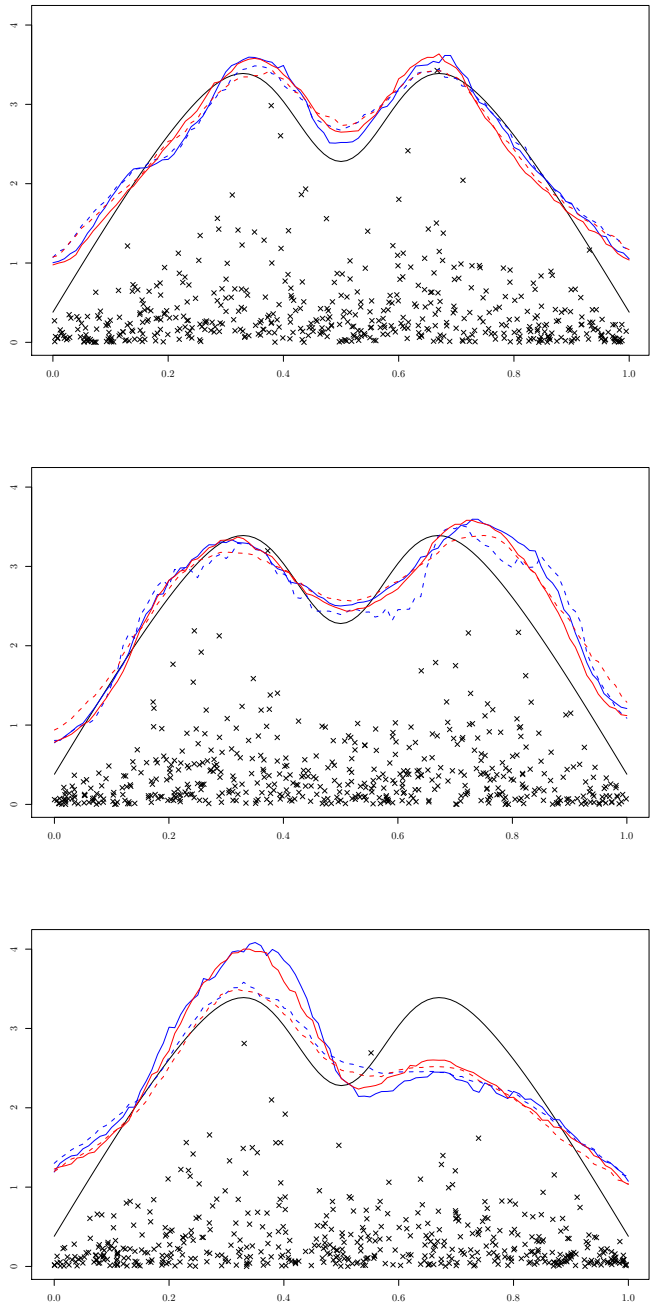


FIGURE 5.4 – Comparaison du vrai quantile $q(1/2n|.)$ (noir) avec les quantiles $\hat{q}^{\mathrm{W}}_{n,2}(1/2n|.)$ (blue) et $\hat{q}^{\mathrm{W}}_{n,3}(1/2n|.)$ (rouge) obtenus en utilisant les critères (5.10) et (5.12) (trait continu) puis la stratégie oracle (5.15) (trait interrompu). La taille de l'échantillon est $n = 500$, la loi conditionnelle de Y sachant $X = x$ est distribuée selon la **loi Pareto**. L'axe des ordonnées est en échelle logarithmique. En bas : la situation correspondante au neuvième décile de l'erreur $\mathbb{D}_{2,3}$, au milieu : la situation correspondante à la médiane de l'erreur $\mathbb{D}_{2,3}$, en haut : la situation correspondante au premier décile de l'erreur $\mathbb{D}_{2,3}$.

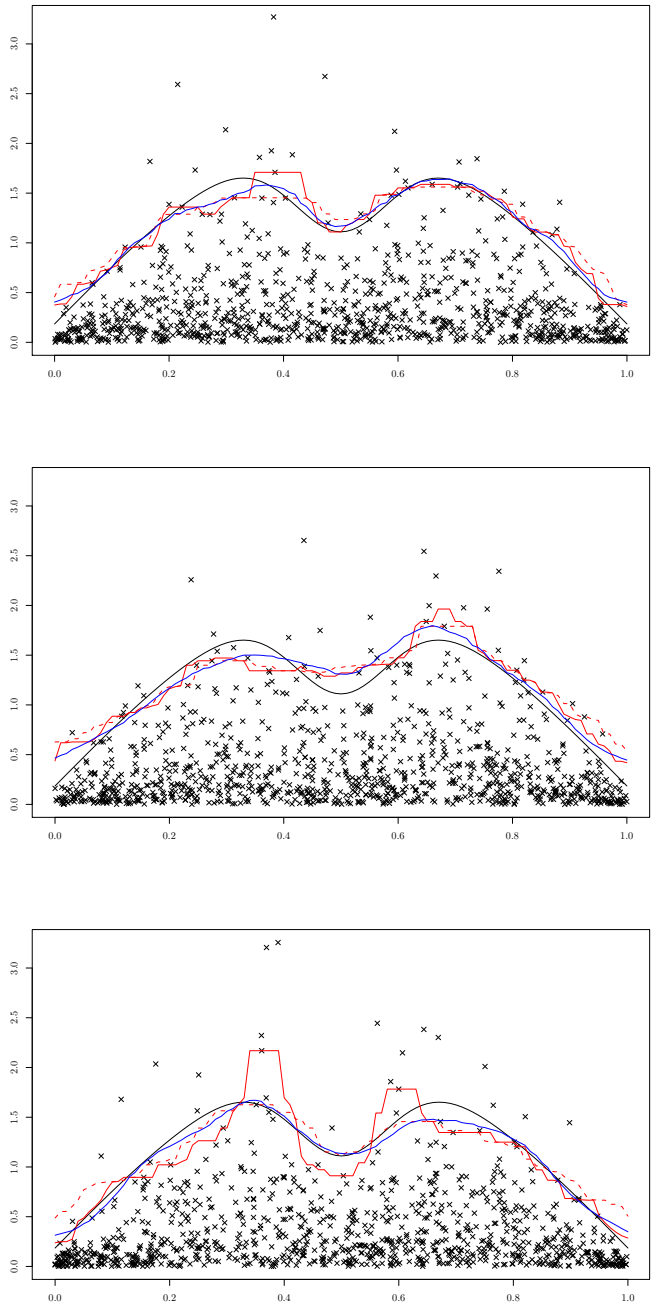


FIGURE 5.5 – Comparaison du vrai quantile $q\big(5\log(n)/n|.\big)$ (noir) avec les quantiles $\hat{q}_n\big(5\log(n)/n|.\big)$ (rouge) obtenus par le critère de Yao (trait continu), la stratégie oracle (5.14) (trait interrompu) et le quantile $\hat{q}^{\mathrm{W}}_{n,2}\big(5\log(n)/n|.\big)$ (bleu) obtenu par la stratégie oracle (5.15). La taille de l'échantillon est $n = 1000$, la loi conditionnelle de Y sachant $X = x$ est distribuée selon la **loi Pareto**. L'axe des ordonnées est en échelle logarithmique. En bas : la situation correspondante au neuvième décile de l'erreur $\mathbb{D}_{cv}$, au milieu : la situation correspondante à la médiane de l'erreur $\mathbb{D}_{cv}$, en haut : la situation correspondante au premier décile de l'erreur $\mathbb{D}_{cv}$.

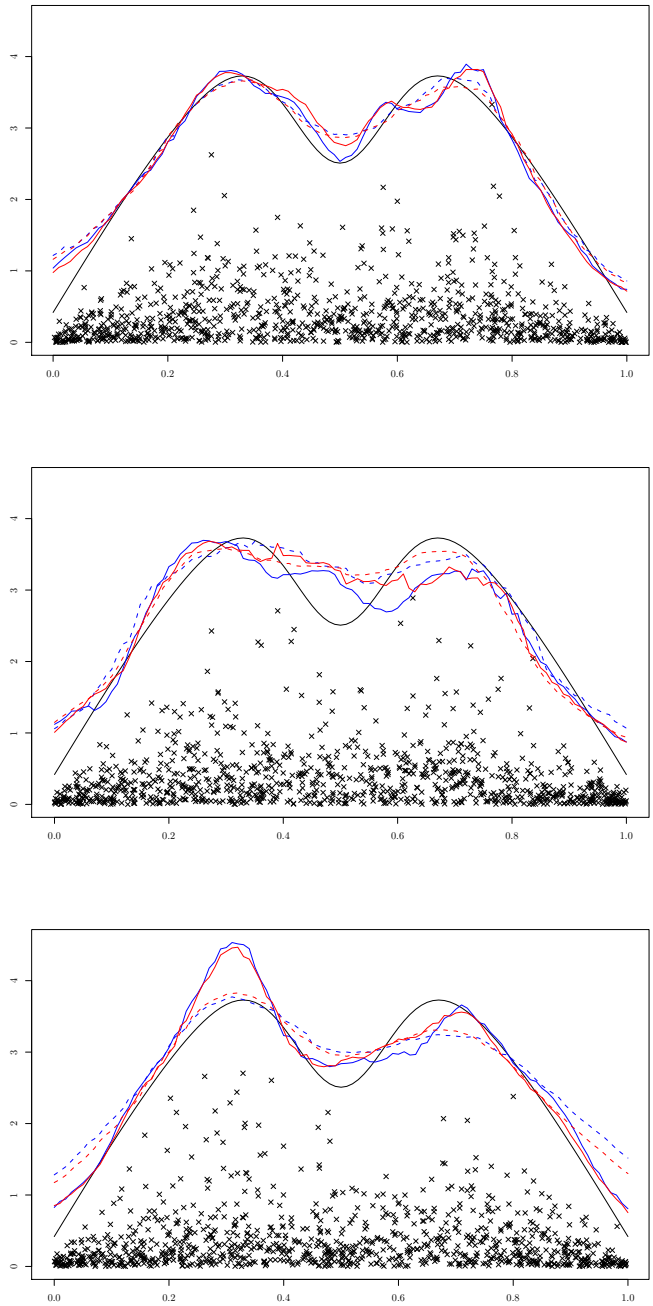


FIGURE 5.6 – Comparaison du vrai quantile $q(1/2n|.)$ (noir) avec les quantiles $\hat{q}^{\mathrm{W}}_{n,2}(1/2n|.)$ (blue) et $\hat{q}^{\mathrm{W}}_{n,3}(1/2n|.)$ (rouge) obtenus en utilisant les critères (5.10) et (5.12) (trait continu) puis la stratégie oracle (5.15) (trait interrompu). La taille de l'échantillon est $n = 1000$, la loi conditionnelle de Y sachant $X = x$ est distribuée selon la **loi Pareto**. L'axe des ordonnées est en échelle logarithmique. En bas : la situation correspondante au neuvième décile de l'erreur $\mathbb{D}_{2,3}$, au milieu : la situation correspondante à la médiane de l'erreur $\mathbb{D}_{2,3}$, en haut : la situation correspondante au premier décile de l'erreur $\mathbb{D}_{2,3}$.

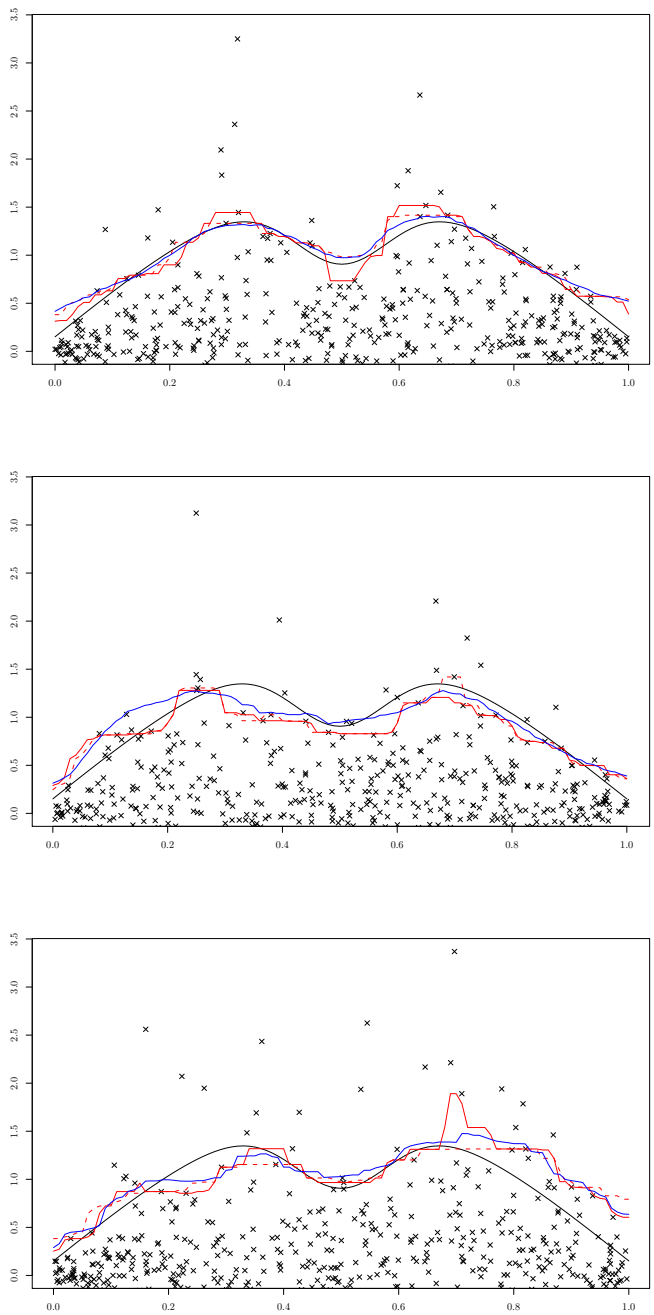


FIGURE 5.7 – Comparaison du vrai quantile $q\big(5\log(n)/n|.\big)$ (noir) avec les quantiles $\hat{q}_n\big(5\log(n)/n|.\big)$ (rouge) obtenus par le critère de Yao (trait continu), la stratégie oracle (5.14) (trait interrompu) et le quantile $\hat{q}^{\mathrm{W}}_{n,1}\big(5\log(n)/n|.\big)$ (bleu) obtenu par la stratégie oracle (5.15). La taille de l'échantillon est $n = 500$, la loi conditionnelle de Y sachant $X = x$ est distribuée selon la **loi Fréchet**. L'axe des ordonnées est en échelle logarithmique. En bas : la situation correspondante au neuvième décile de l'erreur $\mathbb{D}_{cv}$, au milieu : la situation correspondante à la médiane de l'erreur $\mathbb{D}_{cv}$, en haut : la situation correspondante au premier décile de l'erreur $\mathbb{D}_{cv}$.

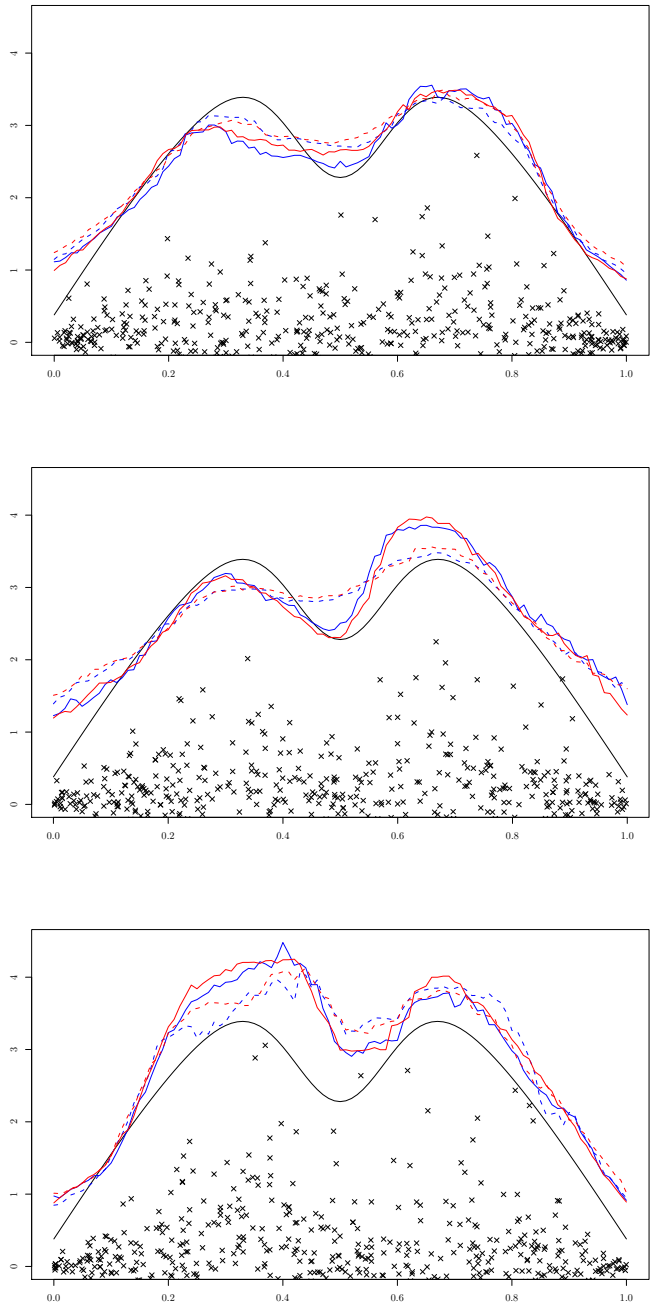


FIGURE 5.8 – Comparaison du vrai quantile $q(1/2n|.)$ (noir) avec les quantiles $\hat{q}^{\mathrm{W}}_{n,1}(1/2n|.)$ (blue) et $\hat{q}^{\mathrm{W}}_{n,3}(1/2n|.)$ (rouge) obtenus en utilisant les critères (5.10) et (5.12) (trait continu) puis la stratégie oracle (5.15) (trait interrompu). La taille de l'échantillon est $n = 500$, la loi conditionnelle de Y sachant $X = x$ est distribuée selon **la loi Fréchet**. L'axe des ordonnées est en échelle logarithmique. En bas : la situation correspondante au neuvième décile de l'erreur $\mathbb{D}_{1,3}$, au milieu : la situation correspondante à la médiane de l'erreur $\mathbb{D}_{1,3}$, en haut : la situation correspondante au premier décile de l'erreur $\mathbb{D}_{1,3}$.

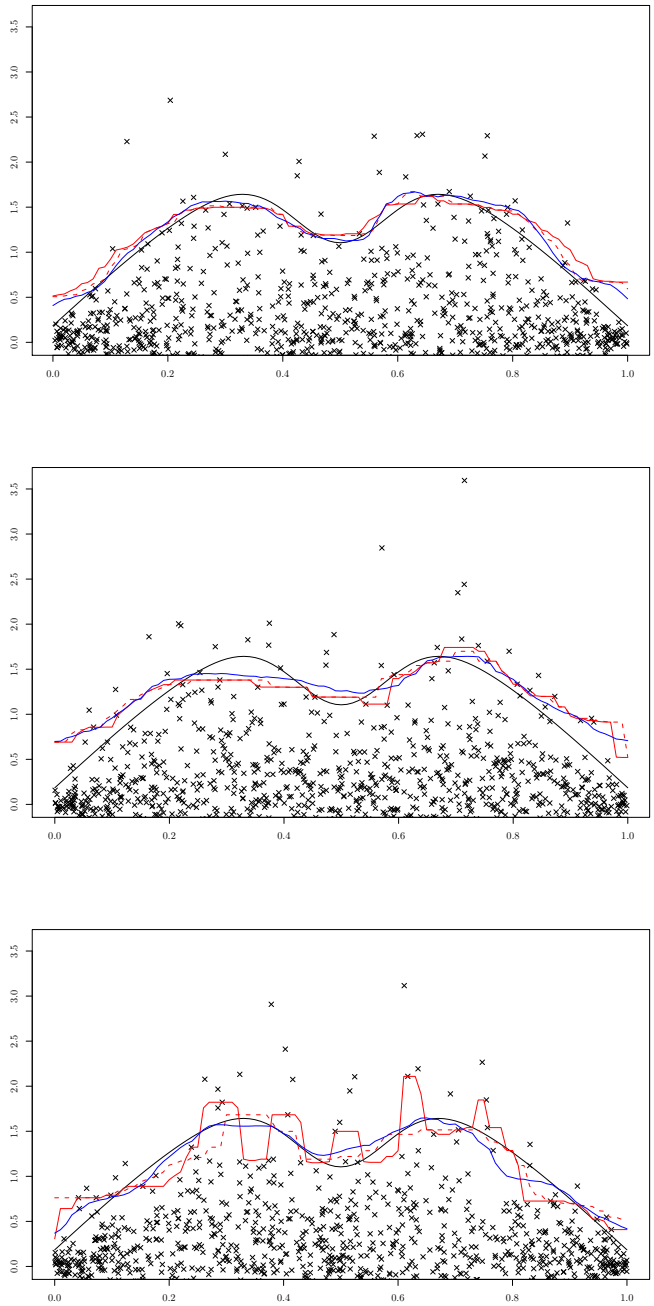


FIGURE 5.9 – Comparaison du vrai quantile $q\big(5\log(n)/n|.\big)$ (noir) avec les quantiles $\hat{q}_n\big(5\log(n)/n|.\big)$ (rouge) obtenus par le critère de Yao (trait continu), la stratégie oracle (5.14) (trait interrompu) et le quantile $\hat{q}^{\mathrm{W}}_{n,1}\big(5\log(n)/n|.\big)$ (bleu) obtenu par la stratégie oracle (5.15). La taille de l'échantillon est $n = 1000$, la loi conditionnelle de Y sachant $X = x$ est distribuée selon la **loi Fréchet**. L'axe des ordonnées est en échelle logarithmique. En bas : la situation correspondante au neuvième décile de l'erreur $\mathbb{D}_{cv}$, au milieu : la situation correspondante à la médiane de l'erreur $\mathbb{D}_{cv}$, en haut : la situation correspondante au premier décile de l'erreur $\mathbb{D}_{cv}$.

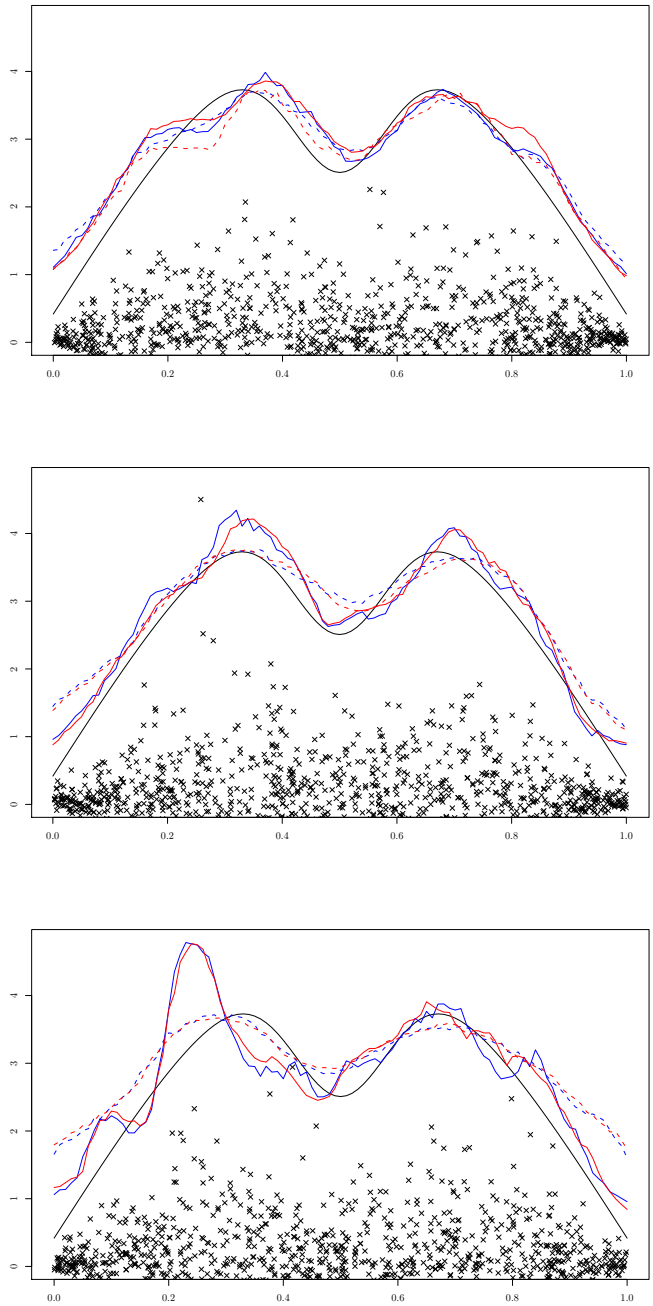
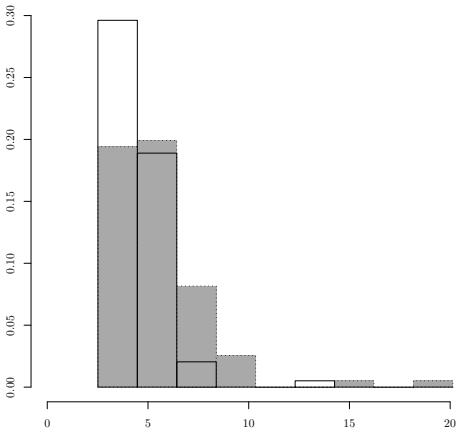
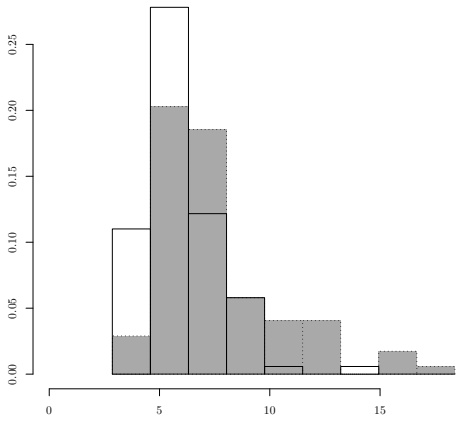


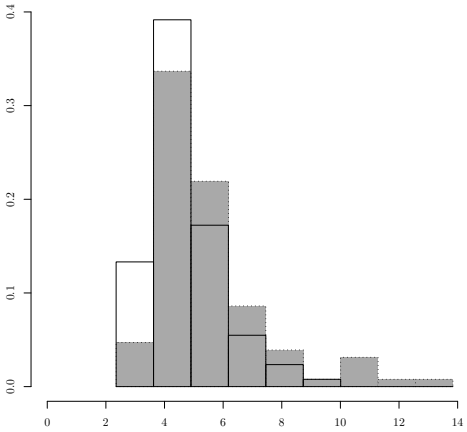
FIGURE 5.10 – Comparaison du vrai quantile $q(1/2n|.)$ (noir) avec les quantiles $\hat{q}^{\mathrm{W}}_{n,1}(1/2n|.)$ (blue) et $\hat{q}^{\mathrm{W}}_{n,3}(1/2n|.)$ (rouge) obtenus en utilisant les critères (5.10) et (5.12) (trait continu) puis la stratégie oracle (5.15) (trait interrompu). La taille de l'échantillon est $n = 1000$, la loi conditionnelle de Y sachant $X = x$ est distribuée selon **la loi Fréchet**. L'axe des ordonnées est en échelle logarithmique. En bas : la situation correspondante au neuvième décile de l'erreur $\mathbb{D}_{1,3}$, au milieu : la situation correspondante à la médiane de l'erreur $\mathbb{D}_{1,3}$, en haut : la situation correspondante au premier décile de l'erreur $\mathbb{D}_{1,3}$.



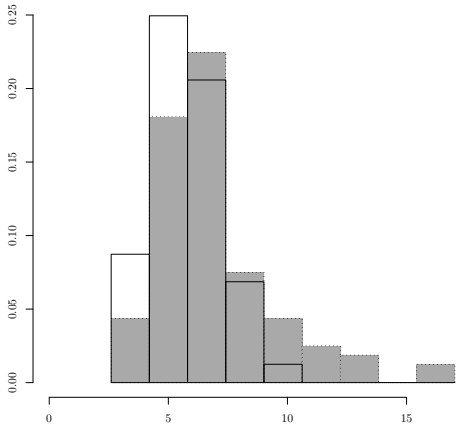
(a) : loi de Pareto avec $n = 500$



(b) : loi de Pareto avec $n = 1000$



(c) : loi de Fréchet avec $n = 500$



(d) : loi de Fréchet avec $n = 1000$

FIGURE 5.11 – Comparaison des histogrammes des erreurs calculées sur $N = 100$ réplications avec la stratégie oracle ($\mathbb{D}^{(r)}_{oracle}$, transparent) et la stratégie de validation croisée ($\mathbb{D}^{(r)}_{cv}$, gris).

5.5.2 Illustration sur données réelles

Comme illustration, on propose une application de notre méthodologie à la télédétection hyperspectrale martienne. L'analyse des signatures spectrales permet l'identification des propriétés physiques, chimiques ou minéralogiques des sols. L'une des questions que se posent les experts est de savoir comment estimer les propriétés physiques associées au sol martien (taille des grains de CO_2, les proportions d'eau, de poussière, de CO_2, etc.) à partir des spectres recueillis par l'instrument OMEGA à bord de la sonde européenne Mars Express. Pour une présentation plus détaillée du contexte physique de l'étude, on pourra se référer à Bernard-Michel *et al.* (2009a). On dispose de $n = 3184$ couples de données notées $\{(S_i, P_i),\ i = 1, \ldots, n\}$ où $S_i \in E$ est un spectre de très grande dimension et $P_i \in [0,1]$ un paramètre physique désignant, dans notre cas, la proportion de CO_2 (gaz carbonique). Dans cette illustration, chaque spectre représente pour chaque longueur d'onde l'intensité de la lumière solaire réfléchie sur un certain nombre de matériaux de la planète Mars.

Compte tenu de l'hypothèse faite sur la dimension de la covariable, nous ne pouvons pas utiliser ces données en l'état. Les techniques de réduction de dimension introduites dans Bernard-Michel *et al.* (2009b) montrent qu'un prédicteur unidimensionnel $X_i =< b, S_i >$ est suffisant pour prédire P_i, avec $b \in E$ et où $< .,. >$ désigne un opérateur de produit scalaire dans E. Les variables d'intérêts seront $Y_i = (1 - P_i)^{-1} - (1 - P_{\min})^{-1}$ où $P_{\min}$ est choisi tel que $Y_i \in [0, \infty[$.

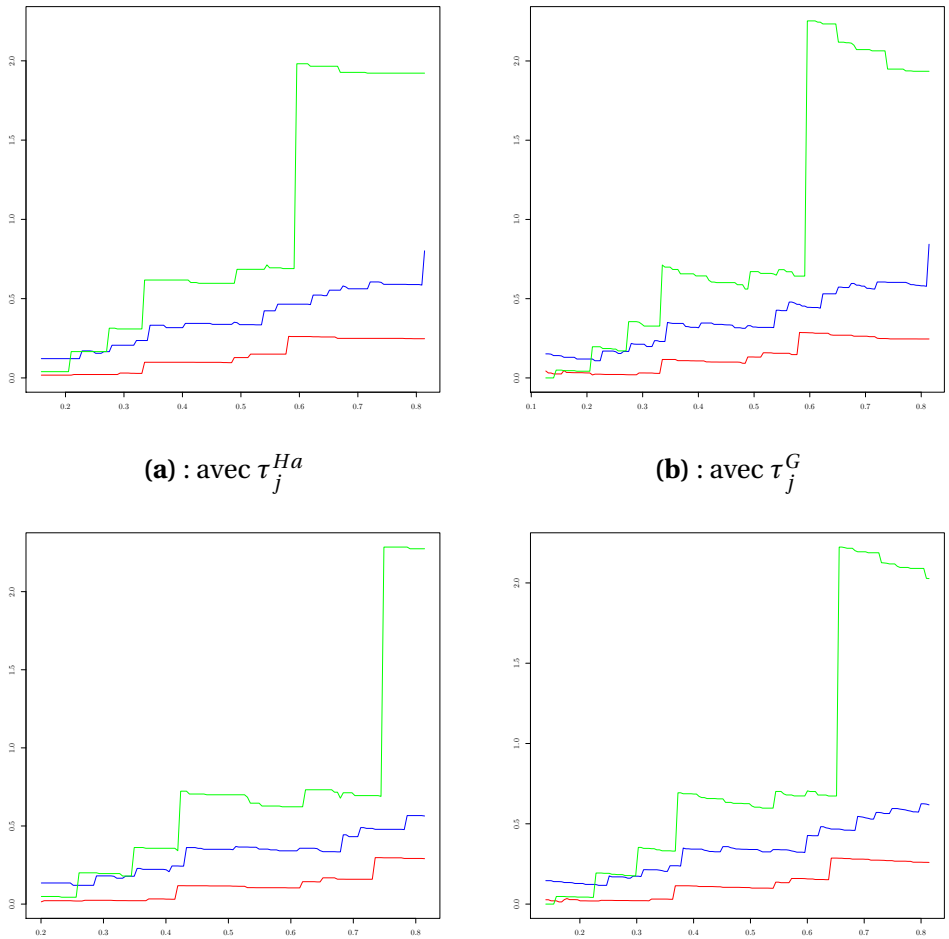
Nous nous sommes intéressés à l'estimation des quantiles extrêmes conditionnels d'ordre $\alpha_n = \zeta \log(n)/n$ avec $\zeta \in \{5, 10, 20\}$. Le paramètre de lissage $h_{cv} \simeq 0.12$ a été choisi suivant le critère (5.10). Notons par $\{Z_j,\ j = 1, \ldots, m(x)\}$ les observations Y_i telles que $X_i \in [x - h_{cv}, x + h_{cv}]$, $i = 1, \ldots, n$. Notre approche repose sur le fait que les $\{Z_j,\ j = 1, \ldots, m(x)\}$ ont une fonction de survie de la forme (5.1). Cette hypothèse peut être vérifiée sur les graphiques QQ-plots obtenus en représentant les droites de coordonnées

$$\left(\log \frac{k}{j}, \log \frac{Z_{m(x)-j+1,m(x)}}{Z_{m(x)-k,m(x)}},\ j = 1, \ldots, k\right).$$

D'après la théorie, l'ajustement de la loi exponentielle, à une facteur d'échelle $\gamma(x)$ près, aux k espacements $\left(\log(Z_{m(x)-j+1,m(x)}) - \log(Z_{m(x)-k+1,m(x)})\right)$ est contrôlé graphiquement en traçant un QQ-plot. Le lecteur curieux pourra consulter (Embrechts *et al.*, 1997, paragraphe 6.2) pour des techniques exploratoires d'analyse de données pour les valeurs extrêmes.

Les QQ-plots obtenus aux points $x = 0.25$, $x = 0.50$ et $x = 0.67$ avec $k = 150$ ont été représentés à la Figure 5.13. L'aspect approximativement linéaire de ces QQ-plots confirme l'adéquation de nos données à l'hypothèse selon laquelle la loi conditionnelle de Y sachant $X = x$ est à queue lourde. Les différentes pentes montrent une forte hétérogénéité de l'échantillon dans le comportement de la queue. La Figure 5.12 montre le comportement des estimateurs à noyau de Hill aux points $x = 0.25$, $x = 0.50$ et $x = 0.67$

avec les suites de poids proposées à la partie 5.4. On peut remarquer que ces estimateurs de l'indice de queue sont sensibles au choix de α_n. Une similitude peut être faite avec le cas non conditionnel (voir chapitres 1 et 2) où se pose le problème du choix de la suite k_n. À la Figure 5.14, nous avons superposé le nuage de points de l'échantillon $\{(X_i, Y_i),\ i = 1, \ldots, n\}$ à nos estimateurs de courbe de niveaux extrêmes.



(a) : avec τ_j^{Ha}

(b) : avec τ_j^{G}

(c) : avec $\tau_j^{BR} = (8/9)^{2j}$

(d) : avec $\tau_j^{BR} = (19/20)^{2j}$

FIGURE 5.12 – Estimateurs à noyau de Hill obtenus en trois points : $x = 0.25$ (rouge) , $x = 0.50$ (bleu) et $x = 0.67$ (vert). En abscisse on a α_n et en ordonnée on a $\hat{\gamma}_n^{\mathrm{H}}(x)$.

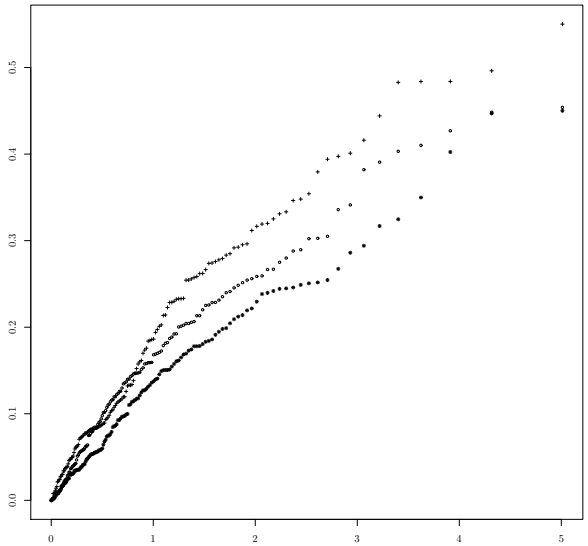


FIGURE 5.13 – QQ-plots obtenues en trois points : $x = 0.25$ ($\star\star\star$), $x = 0.50$ ($\circ\circ\circ$) and $x = 0.67$ ($+++$)

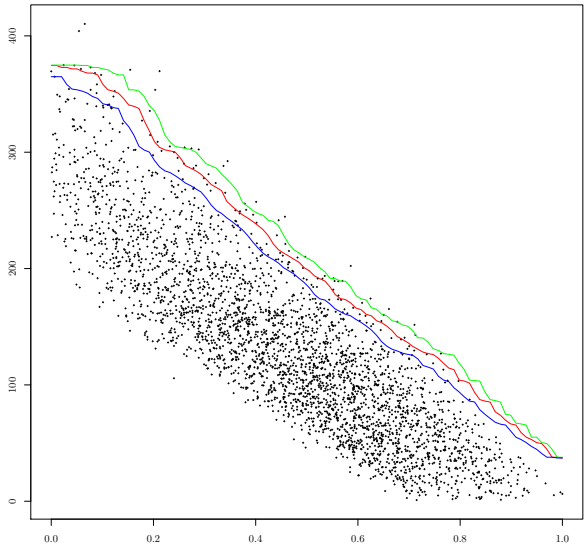


FIGURE 5.14 – Nuage de points $\{(X_i, Y_i),\ i = 1, \ldots, n\}$ et trois estimateurs de courbe de niveaux extrêmes d'ordre $\zeta \log(n)$ ($\zeta = 5$ (vert), $\zeta = 10$ (rouge) et $\zeta = 20$ (bleu))

5.6 Démonstrations

Cette partie se subdivise en deux paragraphes. Le premier est dédié à aux preuves de nos résultats préliminaires et le second à celles des lois limites que nous avons énoncées dans ce chapitre. Avant de nous attaquer aux preuves proprement dites, commençons par donner deux résultats standards sur l'estimation de la densité par la méthode du noyau.

Lemme 5.6.1 (Collomb (1980), **Propositions 2.1 et 2.2**). *Supposons les hypothèses* **(L.3)** *et* **(K)** *satisfaites. Si* $nh_n^p \to \infty$*, alors, pour tout* $x \in \mathbb{R}^p$*,*

(i) $\mathbb{E}(\hat{g}_n(x) - g(x)) = O(h_n)$,

(ii) $var(\hat{g}_n(x)) = \dfrac{g(x)\|K\|_2^2}{nh_n^p}(1+o(1))$.

5.6.1 Preuve des résultats préliminaires

Afin de simplifier les notations, dans tout ce qui suit, on pose $y_{n,j} = a_j y_n$, $\alpha_{n,j} = \tau_j \alpha_n$ et $\beta_{n,j} = \tau_j \beta_n$.

Démonstration du Lemme 5.3.1. Comme $\bar{F}(.|x)$ est une fonction à variations régulières d'indice $-1/\gamma(x)$, la propriété (5.1) et les hypothèses **(L.1)** et **(L.2)** nous permettent d'écrire

$$\begin{aligned}\left|\log\left(\frac{\bar{F}\left(y_n|x\right)}{\bar{F}\left(y_n|x'\right)}\right)\right| &\leq |\log y_n|\left(\left|\frac{1}{\gamma(x)} - \frac{1}{\gamma(x')}\right| + \left|\frac{\log \ell(y_n|x)}{\log y_n} - \frac{\log \ell(y_n|x')}{\log y_n}\right|\right)\\ &\leq (c_\gamma + c_\ell)\log(y_n) d(x,x'),\end{aligned}$$

puisque pour n assez grand, $y_n > 1$. Ainsi,

$$\sup_{d(x,x')\leq h_n}\left|\log\left(\frac{\bar{F}\left(y_n|x\right)}{\bar{F}\left(y_n|x'\right)}\right)\right| = O(h_n \log y_n) \to 0 \text{ quand } n \to \infty.$$

Pour conclure, il suffit de remarquer que $\log(u+1) \sim u$ quand $u \to 0$. □

Démonstration du Lemme 5.3.2. Intéressons nous tout d'abord à la preuve du point (i).
(i) En introduisant la fonction $\varphi(.,x) = \log q(\exp(.)|x)$, nous avons

$$\begin{aligned}\Delta_n &= \log q\left(\beta_n|x\right) - \log q\left(\alpha_n|x\right) + \gamma(x)\log(\beta_n/\alpha_n)\\ &= \varphi(\log\beta_n, x) - \varphi(\log\alpha_n, x) + \gamma(x)\log(\beta_n/\alpha_n).\end{aligned}$$

En vertu de (5.1) et sous l'hypothèse **(F.1)**, la représentation de Karamata (5.5) nous assure que $\varphi(.,x)$ est dérivable. Un développement limité à l'ordre un montre qu'il existe

$\theta_n \in]\beta_n, \alpha_n[$ tel que

$$\begin{aligned}
\Delta_n &= (\gamma(x) + \varphi'(\log\theta_n, x))\log(\beta_n/\alpha_n) \\
&= \left(\gamma(x) + \frac{\bar{F}(q(\theta_n|x)|x)}{\bar{F}'(q(\theta_n|x)|x)q(\theta_n|x)}\right)\log(\beta_n/\alpha_n) \\
&= \left(1 + \frac{1}{\dfrac{\gamma(x)\ell'(q(\theta_n|x)|x)q(\theta_n|x)}{\ell(q(\theta_n|x)|x)} - 1}\right)\gamma(x)\log(\beta_n/\alpha_n) \\
&= \left(1 + \frac{1}{\gamma(x)\varepsilon(q(\theta_n|x)|x) - 1}\right)\gamma(x)\log(\beta_n/\alpha_n) \\
&= -\gamma^2(x)\varepsilon(q(\theta_n|x)|x)\log(\beta_n/\alpha_n)(1+o(1)).
\end{aligned}$$

Puisque $q(.|x)$ and $|\varepsilon(.|x)|$ sont asymptotiquement décroissantes, il s'en suit que $|\varepsilon(q(\theta_n|x)|x)| \leq |\varepsilon(q(\alpha_n|x)|x)|$ et donc que $|\Delta_n| = O(\log(\alpha_n/\beta_n)\varepsilon(q(\alpha_n|x)|x))$. Ce qui conclut le point (i) du Lemme. Intéressons nous maintenant à la preuve du deuxième point.
(ii) Dans la situation considérée, on a

$$0 < \liminf \beta_n/\alpha_n \leq \limsup \beta_n/\alpha_n \leq 1,$$

et (i) entraine que $|\Delta_n| = O(\varepsilon(q(\alpha_n|x)|x)) \to 0$ quand $n \to \infty$. Par conséquent, d'après le développement de Taylor de la fonction exponentielle, on a $\exp(\Delta_n) = 1 + \Delta_n(1+o(1))$. Ce qui achève la preuve du point (ii) et donc du Lemme. □

Démonstration du Lemme 5.3.3. On a

$$\begin{aligned}
\mathbb{P}(\exists i \in \{1,\dots,n\}, (X_i, Y_i) \in R_n(x)) &= 1 - \mathbb{P}(\forall i \in \{1,\dots,n\}, (X_i, Y_i) \notin R_n(x)) \\
&= 1 - (1 - \mathbb{P}((X_1, Y_1) \in R_n(x))^n, \qquad (5.17)
\end{aligned}$$

et compte tenu de l'hypothèse **(L.3)** et du Lemme 5.3.1,

$$\begin{aligned}
\mathbb{P}((X_1, Y_1) \in R_n(x))) &= \int_{B(x,h_n)} \bar{F}\left(q(\alpha_n|u)|u\right) g(u)du \\
&= \bar{F}\left(q(\alpha_n|x)|x\right) g(x)(1 + O(h_n \log y_n)) \int_{B(x,h_n)} du \\
&= v_p h_n^p \bar{F}\left(q(\alpha_n|x)|x\right) g(x)(1 + O(h_n \log y_n)),
\end{aligned}$$

où v_p est le volume de la boule centrée en $x \in \mathbb{R}^p$ et de rayon h_n. Puisque, $\mathbb{P}((X_1, Y_1) \in R_n(x))) \to 0$ quand $n \to \infty$ alors, on peut réécrire (5.17) comme

$$\begin{aligned}
\mathbb{P}(\exists i \in \{1,\dots,n\}, (X_i, Y_i) \in R_n(x)) &= 1 - \exp\left(-v_p g(x) n h_n^p \bar{F}\left(q(\alpha_n|x)|x\right)(1+o(1))\right) \\
&= 1 - \exp\left(-v_p g(x) n h_n^p \alpha_n (1+o(1))\right)
\end{aligned}$$

qui converge vers 1 si et seulement si $nh_n^p\alpha_n \to \infty$. □

5.6.2 Preuve des lois asymptotiques des estimateurs

Démonstration du Lemme 5.3.4. Cette preuve se subdivise en deux points.
(i) Puisque l'échantillon $\{(X_i, Y_i),\ i = 1,\dots,n\}$ est identiquement distribué alors, sous l'hypothèse **(K)** nous avons

$$\begin{aligned}\mathbb{E}(\hat{\psi}_n(y_{n,j},x)) &= \int_{\mathbb{R}^p} K_h(x-t)\bar{F}(y_{n,j}|t)g(t)dt \\ &= \int_S K(u)\bar{F}(y_{n,j}|x-h_nu)g(x-h_nu)du.\end{aligned}$$

Considérons maintenant

$$\begin{aligned}|\mathbb{E}(\hat{\psi}_n(y_{n,j},x)) - \psi(y_{n,j},x)| &\leq \bar{F}(y_{n,j}|x)\int_S K(u)\left|\frac{\bar{F}(y_{n,j}|x-h_nu)}{\bar{F}(y_{n,j}|x)}g(x-h_nu)-g(x)\right|du \\ &\leq \bar{F}(y_{n,j}|x)\int_S K(u)|g(x-h_nu)-g(x)|du \qquad (5.18)\\ &+ \bar{F}(y_{n,j}|x)\int_S K(u)\left|\frac{\bar{F}(y_{n,j}|x-h_nu)}{\bar{F}(y_{n,j}|x)}-1\right|g(x-h_nu)du. \qquad (5.19)\end{aligned}$$

Sous l'hypothèse **(L.3)**, puisque $g(x) > 0$, nous avons

$$(5.18) \leq \bar{F}(y_{n,j}|x)c_g h_n \int_S d(u,0)K(u)du = \psi(y_{n,j},x)O(h_n). \qquad (5.20)$$

En outre, le Lemme 5.3.1 implique que

$$\sup_{u\in S}\left|\frac{\bar{F}(y_{n,j}|x-h_nu)}{\bar{F}(y_{n,j}|x)}-1\right| = O(h_n\log y_{n,j}) = O(h_n\log y_n)$$

et par conséquent, compte tenu de (5.20), nous avons

$$\begin{aligned}(5.19) &= \bar{F}(y_{n,j}|x)O(h_n\log y_n)\int_S K(u)g(x-h_nu)du \\ &= \bar{F}(y_{n,j}|x)g(x)O(h_n\log y_n)(1+o(1)) \\ &= \psi(y_{n,j},x)O(h_n\log y_n). \qquad (5.21)\end{aligned}$$

En combinant (5.20) et (5.21), on conclut la première partie de la preuve.
(ii) Soient $\beta \neq 0$ un vecteur de $\mathbb{R}^J$ et $\Lambda_n(x) = (nh_n^p\psi(y_n,x))^{-1/2}$, considérons la variable aléatoire

$$\begin{aligned}\Psi_n &= \sum_{j=1}^J \beta_j\left(\frac{\hat{\psi}_n(y_{n,j},x)-\mathbb{E}(\hat{\psi}_n(y_{n,j},x))}{\Lambda_n(x)\psi(y_{n,j},x)}\right) \\ &= \sum_{i=1}^n \frac{1}{n\Lambda_n(x)}\left\{\sum_{j=1}^J \frac{\beta_j K_h(x-X_i)\mathbb{1}_{\{Y_i\geq y_{n,j}\}}}{\psi(y_{n,j},x)} - \mathbb{E}\left(\sum_{j=1}^J \frac{\beta_j K_h(x-X_i)\mathbb{1}_{\{Y_i\geq y_{n,j}\}}}{\psi(y_{n,j},x)}\right)\right\} \\ &\overset{def}{=} \sum_{i=1}^n Z_{i,n}.\end{aligned}$$

Il apparait clairement que $\{Z_{i,n},\ i=1,\dots,n\}$ est un ensemble de variables aléatoires centrées, indépendantes et identiquement distribuées de variance

$$\mathrm{var}(Z_{i,n}) = \frac{1}{n^2 h_n^{2p}\Lambda_n^2(x)}\mathrm{var}\left(\sum_{j=1}^{J}\beta_j K\left(\frac{x-X_i}{h_n}\right)\frac{\mathbb{1}_{\{Y_i\geq y_{n,j}\}}}{\psi(y_{n,j},x)}\right) = \frac{1}{n^2 h_n^p \Lambda_n^2(x)}\beta^t B\beta,$$

où B est la matrice $J\times J$ de variance-covariance dont les coefficients sont définis pour $(j,j')\in\{1,\dots,J\}^2$ par

$$\begin{aligned}
B_{j,j'} &= \frac{A_{j,j'}}{\psi(y_{n,j},x)\psi(y_{n,j'},x)},\\
A_{j,j'} &= \frac{1}{h_n^p}\mathrm{cov}\left(K\left(\frac{x-X}{h_n}\right)\mathbb{1}_{\{Y\geq y_{n,j}\}},\ K\left(\frac{x-X}{h_n}\right)\mathbb{1}_{\{Y\geq y_{n,j'}\}}\right)\\
&= \|K\|_2^2\mathbb{E}\left(\frac{1}{h_n^p}Q\left(\frac{x-X}{h_n}\right)\mathbb{1}_{\{Y\geq y_{n,j}\vee y_{n,j'}\}}\right)\\
&- h_n^p\mathbb{E}(K_h(x-X)\mathbb{1}_{\{Y\geq y_{n,j}\}})\mathbb{E}(K_h(x-X)\mathbb{1}_{\{Y\geq y_{n,j'}\}}),
\end{aligned}$$

avec $Q(.)\overset{def}{=}K^2(.)/\|K\|_2^2$ satisfaisant à l'hypothèse **(K)**. Par conséquent, les trois espérances mathématiques ci-dessus sont de même nature. Ainsi, en remarquant que, pour n assez grand $y_{n,j}\vee y_{n,j'} = y_{n,j\vee j'}$, le point **(i)** de la preuve implique que

$$A_{j,j'} = \|K\|_2^2\psi(y_{n,j\vee j'},x)(1+O(h_n\log y_n)) - h_n^p\psi(y_{n,j},x)\psi(y_{n,j'},x)(1+O(h_n\log y_n))$$

qui entraine que

$$B_{j,j'} = \frac{\|K\|_2^2}{\psi(y_{n,j\wedge j'},x)}(1+O(h_n\log y_n)) - h_n^p(1+O(h_n\log y_n)) = \frac{\|K\|_2^2}{\psi(y_{n,j\wedge j'},x)}(1+o(1)),$$

puisque $\psi(y_{n,j\wedge j'},x)\to 0$ quand $n\to\infty$. Grâce à la propriéte (5.1) sur les fonctions à variations régulières, il est facile de voir que $\psi(y_{n,j\wedge j'},x) = a_{j\wedge j'}^{-1/\gamma(x)}\psi(y_n,x)(1+o(1))$. Il en découle que

$$B_{j,j'} = \frac{\|K\|_2^2 C_{j,j'}(x)}{\psi(y_n,x)}(1+o(1))$$

et donc que $\mathrm{var}(Z_{i,n})\sim\|K\|_2^2\beta^t C(x)\beta/n$, pour tout $i=1,\dots,n$. À cette étape de la preuve, on vient de montrer que la variance de Ψ_n converge vers $\|K\|_2^2\beta^t C(x)\beta$. En conséquence, la condition de Lyapounov pour le Théorème central limite appliquée au tableau triangulaire des sommes partielles de $Z_{i,n}$ se résume à

$$\sum_{i=1}^{n}\mathbb{E}\left|Z_{i,n}\right|^3 = n\mathbb{E}\left|Z_{1,n}\right|^3\to 0. \tag{5.22}$$

Remarquons que la variable aléatoire $Z_{1,n}$ est bornée :

$$|Z_{1,n}|\leq\frac{2\|K\|_\infty\sum_{j=1}^{J}|\beta_j|}{n\Lambda_n(x)h_n^p\psi(y_{n,J},x)} = 2\|K\|_\infty a_J^{1/\gamma(x)}\sum_{j=1}^{J}|\beta_j|\Lambda_n(x)(1+o(1))$$

et ainsi,

$$\begin{aligned} n\mathbb{E}\left|Z_{1,n}\right|^3 &\leq 2\|K\|_\infty a_J^{1/\gamma(x)} \sum_{j=1}^J |\beta_j|\Lambda_n(x) n\mathrm{var}(Z_{1,n})(1+o(1)) \\ &= 2\|K\|_\infty \|K\|_2^2 a_J^{1/\gamma(x)} \sum_{j=1}^J |\beta_j|\beta^t C(x)\beta\Lambda_n(x)(1+o(1)) \to 0 \end{aligned}$$

quand $n \to \infty$. Ce qui prouve que Ψ_n converge en loi vers une variable aléatoire Gaussienne centrée de variance $\|K\|_2^2\beta^t C(x)\beta$ pour tout $\beta \neq 0$ in $\mathbb{R}^p$. Ceci achève la démonstration du Lemme. □

Démonstration du Lemme 5.3.1. Gardant à l'esprit les notations du Lemme 5.3.4, tout calcul fait, on a le développement suivant

$$\Lambda_n^{-1}(x)\sum_{j=1}^J \beta_j\left(\frac{\hat{\bar{F}}_n\left(y_{n,j}|x\right)}{\bar{F}(y_{n,j}|x)}-1\right) = \frac{\Delta_{1,n}+\Delta_{2,n}-\Delta_{3,n}}{\hat{g}_n(x)}, \tag{5.23}$$

où

$$\begin{aligned} \Delta_{1,n} &= g(x)\Lambda_n^{-1}(x)\sum_{j=1}^J \beta_j\left(\frac{\hat{\psi}_n(y_{n,j},x)-\mathbb{E}(\hat{\psi}_n(y_{n,j},x))}{\psi(y_{n,j},x)}\right) \\ \Delta_{2,n} &= g(x)\Lambda_n^{-1}(x)\sum_{j=1}^J \beta_j\left(\frac{\hat{\mathbb{E}}(\psi_n(y_{n,j},x))-\psi(y_{n,j},x)}{\psi(y_{n,j},x)}\right) \\ \Delta_{3,n} &= \left(\sum_{j=1}^J \beta_j\right)\Lambda_n^{-1}(x)\left(\hat{g}_n(x)-g(x)\right). \end{aligned}$$

Commençons par remarquer que les hypothèses $nh_n^{d+2}\log^2(y_n)\bar{F}(y_n|x) \to 0$ et $nh_n^p\bar{F}(y_n|x) \to \infty$ impliquent que $h_n \log y_n \to 0$ quand $n \to \infty$. Ainsi, d'après le point (ii) du Lemme 5.3.4, le terme aléatoire $\Delta_{1,n}$ peut se réécrire comme

$$\Delta_{1,n} = g(x)\|K\|_2\sqrt{\beta^t C(x)\beta}\,\xi_n, \tag{5.24}$$

où ξ_n converge vers une variable aléatoire Gaussienne standard. Le terme non aléatoire $\Delta_{2,n}$ est contrôlé par le point (i) du Lemme 5.3.4 :

$$\Delta_{2,n} = O(\Lambda_n^{-1}(x)h_n\log y_n) = O\left(nh_n^{p+2}\bar{F}(y_n|x)\log^2(y_n)\right)^{1/2} = o(1). \tag{5.25}$$

Enfin, $\Delta_{3,n}$ est un terme classique en estimation de densité par la méthode du noyau. Il peut être borné par le Lemme 5.6.1 :

$$\Delta_{3,n} = O(h_n\Lambda_n^{-1}(x)) + O_{\mathrm{P}}\left(\Lambda_n^{-1}(x)(nh_n^p)^{-1/2}\right) = O\left(nh_n^{p+2}\bar{F}(y_n|x)\right)^{1/2} + O_{\mathrm{P}}\left(\bar{F}(y_n|x)\right)^{1/2} = o_{\mathrm{P}}(1). \tag{5.26}$$

En combinant (5.23)–(5.26), il s'en suit que

$$\hat{g}_n(x)\Lambda_n^{-1}(x)\sum_{j=1}^J \beta_j\left(\frac{\hat{\bar{F}}_n\left(y_{n,j}|x\right)}{\bar{F}\left(y_{n,j}|x\right)}-1\right) = g(x)\|K\|_2\sqrt{\beta^t C(x)\beta}\,\xi_n + o_{\mathrm{P}}(1).$$

Finalement, puisque d'après le Lemme 5.6.1 on a $\hat{g}_n(x) \xrightarrow{\mathbb{P}} g(x)$, il en découle que

$$\sqrt{nh_n^p \bar{F}(y_n|x)} \sum_{j=1}^{J} \beta_j \left(\frac{\hat{\bar{F}}_n\left(y_{n,j}|x\right)}{\bar{F}\left(y_{n,j}|x\right)} - 1 \right) = \|K\|_2 \sqrt{\frac{\beta^t C(x)\beta}{g(x)}} \xi_n + o_{\mathrm{P}}(1)$$

et le Théorème est prouvé. □

Démonstration du Théorème 5.3.2. Pour $j = 1, \ldots, J$, on introduit

$$\begin{aligned}
\sigma_{n,j}(x) &= q(\alpha_{n,j}|x)(nh_n^p \alpha_n)^{-1/2} \\
\upsilon_{n,j}(x) &= \alpha_{n,j}^{-1} \gamma(x)(nh_n^p \alpha_n)^{1/2} \\
W_{n,j}(x) &= \upsilon_{n,j}(x) \left(\hat{\bar{F}}_n\left(q\left(\alpha_{n,j}|x\right) + \sigma_{n,j}(x) z_j|x\right) - \bar{F}\left(q\left(\alpha_{n,j}|x\right) + \sigma_{n,j}(x) z_j|x\right) \right) \\
a_{n,j}(x) &= \upsilon_{n,j}(x) \left(\alpha_{n,j} - \bar{F}\left(q\left(\alpha_{n,j}|x\right) + \sigma_{n,j}(x) z_j|x\right) \right)
\end{aligned}$$

et $z_j \in \mathbb{R}$. On veut établir la loi asymptotique de la fonction J-variée définie par

$$\begin{aligned}
\Phi_n(z_1, \ldots, z_J) &= \mathbb{P}\left(\bigcap_{j=1}^{J} \left\{ \sigma_{n,j}^{-1}(x) \left(\hat{q}_n\left(\alpha_{n,j}|x\right) - q\left(\alpha_{n,j}|x\right) \right) \le z_j \right\} \right) \\
&= \mathbb{P}\left(\bigcap_{j=1}^{J} \left\{ \hat{\bar{F}}_n\left(q\left(\alpha_{n,j}|x\right) + \sigma_{n,j}(x) z_j|x\right) \le \alpha_{n,j} \right\} \right) \\
&= \mathbb{P}\left(\bigcap_{j=1}^{J} \left\{ W_{n,j}(x) \le a_{n,j}(x) \right\} \right).
\end{aligned}$$

Premièrement, focalisons nous sur le terme non aléatoire $a_{n,j}(x)$. D'après la propriété (5.1) et l'hypothèse **(F.2)**, la représentation de Karamata (5.5) montre que $\bar{F}(.|x)$ est dérivable. Ainsi, pour chaque $j \in \{1, \ldots, J\}$ il existe $\theta_{n,j} \in]0,1[$ tel que

$$\bar{F}\left(q\left(\alpha_{n,j}|x\right)|x\right) - \bar{F}\left(q\left(\alpha_{n,j}|x\right) + \sigma_{n,j}(x) z_j|x\right) = -\sigma_{n,j}(x) z_j \bar{F}'\left(q_{n,j}|x\right), \tag{5.27}$$

où $q_{n,j} = q\left(\alpha_{n,j}|x\right) + \theta_{n,j}\sigma_{n,j}(x) z_j$. Il est clair que $q\left(\alpha_{n,j}|x\right) \to \infty$ et $\sigma_{n,j}(x)/q\left(\alpha_{n,j}|x\right) \to 0$ quand $n \to \infty$. Par conséquent, $q_{n,j} \to \infty$ et la représentation de Karamata (5.5) entrainent que

$$\lim_{n\to\infty} \frac{q_{n,j} \bar{F}'\left(q_{n,j}|x\right)}{\bar{F}\left(q_{n,j}|x\right)} = -1/\gamma(x). \tag{5.28}$$

De plus, puisque $q_{n,j} \sim q\left(\alpha_{n,j}|x\right)$ as $n \to \infty$ et $\bar{F}(.|x)$ est à variations régulières il s'en suit que $\bar{F}\left(q_{n,j}|x\right) \sim \bar{F}\left(q\left(\alpha_{n,j}|x\right)|x\right) = \alpha_{n,j}$. Compte tenu des résultats trouvés en (5.27) et (5.28), il en découle que

$$a_{n,j}(x) = \frac{\upsilon_{n,j}(x)\sigma_{n,j}(x)\alpha_{n,j} z_j}{\gamma(x) q\left(\alpha_{n,j}|x\right)}(1 + o(1)) = z_j(1 + o(1)). \tag{5.29}$$

Intéressons nous maintenant à la variable aléatoire $W_{n,j}(x)$. En définissant $a_j = \tau_j^{-\gamma(x)}$, $y_{n,j} = q\left(\alpha_{n,j}|x\right) + \sigma_{n,j}(x) z_j$ pour $j = 1, \ldots, J$ et $y_n = q(\alpha_n|x)$, nous avons $y_{n,j} \sim q\left(\alpha_{n,j}|x\right) \sim a_j y_n$ puisque $q(.|x)$ est une fonction à variations régulière d'indice $-\gamma(x)$. En utilisant le même raisonnement, il est facile de montrer que

$\log y_n \sim -\gamma(x)\log\alpha_n$. Par conséquent, le Théorème 5.3.1 s'applique et le vecteur aléatoire

$$\left\{\frac{\sqrt{nh_n^p\bar{F}(y_n|x)}}{v_{n,j}(x)\bar{F}(y_{n,j}|x)}W_{n,j}\right\}_{\{j=1,\dots,J\}} = (1+o(1))\left\{\frac{W_{n,j}}{\gamma(x)}\right\}_{\{j=1,\dots,J\}}$$

converge vers une variable aléatoire Gaussienne centrée de matrice de variance-covariance $\frac{\|K\|_2^2}{g(x)}C(x)$. En se référant au résultat (5.29), nous obtenons que $\Phi_n(z_1,\dots,z_J)$ converge vers la fonction de répartition d'une Gaussienne centrée de matrice de variance-covariance $\frac{\|K\|_2^2\gamma^2(x)}{g(x)}C(x)$ évaluée en $(z_1,\dots,z_J)$, qui est le résultat désiré. □

Démonstration du Théorème 5.3.3. La preuve de ce résultat repose sur la décomposition suivante

$$\frac{\sqrt{nh_n^p\beta_n}}{\log(\beta_n/\alpha_n)}\left(\log(\hat{q}_n^{\mathrm{W}}(\alpha_n|x))-\log(q(\alpha_n|x))\right) = \frac{\sqrt{nh_n^p\beta_n}}{\log(\beta_n/\alpha_n)}\left(Q_{n,1}+Q_{n,2}+Q_{n,3}\right)$$

avec

$$\begin{aligned}
Q_{n,1} &= \sqrt{nh_n^p\beta_n}\left(\hat{\gamma}_n-\gamma(x)\right),\\
Q_{n,2} &= \frac{\sqrt{nh_n^p\beta_n}}{\log(\beta_n/\alpha_n)}\log\left(\hat{q}_n\left(\beta_n|x\right)/q\left(\beta_n|x\right)\right),\\
Q_{n,3} &= \frac{\sqrt{nh_n^p\beta_n}}{\log(\beta_n/\alpha_n)}\left(\log q\left(\beta_n|x\right)-\log q\left(\alpha_n|x\right)+\gamma(x)\log(\beta_n/\alpha_n)\right).
\end{aligned}$$

Traitons séparément les trois termes. *Primo*, d'après l'hypothèse du Théorème, $Q_{n,1}\xrightarrow{\mathscr{L}}\mathscr{N}(0,v^2(x))$. *Secundo*, le Théorème 5.3.2 implique que $\hat{q}_n\left(\beta_n|x\right)/q\left(\beta_n|x\right)\xrightarrow{\mathbb{P}}1$ quand $n\to\infty$ et on a

$$Q_{n,2} = \frac{\sqrt{nh_n^p\beta_n}}{\log(\beta_n/\alpha_n)}\left(\frac{\hat{q}_n\left(\beta_n|x\right)}{q\left(\beta_n|x\right)}-1\right)(1+o(1)) = \frac{O_{\mathrm{P}}(1)}{\log(\beta_n/\alpha_n)}.$$

Conséquemment, $Q_{n,2}\xrightarrow{\mathbb{P}}0$ quand $n\to\infty$. *Tertio*, d'après le point (i) du Lemme 5.3.2, on a $Q_{n,3}=O\left(\sqrt{nh_n^p\beta_n}\varepsilon(q\left(\beta_n|x\right)|x)\right)$ qui converge vers zéro puisque par hypothèse $\varepsilon(q\left(\beta_n|x\right)|x)/\sigma_n \overset{def}{=} \sqrt{nh_n^p\beta_n}\varepsilon(q\left(\beta_n|x\right)|x)\to 0$. Ce qui achève la preuve. □

Démonstration du Corollaire 5.4.1. En posant $\tau_1=4$, $\tau_2=2$, et $\tau_3=1$, le Théorème 5.3.2 montre que, pour $j\in\{1,2,3\}$,

$$\frac{\hat{q}_n(\beta_{n,j}|x)}{q(\beta_{n,j}|x)} = 1+\sigma_n\xi_{n,j} \tag{5.30}$$

où $(\xi_{n,1},\xi_{n,2},\xi_{n,3})^t$ converge vers un vecteur aléatoire Gaussien centré de matrice de variance-covariance

$$|K\|_2^2\gamma^2(x)/g(x)\begin{bmatrix}1/4 & 1/4 & 1/4\\ 1/4 & 1/2 & 1/2\\ 1/4 & 1/2 & 1\end{bmatrix}.$$

D'après la Définition 5.4.1, on a

$$(\log 2)\hat{\gamma}_n^{\mathrm{P}} = \log\left(\hat{q}_n\left(\beta_{n,3}|x\right) - \hat{q}_n\left(\beta_{n,2}|x\right)\right) - \log\left(\hat{q}_n\left(\beta_{n,2}|x\right) - \hat{q}_n\left(\beta_{n,1}|x\right)\right).$$

En remplaçant (5.30) dans $\hat{\gamma}_n^{\mathrm{P}}$, on obtient

$$\begin{aligned}
(\log 2)\hat{\gamma}_n^{\mathrm{P}} &= \log\left((1+\sigma_n\xi_{n,3})q\left(\beta_{n,3}|x\right) - (1+\sigma_n\xi_{n,2})q\left(\beta_{n,2}|x\right)\right) \\
&- \log\left((1+\sigma_n\xi_{n,2})q\left(\beta_{n,2}|x\right) - (1+\sigma_n\xi_{n,1})q\left(\beta_{n,1}|x\right)\right) \\
&= \log\left(\frac{q(\beta_{n,3}|x)}{q(\beta_{n,2}|x)}(1+\sigma_n\xi_{n,3}) - 1 - \sigma_n\xi_{n,2}\right) \\
&- \log\left(1+\sigma_n\xi_{n,2} - \frac{q(\beta_{n,1}|x)}{q(\beta_{n,2}|x)}(1+\sigma_n\xi_{n,1})\right),
\end{aligned}$$

qui se réécrit, compte tenu du Lemme 5.3.2, comme

$$\begin{aligned}
(\log 2)\hat{\gamma}_n^{\mathrm{P}} &= \log\left(2^{\gamma(x)}(1+O(\varepsilon(q(\beta_{n,2}|x)|x)))(1+\sigma_n\xi_{n,3}) - 1 - \sigma_n\xi_{n,2}\right) \\
&- \log\left(1+\sigma_n\xi_{n,2} - 2^{-\gamma(x)}(1+O(\varepsilon(q(\beta_{n,2}|x)|x)))(1+\sigma_n\xi_{n,1})\right).
\end{aligned}$$

Puisque par hypothèse $\varepsilon(q(\beta_{n,2}|x)|x))/\sigma_n \to 0$ alors,

$$\begin{aligned}
(\log 2)\hat{\gamma}_n^{\mathrm{P}} &= \log\left(2^{\gamma(x)} - 1 + \sigma_n(2^{\gamma(x)}\xi_{n,3} - \xi_{n,2} + o_{\mathrm{P}}(1))\right) \\
&- \log\left(1 - 2^{-\gamma(x)} + \sigma_n(\xi_{n,2} - 2^{-\gamma(x)}\xi_{n,1} + o_{\mathrm{P}}(1))\right).
\end{aligned}$$

Tout calcul fait, il s'en suit alors que

$$\begin{aligned}
\sigma_n^{-1}(\log 2)(\hat{\gamma}_n^{\mathrm{P}} - \gamma(x)) &= \sigma_n^{-1}\log\left(1 + \frac{\sigma_n}{2^{\gamma(x)}-1}(2^{\gamma(x)}\xi_{n,3} - \xi_{n,2} + o_{\mathrm{P}}(1))\right) \\
&- \sigma_n^{-1}\log\left(1 + \frac{\sigma_n}{1-2^{-\gamma(x)}}(\xi_{n,2} - 2^{-\gamma(x)}\xi_{n,1} + o_{\mathrm{P}}(1))\right) \\
&= \frac{\xi_{n,1} - (1+2^{\gamma(x)})\xi_{n,2} + 2^{\gamma(x)}\xi_{n,3}}{2^{\gamma(x)}-1} + o_{\mathrm{P}}(1),
\end{aligned}$$

converge vers une variable aléatoire Gaussienne centrée de variance

$$\frac{\|K\|_2^2\gamma^2(x)(2^{2\gamma(x)+1}+1)}{4(2^{\gamma(x)}-1)^2 g(x)}.$$

Ce qui conclut la preuve. □

Démonstration du Corollaire 5.4.2. Soit $c_J = \sum_{j=1}^{J}\log(\tau_1/\tau_j)$, la décomposition suivante

$$\begin{aligned}
\sigma_n^{-1}(\hat{\gamma}_n^{\mathrm{H}}(x) - \gamma(x)) &= c_J^{-1}\sum_{j=1}^{J}\sigma_n^{-1}\left[\log\left(\frac{\hat{q}_n\left(\beta_{n,j}|x\right)}{q\left(\beta_{n,j}|x\right)}\right) - \log\left(\frac{\hat{q}_n\left(\beta_{n,1}|x\right)}{q\left(\beta_{n,1}|x\right)}\right)\right] \\
&+ c_J^{-1}\sum_{j=1}^{J}\sigma_n^{-1}\log\left(\tau_j^{\gamma(x)}\frac{q\left(\beta_{n,j}|x\right)}{q\left(\beta_{n,1}|x\right)}\right) \\
&\overset{def}{=} T_{n,1} + T_{n,2},
\end{aligned}$$

montre que la loi de l'estimateur $\hat{\gamma}_n^{\mathrm{H}}(x)$ dépend du comportement du premier terme. D'après le Théorème 5.3.2, $\hat{q}_n\left(\beta_{n,j}|x\right)/q\left(\beta_{n,j}|x\right) \xrightarrow{\mathbb{P}} 1$ quand $n \to \infty$ uniformément en $j = 1, \ldots, J$. Ainsi,

$$\left\{\sigma_n^{-1}\log\left(\frac{\hat{q}_n\left(\beta_{n,j}|x\right)}{q\left(\beta_{n,j}|x\right)}\right)\right\}_{\{j=1,\ldots,J\}} = (1+o_{\mathrm{P}}(1))\left\{\sigma_n^{-1}\left(\frac{\hat{q}_n\left(\beta_{n,j}|x\right)}{q\left(\beta_{n,j}|x\right)}\right)\right\}_{\{j=1,\ldots,J\}}$$

converge en loi vers un vecteur aléatoire Gaussien centrée de matrice de variance-covariance $\|K\|_2^2\gamma^2(x)\Sigma/g(x)$. En conséquence, $T_{n,1}$ converge vers une variable aléatoire Gaussienne centrée de variance $\beta^t\Sigma\beta\|K\|_2^2\gamma^2(x)/(g(x)c_J^2)$ avec $\beta = (1-J, 1, \ldots, 1)^t \in \mathbb{R}^J$. Tout calcul fait, on trouve que $\beta^t\Sigma\beta = V_J c_J^2$. Le point (ii) du Lemme 5.3.2 montre que $T_{n,2} = \sigma_n^{-1}O(\varepsilon(q(\beta_{n,1}|x)|x))$. Pour conclure, il suffit de remarquer que d'après l'hypothèse du Corollaire on a $\varepsilon(q(\beta_{n,1}|x)|x)/\sigma_n \to 0$ quand $n \to \infty$. □

Conclusions et perspectives

L'objectif principal de ce mémoire était de proposer de nouveaux estimateurs de quantiles extrêmes dans le cadre conditionnel c'est-à-dire dans la situation où la variable d'intérêt Y est mesurée simultanément avec une covariable X. Néanmois, le cas sans covariable a été considéré dans la première partie de la thèse.

Dans le cadre non conditionnel, nous nous sommes interessés à l'étude des valeurs extrêmes d'un échantillon de variables aléatoires réelles indépendantes et identiquement distribuées de fonction de répartition $F \in \mathscr{D}(Fréchet)$. Nous avons proposé et étudié un estimateur des quantiles extrêmes ainsi que de l'indice de queue. Les estimateurs ainsi proposés ont été adaptés au cadre conditionnel.

Par analogie, dans le cadre conditionnel, nous nous sommes focalisés sur l'étude des valeurs extrêmes d'un échantillon d'observations indépendantes dont la loi conditionnelle, de Y en un point x de la covariable X, appartient au $\mathscr{D}(Fréchet)$. La variable d'intérêt Y a toujours été supposée aléatoire et réelle. Concernant la nature des covariables, nous avons considéré deux situations : aléatoire et détermniste.

Lorsque la covariable est déterministe, nous avons proposé trois estimateurs des quantiles extrêmes conditionnels et lorsqu'elle est aléatoire, nous avons introduit un estimateur de petites probabilités conditionnelles, deux estimateurs des quantiles extrêmes conditionnels et deux estimateurs de l'indice de queue conditionnel. Nous avons établi la convergence asymptotique de tous ces estimateurs.

Les estimateurs proposés dans le cadre conditionnel dépendent de deux paramètres : la proportion $\beta_n = k_n/n$ des plus grandes observations et le nombre h_n correspondant au paramètre de lissage. Quand X est déterministe (resp. aléatoire), h_n désigne la taille de la fenêtre mobile (resp. le paramètre de lissage associé à la fonction noyau). On a proposé des méthodes automatiques de sélection desdits paramètres.

Les simulations numériques que nous avons effectuées se sont avérées très encourageantes puisque les résultats obtenus avec les méthodes de sélection proposées se comportent parfois aussi bien que la méthode de référence appelée oracle qui minimise l'erreur avec le vrai quantile. La méthode oracle tend à confirmer que nos méthodes d'estimation sont satisfaisantes.

À court terme, compte tenu de ces satisfactions, on pourrait commencer par établir la normalité asymptotique de nos estimateurs dans le cas des données α-mélangeantes. Il serait intéressant d'étendre l'étude asymptotique de l'estimateur à noyau de la fonction de survie conditionnelle dans un contexte plus général, i.e au $\mathscr{D}(Weibull)$ et/ou $\mathscr{D}(Gumbel)$.

Il serait aussi intéressant de proposer une version lisse (à noyau) de nos estimateurs pour le design fixe. Par ailleurs, puisque les estimateurs à noyau du chapitre 5 sont discontinus par rapport à l'ordre du quantile, on pourrait envisager d'effectuer un lissage par rapport à la variable d'intérêt afin d'obtenir des estimateurs de quantiles extrêmes conditionnels réguliers par rapport à α_n. Autrement dit, il serait bien de montrer qu'il est possible d'estimer des courbes de niveaux extrêmes au moyen d'un estimateur à double noyau de la fonction de survie conditionnelle.

Les estimateurs à noyau de l'indice de queue conditionnel que nous avons proposés sont basés sur une conséquence du résultat de normalité asymptotique de l'estimateur du quantile de régression lorsque l'ordre du quantile ne converge pas trop vite vers un. Pour débuter, on se propose de vérifier si l'on peut appliquer ce résultat de normalité asymptotique pour adapter d'autres estimateurs existants de l'indice de queue au cadre conditionnel. De plus, on envisage d'étudier de nouveaux estimateurs à noyau de l'indice de queue conditionnel.

Contrairement aux estimateurs du chapitre 4, ceux du chapitre 5 ne s'adressent pas aux covariables de dimension infinie. Lorsque la covariable est aléatoire et fonctionnelle, Ferraty *et al.* (2005) ont proposé un estimateur de quantile conditionnel (continu par rapport à $\alpha_n = \alpha \in]0,1[$ fixé) pour des données non nécessairement indépendantes. Ainsi, on pourrait envisager d'étendre leur approche d'estimation à l'étude des quantiles extrêmes, i.e $\alpha_n \to 0$ et proposer de nouveaux estimateurs à noyau de l'indice de queue et des quantiles extrêmes conditionnels dans le cadre fonctionnel.

Au cours de cette thèse, nous avons uniquement établi la convergence ponctuelle de nos estimateurs. **À moyen terme**, il nous parait important d'envisager d'étendre nos résutats de normalité asymptotique à des résultats de convergence de processus indexés par la covariable ou l'ordre des quantiles.

Afin d'améliorer la fiabilité des estimateurs, on pourrait proposer des nouveaux modèles pour pallier les problèmes de corrélation spatio-temporelle assez récurrente

en hydrologie.

Enfin, on envisage d'affiner nos critères de sélection en adaptant l'heuristique de la pente ou les techniques utilisées en sélection de modèles. Ainsi, **à long terme**, on compte proposer un meilleur algorithme de sélection automatique du paramètre β_n (ou k_n dans le cas non conditionnel).

A

Loi limite d'une combinaison linéaire d'espacements entre les logarithmes des plus grandes statistiques d'ordre.

L'étude théorique des estimateurs du Théorème 2 repose en partie sur la représentation exponentielle des espacements entre les logarithmes des plus grandes statistiques d'ordre des k_n plus grandes observations d'un échantillon $\{X_i,\ i = 1,\dots,n\}$ de variables aléatoires réelles indépendantes et identiquement distribuées de fonction de répartition F. Lorsque $F \in \mathscr{D}(Fréchet)$, Feuerverger et Hall (1999) ont proposé d'approcher ces espacements pour $1 \le j \le k_n \le n-1$ par :

$$\Delta_j = j\left(\log X_{n-j+1,n} - \log X_{n-j}\right) \sim \left(\gamma + \varepsilon(n/k_n)\left(\frac{j}{k_n+1}\right)\right) E_j, \tag{A.1}$$

où $(E_1,\dots,E_{k_n})$ est un vecteur de variables aléatoires indépendantes et identiquement distribuées de loi exponentielle standard. De cette représentation, on peut déduire l'estimateur de Hill (1975). Partant de (A.1), Beirlant *et al.* (2002) proposent d'estimer l'indice de queue par

$$\hat{\gamma}_{k_n} = \frac{1}{k_n}\sum_{j=1}^{k_n} K\left(\frac{j}{k_n+1}\right)\Delta_j,$$

où K est une fonction de poids pouvant se réécrire comme $K(t) = \frac{1}{t}\int_0^t u(v)\,dv$, avec $0 < t < 1$ et u une fonction définie sur $]0,1[$. Les auteurs établissent le résultat de normalité asymptotique suivant :

Théorème A.0.1 (Beirlant *et al.* (2002), **Théorème 3.1**). *Supposons l'hypothèse* **(C.1)** *vérifiée (cf. paragraphe 1.5.1). Soit u une fonction définie sur $]0,1[$ et satisfaisant*

$$\left|k_n \int_{(j-1)/k_n}^{j/k_n} u(t)\,dt\right| \le f\left(\frac{j}{k_n+1}\right),$$

où f est une fonction positive définie sur $]0,1[$ telle que

$$\int_0^1 \left(\log(1/u) \vee 1\right) f(u)\,du < \infty,$$

et s'il existe $\delta > 0$ *tel que*

$$\int_0^1 |K|^{2+\delta}(u)\,du < \infty.$$

Si $k_n \to \infty$, $n/k_n \to 0$ *et* $k_n\varepsilon(n/k_n) \to \lambda \in \mathbb{R}$ *quand* $n \to \infty$, *alors*

$$k_n^{1/2}\left[\frac{1}{k_n}\sum_{j=1}^{k_n} K\left(\frac{j}{k_n+1}\right)\Delta_j - \frac{1}{k_n}\sum_{j=1}^{k_n} K\left(\frac{j}{k_n+1}\right)\left(\gamma + \varepsilon(n/k_n)\left(\frac{j}{k_n+1}\right)^{-\rho}\right)\right]$$

converge vers une loi $\mathcal{N}\left(0, \gamma^2 \int_0^1 K^2(u)\,du\right)$.

Bibliographie

ALVES, M., de HAAN, L. et LIN, T. (2003a). Estimation of the parameter controlling the speed of convergence in extreme value theory. *Mathematical Methods of Statistics*, 12:155–176. 92

ALVES, M., GOMES, M. et de HAAN, L. (2003b). A new class of semi-parametric estimators of the second order parameter. *Portugaliae Mathematica*, 60:193–214. 92

ANTOCH, J. et JANSSEN, P. (1989). Nonparametric regression m-quantiles. *Statistics and Probability Letters*, 8(4):355–362. 48

BACRO, J. et BRITO, M. (1994). Weak limiting behaviour of a simple tail Pareto-index estimator. *Journal of Statistical Planning and inference*, 45(1-2):7–19. 24

BALKEMA, A. et de HAAN, L. (1974). Residual life time at a great age. *Annals of Probability*, 2(5):792–804. 17, 21

BASSETT, J. G. et KOENKER, R. (1982). An empirical quantile function for linear models with iid errors. *Journal of the American Statistical Association*, 77(378):407–415. 51

BEIRLANT, J., de WET, T. et GOEGEBEUR, Y. (2004a). Nonparametric estimation of extreme conditional quantiles. *Journal of Statistical Computation and Simulation*, 74(8):567–580. 54

BEIRLANT, J., DIERCKX, G., GUILLLOU, A. et STĂRICĂ, C. (2002). On exponential representations of log-spacings of extreme order statistics. *Extremes*, 5(2):157–180. 24, 32, 33, 42, 63, 129

BEIRLANT, J., DIERCKX, G. et GUILLOU, A. (2005). Estimation of the extreme value index and regression on generalized quantile plots. *Annals of Statistics*, 11(6):949–970. 24

BEIRLANT, J., DIERCKX, G., Y., G. et MATTHYS, G. (1999). Tail index estimation and an exponential regression model. *Extremes*, 2(2):177–200. 24

BEIRLANT, J. et GOEGEBEUR, Y. (2003). Regression with response distributions of Pareto-type. *Computational Statistics and Data Analysis*, 42(4):595–619. 54, 88

BEIRLANT, J. et GOEGEBEUR, Y. (2004). Local polynomial maximum likelihood estimation for Pareto-type distribution. *Journal of Multivariate Analysis*, 89:97–118. 54, 56, 63

BEIRLANT, J., GOEGEBEUR, Y., TEUGELS, J. et SEGERS, J. (2004b). *Statistics of extremes : theory and applications*. Wiley Series in Probability and Statistics. John Wiley and Sons Ltd. 16, 24

BEIRLANT, J. et MATTHYS, G. (2001). Extreme quantile estimation for heavy-tailed distributions. Rapport technique, Department of Mathematics, K. U. Leuven. 54

BEIRLANT, J. et MATTHYS, G. (2003). Estimating the extreme value index and high quantiles with exponential regression models. *Statistica Sinica*, 13(3):853–880. 54

BEIRLANT, J., VYNCKIER, P. et TEUGELS, J. (1996). Excess functions and estimation of the extreme value index. *Bernoulli*, 2(4):293–318. 24, 26

BERLINET, A. (1993). Hierarchies of higher order kernels. *Probability Theory and Related Fields*, 94(4):489–504. 101

BERLINET, A., GANNOUN, A. et MATZNER-LOBER, E. (1998). Proprétés asymptotiques des estimateurs des quantiles. *Comptes Rendus de l'Académie des Sciences*, 326(5):611–614. 46, 49

BERLINET, A., GANNOUN, A. et MATZNER-LOBER, E. (2001). Asymptotic normality of convergent estimates of conditional quantiles. *Statistics*, 18:1400–1415. 50, 52, 95, 97

BERNARD-MICHEL, C., DOUTÉ, S., FAUVEL, M., GARDES, L. et GIRARD, S. (2009a). Retrieval of mars surface physical properties from OMEGA hyperspectral images using regularized sliced inverse regression. *Journal of Geophysical Research - Planets*, 114: E06005. 68, 113

BERNARD-MICHEL, C., GARDES, L. et GIRARD, S. (2009b). Gaussian regularized sliced inverse regression. *Statistics and Computing*, 19(1):85–98. 113

BHATTACHARYYA, P. K. et GANGOPADHYAY, A. K. (1990). Kernel and nearest-neighbor estimation of a conditional quantile. *Annals of Statistics*, 8(3):1400–1415. 51

BINGHAM, N. H., GOLDIE, C. M. et TEUGELS, J. L. (1987). *Regular Variation*, volume 27. Encyclopedia of Mathematics and its Applications, Cambridge University Press. 15, 55, 59, 80, 82

BOIS, P., OBLED, C., de SAINTIGNON, M. et MAILLOUX, H. (1997). *Atlas expérimental des risques de pluies intenses Cévennes - Vivarais*. Pôle grenoblois d'études et de recherche pour la prévention des risques naturels, Grenoble, 2ème édition. 74

BOSQ, D. (2000). *Linear processes in function spaces : theory and applications*, volume 149. Lecture Notes in Statistics, Springer Verlag. 54

BREIMAN, L., STONE, C. J. et KOOPERBERG, C. (1990). Robust confidence bounds for extreme upper quantiles. *Journal of Statistical Computation and Simulation*, 37(3–4):127–149. 22

BUISHAND, T. A., de HAAN, L. et ZHOU, C. (2008). On spatial extremes : With application to a rainfall problem. *Annals of Applied statistics*, 2(2):624–642. 74

CAO-ABAD, R. (1991). Rate of convergence for the wild bootstrap in nonparametric regression. *Annals of Statistics*, 19(4):2226–2231. 101

CARDOT, H., CRAMBES, C. et SARDA, P. (2004). Estimation spline de quantiles conditionnels pour variables explicatives fonctionnelles. *Comptes Rendus de l'Académie des Sciences*, 339(2):141–144. 51

CARDOT, H., CRAMBES, C. et SARDA, P. (2005). Quantile regression when the covariates are functions. *Journal of Nonparametric Statistics*, 17(7):841–856. 51

CHAUDURI, P. (1991). Nonparametric estimates of regression quantiles and their local Bahadur representation. *Annals of Statistics*, 19(2):760–777. 52

CHAVEZ-DEMOULIN, V. et DAVISON, A. C. (2005). Generalized additive modelling of sample extremes. *Journal of the Royal Statistical Society, C*, 54(1):207–222. 54, 63, 88

CHERNOZHUKOV, V. (1998). Nonparametric extreme regression quantiles. Rapport technique, Department of Economics, Stanford University. 54, 88

CHERNOZHUKOV, V. (2001). *Conditional extremes and near-extremes : Estimation, inference, and economic applications*. Thèse de doctorat, Department of Economics, Stanford University. 54, 88

COLE, T. J. (1988). Fitting smoothed centile curves to reference data. *Journal of Royal Statistical Society, A*, 151(3):385–418. 46

COLES, S. G. et TAWN, J. A. (1996). A bayesian analysis of extreme rainfall data. *Applied statistics*, 45(4):463–478. 74

COLLOMB, G. (1980). Estimation non paramétrique de probabilités conditionnelles. *Comptes Rendus de l'Académie des Sciences*, 291:427–430. 47, 48, 90, 94, 116

CONT, R. (2009). La statistique face aux événements rares = statistics dealing with rare events. *Pour la science*, 385:116–123. 1

CSÖRGÖ, S., DEHEUVELS, P. et MASON, D. (1985). Kernel estimates of the tail index of a distribution. *Annals of Statistics*, 13(3):1050–1077. 24

CSÖRGÖ, S. et MASON, D. (1985). Central limit theorems for sums of extreme values. *Mathematical Proceedings of the Cambridge Philosophical Society*, 98(3):547–558. 25

DANIELSSON, J., DENNIS, W. J. et de VRIES, C. G. (1996). The method of moments ratio estimator for the tail shape parameter. *Communication in Statistics, Theory and Methods*, 4(25):711–720. 24

DAOUIA, A., GARDES, L., GIRARD, S. et LEKINA, A. (2010). Kernel estimators of extreme level curves. *Test*. À paraître. 7

DAVIS, R. et RESNICK, S. (1984). Tail estimates motivated by extreme value theory. *Annals of Statistics*, 12(4):1467–1487. 25

DAVISON, A. et RAMESH, N. I. (2000). Local likelihood smoothing of sample extremes. *Journal of the Royal Statistical Society, B*, 62(1):191–208. 54, 63, 88

DAVISON, A. C. et SMITH, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society, B*, 52(3):393–442. 22, 54, 63, 88

de HAAN, L. (1984). Slow variation and characterization of domains of attraction. *In* J. Tiago de OLIVEIRA, éditeur : *Statistical Extremes and Applications*, pages 31–48. Reidel, Dorchrecht. 58

de HAAN, L. (1990). Fighting the Arch-enemy with mathematics. *Statistica Neerlandica*, 44(2):45–68. 2

de HAAN, L. et FERREIRA, A. (2006). *Extreme Value Theory : An Introduction*. Springer Series in Operations Research and Financial Engineering, New York Inc. 16, 19, 24

de HAAN, L. et PENG, L. (1998). Comparison of tail index estimators. *Statistica Neerlandica*, 52(1):60–70. 24

DEHEUVELS, P., HAEUSLER, E. et MASON, D. (1988). Almost sure convergence of the Hill estimator. *Mathematical Proceedings of the Cambridge Philosophical Society*, 104(2): 371–381. 25

DEKKERS, A. et de HAAN, L. (1989). On the estimation of the extreme value index and large quantile estimation. *Annals of Statistics*, 17(4):1795–1832. 23, 26, 28, 58

DEKKERS, A., EINMALH, J. H. J. et de HANN, L. (1989). A moment estimator for the index of an extreme-value distribution. *Annals of Statistics*, 17(4):1833–1855. 24

DIEBOLT, J., GARDES, L., GIRARD, S. et GUILLOU, A. (2008). Bias-reduced estimators of the weibull tail-coefficient. *Test*, 17(2):311–331. 24

DREES, H. (1995). Refined Pickands estimator of the extreme value index. *Annals of Statistics*, 23(6):2059–2080. 28, 31

DREES, H. et KAUFMANN, E. (1998). Selecting the optimal sample fraction in univariate extreme value estimation. *Stochastic Processes and their Applications*, 75(2):149–172. 26

EINMAHL, J. (1990). The empirical distribution function as a tail estimator. *Statistica Neerlandica*, 44(3):79–82. 95

EL-ADLOUNI, S., BOBÉE, B. et OUARDA, T. B. (2007). Caractérisation des distributions à queue lourde pour l'analyse des crues. Rapport technique no r-929, INRS-ETE, Université du Québec. 25

EMBRECHTS, P., KLÜPPELBERG, C. et MIKOSCH, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Springer Verlag. 11, 13, 17, 23, 24, 58, 84, 113

FALK, M. (1995). On testing the extreme value index via the Pot-method. *Annals of Statistics*, 23(6):2013–2035. 24

FALK, M., HÜSLER, J. et REISS, R. D. (2004). *Laws of small numbers : Extremes and rare events*. 2nd edition, Birkhäuser. 60

FAN, J., HU, T. et TRUONG, Y. (1994). Robust nonparametric function estimation. *Scandinavian Journal of Statistics*, 21(4):433–446. 52

FERRATY, F., RABHI, A. et VIEU, P. (2005). Conditional quantiles for dependent functional data with application to the climatic el niño phenomenon. *Indian Journal of Statistics*, 67(2):378–398. 50, 126

FERRATY, F. et VIEU, P. (2000). Dimension fractale et estimation de la régression dans des espaces vectoriels semi-normés. *Comptes Rendus de l'Académie des Sciences*, 330:403–406. 48

FERRATY, F. et VIEU, P. (2006). *Nonparametric Functional Data Analysis : Theory and Practice*. Springer Series in Statistics, Springer. 48, 50, 54, 56, 68

FEUERVERGER, A. et HALL, P. (1999). Estimating a tail exponent by modelling departure from a Pareto distribution. *Annals of Statistics*, 27(2):760–781. 129

FILS, A. et GUILLOU, A. (2004). A new extreme quantile estimator for heavy-tailed distributions. *Comptes Rendus de l'Académie des Sciences*, 338(6):493–498. 34

FISHER, R. et TIPPET, L. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society*, 24:180–190. 10, 11, 17

FRÉCHET, M. (1927). Sur la loi de probabilité de l'écart maximum. *Annales de la Société Polonaise de Mathématique*, 6:93–116. 12

GANNOUN, A. (1989). *Estimation de la médiane conditionnelle*. Thèse de doctorat, Université de Paris VI. 48, 51

GANNOUN, A., GIRARD, S., GUINOT, C. et SARACCO, J. (2002). Trois méthodes non paramétriques pour l'estimation de courbes de référence : Application à l'analyse de propriétés biophysiques de la peau. *Revue de Statistique Appliquée*, 50(1):65–89. 52

GARDES, L. (2002). Estimating the support of a Poisson process via the Faber-Shauder basis and extreme values. *Publications de l'Institut de Statistique de l'Université de Paris*, 46(1–2):43–72. 88

GARDES, L. et GIRARD, S. (2008). A moving window approach for nonparametric estimation of the conditional tail index. *Journal of Multivariate Analysis*, 99(10):2368–2388. 63, 64, 65

GARDES, L. et GIRARD, S. (2010). Conditional extremes from heavy-tailed distributions : an application to the estimation of extreme rainfall return levels. *Extremes*, 13(2):177–204. 56, 74

GARDES, L., GIRARD, S. et LEKINA, A. (2010). Functional nonparametric estimation of conditional extreme quantiles. *Journal of Multivariate Analysis*, 101(2):419–433. 6

GARRIDO, M. (2002). *Modélisation des événements rares et estimation des quantiles extrèmes, méthodes de sélection de modèles pour les queues de distribution*. Thèse de doctorat, Université Grenoble 1. 2

GASSER, T. et HÜLLER, H.-G. (1984). Robust nonparametric function fitting. *Journal of the Royal Statistical Society, B*, 46(1):42–51. 48

GASSER, T. et MÜLLER, H. (1979). Kernel estimation of regression functions. *In Smoothing techniques for curves estimation*, pages 23–68. Springer Verlag, Germany. 101

GEFFROY, J. (1964). Sur un problème d'estimation géométrique. *Publications de l'Institut de Statistique de l'Université de Paris*, 13:191–200. 88

GIJBELS, I. et PENG, L. (2000). Estimation of a support curve via order statistics. *Extremes*, 3(3):251–277. 88

GIRARD, S. et JACOB, P. (2004). Extreme values and kernel estimates of point processes boundaries. *ESAIM : Probability and Statistics*, 8:150–168. 88

GIRARD, S. et JACOB, P. (2008). Frontier estimation via kernel regression on high power-transformed data. *Journal of Multivariate Analysis*, 99(3):403–420. 88

GIRARD, S. et MENNETEAU, L. (2005). Central limit theorems for smoothed extreme value estimates of point processes boundaries. *Journal of Statistical Planning and Inference*, 135(2):433–460. 88

GNEDENKO, B. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Annals of Mathematics*, 44(3):423–453. 10, 11, 17

GOMES, M., de HAAN, L. et PENG, L. (2002). Semi-parametric estimation of the second order parameter in statistics of extremes. *Extremes*, 5(4):207–229. 92

GOMES, M. I. (1999). Generalized jackknife moment estimator of the tail index. *Bulletin of the International Statistical Institute*, 58(1):401–402. 24

GOMES, M. I., FIGUEIREDO, F. et MENDONÇA, S. (2005a). Asymptotically best linear unbiased tail estimators under a second-order regular variation condition. *Journal of Statistical Planning and Inference*, 134(2):409–433. 24

GOMES, M. I. et OLIVEIRA, O. (2001). The bootstrap methodology in statistics of extremes : theory and applications - choice of the optimal sample fraction. *Extremes*, 4(4):331–358. 24

GOMES, M. I., PEREIRA, H. et MIRANDA, M. (2005b). Revisiting the role of the jackknife methodology in the estimation of a positive tail index. *Communications in Statistics–Theory and Methods*, 34(2):319–335. 24

GROENEBOOM, P., LOPUHAA, H. P. et de WOLF P.-P. (2003). Kernel-type estimators for the extreme value index. *Annals of Statistics*, 31:1956–1995. 24

GUIDA, M. et LONGO, M. (1988). Estimation of probability tails based on generalized extreme value distributions. *Reliability Engineering and System Safety*, 20(3):219–242. 17, 19

HAEUSLER, E. et TEUGELS, J. (1985). On asymptotic normality of Hill's estimator for the exponent of regular variation. *Annals of Statistics*, 13(2):743–756. 25

HALL, P., NUSSBAUM, M. et STERN, S. (1997). On the estimation of a support curve of indeterminate sharpness. *Journal of Multivariate Analysis*, 62(2):204–232. 88

HALL, P. et RONDE, N. (2000). Distribution and dependence-function estimation for bivariate extreme-value distributions. *Bernoulli*, 6(5):835–444. 54, 88

HALL, P. et TAJVIDI, N. (2000). Nonparametric analysis of temporal trend when fitting parametric models to extreme-value data. *Statistical Science*, 15(2):153–167. 54, 63

HÄRDLE, W., PARK, B. U. et TSYBAKOV, A. B. (1995). Estimation of a non sharp-support boundaries. *Journal of Multivariate Analysis*, 43(2):205–218. 88

HART, J. D. (1991). Comment to "choosing a kernel regression estimator". *Statistical Sciences*, 6(4):425–427. 48

HE, X. et SHI, P. (1994). Convergence rate of B-spline estimators of non parametric conditional quantiles functions. *Journal of Nonparametric Statistics*, 3(3–4):299–308. 51

HERRMANN, E. (1997). Local bandwidth choice in kernel regression estimation. *Journal of Computational and Graphical Statistics*, 6(1):35–54. 101

HERRMANN, E. (2000). Variance estimation and bandwidth selection for kernel regression. *In Smoothing and Regression : Approaches, computation and application*, pages 71–107. Wiley Series in Probability and Statistics. 101

HILL, B. (1975). A simple general approach to inference about the tail of a distribution. *Annals of Statistics*, 3(5):1163–1174. 17, 23, 25, 32, 64, 98, 99, 129

HORVÁTH, L. et YANDELL, B. S. (1988). Asymptotics of conditional empirical processes. *Journal of Multivariate Analysis*, 26(2):184–206. 47

HOSKING, J. R. M. et WALLIS, J. R. (1987). Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics*, 29(3):339–1349. 22, 24

HOSKING, J. R. M., WALLIS, J. R. et WOOD, E. F. (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted comments. *Technometrics*, 27:251–261. 20

JENKINSON, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *The Quarterly Journal of the Royal Meteorological Society*, 81(384):158–171. 11

JONES, M., MARRON, J. et SHEATHER, S. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91(433):401–407. 101

KOENKER, R. et BASSETT, G. (1978). Regression quantiles. *Econometrica*, pages 33–50. 54

KOENKER, R., NG, P. et PORTNOY, S. (1992). Nonparametric estimation of conditional quantile. *L1-Statistical analysis and related methods*, pages 217–229. ed Y. Dodge, Elsevier : Amsterdam. 52

KOENKER, R., NG, P. et PORTNOY, S. (1994). Quantile smoothing splines. *Biometrika*, 81(4):673–680. 52

KRATZ, M. et RESNICK, S. (1996). The qq-estimator and heavy tails. *Stochastic Models*, 12(4):699–724. 24, 28, 65

KRIGE, D. G. (1951). A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, 52(6):119–139. 74

MARRON, J. S. (1988). Automatic smoothing parameter selection : A survey. *Empirical Economics*, 13(3–4):187–208. 101

MASON, D. (1982). Laws of large numbers for sums of extreme values. *Annals of Probability*, 10(3):754–764. 25

MENNETEAU, L. (2008). Multidimensional limit theorems for smoothed extreme value estimates of point processes boundaries. *ESAIM : Probability and Statistics*, 12:273–307. 88

MINT EL MOUVIT, L. (2000). *Sur l'estimateur linéaire local de la fonction de répartition conditionnelle*. Thèse de doctorat, Université de Montpellier 2. 52

MOLINIÉ, G., YATES, E., CERESETTI, D., ANQUETIN, S., BOUDEVILLAIN, B., CREUTIN, J. et BOIS, P. (2008). Rainfall regimes in a mountainous mediterranean region : Statistical analysis at short time steps. Article soumis. 76

PARZEN, E. (1979). Nonparametric estimation statistical data modeling. *Journal of the American Statistical Association*, 74(365):105–131. 49

PENG, L. (1998). Asymptotically unbiased estimators for the extreme-value index. *Statistics and Probability Letters*, 38(2):107–115. 24

PICKANDS, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics*, 3(1):119–131. 17, 21, 24, 26, 31, 98

POIRAUD-CASANOVA, S. et THOMAS-AGNAN, C. (1998). Quantiles conditionnels. *Journal de la Société de française de Statistique*, 139(4):31–44. 52

PRESCOTT, P. et WALDEN, A. T. (1980). Maximum likelihood estimation of the parameters of generalized extreme-value distribution. *Biometrika*, 67(3):723–724. 20

PRESCOTT, P. et WALDEN, A. T. (1983). Maximum likelihood estimation of the parameters of the three-parameter generalized extreme-value distribution for censored samples. *Journal of Statistical Computation and Simulation*, 16(3–4):241–250. 20

RAMSAY, J. et SILVERMAN, B. (1997). *Functional Data Analysis*. Springer Verlag. 54

RAMSAY, J. et SILVERMAN, B. (2002). *Applied functional Data Analysis*. Springer Verlag. 54

RÉNYI, A. (1953). On the theory of order statistics. *Acta Mathematica Hungarica*, 4(3–4):191–231. 18, 43, 80

RÉNYI, A. et SULANKE, R. (1963). Über die konvexe hülle von n zufällig gewählten punkten. *Z. Wahrsch. Verw. gebiete*, 2:75–84. 88

RÉNYI, A. et SULANKE, R. (1964). Über die konvexe hülle von n zufällig gewählten punkten, ii. *Z. Wahrsch. Verw. gebiete*, 3:138–147. 88

RESNICK, S. (1987). *Extreme Values, regular Variation, and Point Process*. Springer Verlag, New-York. 13, 14, 15

ROSEN, O. et WEISSMAN, I. (1996). Comparison of estimation methods in extreme value theory. *Communication in Statistics-Theory and Methods*, 24(4):759–773. 24

ROUSSAS, G. G. (1969). Nonparametric estimation of the transition distribution function of Markov process. *Annals of Mathematical Statistics*, 40(4):1386–1400. 46, 48

ROYSTON, P. et WRIGHT, E. M. (1998). How to construct normal ranges for fetal variables. *Utrasound in Obstetrics and Gynecology*, 11(1):30–38. 46

S. A. PADOAN, S. A., RIBATET, M. et SISSON, S. A. (2010). Likelihood-based inference for max-stable processes. *Journal of the American Statistical Association*, 105(489):263–277. 74

SAMANTA, M. (1989). Nonparametric estimation of conditional quantiles. *Statistics and Probability Letters*, 7(5):407–412. 46, 49

SCHULTZE, J. et STEINEBACH, J. (1996). On least squares estimates of an exponential tail coefficient. *Statistics and Decisions*, 14(3):353–372. 24, 28, 65

SMITH, R. (1987). Estimating tails of probability distributions. *Annals of Statistics*, 15(3): 1174–1207. 22, 25

SMITH, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, 72(1):67–92. 20, 24

SMITH, R. L. (1989). Extreme value analysis of environmental time series : An application to trend detection in ground-level ozone. *Statistical Science*, 4(4):367–377. 54, 88

STONE, C. J. (1977). Consistent nonparametric regression. *Annals of Statistics*, 5(4):595–645. 47, 52

STUTE, W. (1986). Conditional empirical processes. *Annals of Statistics*, 14(2):638–647. 46, 47, 48

TANGO, T. (1998). Estimation of age-specific reference ranges via smoother avas. *Statistics in Medecine*, 17(11):1231–1243. 46

TRUONG, Y. K. (1989). Asymptotic properties of kernel estimators based on local medians. *Annals of Statistics*, 17(2):606–617. 51

TSOURTI, Z. et PANARETOS, J. (2001). A simulation study on the performance of extreme-value index estimators and proposed robustifying modifications. *5th Hellenic European Conference on Computer Mathematics and its Applications, Athens, Greece*, 2:847–852. 24, 25

TSOURTI, Z. et PANARETOS, J. (2003). Stochastic musings : Perspectives from the pioneers of the late 20th century. pages 141–160. Laurence Erlbaum, John Panaretos édition. 24, 25

TSYBAKOV, A. (1986). Robust reconstruction of functions by the local-approximation method. *Problems of Information Transmission*, 22(2):133–146. 52

TUKEY, J. W. (1961). Curves as parameters, and touch estimation. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1:681–694. 50

VIEU, P. (1994). Bandwidth selection for kernel regression. *In* HÄRDLE, W. et SIMAR, L., éditeurs : *Computer Intensive Methods in Statistics*, pages 134–149. Physica Verlag, Heidelberg, Germany. 101

VIEU, P. (1999). Multiple kernel procedure : an asymptotic support. *Scandinavian Journal of Statistics*, 26(1):61–72. 101

VIHAROS, L. (1993). Asymptotic distributions of linear combinations of extreme values. *Acta Scientiarum Mathematicarum*, 58:211–231. 24

VIHAROS, L. (1995). Limit theorems for linear combinations of extreme values with applications to inference about the tail of a distribution. *Acta Scientiarum Mathematicarum*, 60:761–777. 24

von MISES, R. (1936). La distribution de la plus grande de n valeurs. *Revue de Mathématique Union Interbalcanique*, 1:141–160. 11

WEINSTEIN, S. B. (1973). Theory and application of some classical and generalized asymptotic distributions of extreme values. *IEEE Transactions on Information Theory*, 19(2):148–154. 20

WEISSMAN, I. (1978). Estimation of parameters and large quantiles based on the k-largest observations. *Journal of the American Statistical Association*, 73(364):812–815. 6, 17, 23, 31, 32, 34, 58, 91

YAO, Q. (1999). Conditional predictive regions for stochastic processes. Rapport technique, University of Kent at Canterbury. 101

YU, K. et JONES, M. (1998). Local linear quantile regression. *Journal of the American Statistical Association*, 93:228–237. 52

Résumé : L'objectif de ce travail est de proposer de nouveaux estimateurs de quantiles extrêmes dans le cadre conditionnel c'est-à-dire dans la situation où la variable d'intérêt Y, supposée aléatoire et réelle, est mesurée simultanément avec une covariable X. Pour ce faire, nous nous intéressons à l'étude des valeurs extrêmes d'un échantillon d'observations indépendantes dont la loi conditionnelle de Y en un point x de la covariable X est à « queue lourde ». Selon la nature de la covariable, nous considérons deux situations. *Primo*, lorsque la covariable est déterministe et de dimension finie ou infinie (i.e covariable fonctionnelle), nous proposons d'estimer les quantiles extrêmes par la méthode dite de la « fenêtre mobile ». La loi limite des estimateurs ainsi construits est ensuite donnée en fonction de la vitesse de convergence de l'ordre du quantile vers un. *Secundo*, lorsque la covariable est aléatoire et de dimension finie, nous montrons que sous certaines conditions, il est possible d'estimer les quantiles extrêmes conditionnels au moyen d'un estimateur à « noyau » de la fonction de survie conditionnelle. Ce résultat nous permet d'introduire deux versions lisses de l'estimateur de l'indice de queue conditionnel indispensable lorsque l'on veut extrapoler. Nous établissons la loi asymptotique de ces estimateurs. Par ailleurs, nous considérons le cas sans covariable (non conditionnel) lorsque la fonction de répartition est à « queue lourde ». Nous proposons et étudions un nouvel estimateur des quantiles extrêmes. Afin d'apprécier le comportement de nos nouveaux outils statistiques, des résultats sur simulation ainsi que sur des données réelles sont présentés.

Mots clés : Valeurs extrêmes, Estimation non-paramétrique, Quantiles conditionnels, Lois à queues lourdes, Estimateur à noyau, Fenêtre mobile.

Titre : ESTIMATION NON-PARAMÉTRIQUE DES QUANTILES EXTRÊMES CONDITIONNELS.

Spécialité : MATHÉMATIQUES APPLIQUÉES.

Thèse réalisée à l'Institut National de Recherche en Informatique et Automatique (INRIA), Centre de Recherche Grenoble - Rhône-Alpes.

Abstract : The main goal of this thesis is to propose new estimators of extreme quantiles in the conditional case, that is to say in the situation where the variable of interest Y, supposed to be random and real, is recorded simultaneously with some covariate information X. To this aim, we focus on the case where the conditional distribution of Y given $X = x$ is "heavy-tailed". Two situations are considered. First, when the covariate is deterministic and finite-dimensional or infinite-dimensional (i.e functional covariate), we propose to estimate the extreme quantiles by the "moving window approach". The asymptotic distribution of the proposed estimators is given in the case where the quantile is in the range of data or near and even beyond the sample. Next, when the covariate is random and finite-dimensional, we show that under some conditions, it is possible to estimate these extreme quantiles using a kernel estimator of the conditional survival function. As a consequence, this result allows us to introduce two smooth versions of the conditional tail index estimator necessary to extrapolate. Asymptotic distributions of these estimators are established. Furthermore, we also considered the case without covariate. When the underlying, the cumulative distribution function is "heavy-tailed". A new unconditional extreme quantile estimator is introduced and studied. To assess the behavior of all our new statistical tools, numerical experiments on simulated data are provided and illustrations on real datasets are presented.

Keywords : Extreme-value analysis, Nonparametric estimation, Conditional quantiles, Heavy-tail distributions, Kernel estimator, Moving window.

Titre : NONPARAMETIC ESTIMATION OF CONDITIONAL EXTREME QUANTILES.

Speciality : APPLIED MATHEMATICS.

Thesis fulfilled at the National Institute for Research in Computer Science and Control (French : Institut National de Recherche en Informatique et Automatique, INRIA), Research Center Grenoble - Rhône-Alpes.