Estimation des limites d'extrapolation par les lois de valeurs extrêmes. Application à des données environnementales.

Clément Albert

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTE UNIVERSITE GRENOBLE ALPES

Spécialité : **Mathématiques Appliquées**

Arrêté ministériel : 25 mai 2016

Présentée par

Clément ALBERT

Thèse dirigée par **Stéphane GIRARD, Directeur de recherche, Inria** et codirigée par **Anne DUTFOY, Ingénieure-chercheuse senior, EDF R&D**

préparée au sein du **Laboratoire Jean Kuntzmann** dans **l'École Doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique**

Estimation des limites d'extrapolation par les lois de valeurs extrêmes. Application à des données environnementales

Thèse soutenue publiquement le **17 Décembre 2018**, devant le jury composé de :

Madame Liliane BEL
Professeure, Université Paris-Saclay, Rapporteure
Madame Ivette GOMES
Professeure émérite, Université de Lisbonne, Portugal, Rapporteure
Monsieur Laurent GARDES
Professeur, Université de Strasbourg, Examinateur
Madame Clémentine PRIEUR
Professeure, Université Grenoble-Alpes, Présidente
Madame Anne DUTFOY
Ingénieure-chercheuse senior, EDF R&D, Co-directrice de thèse
Monsieur Stéphane GIRARD
Directeur de recherche, Inria, Directeur de thèse

Remerciements

En premier lieu, je remercie mes directeurs de thèse, Anne Dutfoy et Stéphane Girard, pour m'avoir fait confiance en me proposant ce sujet de thèse mêlant problématiques théoriques et industrielles. Loin de ne faire que proposer, ils ont été de grands superviseurs, à la fois acteurs du projet, mais sachant également me laisser le choix dans la manière d'aborder le problème. J'ai beaucoup apprécié cette liberté. Je vous remercie également pour les connaissances que vous vous êtes efforcés de me transmettre tout au long de ces trois années de thèse ainsi que pour m'avoir toujours poussé à multiplier les expériences, aussi diverses soient-elles. Le peu de fois où je n'ai pas joué le jeu, je l'ai invariablement regretté ! Enfin, j'ai hautement apprécié votre simplicité, qui fait que notre entente a été immédiate et le travail d'autant plus agréable.

Je remercie également Liliane Bel et Ivette Gomes pour avoir gracieusement accepté de rapporter cette thèse. Je suis très honoré de l'intérêt qu'elles ont porté à mon travail. Merci également pour vos divers retours, suggestions et remarques quant à ce manuscrit.

Merci à Clémentine Prieur et Laurent Gardes pour m'avoir fait l'honneur de faire partie de mon jury. Je tiens également à vous remercier pour votre disponibilité et le temps que vous m'avez consacré au cours de ces dernières années, que ce soit dans le cadre de l'enseignement ou de la recherche. J'ai grandement apprécié travailler avec vous.

Je remercie les multiples équipes EDF R&D en lien avec le projet MADONE, en particulier l'équipe PER ICLES (anciennement MRI) pour les échanges fructueux que j'ai pu entretenir avec eux.

Merci également aux membres du LJK qui m'ont invité à présenter lors de séminaires et merci à ceux qui sont dogmatiquement venus y assister.

Je remercie l'Inria pour m'avoir fourni des conditions de travail idéales. En particulier, je tiens chaleureusement à remercier les membres de l'équipe MISTIS (anciens membres compris !) pour leur entrain et leur bonne humeur. Ce fût un réel plaisir que de venir au bureau chaque matin ! Merci encore pour ces moments de partage !

Enfin, merci à ma famille pour leur soutien sans faille !

Résumé

Cette thèse se place dans le cadre de la Statistique des valeurs extrêmes. Elle y apporte trois contributions principales.

L'estimation des quantiles extrêmes se fait dans la littérature en deux étapes. La première étape consiste à utiliser une approximation des quantiles basée sur la théorie des valeurs extrêmes. La deuxième étape consiste à estimer les paramètres inconnus de l'approximation en question, et ce en utilisant les plus grandes valeurs du jeu de données. Cette décomposition mène à deux erreurs de natures différentes, la première étant une erreur déterministe, dite d'approximation ou encore d'extrapolation, la seconde consituant une erreur d'estimation aléatoire.

La première contribution de cette thèse est l'étude théorique de cette erreur d'extrapolation mal connue. Cette étude est menée pour deux types d'estimateurs différents, tous deux cas particuliers de l'approximation par la loi de Pareto généralisée : l'estimateur Exponential Tail dédié au domaine d'attraction de Gumbel et l'estimateur de Weissman dédié à celui de Fréchet. Nous montrons alors que l'erreur en question peut s'interpréter comme le reste d'un développement de Taylor d'ordre un. Des conditions nécessaires et suffisantes sont alors établies de telle sorte que l'erreur tende vers zéro quand la taille de l'échantillon augmente. De manière originale, ces conditions mènent à une division du domaine d'attraction de Gumbel en trois parties distinctes. En comparaison, l'erreur d'extrapolation associée à l'estimateur de Weissman présente un comportement unifié sur tout le domaine d'attraction de Fréchet. Des équivalents de l'erreur sont fournis et leur comportement est illustré numériquement.

La deuxième contribution est la proposition d'un nouvel estimateur des quantiles extrêmes. Le problème est abordé dans le cadre du modèle dit des "lois à queue de type log-Weibull généralisé", où le logarithme de l'inverse du taux de hasard cumulé est supposé à variation régulière étendue. Après une discussion sur les conséquences de cette hypothèse, nous proposons des estimateurs des paramètres du modèle. Ces estimateurs sont alors utilisés afin de construire un nouvel estimateur des quantiles extrêmes. La normalité asymptotique de ce dernier ainsi que celle des paramètres associés est alors établie et leur comportement en pratique est évalué sur données réelles et simulées.

La troisième contribution de cette thèse est la proposition d'outils permettant en pratique de quantifier les limites d'extrapolation d'un jeu de données. Ces outils consistent en des estimateurs des erreurs d'extrapolation associées aux approximations Exponential Tail et Weissman. Ils se basent sur les deux contributions précédentes. Dans un premier temps, nous utilisons l'étude théorique faite des différentes erreurs d'extrapolation pour proposer des équivalents de ces dernières. Ces nouveaux équivalents se veulent généraux et utilisables en pratique. Dans un second temps, nous utilisons les estimateurs proposés du modèle dit des "lois à queue de type log-Weibull généralisé" pour pouvoir estimer lesdits équivalents. Après avoir évalué les performances sur données simulées des nouveaux estimateurs ainsi construits, nous estimons les limites d'extrapolation associées à trois jeux de données réelles constitués de mesures journalières de variables environnementales. Dépendant de l'aléa climatique considéré, nous montrons que ces limites sont plus ou moins contraignantes.

Abstract

This thesis takes place in the extreme value statistics framework. It provides three main contributions to this area. Extreme quantile estimation is a two step approach. First, it consists in proposing an extreme value based quantile approximation. Then, estimators of the unknown quantities are plugged in the previous approximation leading to an extreme quantile estimator.

The first contribution of this thesis is the study of the extrapolation error, which is the error due to the extreme value based approximation of the true quantile. These investigations are carried out using two different kind of estimators, both based on the well-known Generalized Pareto approximation : the Exponential Tail estimator dedicated to the Gumbel maximum domain of attraction and the Weissman estimator dedicated to the Fréchet one. It is shown that the extrapolation error can be interpreted as the remainder of a first order Taylor expansion. Necessary and sufficient conditions are then provided such that this error tends to zero as the sample size increases. Interestingly, in case of the so-called Exponential Tail estimator, these conditions lead to a subdivision of Gumbel maximum domain of attraction into three subsets. In constrast, the extrapolation error associated with Weissman estimator has a common behavior over the whole Fréchet maximum domain of attraction. First order equivalents of the extrapolation error are then derived and their accuracy is illustrated numerically.

The second contribution is the proposition of a new extreme quantile estimator. The problem is addressed in the framework of the so-called "log-Generalized Weibull tail limit" model, where the logarithm of the inverse cumulative hazard rate function is supposed to be of extended regular variation. Based on this model, estimators of the parameters are proposed. Then, a new estimator of extreme quantiles is derived from the latter. Its asymptotic normality is established and its behavior in practice is illustrated on both real and simulated data.

The third contribution of this thesis is the proposition of new mathematical tools allowing the quantification of extrapolation limits associated with a real dataset. These tools consist in some estimators of the extrapolation error. To build them, we take advantages on one hand of the first study we did by proposing first order approximations which are widely applicable in practice. On the other hand, we use the proposed estimators of the "log-Generalized Weibull tail limit" model to estimate the previous approximations. Performances of the obtained estimators are illustrated on simulated data. These estimators are finally used to estimate the extrapolation limits associated with three real datasets consisting in daily measures of some environmental variables. Depending on the climatic phenomena, we show that the extrapolation limits can be more or less stringent.

Table des matières

Introduction générale

Des évènements extrêmes ont lieu en permanence dans le monde. Parmi les plus préoccupants sont les évènements météorologiques, qui peuvent avoir de sérieuses répercussions humaines et matérielles. Citons en trois en particulier, qui ont eu lieu pendant la rédaction de ce manuscrit :

1. Le 23 Juillet 2018, des feux ravagent les abords d'Athènes, en Grèce. Le bilan officiel est de 97 morts et 187 blessés. C'est sans compter les pertes matérielles, plus de 1000 maisons détruites, 300 véhicules et des miliers d'hectares de forêt brûlés.
2. Du 29 Juillet au 19 Août 2018, une série de séismes, d'une magnitude allant de 5.4 à 7 secoue l'île indonésienne de Lombok, causant la mort de 555 personnes, et le déplacement de centaines de milliers d'autres.
3. Le 11 Août 2018, des pluies torrentielles provoquent des inondations dans le sud de la Colombie. Un plan d'urgence est décrété : 30 000 personnes sont évacuées des municipalités touchées, où ces mêmes pluies avaient fait 316 morts en Avril 2017.

L'estimation des évènements extrêmes

Ces évènements se caractérisent par deux aspects. Premièrement, ils sont rares par essence, au sens où leur probabilité d'occurence est très faible (en revanche, le fait qu'un évènement soit rare n'implique pas qu'il soit extrême). Deuxièmement, ils ont d'énormes impacts, qu'ils soient humains, économiques ou financiers. C'est pour cette dernière raison qu'il est d'un grand intérêt de s'en prémunir, soit en tentant de les éviter, soit en atténuant leur impact. Dans tous les cas, cela suppose de pouvoir les prédire et d'en évaluer l'importance.

Pour répondre à cette question et donc décider de combien il faut investir dans la réduction des risques, on pourrait dans un premier temps penser à prédire ces évènements en utilisant des modèles physiques, comme cela est fait en météorologie pour les précipitations par exemple. Cependant, il s'agit de prédire en avance des évènements extrêmes. Or nous ne sommes pas capables à l'heure actuelle de donner des prévisions à plus de quelques semaines, je pense notamment en météorologie.

L'autre moyen de prédire ces évènements consiste en des prédictions probabilistes, en se basant sur un échantillon de données. L'idée est, étant donné un évènement extrême, d'en caractériser la probabilité d'occurence. En ce sens, les mathématiques, et plus particulièrement la théorie probabiliste et la statistique en tant que discipline nous offrent des outils très puissants permettant de répondre à ces questions.

Prédiction probabiliste, un premier exemple

Pour illustrer notre propos, considérons un jeu de données constitué de $n = 21457$ mesures de vitesses de vent (en m/s) au rythme d'une mesure par jour de 1952 à 2011 et notons X la variable aléatoire associée. Les n réalisations sont notées $x_1, ..., x_n$. Un même histogramme des données, répertoriant des vents jusqu'à $44 m/s$, est représenté plusieurs fois Figure 1.

Dans un premier temps, intéressons nous à la probabilité d'observer un vent supérieur à $15 m/s$. Au regard du graphe situé en haut à gauche de la Figure 1,

$$\begin{aligned} \mathbb{P}(\mathrm{X} \geq 15) &\simeq nb(x_i \geq 15)/n \\ &= 8185/21457 \\ &\simeq 38\%. \end{aligned}$$

Par conséquent, il y a un vent de ce type en moyenne une fois tous les $21457/8185 \simeq 2.6$ jours. De la même façon,

$$\mathbb{P}(\mathrm{X} \geq 25) \simeq nb(x_i \geq 25)/n = 1317/21457 \simeq 6\%$$

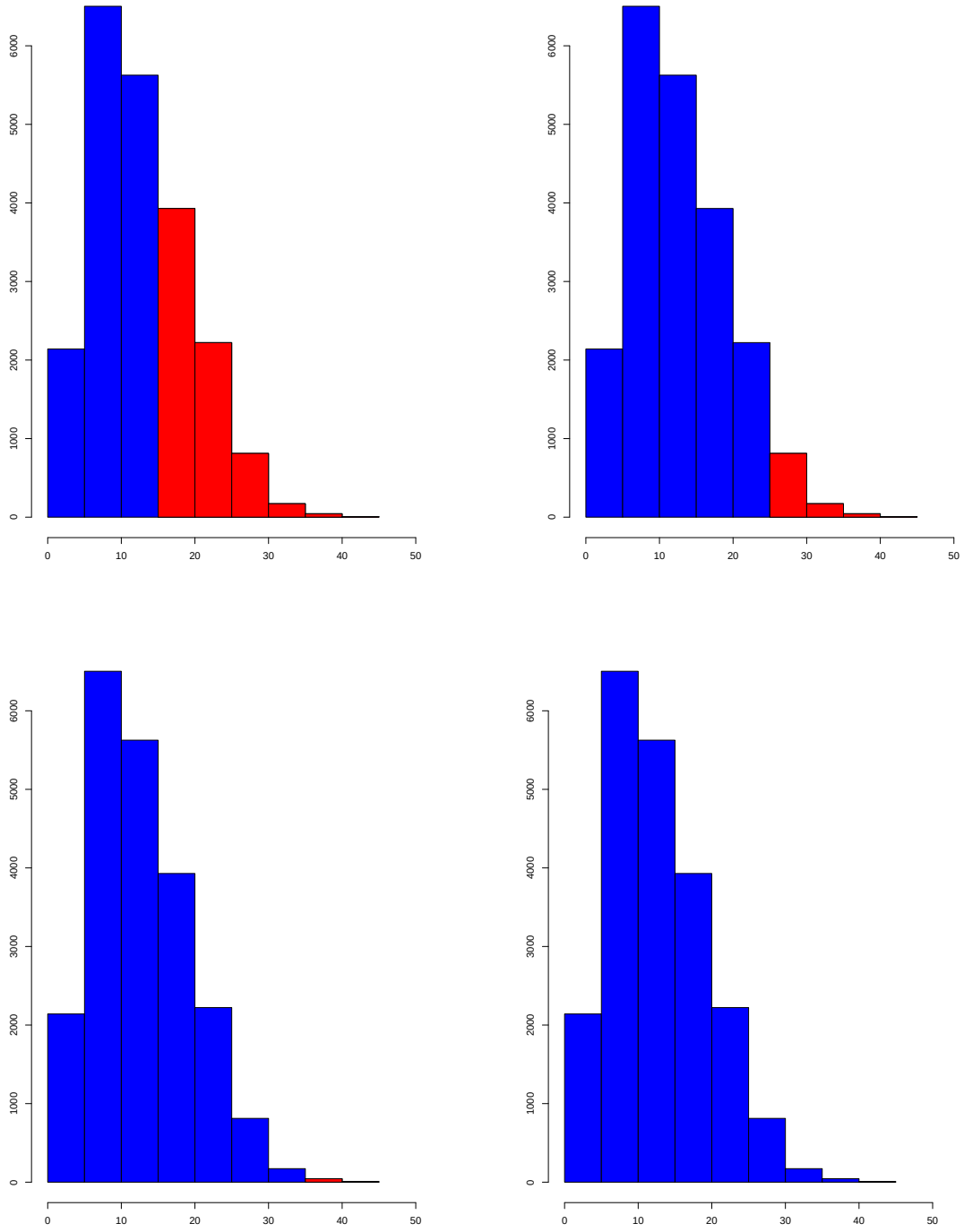

FIGURE 1 – Histogramme des vitesses de vent. En bleu, les vitesses de vent inférieures respectivement à $15m/s$, $25m/s$, $35m/s$ et $45m/s$ (de gauche à droite et de haut en bas). En rouge, les vitesses de vent supérieures à ces mêmes valeurs.

et

$$\mathbb{P}(\mathrm{X} \geq 35) \simeq nb(x_i \geq 35)/n = 66/21457 \simeq 0.3\%$$

de telle sorte qu'un vent de $25m/s$ a lieu en moyenne tous les 16 jours et un vent de $35m/s$ tous les 325 jours (voir graphes situés en haut à droite et en bas à gauche de la Figure 1). Par conséquent, si l'on souhaite construire une infrastructure perenne plus de 325 jours en moyenne, il sera donc nécessaire de faire en sorte qu'elle résiste à un vent de $35m/s$.

Si l'on considère maintenant la probabilité d'observer un vent supérieur à $45m/s$, un raisonnement

analogue nous permet de conclure que

$$\mathbb{P}(\mathrm{X} \geq 45) \simeq nb(x_i \geq 45)/n = 0.$$

Afin de donner une réponse non triviale à ce problème, on ne peut donc pas estimer la probabilité d'occurence d'un tel vent à partir de l'histogramme.

Prédiction probabiliste, un deuxième exemple

Une autre façon de voir le problème est de considérer le problème dual suivant : étant donné un petit nombre p, déterminer le niveau (aussi appelé quantile d'ordre $1-p$ dans le vocabulaire statistique) des digues de telle sorte que la probabilité qu'il y ait une crue lors d'une année donnée soit égale à p. Il s'agit ici de déterminer la hauteur de la vague de probabilité d'occurrence p et non plus la probabilité associée à un évènement extrême donné. Evaluer l'importance de l'évènement est donc crucial. Une sous-évaluation du risque amènera à une exposition au danger plus accrue alors qu'une sur-évaluation entrainera des coûts de construction plus importants.

Admettons que nous soyons en possession de mesures journalières du niveau de l'eau s'étendant sur $1/p$ années. Estimer la hauteur de la plus grande vague par le maximum de l'échantillon dont nous disposons consituerait une bonne approximation de la vraie hauteur inconnue. Cependant, si p est très petit, l'échantillon dont on dispose n'est généralement pas assez grand.

Trois approches stochastiques pour répondre au problème

Ces deux exemples nous permettent ainsi de mettre en évidence deux difficultés pratiques concernant les prédictions probabilistes à partir d'un échantillon de données.

— Premièrement, on dispose d'un grand nombre d'observations pour les évènements plus fréquents et peu pour les évènements extrêmes, voire pas du tout. La difficulté réside donc dans le fait de prédire des évènements extrêmes à l'aide d'évènements fréquents. C'est le contrecoup du premier aspect des évènements extrêmes, qui sont par définition rares.

— Deuxièmement, les échantillons de mesures dont on dispose sont en général petits, couvrant des périodes d'au mieux une centaine d'années de mesures. Cela renforce d'autant plus le fait qu'il est probable que nous n'ayons pas observé d'évènements extrêmes.

Une extrapolation en dehors de l'échantillon se révèle donc généralement nécessaire. L'idée est de faire le meilleur usage des plus grandes valeurs de l'échantillon. Intuitivement parlant, on imagine bien que ce sont elles qui contiennent l'information pertinente. Trois types d'approches existent.

L'approche par simulations La première approche consiste à simuler suivant un modèle physique stochastique. Il s'agit alors de générer suffisamment d'évènements pour arriver à observer plusieurs évènements extrêmes du même calibre que celui d'intérêt, réduisant le problème à un problème classique d'estimation de quantiles (au sens où ces derniers se situent dans l'échantillon). A ce sujet, citons la méthode SHYPRE, ARNAUD et collab. [2016, 2017]; ARNAUD et LAVABRE [1999], qui a pour but de générer de très longues chroniques de hyétogrammes et d'hydrogrammes afin de déterminer des quantiles de pluie, de débit et de cote du plan d'eau d'un barrage.

L'approche paramétrique L'approche paramétrique a pour but d'utiliser les données disponibles afin de modéliser au mieux la loi du hasard régissant le phénomène. Elle consiste à utiliser une famille de lois statistiques paramétriques qui ont tendance à bien s'ajuster à la queue de distribution empirique des données rencontrées dans un domaine d'application précis. Par exemple, dans le cas des pluies, une loi dite exponentielle pourrait être envisagée. A ce sujet, voir EVIN et collab. [2016]. Une fois le modèle ajusté aux données, le calcul des quantiles extrêmes est immédiat, basé sur l'évaluation de la fonction quantile théorique de la loi ajustée au point voulu.

Une approche connexe est d'ajuster ad hoc plusieurs familles de lois paramétriques choisies aux données et de choisir celle qui s'ajuste au mieux aux données.

L'approche utilisant la théorie des valeurs extrêmes Enfin, la dernière approche utilise la théorie des valeurs extrêmes. Cette théorie consiste à extrapoler directement à partir du jeu de données dont on dispose, moyennant des hypothèses de régularité sur les phénomènes observés, c'est à dire à déduire le comportement des évènements extrêmes à partir d'évènements plus fréquents.

Les premiers développements de cette théorie sont attribués à Nicolas Bernoulli en 1709 alors que la première application est due à Fuller en 1914. Le principal essor de la théorie des valeurs extrêmes est cependant dû aux résultats de FISHER et TIPPETT [1928] et GNEDENKO [1943] sur la convergence en loi de la valeur maximale d'une suite de variables aléatoires indépendantes et indentiquement distribuées puis aux résultats de PICKANDS [1975] sur la convergence en loi des excès au dessus d'un seuil.

Cette théorie est maintenant bien développée : nous renvoyons le lecteur à RESNICK [1987], BEIRLANT et collab. [2006], DE HAAN et FERREIRA [2007], GOMES et GUILLOU [2015] ou encore COLES et collab. [2001] pour des synthèses sur la théorie des valeurs extrêmes.

Ses domaines d'application sont nombreux : citons GUMBEL [1954], DE HAAN [1990], KATZ et collab. [2002], WILLEMS et GUILLOU [2006], EL METHNI et collab. [2012], DUTFOY et collab. [2014] pour l'hydrologie, CERESETTI et collab. [2012], COLES et collab. [2003]; COLES et TAWN [1996], GARDES et GIRARD [2010], METHNI et collab. [2014], BECHLER et collab. [2015], GAUME et collab. [2013] pour la météorologie, ou encore BEL et collab. [2008], ROOTZÉN et TAJVIDI [2001] pour la climatologie. La théorie des valeurs extrêmes trouve également des applications en fiabilité, DITLEVSEN [1994], en assurance, BEIRLANT et TEUGELS [1992] et en finance EMBRECHTS et collab. [2013, 1999].

Enfin, la théorie des valeurs extrêmes a l'avantage de fournir des garanties théoriques aussi bien sur le biais que sur la variance des estimateurs des probabilités ou quantiles, et ce aussi bien dans le cadre asymptotique via des approximations que dans le cadre pratique, en utilisant par exemple des estimations d'intervalles de confiance ou des tests d'hypothèses.

C'est dans le cadre de cette troisième approche que nous nous plaçons dans cette thèse.

La prévention des risques au sein d'EDF

Cette thèse est co-financée par EDF R&D dans le cadre du projet MADONE : Méthodes pour les Agressions d'Origine Naturelle Externe. La R&D d'EDF s'est engagée depuis plus d'une dizaine d'années sur le sujet de la modélisation des valeurs extrêmes en mettant au point une méthodologie d'analyse de ces valeurs dans le cadre univarié (valeurs de dimension 1) et multivarié (valeurs vectorielles, de dimension en général 2 ou 3). Cette méthodologie est déroulée pour effectuer de nombreuses études statistiques de valeurs extrêmes à partir de relevés de variables météorologiques (température, débit, vitesse de vent, ...). Elles sont effectuées par plusieurs départements de la R&D, comme le LNHE pour ce qui concerne les variables hydrauliques (débits, houle, ...) et MFEE pour ce qui est du traitement de la température et du vent. Le département MRI apporte son soutien via son expertise en modélisation probabiliste et statistique, dans le cadre univarié et multivarié. Ces études servent à dimensionner les ouvrages EDF (centrales nucléaires, barrages hydrauliques...) aux agressions météorologiques, de type inondation, tempête, sécheresse...

Ces études consistent, à partir d'une loi de valeurs extrêmes calée sur des données, à déterminer les quantiles extrêmes de période de retour centenale voire millénale ou à calculer des probabilités de dépassement de seuil extrême.

Les quantiles extrêmes, de période de retour 100 ans et plus, dépendent du modèle de valeurs extrêmes utilisé. Ils dépendent aussi du nombre de données disponibles pour caler les modèles de valeurs extrêmes et de leur qualité. Afin de rendre plus robuste les prises de décision se pose donc la question de quantifier la sensibilité des valeurs de dimensionnement à des analyses mettant en œuvre des extrapolations par des lois de valeurs extrêmes.

Cette thèse pose la question des limites de crédibilité des extrapolations par des lois de valeurs extrêmes, dans le cadre univarié. Pour d'autres références sur les limites de la théorie des valeurs extrêmes, citons DEGEN et collab. [2007] et MIGNOLA et UGOCCIONI [2005].

Organisation de la thèse

Cette thèse s'articule autour de quatre chapitres principaux.

1. Le Chapitre 1 consiste en une introduction à la théorie des valeurs extrêmes. Se plaçant dans le cadre univarié, il présente les outils utilisés dans cette thèse. La Partie 1.1 commence ainsi par s'intéresser au comportement asymptotique des plus grandes valeurs d'un échantillon de variables aléatoires indépendantes et identiquement distribuées. Les lois limites du maximum de l'échantillon et des excès au dessus d'un seuil y sont comparées. Les différentes caractérisations de ces lois limites, ou domaines d'attraction, font l'objet de la Partie 1.2. Enfin, la Partie 1.3 traite de l'estimation des quantiles extrêmes, brique de base aux travaux de thèse que nous présentons. En particulier, un état de l'art - proposant une revue de la littérature récente - y est mené.

2. Le Chapitre 2 s'intéresse à l'étude de l'erreur (relative) entre le vrai quantile d'une loi et plusieurs de ses approximations basées sur la théorie des valeurs extrêmes. Après avoir, Partie 2.1, brièvement défini cette dernière, la Partie 2.2 présente, sous la forme d'un article soumis pour publication, nos contributions à l'étude d'une telle erreur. Cet article se découpe en six sous-parties distinctes. Le Paragraphe "Introduction" propose une brève déclinaison des différentes notations tandis que le Paragraphe "Theoretical framework" présente le cadre théorique de l'étude. L'application de ce cadre théorique à l'erreur d'extrapolation associée à l'estimateur "Exponential Tail" (ET), dédié au domaine d'attraction de Gumbel, est l'objet du Paragraphe "Application to the ET approximation". Le Paragraphe "Application to Weissman approximation" est quant à lui dédié à la mise en application de ce même cadre théorique à l'erreur d'extrapolation associée au célèbre estimateur de Weissman, lui-même dédié au domaine d'attraction de Fréchet. Le Paragraphe "Numerical illustrations" illustre sur données simulées des équivalents de l'erreur d'extrapolation établis dans les différents paragraphes ci-dessus. Les différentes preuves sont compilées Paragraphe "Proofs of main results". Enfin, la Partie 2.3 propose quelques perspectives de travail qui se dégagent de cette étude.
3. Le Chapitre 3 propose un nouvel estimateur des quantiles extrêmes basé sur le modèle "log-generalized Weibull-tail" introduit par DE VALK [2016b]. La Partie 3.1 constitue un bref résumé des motivations liées à nos travaux. L'ensemble de nos contributions est alors présenté Partie 3.2 sous la forme d'un article soumis pour publication. Le Paragraphe "Introduction" définit le modèle et les notations utilisées à travers l'article. L'inférence des paramètres du modèle ainsi que la proposition d'un nouvel estimateur des quantiles extrêmes sont l'objet du Paragraphe "Inference". Le Paragraphe "Main results" établit quant à lui des résultats de normalité asymptotique. Des illustrations sur données simulées et données réelles sont respectivement présentées Paragraphes "Validation on simulations" et "Illustration on real data". Une comparaison avec l'estimateur de DE VALK et CAI [2018] y est également proposée. Les preuves des résultats sont relayées en annexe "Appendix : Proofs". Enfin, la Partie 3.3 dégage plusieurs perspectives de recherche.
4. Le Chapitre 4 illustre plusieurs exemples d'estimation des limites d'extrapolation sur des données environnementales. La Partie 4.1 exhibe un estimateur de l'erreur d'extrapolation associée à l'approximation Exponential Tail, dédiée au domaine d'attraction de Gumbel. Dans un premier temps, nous nous basons sur l'étude de l'erreur d'extrapolation faite au Chapitre 2 pour proposer un nouvel équivalent de cette dernière. Cet équivalent se veut plus général que ceux proposés au Chapitre 2. Ensuite, nous utilisons les estimateurs introduits Chapitre 3 du modèle des "lois à queue de type log-Weibull généralisé" afin d'estimer l'équivalent en question. Le comportement de l'estimateur de l'erreur d'extrapolation alors obtenu est étudié sur données simulées. Utilisant cet estimateur, nous discutons Partie 4.2 des limites d'extrapolation associées à deux jeux de données, l'un constitué de mesures journalières du débit du Rhône et l'autre de mesures de vitesses instantanées de vent relevées à Reims. L'estimation de l'erreur d'extrapolation associée à l'approximation Weissman, dédiée au domaine d'attraction de Fréchet, fait l'objet de la Partie 4.3. Là aussi, nous proposons un estimateur d'un équivalent de l'erreur d'extrapolation tiré du Chapitre 2, avant d'en tester les performances sur données simulées. L'estimateur ainsi obtenu est alors utilisé dans la Partie 4.4 pour quantifier les limites d'extrapolation associées à un jeu de données constitué de cumuls journaliers de précipitations enregistrés à Valleraugue (Cévennes). Pour terminer, la Partie 4.5 développe plusieurs perspectives de recherche.
5. Nous clôturons le manuscrit par un résumé de nos travaux accompagné de quelques perspectives d'ordre général.

Chapitre 1

Introduction à la théorie des valeurs extrêmes univariées

Sommaire

Résumé

Ce chapitre constitue une introduction à la théorie des valeurs extrêmes dans le cadre univarié. Pour d'autres introductions à la théorie des valeurs extrêmes, nous renvoyons le lecteur à RESNICK *[1987]*, BEIRLANT *et collab. [2006]*, DE HAAN *et* FERREIRA *[2007]*, GOMES *et* GUILLOU *[2015] ou encore* COLES *et collab. [2001] pour les aspects plus statistiques. La Partie 1.1 s'intéresse dans un premier temps au comportement asymptotique du maximum d'un échantillon de variables aléatoires et aux différentes lois limites possibles, ou domaines d'attraction, avant de traiter dans un second temps du comportement asymptotique des excès au-dessus d'un seuil. La Partie 1.2 propose une caractérisation des lois associées aux différents domaines d'attraction. La théorie des fonctions à variation régulière y est résumée. Enfin, la Partie 1.3 s'attache à décrire le principe de l'estimation des quantiles extrêmes, domaine de la théorie des valeurs extrêmes qui sous-tend l'ensemble de ce manuscrit de thèse.*

1.1 Comportement asymptotique des plus grandes valeurs d'un échantillon

La plupart des approches statistiques classiques reposent sur l'étude du comportement moyen des phénomènes observés. La théorie des valeurs extrêmes s'intéresse quant à elle à l'étude des plus grandes valeurs des échantillons de données avec pour but d'en comprendre le comportement. Dans le cas univarié, c'est à dire dans le cas où l'on considère une variable aléatoire scalaire, elle s'attache à répondre aux deux questions suivantes :

(1) Quelle est la probabilité d'observer un évènement d'amplitude supérieure à une valeur x donnée?

(2) Quelle est l'amplitude de l'évènement qui est dépassée avec une faible probabilité p?

La première question s'intéresse au calcul d'une faible probabilité, proche de zéro, associée à un évènement extrême. La deuxième question est la question duale de la première. Elle cherche à quantifier la valeur, aussi appelée quantile dans le vocabulaire statistique, d'un évènement extrême, c'est à dire d'un évènement dont la probabilité de survenir est faible par définition.

Pour répondre aux précédentes questions, on a l'intuition que ce sont les valeurs les plus extrêmes du jeu de données qui contiennent l'information pertinente, et non pas les valeurs centrales, moyennes, comme dans l'approche statistique classique. C'est avec cette idée que la théorie des valeurs extrêmes s'est développée, en s'appuyant notamment sur les résultats de FISHER et TIPPETT [1928] et GNEDENKO [1943] sur la convergence en loi de la valeur maximale d'une suite de variables aléatoires indépendantes et indentiquement distribuées puis sur le résultat de PICKANDS [1975] sur la convergence en loi des excès au-dessus d'un seuil.

Ainsi, nous commençons Paragraphe 1.1.1 par nous intéresser au comportement du maximum d'un échantillon, avant d'étudier Paragraphe 1.1.2 celui des excès au-dessus d'un seuil.

1.1.1 Comportement asymptotique du maximum d'un échantillon

Plaçons-nous dans un cadre statistique et commençons par supposer que l'on dispose d'un échantillon $X_1, X_2, \ldots, X_n$ de variables aléatoires indépendantes de même loi que X, avec X une variable aléatoire continue de fonction de répartition F et de fonction de survie associée $\overline{F}$ définie par

$$\overline{F}(x) = \mathbb{P}(X > x) = 1 - F(x).$$

Nous nous intéressons dans un premier temps au comportement du maximum de l'échantillon, que l'on note $X_{n,n} := \max(X_1, ..., X_n)$. En particulier, on peut se demander quelle est la loi $F_{X_{n,n}}$ du maximum lorsque la taille de l'échantillon augmente : on parlera de comportement asymptotique.

Moyennant quelques calculs de probabilité élémentaires, et en utilisant le fait que les variables aléatoires sont indépendantes et identiquement distribuées, on peut commencer par écrire :

$$\begin{aligned} F_{X_{n,n}}(x) &:= \mathbb{P}(X_{n,n} \leq x) \\ &= \mathbb{P}(X_1 \leq x, X_2 \leq x, ..., X_n \leq x) \\ &= \prod_{i=1}^{n} \mathbb{P}(X_i \leq x) \\ &= F^n(x). \end{aligned} \tag{1.1}$$

Ainsi, la fonction de répartition du maximum est liée à celle de F. Or le comportement de cette dernière est en partie connu. En effet, si l'on définit x^* comme étant le point terminal de la loi F, c'est à dire l'extrémité droite du support,

$$x^* = \sup\{x \in \mathbb{R}, F(x) < 1\},$$

alors on sait que, par définition d'une fonction de répartition :

$$\begin{cases} 0 \leq F(x) < 1 & \text{si} \quad x < x^* \\ F(x) = 1 & \text{si} \quad x \geq x^*. \end{cases}$$

Par conséquent

$$\begin{aligned} \lim_{n\to\infty} F_{X_{n,n}}(x) &= \lim_{n\to\infty} F^n(x) \\ &= \begin{cases} 0 & \text{si} \quad x < x^* \\ 1 & \text{si} \quad x \geq x^*. \end{cases} \end{aligned} \tag{1.2}$$

L'équation (1.2) nous indique donc que la loi du maximum est dégénérée. Ce résultat est très peu informatif sur le comportement du maximum et il préférable d'obtenir une loi non-dégénérée comme limite de l'équation (1.2).

Pour ce faire, l'idée est de considérer non pas le maximum en tant que tel dans l'équation (1.1), mais une renormalisation de ce dernier. La renormalisation la plus simple à laquelle on peut penser consiste en une transformation linéaire, à l'image de celle utilisée dans le Théorème Central Limite (TCL). Nous en rappelons l'énoncé ci-dessous :

Théorème 1 (Théorème Central Limite) *Soit* $X_1, X_2, ..., X_n$ *une suite de variables aléatoires indépendantes et identiquement distribuées de fonction de répartition* F. *Supposons que l'espérance* μ *et l'écart-type* σ *de* F *existent et soient finis avec* $\sigma \neq 0$. *Considérons la somme* $S_n = X_1 + X_2 + ... + X_n$. *Alors*

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{loi} N(0,1), \quad n \to +\infty. \tag{1.3}$$

Le TCL caractérise donc le comportement de la somme linéairement renormalisée de n variables aléatoires iid, en indiquant la loi Normale comme loi limite. La question est de savoir s'il existe un équivalent du TCL non pas pour la somme, mais pour le maximum de n variables aléatoires iid. Le Théorème suivant, tiré de GNEDENKO [1943], donne une réponse positive à cette question :

Théorème 2 (Théorème de Fisher–Tippett–Gnedenko) *Soit* $X_1, X_2, ..., X_n$ *une suite de variables aléatoires indépendantes et identiquement distribuées de fonction de répartition* F. *S'il existe deux suites normalisantes réelles* $(a_n)_{n\geq 1} > 0$ *et* $(b_n)_{n\geq 1} \in \mathbb{R}$ *et une loi non-dégénérée* G_γ *telles que :*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{X_{n,n} - b_n}{a_n} \leq x\right) = \lim_{n\to\infty} F^n(a_n x + b_n) = G_\gamma(x),$$

alors

$$G_\gamma(x) = \begin{cases} \exp\left(-(1+\gamma x)_+^{-1/\gamma}\right) & si\ \gamma \neq 0 \\ \exp\left(-\exp(-x)\right) & si\ \gamma = 0, \end{cases} \tag{1.4}$$

où $\gamma \in \mathbb{R}$ *et* $z_+ = \max(0, z)$.

Ici, la suite normalisante b_n joue le rôle du paramètre de position, de la même façon que μn est l'espérance de S_n dans le TCL (cf (1.3)). La suite a_n joue quant à elle le rôle du paramètre d'échelle, à l'image de $\sigma\sqrt{n}$. Enfin, la loi Normale est remplacée ici par la fonction de répartition G_γ (appelée parfois loi par abus de langage).

Ce théorème joue un rôle aussi important dans la théorie des valeurs extrêmes que le TCL dans la théorie "classique". Il indique que le comportement asymptotique du maximum renormalisé est régi par une seule loi G_γ. Cette loi est appelée loi Généralisée des Valeurs Extrêmes (GEV). Le mot généralisé vient du fait que, contrairement au TCL, la forme de la loi limite limite G n'est pas unique. En effet, la loi GEV en question est adossée à un paramètre de forme γ qui caractérise le comportement de la queue de distribution de F. Ce paramètre γ est appelé l'indice des valeurs extrêmes. On distingue trois formes différentes pour la loi G, selon le signe de γ. On parle de domaines d'attraction :

- si $\gamma = 0$, F est dit appartenir au domaine d'attraction de GUMBEL [1958], et l'on note $F \in DA(Gumbel)$. Ce domaine d'attraction répertorie les lois les plus usuelles, telles que les lois Normale, Exponentielle, Gamma, qui sont des lois à queue de distribution légère (la fonction de survie décroit exponentiellement vite).
- si $\gamma > 0$, F appartient au domaine d'attraction dit de FRÉCHET [1927], et l'on note $F \in DA(Fréchet)$. Ce domaine d'attraction contient des lois à queue lourde telles que les lois de Pareto, Fréchet, Student, dont la décroissance de la fonction de survie est polynomiale.
- Enfin, si $\gamma < 0$, on dit que F appartient au domaine d'attraction de WEIBULL [1951], et l'on écrit $F \in DA(Weibull)$. Ce domaine d'attraction comporte uniquement des lois dont le point terminal x^* est fini.

La Figure 1.1 représente les trois différentes formes de la fonction de répartition et de la densité de G_γ, pour différentes valeurs du paramètre $\gamma \in \{-1, 0, 1\}$.

Enfin, le Tableau 1.1 propose une liste non exhaustive de lois associées aux différents domaines d'attraction (EMBRECHTS et collab. [2013]).

Exemple 1 *Nous prenons l'exemple de la loi Exponentielle standard dont la fonction de répartition est donnée, quel que soit* $x > 0$, *par* $F(x) = 1 - e^{-x}$. *Soit* $X_1, ..., X_n$ *une suite de variables aléatoires iid de fonction de*

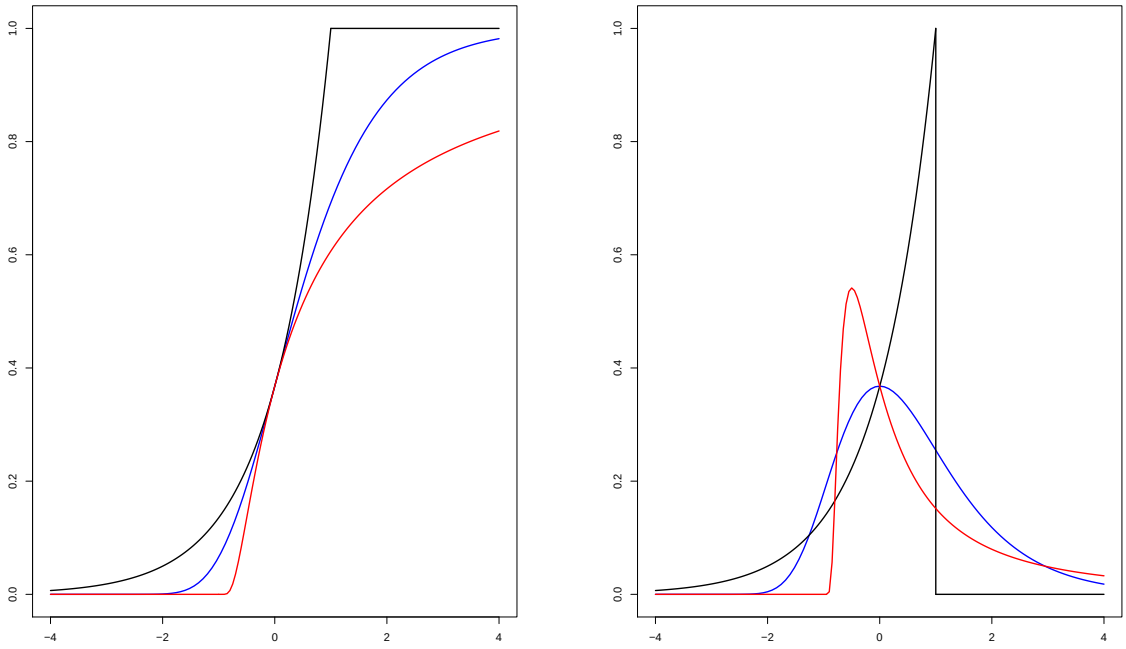

FIGURE 1.1 – A gauche : fonction de répartition G_γ. A droite : densités associées. En noir, $\gamma = -1$, en bleu : $\gamma = 0$, et en rouge : $\gamma = 1$.

Domaines d'attraction	Weibull $\gamma < 0$	Gumbel $\gamma = 0$	Fréchet $\gamma > 0$
Lois	Uniforme	Normale	Pareto
	Beta	Exponentielle	Student
		Log-normale	Burr
		Gamma	Chi-deux
		Weibull	Fréchet
		Logistique	Cauchy
		Gumbel	

TABLEAU 1.1 – Lois standards et leur domaine d'attraction.

répartition F *définie ci-dessus. Le point terminal du support de la loi étant infini, il vient* $X_{n,n} \xrightarrow{\mathbb{P}} \infty$. *Si on normalise alors* $X_{n,n}$ *en utilisant des suites adéquates, il vient :*

$$\begin{aligned}
\mathbb{P}\left(\frac{X_{n,n}-\log(n)}{1} \leq x\right) &= \mathbb{P}(X_{n,n} \leq x+\log(n)) \\
&= (F(x+\log(n))^n \\
&= (1-\exp(-x-\log(n)))^n \\
&= \left(1-\frac{\exp(-x)}{n}\right)^n \\
&\to \exp(-\exp(-x)) \\
&= G_{\gamma=0}(x)
\end{aligned}$$

quand $n \to \infty$.

Par conséquent, la loi Exponentielle est dans le domaine d'attraction de Gumbel.

1.1.2 Comportement asymptotique des excès au-dessus d'un seuil

Jusqu'à présent, nous nous sommes intéressés au comportement asymptotique du maximum de l'échantillon. Nous avons ainsi pu caractériser ses différentes lois limites. Dans la théorie des valeurs extrêmes, l'approche duale consiste à étudier le comportement des dépassements au-dessus d'un grand seuil PICKANDS [1975]. Nous souhaitons en particulier savoir s'il existe un résultat pour ces excès similaires à celui du Théorème 2.

Le cadre statistique est le même que précédemment : supposons que $X_1, ..., X_n$ constitue un échantillon de variables aléatoires iid et définissons un seuil u fixé, de telle sorte que $u < x^*$. Considérons les N_u observations $X_{i_1}, \ldots, X_{i_{N_u}}$ dépassant le seuil u. Les excès au-delà du seuil u sont alors définis par $Y_j := X_{i_j} - u$, avec $j = 1, \ldots, N_u$ (voir Figure 1.2, à gauche).

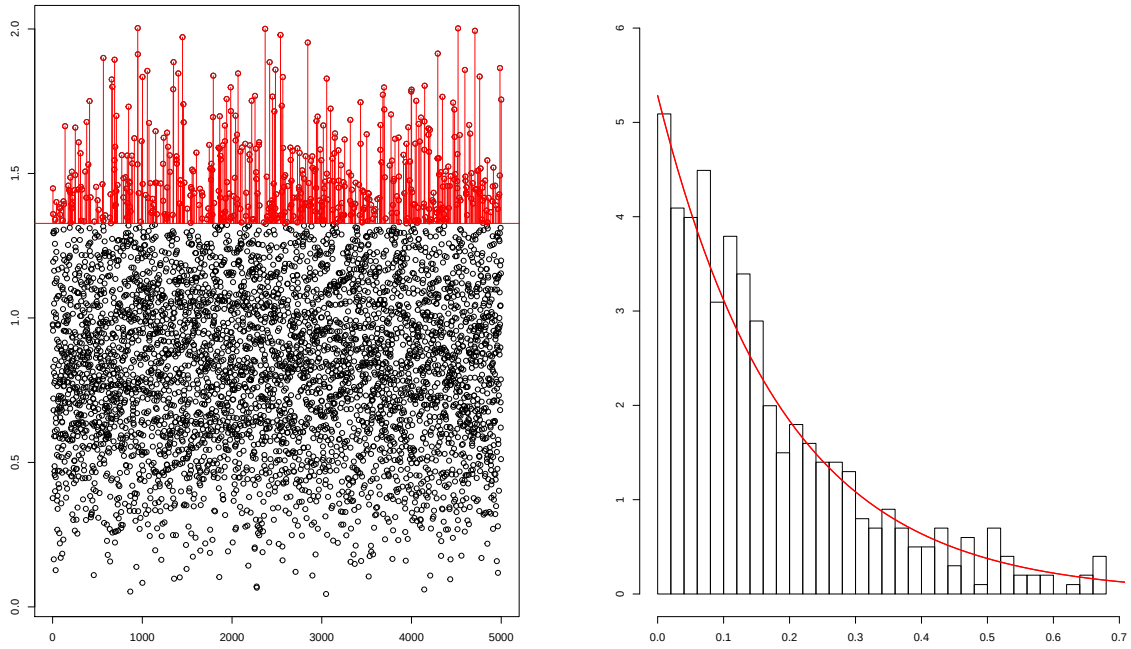

FIGURE 1.2 – A gauche : Illustration de la définition des excès. A droite : Superposition d'une densité exponentielle à l'histogramme des excès.

Le théorème qui suit établit l'existence d'une loi limite pour les excès (sous la condition d'appartenance de F à l'un des domaines d'attraction décrits ci-dessus) et en détaille la forme. De la même façon que nous avions étudié la fonction de répartition du maximum au Paragraphe 1.1.1, ce théorème se base sur la fonction de répartition de l'excès Y au delà d'un seuil u définie par $F_u(y) := \mathbb{P}(Y \leq y | X > u)$.

Remarquons d'ailleurs que la fonction de répartition des excès peut s'écrire de manière générale comme une fonction de la fonction de répartition des observations F :

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

ou de manière équivalente :

$$\begin{aligned} \overline{F}_u(y) &:= \mathbb{P}(Y > y | X > u) \\ &= 1 - F_u(y) \\ &= \frac{\overline{F}(u+y)}{\overline{F}(u)}. \end{aligned} \tag{1.5}$$

Théorème 3 (Théorème de Pickands–Balkema–de Haan) F *appartient au domaine d'attraction de* G_γ *si et seulement si il existe* $\sigma > 0$ *et* $\gamma \in \mathbb{R}$ *tels que la loi des excès* F_u *peut être uniformément approchée par une loi*

de Pareto généralisée notée $H_{\gamma,\sigma}$*, i.e.*

$$\lim_{u\to x^*} \sup_{y\in]0,x^*-u[} \left|F_u(y) - H_{\gamma,\sigma}(y)\right| = 0$$

où

$$H_{\gamma,\sigma}(y) = 1 - \left(1+\gamma\frac{y}{\sigma}\right)^{-1/\gamma} \tag{1.6}$$

définie sur $\{y : y > 0 \quad et \quad (1+\gamma y/\sigma) > 0\}$.

Dans le cas où $\gamma = 0$, cette loi correspond à une simple loi Exponentielle de paramètre $1/\sigma$:

$$H_{0,\sigma}(y) = 1 - \exp\left(-\frac{y}{\sigma}\right), \quad y \geq 0.$$

Notons aussi que $G_{-1,\sigma}(.)$ correspond à la loi Uniforme sur $[0,\sigma]$.

Ainsi, le Théorème 3 présente un potentiel équivalent au Théorème 2. Il indique que, pour une large classe de lois (les mêmes que celles du Théorème 2), la loi de Pareto généralisée H régit le comportement des excès au-dessus d'un seuil suffisamment grand.

Reprenons l'exemple de la loi Exponentielle standard :

Exemple 2 (Loi Exponentielle) *Dans le cas de la loi exponentielle,* $\overline{F}(x) = e^{-x}$ *et, au vu de l'équation (1.5), la fonction de répartition des excès s'écrit :*

$$\overline{F}_u(y) = \frac{e^{-(u+y)}}{e^{-u}} = e^{-y} = H_{0,1}(y) \quad \forall y > 0.$$

Par conséquent, les excès issus d'une loi exponentielle suivent également une loi exponentielle (résultat exact et non asymptotique dans ce cas précis).

La Figure 1.2 (droite) illustre le Théorème 3 en superposant la densité théorique d'une loi exponentielle à l'histogramme des excès issus d'une loi de Weibull.

Une preuve intuitive du Théorème 3 est basée sur le Théorème 2. Pour n assez grand :

$$F^n(u) \simeq \exp\left(-\left(1+\gamma\left(\frac{u-b_n}{a_n}\right)\right)^{-1/\gamma}\right).$$

En passant alors au logarithme des deux côtés de l'équation, il vient :

$$n\log(F(u)) \simeq -\left(1+\gamma\left(\frac{u-b_n}{a_n}\right)\right)^{-1/\gamma}.$$

Un développement limité du logarithme donne alors, pour u assez grand :

$$1 - F(u) \simeq \frac{1}{n}\left(1+\gamma\left(\frac{u-b_n}{a_n}\right)\right)^{-1/\gamma}.$$

Pareillement, pour $y > 0$ on a,

$$1 - F(u+y) \simeq \frac{1}{n}\left(1+\gamma\left(\frac{u+y-b_n}{a_n}\right)\right)^{-1/\gamma}.$$

En remplaçant finalement dans l'expression (1.5) on obtient :

$$F_u(y) \simeq 1 - \left(1+\gamma\frac{y}{\sigma_n}\right)^{-1/\gamma},$$

avec

$$\sigma_n = a_n + \gamma(u - b_n). \tag{1.7}$$

Au vu de ce qui précède, le paramètre d'échelle σ_n de la GPD est une fonction du paramètre d'échelle a_n de la GEV auquel s'ajoute une correction proportionnelle à la différence entre le seuil u et le paramètre de position b_n. Notons qu'il est exactement égal à a_n dans le cas $\gamma = 0$. Le paramètre de forme γ est quant à lui le même pour les deux approches.

Evidemment, toutes les lois ne vérifient pas les hypothèses des Théorèmes 2 et 3, à l'image des lois n'ayant pas de variance pour le Théorème Central Limite. Par la suite, nous nous posons la question de savoir s'il existe des conditions nécessaires et suffisantes simples pour caractériser l'appartenance d'une loi à un domaine d'attraction.

1.2 Caractérisation des domaines d'attraction

On commence par s'intéresser dans cette partie à la théorie des fonctions à variation lente, puis régulière. C'est cette théorie qui servira de support aux différentes caractérisations des domaines d'attraction.

1.2.1 Fonctions à variation lente

Nous commençons dans ce paragraphe par définir la notion de fonction à variation lente.

Définition 1 (BINGHAM et collab. [1987], page 6) *Soit* L *une fonction positive mesurable définie sur un voisinage* $[a,\infty[$ *de l'infini et satisfaisant, quel que soit* $\lambda > 0$ *:*

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

Alors L *est dite être à variation lente.*

Suivant cette définition, on peut vérifier que toute fonction qui admet une limite strictement positive à l'infini est une fonction à variation lente à l'infini. Par ailleurs, la fonction logarithme : $x \longmapsto \log x$, les itérations du logarithme : $x \longmapsto \log\log x$, les puissances du logarithme : $x \longmapsto |\log(x)|^\eta$ avec $\eta \in \mathbb{R}$ ou encore les fonctions de la forme $x \longmapsto \exp(\log(x)^\gamma)$ avec $0 < \gamma < 1$ sont toutes des fontions à variation lente.

Karamata propose en 1930 une caractérisation plus précise des fonctions à variation lente. Cette caractérisation est connue sous le nom de représentation de Karamata :

Théorème 4 (BINGHAM et collab. [1987], Théorème 1.3.1) *La fonction* L *est dite à variation lente à l'infini si et seulement si elle peut être représentée sous la forme suivante :*

$$\forall x \geq a > 0, \quad L(x) = c(x)\exp\left\{\int_a^x \frac{\delta(u)}{u}du\right\},$$

avec c *et* δ *deux fonctions mesurables telles que :*

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

De plus, si la fonction c est constante, alors la fonction L est dérivable de dérivée L' avec, pour tout $x > 0$,

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

En particulier on a :

$$\lim_{x\to\infty} \frac{xL'(x)}{L(x)} = \lim_{x\to\infty} \delta(x) = 0.$$

La fonction L est alors dite normalisée.

Nous terminons ce paragraphe en donnant quelques propriétés fondamentales des fonctions à variation lente :

Proposition 1 (BINGHAM et collab. [1987], Proposition 1.3.6)

1. *Si* L *est une fonction à variation lente à l'infini, alors :*
$$\lim_{x\to\infty} \frac{\log L(x)}{\log(x)} = 0.$$
2. *Si* L *est à variation lente à l'infini et* $\alpha > 0$*, alors :*
$$\lim_{x\to+\infty} x^\alpha L(x) = +\infty \quad et \quad \lim_{x\to+\infty} x^{-\alpha} L(x) = 0.$$
3. *Si* L *est à variation lente à l'infini, alors, pour tout* $\alpha \in \mathbb{R}$ *:*
$$L^\alpha : x \longmapsto [L(x)]^\alpha$$
est une fonction à variation lente à l'infini.
4. *Si* L_1 *et* L_2 *sont à variation lente à l'infini, alors*
$$L_1 + L_2 : x \longmapsto L_1(x) + L_2(x)$$
et
$$L_1 \times L_2 : x \longmapsto L_1(x) \times L_2(x)$$
sont à variation lente à l'infini. Si, de plus, $\lim_{x\to+\infty} L_2(x) = +\infty$*, alors*
$$L_1 \circ L_2 : x \longmapsto L_1[L_2(x)]$$
est aussi à variation lente à l'infini.

1.2.2 Fonctions à variation régulière

La notion de fonction à variation régulière est une extension de celle de fonction à variation lente. Grossièrement parlant, elle caractérise l'ensemble des fonctions qui se comportent comme une fonction puissance à l'infini.

Définition 2 (BINGHAM et collab. [1987], page 18) *Une fonction positive mesurable* $f :]a, +\infty[\to \mathbb{R}_+$ $(a \geq 0)$ *est dite à variation régulière à l'infini d'indice* $\rho \in \mathbb{R}$ *- et on note* $f \in \mathrm{RV}_\rho$ *- si, pour tout* $\lambda > 0$ *:*

$$\lim_{x\to\infty} \frac{f(\lambda x)}{f(x)} = \lambda^\rho. \tag{1.8}$$

Notons que si $\rho = 0$, alors f est une fonction à variation lente à l'infini et possède donc les propriétés décrites Paragraphe 1.2.1.

Il est possible d'élargir cette définition aux fonctions à variation régulière à l'origine (au voisinage de zéro).

Définition 3 *Une fonction positive mesurable* $f :]0, a[\to \mathbb{R}_+$ $(a \geq 0)$ *est dite à variation régulière à l'origine d'indice* $\rho \in \mathbb{R}$ *- et on note* $f \in \mathrm{RV}_\rho^0$ *- si, pour tout* $\lambda > 0$ *:*

$$\lim_{x\to 0^+} \frac{f(\lambda x)}{f(x)} = \lambda^\rho.$$

Remarquons que $f \in \mathrm{RV}_\rho^0$ est équivalent à dire que la fonction $x \longmapsto f(1/x)$ est à variation régulière à l'infini d'indice $-\rho$. Par la suite, nous considérons qu'une fonction est à variation régulière à l'infini si rien n'est précisé.

Il est important de noter que toute fonction à variations régulière d'indice $\rho \in \mathbb{R}$ peut s'écrire sous la forme d'un produit d'une fonction lente et d'une fonction puissance. C'est l'objet du Théorème qui suit :

Théorème 5 (BINGHAM et collab. [1987], Théorème 1.4.1) *Soient* $f :]a, +\infty[\to \mathbb{R}_+$ $(a \geq 0)$ *une fonction positive mesurable et* $\rho \in \mathbb{R}$*. Les assertions suivantes sont équivalentes :*

1. $f \in \mathrm{RV}_\rho$
2. *Il existe* $L \in \mathrm{RV}_0$ *telle que :*
$$f(x) = x^\rho L(x). \tag{1.9}$$

En utilisant la Définition 2 ou le Théorème 5, il est possible de montrer que pour $\rho \in \mathbb{R}$ et $\gamma \in \mathbb{R}$, les fonctions : $x \longmapsto x^\rho$, $x \longmapsto x^\rho \log(1+x)$, $x \longmapsto [x\log(1+x)]^\rho$, $x \longmapsto x^\rho (\log x)^\gamma$ et $x \longmapsto x^\rho (\log x \log x)^\gamma$ sont à variation régulière à l'infini d'indice ρ.

Le reste de ce paragraphe est dédié aux propriétés des fonctions à variation régulière qui nous sont utiles tout au long de la thèse :

Proposition 2 (BINGHAM et collab. [1987], Proposition 1.5.7) *Soient* ρ, ρ_1 *et* ρ_2 *des réels.*

1. *Si* f *est une fonction à variation régulière à l'infini d'indice* $\rho \in \mathbb{R}$*, alors :*
$$\lim_{x\to+\infty} f(x) = \begin{cases} 0 & si \quad \rho < 0 \\ +\infty & si \quad \rho > 0. \end{cases}$$
2. *Si* $f \in \mathrm{RV}_\rho$*, alors :*
$$\lim_{x\to\infty} \frac{\log f(x)}{\log(x)} = \rho.$$
3. *Si* $f \in \mathrm{RV}_\rho$ *et* $\alpha \in \mathbb{R}$*, alors :*
$$f^\alpha : x \longmapsto [f(x)]^\alpha \in \mathrm{RV}_{\alpha\rho}.$$
4. *Si* $f_1 \in \mathrm{RV}_{\rho_1}$ *et* $f_2 \in \mathrm{RV}_{\rho_2}$*, alors*
$$f_1 + f_2 : x \longmapsto f_1(x) + f_2(x) \in \mathrm{RV}_{\max(\rho_1,\rho_2)}$$
et si, de plus, $\lim_{x\to+\infty} f_2(x) = +\infty$*, alors*
$$f_1 \circ f_2 : x \longmapsto f_1[f_2(x)] \in \mathrm{RV}_{\rho_1\rho_2}.$$

Le deuxième résultat que nous exposons concerne les propritétés de l'inverse généralisé d'une fonction à variation régulière. Rappelons que l'inverse généralisé d'une fonction f croissante est définie par

$$f^{\leftarrow}(x) = \inf\{y, f(y) > x\}. \tag{1.10}$$

Notons que l'inverse généralisé correspond à l'inverse classique f^{-1} lorsque la fonction considérée est continue et strictement croissante.

Proposition 3 *Supposons que f est une fonction à variation régulière d'indice $\rho \in \mathbb{R}$.*

1. *Si $\rho > 0$, alors $f^{\leftarrow}(\cdot)$ est à variation régulière d'indice $1/\rho$.*
2. *Si $\rho < 0$, alors $f^{\leftarrow}(1/\cdot)$ est à variation régulière d'indice $-1/\rho$.*

Il est possible de trouver une preuve dans BINGHAM et collab. [1987], Théorème 1.5.12.

Les deux prochains résultats concernent respectivement la primitive et la dérivée d'une fonction à variation régulière.

Théorème 6 (BINGHAM et collab. [1987], Théorème 1.5.11) *Soient f une fonction définie sur le voisinage $]a, +\infty[$ $(a \geq 0)$ de l'infini et $\rho \in \mathbb{R}$.*

1. *Si $f \in \mathrm{RV}_\rho$ et $\rho \geq -1$, alors il existe $b > a$ tel que, quel que soit $x \geq b$:*

$$x \longmapsto \int_b^x f(t)\,dt \in \mathrm{RV}_{\rho+1}.$$

2. *Si $f \in \mathrm{RV}_\rho$ et $\rho < -1$, alors, quel que soit $x > a$:*

$$x \longmapsto \int_x^{+\infty} f(t)\,dt \in \mathrm{RV}_{\rho+1}.$$

Le Théorème 6 indique que, si f est une fonction à variation régulière d'indice ρ, alors une primitive de f est à variation régulière d'indice $\rho + 1$. Le résultat suivant est connu sous le nom de Théorème de la densité monotone.

Théorème 7 (BINGHAM et collab. [1987], Théorème 1.7.2) *Soit $\mathrm{F}(x) = \int_0^x f(t)\,dt$. Si F est à variation régulière d'indice ρ, et si f est monotone à l'infini, alors*

$$|f| \in \mathrm{RV}_{\rho-1}.$$

Si de plus, $\mathrm{F}(x) \sim cx^\rho \mathrm{L}(x)$ $(x \to \infty)$, où $c > 0$, $\rho \neq 0$, $\mathrm{L} \in \mathrm{RV}_0$, alors

$$f(x) \sim c\rho x^{\rho-1}\mathrm{L}(x) \quad x \to \infty.$$

Le Théorème 7 constitue le pendant du Théorème 6. Il indique que la dérivée d'une fonction à variation régulière d'indice $\rho \neq 0$ est également une fonction à variation régulière, mais d'indice $\rho - 1$. Notons que contrairement au Théorème 6, le Théorème 7 nécessite une hypothèse de monotonie à l'infini de la fonction f. De plus, il n'est plus vrai dans le cas $\rho = 0$. Sa mise en pratique demande par conséquent de prendre quelques précautions.

Ayant caractérisé la dérivée et la primitive d'une fonction à variation régulière, nous terminons par exposer deux résultats : le premier est un corollaire du Théorème 7. Il concerne la limite de $xf'(x)/f(x)$ avec $f \in \mathrm{RV}_\rho$. Le deuxième résultat est connu sous le nom de bornes de Potter. Comme l'indique son nom, il propose des bornes sur les fonctions à variation régulière.

Corollaire 1 *Soit $\mathrm{F}(x) = \int_0^x f(t)\,dt$ (de dérivée f). Si $\mathrm{F} \in \mathrm{RV}_\rho$, $\rho \in \mathbb{R}$, et si f est monotone à l'infini, alors*

$$\lim_{x \to +\infty} \frac{xf(x)}{\mathrm{F}(x)} = \rho.$$

Ce résultat est très utile en pratique si la densité d'une fonction de répartition F existe.

Théorème 8 (BINGHAM et collab. [1987], Théorème 1.5.6) *Supposons que $f \in \mathrm{RV}_\rho$ et soient $\delta_1 > 0$, $\delta_2 > 0$. Alors, il existe $t_0 \in \mathbb{R}$ tel que, quel que soit $t \geq t_0$, $tx \geq t_0$,*

$$(1-\delta_1)\lambda^\rho \min(\lambda^{\delta_2}, \lambda^{-\delta_2}) < \frac{f(\lambda x)}{f(x)} < (1+\delta_1)\lambda^\rho \max(\lambda^{\delta_2}, \lambda^{-\delta_2}).$$

Le Théorème 8 peut être utilisé pour prouver que les fonctions à variation régulière conservent les équivalences :

Proposition 4 *Soit $f \in \mathrm{RV}_\rho$, $\rho \in \mathbb{R}$, $u_n \to \infty$ et $v_n \overset{n\to\infty}{\sim} u_n$. Alors*

$$f(v_n) \overset{n\to\infty}{\sim} f(u_n).$$

1.2.3 Fonctions à variation régulière étendue

Il est utile d'étendre la notion de fonction à variation régulière. En particulier, la notion de fonctions à variation régulière étendue s'avère particulièrement utile pour démontrer par exemple des résultats de consistance d'estimateurs de l'indice des valeurs extrêmes γ.

Définition 4 (DE HAAN et FERREIRA [2007], Définition B.2.3) *Une fonction mesurable f est dite être à variation régulière étendue d'indice γ - et on note $f \in \mathrm{ERV}_\gamma$ - s'il existe une fonction positive a (appelée la fonction auxiliaire) et un réel γ tels que, pour tout $x > 0$,*

$$\lim_{t\to\infty} \frac{f(tx)-f(t)}{a(t)} = \mathrm{L}_\gamma(x) := \int_1^x u^{\gamma-1}\,du, \tag{1.11}$$

$$\mathrm{L}_\gamma(x) = \begin{cases} \dfrac{x^\gamma - 1}{\gamma} & \text{si } \gamma \neq 0, \\ \log x & \text{si } \gamma = 0. \end{cases}$$

De plus, (1.11) est vérifiée avec a une fonction mesurable à variation régulière d'indice γ.

Une fonction à variation régulière étendue est donc une fonction dont la limite

$$\lim_{t\to\infty} \frac{f(tx)-f(t)}{a(t)}$$

existe et n'est pas constante. Si cette dernière existe, elle prend la forme donnée par (1.11).

Remarque 1 *Un choix de fonction auxiliaire est $a(t) = tf'(t)$ si f est dérivable.*

On retrouve dans la théorie des valeurs extrêmes ce genre de limite sous le nom de condition du premier ordre (cf 14). Il existe également des conditions du second ordre, permettant de contrôler la vitesse de convergence dans (1.11). Ces dernières conditions font appel à la notion de fonction à variation régulière étendue du second ordre.

Définition 5 (DE HAAN et FERREIRA [2007], Définition B.3.4) *Une fonction mesurable f est dite être à variation régulière étendue d'ordre deux s'il existe des fonctions positives a et A avec $\lim_{t\to\infty} \mathrm{A}(t) = 0$ telles que, quel que soit $x > 0$,*

$$\mathrm{H}(x) := \lim_{t\to\infty} \frac{\frac{f(tx)-f(t)}{a(t)} - \frac{x^\gamma-1}{\gamma}}{\mathrm{A}(t)} \tag{1.12}$$

existe avec H une fonction qui n'est pas un multiple de $(x^\gamma-1)/\gamma$.

Les fonctions a et A sont respectivement appelées fonctions auxiliaires du premier et du deuxième ordre. Nous caractérisons la forme de la fonction H ci-dessous.

Théorème 9 (DE HAAN et FERREIRA [2007], Théorème B.3.1 et Remarque B.3.5) *Supposons qu'il existe une fonction mesurable f et des fonctions positives $\tilde{a}$ et $\tilde{\mathrm{A}}$ telles que la limite (1.12) existe pour tout $x > 0$ et n'est pas un multiple de $(x^\gamma-1)/\gamma$. Alors il existe des fonctions positives a et A et un réel $\rho \leq 0$ tels que, pour tout $x > 0$,*

$$\lim_{t\to\infty} \frac{\frac{f(tx)-f(t)}{a(t)} - \frac{x^\gamma-1}{\gamma}}{\mathrm{A}(t)} = \int_1^x s^{\gamma-1} \int_1^s u^{\rho-1}\,du\,ds. \tag{1.13}$$

Le paramètre ρ est appelé paramètre du second ordre. L'estimation du paramètre ρ et de la fonction auxiliaire A est l'objet de GOMES et collab. [2002], BEIRLANT et collab. [2002] ou encore E. DEME [2013].

Remarque 2

1. *La fonction A, qui décrit la vitesse de convergence dans (1.12) est à variation régulière d'indice $\rho \leq 0$. Par conséquent :*
   - *Si $\rho < 0$, la convergence est rapide, associée à une vitesse polynomiale ;*
   - *Si $\rho = 0$, la convergence est bien plus lente, associée à une vitesse logarithmique.*

2. *La forme générale de* H *est donnée par (1.13). Il est possible d'en détailler la forme dans certains cas. Dans le cas où* $\gamma \neq 0$, $\rho \neq 0$ *et* $\gamma + \rho \neq 0$, H *s'écrit :*

$$H_{\gamma,\rho}(x) := \frac{1}{\rho}\left(\frac{x^{\gamma+\rho}-1}{\gamma+\rho} - \frac{x^{\gamma}-1}{\gamma}\right).$$

Dans les autres cas :

$$\begin{cases} \frac{1}{\rho}\left(\frac{x^{-\rho}-1}{\rho} + \log x\right), & \rho = -\gamma \neq 0 \\ \frac{1}{\gamma}\left(x^{\gamma}\log x - \frac{x^{\gamma}-1}{\gamma}\right), & \rho = 0 \neq \gamma \\ \frac{1}{\rho}\left(\frac{x^{\rho}-1}{\rho} - \log x\right), & \rho \neq 0 = \gamma \\ \frac{1}{2}(\log x)^2, & \rho = 0 = \gamma. \end{cases}$$

La théorie des fonctions à variation régulière ayant été introduite, nous présentons maintenant des résultats permettant de caractériser le domaine d'attraction d'une loi F et ses constantes de normalisation.

1.2.4 Domaine d'attraction de Fréchet

Le premier résultat détaille plus précisément la décroissance polynomiale d'une fonction de survie appartenant au domaine d'attraction de Fréchet.

Théorème 10 (DE HAAN et FERREIRA [2007], Théorème 1.2.1) *Une fonction de répartition* F *appartient au domaine d'attraction de Fréchet si et seulement si* $x^* = \infty$ *et, quel que soit* $x > 0$ *:*

$$\lim_{t\to\infty} \frac{\overline{F}(tx)}{\overline{F}(t)} = x^{-1/\gamma}, \quad \gamma > 0.$$

Autrement dit, au vu de la Définition 2, F $\in$DA(Fréchet) si et seulement si la fonction de survie $\overline{F}$ est à variation régulière d'indice $-1/\gamma$: $\overline{F} \in RV_{-1/\gamma}$. En utilisant la caractérisation donnée par le Théorème 5, F $\in$DA(Fréchet) est finalement équivalent à

$$F(x) = 1 - x^{-1/\gamma}L(x) \quad \text{avec} \quad L \in RV_0. \tag{1.14}$$

La preuve des résultats précédents peut-être trouvée dans DE HAAN et FERREIRA [2007], page 25.

La caractérisation des fonctions du domaine d'attraction de Fréchet est donc particulièrement simple. Ce dernier domaine contient des lois très similaires entre elles et c'est pourquoi cela constitue un domaine relativement petit comparé au domaine d'attraction de Gumbel par exemple (nous y revenons ci-après). Nous citons ci-dessous quelques exemples de lois appartenant au domaine d'attraction de Fréchet (cf Tableau 1.1).

Exemple 3

— *La loi de Pareto, dont la fonction de survie est donnée par* $\overline{F}(x) = x^{-\alpha}$ *(*$x > 0$, $\alpha > 0$*) appartient au domaine d'attraction de Fréchet avec un indice* $\gamma = 1/\alpha$.

— *La loi de Fréchet définie par* $\overline{F}(x) = 1 - \exp(-x^{-\alpha}) \overset{x\to\infty}{\sim} x^{-\alpha}$ *(*$x > 0$, $\alpha > 0$*) appartient également au domaine de Fréchet avec un indice* $\gamma = 1/\alpha$.

Il est intéressant en pratique de reformuler le Théorème 10 à travers la fonction quantile $U(\cdot) := F(1 - 1/\cdot) = \overline{F}(1/\cdot) = q(1/\cdot)$:

Corollaire 2 (DE HAAN et FERREIRA [2007], Corollaire 1.2.10) *Une fonction de répartition* F *appartient au domaine d'attraction de Fréchet si et seulement si, pour* $\gamma > 0$ *:* $x^* = \infty$ *et, quel que soit* $x > 0$ *:*

$$\lim_{t\to\infty} \frac{U(tx)}{U(t)} = x^{\gamma}. \tag{1.15}$$

La caractérisation (1.15) fait aussi intervenir la notion de fonction à variation régulière définie Paragraphe 1.2.2. Elle nous indique que F $\in$DA(Fréchet) si et seulement si U est à variation régulière d'indice γ.

Pour terminer, nous donnons une caractérisation des suites normalisantes (a_n) et (b_n) lorsque F $\in$DA(Fréchet) :

Proposition 5 (DE HAAN et FERREIRA [2007], Corollaire 1.2.4) *Si* F *est dans le domaine d'attraction de Fréchet, alors un choix possible de suites normalisantes est :*

$$\begin{aligned} a_n &:= \overline{\mathrm{F}}^{\leftarrow}(1/n) = \mathrm{U}(n), \\ b_n &:= 0. \end{aligned}$$

Une preuve de cette caractérisation peut être trouvée dans DE HAAN et FERREIRA [2007], page 27.

Pour conclure, le domaine d'attraction de Fréchet admet uniquement des lois à queue lourde, c'est à dire dont la décroissance de la fonction de survie est polynomiale, faisant de lui un domaine très homogène. Il est particulièrement utilisé dans les applications : on citera la météorologie, GARDES et GIRARD [2010] et METHNI et collab. [2014] ou encore l'hydrologie, ANDERSON et MEERSCHAERT [1998] et EL METHNI et collab. [2012].

1.2.5 Domaine d'attraction de Weibull

Le premier résultat donne une caractérisation de la fonction de survie d'une loi appartenant au domaine d'attraction de Weibull :

Théorème 11 (DE HAAN et FERREIRA [2007], Théorème 1.2.1) *Une fonction de répartition* F *appartient au domaine d'attraction de Weibull si et seulement si x^* est fini et, quel que soit $x > 0$,*

$$\lim_{t \downarrow 0} \frac{1 - \mathrm{F}(x^* - tx)}{1 - \mathrm{F}(x^* - t)} = x^{-1/\gamma}, \quad \gamma < 0.$$

En fait, ce résultat est très similaire à la caractérisation des lois du domaine d'attraction de Fréchet. GNEDENKO [1943] montre qu'il existe un lien entre ces deux caractérisations.

Théorème 12 *Une fonction de répartition* F *appartient au domaine d'attraction de Weibull (avec un indice des valeurs extrêmes $\gamma < 0$) si et seulement si x^* est fini et si la fonction de répartition* F_* *définie par :*

$$\mathrm{F}_*(x) = \mathrm{F}(x^* - 1/x), \quad x > 0$$

appartient au domaine d'attraction de Fréchet avec un indice des valeurs extrêmes $-\gamma > 0$, c'est à dire

$$\forall x > 0, \quad \lim_{t \to \infty} \frac{\overline{\mathrm{F}}(x^* - 1/(tx))}{\overline{\mathrm{F}}(x^* - 1/t)} = \lim_{t \to \infty} \frac{\overline{\mathrm{F}}_*(tx)}{\overline{\mathrm{F}}_*(t)} = x^{+1/\gamma}.$$

A partir du Théorème 12, il est possible de retrouver la caractérisation du Théorème 11 en utilisant la Définition 3. Finalement, au vu de la Définition 5, une fonction de répartition F du domaine d'attraction de Weibull s'écrit, pour $x \leq x^*$:

$$\mathrm{F}(x) = 1 - (x^* - x)^{-1/\gamma} \mathrm{L}((x^* - x)^{-1}), \tag{1.16}$$

avec L une fonction à variation lente. Une fonction de survie $\overline{\mathrm{F}}$ appartenant au domaine d'attraction de Weibull est donc une fonction qui décroît de manière polynomiale au voisinage d'un point terminal fini. Le résultat suivant caractérise les suites normalisantes (a_n) et (b_n) :

Proposition 6 (DE HAAN et FERREIRA [2007], Corollaire 1.2.4) *Si* F *est dans le domaine d'attraction de Weibull, alors un choix possible de suites normalisantes est :*

$$\begin{aligned} a_n &:= x^* - \overline{\mathrm{F}}^{\leftarrow}(1/n) = x^* - \mathrm{U}(n), \\ b_n &:= x^*. \end{aligned}$$

Les applications dédiées à ce domaine d'attraction sont variées : citons EINMAHL et MAGNUS [2008] en sport ou encore AARSSEN et DE HAAN [1994] pour l'étude de l'espérance de vie maximale. L'estimation du point terminal des lois appartenant au domaine d'attraction de Weibull est l'objet de FALK [1995], GIRARD et collab. [2012] ou encore HALL [1982].

1.2.6 Domaine d'attraction de Gumbel

Le domaine d'attraction de Gumbel correspond au cas $\gamma = 0$ dans le Théorème 2. C'est un des domaines les plus complexes à étudier dans le sens où il comporte une très grande variété de lois. Parmi elles, on peut citer des lois ayant un point terminal fini, mais aussi des lois dont la fonction de survie décroît exponentiellement et enfin des lois à queue plus lourde.

Ainsi, a contrario des deux autres domaines d'attraction, le domaine d'attraction de Gumbel n'a pas de caractérisation simple pour décrire toutes ses lois, comme le montre la condition nécessaire et suffisante suivante.

Théorème 13 (DE HAAN et FERREIRA [2007], Théorème 1.2.1) *La fonction de répartition* F *appartient au domaine d'attraction de Gumbel si et seulement si il existe une fonction* g *positive telle que, quel que soit* $x \in \mathbb{R}$,

$$\lim_{t \to x^*} \frac{1 - F(t + xg(t))}{1 - F(t)} = e^{-x} \tag{1.17}$$

avec x^* *qui peut être fini ou infini. Si (1.17) est vérifié, alors* $\int_t^{x^*} (1 - F(s))\,ds < \infty$ *pour* $t < x^*$ *et (1.17) est aussi vérifié avec*

$$g(t) := \frac{\int_t^{x^*} (1 - F(s))\,ds}{1 - F(t)}.$$

Il est difficile de tirer profit de cette caractérisation en pratique, qui est le reflet de la complexité du domaine d'attraction de Gumbel. Cependant, il existe dans la littérature des conditions nécessaires ainsi que des conditions suffisantes simples. Ces conditions sont la plupart données par rapport à la fonction quantile U et restent applicables à l'ensemble du domaine d'attraction de Gumbel. Nous en présentons une de chaque ci-dessous :

Proposition 7 (DE HAAN et FERREIRA [2007], Lemme 1.2.9 et Corollaire 1.1.10)

— *Si* F ∈ *DA(Gumbel), alors :*

$$\lim_{t \to \infty} \frac{U(tx)}{U(t)} = 1. \tag{1.18}$$

— *Supposons* F *dérivable. Si*

$$\lim_{t \to \infty} \frac{U'(tx)}{U'(t)} = x^{-1}, \tag{1.19}$$

alors F ∈ *DA(Gumbel).*

Au vu de l'introduction à la théorie des fonctions à variation régulière faite Paragraphe 1.2.2, ces conditions sont facilement interprétables. La première indique que, si F ∈DA(Gumbel), alors U est une fonction à variation lente. Réciproquement, si U′ est à variation régulière d'indice $\rho = -1$, alors F ∈DA(Gumbel). Moyennant l'existence d'une dérivée de F, ces conditions sont nettement plus simples à mettre en oeuvre en pratique que le Théorème 13.

La proposition suivante nous renseigne quant aux constantes de normalisation d'une loi appartenant au domaine d'attraction de Gumbel.

Proposition 8 (DE HAAN et FERREIRA [2007], Corollaire 1.2.4) *Si* F *est dans le domaine d'attraction de Gumbel, alors un choix possible de suites normalisantes est :*

$$\begin{aligned} a_n &:= g(U(n)), \\ b_n &:= U(n), \end{aligned} \tag{1.20}$$

avec g *définie comme dans le Théorème 13.*

Comme nous pouvons le constater ci-dessus, il n'est pas possible d'obtenir une caractérisation simple de l'ensemble du domaine d'attraction de Gumbel. Dans la littérature cependant, plusieurs caractérisations simples sont proposées pour des sous-familles de lois issues de ce dernier. Nous décrivons trois de ces familles ci-dessous, ainsi que leur caractérisation :

Lois à queue de type Weibull Les lois à queue de type Weibull sont des lois appartenant au domaine d'attraction de Gumbel et dont la fonction de survie décroît exponentiellement vers zéro. Cette famille regroupe une grande variété de lois usuelles, dont les lois Exponentielle, Normale, Gamma, Weibull et Logistique (cf GALAMBOS [1977]). Nous en donnons une définition ci-dessous :

Définition 6 *Une fonction de répartition* F *est dite à queue de type Weibull s'il existe* $\beta > 0$ *tel que pour tout* $x > 0$, *sa fonction de survie associée* $\overline{F}$ *vérifie :*

$$\lim_{t \to \infty} \frac{-\log \overline{F}(tx)}{-\log \overline{F}(t)} = x^{1/\beta}.$$

Autrement dit, une loi à queue de type Weibull vérifie $\log(1/\overline{F}) \in RV_{1/\beta}$. Une caractérisation de la fonction de survie d'une loi à queue de type Weibull est ainsi donnée par :

$$\overline{F}(x) = \exp\left(-x^{1/\beta} \ell(x)\right) \quad \text{avec} \quad \ell \in RV_0. \tag{1.21}$$

Contrairement aux approches précédentes, dans lesquelles l'indice des valeurs extrêmes γ gère la forme de la queue de distribution, ici $\gamma = 0$ et c'est $\beta > 0$, appelé coefficient de queue de type Weibull, qui régit la décroissance de la queue. Une valeur du paramètre de forme β proche de zéro correspond à une décroissance rapide de la queue de distribution. Réciproquement une valeur de β grande correspond à une décroissance lente de la queue de distribution.

L'estimation de l'indice de queue est l'objet de BERRED [1991], BRONIATOWSKI [1993], BEIRLANT et collab. [1995], GIRARD [2004], GARDES et GIRARD [2005], DIEBOLT et collab. [2008a] et GOEGEBEUR et collab. [2010]. Une comparaison de ces différents estimateurs est proposée dans GARDES et GIRARD [2006] et une synthèse sur les lois à queue de type Weibull est proposée dans GARDES et GIRARD [2013]. Les lois à queue de type Weibull sont par exemple utilisées en hydrologie EL METHNI et collab. [2012] ou en assurance BEIRLANT et collab. [1995]; BEIRLANT et TEUGELS [1992].

Lois à queue de type log-Weibull Une variable aléatoire X suit une loi à queue de type log-Weibull si log(X) suit une loi à queue de type Weibull. Nous donnons ci-dessous une caractérisation des fonctions de survie des lois à queue de type log-Weibull :

Définition 7 *Une fonction de répartition* F *est dite à queue de type Log-Weibull s'il existe* $\beta > 0$ *tel que pour tout* $x > 0$, *sa fonction de survie associée* $\overline{F}$ *vérifie :*

$$\lim_{t\to\infty} \frac{-\log\overline{F}(e^{tx})}{-\log\overline{F}(e^{t})} = x^{1/\beta}.$$

Une loi à queue de type Log-Weibull vérifie donc $\log(1/\overline{F}(\exp(\cdot))) \in \mathrm{RV}_{1/\beta}$. Autrement dit,

$$\overline{F}(x) = \exp\left(-(\log x)^{1/\beta}\ell(\log x)\right) \quad \text{avec} \quad \ell \in \mathrm{RV}_0. \tag{1.22}$$

La décroissance de la fonction de survie des lois à queue de type Log-weibull est ainsi plus lente que celle des lois à queue de type Weibull, la plus célèbre loi de ce type étant la loi Lognormale.

Une estimation du coefficient à queue de type log-Weibull est proposée dans GARDES et collab. [2011].

Lois à point terminal fini Le domaine d'attraction de Gumbel contient également des lois dont le point terminal est fini. Ces lois sont cependant différentes des lois du domaine d'attraction de Weibull dans la manière dont elle converge vers le point terminal. Nous en proposons un exemple ci-dessous :

Exemple 4 *Un exemple de loi à point terminal fini du domaine d'attraction de Gumbel est donné par :*

$$\overline{F}(x) = \exp\left(-\left(\frac{x^*}{x^* - x}\right)^{1/\beta} \mathrm{L}(x)\right),$$

avec L *une fonction à variation lente,* $x < x^*$ *et* $\beta > 0$.

Ainsi, contrairement aux lois du domaine d'attraction de Weibull, les fonctions de survie à point terminal fini du domaine d'attraction de Gumbel présentent une décroissance exponentielle au voisinage de leur point terminal.

En résumé, le domaine d'attraction de Gumbel est très vaste. Cette diversité - et le fait qu'il contient la plupart des lois usuelles - fait de lui un domaine privilégié dans les applications. En particulier, l'application historique du domaine d'attraction de Gumbel est l'hydrologie, citons GUMBEL [1941, 1954, 1958] et DE HAAN [1990].

1.2.7 Une caractérisation générale des domaines d'attraction

Nous terminons cette partie en proposant une deuxième caractérisation générale des domaines d'attraction, complémentaire de celle proposée au Théorème 2, mais cette fois-ci en terme de la fonction quantile de queue U. Cette caractérisation se base sur la notion précédemment définie de fonction à variation régulière étendue (cf Paragraphe 1.2.3) :

Théorème 14 (DE HAAN et FERREIRA [2007], Théorème 1.1.6) *Pour* $\gamma \in \mathbb{R}$, *les assertions suivantes sont équivalentes :*

1. *Il existe deux suites normalisantes réelles* $(a_n)_{n\geq 1} > 0$ *et* $(b_n)_{n\geq 1} \in \mathbb{R}$ *telles que :*

$$\lim_{n\to\infty} F^n(a_n x + b_n) = G_\gamma(x) = \exp\left(-(1+\gamma x)_+^{-1/\gamma}\right).$$

2. *Il existe une fonction positive a telle que, quel que soit $x > 0$,*

$$\lim_{t\to\infty} \frac{U(tx) - U(t)}{a(t)} = L_\gamma(x). \tag{1.23}$$

Autrement dit, F est dans le domaine d'attraction de G_γ, $\gamma \in \mathbb{R}$ si et seulement si sa fonction quantile de queue associée est à variation régulière étendue de paramètre γ.

Ayant listé les différentes caractérisations des domaines d'attraction, nous sommes prêts à aborder la notion d'estimation de quantiles extrêmes. C'est l'objet de la partie qui suit.

1.3 Estimation de quantiles extrêmes

Après avoir défini la notion de quantile, nous décrivons les différentes stratégies proposées dans la littérature pour aborder le problème de l'estimation des quantiles extrêmes. Ces dernières se basent sur trois approches différentes, que nous décrivons à tour de rôle.

1.3.1 Définition

Un quantile est une valeur du support de la loi qui est dépassée avec une probabilité p :

Définition 8 *Le quantile $q(p)$ d'ordre $1-p$ associé à F est défini par :*

$$q(p) := \overline{F}^{\leftarrow}(p) = \inf\{x : \overline{F}(x) \leq p\} \quad avec \quad p \in]0,1[, \tag{1.24}$$

où $\overline{F}^{\leftarrow}$ représente l'inverse généralisé de $\overline{F}$ (cf équation (1.10)).

Soit $X_1,...,X_n$ une suite de variables aléatoires iid de fonction de répartition F et $X_{1,n},...,X_{n,n}$ leurs statistiques d'ordre associées. Définissons la fonction de répartition empirique F_n :

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \leq x\}, \tag{1.25}$$

où $\sum_{i=1}^{n} \mathbb{1}\{X_i \leq x\}$ représente le nombre d'éléments inférieurs ou égaux à x dans l'échantillon. Une manière d'estimer $q(p)$ est d'inverser la fonction de répartition empirique :

$$\hat{q}(p) = F_n^{\leftarrow}(1-p) = \inf\{t \in \mathbb{R} : F_n(t) \geq 1-p\}.$$

Moyennant la définition de la fonction de répartition empirique, l'estimateur ci-dessus correspond aussi à la $\lfloor np \rfloor$ ième plus grande observation de l'échantillon. Une autre manière équivalente de définir l'estimateur du quantile est alors :

$$\hat{q}(p) = X_{n-\lfloor np \rfloor, n}, \tag{1.26}$$

où $\lfloor . \rfloor$ représente la fonction partie entière. Van Der Vaart et Wellner [1996] montre que $\hat{q}(p)$ est consistant.

Cependant, dans l'exemple donné par la Question (2) du Paragraphe 1.1, nous sommes intéressés par le quantile d'ordre $1-p$, avec p qui est en fait bien plus petit que $1/n$, c'est à dire par le cas où le nombre moyen d'observations np au-dessus de $q(p)$ est égal à un nombre très petit. Cela signifie que nous cherchons une valeur $q(p)$ qui est à droite de toutes (ou presque toutes) les observations. Autrement dit, nous souhaitons extrapoler en dehors de la portée des observations disponibles. Puisque c'est le but central de notre problème, nous voulons préserver dans les développements asymptotiques le fait que np doit être bien plus petit que n'importe quelle constante positive. Par conséquent, nous sommes contraints de supposer que p dépend de n, $p = p_n$ tel que $\lim_{n\to\infty} p_n = 0$.

Définition 9 *Un quantile extrême $q(p_n)$ est un quantile d'ordre $1-p_n$ qui vérifie $p_n \to 0$ quand $n \to \infty$.*

Un quantile extrême est donc simplement un quantile dont l'ordre tend vers zéro quand la taille de l'échantillon augmente. Dans le contexte de la finance ou de l'assurance, un quantile extrême s'interprète comme la "Value-at-Risk" d'une perte extrême, voir Embrechts [2000]; McNeil et collab. [2005] pour des liens entre théorie des valeurs extrêmes et théorie des risques. Dans les applications environnementales, un quantile extrême coïncide avec le niveau de retour associé à un évènement climatique exceptionnel (pluies extrêmes Coles et collab. [2003], vitesses de vent extrêmes Jagger et Elsner [2006], hauteurs de vague extrêmes Muir et El-Shaarawi [1986], crues Katz et collab. [2002],...).

La difficulté dans l'estimation des quantiles extrêmes réside en réalité dans la vitesse de convergence de p_n vers zéro. Pour s'en convaincre, on peut se demander quelle est la probabilité qu'un quantile extrême soit plus grand que le maximum de l'échantillon. Sous l'hypothèse que les X_i soient indépendants et identiquement distribués et que $p_n \to 0$, cette probabilité s'écrit :

$$\begin{aligned}
\mathbb{P}(X_{n,n} < q(p_n)) &= \mathbb{P}\left(\bigcap_{i=1}^{n}\{X_i \leq q(p_n)\}\right) \\
&= \prod_{i=1}^{n}\mathbb{P}(X_i \leq q(p_n)) \\
&= F^n(q(p_n)) \\
&= (1-p_n)^n \\
&= \exp(n\log(1-p_n)) \\
&= \exp(-np_n(1+o(1)).
\end{aligned}$$

Ces quelques développements mathématiques nous permettent de voir que la probabilité qu'un quantile extrême soit plus grand que le maximum de l'échantillon dépend du produit np_n. Deux cas se présentent : soit $np_n \to \infty$, soit $np_n \to 0$.

Cas intermédiaire : Dans le cas où $np_n \to \infty$, alors $\mathbb{P}(X_{n,n} < q(p_n)) \to 0$. Par conséquent, le quantile d'intérêt se trouve presque sûrement à l'intérieur de l'échantillon. On parle de quantile intermédiaire (Par analogie, p_n est appelé un niveau intermédiaire dans ce cas). Cela correspond au cas où p_n tend vers zéro lentement, plus lentement que n ne croit vers l'infini. Il est alors possible d'estimer $q(p_n)$ par une simple statistique d'ordre supérieure. Plus précisément, $q(p_n)$ est estimé par

$$\hat{q}(p_n) = X_{n-\lfloor np_n \rfloor, n},$$

c'est à dire la $\lfloor np_n \rfloor$ ième plus grande observation de l'échantillon, voir équation (1.26). Sous des conditions de type von Mises (voir DE HAAN et FERREIRA [2007, Corollaire 1.1.10]), cet estimateur est asymptotiquement gaussien (voir DE HAAN et FERREIRA [2007, Théorème 2.2.1]). Dans la suite, α_n représente un ordre intermédiaire et $q(\alpha_n)$ son quantile associé.

Cas extrême : Dans le cas où $np_n \to 0$, alors $\mathbb{P}(X_{n,n} < q(p_n)) \to 1$. Autrement dit, le quantile d'intérêt se trouve cette fois-ci presque sûrement en dehors de l'échantillon. Dans un tel cas, une simple inversion de la fonction de répartition empirique ne suffit plus pour estimer $q(p_n)$. En effet, $\widehat{F}_n(x) = 1$ pour $x \geq X_{n,n}$. Une extrapolation en dehors de l'échantillon est nécessaire en vue de donner une estimation non triviale de $q(p_n)$. Comme on va le voir par la suite, cette extrapolation se fait à partir des plus grandes données observées de l'échantillon.

Une illustration des notions de quantiles intermédiaires et extrêmes est donnée Figure 1.3. Sur cette figure, la courbe noire correspond à la fonction de répartition d'une loi de Weibull. Des simulations suivant cette dernière ont été reportées sur l'axe des abcisses, qui correspond au support de la loi, et donc à l'axe des quantiles. Le défi de l'estimation de quantiles extrêmes consiste à estimer le quantile rouge, situé en dehors de l'échantillon. Les approches utilisées se basent sur le quantile intermédiaire bleu.

Il existe en théorie des valeurs extrêmes trois approches différentes pour l'estimation des quantiles extrêmes. Ces approches sont basées sur les approximations des queues de distribution issues des Théorèmes 2 et 3 ainsi que des caractérisations des domaines d'attraction du Paragraphe 1.2. Nous présentons ces approches dans les paragraphes qui suivent, où l'on suppose que F est une fonction de répartion appartenant à l'un des domaines d'attraction précédemment introduits.

1.3.2 Une approche basée sur le théorème des valeurs extrêmes : la méthode des maxima par blocs

La première approche se base sur le Théorème 2. D'après ce dernier, pour n assez grand, il vient :

$$\mathbb{P}\left(\frac{X_{n,n}-b_n}{a_n} \leq x\right) = F^n(a_n x + b_n) \simeq G_\gamma(x), \tag{1.27}$$

En passant au logarithme des deux côtés de l'équation (1.27) et en utilisant le lien entre F et $\overline{F}$, il vient

$$n\log\left(1-\overline{F}(a_n x + b_n)\right) \simeq \log G_\gamma(x).$$

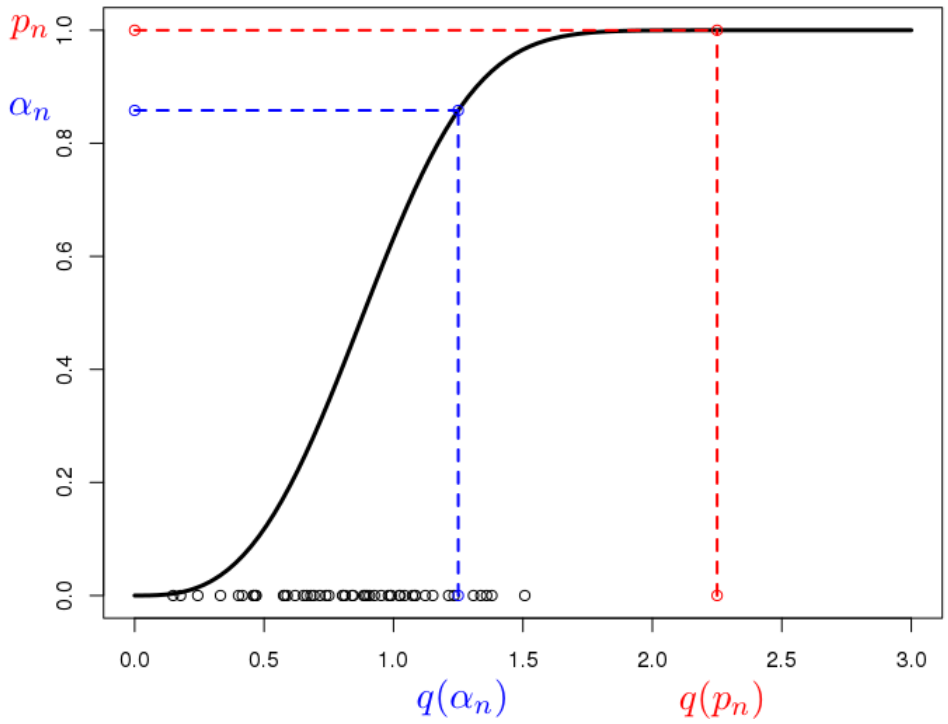

FIGURE 1.3 – Illustration des notions de quantiles intermédiaires (en bleu) et extrêmes (en rouge). La courbe noire représente la fonction de répartition F. Les points noirs en abcisse représentent des réalisations de cette loi.

Un développement limité du logarithme et un changement de variable donnent alors

$$\overline{\mathrm{F}}(x) \simeq -\frac{1}{n}\log \mathrm{G}_\gamma\left(\frac{x-b_n}{a_n}\right).$$

En utilisant alors l'expression analytique de G_γ (cf (1.4)), on obtient finalement une approximation de la queue de distribution de F :

$$\overline{\mathrm{F}}(x) \simeq \frac{1}{n}\left(1+\gamma\frac{x-b_n}{a_n}\right)_+^{-1/\gamma}. \tag{1.28}$$

En particulier, dans le cas où $\gamma = 0$:

$$\overline{\mathrm{F}}(x) \simeq \frac{1}{n}\exp\left(-\frac{x-b_n}{a_n}\right). \tag{1.29}$$

Cette dernière équation nous indique que, dans le cas où $\gamma = 0$, la fonction de survie peut s'approcher par une loi Exponentielle de paramètres appropriés. La Figure 1.4 illustre cette approximation. Sur cette dernière, la courbe noire pleine représente la fonction de répartition d'une loi de Weibull de paramètre de forme égal à trois. La courbe rouge représente l'approximation donnée équation (1.29) (ou plutôt 1 moins cette approximation), c'est à dire la fonction de répartition d'une loi Exponentielle de paramètre donnés par (1.20), avec $n = 20$. On s'aperçoit que la courbe rouge constitue une excellente approximation de la fonction de répartition, notamment dans la queue de distribution, pour de grandes valeurs des quantiles. D'autre part, on a simulé $n = 20$ observations issues une loi de Weibull de paramètre de forme égal à trois. Ces observations, représentées par les cercles sur l'axe des abcisses, nous permettent de construire la fonction de répartition empirique, qui constitue la courbe noire, continue par morceaux sur le graphique. Cela nous permet de remarquer que, bien que la fonction de répartition empirique donne une bonne estimation de la fonction de répartition pour des valeurs moyennes du quantile, dès lors que l'on sort de l'échantillon pour considérer des quantiles plus grands, la qualité d'estimation se déteriore complètement (la fonction de répartition est estimée à 0.95 au point 1.2 et à 1 au point 1.5).

La théorie des valeurs extrêmes nous fournit donc une excellente approximation de la fonction de répartition F en queue. Cependant, rappelons que l'on s'intéresse dans cette partie à l'estimation de quantiles extrêmes, quantiles définis par $q(p_n) = \overline{\mathrm{F}}^{\leftarrow}(p_n)$. Ainsi, l'approximation (1.28) n'est pas suffisante. Il convient d'inverser l'équation associée afin d'obtenir une approximation du quantile :

$$q(p_n) \simeq b_n + \frac{a_n}{\gamma}\left(\left(\frac{1}{np_n}\right)^\gamma - 1\right). \tag{1.30}$$

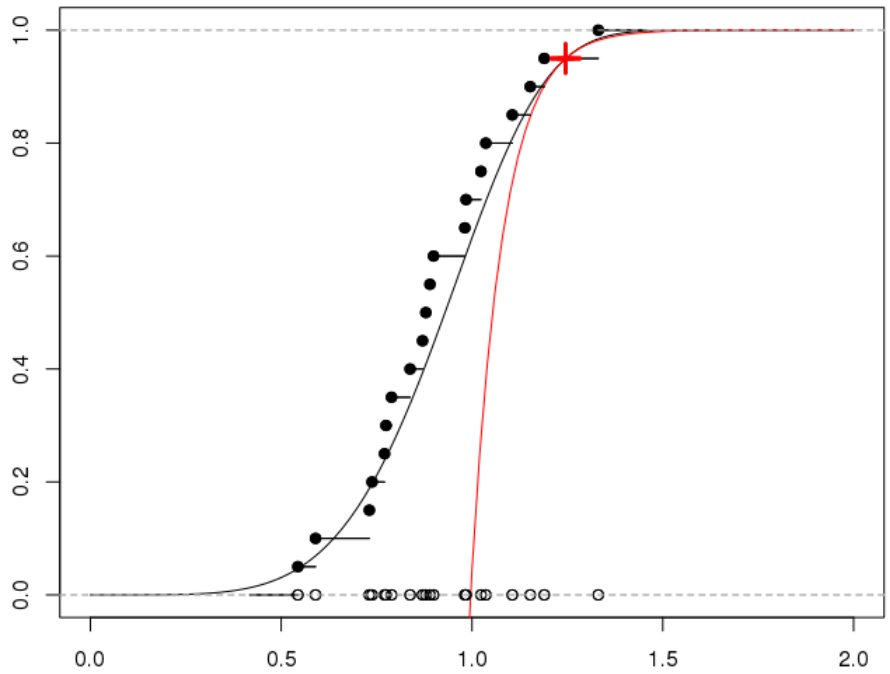

FIGURE 1.4 – Fonction de répartition d'une loi Weibull de paramètre de forme égal à trois (en noir, trait plein), approximation (1.29) associée (en rouge) et fonction de répartition empirique (en noir, fonction continue par morceaux, voir équation (1.25)).

Dans le cas où $\gamma = 0$, cette approximation se réduit à :

$$q(p_n) \simeq b_n - a_n \log(np_n). \tag{1.31}$$

D'après ces deux équations, l'extrapolation est faite à partir du point b_n, que l'on peut choisir comme étant le quantile d'ordre $1/n$ (cf Proposition 8), auquel on ajoute une correction qui dépend de np_n. Cette dernière quantité correspond au rapport entre l'ordre jusqu'où on souhaite extrapoler p_n et l'ordre à partir duquel on extrapole, à savoir $1/n$, l'ordre du maximum.

L'approximation (1.31) est l'objet de la Figure 1.5. Sur cette figure, la courbe noire représente la fonction quantile et la courbe rouge ladite approximation, en fonction de $1-p$. Les observations simulées Figure 1.4 ont été reportées sur le graphe. Cela nous permet de constater que l'approximation de la fonction quantile est encore une fois très bonne, et ce même pour des ordres p proches de zéro. Une légère divergence est tout de même à constater pour des valeurs p très proches de zéro. Ce phénomène est analysé Chapitre 3.

Etant maintenant doté d'une approximation de la fonction quantile, il reste à estimer les constantes normalisantes a_n, b_n et l'indice des valeurs extrêmes γ, qui sont inconnus en pratique. Moyennant $\hat{a}_n$, $\hat{b}_n$ et $\hat{\gamma}_n$ des estimateurs appropriés de a_n, b_n et γ,

Définition 10 *L'estimateur des quantiles extrêmes basé sur le théorème des valeurs extrêmes est défini par :*

$$\hat{q}_{\text{GEV}}(p_n) = \hat{b}_n + \frac{\hat{a}_n}{\hat{\gamma}_n}\left(\left(\frac{1}{np_n}\right)^{\hat{\gamma}_n} - 1\right). \tag{1.32}$$

Afin d'estimer les paramètres a_n, b_n et γ de la loi GEV, le seul maximum de l'échantillon - dont on sait qu'il suit approximativement une GEV d'après le Théorème 2 - ne saurait suffire. Un échantillon complet de réalisations distribuées selon une GEV est nécessaire.

Pour ce faire, GUMBEL [1958] propose l'approche des maxima par bloc. Disposant d'un échantillon de variables aléatoires iid $X_1, X_2, ..., X_n$, cette dernière consiste à le diviser en m blocs de tailles égales, puis à extraire les valeurs maximales de chacun de ces blocs, afin d'obtenir un échantillon de maxima, notés $Z_1, \ldots, Z_m$. Pour peu que m soit assez grand, le Théorème 2 nous indique que les réalisations des variables $Z_1, \ldots, Z_m$ se comportent à peu près comme des réalisations d'un loi GEV.

Disposant maitenant d'un échantillon de réalisations quasiment issu d'une loi GEV, il devient possible d'estimer les paramètres de cette dernière. Nous présentons ci-dessous plusieurs méthodes d'estimation utilisées dans la littérature.

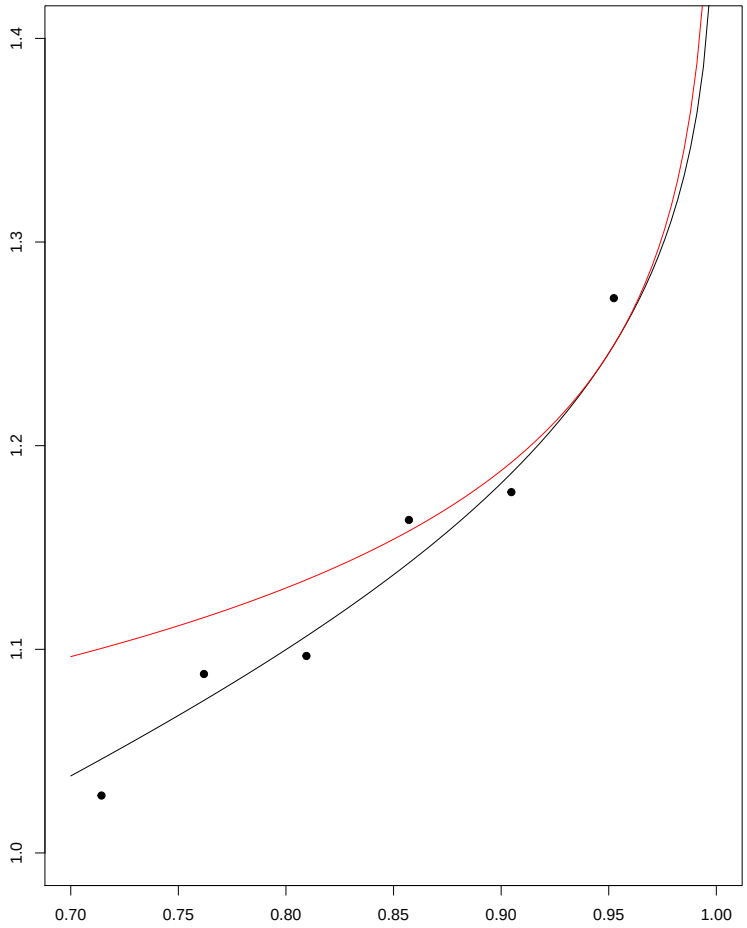

FIGURE 1.5 – Fonction quantile d'une loi Weibull de paramètre de forme égal à trois (en noir) et son approximation (en rouge) basée sur l'équation (1.31) en fonction de l'ordre $1-p$. Les points noirs correspondent à des réalisations de la loi.

Estimation par maximum de vraisemblance L'estimation des paramètres de la GEV par maximum de vraisemblance est proposée par PRESCOTT et WALDEN [1980, 1983]. Celle-ci consiste à maximiser la fonction de log-vraisemblance suivante en les paramètres γ, a_m et b_m :

$$\begin{aligned}\log\big(\mathscr{L}(\gamma, a_m, b_m)\big) &= -m\log(a_m) - \left(1+\frac{1}{\gamma}\right)\sum_{i=1}^{m}\log\left(1+\gamma\left(\frac{Z_i-b_m}{a_m}\right)_+\right) \\ &- \sum_{i=1}^{m}\left(1+\gamma\left(\frac{Z_i-b_m}{a_m}\right)\right)_+^{-1/\gamma}.\end{aligned}$$

L'expression ci-dessus est obtenue à partir de la Définition 1.4. Dans le cas $\gamma = 0$, elle se réécrit

$$\log\big(\mathscr{L}(0, a_m, b_m)\big) = -m\log(a_m) - \sum_{i=1}^{m}\exp\left(-\frac{Z_i-b_m}{a_m}\right) - \sum_{i=1}^{m}\left(\frac{Z_i-b_m}{a_m}\right).$$

N'ayant pas de formes explicites, les estimateurs du maximum de vraisemblance sont alors obtenus à l'aide d'algorithmes d'optimisation, en général des algorithmes de gradient, de type Newton-Raphson, la fonction de log-vraisemblance étant concave. De tels algorithmes sont décrits dans HOSKING et collab. [1985] et MACLEOD [1989].

SMITH [1985] montre que les estimateurs ainsi obtenus présentent des propriétés de consistance et de normalité asymptotique dans le cas où $\gamma > -1/2$. ZHOU [2009], DOMBRY [2015] et ZHOU [2010] étendent ces résultats au cas $\gamma > -1$. ZHOU [2010] prouve également que ces estimateurs ne sont plus consistants dès lors que $\gamma < -1$.

Estimation par la méthode des moments pondérés Les moments pondérés d'une variable aléatoire Z qui possède une fonction de répartition G sont définis (cf GREENWOOD et collab. [1979]) par :

$$M_{p,r,s} = \mathbb{E}[Z^p (G(Z))^r (1-G(Z))^s].$$

En particulier, $M_{p,0,0}$ est le moment d'ordre p classique. Aussi, pour $p=1$ et $s=0$:

$$M_{1,r,0} = \mathbb{E}[Z(G(Z))^r].$$

Cette dernière expression est particulièrement utile dans le cas de la GEV et HOSKING et collab. [1985] montrent que, pour $\gamma < 1$,

$$M_{1,r,0} = \frac{1}{r+1}\left(b_m - \frac{a_m}{\gamma}(1-(r+1)^\gamma \Gamma(1-\gamma))\right),$$

où Γ est la fonction gamma d'Euler définie pour tout $t>0$ par :

$$\Gamma(t) = \int_0^\infty x^{t-1}\exp(-x)dx.$$

Ils proposent alors de définir les estimateurs des moments pondérés comme les solutions du système suivant :

$$\begin{aligned} M_{1,0,0} &= b_m - \frac{a_m}{\gamma}(1-\Gamma(1-\gamma)), \\ 2M_{1,1,0} - M_{1,0,0} &= -\frac{a_m}{\gamma}(1-2^\gamma)\Gamma(1-\gamma), \\ \frac{3M_{1,2,0}-M_{1,0,0}}{2M_{1,1,0}-M_{1,0,0}} &= \frac{3^\gamma - 1}{2^\gamma - 1}, \end{aligned}$$

où les $M_{1,r,0}$, $r \in \{0,1,2\}$ sont respectivement remplacés par leur estimateur empirique

$$\hat{M}_{1,r,0} = \frac{1}{m}\sum_{i=1}^{m} Z_{i,m}\left(\frac{i-1}{m}\right)^r.$$

avec $Z_{1,m}, Z_{2,m}, ..., Z_{m,m}$ les statistiques d'ordre associées à $Z_1, \ldots, Z_m$.

Dans le cas d'échantillons de petite ou de moyenne taille, la méthode des moments pondérés donne de meilleurs résultats que la méthode du maximum de vraisemblance voir HOSKING et collab. [1985]. De plus les estimateurs des moments pondérés sont plus simples à calculer : une approximation polynomiale d'ordre 2 permet d'inverser la dernière équation du système et de proposer des estimateurs dont l'expression est analytique. C'est pourquoi les estimateurs des moments pondérés sont particulièrement appréciés en hydrologie et climatologie, citons DIEBOLT et collab. [2008b] et KATZ et collab. [2002]. Du côté théorique, FERREIRA et HAAN [2015] en prouvent la normalité asymptotique pour $\gamma < 1/2$.

1.3.3 Une approche basée sur le Théorème de Pickands : la méthode des dépassements de seuils

L'approche que nous décrivons ci-dessous se base sur le Théorème 3 - qui, rappelons le, établit que la loi des excès au-dessus d'un seuil u_n peut être approchée par une loi de Pareto généralisée - et sur le lien qui unit la fonction de survie $\bar{F}$ à celle des excès $\bar{F}_{u_n}$, voir équation (1.5). Rappelons en effet que, d'après l'équation précédente,

$$\bar{F}(y+u_n) = \bar{F}(u_n)\bar{F}_{u_n}(y). \tag{1.33}$$

Si l'on pose alors $x = y + u_n$, il vient :

$$\bar{F}(x) = \bar{F}(u_n)\bar{F}_{u_n}(x-u_n). \tag{1.34}$$

Cette relation est très intéressante dans le sens où elle lie la queue de distribution de F au point x, un quantile extrême, avec la queue de distribution de F au point u_n, c'est à dire un quantile intermédiaire, se situant dans l'échantillon. Ce lien se fait via le produit entre $\bar{F}(u_n)$ et $\bar{F}_{u_n}(x-u_n)$. En utilisant alors le Théorème 3, pour un seuil u_n suffisamment grand tel que $u_n = q(\alpha_n)$, il vient :

$$\bar{F}(x) \simeq \alpha_n\left(1+\gamma\frac{x-u_n}{\sigma_n}\right)^{-1/\gamma} \tag{1.35}$$

ou encore, dans le cas $\gamma = 0$,

$$\bar{F}(x) \simeq \alpha_n \exp\left(-\frac{x-u_n}{\sigma_n}\right). \tag{1.36}$$

Il est intéressant de noter que l'approximation (1.36) est une généralisation de l'approximation (1.29) obtenue dans le cadre de l'approche basée sur le théorème des valeurs extrêmes. En effet, $\sigma_n = a_n$ dans le cas où $\gamma = 0$ au vu de l'équation (1.7). De plus, si l'on décide de fixer $\alpha_n = 1/n$, alors le seuil $u_n = q(1/n) = b_n$. Ainsi, la Figure 1.4 constitue une illustration de l'approximation (1.36) dans le cas où le seuil est égal au quantile d'ordre $1/n = 1/20$ ($u \simeq 1.25$).

La Figure 1.6 illustre l'approximation (1.36) pour d'autres choix de seuil : $u \in \{1, 1.25, 1.5\}$.

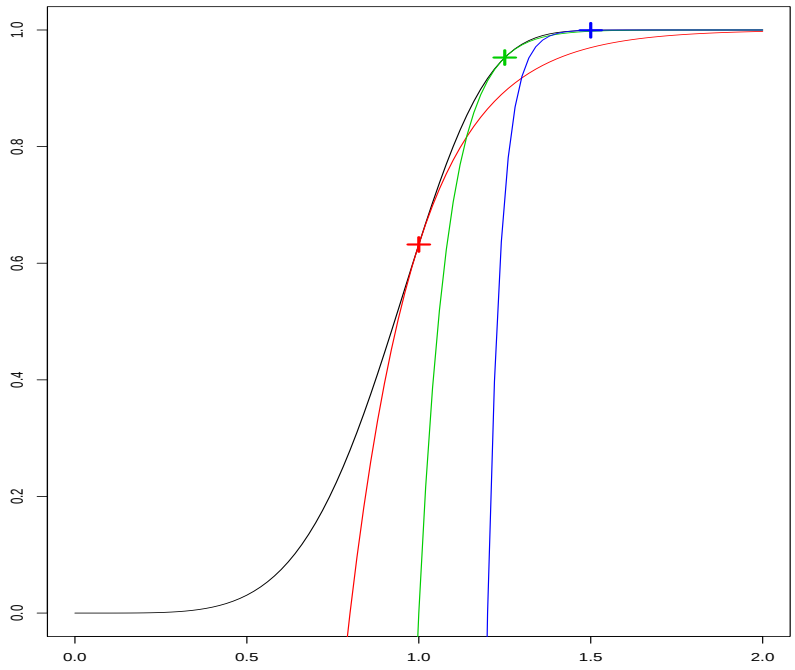

FIGURE 1.6 – Fonction de répartition d'une loi Weibull($a = 5$, $b = 1$) (en noir) et son approximation basée sur l'équation (1.31) pour plusieurs choix du seuil : en rouge, $u = 1$, en vert $u = 1.25$ et en bleu $u = 1.5$.

Comme dans le paragraphe précédent, on souhaite maintenant inverser les équations (1.35) et (1.36) afin d'obtenir une approximation de la fonction quantile. Se faisant, on obtient :

$$q(p_n) \simeq q(\alpha_n) + \frac{\sigma_n}{\gamma}\left[\left(\frac{\alpha_n}{p_n}\right)^{\gamma} - 1\right] \tag{1.37}$$

approximation qui se réduit dans le cas $\gamma = 0$ à

$$q(p_n) \simeq q(\alpha_n) + \sigma_n \log\left(\frac{\alpha_n}{p_n}\right). \tag{1.38}$$

Encore une fois, il est intéressant de remarquer que l'extrapolation se fait à partir du quantile intermédiaire $q(\alpha_n)$ jusqu'au quantile extrême $q(p_n)$ grâce à une correction α_n/p_n qui dépend du ratio des ordres. Dans ce sens, l'approche par le Théorème de Pickands est encore une généralisation de l'approche par le Théorème des valeurs extrêmes.

Etant donné des estimateurs $\hat{q}(\alpha_n)$, $\hat{\sigma}_n$ et $\hat{\gamma}_n$ des quantités inconnues $u_n = q(\alpha_n)$, σ_n et γ,

Définition 11 *L'estimateur des quantiles extrêmes basé sur le Théorème de Pickands est défini par :*

$$\hat{q}_{\mathrm{GPD}}(p_n) = \hat{q}(\alpha_n) + \frac{\hat{\sigma}_n}{\hat{\gamma}_n}\left(\left(\frac{\alpha_n}{p_n}\right)^{\hat{\gamma}_n} - 1\right). \tag{1.39}$$

Par la suite, nous proposons des estimateurs des quantités $u_n = q(\alpha_n)$, σ_n et γ. L'estimation de ces quantités nécessite un échantillon d'excès au-dessus du seuil u_n, de la même façon qu'un échantillon de maxima était requis dans le cas de l'approche des maxima par bloc. On utilise en pratique la méthode POT ("Peaks Over Threshold", dite des dépassements de seuil en français) qui consiste à sélectionner uniquement les observations dépassant un certain seuil u_n donné, puis à y retrancher la valeur du seuil afin d'obtenir un échantillon d'excès noté $Y_1, \ldots, Y_{k_n}$ qui servira à l'inférence des paramètres.

L'estimateur Exponential Tail L'estimateur Exponential Tail est un estimateur proposé par BREIMAN et collab. [1990]. Celui-ci repose sur l'équation (1.38), dans le cas particulier où l'on a affaire à une fonction de répartition F appartenant au domaine d'attraction de Gumbel, c'est à dire dont l'indice des valeurs extrêmes associé est égal à zéro :

$$q(p_n) \simeq q(\alpha_n) + \sigma_n \log\left(\frac{\alpha_n}{p_n}\right). \tag{1.40}$$

En se basant sur cette équation, BREIMAN et collab. [1990] proposent tout d'abord d'estimer le seuil $u_n = q(\alpha_n)$ par une valeur de l'échantillon ordonné. En effet, $q(\alpha_n)$ étant un quantile d'ordre intermédiaire, il se trouve à l'intérieur de l'échantillon. En particulier, si l'on définit k_n le nombre d'excès au-dessus du seuil u_n et que l'on pose $\alpha_n = k_n/n$, alors α_n représente la proportion d'excès dans l'échantillon. Il est alors naturel d'estimer $q(\alpha_n)$ par la $\lfloor n\alpha_n \rfloor$ ième plus grande observation de l'échantillon : $\hat{q}(\alpha_n) = X_{n-\lfloor n\alpha_n \rfloor,n} = X_{n-k_n,n}$ (cf Paragraphe 1.3.1). Notons que cette estimation de $q(\alpha_n)$ est commune à toutes les méthodes que nous présentons et c'est pourquoi nous discutons uniquement de l'estimation des quantités σ_n et γ dans les paragraphes qui suivent.

BREIMAN et collab. [1990] proposent ensuite d'estimer σ_n par la moyenne des excès au-dessus du seuil :

$$\hat{\sigma}_n = \frac{1}{k_n - 1} \sum_{i=1}^{k_n - 1} (X_{n-i,n} - X_{n-k_n,n}) = \frac{1}{k_n} \sum_{i=1}^{k_n} Y_i.$$

Définition 12 *L'estimateur Exponential Tail de* BREIMAN *et collab. [1990] est défini par :*

$$\hat{q}_{ET}(p_n; \alpha_n) = X_{n-k_n,n} + \frac{1}{k_n - 1} \sum_{i=1}^{k_n - 1} (X_{n-i,n} - X_{n-k_n,n}) \log(\alpha_n / p_n).$$

Il correspond à l'équation (1.40) dans laquelle on a remplacé $q(\alpha_n)$ et σ_n par les estimateurs définis ci-dessus. BREIMAN et collab. [1990] proposent également une extension de cet estimateur, appelé l'estimateur Quadratic Tail.

Par la suite, nous revenons au cas général $\gamma \neq 0$ et considérons d'autres méthodes permettant d'estimer σ_n et γ dans (1.39).

Estimation par maximum de vraisemblance Le principe est le même que dans le cas GEV : l'idée est de maximiser en les paramètres σ et γ la log-vraisemblance donnée par

$$\log\big(\mathscr{L}(\gamma, \sigma)\big) = -k_n \log(\sigma) - \left(1 + \frac{1}{\gamma}\right) \sum_{i=1}^{k_n} \log\left(1 + \frac{\gamma}{\sigma} Z_i\right)_+.$$

Là aussi, les solutions n'étant pas explicites, l'optimisation se fait via un algorithme de type Newton-Raphson (voir HOSKING et WALLIS [1987]).

L'existence et la consistance des estimateurs du maximum de vraisemblance est l'objet de ZHOU [2009]. La normalité asymptotique de ces estimateurs est établie par DREES et collab. [2004].

Estimation utilisant les moments HOSKING et WALLIS [1987] proposent une méthode d'estimation des paramètres de la GPD basée sur l'utilisation de ses différents moments. En effet, dans le cas où $\gamma < 1/2$, l'espérance et la variance d'une variable aléatoire issue d'une GPD existent et plus précisemment, on a :

$$\mathbb{E}(Y) = \frac{\sigma}{1 - \gamma} \quad \text{et} \quad \mathrm{Var}(Y) = \frac{\sigma^2}{(1-\gamma)^2(1-2\gamma)}.$$

Il est alors aisé d'exprimer les paramètres σ et γ comme une fonction de ces derniers :

$$\gamma = \frac{1}{2}\left(1 - \frac{\mathbb{E}(Y)^2}{\mathrm{Var}(Y)}\right) \quad \text{et} \quad \sigma = \frac{\mathbb{E}(Y)}{2}\left(1 + \frac{\mathbb{E}(Y)^2}{\mathrm{Var}(Y)}\right).$$

Enfin, il suffit de remplacer ces moments par leur estimateur empirique associé :

$$\overline{Y} := \frac{1}{k_n} \sum_{i=1}^{k_n} Y_i \quad \text{et} \quad s^2(Y) := \frac{1}{k_n - 1} \sum_{i=1}^{k_n} (Y_i - \overline{Y})^2.$$

afin d'obtenir des estimateurs de σ et γ :

$$\hat{\gamma}_n^m = \frac{1}{2}\left(1 - \frac{\overline{Y}^2}{s^2(Y)}\right) \quad \text{et} \quad \hat{\sigma}_n^m = \frac{\overline{Y}}{2}\left(1 + \frac{\overline{Y}^2}{s^2(Y)}\right).$$

Hosking et Wallis [1987] en prouvent la normalité asymptotique sous la condition $\gamma < 1/4$, ce qui représente un domaine assez limité en pratique. La méthode des moments pondérés présentée ci-dessous pallie en partie ce problème.

Estimation utilisant les moments pondérés Suite à cette constatation, Hosking et Wallis [1987] proposent une méthode similaire, basée sur des moments pondérés. Le moment pondéré d'ordre s est défini pour une loi GPD par :

$$M_{1,0,s} := \mathbb{E}(Y(1 - G_{\gamma,\sigma}(Y))^s) = \frac{\sigma}{(1+s)(1+s-\gamma)} \quad \text{avec} \quad \gamma < 1/s.$$

Les estimateurs des moments pondérés sont alors solutions du système suivant :

$$\begin{aligned} M_{1,0,0} &= \frac{\sigma}{1-\gamma}, \\ M_{1,0,1} &= \frac{\sigma}{2(2-\gamma)}, \end{aligned}$$

qui sont explicites et données par

$$\sigma = \frac{2M_{1,0,0}M_{1,0,1}}{M_{1,0,0} - 2M_{1,0,1}} \quad \text{et} \quad \gamma = \frac{M_{1,0,0} - 4M_{1,0,1}}{M_{1,0,0} - 2M_{1,0,1}}.$$

Il ne reste alors qu'à remplacer les quantités $M_{1,0,0}$ et $M_{1,0,1}$ par leur équivalent empirique :

$$\hat{M}_{1,0,s} := \frac{1}{k_n}\sum_{i=1}^{k_n}\left(1 - \frac{i}{1+k_n}\right)^s Y_{i,k_n}$$

pour $s \in \{0,1\}$ avec $Y_{1,k_n}, \ldots, Y_{k_n,k_n}$ les statistiques d'ordre associées à $Y_1, \ldots, Y_{k_n}$. En procédant ainsi, Hosking et Wallis [1987] montrent que le domaine de validité concernant la normalité asymptotique desdits estimateurs est étendu à $\gamma < 1/2$. Parmi les autres travaux utilisant les estimateurs des moments pondérés, on citera Diebolt et collab. [2004, 2007].

Estimateur de Dekkers et collab. [1989] L'estimateur de γ que nous introduisons par la suite est également un estimateur basé sur des moments. Proposé par Dekkers et collab. [1989], il a pour vocation à généraliser le très célèbre estimateur de Hill (estimateur décrit Paragraphe 1.3.4) à tous les domaines d'attraction. En effet, l'estimateur de Hill [1975], défini par

$$\hat{\gamma}_n^H := \frac{1}{k_n}\sum_{i=0}^{k_n-1}\left(\log X_{n-i,n} - \log X_{n-k_n,n}\right),$$

est un estimateur de l'indice des valeurs extrêmes dédié au cas $\gamma > 0$.

Pour lever cette restriction, Dekkers, Einmahl et de Haan proposent de combiner un estimateur de $\gamma_+ := \max(0,\gamma)$ avec un estimateur de $\gamma_- := \min(0,\gamma)$.

Définissons les statistiques suivantes (aussi appelées moments) :

$$m_n^{(\alpha)} := \frac{1}{k_n}\sum_{i=0}^{k_n-1}\left(\log X_{n-i,n} - \log X_{n-k_n,n}\right)^\alpha, \tag{1.41}$$

avec $(k_n)_{n\geq 1}$ une suite d'entiers tel que $1 < k_n \leq n$ et $\alpha \in \{1,2\}$.

Comme estimateur de γ_+, Dekkers et al proposent de reprendre l'estimateur de Hill, qui n'est autre que $m_n^{(1)}$:

$$\hat{\gamma}_{n,+} := m_n^{(1)} = \hat{\gamma}_n^H.$$

Comme estimateur de γ_-, les auteurs remarquent que, si la fonction de répartition associée F vérifie la condition d'appartenance à un domaine d'attraction (cf Théorème 2), alors

$$\frac{\left(m_n^{(1)}\right)^2}{m_n^{(2)}} \xrightarrow{\mathbb{P}} \frac{1 - 2\gamma_-}{2(1-\gamma_-)}.$$

Ils proposent ainsi d'estimer γ_- par

$$\hat{\gamma}_{n,-} := 1 - \frac{1}{2}\left(1 - \frac{\left(m_n^{(1)}\right)^2}{m_n^{(2)}}\right)^{-1}.$$

Finalement, la somme des estimateurs de γ_+ et γ_- constitue l'estimateur des moments de Dekkers, Einmahl et de Haan :

$$\hat{\gamma}_n^{\mathrm{M}} = m_n^{(1)} + 1 - \frac{1}{2}\left(1 - \frac{\left(m_n^{(1)}\right)^2}{m_n^{(2)}}\right)^{-1}. \tag{1.42}$$

En parallèle de cet estimateur, ils proposent d'estimer $\sigma > 0$ par

$$\hat{\sigma}_n = \mathrm{X}_{n-k_n,n}\, m_n^{(1)}(1 - \hat{\gamma}_{n,-}).$$

En faisant alors du plug-in des précédents estimateurs de σ et γ dans l'équation (1.39), on obtient un estimateur du quantile extrême $q(p_n)$:

Définition 13 *L'estimateur des moments de* DEKKERS *et collab. [1989] est défini par :*

$$\hat{q}_{\mathrm{M}}(p_n) = \mathrm{X}_{n-k_n,n} + \mathrm{X}_{n-k_n,n}\frac{m_n^{(1)}(1 - \hat{\gamma}_n^{\mathrm{M}} + m_n^{(1)})}{\hat{\gamma}_n^{\mathrm{M}}}\left(\left(\frac{k_n}{np_n}\right)^{\hat{\gamma}_n^{\mathrm{M}}} - 1\right). \tag{1.43}$$

Les auteurs en prouvent la consistance et la normalité asymptotique (cf par exemple DE HAAN et FERREIRA [2007, Théorème 4.3.1]). Dans le Chapitre 3, nous reprenons les idées développées par Dekkers et al pour proposer un nouvel estimateur des quantiles extrêmes.

La troisième approche que nous décrivons est l'approche semi-paramétrique. Contrairement aux deux approches précédentes, elle se base sur les caractérisations des différents domaines d'attraction (cf Paragraphe 1.2).

1.3.4 L'approche semi-paramétrique

L'approche semi-paramétrique consiste généralement à se placer dans un domaine d'attraction et à utiliser une caractérisation des fonctions associées afin de proposer des estimateurs des quantiles extrêmes. Nous commençons ainsi par nous placer dans le domaine d'attraction de Fréchet. Dans ce cadre, nous présentons l'estimateur de Weissman. L'estimateur de Hill y est également discuté. Puis vient la proposition d'un estimateur des quantiles extrêmes dédié à une famille de lois du domaine d'attraction de Gumbel, les lois à queue de type Weibull. Enfin, nous proposons un estimateur des quantiles extrêmes basé sur un nouveau genre de modèle, dit des lois à queue de type log-Weibull généralisée.

Lois à queue lourde Nous nous restreignons dans ce paragraphe à une fonction de répartition F appartenant au domaine d'attraction de Fréchet. Rappelons qu'une fonction de répartition
$\mathrm{F} \in \mathrm{DA}(\text{Fréchet})$ vérifie $\overline{\mathrm{F}} \in \mathrm{RV}_{-1/\gamma}$:

$$\overline{\mathrm{F}}(x) = x^{-1/\gamma}\mathrm{L}(x) \quad \text{avec} \quad \mathrm{L} \in \mathrm{RV}_0. \tag{1.44}$$

Ainsi, pour obtenir une caractérisation des fonctions quantiles appartenant au domaine d'attraction de Fréchet, il suffit d'inverser l'équation précédente (cf Définition 1.24). En utilisant alors la Proposition 3, il vient que $q \in \mathrm{RV}_{-\gamma}$:

$$q(p) = p^{-\gamma}\ell(p^{-1}) \quad \text{avec} \quad \ell \in \mathrm{RV}_0, \tag{1.45}$$

où $p \in]0,1[$.

L'estimateur de WEISSMAN [1978] tire profit de l'équation (1.45). Pour tout $\gamma > 0$, on peut écrire :

$$q(p_n) = p_n^{-\gamma}\ell(p_n^{-1}), \tag{1.46}$$
$$q(\alpha_n) = \alpha_n^{-\gamma}\ell(\alpha_n^{-1}). \tag{1.47}$$

En divisant alors (1.47) par (1.46) et en utilisant la Définition 1, il vient, pour α_n suffisamment petit et $p_n < \alpha_n$:

$$q(p_n) \simeq q(\alpha_n)\left(\frac{\alpha_n}{p_n}\right)^{\gamma}. \tag{1.48}$$

L'idée est alors de choisir α_n et p_n de telle sorte que $q(p_n)$ soit en dehors de l'échantillon mais que $q(\alpha_n)$ reste dans l'échantillon (cf page 25). L'extrapolation est alors faite à partir de ce dernier, facile à estimer par inversion de la fonction de survie empirique, auquel est multiplié une correction proportionnelle au rapport des ordres des deux quantiles.

Notons que l'approximation (1.48) est un cas particulier de l'approche GPD où $\sigma = \gamma q(\alpha_n)$.

Finalement, il ne reste plus qu'à remplacer $q(\alpha_n)$ et γ par des estimateurs appropriés afin d'obtenir l'estimateur de Weissman :

Définition 14 *L'estimateur de* WEISSMAN *[1978] est défini par :*

$$\hat{q}_{\mathrm{W}}(p_n) = \hat{q}(\alpha_n)\left(\frac{\alpha_n}{p_n}\right)^{\hat{\gamma}_n}. \tag{1.49}$$

Weissman propose d'estimer $q(\alpha_n)$ par son équivalent empirique $X_{n-\lfloor n\alpha_n\rfloor,n}$ et γ en utilisant l'estimateur de HILL [1975] que nous présentons au paragraphe qui suit. Moyennant ces choix, plus de détails sur les propriétés de l'estimateur de Weissman peuvent être trouvés dans WEISSMAN [1978] ou encore dans DE HAAN et FERREIRA [2007, Théorème 4.3.8].

L'estimateur de HILL [1975] est initialement introduit comme un estimateur de l'indice des valeurs extrêmes dans le cas où $\gamma > 0$. Pour construire l'estimateur de Hill, on peut se baser sur l'équation (1.48), que l'on peut réécrire :

$$\log(q(p_n) - \log(q(\alpha_n))) \simeq \gamma \log(\alpha_n / p_n).$$

En choisissant alors $\alpha_n = k_n/n$ et en considérant plusieurs valeurs de $p_n = i/n$, avec $i = 1, \ldots, k_n - 1$ et $p_n < \alpha_n$, on obtient :

$$\log(q(i/n)) - \log(q(k_n/n)) \simeq \gamma \log(k_n/i).$$

Notons que, tout comme $q(\alpha_n)$, $q(p_n)$ représente un quantile de l'échantillon avec ce choix. Si l'on estime maintenant les quantiles intermédiaires par leur équivalent empirique, il vient :

$$\log(X_{n-i+1,n}) - \log(X_{n-k_n+1,n}) \simeq \gamma \log(k_n/i). \tag{1.50}$$

On somme alors de part et d'autres sur $i = 1, \ldots, k_n - 1$:

$$\gamma \simeq \frac{\sum_{i=1}^{k_n-1} \log(X_{n-i+1,n}) - \log(X_{n-k+1,n})}{\sum_{i=1}^{k_n-1} \log(k_n/i)}.$$

En utilisant finalement la formule de Stirling, on montre alors que le dénominateur est équivalent à k_n, ce qui nous permet d'obtenir l'estimateur de Hill :

Définition 15 *L'estimateur de* HILL *[1975] est défini par :*

$$\hat{\gamma}_n^{\mathrm{H}} = \frac{1}{k_n - 1} \sum_{i=1}^{k_n-1} \log\left(X_{n-i+1,n}\right) - \log\left(X_{n-k_n+1,n}\right).$$

DE HAAN et FERREIRA [2007, page 69] et BEIRLANT et collab. [2006, page 101] proposent d'autres façons de construire l'estimateur de Hill. L'étude du comportement asymptotique de l'estimateur de Hill est le sujet de nombreux articles. MASON [1982] en démontre la consistance faible et DEHEUVELS et collab. [1988] la consistance forte. Pour la preuve de la normalité asymptotique de l'estimateur de Hill, citons HAEUSLER et TEUGELS [1985] ou encore DE HAAN et RESNICK [1998].

Pour ce qui est de l'estimation de quantiles extrêmes du domaine d'attraction de Gumbel, les approches semi-paramétriques se basent sur des sous-familles de lois, en particulier celles dites à queue de type Weibull.

Lois à queue de type Weibull Rappelons que les lois à queue de type Weibull présentent la caractérisation suivante (cf (1.21)) :

$$\overline{\mathrm{F}}(x) = \exp\left(-x^{1/\beta}\mathrm{L}(x)\right) \quad \text{avec} \quad \mathrm{L} \in \mathrm{RV}_0.$$

Au vu de la Proposition 3, il est par conséquent facile d'obtenir la caractérisation de la fonction quantile associée :

$$q(p_n) = (-\log(p_n))^{\beta}\mathrm{L}(-\log(p_n)) \quad \text{avec} \quad \mathrm{L} \in \mathrm{RV}_0. \tag{1.51}$$

Cependant, L est inconnue en pratique et l'idée est de procéder de la même manière que pour l'estimateur de Weissman. Pour ce faire, on commence par passer au logarithme puis on divise par $\log_2(1/p_n)$ des deux côtés de l'équation (1.51) :

$$\frac{\log q(p_n)}{\log_2(1/p_n)} = \beta + \frac{\log \mathrm{L}(\log(1/p_n))}{\log_2(1/p_n)},$$

avec $\log_2(.) = \log(\log(.))$. En remarquant que $\log \mathrm{L}(\log(1/p_n))/(\log_2(1/p_n)) \underset{n\to\infty}{\longrightarrow} 0$ au vu de la Proposition 1, il vient, lorsque $n \to \infty$:

$$\log q(p_n) \simeq \beta \log_2(1/p_n). \tag{1.52}$$

De la même façon, en considérant un ordre intermédiaire α_n, on écrit :

$$\log(q(\alpha_n)) \simeq \beta \log_2(1/\alpha_n). \tag{1.53}$$

En soustrayant alors les équations (1.52) et (1.53), puis en passant à l'exponentielle, il vient :

$$q(p_n) \simeq q(\alpha_n)\left(\frac{\log(1/p_n)}{\log(1/\alpha_n)}\right)^{\beta}. \tag{1.54}$$

Moyennant des estimateurs appropriés des quantités $q(\alpha_n)$ et β, Gardes et Girard [2005] proposent l'estimateur des quantiles extrêmes suivant :

Définition 16 *L'estimateur des quantiles extrêmes de Gardes et Girard [2005] pour les lois à queue de type Weibull est défini par :*

$$\hat{q}_{\mathrm{G}}(p_n) = \hat{q}(\alpha_n)\left(\frac{\log(1/p_n)}{\log(1/\alpha_n)}\right)^{\hat{\beta}_n}. \tag{1.55}$$

Il reste à proposer des estimateurs de $q(\alpha_n)$ et β. Encore une fois, $q(\alpha_n)$ étant un quantile intermédiaire, il est estimé par $\mathrm{X}_{n-k_n,n}$ après avoir remplacé α_n par k_n/n. Il reste à estimer β.

Pour proposer un estimateur de β, Girard [2004] propose de repartir de l'expression de la fonction quantile (cf (1.51)) et de considérer deux suites intermédiaires α_n et α'_n :

$$\begin{aligned} q(\alpha'_n) &= (\log(1/\alpha'_n))^{\beta}\mathrm{L}(\log(1/\alpha'_n)), \\ q(\alpha_n) &= (\log(1/\alpha_n))^{\beta}\mathrm{L}(\log(1/\alpha_n)). \end{aligned}$$

En passant au logarithme puis en soustrayant ces deux équations, il vient :

$$\log(q(\alpha'_n)) - \log(q(\alpha_n)) = \beta(\log_2(1/\alpha'_n) - \log_2(1/\alpha_n)) + \log\left(\frac{\mathrm{L}(\log(1/\alpha'_n))}{\mathrm{L}(\log(1/\alpha_n))}\right).$$

Il suffit alors d'utiliser les propriétés des fonctions à variation lente (cf Définition 1) et de remplacer α_n et α'_n par k_n/n et i/n, avec k_n une suite intermédiaire. On obtient l'approximation suivante :

$$\log(q(i/n)) - \log(q(k_n/n)) \simeq \beta(\log_2(n/i) - \log_2(n/k_n)).$$

En sommant de part et d'autre sur i, il vient :

$$\sum_{i=1}^{k_n}\left[\log(q(i/n)) - \log(q(k_n/n))\right] \simeq \beta \sum_{i=1}^{k_n}\left[\log_2(n/i) - \log_2(n/k_n)\right].$$

Finalement, Girard [2004] propose l'estimateur suivant :

Définition 17 *L'estimateur de Girard [2004] pour l'indice de queue de type Weibull est défini par :*

$$\hat{\beta}^{\mathrm{G}}_n = \frac{1}{\sum_{i=1}^{k_n}\left[\log_2(n/i) - \log_2(n/k_n)\right]} \sum_{i=1}^{k_n}\left[\log(\mathrm{X}_{n-i+1,n}) - \log(\mathrm{X}_{n-k_n,n})\right]. \tag{1.56}$$

Notons que cet estimateur diffère de l'estimateur de Hill (cf Définition 15) pour l'indice des valeurs extrêmes uniquement par son terme de normalisation au dénominateur. De nombreux autres estimateurs de l'indice de queue de type Weibull ont été proposés dans la littérature. Nous renvoyons le lecteur au Paragraphe 1.2.6.

Les trois domaines d'attraction donnent des représentations très différentes du comportement des valeurs extrêmes. Choisir un domaine d'attraction et estimer les paramètres n'est pas totalement satisfaisant à deux titres : en premier lieu, cela requiert une méthode permettant de décider dans quel domaine d'attraction se placer. Deuxièmement, une fois le domaine d'attraction choisi, l'inférence associée est présumée correcte et ne prend pas en compte l'incertitude liée au choix fait, même si cette dernière peut-être substantielle. Pour pallier ce problème, de nouvelles approches ont été proposées par El Methni et collab. [2012], puis par de Valk [2016b]; de Valk et Cai [2018]. Ces approches se détachent du choix du domaine d'attraction en proposant des caractérisations qui englobent plusieurs domaines d'attraction (ou plusieurs sous-familles de lois issues de domaines d'attraction différents). Nous en détaillons les idées dans les paragraphes qui suivent.

Lois à queue lourde et à queue de type Weibull El Methni et collab. [2012] proposent un nouvel estimateur des quantiles extrêmes pouvant s'appliquer aussi bien à des lois du domaine d'attraction de Fréchet que des lois à queue de type Weibull, évitant ainsi de faire un choix entre les domaines d'attraction de Gumbel et de Fréchet. Pour ce faire, l'idée proposée est de combiner les caractérisations de la décroissance des fonctions de survie, dans un cas polynomiale et l'autre exponentielle.

Le modèle, proposé par Gardes et collab. [2011], est le suivant :

$$\overline{\mathrm{F}}(x) = \exp(-\mathrm{L}_\tau^{\leftarrow}(\log \mathrm{H}(x))) \tag{1.57}$$

où

$$\mathrm{L}_\tau(x) = \int_1^x u^{\tau-1}\,du,$$

avec $\tau \in [0,1]$ et H une fonction croissante telle que $\mathrm{H}^{\leftarrow} \in \mathrm{RV}_\theta$, $\theta > 0$.

Le paramètre τ régit la lourdeur de la queue de distribution. Les auteurs montrent que,

— si $\tau = 1$, alors F appartient au domaine d'attraction de Fréchet ;

— si $\tau = 0$, alors F est une loi à queue de type Weibull de paramètre θ ;

— si $\tau \in]0,1[$, alors F correspond à tout un panel de loi du domaine d'attraction de Gumbel, dont la queue est plus lourde que celle d'une loi de type Weibull.

Utilisant ce modèle, les auteurs proposent alors un nouvel estimateur des quantiles extrêmes en remarquant que, sous (1.57),

$$\begin{aligned} q(x) &:= \overline{\mathrm{F}}^{\leftarrow}(x) \\ &= \mathrm{H}^{\leftarrow}\left(\exp(\mathrm{L}_\tau(-\log x))\right) \\ &= \left(\exp \mathrm{L}_\tau(-\log x)\right)^\theta \ell\left(\exp \mathrm{L}_\tau(-\log x)\right) \end{aligned}$$

et donc :

$$\log q(p_n) - \log q(\alpha_n) = \theta(\mathrm{L}_\tau(-\log p_n) - \mathrm{L}_\tau(-\log \alpha_n)) + \log\left(\frac{\ell(\exp \mathrm{L}_\tau(-\log p_n))}{\ell(\exp \mathrm{L}_\tau(-\log \alpha_n))}\right). \tag{1.58}$$

Mais ℓ est une fonction à variation lente et par conséquent,

$$q(p_n) \simeq q(\alpha_n) \exp\left\{\theta\left[\mathrm{L}_\tau(-\log p_n) - \mathrm{L}_\tau(-\log \alpha_n)\right]\right\}. \tag{1.59}$$

En remplaçant alors les quantités θ et τ par des estimateurs appropriés, les auteurs proposent l'estimateur suivant.

Définition 18 *L'estimateur des quantiles extrêmes proposé par* El Methni *et collab. [2012] est défini par :*

$$\hat{q}_{\mathrm{E}}(p_n) = \hat{q}(\alpha_n) \exp\left\{\hat{\theta}_n\left[\mathrm{L}_{\hat{\tau}_n}(\log 1/p_n) - \mathrm{L}_{\hat{\tau}_n}(\log 1/\alpha_n)\right]\right\}. \tag{1.60}$$

Le paramètre θ est alors estimé comme suit.

Définition 19 *L'estimateur de* El Methni *et collab. [2012] pour* θ *dans le modèle (1.57) est défini par :*

$$\hat{\theta}_n^{\mathrm{G}} = \frac{\hat{\gamma}_n^{\mathrm{H}}(k_n)}{\mu_{\hat{\tau}_n}(\log n/k_n)}, \tag{1.61}$$

où $\hat{\gamma}_n^{\mathrm{H}}$ *est l'estimateur de Hill (cf 15) et* $\mu_\tau(t)$ *est défini, quel que soit* $t > 0$ *par :*

$$\mu_\tau(t) = \int_0^\infty \left[\mathrm{L}_\tau(x+t) - \mathrm{L}_\tau(t)\right] e^{-x}\,dx.$$

Pour estimer θ, l'idée est de repartir de l'équation (1.58) en posant $t = -\log \alpha_n$ et $x + t = -\log p_n$:

$$\log q(e^{-(x+t)}) - \log q(e^{-t}) \simeq \theta\left[\mathrm{L}_\tau(x+t) - \mathrm{L}_\tau(t)\right].$$

En intégrant des deux côtés par rapport à x sur $]0,+\infty[$, il vient :

$$\int_0^\infty \left[\log q(e^{-(x+t)}) - \log q(e^{-t})\right] e^{-x}\,dx \simeq \theta \int_0^\infty \left[\mathrm{L}_\tau(x+t) - \mathrm{L}_\tau(t)\right] e^{-x}\,dx.$$

Finalement,

$$\theta \simeq \frac{\int_0^\infty \left[\log q(e^{-(x+t)}) - \log q(e^{-t})\right] e^{-x}\,dx}{\int_0^\infty \left[\mathrm{L}_\tau(x+t) - \mathrm{L}_\tau(t)\right] e^{-x}\,dx}.$$

L'estimateur (1.61) est alors obtenu en remplaçant e^{-t} par k_n/n (avec (k_n) une suite intermédiaire d'entiers vérifiant $\lim_{n\to\infty} k_n = \infty$ et $\lim_{n\to\infty} k_n/n = 0$) et q par son équivalent empirique.

Pour τ, El Methni et collab. [2012] proposent l'estimateur suivant :

Définition 20 *L'estimateur de* El Methni *et collab. [2012] pour* τ *dans le modèle (1.57) est défini par :*

$$\hat{\tau}_n^{\mathrm{E}} = \psi^{-1}\left(\frac{\hat{\gamma}_n^{\mathrm{H}}(k_n)}{\hat{\gamma}_n^{\mathrm{H}}(k'_n)}; \log n/k_n; \log n/k'_n\right) \tag{1.62}$$

avec

$$\psi(x;t;t') = \frac{\mu_x(t)}{\mu_x(t')}. \tag{1.63}$$

Pour proposer (1.62), les auteurs commencent par considérer deux suites intermédiaires d'entiers (k_n) et (k'_n) avec $1 < k_n < k'_n$ pour tout n et utilise le fait que $\hat{\theta}(k_n) \xrightarrow{\mathbb{P}} \theta$ et $\hat{\theta}(k'_n) \xrightarrow{\mathbb{P}} \theta$ pour écrire :

$$\frac{\hat{\theta}_n(k_n)}{\hat{\theta}_n(k'_n)} = \frac{\hat{\gamma}_n^{\mathrm{H}}(k_n)}{\hat{\gamma}_n^{\mathrm{H}}(k'_n)} \frac{\mu_\tau(\log n/k'_n)}{\mu_\tau(\log n/k_n)} \xrightarrow{\mathbb{P}} 1.$$

L'estimateur (1.62) est alors obtenu en définissant pour tout $t > t' > 0$ la fonction ψ donnée par (1.63) et en remarquant que cette fonction est une bijection de $\mathbb{R}$ dans $(-\infty, k'_n/k_n)$.

Avec ce choix d'estimateurs, El Methni et collab. [2012] prouvent la normalité asymptotique de (1.60).

Lois à queue de type log-Weibull généralisé Dans la même idée que l'estimateur précédent, de Valk et Cai [2018] propose un nouvel estimateur des quantiles extrêmes permettant d'extrapoler des quantiles issus de lois très diverses. L'estimateur en question se base sur un nouveau modèle introduit et discuté par de Valk [2016b] lui-même. Au lieu de considérer la caractérisation habituelle d'appartenance à un domaine d'attraction (cf Théorème 14), qui indique qu'une fonction de répartition $F \in \mathrm{DA}(G_\gamma)$ si et seulement si U est à variation régulière étendue ($U \in \mathrm{ERV}_\gamma$), de Valk [2016b] propose de définir une nouvelle relation limite basée sur la fonction suivante : $V(x) := \log U(e^x) = \log \overleftarrow{1/F}(e^{-x}) = \log q(e^{-x})$. Son modèle est le suivant : il suppose que $V \in \mathrm{ERV}_\theta$ (cf Définition 4) en lieu et place de U, ou plus précisémment (cf équation (1.11)) :

$$\lim_{x\to\infty} \frac{V(tx) - V(x)}{a(x)} = L_\theta(t) \quad t > 0, \tag{1.64}$$

avec θ un réel appelé l'indice de queue de type log-Weibull généralisé et a une fonction positive mesurable.

Il montre dans de Valk [2016b] que le modèle en question englobe une grande variété de lois, parmi lesquelles on compte l'ensemble des lois du domaine d'attraction de Fréchet, la majeure partie du domaine d'attraction de Gumbel (lois à queue de type Weibull, lois à queue de type log-Weibull et lois à point terminal fini) et enfin des lois dites à queue super lourde qui ne vérifient pas les conditions du Théorème 2.

Il montre également que ce modèle permet d'extrapoler des quantiles plus extrêmes que le modèle classique $U \in \mathrm{ERV}$, au sens où

$$p_n \in [n^{-\tau_2}, n^{-\tau_1}],$$

$1 < \tau_1 < \tau_2$.

Par ailleurs, il explique que ce modèle généralise celui de Gardes et collab. [2011], qui implique une limite log-Weibull généralisée de type (1.64) avec $\theta \in [0,1]$ et $a(x) = c\gamma^\theta$, $c > 0$.

De ce modèle, il tire l'estimateur suivant :

Définition 21 *L'estimateur des quantiles extrêmes proposé par* de Valk *et* Cai *[2018] est défini par*

$$\hat{q}_V(p_n) = \hat{q}(\alpha_n) \exp\left(\hat{a}_n(\log 1/\alpha_n) L_{\hat{\theta}_n}\left(\frac{\log 1/p_n}{\log 1/\alpha_n}\right)\right) \tag{1.65}$$

Pour comprendre le raisonnement derrière (1.65), il faut repartir de l'équation (1.64). Pour x assez grand, on a :

$$\frac{V(tx) - V(x)}{a(x)} \simeq L_\theta(t) \tag{1.66}$$

ou encore

$$V(tx) \simeq V(x) + a(x) L_\theta(t). \tag{1.67}$$

Or $V(x) := \log q(\exp(-x))$ et l'équation (3.2) se réecrit

$$q(e^{-tx}) \simeq q(e^{-x}) \exp\left(a(x) L_\theta(t)\right).$$

En remplaçant alors e^{-tx} par p_n et e^{-x} par α_n, il vient :

$$q(p_n) \simeq q(\alpha_n) \exp\left(a(\log 1/\alpha_n) L_\theta\left(\frac{\log 1/p_n}{\log 1/\alpha_n}\right)\right). \tag{1.68}$$

Pour θ, de Valk et Cai [2018] proposent l'estimateur suivant :

Définition 22 *L'estimateur proposé par* DE VALK *et* CAI *[2018] de* θ *dans le modèle (1.64) est défini par :*

$$\hat{\theta}_n^{\mathrm{V}} := 1 + \frac{\sum_{i=1}^{k_n-1} \left(\log \hat{\gamma}_n^{\mathrm{H}}(i) - \log \hat{\gamma}_n^{\mathrm{H}}(k_n)\right)}{\sum_{i=1}^{k_n-1} \left(\log(\vartheta_{i+1,n}) - \log(\vartheta_{k_n+1,n})\right)},$$

avec $\hat{\gamma}_n^{\mathrm{H}}$ *l'estimateur de Hill décrit équation (15),* $\vartheta_{i,n} := \sum_{j=i}^{n} j^{-1}$ *et* (k_n) *une suite intermédiaire d'entiers. pour tout* n*.*

Pour a, les auteurs proposent :

Définition 23 *L'estimateur proposé par* DE VALK *et* CAI *[2018] de a dans le modèle (1.64) est défini par :*

$$\hat{a}_n^{\mathrm{V}}(\log 1/\alpha_n) = \frac{\hat{\gamma}_n^{\mathrm{H}}(l_n)}{\frac{1}{l_n}\sum_{j=1}^{l_n} \mathrm{L}_{\hat{\theta}_n}\left(\frac{\vartheta_{j,n}}{\vartheta_{l_n+1,n}},\right)} \tag{1.69}$$

où (l_n) *est une suite intermédiaire d'entiers telle que* $1 \le l_n \le k_n$ *pour tout* n*.*

Nous donnons les principales pistes permettant de construire ces deux estimateurs ci-dessous. Nous commençons par ce dernier. L'idée derrière l'estimateur (1.69) est d'utiliser à nouveau l'approximation (3.1) :

$$a(x)\mathrm{L}_\theta(t) \simeq \mathrm{V}(tx) - \mathrm{V}(x). \tag{1.70}$$

En remplaçant V par $\log q(\exp(-x))$, x par $\log n/l_n$ et tx par $\log n/l_n'$, il vient :

$$a(\log n/l_n)\mathrm{L}_\theta\left(\frac{\log n/l_n'}{\log n/l_n}\right) \simeq \log q(l_n'/n) - \log q(l_n/n). \tag{1.71}$$

En sommant alors sur l_n', on obtient :

$$a(\log n/l_n)\sum_{j=1}^{l_n} \mathrm{L}_\theta\left(\frac{\log n/j}{\log n/l_n}\right) \simeq \sum_{j=1}^{l_n} \log q(j/n) - \log q(l_n/n). \tag{1.72}$$

L'estimateur (1.69) est finalement obtenu en remarquant que $\log \frac{n}{j+1} < \vartheta_{j,n} := \sum_{i=j}^{n} i^{-1} < \log \frac{n}{j}$. Notons que l'estimateur (1.69) de $a(\log 1/\alpha_n)$ est très similaire à l'estimateur de θ proposé par GARDES et collab. [2011] dans le modèle (1.57) (cf (1.61)). Seule la construction du dénominateur les différencie.

Pour estimer θ, il faut remarquer que $a(x) = x^\theta \ell(x)$ (cf Définition 4). En considérant alors $0 < \alpha_n' < \alpha_n < 1$ des ordres intermédiaires, il vient :

$$\begin{aligned} a(\log 1/\alpha_n') &= (\log 1/\alpha_n')^\theta \ell(\log 1/\alpha_n') \\ a(\log 1/\alpha_n) &= (\log 1/\alpha_n)^\theta \ell(\log 1/\alpha_n). \end{aligned}$$

En passant au logarithme, puis en soustrayant les deux équations précédentes, on obtient :

$$\log a(\log 1/\alpha_n') - \log a(\log 1/\alpha_n) \simeq \theta\left[\log_2 1/\alpha_n' - \log_2 1/\alpha_n\right].$$

En remplaçant alors α_n' par i/n et α_n par k_n/n, puis en sommant sur i, il vient :

$$\theta \simeq \frac{\sum_{i=1}^{k_n-1}\left[\log a(\log(i/n)) - \log a(\log(k_n/n))\right]}{\sum_{i=1}^{k_n-1}\left[\log_2(i/n) - \log_2(k_n/n)\right]}. \tag{1.73}$$

Mais $\log \frac{n}{i+1} < \vartheta_{i,n} < \log \frac{n}{i}$ et

$$\begin{aligned} a(\vartheta_{i+1,n}) &\sim \frac{\sum_{j=1}^{i}\left[\log q(\vartheta_{j,n}) - \log q(\vartheta_{i+1,n})\right]}{\sum_{j=1}^{i}\mathrm{L}_\theta\left(\frac{\vartheta_{j,n}}{\vartheta_{i+1,n}}\right)} \\ &\sim \vartheta_{i+1,n}\frac{1}{i}\sum_{j=1}^{i}\left[\log q(\vartheta_{j,n}) - \log q(\vartheta_{i+1,n})\right] \end{aligned}$$

d'après (1.72). D'où

$$\theta \simeq 1 + \frac{\sum_{i=1}^{k_n-1}\left(\log\left(\frac{1}{i}\sum_{j=1}^{i}\left[\log q(\vartheta_{j,n}) - \log q(\vartheta_{i+1,n})\right]\right) - \log\left(\frac{1}{k_n}\sum_{j=1}^{k_n}\left[\log q(\vartheta_{j,n}) - \log q(\vartheta_{k_n+1,n})\right]\right)\right)}{\sum_{i=1}^{k_n-1}\left(\log(\vartheta_{i+1,n}) - \log(\vartheta_{k_n+1,n})\right)}$$

et le résultat voulu.

1.3.5 En pratique : niveau/période de retour et retour sur l'hypothèse d'indépendance

Ayant introduit les notions de quantiles extrêmes, il est facile de voir que, statistiquement parlant, la Question (2) du Paragraphe 1.1 se rapporte à l'estimation d'un quantile extrême alors que la Question (1) du Paragraphe 1.1 se rapporte à la probabilité associée à un tel quantile. Cependant, en pratique, les experts s'intéressent plutôt aux questions suivantes :

(1) Quelle est la probabilité d'observer un évènement d'amplitude supérieure à une valeur x au cours d'une année donnée ?

(2) Quelle est la valeur dépassée par le maximum annuel avec une faible probabilité p donnée ?

Les questions précédentes diffèrent légèrement des Questions (1) et (2) du Paragraphe 1.1 dans le sens où elle se rapportent à l'estimation d'un quantile issu de la loi du maximum annuel d'une série de données. Or jusqu'à présent, nous avons considéré des quantiles extrêmes de la loi F des données. Le paragraphe qui suit s'attache à faire le lien entre les deux quantités précédentes.

Niveau de retour/ Période de retour

Définition 24 *On appelle niveau de retour associé à la période de retour* $\mathrm{T} = 1/r$ *le quantile extrême d'ordre* $1-r$ *de la loi du maximum* $\mathrm{X}_{m,m}$ *sur une période donnée (*$m = 365$ *si l'on considère le maximum annuel de données journalières).*

Au vu de sa définition, on s'attend à ce que

— le niveau de retour $x_{1/\mathrm{T}}$ soit dépassé en moyenne toutes les $\mathrm{T} = 1/r$ périodes ;

— pour tout année considérée, le maximum annuel dépasse le niveau $x_{1/\mathrm{T}}$ avec probabilité $r = 1/\mathrm{T}$.

Mathématiquement parlant, le niveau de retour $z(r)$ est obtenu en résolvant l'équation suivante :

$$\mathbb{P}(\mathrm{X}_{m,m} \leq z(r)) = 1 - r.$$

Pour ce faire, on peut utiliser le Théorème 2 qui indique que la fonction de répartition du maximum renormalisé peut-être approchée par une loi GEV pour m suffisamment grand :

$$\mathbb{P}(\mathrm{X}_{m,m} \leq z(r)) \simeq \mathrm{G}_{\gamma,b_m,a_m}(z(r)),$$

où $\mathrm{G}_{\gamma,b_m,a_m}(x) := \mathrm{G}_\gamma((x-b_m)/a_m)$. Par conséquent, il suffit maintenant de résoudre $\mathrm{G}_{\gamma,b_m,a_m}(z(r)) = 1-r$ pour obtenir l'approximation suivante du niveau de retour $z(r)$:

$$z(r) \simeq \begin{cases} b_m - \dfrac{a_m}{\gamma}\left[1 - (-\log(1-r))^{-\gamma}\right] & \text{si} \quad \gamma \neq 0 \\ b_m - a_m \log(-\log(1-r)) & \text{si} \quad \gamma = 0. \end{cases}$$

Pour approfondir l'interprétation des quantités que sont les niveaux de retour, il est intéressant d'établir un lien entre ces derniers et les quantiles extrêmes d'ordre $1-p$ de la distribution des données.

Lien entre niveaux de retour et quantiles extrêmes de la distribution des données. Rappelons qu'un quantile extrême de fonction de répartition F est approché par (cf équations (1.30) et (1.31)) :

$$q(p_m) \simeq \begin{cases} b_m + \frac{a_m}{\gamma}\left(\left(\frac{1}{m p_m}\right)^\gamma - 1\right) & \text{si} \quad \gamma \neq 0 \\ b_m - a_m \log(m p_m) & \text{si} \quad \gamma = 0. \end{cases}$$

En imposant $z(r) = q(p_m)$ avec

$$\begin{aligned} \mathbb{P}(\mathrm{X}_{m,m} \leq z_q) &= 1 - r \\ \mathbb{P}(\mathrm{X} \leq q(p_m)) &= 1 - p_m \end{aligned}$$

et en remarquant que $\mathbb{P}(\mathrm{X} \leq y) = \mathbb{P}^{1/m}(\mathrm{X}_{m,m} \leq y)$ pour tout y, il vient

$$1 - r = (1 - p_m)^m$$

ou encore $r \sim m p_m$ lorsque p_m est petit.

Par conséquent, il est possible d'interpréter le niveau de retour comme suit :

— Le niveau de retour $z(r)$ correspond au niveau qu'on s'attend à voir dépassé en moyenne une fois toutes les $1/p_m \sim m/r$ observations (ou jours si les données sont journalières) ;

— Le quantile extrême d'ordre $1-p_m$ de la distribution des données correspond au niveau qu'on s'attend à voir dépassé en moyenne une fois toutes les $T = 1/r \sim 1/(mp_m)$ périodes.

Un niveau de retour à 1 an revient à calculer le quantile d'ordre 1/365. Une crue centennale correspond ainsi à un niveau de retour associé à une période de retour T = 100 ans.

En conclusion, le niveau de retour de période T correspond approximativement (lorsque T est grand) au quantile extrême d'ordre $1/(m\mathrm{T})$. L'estimation des niveaux de retour est donc similaire à celles des quantiles extrêmes. Nous renvoyons le lecteur à la Partie 1.3 pour plus de détails.

Retour sur l'hypothèse d'indépendance Jusqu'à présent, on s'est placé dans le cadre où les observations sont indépendantes entre elles. Or, plus les évènements/mesures sont proches dans le temps, plus on a de risques que les données présentent de la dépendance. Il s'agit donc d'une hypothèse peu vérifiée en pratique.

Pour pallier ce problème, on peut commencer par s'intéresser à l'étude d'une série stationnaire présentant de la dépendance. Nous rappelons tout d'abord la définition d'une série stationnaire.

Définition 25 *Une série* $\mathrm{X}_1,\mathrm{X}_2,...$ *est dite stationnaire si pour tout ensemble d'entiers* $i_1,...,i_k$ *et pour tout entier* m*, la distribution jointe de* $(\mathrm{X}_{i_1},...,\mathrm{X}_{i_k})$ *est identique à celle de* $(\mathrm{X}_{i_1+m},...,\mathrm{X}_{i_k+m})$.

Soit maintenant $\tilde{\mathrm{X}}_1,\tilde{\mathrm{X}}_2,...,\tilde{\mathrm{X}}_n$ une suite stationnaire de marginale F vérifiant la condition $\mathrm{D}(u_n)$ suivante :

Définition 26 (Coles et collab. [2001], Définition 5.1) *Une série stationnaire* $\tilde{\mathrm{X}}_1,\tilde{\mathrm{X}}_2,...,\tilde{\mathrm{X}}_n$ *satisfait la condition* $\mathrm{D}(u_n)$ *si pour tout* $i_1 < ... < i_p < j_1 < ... j_q$ *avec* $j_1 - i_p > l$*, on a*

$$|\mathbb{P}(\tilde{\mathrm{X}}_{i_1} \leq u_n,...,\tilde{\mathrm{X}}_{i_p} \leq u_n,\tilde{\mathrm{X}}_{j_1} \leq u_n,...,\tilde{\mathrm{X}}_{j_q} \leq u_n) - \mathbb{P}(\tilde{\mathrm{X}}_{i_1} \leq u_n,...,\tilde{\mathrm{X}}_{i_p} \leq u_n)\mathbb{P}(\tilde{\mathrm{X}}_{j_1} \leq u_n,...,\tilde{\mathrm{X}}_{j_q} \leq u_n)| < \alpha(n,l), \tag{1.74}$$

où $\alpha(n,l) \to 0$ *avec* $l = l_n$ *telle que* $l_n/n \to 0$ *quand* $n \to +\infty$ *et* u_n *une suite de seuils,* $u_n \underset{n\to\infty}{\longrightarrow} +\infty$.

Notons que pour des variables indépendantes, la différence en probabilité dans (1.74) est exactement zéro quelle que soit la suite u_n considérée. Dans le cas où les variables présentent de la dépendance, la condition précédente assure que, pour un ensemble de variables suffisamment éloignées (au sens donné par la Définition ci-dessus), la différence en probabilité dans (1.74) est suffisamment proche de zéro pour que cela n'ait pas d'effet sur la loi limite donnée dans le Théorème 2. Cette idée est développée par le résultat suivant :

Théorème 15 (Coles et collab. [2001], Théorème 5.1) *Soit* $\tilde{\mathrm{X}}_1,\tilde{\mathrm{X}}_2,...,\tilde{\mathrm{X}}_n$ *une suite stationnaire. S'il existe des suites normalisantes* a_n *et* $b_n > 0$ *telles que*

$$\lim_{n\to+\infty} \mathbb{P}\left(\frac{\tilde{\mathrm{X}}_{n,n} - a_n}{b_n} \leq z\right) = \tilde{\mathrm{G}}_\gamma(z),$$

avec $\tilde{\mathrm{G}}_\gamma$ *non-dégénérée et si la condition* $\mathrm{D}(u_n)$ *est vérifiée avec* $u_n = b_n z + a_n$ *pour tout* z*, alors* $\tilde{\mathrm{G}}_\gamma$ *correspond à la loi généralisée des valeurs extrêmes.*

La preuve de ce théorème peut-être trouvée dans Leadbetter [1983]. Ce théorème est important dans le sens où il indique qu'il est possible d'obtenir pour le maximum d'une série stationnaire des résultats analogues à ceux du cas indépendant pour peu que la série présente une dépendance à court terme (dans le sens où elle vérifie (1.74)).

Cependant, la dépendance affecte les paramètres de la loi limite en question. C'est l'objet du théorème suivant, qui compare les paramètres des lois GEV d'une série iid et d'une série stationnaire.

Théorème 16 (Coles et collab. [2001], Théorème 5.2) *Soient* $\tilde{\mathrm{X}}_1,\tilde{\mathrm{X}}_2,...,\tilde{\mathrm{X}}_n$ *une suite stationnaire de marginale* F *et* $\mathrm{X}_1,\mathrm{X}_2,...,\mathrm{X}_n$ *une suite de variables aléatoires indépendantes et identiquement distribuées suivant* F*. Sous certaines conditions de régularité,*

$$\lim_{n\to+\infty} \mathbb{P}\left(\frac{\mathrm{X}_{n,n} - a_n}{b_n} \leq z\right) = \mathrm{G}_\gamma(z),$$

où a_n *et* $b_n > 0$ *sont des suites normalisantes, avec* $\tilde{\mathrm{G}}_\gamma$ *une loi non-dégénérée est équivalent à :*

$$\lim_{n\to+\infty} \mathbb{P}\left(\frac{\tilde{\mathrm{X}}_{n,n} - a_n}{b_n} \leq z\right) = \tilde{\mathrm{G}}_\gamma(z),$$

avec

$$\tilde{\mathrm{G}}_\gamma(z) = \mathrm{G}_\gamma^\omega(z) \tag{1.75}$$

et où ω *est une constante telle que* $0 < \omega \leq 1$.

Au vu de cette dernière relation, il est possible de vérifier que $\tilde{G}_\gamma(z)$ est toujours une loi GEV, en cohérence avec le Théorème 15 :

$$\begin{aligned}\tilde{G}_\gamma(z) &= G_\gamma^\omega(z) \\ &= \exp\left\{-\left[1+\gamma\left(\frac{z-b_n}{a_n}\right)\right]^{-1/\gamma}\right\}^\omega \\ &= \exp\left\{-\omega\left[1+\gamma\left(\frac{z-b_n}{a_n}\right)\right]^{-1/\gamma}\right\} \\ &= \exp\left\{-\left[1+\gamma\left(\frac{z-b_n^*}{a_n^*}\right)\right]^{-1/\gamma}\right\},\end{aligned}$$

avec $b_n^* = b_n - \frac{a_n}{\gamma}\left(1-\omega^\gamma\right)$ et $a_n^* = a_n\omega^\gamma$.

Le Théorème 16 nous indique donc que la loi limite du maximum d'une série stationnaire - si elle existe, c'est à dire si (1.74) est vérifiée au vu du Théorème 15 - est non seulement une loi généralisée des valeurs extrêmes, mais aussi que cette dernière est liée à la loi du maximum d'une série iid. Cette liaison se traduit par le remplacement de la loi $G_\gamma(z)$ par $G_\gamma^\omega(z)$, où ω est appelé l'indice extrémal (nous en détaillons l'interprétation ci-dessous). Ainsi, travailler avec une série stationnaire revient (sous certaines conditions) à se ramener au cadre iid mais en ne travaillant plus qu'avec $n\omega$ observations au lieu de n, ce qui consiste en une perte d'information puisque $0 < \omega \leq 1$. Cette perte d'information est plus ou moins importante selon la valeur de ω :

— Plus ω est proche de zéro, plus cette perte sera conséquente ;

— Le cas $\omega = 1$ correspond au cas d'une série indépendante.

Les détails sur les conditions de régularité nécessaire à l'établissement du Théorème 16 peuvent être trouvés dans LEADBETTER [1983].

Dans le paragraphe qui suit, nous explicitons l'impact de la dépendance sur le niveau de retour associé à une période de retour T.

Approche des maxima par bloc Au vu des résultats précédents, l'approche des maxima par bloc reste valide dans le cas d'une série stationnaire, et ce via l'introduction de l'indice extrémal. La procédure est la même que dans le cas indépendant. Seule une perte d'information est à déplorer par rapport à ce dernier. Ainsi, dans le cas indépendant, le niveau de retour $z(r)$ était approché par

$$z(r) \simeq b_m - \frac{a_m}{\gamma}\left[1-(-\log(1-r))^{-\gamma}\right] \quad \gamma \neq 0.$$

Dans le cas dépendant, cette expression devient alors :

$$z(r) \simeq b_m - \frac{a_m}{\gamma}\left[1-\left(-\frac{1}{\omega}\log(1-r)\right)^{-\gamma}\right] \quad \gamma \neq 0.$$

Mais

$$\frac{1}{\omega}\log(1-r) = \log\left((1-r)^{1/\omega}\right) = \log(1-r'),$$

avec

$$r' = 1-(1-r)^{1/\omega} \simeq \frac{r}{w}.$$

Par conséquent, le quantile extrêmes d'ordre r dans le cas dépendant correspond approximativement (lorsque r est petit, ce qui est par définition le cas d'un quantile extrême) au quantile d'ordre r/ω dans le cas indépendant. Le paramètre ω étant compris entre 0 et 1, $r < r/\omega$ et la portée des extrapolations se voit réduite dans le cas dépendant.

Approche des excès au-delà d'un seuil L'approche des excès, dite POT, doit être adaptée. Une solution consiste à ne retenir qu'une seule réalisation (la plus grande) dans un "cluster" de dépassements. On peut ensuite considérer ces réalisations comme indépendantes et y ajuster une GPD de façon classique (cf Paragraphe 1.3.3). Dans le cas indépendant, nous avions :

$$q(p_n) \simeq q(\alpha_n) + \frac{\sigma_n}{\gamma_n}\left(\left(\frac{\alpha_n}{p_n}\right)^{\gamma_n} - 1\right) \quad \gamma \neq 0$$

et nous remplaçions α_n par k_n/n où k_n représentait le nombre d'excès. Dans le cas où nous sommes en présence de dépendance, il convient de remplacer k_n par le nombre N_{k_n} de clusters de dépassements :

$$q(p_n) \simeq q(N_{k_n}/n) + \frac{\sigma_n}{\gamma_n}\left(\left(\frac{N_{k_n}}{np_n}\right)^{\gamma_n} - 1\right) \quad \gamma \neq 0.$$

Lien entre p et T la période de retour dans le cas dépendant Rappelons que, d'après le Paragraphe "Lien entre niveaux de retour et quantiles extrêmes de la distribution des données", le niveau de retour de période T correspond au quantile extrême d'ordre $1/(m\mathrm{T})$. Or, le quantile extrême d'ordre r dans le cas dépendant correspond au quantile d'ordre r/ω dans le cas indépendant d'après le Paragraphe "Approche des maxima par bloc". En conséquence, le niveau de retour de période T correspond au quantile extrême d'ordre $1/(m\omega\mathrm{T})$ dans le cas dépendant.

Chapitre 2

Etude de l'erreur d'extrapolation associée à l'estimation des quantiles extrêmes

Sommaire

Résumé

Nous étudions dans ce chapitre le comportement asymptotique de "l'erreur d'extrapolation" (relative) associée avec plusieurs estimateurs des quantiles extrêmes basés sur la théorie des valeurs extrêmes. Cette étude est l'objet d'un article soumis pour publication, ALBERT *et collab. [2018b]. La Partie 2.1 propose une définition de ladite erreur. Une introduction des notations ainsi que la présentation des différentes motivations y sont menées. Un résumé en français de nos contributions y est également rédigé. L'article en question constitue la Partie 2.2 de ce chapitre. Nous y montrons que l'erreur d'extrapolation peut s'interpréter comme le reste d'un développement de Taylor d'ordre un. Des conditions nécessaires et suffisantes sont alors données de telle sorte que l'erreur d'extrapolation relative tende vers zéro quand la taille de l'échantillon augmente. Ceci est fait pour deux estimateurs que sont l'estimateur Exponential Tail dédié au domaine d'attraction de Gumbel et l'estimateur de Weissman dédié au domaine d'attraction de Fréchet. Dans le cas du premier, nous montrons que les précédentes conditions mènent à une sous-division du domaine d'attraction de Gumbel en trois parties. Contrairement à ce dernier, nous montrons que l'erreur d'extrapolation relative associée à l'estimateur de Weissman présente un comportement commun sur tout le domaine d'attraction de Fréchet. Des équivalents de l'erreur d'extrapolation relative sont alors donnés et illustrés numériquement. Enfin, la Partie 2.3 proposent quelques perspectives à donner à ces travaux.*

2.1 Motivations

Nous nous intéressons dans ce chapitre aux limites d'extrapolation dans le cadre de l'estimation des quantiles extrêmes. Rappelons qu'un quantile extrême est défini par (cf Définition 9)

$$q(p_n) := \overline{F}^{\leftarrow}(p_n) = \inf\{x : \overline{F}(x) \leq p_n\} \quad \text{avec} \quad p_n \to 0, \tag{2.1}$$

quand n tend vers l'infini, où n représente la taille de l'échantillon et $\overline{F}^{\leftarrow}$ l'inverse généralisée de $\overline{F}$, la fonction de survie. Définissons l'erreur relative, c'est à dire de la différence renormalisée entre le vrai quantile $q(p_n)$ et un estimateur de ce dernier $\hat{q}(p_n)$:

$$\epsilon(p_n) := \frac{q(p_n) - \hat{q}(p_n)}{q(p_n)}.$$

Décomposition de l'erreur Il est possible de montrer que l'erreur $\epsilon(p_n)$ se décompose comme la somme de deux termes :

$$\epsilon(p_n) \quad = \quad \epsilon_{est}(p_n) + \epsilon_{ext}(p_n),$$

avec

$$\epsilon_{est}(p_n) \quad := \quad \frac{\tilde{q}(p_n) - \hat{q}(p_n)}{q(p_n)}, \tag{2.2}$$

$$\epsilon_{ext}(p_n) \quad := \quad \frac{q(p_n) - \tilde{q}(p_n)}{q(p_n)} \tag{2.3}$$

et $\tilde{q}(p_n)$ une quelconque approximation quantile issue de la théorie des valeurs extrêmes (cf (1.30), (1.31), (1.37), (1.38), (1.40), (1.48), (1.54), (1.59) et (3.3)).

Le premier terme de cette décomposition, $\epsilon_{est}(p_n)$ (cf (2.2)), est une erreur d'estimation aléatoire. Il correspond à l'erreur due à l'estimation des paramètres dans les approximations précédentes. Dans le cas des approximations (1.30) et (1.37) par exemple, ce terme correspond à l'erreur d'estimation des paramètres a_n, b_n et γ par les différentes méthodes d'estimation que sont le maximum de vraisemblance, la méthode des moments ou encore des moment pondérés (cf Paragraphes 1.3.2 et 1.3.3). Dans le cas des approximations (1.48) et (1.54), il correspond à l'erreur associée à l'estimation de $q(\alpha_n)$ (avec α_n un ordre intermédiaire, cf Paragraphe 1.3.1) et à celle du paramètre de forme γ ou β (cf Paragraphe 1.3.4). De manière générale, ce terme d'erreur (renormalisé ou non) est intensivement étudié dans la littérature : citons DE HAAN et ROOTZÉN [1993], DIEBOLT et GIRARD [2003] et GOMES et PESTANA [2007]. C'est aussi un terme d'erreur bien connu dans la pratique et qui est géré à l'aide d'intervalles de confiance sur les paramètres et les quantiles extrêmes estimés (cf COLES et collab. [2001], Paragraphes 3.3.3 et 4.3.3).

La deuxième erreur (cf (2.3)) est une erreur déterministe due à l'approximation - basée sur la théorie des valeurs extrêmes - du vrai quantile par des quantités faisant intervenir les paramètres mentionnés ci-dessus (a_n, b_n, etc). Parmi ces quantités, citons encore une fois (1.30), (1.31), (1.37), (1.38), (1.40), (1.48), (1.54), (1.59) et (3.3). Grossièrement parlant, cette erreur quantifie à quel point le maximum a convergé vers sa loi limite dans le cas d'une approximation basée sur le Théorème 2. D'un point de vue des approximations utilisant le Théorème 3, cette erreur quantifie la convergence des excès vers une loi de Pareto généralisée. Dans tous les cas, elle représente une erreur qui est le plus souvent négligée en pratique. Des travaux théoriques se sont tout de même intéressés à cette dernière : citons par exemple GIRARD et DIEBOLT [1999] ou encore BEIRLANT et collab. [2003].

Erreur d'extrapolation Dans ce chapitre, nous nous intéressons plus particulièrement à l'erreur de deuxième type, que nous appelons par la suite erreur d'extrapolation. La Figure 1.5 illustre l'intérêt que l'on peut porter à l'étude d'une telle erreur. Rappelons que cette figure représente l'approximation quantile (en rouge) obtenue par l'approche GPD de la vraie fonction quantile d'une loi de Weibull (représentée en noir), appartenant au domaine d'attraction de Gumbel. Ainsi, la différence entre les deux courbes sur cette Figure correspond au numérateur dans (2.3). On pourrait s'attendre à ce que cette différence tende vers zéro lorsque l'ordre du quantile, p_n, en abcisse, devient petit. C'était ce qu'il se passait sur la Figure 1.4 avec l'approximation de la fonction de survie. Cependant, sur la Figure 1.5, il n'en est rien. On voit très légèrement apparaître sur la droite du graphe un début de divergence entre les courbes rouge et noire. L'approximation correspondant à la courbe en rouge sur la Figure 1.5 ne semble donc pas appropriée dans le cas d'une loi de Weibull. D'où les questions : dans quel cas l'approximation quantile converge-t-elle vers la vraie fonction quantile ? Qu'est ce qui nous garantit que l'erreur d'extrapolation tende vers zéro ?

Dans la suite du Chapitre, nous répondons à cette question en caractérisant le comportement asymptotique de l'erreur d'extrapolation. Le but est de quantifier jusqu'où il est possible d'extrapoler de manière consistante. Plus spécifiquement, nous donnons des conditions nécessaires et suffisantes sur le couple (p_n, α_n) de telle sorte que l'erreur d'extrapolation relative tende vers zéro quand n tend vers l'infini.

Ceci est fait pour deux approximations quantiles que sont l'approximation Exponential Tail dédiée au domaine d'attraction de Gumbel (cf (1.40)) et l'approximation Weissman (cf (1.48)) dédiée au domaine d'attraction de Fréchet. Nous montrons que dans le premier cas, les conditions que nous obtenons sur l'erreur mènent à une division du domaine d'attraction de Gumbel en trois sous-familles distinctes. Dans le cas de l'approximation Weissman, il n'en est rien, l'erreur présente le même comportement sur tout le domaine d'attraction de Fréchet.

Plus généralement, nous montrons qu'il existe un cadre unifié permettant de traiter les deux erreurs précédentes. Nous en détaillons l'idée ci-dessous :

2.1.1 L'approximation Exponential Tail

Rappelons que l'approximation Exponential tail (cf (1.40)) correspond à l'approximation GPD (cf (1.37)) dans le cas où $\gamma = 0$:

$$q(p_n) \approx q(\alpha_n) + \sigma_n \log\left(\frac{\alpha_n}{p_n}\right).$$

Dans ce cas précis, l'erreur d'extrapolation s'écrit :

$$\varepsilon_{\mathrm{ET}}(p_n;\alpha_n) := \varepsilon_{ext}(p_n) = \frac{q(p_n) - q(\alpha_n) - \sigma_n \log\left(\frac{\alpha_n}{p_n}\right)}{q(p_n)}. \tag{2.4}$$

L'idée est alors de réécrire l'erreur précédente en fonction de la fonction de hasard cumulée définie par $\mathrm{H}(\cdot) := -\log(1 - \mathrm{F}(\cdot))$. Moyennant cette définition, il est aisé de voir que

$$q(x) = \mathrm{H}^{\leftarrow}(-\log x)$$

et donc :

$$\varepsilon_{\mathrm{ET}}(p_n;\alpha_n) = \frac{\mathrm{H}^{\leftarrow}(\log 1/p_n) - \mathrm{H}^{\leftarrow}(\log 1/\alpha_n) - \sigma_n \log\left(\frac{\alpha_n}{p_n}\right)}{\mathrm{H}^{\leftarrow}(\log 1/p_n)}.$$

En posant alors

$$x_n = \log(1/\alpha_n),$$

$$y_n = \log(1/p_n)$$

et

$$\sigma_n = \left(\mathrm{H}^{\leftarrow}\right)'(-\log\alpha_n),$$

il vient :

$$\varepsilon_{\mathrm{ET}}(p_n;\alpha_n) := \frac{\mathrm{H}^{\leftarrow}(y_n) - \mathrm{H}^{\leftarrow}(x_n) - (y_n - x_n)\left(\mathrm{H}^{\leftarrow}\right)'(x_n)}{\mathrm{H}^{\leftarrow}(y_n)}.$$

Nous reconnaissons là un développement de Taylor d'ordre un de la fonction $\mathrm{H}^{\leftarrow}$. L'erreur s'interprète donc comme un reste d'ordre deux. Ce reste d'ordre deux est alors étudié dans le cadre du modèle proposé par DE VALK [2016b] (cf Paragraphe 1.3.4), en supposant que $\mathrm{V}(\cdot) := \log \mathrm{H}^{\leftarrow}(\cdot)$ est à variation régulière étendue d'ordre θ_1 (cf Paragraphe 1.2.3), ou plus précisément en supposant que la dérivée de la fonction $\log \mathrm{H}^{\leftarrow}$ est variation régulière d'ordre θ_1 (qui est une condition plus forte que le modèle proposé par DE VALK [2016b]).

Nous montrons par ailleurs (voir les Propositions 3 et 4 du Paragraphe 2.2) que cette hypothèse, bien que plus forte que celle proposée dans le cadre du modèle de DE VALK [2016b], n'est nullement restrictive en terme de couverture de lois. Un résumé graphique de ces propositions, montrant les différents domaines d'attraction et familles de lois vérifiant l'hypothèse en question, est proposé Figure 2.1.

2.1.2 L'approximation Weissman

Rappelons que l'approximation Weissman (cf (1.48)) est donnée par :

$$q(p_n) \simeq q(\alpha_n)\left(\frac{\alpha_n}{p_n}\right)^{\gamma}.$$

L'erreur d'extrapolation associée s'écrit par conséquent comme suit :

$$\begin{aligned}\varepsilon_W(p_n;\alpha_n) := \epsilon_{ext}(p_n) &= 1-\frac{q(\alpha_n)}{q(p_n)}\left(\frac{\alpha_n}{p_n}\right)^\gamma \\ &= 1-\exp\left(\log q(\alpha_n)-\log q(p_n)+\gamma\log\left(\frac{\alpha_n}{p_n}\right)\right) \\ &= 1-\exp\left(\frac{\log q(\alpha_n)-\log q(p_n)+\gamma(\log\alpha_n-\log p_n)}{\log q(p_n)}\log q(p_n)\right).\end{aligned}$$

Contrairement au paragraphe précédent, où la fonction d'intérêt était la fonction $H^{\leftarrow}$, l'idée est ici de réécrire l'erreur précédente comme fonction de $\log H^{\leftarrow}$. En remarquant que $H^{\leftarrow}(x) = q(e^{-x})$, il vient :

$$\varepsilon_W(p_n;\alpha_n) = 1-\exp\left(\frac{\log H^{\leftarrow}(-\log\alpha_n)-\log H^{\leftarrow}(-\log p_n)+\gamma(\log\alpha_n-\log p_n)}{\log H^{\leftarrow}(-\log p_n)}\log q(p_n)\right).$$

En posant finalement

$$x_n = \log(1/\alpha_n),$$
$$y_n = \log(1/p_n)$$

et

$$\gamma = \gamma_n = \left(\log H^{\leftarrow}\right)'(x_n),$$

l'erreur d'extrapolation s'écrit :

$$\varepsilon_W(p_n;\alpha_n) = 1-\exp\left(-\frac{\log H^{\leftarrow}(y_n)-\log H^{\leftarrow}(x_n)-\left(\log H^{\leftarrow}\right)'(x_n)(y_n-x_n)}{\log H^{\leftarrow}(y_n)}\log q(p_n)\right).$$

Il est là encore possible de reconnaître un développement de Taylor à l'ordre un, mais cette fois-ci de la fonction $\log H^{\leftarrow}$. Ce terme est alors étudié sous l'hypothèse plus classique d'appartenance au domaine d'attraction de Fréchet, voir Théorème 10 du Paragraphe 1.2.4.

De manière générale, nous montrons qu'il existe un cadre permettant d'étudier de manière unifiée les erreurs $\varepsilon_{ET}(p_n;\alpha_n)$ et $\varepsilon_W(p_n;\alpha_n)$ associées respectivement aux approximations ET et Weissman, moyennant la définition d'une fonction d'intérêt φ, jouant tour à tour le rôle des fonctions $H^{\leftarrow}$ et $\log H^{\leftarrow}$.

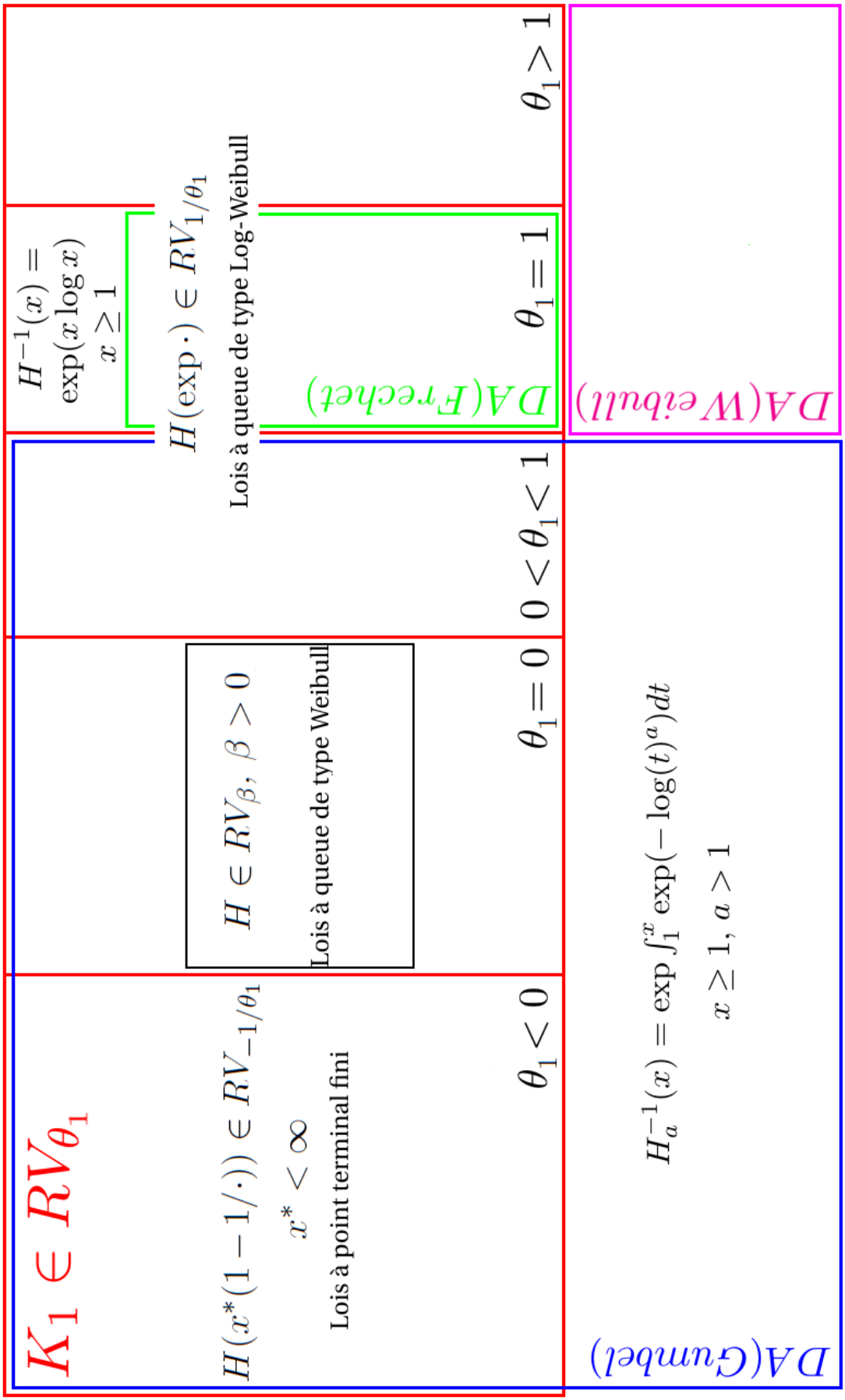

FIGURE 2.1 – Domaine d'applicabilité de l'hypothèse **(A3)** : $K_1 \in RV_{\theta_1}$. Notations définies Paragraphe 2.2 du Chapitre 2.

2.2 Comportement asymptotique de l'erreur d'extrapolation associée à l'estimation des quantiles extrêmes

Les résulats de cette étude sont présentés ci-dessous sous la forme d'un article soumis pour publication, voir ALBERT et collab. [2018b].

Asymptotic behavior of the extrapolation error associated with the estimation of extreme quantiles

Clément Albert[(1)], Anne Dutfoy[(2)] and Stéphane Girard[(1, *)]

[(1)] *Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France*

[(2)] *EDF R&D dept. Périclès, 91120 Palaiseau, France*

Abstract

We investigate the asymptotic behavior of the (relative) extrapolation error associated with some estimators of extreme quantiles based on extreme-value theory. It is shown that the extrapolation error can be interpreted as the remainder of a first order Taylor expansion. Necessary and sufficient conditions are then provided such that this error tends to zero as the sample size increases. Interestingly, in case of the so-called Exponential Tail estimator, these conditions lead to a subdivision of Gumbel maximum domain of attraction into three subsets. In constrast, the extrapolation error associated with Weissman estimator has a common behavior over the whole Fréchet maximum domain of attraction. First order equivalents of the extrapolation error are then derived and their accuracy is illustrated numerically.

Keywords: Extrapolation error, Extreme quantiles, Extreme-value theory.

AMS 2000 subject classification: 62G32, 62G20.

1 Introduction

The starting point of this work is the study of the asymptotic behavior of the Exponential Tail (ET) estimator, a nonparametric estimator of the extreme quantiles from an unknown distribution. Theoretical developments can be found in [5] while numerical aspects are investigated in [11]. Given a n-sample $X_1, \ldots, X_n$ from a cumulative distribution function F with associated survival distribution function $\bar{F}$, an extreme quantile is a $(1-p_n)$th quantile $q(p_n)$ of F essentially larger than the maximal observation, *i.e.* such that $\bar{F}(q(p_n)) = p_n$ with $p_n \leq 1/n$. The estimation of extreme quantiles requires specific methods. Among them, the Peaks Over Threshold (POT) method relies on an approximation of the distribution of excesses over a given threshold [25]. More precisely, let us introduce a deterministic threshold u_n such that $\bar{F}(u_n) = \alpha_n$ or equivalently $u_n = q(\alpha_n)$ with $\alpha_n \to 0$ and $n\alpha_n > 1$ as $n \to \infty$. The excesses above u_n are defined as $Y_i = X_i - u_n$ for all $X_i > u_n$. The survival distribution function of an excess is given

*Corresponding author, Stephane.Girard@inria.fr

by $\bar{F}_{u_n}(x) = \bar{F}(u_n + x)/\bar{F}(u_n)$. Pickands theorem [14, 24] states that, under mild conditions, $\bar{F}_{u_n}$ can be approximated by a Generalized Pareto Distribution (GPD). As a consequence, the extreme quantile $q(p_n)$ can be in turn approximated by the deterministic term

$$\tilde{q}_{\mathrm{GPD}}(p_n; \alpha_n) = q(\alpha_n) + \frac{\sigma_n}{\gamma_n}\left[\left(\frac{\alpha_n}{p_n}\right)^{\gamma_n} - 1\right], \tag{1}$$

where σ_n and γ_n are respectively the scale and shape parameters of the GPD distribution. Then, the POT method consists in estimating these two unknown parameters. The ET method corresponds to the important particular case where F belongs to Gumbel Maximum Domain of Attraction, MDA(Gumbel). In such a situation, $\gamma_n = 0$ and the GPD distribution reduces to an Exponential distribution with scale parameter σ_n. Thus, approximation (1) can be rewritten as

$$\tilde{q}_{\mathrm{ET}}(p_n; \alpha_n) = q(\alpha_n) + \sigma_n \log(\alpha_n/p_n) \tag{2}$$

and the associated estimator [5] is

$$\hat{q}_{\mathrm{ET}}(p_n; \alpha_n) = \hat{q}(\alpha_n) + \hat{\sigma}_n \log(\alpha_n/p_n)$$

where $\hat{q}(\alpha_n) = X_{n-k_n+1,n}$ with $k_n = \lfloor n\alpha_n \rfloor$ and

$$\hat{\sigma}_n = \frac{1}{k_n}\sum_{i=1}^{k_n} X_{n-i+1,n} - X_{n-k_n+1,n}.$$

Let us recall that $X_{1,n} \leq \ldots \leq X_{n,n}$ denote the order statistics associated with $X_1, \ldots, X_n$. The error $(q(p_n) - \hat{q}_{\mathrm{ET}}(p_n; \alpha_n))$ can be expanded as a sum of two terms:

$$q(p_n) - \hat{q}_{\mathrm{ET}}(p_n; \alpha_n) = (\tilde{q}_{\mathrm{ET}}(p_n; \alpha_n) - \hat{q}_{\mathrm{ET}}(p_n; \alpha_n)) + (q(p_n) - \tilde{q}_{\mathrm{ET}}(p_n; \alpha_n)),$$

the first one being a random estimation error

$$\tilde{q}_{\mathrm{ET}}(p_n; \alpha_n) - \hat{q}_{\mathrm{ET}}(p_n; \alpha_n) = q(\alpha_n) - \hat{q}(\alpha_n) + (\sigma_n - \hat{\sigma}_n)\log(\alpha_n/p_n) \tag{3}$$

and the second one being a deterministic extrapolation error

$$q(p_n) - \tilde{q}_{\mathrm{ET}}(p_n; \alpha_n) = q(p_n) - q(\alpha_n) - \sigma_n \log(\alpha_n/p_n). \tag{4}$$

The asymptotic behavior of the estimation error (3) is driven by the asymptotic distributions of $\hat{q}(\alpha_n)$ and $\hat{\sigma}_n$ which can be found for instance in [12] or in [7], Theorem 2.4.1 and Theorem 3.4.2 respectively.

In this paper, we focus on the asymptotic behavior of the extrapolation error (4). Indeed, in view of (2), it appears that the ET method extrapolates in the distribution tail from $q(\alpha_n)$ to $q(p_n)$ thanks to an additive correction proportional to $\log(\alpha_n/p_n)$. Our goal is thus to quantify to what extent this extrapolation can be performed in a consistent way. More specifically, we provide necessary and sufficient conditions on the pair (p_n, α_n) such that the relative extrapolation error

$$\varepsilon_{\mathrm{ET}}(p_n; \alpha_n) := (q(p_n) - \tilde{q}_{\mathrm{ET}}(p_n; \alpha_n))/q(p_n) \tag{5}$$

tends to zero as $n \to \infty$. These conditions depend on the underlying distribution function F and they lead to a subdivision of MDA(Gumbel) into three sub-domains depending on the restrictions they impose on the extrapolation range. Related works include [6, 20] who exhibited penultimate approximations for F^n together with convergence rates for distributions in MDA(Gumbel). These results were extended to other maximum domains of attraction in [21, 22] while penultimate approximations were established for the distribution of the excesses [27]. The relative extrapolation error induced by the approximation of $\bar{F}_{u_n}$ by a the survival distribution function of a GPD is studied in [3].

Here, similarly to [3], we focus on the approximation of quantiles rather than approximations of distribution functions. Let us also highlight that these investigations are not limited to the ET method. To illustrate this, let us introduce $x(n) = \log(1/\alpha_n)$, $y(n) = \log(1/p_n)$ and $\varphi(\cdot) = (\bar{F})^{-1}(1/\exp(\cdot))$. The extrapolation error (4) can thus be interpreted as the remainder of a first order Taylor expansion:

$$q(p_n) - \tilde{q}_{\mathrm{ET}}(p_n; \alpha_n) = \varphi(y(n)) - \varphi(x(n)) - \sigma_n(y(n) - x(n)) \text{ where } \sigma_n = \varphi'(x(n)). \tag{6}$$

We shall show that Weissman estimator [26] dedicated to MDA(Fréchet) can also enter this framework thanks to adapted definitions of functions x, y and φ. In this case, the necessary and sufficient conditions on the extrapolation range are automatically fulfilled for most distributions in MDA(Fréchet) which is a very different situation from MDA(Gumbel).

The paper is organized as follows: The asymptotic behavior of the remainder associated with the first order Taylor expansion (6) is investigated in Section 2. The applications to ET and Weissman approximations are detailed in Section 3 and Section 4 respectively. As a conclusion, some numerical illustrations are presented in Section 5. Proofs are postponed to Section 6 and auxiliary results can be found in the Appendix.

2 Theoretical framework

The following functions are introduced.

(A1) x and y are two functions $\mathbb{R}_+ \to \mathbb{R}_+$ such that $0 < x(t) \leq y(t)$ for t large enough, $x(t) \to \infty$ as $t \to \infty$ and $0 < \liminf_{t\to\infty} x(t)/y(t) \leq \limsup_{t\to\infty} x(t)/y(t) \leq 1$.

(A2) φ is a twice differentiable, increasing function.

Motivated by (5) and (6), we introduce

$$\Delta(t) = \frac{\varphi(y(t)) - \varphi(x(t)) - (y(t) - x(t))\varphi'(x(t))}{\varphi(y(t))}, \tag{7}$$

for all $t > 0$. The goal of this section is to establish necessary and sufficient conditions on $\delta(t) := (y(t) - x(t))/y(t)$ so that $\Delta(t) \to 0$ as $t \to \infty$. The following two functions are of the utmost importance in this study:

$$K_1(s) = \frac{s\varphi'(s)}{\varphi(s)},\ K_2(s) = \frac{s^2\varphi''(s)}{\varphi(s)},\ s \geq 0.$$

The study of Δ relies on the assumption that K_1 is regularly-varying at infinity with index $\theta_1 \leq 1$. This property is denoted for short by

(A3) $K_1 \in RV_{\theta_1}$, $\theta_1 \leq 1$

and means that K_1 is ultimately positive such that

$$\frac{K_1(ts)}{K_1(s)} \to t^{\theta_1} \text{ as } s \to \infty \text{ for all } t > 0.$$

We refer to [4] for a general account on regular variation theory. This assumption is discussed in Section 3 and Section 4 while applying this general framework to the particular cases of ET and Weissman estimators. Finally, a monotonicity assumption is also considered:

(A4) K_1' is ultimately monotone.

Under **(A4)**, K_1 is also ultimately monotone and therefore the limits of $K_1(s)$ and $K_2(s)$ when $s \to \infty$ exist in $\bar{\mathbb{R}}$. The following notations are thus introduced:

$$\lim_{s\to\infty} K_1(s) = \ell_1 \in \bar{\mathbb{R}}_+ \text{ and } \lim_{s\to\infty} K_2(s) = \ell_2 \in \bar{\mathbb{R}}.$$

We are now in position to state our first main result:

Proposition 1 (Role of ℓ_1 for $\Delta \to 0$) *Suppose* **(A1)**–**(A4)** *hold.*

(i) If $\ell_1 \in \{0, 1\}$ then $\ell_2 = 0$ and $\Delta(t) \to 0$ as $t \to \infty$.

(ii) If $\ell_1 \in (0, \infty) \setminus \{1\}$ then $\ell_2 \in (0, \infty)$ and $\Delta(t) \to 0$ if and only if $\delta(t) \to 0$ as $t \to \infty$.

(iii) If $\ell_1 = \infty$ then $|\ell_2| = \infty$ and $\Delta(t) \to 0$ if and only if $\delta^2(t)K_2(y(t)) \to 0$ as $t \to \infty$.

Three cases appear. If $\ell_1 \in \{0, 1\}$ then $\Delta(t) \to 0$ as $t \to \infty$ as soon as **(A1)** holds. If $0 < \ell_1 < \infty$ and $\ell_1 \neq 1$ then a necessary and sufficient condition for $\Delta(t) \to 0$ is $\delta(t) \to 0$ as $t \to \infty$. If $\ell_1 = \infty$ then the necessary and sufficient condition for $\Delta(t) \to 0$ is $\delta^2(t)K_2(y(t)) \to 0$ as $t \to \infty$. Clearly, this condition implies $\delta(t) \to 0$ since, in this situation, $|\ell_2| = \infty$.

Finally, letting $c(a, b) = \int_0^1 (1 - au)^{b-2} u du$, $a \geq 0$, $b \geq 0$, first order approximations of Δ can be provided in each situation.

Proposition 2 (First order approximations of Δ) *Suppose* **(A1)**–**(A4)** *hold.*

(i) Assume $\ell_1 \in \{0, 1\}$ (and thus $\ell_2 = 0$).

(a) If $\delta(t) \to 0$ as $t \to \infty$, then

$$\Delta(t) \sim \delta^2(t) \int_0^1 K_2(y(t)(1 - \delta(t)u)) u du \text{ as } t \to \infty.$$

(b) If $\delta(t) \to \delta_\infty \in (0,1)$ *as* $t \to \infty$*, then*

$$\Delta(t) \sim \delta_\infty^2 \int_0^1 K_2(y(t)(1-\delta(t)u))(1-\delta_\infty u)^{\ell_1-2} u du \text{ as } t \to \infty.$$

(ii) Assume $0 < \ell_1 < \infty$ *and* $\ell_1 \neq 1$.

(a) If $\delta(t) \to 0$ *as* $t \to \infty$*, then*

$$\Delta(t) \sim \frac{\ell_1(\ell_1-1)}{2}\delta^2(t) \text{ as } t \to \infty.$$

(b) If $\delta(t) \to \delta_\infty \in (0,1)$ *as* $t \to \infty$*, then*

$$\Delta(t) \to c(\delta_\infty, \ell_1)\ell_1(\ell_1-1)\delta_\infty^2 \text{ as } t \to \infty.$$

(iii) Assume $\ell_1 = \infty$.

(a) If $\delta(t)K_1(y(t)) \to 0$ *as* $t \to \infty$*, then*

$$\Delta(t) = \frac{1}{2}\delta^2(t)K_1^2(y(t)) \sim \frac{1}{2}\delta^2(t)K_2(y(t)) \text{ as } t \to \infty.$$

(b) If $\delta(t)K_1(y(t)) \to a \in (0,\infty]$ *as* $t \to \infty$*, then*

$$\Delta(t) \to \int_0^a u\exp(-u)du \text{ as } t \to \infty.$$

In situation (i) where $\ell_1 \in \{0,1\}$, $\Delta \to 0$ in both cases $\delta \to 0$ and $\delta \to \delta_\infty \neq 0$, and the convergence is the fastest in the case $\delta \to 0$. In situation (ii) where $0 < \ell_1 < \infty$ and $\ell_1 \neq 1$, Δ is asymptotically proportional to δ^2. In situation (iii) where $\ell_1 = \infty$, $\Delta \to 0$ in the only case where $\delta K_1(y) \to 0$ and Δ is asymptotically proportional to $(\delta K_1(y))^2$ or equivalently to $\delta^2 K_2(y)$.

3 Application to the ET approximation

Recall that $y(n) = \log(1/p_n)$, $x(n) = \log(1/\alpha_n)$ with $0 < p_n \leq 1/n \leq \alpha_n < 1$. Introduce

$$\tau_n = \frac{\log(1/p_n)}{\log(n)} \text{ and } \tau_n' = \frac{\log(1/\alpha_n)}{\log(n)}$$

so that $p_n = n^{-\tau_n}$, $\tau_n \geq 1$, $\alpha_n = n^{-\tau_n'}$, $\tau_n' \leq 1$ and $\delta(n) = (y(n)-x(n))/y(n) = 1 - \tau_n'/\tau_n$. In the sequel, F is assumed to be increasing and twice differentiable and the cumulative hazard rate function is denoted by $H(\cdot) = -\log \bar{F}(\cdot)$. Following the ideas of Section 1, we let $\varphi(\cdot) = (\bar{F})^{-1}(1/\exp(\cdot)) = H^{-1}(\cdot)$ so that $\varepsilon_{\mathrm{ET}}(p_n;\alpha_n) = \Delta(n)$. In this context, the assumption $K_1 \in RV_{\theta_1}$, $\theta_1 \in \mathbb{R}$ is a sufficient condition for H^{-1} is extended regularly varying, see [7], Section B.2 for details on extended regular variation. This assumption has been introduced and discussed by Cees de Valk *et. al.* in a series of papers [8, 9, 10]. The next result provides a characterization of the tail behavior of F according to the sign of θ_1. We refer to [9], Theorem 1 for a characterization under the weaker assumption of extended regular variation.

Proposition 3 (Characterizations)
Suppose F is increasing, twice differentiable and K_1' is ultimately monotone. Let $x^* := \sup\{x : F(x) < 1\}$ be the endpoint of F.

(i) If $H \in RV_\beta$, $\beta > 0$, then $K_1 \in RV_0$ and $\ell_1 = 1/\beta$.

(ii) $K_1 \in RV_{\theta_1}$, $\theta_1 > 0$ (and thus $\ell_1 = \infty$) if and only if $x^* = \infty$ and $H(\exp \cdot) \in RV_{1/\theta_1}$.

(iii) $K_1 \in RV_{\theta_1}$, $\theta_1 < 0$ (and thus $\ell_1 = 0$) if and only if $x^* < \infty$ and $H(x^*(1 - 1/\cdot)) \in RV_{-1/\theta_1}$.

In the case (i) where H is regularly varying with index $\beta > 0$, necessarily $\theta_1 = 0$ and F is referred to as a Weibull tail-distribution, see for instance [2, 16, 19]. Such distributions encompass Gaussian, Gamma, Exponential and strict Weibull distributions. In the case (ii) where $H(\exp \cdot)$ is regularly varying, F is called a log-Weibull tail-distribution, see [1, 13, 18], the most popular example being the lognormal distribution. The case (iii) corresponds to distributions with a Weibull tail behavior in the neighborhood of a finite endpoint.

Besides, let us highlight that the domain of attraction associated with F depends on the position of θ_1 with respect to 1. Note that [9], Proposition 1 provides a similar classification under the weaker assumption of extended regular variation.

Proposition 4 (Domains of attraction)
Suppose F is increasing, twice differentiable and K_1' is ultimately monotone.

(i) If $K_1 \in RV_{\theta_1}$, $\theta_1 < 1$, then $F \in$MDA(Gumbel).

(ii) If $F \in$ MDA(Fréchet) then $K_1 \in RV_1$.

(iii) If $K_1 \in RV_{\theta_1}$, $\theta_1 > 1$, then F does not belong to any domain of attraction.

These results justify the assumption $\theta_1 \leq 1$ introduced in **(A3)**: MDA(Gumbel) is associated with $\theta_1 < 1$ while MDA(Fréchet) is associated with $\theta_1 = 1$. However, there is no perfect one-to-one correspondence as illustrated by the following two examples:

- Consider the distribution defined by $H_a^{-1}(x) = \exp \int_1^x \exp(-\log(t)^a)dt$, $x \geq 1$, $a > 1$. From [7], Corollary 1.1.10, this distribution belongs to MDA(Gumbel) while $K_1(x) = x \exp(-(\log x)^a)$ is not regularly varying.
- Consider the distribution defined by $H^{-1}(x) = \exp(x \log x)$, $x \geq 1$. From [7], Corollary 1.2.10, this distribution does not belong to MDA(Fréchet) while $K_1(x) \sim x \log x$ is regularly varying with index $\theta_1 = 1$.

The situation $\theta_1 > 1$ which does not correspond to any domain of attraction is sometimes referred to as super-heavy tails, see [1] or [4], Section 8.8 for further developments on this topic. Applying Proposition 1 to the ET framework yields:

Theorem 1 (Necessary and sufficient conditions on (α_n, p_n) for $\varepsilon_{\mathbf{ET}}(p_n;\alpha_n) \to 0$) *Suppose F is increasing, twice differentiable and* **(A3)**, **(A4)** *hold. Let $0 < p_n \leq 1/n \leq \alpha_n < 1$ such that $\limsup \delta_n < 1$.*

(i) If $\ell_1 \in \{0,1\}$ then $\varepsilon_{ET}(p_n;\alpha_n) \to 0$ as $n \to \infty$.

(ii) If $\ell_1 \in (0,\infty) \setminus \{1\}$ then $\varepsilon_{ET}(p_n;\alpha_n) \to 0$ if and only if $\tau_n \to 1$ and $\tau_n' \to 1$ as $n \to \infty$.

(iii) If $\ell_1 = \infty$ then $\varepsilon_{ET}(p_n;\alpha_n) \to 0$ if and only if $(\tau_n - \tau_n')^2 K_2(\log n) \to 0$ as $n \to \infty$.

First, if $F \in$MDA(Fréchet) then $\theta_1 = 1$ in view of Proposition 4(ii) and thus $\ell_1 = \infty$. From Theorem 1(iii), it is possible to extrapolate even though the ET method has not be designed for this situation: $\varepsilon_{\mathrm{ET}}(p_n;\alpha_n) \to 0$ under the restriction on (α_n, p_n) that $(\tau_n - \tau_n')^2 K_2(\log n) \to 0$ as $n \to \infty$. Second, it appears that, from the extrapolation error point of view, three sub-domains of MDA(Gumbel) can be exhibited:

- MDA$_1$(Gumbel) defined by $\ell_1 \in \{0,1\}$ and where the relative extrapolation error tends to zero without restriction on the order p_n of the extreme quantile. As illustrated by Proposition 3(iii), the case $\ell_1 = 0$ includes distributions with a finite endpoint. The case $\ell_1 = 1$ encompasses Weibull tail-distributions with shape parameter $\beta = 1$ (Proposition 3(i)), *i.e* close to the Exponential distribution (the Gamma distribution for instance) as well as the class E defined in [6].

- MDA$_2$(Gumbel) defined by $\ell_1 \in (0,\infty) \setminus \{1\}$ and where the relative extrapolation error tends to zero for extreme quantiles close to the maximal observation in the sense that $\log(p_n) \sim \log(1/n)$ as $n \to \infty$. Extreme orders such as $p_n = n^{-\tau}$, $\tau > 1$ are thus not permitted. As illustrated by Proposition 3(i), this situation encompasses Weibull tail-distributions with shape parameter $\beta \neq 1$ *i.e* far from the Exponential distribution (the Gaussian distribution for instance).

- MDA$_3$(Gumbel) defined by $\ell_1 = \infty$ and where the relative extrapolation error tends to zero under strong restrictions on the order p_n of the extreme quantile: $\log(p_n)/\log(1/n) = 1 + o(|K_2(\log n)|^{1/2})$ as $n \to \infty$. As illustrated by Proposition 3(ii), this case corresponds to log-Weibull tail-distributions (including the lognormal distribution for instance).

We refer to Table 1 for examples of distributions in each sub-domain. Note that these three sub-domains do not cover the whole MDA(Gumbel) since they require the existence of ℓ_1 and thus K_1. To conclude this part, one may obtain first order approximations of the relative extrapolation error $\varepsilon_{\mathrm{ET}}(p_n;\alpha_n)$ thanks to Proposition 2. The results are collected in Theorem 2 below. Remark that the assumption $|K_2|$ is regularly varying is needed only in the case $\ell_1 = 1$, since, in other situations it is a consequence of **(A3)**, see Lemma 2.

Theorem 2 (First order approximations of $\varepsilon_{\mathbf{ET}}(p_n;\alpha_n)$)
Suppose the assumptions of Theorem 1 hold.

(i) Assume $F \in MDA_1(Gumbel)$. If $\ell_1 = 1$, let us suppose that there exists $\theta_2 \leq 0$ such that $|K_2| \in RV_{\theta_2}$.

(a) If $\delta(n) \to 0$ then $\varepsilon_{ET}(p_n;\alpha_n) \sim \frac{1}{2}(\tau_n - \tau_n')^2 K_2(\log n) \sim \frac{1}{2}\delta^2(n)K_2(\log n)$.

(b) If $\delta(n) \to \delta_\infty \in (0,1)$ then $\varepsilon_{ET}(p_n;\alpha_n) \sim \delta_\infty^2 c(\delta_\infty, \ell_1+\theta_2)K_2(\tau_n \log n)$.

(ii) Assume $F \in MDA_2(Gumbel)$

(a) If $\delta(n) \to 0$ then $\varepsilon_{ET}(p_n;\alpha_n) \sim \frac{\ell_1(\ell_1-1)}{2}(\tau_n - \tau_n')^2 \sim \frac{\ell_1(\ell_1-1)}{2}\delta^2(n)$.

(b) If $\delta(n) \to \delta_\infty \in (0,1)$ then $\varepsilon_{ET}(p_n;\alpha_n) \to \delta_\infty^2 \ell_1(\ell_1-1)c(\delta_\infty,\ell_1)$.

(iii) Assume $F \in MDA_3(Gumbel)$

(a) If $\delta(n)K_1(\log n) \to 0$ then $\varepsilon_{ET}(p_n;\alpha_n) \sim \frac{1}{2}(\tau_n - \tau_n')^2 K_1^2(\log n) \sim \frac{1}{2}\delta^2(n)K_2(\log n)$.

(b) If $\delta(n)K_1(\log n) \to a \in (0,\infty]$ then $\varepsilon_{ET}(p_n;\alpha_n) \to \int_0^a u\exp(-u)du$.

Before commenting the asymptotic behavior of $\varepsilon_{\mathrm{ET}}(p_n;\alpha_n)$, we would like to compare our results with [3].

Remark 1 *The extrapolation error associated with the GPD approximation has been studied in [3]. Focusing on MDA(Gumbel), the asymptotic equivalents provided by [3], Theorem 2 can be compared to our results. However, we would like to stress that [3], Theorem 2 only holds in the case where $\delta(n) \to 0$ as $n \to \infty$ and for the particular case of "Weibull type distributions" implying in particular that $\ell_1 \neq 0$. It can be shown that the asymptotic equivalents provided by [3], Theorem 2.1 and Theorem 2.3 coincide with the ones of Theorem 2(i,a) and Theorem 2(iii,a) respectively. However, the first order approximation of $\varepsilon_{ET}(p_n;\alpha_n)$ stated in [3], Theorem 2.2 do not coincide with Theorem 2(ii,a) and seems to be wrong. This can be easily checked on the distribution defined by $H^{-1}(x) = x + 1/x$ where $\varepsilon_{ET}(p_n;\alpha_n) \sim (\delta(n)/\log n)^2$ as $n \to \infty$.*

The only situation where $\delta(n) \to \delta_\infty \neq 0$ and $\varepsilon_{\mathrm{ET}}(p_n;\alpha_n) \to 0$ as $n \to \infty$ occurs for $F \in \mathrm{MDA}_1(\mathrm{Gumbel})$. In this particular case, it is possible to choose extreme orders such that $p_n = n^{-\tau}$, $\tau > 1$, and the relative extrapolation error tends to zero at a logarithmic rate. As expected, in the three situations (i,ii,iii)-(a) where $\delta(n) \to 0$ and $\varepsilon_{\mathrm{ET}}(p_n;\alpha_n) \to 0$ as $n \to \infty$, the convergence is the fastest in $\mathrm{MDA}_1(\mathrm{Gumbel})$ and the slowest in $\mathrm{MDA}_3(\mathrm{Gumbel})$. Let us also highlight that the rate of convergence is independent from the distribution in $\mathrm{MDA}_2(\mathrm{Gumbel})$. To illustrate these results, let us focus on the distributions introduced in Table 1. Clearly, in all six cases, $F \in \mathrm{DA}(\mathrm{Gumbel})$, K_1 and $|K_2|$ are regularly varying so that the assumptions of Theorem 2 are fulfilled. Let us consider the case where $p_n = 1/(n\log n)$ and $\alpha_n = (\log n)/n$ leading to

$$\tau_n = 1 + \frac{\log\log n}{\log n}\,, \tau_n' = 1 - \frac{\log\log n}{\log n} \text{ and } \delta(n) \sim 2\frac{\log\log n}{\log n}, \tag{8}$$

as $n \to \infty$. Let us stress that these choices entail $\delta(n) \to 0$ and $\delta(n)K_1(\log n) \to 0$ as $n \to \infty$ so that Theorem 2(i,ii,iii)-(a) can be applied and $\varepsilon_{\mathrm{ET}}(p_n; \alpha_n) \to 0$ as $n \to \infty$ for all six distributions. The associated first order approximations of $\varepsilon_{\mathrm{ET}}(p_n; \alpha_n)$ are provided in Table 2. It appears that, in most cases, the convergence of the relative extrapolation error to zero is rather slow. The log-Weibull($\beta > 1$) distribution corresponds to the worst case, since arbitrary low rates of convergence can be obtained by letting $\beta \overset{>}{\to} 1$. At the opposite, the Finite endpoint($\beta > 0$) distribution is the most favorable case, letting $\beta \overset{>}{\to} 0$ could lead to arbitrary high logarithmic rates of convergence.

As a conclusion, in MDA(Gumbel), the extrapolation abilites of the ET method are poor. To overcome this limitation, two main approaches are usually considered. The first one is to focus on a subset of distributions, for instance Weibull tail-distributions in MDA_2(Gumbel), where adapted estimators can outperform the ET method, see [15] for an illustration. The second one is to rely on new assumptions on the distribution tail, such as the log-generalized Weibull tail limit [10].

4 Application to Weissman approximation

When $F \in$ MDA(Fréchet), $\gamma_n > 0$ and the GPD approximation (1) can be simplified by letting $\sigma_n = \gamma_n q(\alpha_n)$, see [7], Theorem 1.2.5, leading to

$$\tilde{q}_{\mathrm{W}}(p_n; \alpha_n) = q(\alpha_n) \left(\frac{\alpha_n}{p_n}\right)^{\gamma_n}, \tag{9}$$

which is called Weissman approximation in the sequel. Weissman estimator [26] is then obtained by replacing the intermediate quantile $q(\alpha_n)$ and the tail index γ_n by appropriate estimators:

$$\hat{q}_{\mathrm{W}}(p_n; \alpha_n) = \hat{q}(\alpha_n) \left(\frac{\alpha_n}{p_n}\right)^{\hat{\gamma}_n}.$$

The most common choices are $\hat{q}(\alpha_n) = X_{n-k_n+1,n}$, see Section 1, and Hill estimator [23]:

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

Taking the logarithm of (9) yields

$$\log q(p_n) - \log \tilde{q}_{\mathrm{W}}(p_n; \alpha_n) = \log q(p_n) - \log q(\alpha_n) - \gamma_n \log(\alpha_n / p_n)$$

and thus, similarly to the ET case (4), the extrapolation error can be interpreted as a first order Taylor remainder. To this end, recall that $y(n) = \log(1/p_n)$, $x(n) = \log(1/\alpha_n)$ with $0 < p_n \leq 1/n \leq \alpha_n < 1$ and introduce $\varphi(\cdot) = \log(\bar{F})^{-1}(1/\exp(\cdot)) = \log U(\exp \cdot)$ where U is the tail quantile function, so that

$$\log q(p_n) - \log \tilde{q}_{\mathrm{W}}(p_n; \alpha_n) = \varphi(y(n)) - \varphi(x(n)) - \gamma_n(y(n) - x(n)) \text{ where } \gamma_n = \varphi'(x(n)).$$

The quantity of interest is

$$\varepsilon_{\mathrm{W}}(p_n;\alpha_n) := (q(p_n) - \tilde{q}_{\mathrm{W}}(p_n;\alpha_n))/q(p_n) = 1 - \exp(-\Delta(n)\log q(p_n)), \tag{10}$$

where $\Delta(n)$ is defined in (7). Here, the property $K_1 \in RV_{\theta_1}$ is a direct consequence of the assumption $F \in$ MDA(Fréchet). Indeed, from [7], Corollary 1.2.1, $F \in$ MDA(Fréchet) is equivalent to $U \in RV_\gamma$ for some $\gamma > 0$ which can be rewritten as

$$U(t) = t^\gamma L(t), \text{ with } L \in RV_0 \text{ and } \gamma > 0. \tag{11}$$

The function L is said to be slowly-varying [4]. Classical properties of slowly-varying functions yield, as $t \to \infty$,

$$\begin{aligned} \varphi(t) &= \gamma t + \log L(\exp t) \sim \gamma t, \\ \varphi'(t) &= \gamma + \eta(\exp t) \to \gamma, \end{aligned} \tag{12}$$

where $\eta(t) = tL'(t)/L(t)$ is called the auxiliary function associated with L. It follows that $K_1(t) \to 1$ as $t \to \infty$ and thus $\ell_1 = 1$ and $K_1 \in RV_0$. This means that only the case (i) of Proposition 1 and Proposition 2 has to be considered. In particular $\Delta(n) \to 0$ as $n \to \infty$ without further assumption, which is a very different situation from Section 3:

Theorem 3 (First order approximation of $\varepsilon_{\mathbf{W}}(p_n;\alpha_n)$)
Let $0 < p_n \le 1/n \le \alpha_n < 1$ such that $\limsup \delta_n < 1$. Suppose F is increasing, twice differentiable and (11) holds. Let $\eta(t) = tL'(t)/L(t)$ be the auxiliary function associated with L. Suppose K_1', η are asymptotically monotone and $|\eta| \in RV_\rho$ with $\rho < 0$.

(i) If $\delta(n) \to \delta_\infty \in (0,1)$ then $\varepsilon_{W}(p_n;\alpha_n) \to 0$ as $n \to \infty$.

(ii) Moreover, as $n \to \infty$,

$$\varepsilon_{W}(p_n;\alpha_n) \sim -\frac{\delta_\infty}{1-\delta_\infty}\log(1/\alpha_n)\eta(1/\alpha_n).$$

The assumption $|\eta| \in RV_\rho$, $\rho < 0$, is recurrent in extreme-value statistics to control the bias of estimators, ρ being known as the second-order parameter, see *e.g.* [17]. This assumption holds for most heavy-tailed distributions such as Burr, Fréchet, Pareto or Student distributions. Let us also remark that one can choose extreme orders such that $p_n = n^{-\tau}$, $\tau > 1$ as in MDA$_1$(Gumbel), see Theorem 2(i)-(b), and still obtain $\varepsilon_{\mathrm{W}}(p_n;\alpha_n) \to 0$ as $n \to \infty$. However, here, the relative extrapolation error converges to zero at a polynomial rate, depending on ρ.

5 Numerical illustrations

First, the quality of the first order approximations associated with the ET relative extrapolation error given in Table 2 is assessed graphically. Recall that these results are obtained by applying Theorem 2 to sequences (τ_n) and (τ_n') given in (8) and distributions described in Table 1:

Finite endpoint($\beta = 5$), Gamma($a = 0.1$), Weibull($\beta = 5$), Gaussian, log-Weibull($\beta = 3$) and lognormal($\sigma = 0.5$). The exact relative extrapolation error $\varepsilon_{\mathrm{ET}}(p_n; \alpha_n)$ as well as the corresponding first order approximation provided by Theorem 2 are computed as functions of $\log n$. The results are displayed on Figures 1–3. It appears that, for all six distributions, the relative extrapolation error converges towards zero as predicted by Theorem 2, even though the convergence can be very slow in DA_3(Gumbel), see Figure 3. In all cases, the asymptotic sign of $\varepsilon_{\mathrm{ET}}(p_n; \alpha_n)$ is coherent with the first order equivalent given in Table 2: Positive for Gamma($a < 1$), log-Weibull($\beta > 1$) and lognormal distributions, negative for Finite endpoint($\beta > 0$), Weibull($\beta > 1$) and Gaussian distributions. Finally, the first order equivalent provides a reasonable approximation of the error behavior in all situations.

To conclude, Figure 4 displays the exact relative extrapolation error $\varepsilon_{\mathrm{W}}(p_n; \alpha_n)$ associated with Weissman estimator together with its corresponding first order approximation provided by Theorem 3(ii) as a function of $\log n$. These results are obtained by choosing sequences $p_n = n^{-5/4}$ and $\alpha_n = n^{-3/4}$ such that $\delta(n) = 2/5$ and by considering a Burr distribution defined by $U(t) := (t^{1/k} - 1)^k$, $t \geq 1$, $k > 0$, with extreme-value index $\gamma = 1$ and auxiliary function $\eta(t) = 1/(t^{1/k} - 1)$. Clearly, η is regularly varying with index $\rho = -1/k$. In both cases $k = 3$ (top) and $k = 4$ (bottom), it appears that the relative extrapolation error converges to zero even though $\delta(n)$ is constant. This graphical assessment is in agreement with Theorem 3(i). As expected, both errors are negative since the auxiliary function η is positive. It also appears that, the smaller k, the faster the convergence is. This is in accordance with $\eta \in RV_{-1/k}$. Finally, the first order equivalent also provides a reasonable approximation of the error behavior in the Burr case.

6 Proofs of main results

Proof of Proposition 1. (i) If $\ell_1 = 0$ then Lemma 2(i) shows that $\ell_2 = 0$. If $\ell_1 = 1$ then, from Lemma 2(ii), $\ell_2 = 0$. Lemma 5(i) concludes the proof.
(ii) If $0 < \ell_1 < \infty$ and $\ell_1 \neq 1$ then Lemma 2(iii) entails that ℓ_2 is finite and non zero. Lemma 5(i,ii) concludes the proof.
(iii) If $\ell_1 = \infty$ then $K_2(x) \sim K_1^2(x)$ as $x \to \infty$, see Lemma 2(iv). Besides, K_2 is regularly varying of order $2\theta_1$ and thus $K_2(x(t))$ and $K_2(y(t))$ are of the same order as $t \to \infty$ under **(A1)**. Lemma 5(i,ii) concludes the proof. ∎

Proof of Proposition 2. (i) If $\ell_1 \in \{0, 1\}$ and $\delta(t) \to \delta_\infty \in [0, 1)$ as $t \to \infty$, then the result is a straightforward consequence of Lemma 4.
(ii) Assume $0 < \ell_1 < \infty$ and $\ell_1 \neq 1$. Then Lemma 2(iii) entails $\ell_2 = \ell_1(\ell_1 - 1)$, and Lemma 4 yields

$$\Delta(t) \sim \delta^2(t) \int_0^1 K_2(y(t)(1 - \delta(t)u))(1 - \delta(t)u)^{\ell_1 - 2} u du.$$

When $\delta(t) \to \delta_\infty \in [0,1)$ as $t \to \infty$, Lebesgue's dominated convergence theorem entails

$$\int_0^1 K_2(y(t)(1-\delta(t)u))(1-\delta(t)u)^{\ell_1-2}udu \to \ell_1(\ell_1-1)\int_0^1 (1-\delta_\infty u)^{\ell_1-2}udu,$$

and the result is proved.
(iii) Assume $\ell_1 = \infty$. Then Lemma 2(iv) entails that $K_2(x) \sim K_1^2(x)$ as $x \to \infty$. As a consequence, Lemma 3(i) and Lebesgue's dominated convergence theorem yield

$$\Delta(t) \sim \delta^2(t)\int_0^1 \frac{K_1^2(y(t)(1-\delta(t)u))}{(1-\delta(t)u)^2}\exp\left(K_1(y(t))L_{\theta_1}(1-\delta(t)u)(1+o(1))\right)udu.$$

The regular variation property **(A3)** entails

$$\Delta(t) \sim \delta^2(t)K_1^2(y(t))\int_0^1 (1-\delta(t)u)^{2\theta_1-2}\exp\left(K_1(y(t))(L_{\theta_1}(1-\delta(t)u)(1+o(1))\right)udu.$$

Two main situations are considered:

1. If $\delta(t) \to 0$ as $t \to \infty$, then $L_{\theta_1}(1-\delta(t)u) = -\delta(t)u(1+o(1))$. Letting $A(t) = \delta(t)K_1(y(t))$, it follows

$$\Delta(t) \sim A^2(t)\int_0^1 \exp(-A(t)u(1+o(1)))udu \sim \Phi(A(t)(1+o(1)))A^2(t),$$

with $\Phi(\cdot) = \Psi_1(\cdot;1)$, see Lemma 1. Three sub-cases arise: (a) If $A(t) \to 0$ as $t \to \infty$, then $\Phi(A(t)) \to 1/2$ in view of Lemma 1(i) and

$$\Delta(t) \sim \frac{1}{2}\delta^2(t)K_1^2(y(t)).$$

(b) If $A(t) \to a \in (0,\infty)$ then $\Delta(t) \to a^2\Phi(a) = \int_0^a u\exp(-u)du$ as $t \to \infty$ in view of the continuity of Φ, see Lemma 1(i). If $A(t) \to \infty$, then $\Phi(A(t)) \sim 1/A^2(t)$ from Lemma 1(ii) and therefore $\Delta(t) \to 1 = \int_0^\infty u\exp(-u)du$ as $t \to \infty$.

2. If $\delta(t) \to \delta_\infty \in (0,1)$, then $A(t) \to \infty$ as $t \to \infty$. Two successive change of variables yield

$$\begin{aligned}
\Delta(t) &\sim \delta_\infty^2 K_1^2(y(t))\int_0^1 (1-\delta_\infty u)^{2\theta_1-2}\exp\left(K_1(y(t))L_{\theta_1}(1-\delta(t)u)(1+o(1))\right)udu \\
&\sim K_1^2(y(t))\int_{1-\delta_\infty}^1 (1-v)v^{2\theta_1-2}\exp\left(K_1(y(t))L_{\theta_1}(v)(1+o(1))\right)dv \\
&\sim K_1^2(y(t))\int_{L_{\theta_1}(1-\delta_\infty)}^0 (L_{\theta_1}^{-1}(w))^{\theta_1-1}(1-L_{\theta_1}^{-1}(w))\exp\left(K_1(y(t))w(1+o(1))\right)dw.
\end{aligned}$$

Let us introduce $\xi(w) = (L_{\theta_1}^{-1}(w))^{\theta_1-1}(1-L_{\theta_1}^{-1}(w))$ for all $w \in [L_{\theta_1}(1-\delta_\infty),0]$. Routine calculations show that $\xi(0) = 0$ and $\xi'(0) = -1$. A second order Taylor expansion thus yields $\xi(w) = -w + w^2\xi''(\eta_w)/2$ with $\eta_w \in [w,0] \subset [L_{\theta_1}(1-\delta_\infty),0]$. Replacing, we get

$$\begin{aligned}
\Delta(t) &= -K_1^2(y(t))\int_{L_{\theta_1}(1-\delta_\infty)}^0 w\exp\left(K_1(y(t))w(1+o(1))\right)dw(1+o(1)) + R(t) \\
&= K_1^2(y(t))\Psi_1\left(K_1(y(t))(1+o(1));-L_{\theta_1}(1-\delta_\infty)\right) + R(t),
\end{aligned}$$

where Ψ_1 is defined in Lemma 1 and

$$R(t) = \frac{1}{2}K_1^2(y(t)) \int_{L_{\theta_1}(1-\delta_\infty)}^{0} w^2 \xi''(\eta_w) \exp\left(K_1(y(t))w(1+o(1))\right) dw(1+o(1)).$$

Remarking that $|\xi''|$ is bounded on compact sets, there exists $M > 0$ such that

$$|R(t)| \leq M K_1^2(y(t)) \Psi_2\left(K_1(y(t))(1+o(1)); -L_{\theta_1}(1-\delta_\infty)\right),$$

where Ψ_2 is defined in Lemma 1. As a consequence of Lemma 1(ii), $R(t) = O(1/K_1(y(t)))$ and $\Delta(t) \to 1$ as $t \to \infty$. Let us remark that this case is similar to the situation $\delta(t) \to 0$ and $A(t) \to \infty$. ■

Proof of Proposition 3. (i) If $H \in RV_\beta$, $\beta > 0$ then the monotone density theorem ([4], Proposition 1.7.2) yields

$$H(t) \sim \frac{1}{\beta} t H'(t) \text{ as } t \to \infty.$$

Letting $x = H(t)$, we have

$$x \sim \frac{1}{\beta} \frac{H^{-1}(x)}{(H^{-1})'(x)}$$

or equivalently $K_1(x) \to 1/\beta$ as $x \to \infty$. It follows that $\ell_1 = 1/\beta$ and $K_1 \in RV_0$.
(ii, $\Longleftarrow$) Let us assume that $H(\exp \cdot) \in RV_{1/\theta_1}$, $\theta_1 > 0$. Then, $\log H^{-1} \in RV_{\theta_1}$ and the monotone density theorem ([4], Proposition 1.7.2) implies $(\log H^{-1})' \in RV_{\theta_1 - 1}$ *i.e.* $K_1 \in RV_{\theta_1}$.
(ii, $\Longrightarrow$) Conversely, assume $K_1 \in RV_{\theta_1}$, $\theta_1 > 0$. Then, necessarily $\ell_1 = \infty$. From [4], Theorem 1.5.8, we have for all x_0 sufficiently large,

$$\log H^{-1}(x) - \log H^{-1}(x_0) = \int_{x_0}^{x} (\log H^{-1}(t))' dt = \int_{x_0}^{x} \frac{K_1(t)}{t} dt =\sim \frac{1}{\theta_1} K_1(x), \tag{13}$$

as $x \to \infty$. It is thus clear that $\log H^{-1} \in RV_{\theta_1}$ and therefore $H(\exp \cdot) \in RV_{1/\theta_1}$.
(iii, $\Longleftarrow$) Let us assume that $x^* < \infty$ and $h(\cdot) := H(x^*(1 - 1/\cdot)) \in RV_{-1/\theta_1}$, $\theta_1 < 0$. Consequently, $H^{-1}(\cdot) = x^*(1 - 1/h^{-1}(\cdot))$ where $h^{-1} \in RV_{-\theta_1}$ and $h^{-1}(x) \to \infty$ as $x \to \infty$. Straightforward calculations and the monotone density theorem ([4], Proposition 1.7.2) lead to

$$\begin{aligned}
\log H^{-1}(x) &= \log x^* + \log\left(1 - \frac{1}{h^{-1}(x)}\right) \\
(\log H^{-1})'(x) &= \frac{(h^{-1})'(x)}{h^{-1}(x)(h^{-1}(x) - 1)} \\
K_1(x) &= \frac{x(h^{-1})'(x)}{h^{-1}(x)(h^{-1}(x) - 1)} \sim -\frac{\theta_1}{h^{-1}(x)},
\end{aligned}$$

and therefore $K_1 \in RV_{\theta_1}$, $\theta_1 < 0$.
(iii, $\Longrightarrow$) Conversely, assume $K_1 \in RV_{\theta_1}$, $\theta_1 < 0$. Thus $(\log H^{-1})' \in RV_{\theta_1 - 1}$ and [4], Theorem 1.5.8 yields first, for all x sufficiently large,

$$\log x^* - \log H^{-1}(x) = \int_{x}^{\infty} (\log H^{-1})'(t) dt < \infty$$

and thus $x^* < \infty$. Second, one also has

$$\frac{K_1(x)}{\int_x^\infty (\log H^{-1})'(t)dt} \to -\theta_1$$

as $x \to \infty$. Combining the two above results yield

$$\frac{K_1(x)}{\log x^* - \log H^{-1}(x)} = -\theta_1(1+o(1))$$

and consequently

$$H^{-1}(x) = x^* \exp\left(\frac{1}{\theta_1}K_1(x)(1+o(1))\right) = x^*\left(1+\frac{1}{\theta_1}K_1(x)(1+o(1))\right)$$

since $K_1(x) \to 0$ as $x \to \infty$. Applying [4], Theorem 1.5.12 yields $H(x^*(1-1/\cdot)) \in RV_{-1/\theta_1}$ and concludes the proof. ■

Proof of Proposition 4. (i) Assume $K_1 \in RV_{\theta_1}$, $\theta_1 < 1$ and let $U(\cdot) = H^{-1}(\log \cdot)$ be the tail quantile function. For all $x > 0$ and $t > 0$, consider

$$\frac{U'(tx)}{U'(t)} = \frac{1}{x}\frac{\left(H^{-1}\right)'(\log tx)}{\left(H^{-1}\right)'(\log t)} = \frac{1}{x}\frac{\log t}{\log tx}\frac{H^{-1}(\log tx)}{H^{-1}(\log t)}\frac{K_1(\log tx)}{K_1(\log t)}.$$

Since $K_1 \in RV_{\theta_1}$ and the logarithm is a slowly-varying function, $K_1(\log \cdot) \in RV_0$ and thus

$$\frac{U'(tx)}{U'(t)} = \frac{1}{x}\frac{H^{-1}(\log tx)}{H^{-1}(\log t)}(1+o(1))$$

as $t \to \infty$. Besides,

$$\begin{aligned}
\frac{H^{-1}(\log tx)}{H^{-1}(\log t)} &= \exp(\log H^{-1}(\log tx) - \log H^{-1}(\log t)) \\
&= \exp\left(\int_{\log t}^{\log tx} (\log H^{-1})'(u)du\right) \\
&= \exp\left(\int_{\log t}^{\log tx} \frac{K_1(u)}{u}du\right) \\
&= \exp\left(\log x \int_0^1 \frac{K_1(\log t + v\log x)}{\log t + v\log x}dv\right),
\end{aligned}$$

and the regular variation property of K_1 implies that

$$\frac{K_1(\log t + v\log x)}{\log t + v\log x} = \frac{K_1(\log t)}{\log t}(1+o(1))$$

as $t \to \infty$ uniformly locally on $v \in [0,1]$. It follows that

$$\frac{H^{-1}(\log tx)}{H^{-1}(\log t)} = \exp\left(\log x \frac{K_1(\log t)}{\log t}(1+o(1))\right) \to 1$$

as $t \to \infty$ since $K \in RV_{\theta_1}$ with $\theta_1 < 1$. As a conclusion, $U'(tx)/U'(t) \to 1/x$ as $t \to \infty$ for all $x > 0$ and thus $U' \in RV_{-1}$. This implies that $F \in$ MDA(Gumbel), see [7], Corollary 1.1.10.
(ii) Assume $F \in$ MDA(Fréchet). From [7], Corollary 1.2.10, there exists $\gamma > 0$ such that the tail quantile function $U \in RV_\gamma$. Since $H^{-1}(\cdot) = U(\exp \cdot)$, it follows that

$$K_1(x) = x \frac{\exp(x) U'(\exp x)}{U(\exp x)} \sim \gamma x$$

as $x \to \infty$ from the monotone density theorem [4], Theorem 1.7.2. It is thus clear that $K_1 \in RV_1$.
(iii) Assume $K_1 \in RV_{\theta_1}$, $\theta_1 > 1$. First, Proposition 3(ii) implies that $x^* = \infty$ and thus $F \notin$ MDA(Weibull). Second, Proposition 4(ii) shows that $F \in$ MDA(Fréchet) entails $K_1 \in RV_1$. It is thus clear that $F \notin$ MDA(Fréchet). Finally, it remains to show that $F \notin$ MDA(Gumbel). To this end, consider for all $x > 0$ and $t \to \infty$,

$$\frac{U(tx)}{U(t)} = \frac{H^{-1}(\log tx)}{H^{-1}(\log t)} = \exp\left\{ \frac{1}{\theta_1} \left(K_1(\log tx) - K_1(\log t)\right)(1 + o(1)) \right\}$$

from (13) in the proof of Proposition 3(ii, $\Longrightarrow$). A first order Taylor expansion yields

$$\frac{U(tx)}{U(t)} = \exp\left\{ \frac{\log x}{\theta_1} K_1'(\log t + \eta \log x)(1 + o(1)) \right\} = \exp\left\{ \frac{\log x}{\theta_1} K_1'(\log t)(1 + o(1)) \right\}$$

where $\eta \in (0, 1)$ since $K_1' \in RV_{\theta_1 - 1}$. Recalling that $\theta_1 > 1$, it is then clear that $K_1'(\log t) \to \infty$ as $t \to \infty$ and therefore $U(tx)/U(t) \to 0$ as $t \to \infty$ if $x < 1$ while $U(tx)/U(t) \to \infty$ as $t \to \infty$ if $x > 1$. Finally [7], Lemma 1.2.9 shows that $F \notin$ MDA(Gumbel) since $U(tx)/U(t)$ does not converge to 1 as $t \to \infty$. ∎

Proof of Theorem 1. The proof relies on the application of Proposition 1. Condition **(A1)** is fulfilled under the assumptions $0 < p_n \leq 1/n \leq \alpha_n < 1$ and $\limsup \delta_n < 1$.
(i) is a straightforward consequence of Proposition 1(i).
(ii) is based on the remark that $\delta(n) \to 0$ if and only if $\tau_n \to 1$ and $\tau_n' \to 1$ since, by assumption, $\tau_n' \leq 1 \leq \tau_n$.
(iii) Since $\ell_2 = \infty$, $\delta^2(n) K_2(\tau_n \log n) \to 0$ implies $\delta(n) \to 0$ and thus $\tau_n \to 1$ and $\tau_n' \to 1$ in view of the above remark. Thus, $\delta^2(n) \sim (\tau_n - \tau_n')^2$ as $n \to \infty$. Besides, Lemma 2(iv) entails that K_2 is regularly varying when $\ell_2 = \infty$. As a consequence, $K_2(\tau_n \log n) \sim K_2(\log n)$ as $n \to \infty$ and the result is proved. ∎

Proof of Theorem 2. The proof relies on the application of Proposition 2.
(i) is a consequence of Proposition 2(i). Let us highlight that, when $\delta(n) \to 0$ then $x(n) \sim y(n)$ and thus $K_2(x_n) \sim K_2(y_n) \sim K_2(\log n)$ in view of the property $|K_2|$ is regularly varying. Moreover, $\delta^2(n) \sim (\tau_n - \tau_n')^2$ as already seen in the proof of Theorem 1.
(ii) is a straightforward consequence of Proposition 2(ii).
(iii) is a consequence of Proposition 2(iii). If $\delta(n) K_1(\log n) \to a \in [0, \infty)$ then, necessarily, $\delta(n) \to 0$ and thus $y(n) \sim \log n$ leading to $K_1(y(n)) \sim K_1(\log n)$. In the case where $\delta(n) K_1(\log n) \to \infty$, one still has $1 \leq \liminf \tau_n \leq \limsup \tau_n < \infty$ and thus $K_1(y(n))$ and $K_1(\log n)$ are asymptotically of the same order, the result is proved. ∎

Proof of Theorem 3. The proof relies on the application of Lemma 4 with $\ell_1 = 1$:

$$\Delta(n) \sim \delta^2(n) \int_0^1 K_2\left(y(n)(1-\delta(n)u)\right)(1-\delta(n)u)^{-1} u du.$$

From (12), $\varphi''(t) = \exp(t)\eta'(\exp(t))$ and consequently

$$K_2(t) \sim \frac{1}{\gamma} t \exp(t)\eta'(\exp(t)) \text{ as } t \to \infty.$$

Moreover, η asymptotically monotone and $|\eta| \in RV_\rho$ imply $x\eta'(x)/\eta(x) \to \rho$ as $x \to \infty$, leading to, as $t \to \infty$,

$$K_2(t) \sim \frac{\rho}{\gamma} t\eta(\exp(t)).$$

It follows, when $\delta(n) \to \delta_\infty \in (0,1)$,

$$\Delta(n) \sim y(n)\frac{\delta_\infty^2 \rho}{\gamma} \int_0^1 u\eta\left(e^{y(n)(1-\delta(n)u)}\right) du.$$

Since $|\eta| \in RV_\rho$, Potter's bounds (see for instance [7], Proposition B.1.9 (5.)) entail that there exists $0 < \epsilon < |\rho|$ such that

$$(1-\epsilon)e^{y(n)\delta(n)(1-u)(\rho-\epsilon)} \leq \frac{|\eta|\left(e^{y(n)(1-\delta(n)u)}\right)}{|\eta|\left(e^{x(n)}\right)} \leq (1+\epsilon)e^{y(n)\delta(n)(1-u)(\rho+\epsilon)}.$$

Recalling that η is asymptotically monotone with a constant sign yields

$$\Delta(n) \sim \frac{\delta_\infty^2 \rho}{\gamma} \eta\left(e^{x(n)}\right) I_n y_n,$$

where $I_n^- \leq I_n \leq I_n^+$ and

$$I_n^- = (1-\epsilon)\int_0^1 u e^{y(n)\delta(n)(1-u)(\rho-\epsilon)} du$$
$$I_n^+ = (1+\epsilon)\int_0^1 u e^{y(n)\delta(n)(1-u)(\rho+\epsilon)} du.$$

Straightforward calculations show that, for all $x < 0$,

$$\int_0^1 u e^{y(n)\delta(n)(1-u)x} du \sim -\frac{1}{y(n)\delta_\infty x},$$

as $n \to \infty$, since $\delta(n) \to \delta_\infty \in (0,1)$ and thus $y(n)\delta(n) \to \infty$ as $n \to \infty$.
Consequently, $I_n^- \sim (1-\epsilon)/(y(n)\delta_\infty(\epsilon-\rho))$, $I_n^+ \sim (1+\epsilon)/(y(n)\delta_\infty(-\epsilon-\rho))$ and thus $I_n y(n) \in \left[\frac{(1-\epsilon)}{\delta_\infty(\epsilon-\rho)}(1+o(1)); \frac{(1+\epsilon)}{\delta_\infty(-\epsilon-\rho)}(1+o(1))\right]$. Letting $\epsilon \to 0$ entails $I_n y(n) \to -1/(\rho\delta_\infty)$ as $n \to \infty$ and thus

$$\Delta(n) \sim -\frac{\delta_\infty}{\gamma}\eta\left(e^{x(n)}\right).$$

Remarking that $\log q(p_n) = \varphi(y(n)) \sim \gamma y(n) \sim \frac{\gamma}{1-\delta_\infty} x(n)$ and taking account of (10) yield, when $\delta(n) \to \delta_\infty \in (0,1)$,

$$\varepsilon_{\mathrm{W}}(p_n;\alpha_n) = 1 - \exp\left(-\frac{\gamma}{1-\delta_\infty}\Delta(n)x(n)(1+o(1))\right).$$

Finally, since $|\eta| \in RV_\rho$, $\rho < 0$, $\Delta(n)x(n) \sim -\dfrac{\delta_\infty}{\gamma} x(n)\eta(e^{x(n)}) \to 0$ as $n \to \infty$ and thus

$$\varepsilon_{\mathrm{W}}(p_n; \alpha_n) \sim -\frac{\delta_\infty}{1-\delta_\infty} x(n)\eta(e^{x(n)}) \sim -\frac{\delta_\infty}{1-\delta_\infty} \log(1/\alpha_n)\eta\left(1/\alpha_n\right),$$

which concludes the proof of (i) and (ii). ■

References

[1] I. Alves, L. de Haan, and C. Neves. A test procedure for detecting super-heavy tails. *Journal of Statistical Planning and Inference*, 139(2):213–227, 2009.

[2] J. Beirlant, M. Broniatowski, J. Teugels, and P. Vynckier. The mean residual life function at great age: Applications to tail estimation. *Journal of Statistical Planning and Inference*, 45(1-2):21–48, 1995.

[3] J. Beirlant, J-P. Raoult, and R. Worms. On the relative approximation error of the generalized Pareto approximation for a high quantile. *Extremes*, 13:335–360, 2003.

[4] N.H. Bingham, C.M. Goldie, and J.L. Teugels. *Regular Variation*, volume 27 of *Encyclopedia of Mathematics and its application*. Cambridge University Press, 1987.

[5] L. Breiman, C.J. Stone, and C. Kooperberg. Robust confidence bounds for extreme upper quantiles. *Journal of Statistical Computation and Simulation*, 37:127–149, 1990.

[6] J. Cohen. Convergence rates for the ultimate and pentultimate approximations in extreme-value theory. *Advances in Applied Probability*, 14(4):833–854, 1982.

[7] L. de Haan and A. Ferreira. *Extreme value theory: an introduction*. Springer Science & Business Media, 2007.

[8] C. de Valk. Approximation and estimation of very small probabilities of multivariate extreme events. *Extremes*, 19(4):687–717, 2016.

[9] C. de Valk. Approximation of high quantiles from intermediate quantiles. *Extremes*, 19(4):661–686, 2016.

[10] C. de Valk and J.-J. Cai. A high quantile estimator based on the log-generalized Weibull tail limit. *Econometrics and Statistics*, 6:107–128, 2018.

[11] J. Diebolt, M.-A. El-Aroui, V. Durbec, and B. Villain. Estimation of extreme quantiles: Empirical tools for methods assessment and comparison. *International Journal of Reliability, Quality and Safety Engineering*, 7(01):75–94, 2000.

[12] J. Diebolt and S. Girard. A note on the asymptotic normality of the ET method for extreme quantile estimation. *Statistics and Probability Letters*, 62(4):397–406, 2003.

[13] J. El Methni, L. Gardes, S. Girard, and A. Guillou. Estimation of extreme quantiles from heavy and light tailed distributions. *Journal of Statistical Planning and Inference*, 142(10):2735–2747, 2012.

[14] J. Galambos. *The asymptotic theory of extreme order statistics*. R.E. Krieger publishing compagny, 1987.

[15] L. Gardes and S. Girard. Estimating extreme quantiles of Weibull tail-distributions. *Communication in Statistics - Theory and Methods*, 34:1065–1080, 2005.

[16] L. Gardes and S. Girard. Estimation of the Weibull tail-coefficient with linear combination of upper order statistics. *Journal of Statistical Planning and Inference*, 138(5):1416–1427, 2008.

[17] L. Gardes and S. Girard. Conditional extremes from heavy-tailed distributions: An application to the estimation of extreme rainfall return levels. *Extremes*, 13(2):177–204, 2010.

[18] L. Gardes, S. Girard, and A. Guillou. Weibull tail-distributions revisited: a new look at some tail estimators. *Journal of Statistical Planning and Inference*, 141(1):429–444, 2011.

[19] Y. Goegebeur, J. Beirlant, and T. De Wet. Generalized kernel estimators for the Weibull-tail coefficient. *Communications in Statistics-Theory and Methods*, 39(20):3695–3716, 2010.

[20] M.I. Gomes. Penultimate limiting forms in extreme value theory. *Annals of the Institute of Statistical Mathematics*, 36(1):71–85, 1984.

[21] M.I. Gomes and L. de Haan. Approximation by penultimate extreme value distributions. *Extremes*, 2(1):71–85, 1999.

[22] M.I. Gomes and D.D. Pestana. Nonstandard domains of attraction and rates of convergence. *In: New Perspecties in Theoretical and Applied Statistics*, pages 467–477, 1987.

[23] B. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174, 1975.

[24] J. Pickands. Statistical inference using extreme order statistics. *The Annals of Statistics*, 3:119–131, 1975.

[25] R.L. Smith. Estimating tails of probability distributions. *The Annals of Statistics*, 15(3):1174–1207, 1987.

[26] I. Weissman. Estimation of parameters and large quantiles based on the k largest observations. *Journal of the American Statistical Association*, 73(364):812–815, 1978.

[27] R. Worms. Penultimate approximation for the distribution of the excesses. *ESAIM: Probability and Statistics*, 6:21–31, 2002.

Acknowledgments.

The authors thank two anonymous referees and the associate editor for their helpful suggestions and comments which contributed to an improved presentation of the results of this paper.

Appendix: Auxiliary results

We begin with an elementary result.

Lemma 1 *For all* $(a,b,t) \in \mathbb{R}_+^3$*, let*

$$\Psi_a(t;b) = \int_0^b u^a \exp(-tu)du.$$

(i) $\Psi_a(\cdot;b)$ *is continuous, non-increasing on* $\mathbb{R}_+$*,* $\Psi_a(0;b) = b^{a+1}/(a+1)$ *and* $\Psi_a(t;b) \to 0$ *as* $t \to \infty$*.*

(ii) $\Psi_1(t,b) \sim 1/t^2$ *and* $\Psi_2(t,b) \sim 1/t^3$ *as* $t \to \infty$*.*

The proof is straightforward. Lemma 2 below shows that $K_1 \in RV_{\theta_1}$ implies $|K_2| \in RV_{\theta_2}$ when $\ell_1 \neq 1$. In the case where $\ell_1 = 1$, the logistic distribution defined by $H^{-1}(x) = \log(\exp(x) - 1)$, $x > 0$ is a case where $K_2(x) \sim -x\exp(-x)$ is not regularly varying as $x \to \infty$.

Lemma 2 *Assume* **(A3)**, **(A4)** *hold.*

(i) If $\ell_1 = 0$ *then* $\theta_1 \leq 0$*,* $\ell_2 = 0$*,* $-K_2 \in RV_{\theta_1}$ *and* $K_2(t) \sim (\theta_1 - 1)K_1(t)$ *as* $t \to \infty$*.*

(ii) If $\ell_1 = 1$ *then* $\theta_1 = 0$ *and* $\ell_2 = 0$*.*

(iii) If $0 < \ell_1 < \infty$ *and* $\ell_1 \neq 1$ *then* $\theta_1 = 0$*,* $\ell_2 = \ell_1(\ell_1 - 1) \neq 0$ *and* $|K_2| \in RV_0$*.*

(iv) If $\ell_1 = \infty$ *then* $\theta_1 \geq 0$*,* $\ell_2 = \infty$*,* $K_2 \in RV_{2\theta_1}$ *and* $K_2(t) \sim K_1^2(t)$ *as* $t \to \infty$*.*

Proof. The proof relies on the following four facts: First, for all $x \in \mathbb{R}$,

$$K_2(x) = K_1^2(x) + K_1(x)\left(\frac{xK_1'(x)}{K_1(x)} - 1\right),$$

or, equivalently,

$$\frac{K_2(x)}{K_1^2(x)} = 1 + \frac{1}{K_1(x)}\left(\frac{xK_1'(x)}{K_1(x)} - 1\right). \tag{14}$$

Second, $xK_1'(x)/K_1(x) \to \theta_1$ as $x \to \infty$ from the monotone density theorem ([4], Proposition 1.7.2). Third, it straightforwardly follows that $\ell_2 = \ell_1(\ell_1 + \theta_1 - 1)$. Finally, for all positive function K, $K(x) \to c > 0$ as $x \to \infty$ implies $K \in RV_0$. ∎

The next lemma establishes the links between δ and Δ through K_1 and K_2.

Lemma 3 *Suppose* **(A1)**–**(A4)** *hold.*

(i) For all $t > 0$*:*

$$\Delta(t) = \delta^2(t)\int_0^1 \frac{K_2(y(t)(1-\delta(t)u))}{(1-\delta(t)u)^2}\exp\left(K_1(y(t))L_{\theta_1}(1-\delta(t)u)(1+o(1))\right)udu,$$

where $L_{\theta_1}(x) = \int_1^x u^{\theta_1 - 1}du$ *for all* $x \in \mathbb{R}$*.*

(ii) If, moreover, $\ell_1 \neq 1$, then, for all $t > 0$:

$$\begin{aligned} |\Delta(t)| &\leq \max(|K_2(y(t))|, |K_2(x(t))|)\ \frac{\delta^2(t)}{(1-\delta(t))^2}\Phi\left(\delta(t)K_1(y(t))(1+o(1))\right) \text{ and} \\ |\Delta(t)| &\geq \min(|K_2(y(t))|, |K_2(x(t))|)\ \delta^2(t)\Phi\left(\delta(t)K_1(y(t))(1-\delta(t))^{\theta_1-1}(1+o(1))\right), \end{aligned}$$

where $\Phi(s) = \Psi_1(s;1) = \int_0^1 u\exp(-us)du$ for all $s \geq 0$.

Proof. (i) Under **(A2)**, a second order Taylor expansion with integral remainder yields

$$\begin{aligned} \Delta(t) &= \int_{x(t)}^{y(t)} \frac{K_2(s)}{s^2}\frac{\varphi(s)}{\varphi(y(t))}(y(t)-s)ds \\ &= \delta^2(t)\int_0^1 \frac{K_2(y(t)(1-\delta_n u))}{(1-\delta(t)u)^2}\frac{\varphi(y(t)(1-\delta(t)u))}{\varphi(y(t))}udu, \end{aligned}$$

thanks to the change of variable $u = (y(t)-s)/(y(t)-x(t))$. Besides,

$$\begin{aligned} \frac{\varphi(y(t)(1-\delta(t)u))}{\varphi(y(t))} &= \exp\left(\log\varphi(y(t)(1-\delta(t)u)) - \log\varphi(y(t))\right) \\ &= \exp\left(\int_{y(t)}^{y(t)(1-\delta(t)u)} (\log\varphi(s))' ds\right) \\ &= \exp\left(\int_{y(t)}^{y(t)(1-\delta(t)u)} \frac{K_1(s)}{s} ds\right) \\ &= \exp\left(\int_1^{1-\delta(t)u} \frac{K_1(vy(t))}{v} dv\right) \\ &= \exp\left(K_1(y(t))\int_1^{1-\delta(t)u} \frac{K_1(vy(t))}{K_1(y(t))}\frac{dv}{v}\right). \end{aligned}$$

Since $1-\delta(t)u \in [1-\delta(t), 1]$, **(A3)** yields $K_1(vy(t))/K_1(y(t)) \to v^{\theta_1}$ uniformly locally as $t \to \infty$ and consequently $y(t) \to \infty$. Condition **(A1)** then leads to

$$\frac{\varphi(y(t)(1-\delta(t)u))}{\varphi(y(t))} = \exp\left(K_1(y(t))L_{\theta_1}(1-\delta(t)u)(1+o(1))\right).$$

It thus follows that

$$\Delta(t) = \delta^2(t)\int_0^1 \frac{K_2(y(t)(1-\delta(t)u))}{(1-\delta(t)u)^2}\exp\left(K_1(y(t))L_{\theta_1}(1-\delta(t)u)(1+o(1))\right)udu$$

and the first part of the result is proved.
(ii) From Lemma 2, when $\ell_1 \neq 1$ the sign of K_2 is ultimately constant so that

$$|\Delta(t)| = \delta^2(t)\int_0^1 \frac{|K_2(y(t)(1-\delta(t)u))|}{(1-\delta(t)u)^2}\exp\left(K_1(y(t))L_{\theta_1}(1-\delta(t)u)(1+o(1))\right)udu.$$

Let us remark that, for all $u \in [0,1]$ and $\theta_1 \leq 1$, one has $1-\delta(t) \leq 1-\delta(t)u \leq 1$ and

$$-(1-\delta(t))^{\theta_1-1}\delta(t)u \leq L_{\theta_1}(1-\delta(t)u) \leq -\delta(t)u.$$

It is thus clear that

$$\begin{aligned}|\Delta(t)| &\leq \frac{\delta^2(t)}{(1-\delta(t))^2}\int_0^1 |K_2(y(t)(1-\delta(t)u))|\exp\left(-\delta(t)K_1(y(t))u(1+o(1))\right)udu,\\ |\Delta(t)| &\geq \delta^2(t)\int_0^1 |K_2(y(t)(1-\delta(t)u))|\exp\left(-\delta(t)K_1(y(t))(1-\delta(t))^{\theta_1-1}u(1+o(1))\right)udu.\end{aligned}$$

Besides, Lemma 2 entails that $|K_2|$ is regularly varying when $\ell_1 \neq 1$. Therefore, $|K_2|$ is ultimately monotone and it follows that, for t large enough,

$$m(t) \leq |K_2(y(t)(1-\delta(t)u))| \leq M(t),$$

where $m(t) := \min(|K_2(y(t))|, |K_2(x(t))|)$ and $M(t) := \max(|K_2(y(t))|, |K_2(x(t))|)$, leading to

$$\begin{aligned}|\Delta(t)| &\leq M(t)\frac{\delta^2(t)}{(1-\delta(t))^2}\int_0^1 u\exp\left(-\delta(t)K_1(y_n)u(1+o(1))\right)du \text{ and}\\ |\Delta(t)| &\geq m(t)\delta^2(t)\int_0^1 u\exp\left(-\delta(t)K_1(y(t))(1-\delta(t))^{\theta_1-1}u(1+o(1))\right)du.\end{aligned}$$

Introducing for all $s \geq 0$, $\Phi(s) = \int_0^1 u\exp(-us)du$, the above bounds can be rewritten as

$$\begin{aligned}|\Delta(t)| &\leq M(t)\frac{\delta^2(t)}{(1-\delta(t))^2}\Phi\left(\delta(t)K_1(y(t))(1+o(1))\right) \text{ and}\\ |\Delta(t)| &\geq m(t)\delta^2(t)\Phi\left(\delta(t)K_1(y(t))(1-\delta(t))^{\theta_1-1}(1+o(1))\right),\end{aligned}$$

which concludes the proof. ∎

In the case where $\ell_1 < \infty$, the asymptotic equivalent provided in Lemma 3(i) can be simplified as follows:

Lemma 4 *Suppose* **(A1)**–**(A4)** *hold and* $\ell_1 < \infty$. *Then,*

$$\Delta(t) = \delta^2(t)\int_0^1 K_2(y(t)(1-\delta(t)u))(1-\delta(t)u)^{\ell_1-2}udu(1+o(1)),$$

as $t \to \infty$.

Proof. If $\ell_1 = 0$ then Lemma 3(i) yields

$$\Delta(t) = \delta^2(t)\int_0^1 K_2(y(t)(1-\delta(t)u))(1-\delta(t)u)^{-2}udu(1+o(1)).$$

In the situation where $0 < \ell_1 < \infty$, Lemma 2(iii) entails $\theta_1 = 0$ and Lemma 3(i) yields

$$\Delta(t) = \delta^2(t)\int_0^1 K_2(y(t)(1-\delta(t)u))(1-\delta(t)u)^{\ell_1-2+o(1)}udu(1+o(1)),$$

and the result is proved. ∎

As a consequence of the two above results, a sufficient condition as well as a necessary condition can be established such that $\Delta(t) \to 0$ as $t \to \infty$.

Lemma 5 *Suppose* **(A1)**–**(A4)** *hold.*

(i) If $\delta^2(t)\max(|K_2(y(t))|, |K_2(x(t))|) \to 0$ *then* $\Delta(t) \to 0$ *as* $t \to \infty$.

(ii) If $\Delta(t) \to 0$ *then* $\delta^2(t)\min(|K_2(y(t))|, |K_2(x(t))|) \to 0$ *as* $t \to \infty$.

Proof. Let us first note that when $\ell_1 = 1$ then $\ell_2 = 0$ from Lemma 2(ii). It is thus clear in view of Lemma 4 that $\Delta(t) \to 0$ as $t \to \infty$ without restriction on $\delta(t)$. In the following, we thus focus on the case where $\ell_1 \neq 1$. Lemma 2 entails that $|K_2|$ is regularly varying since $\ell_1 \neq 1$. Therefore, $|K_2|$ is ultimately monotone. Let us focus on the situation where $|K_2|$ is ultimately non decreasing and introduce $A(t) = \delta(t)K_1(y(t))$ for all $t > 0$.
(i) Assume that $\delta^2(t)|K_2(y(t))| \to 0$ as $t \to \infty$. From Lemma 1(i), $0 \leq \Phi(s) \leq 1/2$ for all $s \geq 0$ and thus Lemma 3(ii) entails

$$|\Delta(t)| \leq \frac{\delta^2(t)|K_2(y(t))|}{2(1-\delta(t))^2} \to 0 \tag{15}$$

as $t \to \infty$ in view of **(A1)**.
(ii) From Lemma 3(ii), one has

$$|\Delta(t)| \geq |K_2(x(t))|\delta^2(t)\Phi\left(A(t)(1-\delta(t))^{\theta_1-1}(1+o(1))\right) \geq |K_2(x(t))|\delta^2(t)\Phi\left(cA(t)\right)$$

for t large enough and some $c > 0$ since Φ is non-increasing, see Lemma 1(i). For all $s \geq 0$, let $\psi(s) = \int_0^s x\exp(-x)dx = s^2\Phi(s)$. Consider $s_0 \geq c(3-2\theta_1)$ with $\theta_1 \leq 1$ and remark that $\Phi(s) \geq \Phi(s_0)$ for all $0 \leq s \leq s_0$ and $\psi(s) \geq \psi(s_0)$ for all $s \geq s_0$. As a consequence, for all $s > 0$,

$$\Phi(s) \geq \frac{\psi(s_0)}{s_0^2}\mathbb{I}\{s \leq s_0\} + \frac{\psi(s_0)}{s^2}\mathbb{I}\{s \geq s_0\},$$

and thus

$$\begin{aligned}
|\Delta(t)| &\geq \frac{\psi(s_0)}{s_0^2}|K_2(x(t))|\delta^2(t)\mathbb{I}\{A(t) \leq s_0/c\} \\
&+ \frac{\psi(s_0)}{c^2}\frac{|K_2(x(t))|}{K_1^2(y(t))}\mathbb{I}\{A_n \geq s_0/c\} \\
&\geq \frac{\psi(s_0)}{s_0^2}|K_2(x(t))|\delta^2(t)\mathbb{I}\{A(t) \leq s_0/c\} \\
&+ \frac{\psi(s_0)}{c^2}\frac{|K_2(x(t))|}{K_1^2(x(t))}\frac{K_1^2(x(t))}{K_1^2(y(t))}\mathbb{I}\{A(t) \geq s_0/c\}.
\end{aligned} \tag{16}$$

Since K_1 is regularly varying, $K_1(x(t))/K_1(y(t)) \sim (1-\delta(t))^{\theta_1} \geq c' > 0$ as $t \to \infty$ in view of **(A1)** and

$$|\Delta(t)| \geq \frac{\psi(s_0)}{s_0^2}|K_2(x(t))|\delta^2(t)\mathbb{I}\{A(t) \leq s_0/c\} + \psi(s_0)\left(\frac{c'}{c}\right)^2\frac{|K_2(x(t))|}{K_1^2(x(t))}\mathbb{I}\{A(t) \geq s_0/c\}.$$

Remarking that, (14) in the proof of Lemma 2 implies that, for t large enough,

$$\frac{K_2(x(t))}{K_1^2(x(t))} = 1 + \frac{1}{K_1(x(t))}\left(\frac{x(t)K_1'(x(t))}{K_1(x(t))} - 1\right) = 1 + \frac{\delta(t)}{A(t)}(\theta_1 - 1 + o(1))$$

which yields when $A(t) \geq s_0/c$,

$$\left|\frac{K_2(x(t))}{K_1^2(x(t))} - 1\right| \leq \frac{c\delta(t)}{s_0}|\theta_1 - 1 + o(1)| \leq \frac{c}{s_0}(3/2 - \theta_1) \leq \frac{1}{2}.$$

It thus follows that

$$\frac{|K_2(x(t))|}{K_1^2(x(t))}\mathbb{I}\{A(t) \geq s_0/c\} \geq \frac{1}{2}\mathbb{I}\{A(t) \geq s_0/c\}$$

and therefore,

$$|\Delta(t)| \geq \frac{\psi(s_0)}{s_0^2}|K_2(x(t))|\delta^2(t)\mathbb{I}\{A(t) \leq s_0/c\} + \frac{\psi(s_0)}{2}\left(\frac{c'}{c}\right)^2 \mathbb{I}\{A(t) \geq s_0/c\}.$$

As a conclusion, $|\Delta(t)| \to 0$ implies $|K_2(x(t))|\delta^2(t)\mathbb{I}\{A(t) \leq s_0/c\} \to 0$ and $\mathbb{I}\{A(t) \geq s_0/c\} \to 0$ as $t \to \infty$. Consequently, $A(t) \leq s_0/c$ eventually and $\delta^2(t)K_2(x(t)) \to 0$ as $t \to \infty$.
Let us now consider the situation where $|K_2|$ is ultimately non increasing.
(i) The proof is similar, the upper bound (15) is replaced by

$$|\Delta(t)| \leq \frac{\delta^2(t)|K_2(x(t))|}{2(1-\delta(t))^2}. \tag{17}$$

(ii) The lower bound (16) is replaced by

$$|\Delta(t)| \geq \frac{\psi(s_0)}{s_0^2}|K_2(y(t))|\delta^2(t)\mathbb{I}\{A(t) \leq s_0/c\} + \frac{\psi(s_0)}{c^2}\frac{|K_2(y(t))|}{K_1^2(y(t))}\mathbb{I}\{A(t) \geq s_0/c\}$$

and the end of the proof is similar. ■

	$\bar{F}(x)$	θ_1	θ_2	$K_1(x)$	$K_2(x)$	ℓ_1
DA_1(Gumbel)						
Finite endpoint ($\beta > 0$)	$\exp\left(-(-\log x)^{-\beta}\right)$ $x \in (0,1)$	$-1/\beta$	$-1/\beta$	$\frac{1}{\beta}x^{-1/\beta}$	$-\frac{1+\beta}{\beta^2}x^{-1/\beta}(1+o(1))$	0
Gamma ($a > 0$)	$\frac{1}{\Gamma(a)}\int_x^\infty t^{a-1}e^{-t}dt$ $x \geq 0$	0	-1	$1+o(1)$	$\frac{1-a}{x}(1+o(1))$	1
DA_2(Gumbel)						
Weibull ($\beta \neq 1$)	$\exp(-x^\beta)$ $x \geq 0$	0	0	$\frac{1}{\beta}$	$\frac{1-\beta}{\beta^2}$	$1/\beta$
Gaussian	$\frac{1}{\sqrt{2\pi}}\int_x^\infty \exp\left(-\frac{t^2}{2}\right)dt$	0	0	$\frac{1}{2}+o(1)$	$-\frac{1}{4}+o(1)$	1/2
DA_3(Gumbel)						
Log-Weibull ($\beta > 1$)	$\exp(-(\log x)^\beta)$ $x \geq 1$	$1/\beta$	$2/\beta$	$\frac{1}{\beta}x^{1/\beta}$	$\frac{1}{\beta^2}x^{2/\beta}(1+o(1))$	$+\infty$
Lognormal ($\sigma > 0$)	$\frac{1}{\sigma\sqrt{2\pi}}\int_x^\infty \frac{1}{t}\exp\left(-\frac{(\log t)^2}{2\sigma^2}\right)dt$ $x \geq 0$	1/2	1	$\frac{\sigma}{\sqrt{2}}x^{1/2}(1+o(1))$	$\frac{\sigma^2}{2}x(1+o(1))$	$+\infty$

Table 1: Examples of distributions in DA(Gumbel).

Distribution	First order approximation of $\varepsilon_{\mathrm{ET}}(p_n;\alpha_n)$
$\mathbf{DA}_1$(Gumbel)	
Finite endpoint($\beta>0$)	$-\dfrac{2(1+\beta)}{\beta^2}\dfrac{(\log\log n)^2}{(\log n)^{2+1/\beta}}$
Gamma($a>0$)	$2(1-a)\dfrac{(\log\log n)^2}{(\log n)^3}$
$\mathbf{DA}_2$(Gumbel)	
Weibull($\beta\neq 1$)	$\dfrac{2(1-\beta)}{\beta^2}\dfrac{(\log\log n)^2}{(\log n)^2}$
Gaussian	$-\dfrac{1}{2}\dfrac{(\log\log n)^2}{(\log n)^2}$
$\mathbf{DA}_3$(Gumbel)	
Log-Weibull($\beta>1$)	$\dfrac{2}{\beta^2}\dfrac{(\log\log n)^2}{(\log n)^{2-2/\beta}}$
Lognormal	$\sigma^2\dfrac{(\log\log n)^2}{\log n}$

Table 2: First order approximations of $\varepsilon_{\mathrm{ET}}(p_n;\alpha_n)$ with $p_n = 1/(n\log n)$ and $\alpha_n = (\log n)/n$ associated with the distributions described in Table 1.

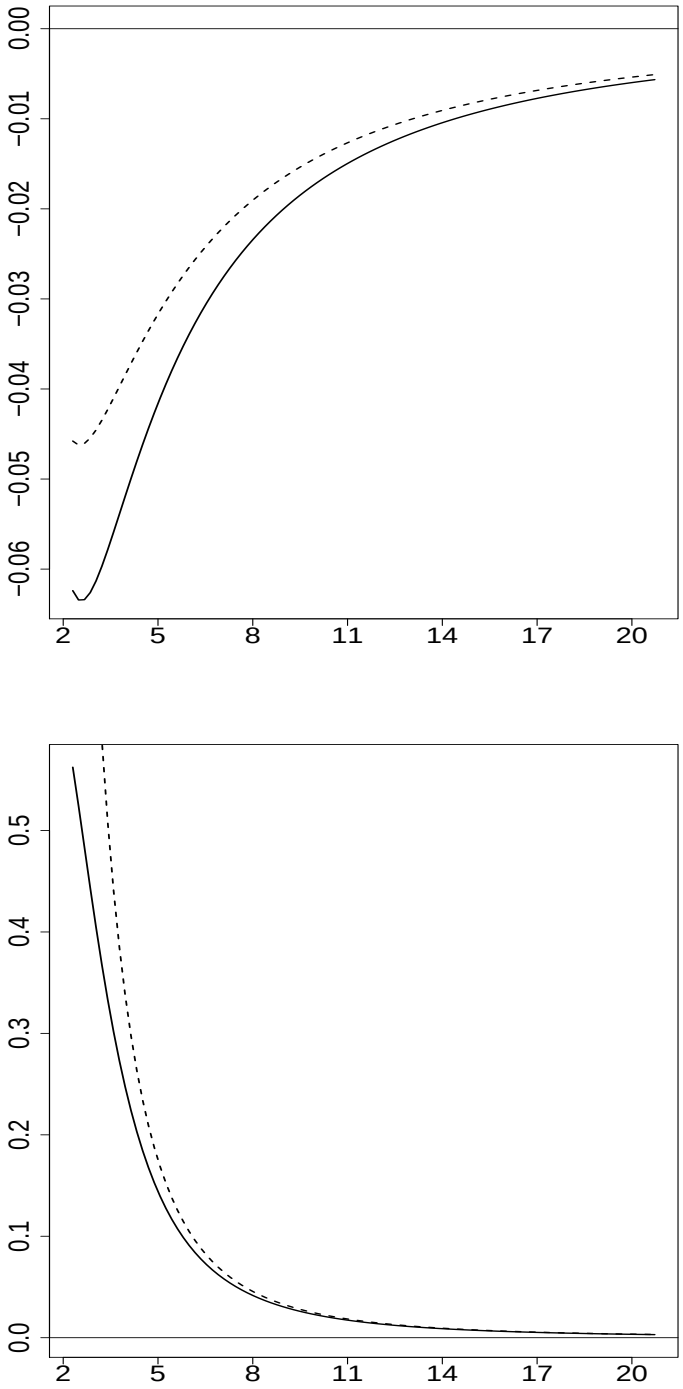

Figure 1: Extrapolation error in DA_1(Gumbel). Vertically: Extrapolation error $\varepsilon_{\mathrm{ET}}(p_n; \alpha_n)$ (solid line) and its first order approximation $\frac{1}{2}\eta_n^2 K_2(\log n)$ (dashed line) provided by Theorem 2(i)-(a). Horizontally: $\log n$. Top: Finite endpoint($\beta = 5$) distribution, bottom: Gamma($a = 0.1$) distribution.

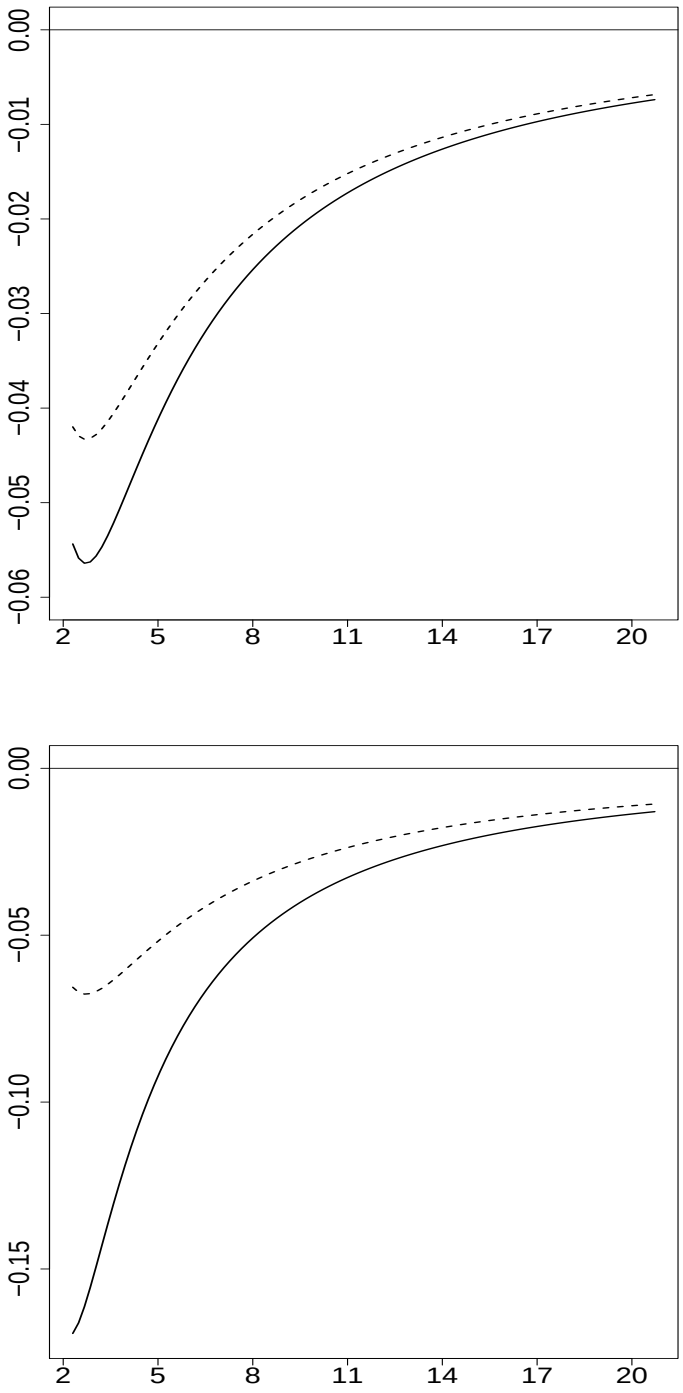

Figure 2: Extrapolation error in DA_2(Gumbel). Vertically: Extrapolation error $\varepsilon_{\mathrm{ET}}(p_n; \alpha_n)$ (solid line) and its first order approximation $\frac{\ell_1(\ell_1-1)}{2}\eta_n^2$ (dashed line) provided by Theorem 2(ii)-(a). Horizontally: $\log n$. Top: Weibull($\beta = 5$) distribution, bottom: Gaussian distribution.

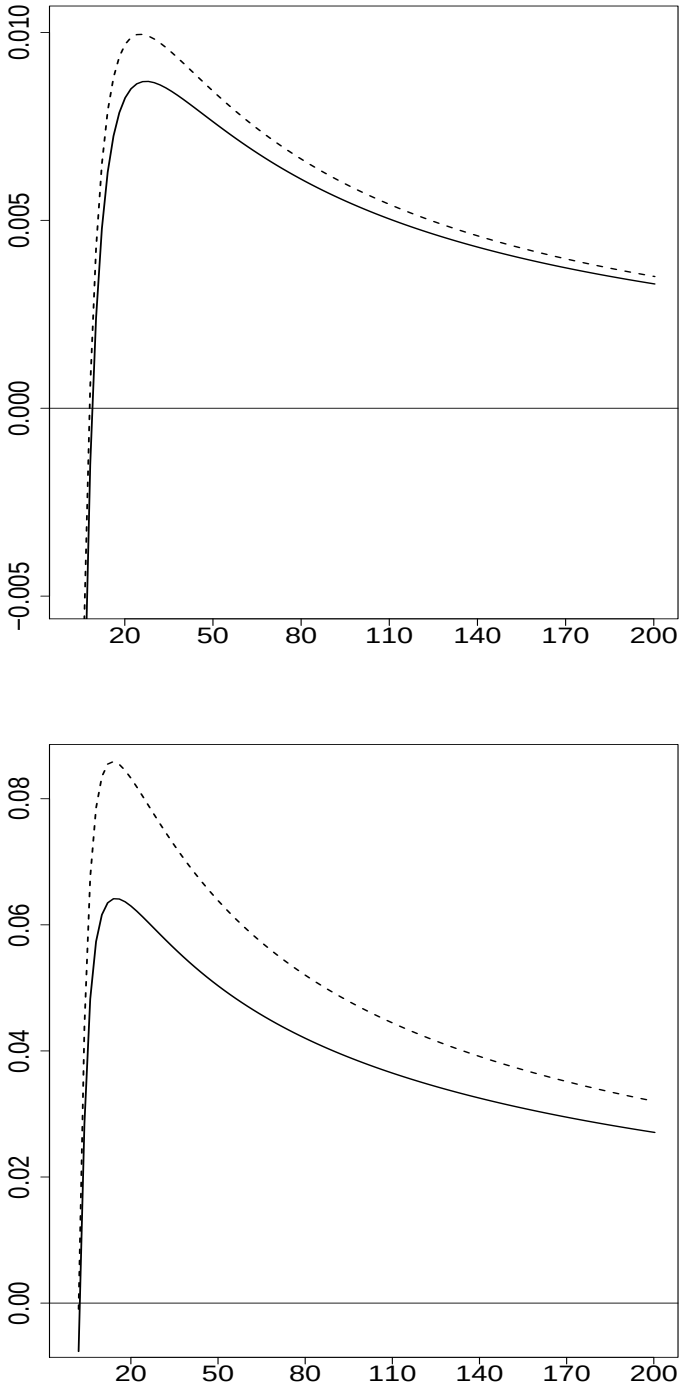

Figure 3: Extrapolation error in DA_3(Gumbel). Vertically: Extrapolation error $\varepsilon_{\mathrm{ET}}(p_n; \alpha_n)$ (solid line) and its first order approximation $\frac{1}{2}\eta_n^2 K_2(\log n)$ (dashed line) provided by Theorem 2(iii)-(a). Horizontally: $\log n$. Top: log-Weibull($\beta = 3$) distribution, bottom: lognormal($\sigma = 0.5$) distribution.

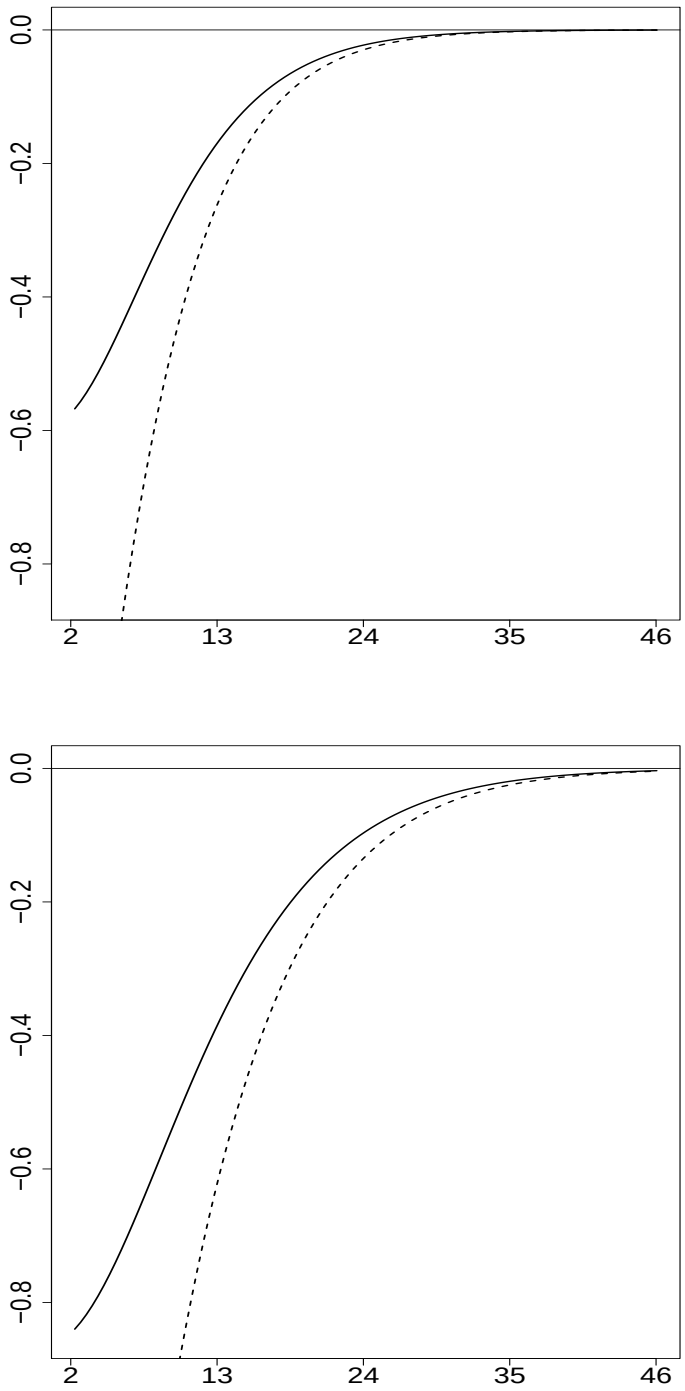

Figure 4: Extrapolation error in DA(Fréchet). Vertically: Extrapolation error $\varepsilon_{\mathrm{W}}(p_n;\alpha_n)$ (solid line) and its first order approximation $-\dfrac{\delta_\infty}{1-\delta_\infty}\log(1/\alpha_n)\eta(1/\alpha_n)$ (dashed line) provided by Theorem 3(ii). Horizontally: $\log n$. Top: $\mathrm{Burr}(k=3)$ distribution (see Section 5), bottom: $\mathrm{Burr}(k=4)$ distribution.

2.3 Perspectives

Dans ce chapitre, nous avons étudié les erreurs d'extrapolation associées à deux approximations issues de la théorie des valeurs extrêmes. Dans le cas de l'approximation ET dédiée au domaine d'attraction de Gumbel, nous avons montré que l'étude de l'erreur d'extrapolation menait à une sous-division du domaine d'attraction de Gumbel en trois parties. En particulier, nous avons pu en déduire que l'approximation ET était pensée autour de la loi Exponentielle, l'erreur d'extrapolation tendant vers zéro sans aucune restriction pour des lois proches de cette dernière loi. Dans le cas de l'approximation Weissman, nous avons montré que l'erreur d'extrapolation tendait vers zéro quelle que soit la loi appartenant au domaine d'attraction de Fréchet considérée, contrastant ainsi avec les conclusions obtenues dans le cas ET. Nous donnons ci-dessous quelques perspectives à donner à ces travaux.

Autres sources d'erreur Rappelons que l'erreur relative $(\hat{q}_{\mathrm{ET}}(p_n;\alpha_n)-q(p_n))/q(p_n)$ peut se décomposer comme la somme de deux termes :

$$\epsilon(p_n) \quad = \quad \epsilon_{est}(p_n)+\epsilon_{ext}(p_n),$$

avec

$$\begin{aligned}\epsilon_{est}(p_n) &:= \frac{\tilde{q}(p_n)-\hat{q}(p_n)}{q(p_n)},\\ \epsilon_{ext}(p_n) &:= \frac{q(p_n)-\tilde{q}(p_n)}{q(p_n)},\end{aligned}$$

le premier étant une erreur d'estimation aléatoire et le second une erreur d'extrapolation déterministe. Dans ce chapitre, nous nous sommes intéressés à ce deuxième terme. Une des perspectives de ces travaux est de proposer une étude de cette première erreur. Nous en détaillons les principaux aspects ci-dessous, en commençant par l'erreur d'estimation associée à l'approximation Weissman.

Dans le cadre de l'approximation Weissman, l'erreur d'estimation se réécrit :

$$\begin{aligned}\delta_{\mathrm{W}}(p_n;\alpha_n) &:= \frac{\tilde{q}_{\mathrm{W}}(p_n;\alpha_n)-\hat{q}_{\mathrm{W}}(p_n;\alpha_n)}{q(p_n)}\\ &= \frac{q(\alpha_n)}{q(p_n)}\left(\frac{\alpha_n}{p_n}\right)^{\gamma}\left[1-\frac{\hat{q}(\alpha_n)}{q(\alpha_n)}\left(\frac{\alpha_n}{p_n}\right)^{\hat{\gamma}_n-\gamma}\right].\end{aligned} \tag{2.5}$$

Le comportement asymptotique de (2.5) pourrait par exemple être déduit de DE HAAN et FERREIRA [2007, Théorème 4.3.8]. Dans ce dernier résultat, les auteurs supposent cependant que la fonction quantile de queue U vérifie une condition du second-ordre (cf Paragraphe 1.2.3), en plus de supposer que F ∈MDA(Fréchet).

Dans le cadre de l'estimateur ET, l'erreur d'estimation s'écrit

$$\begin{aligned}\delta_{\mathrm{ET}}(p_n;\alpha_n) &:= \frac{\tilde{q}_{\mathrm{ET}}(p_n;\alpha_n)-\hat{q}_{\mathrm{ET}}(p_n;\alpha_n)}{q(p_n)}\\ &= -\frac{\hat{q}(\alpha_n)-q(\alpha_n)+(\hat{\sigma}_n-\sigma_n)\log(\alpha_n/p_n)}{q(p_n)}.\end{aligned} \tag{2.6}$$

Le comportement asymptotique de (2.6) fait l'objet de DE HAAN et ROOTZÉN [1993, Proposition 1] ou encore de DIEBOLT et GIRARD [2003, Théorème 1] (voir également la Remarque 2.1). Ces études sont cependant faites sous l'hypothèse classique d'appartenance au domaine d'attraction de Gumbel pour le premier ou des hypothèses du second-ordre sur la fonction U pour le deuxième. Or l'étude que nous proposons ci-dessus dans le cadre de l'estimateur ET se base sur l'hypothèse que $\mathrm{V}(\cdot):=\log \mathrm{U}(\exp(\cdot))=\log \mathrm{H}^{-1}(\cdot)$ est à variation régulière étendue, voir Page 54.

Une des perspectives de ces travaux serait de proposer une étude du comportement asymptotique de l'erreur d'estimation associée à ET sous l'hypothèse que V est à variation régulière étendue. Pour ce faire, une idée serait par exemple d'adapter l'étude de l'erreur d'estimation proposée par DIEBOLT et GIRARD [2003] à nos hypothèses de modèle.

Dans DIEBOLT et GIRARD [2003], les auteurs prouvent que :

$$\Delta_n := \frac{k_n^{1/2}}{\sigma_n \log(\alpha_n/p_n)}(\hat{q}_{\mathrm{ET}}(p_n;\alpha_n)-\tilde{q}_{\mathrm{ET}}(p_n;\alpha_n)) \xrightarrow{d} \mathrm{N}(0,1).$$

Pour ce faire, les auteurs proposent de découper la quantité précédente en plusieurs termes :

$$\begin{aligned}\Delta_n &= \frac{k_n^{1/2}}{\sigma_n \log(\alpha_n/p_n)}(\hat{q}(\alpha_n) - q(\alpha_n)) \\ &+ k_n^{1/2}\left(\frac{\hat{\sigma}_n}{\sigma_n} - \mu(u_n)\right) \\ &+ k_n^{1/2}(\mu(u_n) - 1),\end{aligned}$$

avec

$$\mu(u_n) = \int_0^\infty \overline{F}_{u_n}(\sigma_n y)\, dy.$$

Ils étudient ensuite séparemment les trois termes. Le premier terme correspond à l'erreur d'estimation aléatoire du quantile intermédiaire $q(\alpha_n)$ par la statistique d'ordre intermédiaire $X_{n-k_n+1,n}$. Le deuxième terme correspond à l'erreur d'estimation aléatoire de la constante normalisante σ_n par la moyenne des excès $\hat{\sigma}_n = \frac{1}{k_n}\sum_{i=1}^{k_n}(X_{n-i+1,n} - X_{n-k_n+1,n})$. Enfin, le troisième terme correspond à l'erreur d'approximation déterministe de la loi des excès par une exponentielle.

Il conviendrait alors d'étudier ces deux derniers termes en supposant que V est à variation régulière étendue, le premier terme étant étudié Théorème 1 du Chapitre 3. Par ailleurs, il serait également intéressant de comparer l'erreur d'extrapolation et le troisième terme, qui est lui aussi déterministe.

Influence des suites normalisantes La décomposition en deux termes de la différence entre l'estimateur du quantile et ce dernier :

$$\hat{q}(p_n) - q(p_n) = (\hat{q}(p_n) - \tilde{q}(p_n)) + (\tilde{q}(p_n) - q(p_n))$$

pose la question du choix de $\tilde{q}(p_n)$. Si tous les centrages sont équivalents quand il s'agit d'étudier l'erreur globale, ce n'est pas le cas quand il s'agit uniquement d'étudier $\tilde{q}(p_n) - q(p_n)$.

Par exemple, dans le cas de l'approximation ET, rappelons que (cf équation (1.40))

$$\tilde{q}(p_n) := q(\alpha_n) + \sigma_n \log(\alpha_n/p_n).$$

Dans l'article ci-dessus, nous avons choisi de poser $\sigma_n = \varphi'(x(n))$ (cf Page 52), avec $x(n) = \log(1/\alpha_n)$, $\varphi(\cdot) = H^{-1}(\cdot)$ et $H(\cdot) := -\log(1 - F(\cdot))$ (cf Page 54), menant à

$$q(\alpha_n) = H^{-1}(\log 1/\alpha_n)$$

et

$$\sigma_n = (H^{-1})'(\log 1/\alpha_n).$$

Or nous aurions très bien pu choisir

$$q(\alpha_n) = H^{-1}(\log 1/\alpha_n) = U(1/\alpha_n) := b_n^*$$

et

$$\sigma_n = g(b_n^*),$$

avec

$$g(t) := \frac{\int_t^{x^*}(1 - F(s))\, ds}{1 - F(t)},$$

ce qui aurait constitué un choix plus logique au vu de la Proposition 8.

En fait, il est possible de montrer que ces deux choix sont équivalents quand n tend vers l'infini. C'est l'objet du Lemme 1 (Paragraphe 2.4) qui, sous nos hypothèses de modèle, stipule que :

$$g(t) \sim \frac{1}{H'(t)},\ t \to \infty. \tag{2.7}$$

Par conséquent,

$$\begin{aligned} g(b_n^*) &= g(H^{-1}(-\log\alpha_n)) \\ &\sim \frac{1}{H'(H^{-1}(-\log\alpha_n))} \\ &= (H^{-1})'(-\log\alpha_n).\end{aligned}$$

Ainsi, $(H^{-1})'(\log 1/\alpha_n)$ n'est pas exactement égal mais asymptotiquement équivalent à $g(b_n^*)$. Il serait intéressant de connaître la vitesse de convergence dans l'équation (2.7) afin de juger comment le choix du centrage affecte l'étude de l'erreur d'extrapolation.

Similairement, dans le cas de l'approximation Weissman, nous avons choisi

$$\gamma_n = \varphi'(x(n)) = \gamma + \eta(\exp x(n)) = \gamma + \eta(1/\alpha_n) \underset{n\to\infty}{\longrightarrow} \gamma$$

(cf Pages 58-59), avec cette fois-ci $\varphi(\cdot) := \log U(\exp(\cdot)) = V(\cdot)$ et U la fonction quantile de queue. Nous aurions pu choisir directement $\gamma_n = \gamma$, déplaçant le problème de l'étude de l'erreur d'extrapolation au problème de l'étude du terme $\hat{q}(p_n) - \tilde{q}(p_n)$, qui correspond au numérateur de l'erreur d'estimation.

Utilisation d'autres métriques Nous nous sommes intéressés à l'erreur relative d'extrapolation définie de manière générale par :

$$\frac{\varphi(y) - \varphi(x) - (y-x)\varphi'(x)}{\varphi(y)}.$$

Bien qu'étant largement étudiée, cette métrique présente l'inconvénient de ne pas être invariante par translation. Un autre choix possible serait de normaliser

$$\varphi(y) - \varphi(x) - (y-x)\varphi'(x)$$

par

$$\varphi(y) - \varphi(x).$$

Dans le cas ET, cela reviendrait à quantifier le gain de l'approximation $\tilde{q}_{\mathrm{ET}}(p_n;\alpha_n) = q(\alpha_n) + \sigma_n \log(\alpha_n/p_n)$ par rapport à l'approximation triviale $q(p_n) \simeq q(\alpha_n)$. Ce travail pourrait constituer une perspective intéressante.

Etude de l'erreur associée à d'autres approximations Nous avons étudié l'erreur d'extrapolation dans deux configurations, toutes deux cas particuliers de l'approximation GPD. Proposer des études dans le cas d'autres approximations pourrait constituer une piste de travail intéressante : citons les approximation (1.37), (1.54), (1.59) et (3.3).

Pour cette première, qui n'est rien d'autre que l'approximation GPD dans le cas général $\gamma \neq 0$, des études de l'erreur globale ont été proposées dans DE HAAN et RESNICK [1996] et [Théorème 4.3.1]DE HAAN et FERREIRA [2007]. Il s'agirait d'adapter ces études à nos hypothèses de modèle. Cela nous permettrait également de traiter le domaine d'attraction de Weibull.

2.4 Annexe

Soient K_1 et K_2 comme définies dans la partie "Application to the ET approximation" du Paragraphe 2.2, $g(t) := \int_t^{x^*} (1-F(s))ds/(1-F(t))$, x^* étant le point terminal associé à la fonction de répartition F, et H la fonction de hasard cumulée définie par $H(\cdot) := -\log(1-F(\cdot))$.

Lemme 1 *Supposons que* K_1 *et* $|K_2|$ *soient des fonctions à variation régulière d'ordres respectifs* $\theta_1 < 1$ *et* $\theta_2 \in \mathbb{R}$. *Alors :*

$$g(t) \sim \frac{1}{H'(t)},\ t \to \infty.$$

Preuve. Preuve du Lemme 1 :

$$\begin{aligned}
g(t) &:= \int_t^{x^*} (1-F(s))\,ds/(1-F(t)) \\
&= \int_t^{x^*} e^{H(t)-H(s)}\,ds \\
&= \int_0^{x^*} e^{H(t)-H(y+t)}\,dy, \qquad y = s-t \\
&= \int_0^{+\infty} e^{-u}\left(H^{-1}\right)'(u+H(t))\,du, \qquad u = H(y+t)-H(t).
\end{aligned}$$

Posons $t = \mathrm{H}^{-1}(s)$. Alors

$$\begin{aligned} g(\mathrm{H}^{-1}(s)) &= \int_0^{+\infty} e^{-u} \left(\mathrm{H}^{-1}\right)'(u+s)\,\mathrm{d}u \\ &= \int_0^{+\infty} e^{-u} \frac{\left(\mathrm{H}^{-1}\right)'(u+s)}{\left(\mathrm{H}^{-1}\right)'(s)}\,\mathrm{d}u \left(\mathrm{H}^{-1}\right)'(s). \end{aligned}$$

Mais

$$\begin{aligned} \frac{\left(\mathrm{H}^{-1}\right)'(u+s)}{\left(\mathrm{H}^{-1}\right)'(s)} &= \exp\left(\log\left(\mathrm{H}^{-1}\right)'(u+s) - \log\left(\mathrm{H}^{-1}\right)'(s)\right) \\ &= \exp\int_s^{u+s} \left(\log\left(\mathrm{H}^{-1}\right)'\right)'(v)\,\mathrm{d}v \\ &= \exp\int_s^{u+s} \frac{\left(\mathrm{H}^{-1}\right)''(v)}{\left(\mathrm{H}^{-1}\right)'(v)}\,\mathrm{d}v \\ &= \exp\int_s^{u+s} \frac{\mathrm{K}_2(v)}{v\mathrm{K}_1(v)}\,\mathrm{d}v \\ &= \exp\int_1^{1+\frac{u}{s}} \frac{\mathrm{K}_2(ws)}{w\mathrm{K}_1(ws)}\,\mathrm{d}w, \end{aligned}$$

avec $w = v/s$. D'où

$$g(\mathrm{H}^{-1}(s)) = \left(\mathrm{H}^{-1}\right)'(s) \int_0^{+\infty} \exp\left(\int_1^{1+\frac{u}{s}} \frac{\mathrm{K}_2(ws)}{w\mathrm{K}_1(ws)}\,\mathrm{d}w - u\right)\mathrm{d}u.$$

Le reste de la démonstration consiste à montrer que l'intégrale ci-dessus tend vers 1 lorsque $s \to \infty$.

$$\begin{aligned} &\int_1^{1+\frac{u}{s}} \frac{\mathrm{K}_2(ws)}{w\mathrm{K}_1(ws)}\,\mathrm{d}w \\ = \ &\frac{\mathrm{K}_2(s)}{\mathrm{K}_1(s)} \int_1^{1+\frac{u}{s}} \frac{1}{w} \frac{\mathrm{K}_2(ws)}{\mathrm{K}_2(s)} \frac{\mathrm{K}_1(s)}{\mathrm{K}_1(ws)}\,\mathrm{d}w \end{aligned} \tag{2.8}$$

Le Théorème de Potter (voir Théorème 8 du Chapitre 1) nous dit que pour n'importe quelles constantes $\mathrm{A} > 1$, $\delta > 0$, il existe $\mathrm{S} = \mathrm{S}(\mathrm{A}, \delta)$ tel que :

$$\begin{aligned} &\left|\frac{\mathrm{K}_2(s)}{\mathrm{K}_1(s)}\right| \int_1^{1+\frac{u}{s}} \frac{1}{\mathrm{A}} w^{\theta_2-\theta_1-1-\delta}\,\mathrm{d}w \\ &\qquad \leq |(2.8)| \leq \\ &\left|\frac{\mathrm{K}_2(s)}{\mathrm{K}_1(s)}\right| \int_1^{1+\frac{u}{s}} \mathrm{A} w^{\theta_2-\theta_1-1+\delta}\,\mathrm{d}w \quad , \end{aligned}$$

$s > \mathrm{S}$. Or pour tout $\theta \neq 0$,

$$\int_1^{1+\frac{u}{s}} w^{\theta}\,\mathrm{d}w = \frac{\left(1+\frac{u}{s}\right)^{\theta} - 1}{\theta}.$$

On définit la fonction h_θ telle que :

$$\int_0^{+\infty} \exp\left(\left|\frac{\mathrm{K}_2(s)}{\mathrm{K}_1(s)}\right| \frac{\left(1+\frac{u}{s}\right)^{\theta} - 1}{\theta} - u\right)\mathrm{d}u =: \int_0^{+\infty} h_\theta(s,u)\,\mathrm{d}u.$$

Or, lorsque $s \to +\infty$,

$$\frac{\mathrm{K}_2(s)}{\mathrm{K}_1(s)} \frac{\left(1+\frac{u}{s}\right)^{\theta} - 1}{\theta} \sim u \frac{\mathrm{K}_2(s)}{s\mathrm{K}_1(s)}.$$

Mais (voir preuve du Lemme 2, partie "Appendix : Auxiliary results" du Paragraphe 2.2),

$$\begin{aligned} \frac{\mathrm{K}_2(s)}{s\mathrm{K}_1(s)} &= \frac{\mathrm{K}_1(s)}{s} + \frac{1}{s}\left(\frac{s\mathrm{K}_1'(s)}{\mathrm{K}_1(s)} - 1\right) \\ &\overset{s\to+\infty}{\longrightarrow} 0 \end{aligned}$$

si $\theta_1 < 1$, ce qui est le cas par hypothèse, donc

$$h_\theta(s,u) \quad \overset{s\to+\infty}{\longrightarrow} e^{-u}$$

quel que soit $u \in \mathbb{R}$. Pour conclure la preuve, il reste alors à montrer qu'il est possible de passer à la limite dans l'intégrale afin de pouvoir écrire

$$\begin{aligned}\int_0^{+\infty} h_\theta(s,u)\,\mathrm{d}u \quad &\overset{s\to+\infty}{\longrightarrow} \quad \int_0^{+\infty} e^{-u}\,\mathrm{d}u \\ &= \quad 1,\end{aligned}$$

ce qui est le résultat voulu. Or, toujours au vu de la limite de $\dfrac{K_2(s)}{sK_1(s)}$, il existe s^*, tel que, quel que soit $s > s^*$,

$$\left| u\frac{K_2(s)}{sK_1(s)} \right| < \frac{1}{4}u.$$

En utilisant alors le fait que c'est un équivalent de $u\frac{K_2(s)}{sK_1(s)}$, on écrit :

$$\frac{K_2(s)}{K_1(s)} \left| \frac{\left(1+\frac{u}{s}\right)^\theta - 1}{\theta} \right| < \frac{1}{2}u.$$

Finalement, il existe s^*, tel que, quel que soit $s > s^*$,

$$|h_\theta(s,u)| = \exp\left(\left| \frac{K_2(s)}{K_1(s)} \right| \frac{\left(1+\frac{u}{s}\right)^\theta - 1}{\theta} - u \right) < \exp\left(-\frac{1}{2}u\right)$$

qui est intégrable, donc d'après le théorème de convergence dominée,

$$\begin{aligned}\int_0^{+\infty} h_\theta(s,u)\,\mathrm{d}u \quad &\overset{s\to+\infty}{\longrightarrow} \quad \int_0^{+\infty} e^{-u}\,\mathrm{d}u \\ &= \quad 1.\end{aligned}$$

En conclusion,

$$g(H^{-1}(s)) \quad \sim \quad \left(H^{-1}\right)'(s),\ s \to \infty$$

ou bien encore

$$g(t) \quad \sim \quad \frac{1}{H'(t)},\ t \to \infty.$$

□

Chapitre 3

Un nouvel estimateur des quantiles extrêmes basé sur le modèle "Log Weibull-tail généralisé"

Sommaire

Résumé

Nous proposons dans ce chapitre un nouvel estimateur des quantiles extrêmes basé sur le modèle des lois à queues de type log-Weibull généralisé, introduit par DE VALK *[2016b]. La proposition de cet estimateur et son étude sont l'objet d'un article soumis pour publication,* ALBERT *et collab. [2018a]. La Partie 3.1 pose les notations qui serviront à l'étude et s'attache à résumer les contributions de l'article en question. Ce dernier constitue la Partie 3.2. Nous y décrivons l'estimateur proposé et établissons sa normalité asymptotique. Ses performances sont alors comparées, à la fois sur données réelles et simulées, à l'estimateur proposé par* DE VALK *et* CAI *[2018]. Nous montrons que l'estimateur proposé est plus efficace dans certaines situations. Des perspectives de travail sont finalement données Partie 3.3.*

3.1 Motivations

Soit X une variable aléatoire de fonction de répartition F et de fonction de survie $\overline{F}(\cdot) := 1 - F(\cdot)$. Moyennant un échantillon $X_1, ..., X_n$ iid de même loi que X, nous nous intéressons à l'estimation du quantile extrême $q(p_n)$ de F. D'après la Définition 9, rappelons qu'un tel quantile est défini par $q(p_n) := \overline{F}^{\leftarrow}(p_n)$ avec $p_n \to 0$ lorsque $n \to \infty$.

Dans le cas où $np_n \to \infty$, le quantile $q(p_n)$ se trouve presque sûrement à l'intérieur de l'échantillon et il est alors possible de l'estimer par une simple valeur de l'échantillon ordonné (cf Paragraphe 1.3.1). Dans le cas où $np_n \to 0$, le quantile $q(p_n)$ se trouve cette fois-ci presque sûrement en dehors de l'échantillon et une simple statistique d'ordre ne suffit plus à son estimation (cf Paragraphe 1.3.1). Pour répondre à l'estimation de tels quantiles, une extrapolation en dehors de l'échantillon est nécessaire. Un état de l'art de ces méthodes d'extrapolation, basées sur la théorie des valeurs extrêmes, peut être trouvé dans DE HAAN et FERREIRA [2007, Chapitre 4] ou encore dans EMBRECHTS et collab. [2013, Chapitre 6].

La plupart de ces méthodes se basent sur la condition d'appartenance à un domaine d'attraction, c'est à dire sur l'hypothèse que la fonction quantile de queue $U(\cdot) := q(1/\cdot)$ est à variation régulière étendue (cf Théorème 14). Dans une série d'articles récents, DE VALK [2016a,b], Cees de Valk propose et discute une méthode alternative, consistant à mettre l'hypothèse de variation régulière étendue sur la fonction $V(\cdot) := \log U(\exp(\cdot)) = \log q(1/\exp(\cdot))$ plutôt que la fonction U (voir (1.64)). Il montre que cette hypothèse permet d'extrapoler plus loin (au sens où cette dernière autorise des suites p_n qui tendent vers zéro plus rapidement que sous les hypothèses des approches usuelles) tout en gardant une large portée d'applicabilité en termes de lois (voir le sous-paragraphe "Lois à queue de type log-Weibull généralisé" du Paragraphe 1.3.4 ou encore le Paragraphe 2.2, partie "Application to the ET approximation").

Dans l'article ci-dessous, nous proposons un estimateur des quantiles extrêmes basé sur le modèle développé par Cees de Valk. Sa construction est analogue à celle de l'estimateur de DE VALK et CAI [2018]. Nous en rappelons les étapes ci-dessous (cf équations (3.1-3.3)). Par définition des fonctions à variation régulière étendue d'indice θ (cf Définition 4), pour x assez grand, on a, quand $t \to \infty$:

$$\frac{V(tx) - V(x)}{a(x)} \simeq L_\theta(t) \tag{3.1}$$

ou encore

$$V(tx) \simeq V(x) + a(x)L_\theta(t). \tag{3.2}$$

Or $V(x) := \log q(e^{-x})$ et l'équation (3.2) se réecrit

$$q(e^{-tx}) \simeq q(e^{-x}) \exp\left(a(x)L_\theta(t)\right).$$

En remplaçant alors e^{-tx} par p_n et e^{-x} par α_n, il vient :

$$q(p_n) \simeq q(\alpha_n) \exp\left(a(\log 1/\alpha_n)L_\theta\left(\frac{\log 1/p_n}{\log 1/\alpha_n}\right)\right). \tag{3.3}$$

Puisque α_n est un niveau intermédiaire, $q(\alpha_n)$ est simplement estimé par $\hat{q}(\alpha_n) = X_{n-k_n,n}$ (cf Paragraphe 1.3.1), de manière analogue à l'estimateur proposé par DE VALK et CAI [2018]. L'estimateur que nous proposons se différencie dans l'estimation des quantités $a(\log 1/\alpha_n)$ et θ, inspirée, pour ce dernier, par l'estimateur des moments de DEKKERS et collab. [1989] (cf Paragraphe 1.3.3). En particulier, nous nous basons sur les statistiques suivantes :

$$M_n^{(j)} := \frac{1}{k_n} \sum_{i=0}^{k_n-1} \left[\log_2(X_{n-i,n}) - \log_2(X_{n-k_n,n})\right]^j,$$

pour tout $j \in \mathbb{N}$ avec $\log_2 := \log\log$. Contrairement à l'estimateur de DEKKERS et collab. [1989] du Paragraphe 1.3.3, c'est l'accroissement des doubles logarithmes et non pas des simples logarithmes qui est la clé de voûte des estimateurs que nous proposons dans l'article Partie 3.2.

Moyennant des estimateurs de $a(\log 1/\alpha_n)$ et θ basés sur ces statistiques, nous montrons alors que l'estimateur des quantiles extrêmes associé est asymptotiquement gaussien. Ceci est fait dans le cas intermédiaire et extrême. Le comportement en pratique dudit estimateur est alors évalué sur données réelles et simulées. Une comparaison avec l'estimateur de DE VALK et CAI [2018] est également proposée.

3.2 Un nouvel estimateur des quantiles extrêmes basé sur le modèle des queues de type log-Weibull généralisé

Les résulats sont présentés ci-dessous sous la forme d'un article soumis pour publication, voir ALBERT et collab. [2018a]. Les notations adoptées dans l'article sont légèrement différentes de celles introduites

Chapitre 1. Ainsi,

$$S(\cdot) := 1 - F(\cdot)$$

joue le rôle de la fonction de survie précédemment notée $\overline{F}$,

$$Q(\cdot) := \overline{F}^{\leftarrow}(\cdot)$$

joue le rôle de la fonction quantile q et β_n le rôle d'un ordre extrême, précédemment noté p_n.

An extreme quantile estimator for the log-generalized Weibull-tail model

Clément Albert[(1)], Anne Dutfoy[(2)], Laurent Gardes[(3)] and Stéphane Girard[(1, *)]

[(1)] *Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France*

[(2)] *EDF R&D dept. Périclès, 91120 Palaiseau, France*

[(3)] *Université de Strasbourg & CNRS, IRMA, UMR 7501, 7 rue René Descartes, 67084 Strasbourg Cedex, France*

Abstract

We propose a new estimator for extreme quantiles under the log-generalized Weibull-tail model, introduced in [de Valk, C. (2016). Approximation of high quantiles from intermediate quantiles, *Extremes*, 19(4), 661–686]. This model relies on a new regular variation condition which, in some situations, permits to extrapolate further into the tails than the classical assumption in extreme-value theory. The asymptotic normality of the estimator is established and its finite sample properties are illustrated both on simulated and real datasets.

Keywords: Extreme quantiles, Extreme-value theory, Extended regular variation.

AMS 2000 subject classification: 62G32, 62G20.

1 Introduction

Let X be a random variable with distribution function $F(\cdot) = \mathbb{P}(X \leq \cdot)$ and survival function $S := 1 - F$. Starting from a $n-$sample from X, our goal is to estimate extreme quantiles from S of level $1 - \beta_n$ with $\beta_n \to 0$ as $n \to \infty$. Recall that a quantile of level $1 - \beta$ is given by $Q(\beta) := \inf\{y;\ S(y) \leq \beta\}$. The rate of convergence of β_n to zero drives the difficulty of the estimation problem. Indeed, if $n\beta_n \to 0$ as $n \to \infty$, then $Q(\beta_n)$ is asymptotically almost surely larger than the sample maxima. In finance or insurance contexts, an extreme quantile is interpreted as the Value-at-Risk associated with an extreme loss, see [10, 19] for links between extreme-value theory and risk theory. In environmental applications, an extreme quantile coincides with the return level associated with an exceptional climatic event (extreme rainfalls [6], extreme wind velocities [16], extreme wave heights [18], river peak flows [17],...).

Dedicated methods have been designed to address the estimation of extreme quantiles, see [9, Chapter 6] or [15, Chapter 4], for an overview. Most of them rely on an extended regular variation assumption on the function Q. Recently, an alternative method has been initiated by Cees de Valk in a series of papers [21, 22], the goal being to estimate "more" extreme quantiles *i.e.* associated with sequences β_n tending to zero at a faster rate than in the previously mentioned

*Corresponding author, Stephane.Girard@inria.fr

approaches [9, 15]. The idea is to put the extended regular variation assumption on the function $V(\cdot) := \ln Q(1/\exp\cdot)$ rather than on $Q(\cdot)$, see Paragraph 1.1 for technical details and Paragraph 1.2 for examples. Dedicated estimation methods are introduced in [23]. The goal of this work is to contribute to the popularity of this model by proposing alternative estimators, which are more efficient than the initial ones [23] in some situations.

1.1 Tail model

Let X be a random variable with survival function S. For the sake of simplicity, we assume in what follows that $S(1) = 1$ *i.e.* X is almost surely larger than 1. The tail model considered in this work is given by

$$S(x) = \exp(-V^{\leftarrow}(\ln x)),\ x \geq 1, \tag{1}$$

where $V^{\leftarrow}(\cdot) := \inf\{y;\ V(y) \geq \cdot\}$ is the generalized inverse of $V(\cdot) = \ln Q(1/\exp\cdot)$ with Q the quantile function. The function V is supposed to be of extended regular variation with index $\theta \in \mathbb{R}$. More specifically, there exists a positive function a (called the auxiliary function) such that

$$\lim_{x\to\infty} \frac{V(tx) - V(x)}{a(x)} = \int_1^t u^{\theta-1}du =: L_\theta(t),\ \text{for all } t > 0. \tag{2}$$

The class of extended regularly varying functions is denoted by $\mathcal{ERV}(\theta)$. Model (1) is referred to as the "log-generalized Weibull-tail model" [21, 22, 23]. From [15, Corollary 1.1.10], a sufficient condition for (2) is

(A1) V is differentiable with derivative V' satisfying

$$\lim_{x\to\infty} \frac{V'(tx)}{V'(x)} = t^{\theta-1}. \tag{3}$$

Such a function V' is said to be regularly varying with index $\theta - 1$ and this property is denoted by $V' \in \mathcal{RV}(\theta - 1)$. We refer to [5] for a general account on regular variation theory. Moreover, under **(A1)**, a possible choice of auxiliary function in (2) is $a(x) = xV'(x)$.

1.2 Properties and examples

Condition **(A1)** generalizes the tail model introduced in [8, 12] where it is assumed that the function V in (2) is asymptotically proportional to L_τ for some $\tau \in [0, 1]$. One can then easily show that such a tail parameter τ coincides with the index θ of extended regular variation in the situation where $\theta \in [0, 1]$. In terms of Maximum Domain of Attraction (MDA), the following result has been established in [2, Proposition 4]:

Lemma 1 *Assume F is twice differentiable.*

(i) If **(A1)** *holds with $\theta < 1$ then $F \in$ MDA(Gumbel).*

(ii) If $F \in$ MDA(Fréchet) then **(A1)** *holds with $\theta = 1$.*

(iii) If **(A1)** *holds with $\theta > 1$ then F does not belong to any MDA.*

It thus appears that model **(A1)** with $\theta \leq 1$ is of particular interest since it is associated with most distributions in MDA(Gumbel) $\cup$ MDA(Fréchet). The situation $\theta > 1$ which does not correspond to any domain of attraction is sometimes referred to as super-heavy tails, see for instance [3]. The following examples are taken from [2, Proposition 3]:

Example 1 *Let $x^* := \sup\{x \geq 1,\, F(x) < 1\}$ be the endpoint of F. Then, under some monotonicity assumptions:*

(i) If $V^{\leftarrow}(\ln \cdot) \in \mathcal{RV}(1/\beta)$, $\beta > 0$, then **(A1)** *holds with $\theta = 0$. In this case, F is referred to as a Weibull tail-distribution, see for instance [4, 11, 14]. Such distributions encompass Gaussian, Gamma, Exponential and strict Weibull distributions.*

(ii) $V^{\leftarrow} \in \mathcal{RV}(1/\beta)$, $0 < \beta < 1$ if and only if **(A1)** *holds with $\theta = \beta > 0$. Here, F is called a log-Weibull tail-distribution, see [3, 8, 12], the most popular example being the lognormal distribution.*

(iii) $1 \leq x^* < \infty$ and $V^{\leftarrow}(\ln x^* + \ln(1 - 1/\cdot)) \in \mathcal{RV}_{-1/\beta}$, $\beta < 0$ if and only if **(A1)** *holds with $\theta = \beta < 0$. This case corresponds to distributions with a Weibull tail behavior in the neighborhood of a finite endpoint.*

We also refer to Table 1 for examples of distributions corresponding to the three above families: $\theta = 0$, $\theta > 0$ and $\theta < 0$.

1.3 Outline

The inference aspects associated with model (1) are examined in Section 2: Estimators for extreme quantiles are introduced as well as estimators for the extended regular variation index θ and the auxiliary function a. The asymptotic distributions of these estimators are established in Section 3. Their finite sample performance are investigated in Section 4 on simulated data and compared to the proposals introduced in [23]. Finally, an illustration on real data is presented in Section 5. Proofs are postponed to Section 6.

2 Inference

Let $X_1, \ldots, X_n$ be n independent copies of a random variable X distributed as in (1). The associated ordered statistics are denoted by $X_{1,n} \leq \ldots \leq X_{n,n}$ throughout the paper. Starting from this random sample, we focus on the estimation of extreme quantiles *i.e.* $Q(u) := S^{\leftarrow}(u) = \exp(V(\ln(1/u)))$ when $u \to 0$. Two situations for the level u are considered.

Intermediate case. If $u = \alpha_n$ where α_n is an intermediate level satisfying $\alpha_n \to 0$ and $n\alpha_n \to \infty$ as $n \to \infty$, a natural estimator is obtained by replacing Q by its empirical counterpart $\hat{Q}_n$. More precisely, $Q(\alpha_n)$ is estimated by $\hat{Q}_n(\alpha_n) = X_{n-\lfloor n\alpha_n \rfloor, n}$.

Extreme case. If $u = \beta_n$ where β_n is an extreme level such that $n\beta_n \to c \geq 0$ as $n \to \infty$, a simple order statistic cannot be used. Extrapolation beyond the sample should be performed.

Starting from an intermediate level $\alpha_n := k_n/n$ where $k_n \to \infty$ and $k_n/n \to 0$, we propose to estimate $Q(\beta_n)$ by

$$\check{Q}_n(\beta_n) := X_{n-k_n,n} \exp\left(\hat{a}_n(\ln(n/k_n)) L_{\hat{\theta}_n}\left(\frac{\ln \beta_n}{\ln(k_n/n)}\right)\right), \tag{4}$$

where $\hat{\theta}_n$ and $\hat{a}_n(\ln(n/k_n))$ are suitable estimators of θ and $a(\ln(n/k_n))$. The rationale behind (4) is based on (2) which basically means that for α close to 0 and for all $t > 0$,

$$\ln Q(t\alpha) \approx \ln Q(\alpha) + a(\ln(1/\alpha)) L_\theta\left(1 + \frac{\ln(t)}{\ln(\alpha)}\right).$$

Estimator (4) is then obtained by taking $\alpha = k_n/n$ and $t = n\beta_n/k_n$ and by replacing the unknown quantities $Q(k_n/n)$, $a(\ln(n/k_n))$ and θ by their corresponding estimators. Since k_n/n is an intermediate level, $Q(k_n/n)$ is estimated by $\hat{Q}_n(k_n/n) = X_{n-k_n,n}$.

Parameters estimation. Let us now propose new estimators of θ and $a(\ln(n/k_n))$. To this end, for $j \in \{1, 2\}$, consider the statistic

$$M_n^{(j)} := \frac{1}{k_n} \sum_{i=0}^{k_n-1} \left(\ln_2(X_{n-i,n}) - \ln_2(X_{n-k_n,n})\right)^j,$$

where $\ln_2 := \ln \ln$, as well as the functions

$$\mu_b(x, \zeta) := \int_0^1 L_\zeta^b\left(1 + \frac{\ln(1/s)}{x}\right) ds \text{ and } \Psi_x(\zeta) := \frac{\mu_1^2(x,\zeta)}{\mu_2(x,\zeta)},$$

defined for $x > 0$, $b \in \mathbb{N} \setminus \{0\}$ and $\zeta < 1$. Let us mention that $\mu_1(x, 0) = \mathrm{e}^x E_1(x)$ where $E_1(x) := \int_x^\infty u^{-1}\mathrm{e}^{-u}du$ is the exponential integral, see for instance [1, eq 5.1.1]. Furthermore, it can be shown (see Lemma 5) that Ψ_x is a decreasing function, at least for x large enough, and thus its generalized inverse $\Psi_x^{\leftarrow}$ is well defined for x large enough. The following statistics are then introduced:

$$\hat{\theta}_{n,+}^{(M)} := \frac{M_n^{(1)}}{\mu_1(\ln(n/k_n), 0)}, \tag{5}$$

$$\hat{\theta}_{n,-}^{(M)} := \Psi_{\ln(n/k_n)}^{\leftarrow}\left(\frac{(M_n^{(1)})^2}{M_n^{(2)}}\right), \tag{6}$$

$$\hat{\theta}_n^{(M)} := \hat{\theta}_{n,+}^{(M)} + \hat{\theta}_{n,-}^{(M)}, \tag{7}$$

$$\hat{a}_n^{(M)}(\ln(n/k_n)) := \frac{\ln X_{n-k_n,n}}{\mu_1(\ln(n/k_n), \hat{\theta}_{n,-}^{(M)})} M_n^{(1)}. \tag{8}$$

We conclude this section by giving the main ideas leading to the estimators (7) and (8) of respectively θ and $a[\ln(n/k_n)]$. The estimator (7) is similar in spirit to the moment estimator introduced in [7]. Its construction is based on the following two results. Letting $\theta_+ := \theta \vee 0$ and $\theta_- := \theta \wedge 0$, for any increasing function $V \in \mathcal{ERV}(\theta)$,

$$\lim_{x\to\infty} \frac{V(x)}{a(x)} \ln \frac{V(tx)}{V(x)} = L_{\theta_-}(t), \tag{9}$$

locally uniformly in $(0,\infty)$, see [15, Lemma B.3.16]. Moreover, one has (see for instance [15, Eq. 3.5.5]),

$$\lim_{x\to\infty}\frac{a(x)}{V(x)}=\theta_+.$$

Plugging $x:=\ln(1/\alpha)$ and $t:=1+\ln(s)/\ln(\alpha)$ in (9) yields the approximation

$$\ln_2 Q(s\alpha)-\ln_2 Q(\alpha)\approx\theta_+ L_0\left(1+\frac{\ln s}{\ln\alpha}\right), \tag{10}$$

as $\alpha\to 0$ and for all $s\in(0,1)$. Integrating with respect to s on $(0,1)$ leads to

$$\int_0^1 \ln_2 Q(s\alpha)-\ln_2 Q(\alpha)ds\Big/\int_0^1 L_0\left(1+\frac{\ln s}{\ln\alpha}\right)ds\approx\theta_+.$$

Considering $\alpha=k_n/n$ where k_n is an intermediate sequence such that $k_n\to\infty$ and $k_n/n\to 0$ and replacing Q by its empirical estimator $\hat{Q}_n$ lead to the estimator (5) of θ_+. Similarly, remark that (10) leads to the approximation

$$\left(\int_0^1 \ln_2 Q(s\alpha)-\ln_2 Q(\alpha)ds\right)^2\Big/\int_0^1\left(\ln_2 Q(s\alpha)-\ln_2 Q(\alpha)\right)^2 ds\approx\Psi_{\ln(1/\alpha)}(\theta_-),$$

as $\alpha\to 0$. Replacing again in the previous approximation α by k_n/n and Q by its empirical counterpart suggests to estimate θ_- by (6). Finally, estimator (8) is obtained by remarking that, from (10):

$$\frac{\ln Q(\alpha)}{a(\ln(1/\alpha))}\int_0^1\ln\frac{\ln Q(s\alpha)}{\ln Q(\alpha)}ds\approx\mu_1(\ln(1/\alpha),\theta_-),$$

for α close to 0. Replacing α by k_n/n, Q by $\hat{Q}_n$ and θ_- by $\hat{\theta}_{n,-}^{(M)}$ gives (8).

3 Main results

3.1 Quantile estimation: Intermediate case

Let us first focus on the asymptotic behavior of the quantile estimator in the intermediate case.

Theorem 1 *Under model (1), assume that* **(A1)** *holds. For all intermediate level α_n (i.e. such that $\alpha_n\to 0$ and $n\alpha_n\to\infty$ as $n\to\infty$), one has*

$$\frac{(n\alpha_n)^{1/2}\ln(1/\alpha_n)}{a(\ln(1/\alpha_n))}\ln\left(\frac{\hat{Q}_n(\alpha_n)}{Q(\alpha_n)}\right)\xrightarrow{d}\mathcal{N}(0,1).$$

First, remark that introducing $k_n=\lfloor n\alpha_n\rfloor$ and choosing $a(t)=tV'(t)$ (see Paragraph 1.1), the above asymptotic normality result can be rewritten as

$$\frac{k_n^{1/2}}{V'(\ln(n/k_n))}\ln\left(\frac{\hat{Q}_n(k_n/n)}{Q(k_n/n)}\right)\xrightarrow{d}\mathcal{N}(0,1).$$

If, moreover,

$$k_n^{1/2}/V'(\ln(n/k_n))\to\infty\text{ as }n\to\infty, \tag{11}$$

then

$$\frac{k_n^{1/2}}{(n/k_n)U'(n/k_n)}\left(\hat{Q}_n(k_n/n)-Q(k_n/n)\right)\xrightarrow{d}\mathcal{N}(0,1),$$

where $U(\cdot)=Q(1/\cdot)$ is the tail quantile function. This result coincides with [15, Theorem 2.2.1] established under a von Mises' condition for the maximum domain of attraction of an extreme-value distribution. Clearly, (11) holds when $\theta<1$ since, in this case, $V'(t)\to 0$ as $t\to\infty$. Moreover, if $F\in$ MDA(Fréchet) then $\theta=1$ from Lemma 1(ii) and $U\in\mathcal{RV}(\gamma)$ for some $\gamma>0$. It thus follows that $V'(\ln t)=tU'(t)/U(t)\to\gamma$ as $t\to\infty$ and (11) is verified. The case $\theta>1$ is not relevant here, since, in this case, F does not belong to any domain of attraction, see Lemma 1(iii).
Second, under additional conditions,

$$\frac{\hat{a}_n^{(M)}(\ln(1/\alpha_n))}{a(\ln(1/\alpha_n))}\xrightarrow{\mathbb{P}}1,$$

see Theorem 4 below, and thus

$$\frac{(n\alpha_n)^{1/2}\ln(1/\alpha_n)}{\hat{a}_n^{(M)}(\ln(1/\alpha_n))}\ln\left(\frac{\hat{Q}_n(\alpha_n)}{Q(\alpha_n)}\right)\xrightarrow{d}\mathcal{N}(0,1)$$

which provides a way for constructing asymptotic confidence intervals for intermediate quantiles $Q(\alpha_n)$ based on $\hat{Q}_n(\alpha_n)$. Letting u_ζ the $(1+\zeta)/2$th quantile from a standard Gaussian distribution,

$$\left[\hat{Q}_n(\alpha_n)\exp\left(-\frac{\hat{a}_n^{(M)}(\ln(1/\alpha_n))}{(n\alpha_n)^{1/2}\ln(1/\alpha_n)}u_\zeta\right);\hat{Q}_n(\alpha_n)\exp\left(\frac{\hat{a}_n^{(M)}(\ln(1/\alpha_n))}{(n\alpha_n)^{1/2}\ln(1/\alpha_n)}u_\zeta\right)\right]$$

is an asymptotic confidence interval for $Q(\alpha_n)$ of confidence level ζ.

3.2 Quantile estimation: Extreme case

Our next goal is to establish the asymptotic normality of $\check{Q}_n(\beta_n)$ for an extreme level β_n satisfying $n\beta_n\to c\geq 0$. A second-order condition is needed on $V\in\mathcal{ERV}(\theta)$ to control the rate of convergence in (2):

(A2) There exist a function $\tilde{A}$ with $\tilde{A}(x)\to 0$ as $x\to\infty$ and $\rho<0$ such that for all $t>0$,

$$\lim_{x\to\infty}\frac{1}{\tilde{A}(x)}\left(\frac{V(tx)-V(x)}{a(x)}-L_\theta(t)\right)=H_{\theta,\rho}(t):=\int_1^t u^{\theta-1}L_\rho(u)du.$$

locally uniformly for $t>0$.

Note that **(A2)** also provides the rate of convergence in (9). Indeed, from [15, Lemma B.3.16], condition **(A2)** with $\theta\neq\rho$ entails that there exists a function A with $A(x)\to 0$ as $x\to\infty$ such that

$$\lim_{x\to\infty}\frac{1}{A(x)}\left(\frac{V(x)}{a(x)}\ln\frac{V(tx)}{V(x)}-L_{\theta_-}(t)\right)=H_{\theta_-,\rho'}(t). \tag{12}$$

The function $|A|$ is regularly varying with index $\rho'\leq 0$ where, according to [15, Lemma B.3.16],

$$\rho'=\begin{cases}\rho & \text{if } \theta<\rho,\\ \theta & \text{if } \rho<\theta\leq 0,\\ -\theta & \text{if } (0<\theta<-\rho \text{ and } l\neq 0),\\ \rho & \text{if } (0<\theta<-\rho \text{ and } l=0) \text{ or } (\theta\geq-\rho),\end{cases} \tag{13}$$

with, for $\theta > 0$,

$$l := \lim_{x\to\infty} \left(V(x) - \frac{a(x)}{\theta} \right).$$

Let us also introduce the positive function B defined by $B(x) := \max(|\tilde{A}(x)|, |A(x)|/x)$. It is easily checked that B is regularly varying with index $\rho'' := \max(\rho, \rho' - 1)$. We are now in position to establish the asymptotic distribution of $\check{Q}_n(\beta_n)$ for general estimators of θ and $a(\ln(n/k_n))$ satisfying the condition:

(A3) There exist a sequence $\sigma_n \to 0$ and a random vector (B, Θ, Λ) such that

$$\sigma_n^{-1} \left\{ \frac{\ln X_{n-k_n,n} - \ln Q(k_n/n)}{a(\ln(n/k_n)) H_{\theta,0}(d_n)}, \hat{\theta}_n - \theta, \frac{L_\theta(d_n)}{H_{\theta,0}(d_n)} \left(\frac{\hat{a}_n(\ln(n/k_n))}{a(\ln(n/k_n))} - 1 \right) \right\} \xrightarrow{d} (\Omega, \Theta, \Lambda),$$

where $d_n := \ln(1/\beta_n)/\ln(n/k_n)$.

Theorem 2 *Under model (1), assume conditions* **(A2)**, **(A3)** *hold. Let (k_n) and (β_n) be two sequences such that $n\beta_n \to c \geq 0$, $k_n \to \infty$, $k_n/n \to 0$, $d_n \to d \in [1, \infty]$, $\sigma_n \ln(d_n) \to 0$ and $\sigma_n^{-1} \tilde{A}(\ln(n/k_n)) \to 0$ as $n \to \infty$. Then,*

$$\frac{\sigma_n^{-1}}{a(\ln(n/k_n)) H_{\theta,0}(d_n)} \ln \left(\frac{\check{Q}_n(\beta_n)}{Q(\beta_n)} \right) \xrightarrow{d} \Omega + \Theta + \Lambda.$$

Under the conditions of Theorem 2, three situations can arise for the extreme quantile level β_n. The first one is when $d_n \to 1$ which corresponds to the least extreme case. This condition is achieved for instance when $n\beta_n \to c > 0$. In this situation, a Taylor expansion yields

$$H_{\theta,0}(d_n) \overset{d_n\to 1}{\sim} (d_n - 1)^2/2 \to 0. \tag{14}$$

The second case corresponds to the situation where $d_n \to d \in (1, \infty)$. Here,

$$H_{\theta,0}(d_n) \overset{d_n\to d}{\longrightarrow} H_{\theta,0}(d) > 0. \tag{15}$$

Note that for these two situations, $\sigma_n \ln(d_n) \to 0$ is a consequence of the assumption $\sigma_n \to 0$ as $n \to \infty$. Finally, the most extreme case occurs when $d_n \to \infty$ leading to

$$H_{\theta,0}(d_n) \overset{d_n\to\infty}{\sim} \begin{cases} d_n^\theta \ln(d_n)/\theta & \text{if } \theta > 0, \\ \ln^2(d_n)/2 & \text{if } \theta = 0, \\ 1/\theta^2 & \text{if } \theta < 0. \end{cases} \tag{16}$$

As expected, the rate of convergence in Theorem 2 is getting worse when the quantile level β_n is getting more extreme. Let us also highlight that, when $\theta < 0$, the rates of convergence in situations $d_n \to d > 1$ and $d_n \to \infty$ are of the same order.

To conclude this section, let us give the following consistency result.

Proposition 1 *Under the conditions of Theorem 2,*

$$\frac{\hat{a}_n(\ln(n/k_n))}{a(\ln(n/k_n))} \xrightarrow{\mathbb{P}} 1 \text{ and } \frac{H_{\hat{\theta}_n,0}(d_n)}{H_{\theta,0}(d_n)} \xrightarrow{\mathbb{P}} 1. \tag{17}$$

and therefore

$$\frac{\sigma_n^{-1}}{\hat{a}_n(\ln(n/k_n)) H_{\hat{\theta}_n,0}(d_n)} \ln \left(\frac{\check{Q}_n(\beta_n)}{Q(\beta_n)} \right) \xrightarrow{d} \Omega + \Theta + \Lambda.$$

Proposition 1 can be used to construct asymptotic confidence intervals for extreme quantiles $Q(\beta_n)$ based on $\check{Q}_n(\beta_n)$, see (19) below.

3.3 Parameters estimation

First, the asymptotic distribution of the estimator of θ proposed in (7) is provided.

Theorem 3 *Under model (1), assume that condition* **(A2)** *holds with* $\theta \neq \rho$. *Let* (k_n) *be a sequence such that* $k_n/\ln^2(n) \to \infty$, $k_n/n \to 0$ *and* $k_n A^2(\ln(n/k_n))/\ln^2(n/k_n) \to 0$ *as* $n \to \infty$. *Then,*

$$\frac{k_n^{1/2}}{\ln(n/k_n)}\left(\hat{\theta}_n^{(M)} - \theta\right) \xrightarrow{d} \mathcal{N}(0,1).$$

It is shown in the proof of Theorem 3 that the negative part $\hat{\theta}_{n,-}^{(M)}$ of the estimator converges slower than the positive part $\hat{\theta}_{n,+}^{(M)}$, see (25) and (27). As a consequence, $\hat{\theta}_n^{(M)}$ inherits its asymptotic normality from $\hat{\theta}_{n,-}^{(M)}$. This phenomenon can be explained by the fact that $\hat{\theta}_{n,-}^{(M)}$ is obtained through the inversion of the function $\Psi_{\ln(n/k_n)}$. The rate of convergence of $\hat{\theta}_{n,-}^{(M)}$ thus depends on the first derivative of $\Psi_{\ln(n/k_n)}$ which converges to 0 as $n \to \infty$, see Lemma 5(ii). Note also that from [13, Lemma 1], condition $k_n A^2(\ln(n/k_n))/\ln^2(n/k_n) \to 0$ implies $\ln(k_n)/\ln(n) \to 0$ as $n \to \infty$ and thus $\ln(n/k_n) \sim \ln(n)$. Second, the asymptotic distribution of the estimator of $a(\ln(n/k_n))$ proposed in (8) is established in the following theorem.

Theorem 4 *Under model (1), assume that condition* **(A2)** *holds with* $\theta \neq \rho$. *Let* (k_n) *be a sequence such that* $k_n/\ln^2(n) \to \infty$, $k_n/n \to 0$ *and* $k_n B^2(\ln(n/k_n)) \to 0$ *as* $n \to \infty$. *Then,*

$$k_n^{1/2}\left(\frac{\hat{a}_n^{(M)}(\ln(n/k_n))}{a(\ln(n/k_n))} - 1\right) \xrightarrow{d} \mathcal{N}(0,2).$$

Note that, if $\rho > -1$ and $k_n \tilde{A}^2(\ln(n)) \to 0$, then $k_n/\ln^2(n) \to 0$. Hence, Theorem 4 does not apply when $\rho \in (-1, 0)$. Let us stress that this limitation also appears in [8, Theorem 1]. As a straightforward consequence of Theorems 1 – 4, the asymptotic normality of the extreme quantile estimator $\check{Q}_n^{(M)}(\beta_n)$ is obtained by considering $\hat{\theta}_n = \hat{\theta}_n^{(M)}$ and $\hat{a}_n(\ln(n/k_n)) = \hat{a}_n^{(M)}(\ln(n/k_n))$ in (4).

Corollary 1 *Under model (1), assume that* **(A2)** *holds with* $\theta \neq \rho$. *Let* (k_n) *and* (β_n) *be two sequences such that* $n\beta_n \to c \geq 0$, $k_n/n \to 0$, $k_n B^2(\ln(n/k_n)) \to 0$, $d_n \to d \in [1, \infty]$ *and* $(\ln(n)\max(1, \ln(d_n)))^2/k_n \to 0$ *as* $n \to \infty$. *Then,*

$$\frac{k_n^{1/2}/\ln(n/k_n)}{a(\ln(n/k_n))H_{\theta,0}(d_n)} \ln\left(\frac{\check{Q}_n^{(M)}(\beta_n)}{Q(\beta_n)}\right) \xrightarrow{d} \mathcal{N}(0,1).$$

The proof of this result consists in showing that the estimators $\hat{\theta}_n^{(M)}$ and $\hat{a}_n^{(M)}(\ln(n/k_n))$ satisfy condition **(A3)** with $(\Omega, \Theta, \Lambda) = (0, \Theta, 0)$ where Θ is a standard Gaussian random variable. Hence, in this situation, only the estimator of θ contributes to the asymptotic distribution of $\check{Q}_n^{(M)}(\beta_n)$. It appears that $\ln(n/k_n)a(\ln(n/k_n))H_{\theta,0}(d_n)/k_n^{1/2} \to 0$ is a sufficient condition to ensure that $\check{Q}_n^{(M)}(\beta_n)$ is a relatively consistent estimator of $Q(\beta_n)$, *i.e.* such that $\check{Q}_n^{(M)}(\beta_n)/Q(\beta_n) \xrightarrow{\mathbb{P}} 1$. Recalling that $\ln(n/k_n) \sim \ln n$ as $n \to \infty$, that $B \in \mathcal{RV}(\rho'')$ and $a \in \mathcal{RV}(\theta)$, we end up with a set of three conditions on the sequences (k_n) and (β_n): $k_n B^2(\ln n) \to 0$, $(\ln(n)\max(1, \ln(d_n)))^2/k_n \to 0$ and $(\ln(n)a(\ln n)H_{\theta,0}(d_n))^2/k_n \to 0$ as $n \to \infty$. Let us illustrate how these conditions may limit the extrapolation range β_n depending on the index θ of extended regular variation in three typical situations:

- Let $\beta_n = c/n$, $c \in (0,1)$. Here $d_n \to 1$ as $n \to \infty$, this is the least extreme case considered in Subsection 3.2, and, in view of (14), $H_{\theta,0}(d_n) \sim (\ln(k_n)/\ln(n))^2/2$. Two constraints arise on the distribution parameters: $\rho \leq -1$ and $\theta \leq 2 - \rho''$. The first one, $\rho \leq -1$, was already imposed by Theorem 4. The second one is fulfilled as soon as $\theta \leq 2$ including MDA(Fréchet), see Lemma 1(ii), finite endpoint, Weibull-tail, log-Weibull tail distributions defined in Example 1 and some super-heavy tail distributions. As an example, all distributions of Table 1 satisfy the above constraints.

- Let $\beta_n = n^{-\tau}$, $\tau > 1$. Here, $d_n = \tau$, this is the second extreme case considered in Subsection 3.2, and, as a particular case of (15), $H_{\theta,0}(d_n)$ is constant. The constraints are: $\rho \leq -1$ and $\theta \leq -1 - \rho''$. The condition on θ is fulfilled by finite endpoint distributions of Example 1(iii), Weibull-tail distributions (Example 1(i)) and some log-Weibull tail distributions (Example 1(ii)). In MDA(Fréchet), $\theta = 1$ and thus the condition on the second order parameters is strengthened: $\rho \leq -2$ and $\rho' \leq -1$. As an example, in Table 1, Lognormal and Pareto-like distributions do not satisfy the above constraints.

- Let $\beta_n = \exp(-cn)$, $c > 0$. Here $d_n \to \infty$ as $n \to \infty$, this is the most extreme case considered in Subsection 3.2. In view of (16), three subcases have to be considered. If $\theta < 0$ then $H_{\theta,0}(d_n)$ is asymptotically constant and the conditions are $\rho \leq -2$ and $\rho' \leq -1$. If $\theta = 0$ then necessarily $\rho' = 0$ in view of (13), $H_{\theta,0}(d_n) \sim (\ln n)^2/2$ and it is not possible to find sequences satisfying the constraints. If $\theta > 0$ then $H_{\theta,0}(d_n) \sim (c^\theta/\theta) n^\theta (\ln n)^{1-\theta}$ and it is not possible either to find sequences satisfying the constraints.

It thus appears that only the first two cases $\beta_n = c/n$, $c \in (0,1)$ and $\beta_n = n^{-\tau}$, $\tau > 1$ are of practical interest. The third situation $\beta_n = \exp(-cn)$, $c > 0$ can be addressed only when $\theta < 0$, *i.e.* for finite endpoint distributions. For such distributions, the estimation of very extreme quantiles boils down to estimating the endpoint. In the first two cases, a possible choice of the intermediate sequence when $\rho'' < -1$ is $k_n = (\ln n)^{-2\rho''-\varepsilon}$ where $\varepsilon > 0$ is arbitrarily small. Moreover, in the second case where $\beta_n = n^{-\tau}$, $\tau > 1$, it is possible to compare the asymptotic standard deviation of $\ln \check{Q}_n^{(M)}(\beta_n)$, denoted by σ_n, to the one associated with the estimator introduced in [23], denoted by σ'_n. Our Corollary 1 and [23, Corollary 2] yield:

$$\begin{aligned} \sigma_n &\sim H_{\theta,0}(\tau) k_n^{-1/2} (\ln n) a(\ln n), \\ \sigma'_n &\sim (L_\theta^2(\tau) + H_{\theta,0}^2(\tau))^{1/2} k_n^{-1/2} (\ln n) a(\ln n). \end{aligned}$$

As a consequence, the asymptotic standard deviations are equivalent up to a multiplicative constant:

$$\frac{\sigma_n}{\sigma'_n} \to \left(1 + \frac{L_\theta^2(\tau)}{H_{\theta,0}^2(\tau)}\right)^{-1/2} =: \Lambda_\theta(\tau) \leq 1 \text{ as } n \to \infty. \tag{18}$$

The behavior of $\Lambda_\theta(\tau)$ with respect to θ and τ is illustrated on Figure 1. It appears that $\Lambda_\theta(\tau)$ is an increasing function of τ and θ. As expected $\Lambda_\theta(\tau) \leq 1$ meaning that $\check{Q}_n^{(M)}(\beta_n)$ is asymptotically more efficient than [23]'s competitor, especially when θ is small.
Finally, in view of Proposition 1, the unknown quantities $H_{\theta,0}(d_n)$ and $a(\ln(n/k_n))$ can be replaced by their corresponding estimators $H_{\hat{\theta}_n^{(M)},0}(d_n)$ and $\hat{a}_n^{(M)}(\ln(n/k_n))$ without changing the asymptotic

distribution in Corollary 1. As mentioned before, the obtained result can then lead to asymptotic confidence intervals. Letting u_ζ the $(1+\zeta)/2$th quantile from a standard Gaussian distribution,

$$\check{Q}_n^{(M)}(\beta_n)\left[\exp\left(-\frac{\hat{a}_n^{(M)}(\ln(n/k_n))H_{\hat{\theta}_n^{(M)},0}(d_n)}{k_n^{1/2}/\ln(n/k_n)}u_\zeta\right);\exp\left(\frac{\hat{a}_n^{(M)}(\ln(n/k_n))H_{\hat{\theta}_n^{(M)},0}(d_n)}{k_n^{1/2}/\ln(n/k_n)}u_\zeta\right)\right] \quad (19)$$

is an asymptotic confidence interval for $Q(\beta_n)$ of confidence level ζ.

4 Validation on simulations

The finite-sample behavior of the quantile estimator $\check{Q}_n^{[1]}(\beta_n) := \check{Q}_n(\beta_n)$ defined in (4) is investigated on $N = 500$ simulated random samples of size $n = 5000$, in the case where $\beta_n = n^{-2} = 4.10^{-8}$.

Estimators. Three competitors are considered:

1. The first one, $\check{Q}_n^{[2]}(\beta_n)$ is deduced from (4) by letting $\hat{\theta}_{n,-}^{(M)} := 0$ in $\hat{a}_n^{(M)}(\ln(n/k_n))$ and $\hat{\theta}_n^{(M)}$, see (8) and (7). The resulting estimator $\check{Q}_n^{[2]}(\beta_n)$ should perform well for estimating extreme quantiles from distributions with associated $\theta \geq 0$.

2. Similarly, the second one is also obtained by letting $\hat{\theta}_n^{(M)} := 0$ in (4) and $\hat{\theta}_{n,-}^{(M)} := 0$ in (8). We thus obtain:

$$\check{Q}_n^{[3]}(\beta_n) := X_{n-k_n,n}\exp\left(\hat{a}_n(\ln(n/k_n))\ln\left(\frac{\ln\beta_n}{\ln(k_n/n)}\right)\right),$$

   which is exactly the estimator dedicated to extreme quantiles from Weibull-tail distributions introduced in [13]. It should perform well for estimating extreme quantiles from distributions with associated $\theta = 0$.

3. Finally, the third estimator was introduced in [23]:

$$\check{Q}_n^{[4]}(\beta_n) := X_{n-\ell_n,n}\exp\left(\hat{a}_{\ell_n,n}^{[4]}L_{\hat{\theta}_{k_n,n}^{[4]}}\left(\frac{\ln(1/\beta_n)}{\nu_{\ell_n+1,n}}\right)\right),$$

   with

$$\begin{aligned}
\hat{a}_{\ell_n,n}^{[4]} &:= \frac{\hat{\gamma}_{\ell_n,n}}{\frac{1}{\ell_n}\sum_{j=1}^{\ell_n}L_{\hat{\theta}_{k_n,n}^{[4]}}\left(\frac{\nu_{j,n}}{\nu_{\ell_n+1,n}}\right)},\\
\hat{\theta}_{k_n,n}^{[4]} &:= 1+\frac{\sum_{i=1}^{k_n-1}(\ln\hat{\gamma}_{i,n}-\ln\hat{\gamma}_{k_n,n})}{\sum_{i=1}^{k_n-1}(\ln\nu_{i+1,n}-\ln\nu_{k_n+1,n})},\\
\hat{\gamma}_{i,n} &:= \frac{1}{i}\sum_{j=1}^{i}(\ln X_{n-j+1,n}-\ln X_{n-i,n}),
\end{aligned}$$

   where $\nu_{i,n} := \sum_{j=i}^{n} j^{-1}$ and $\ell_n = k_n/\nu_{k_n+1,n}^2$.

Distribution functions. The estimators are compared on the 8 distributions described in Table 1: Gamma($a = 1.5$, s), Weibull($k = 0.5$, λ_1), Gaussian(μ_1, $\sigma = 1$), Lognormal(μ_2, $\sigma = 1$), Burr(λ_2, $c = 0.5$, $k = 0.5$), Pareto-like, super heavy-tail and finite endpoint (x^*). Note that the Pareto-like distribution is taken from [23]. The position parameters μ_1, μ_2 as well as the scaling parameters s, λ_1, λ_2 and the finite endpoint x^* are chosen such that the simulated data points are all larger than 1.

Results. The log ratio errors $\check{\nu}_n^{[q]} := \ln\left(\check{Q}_n^{[q]}(\beta_n)/Q_n(\beta_n)\right)$ are computed for all 4 estimators ($q = 1, \ldots, 4$), for each of the 500 datasets from the 8 distributions. The bias of each estimator is then estimated (on a logarithmic scale) by averaging the $\check{\nu}_n^{[q]}$ over the $N = 500$ replications. Similarly, the mean-squared error (MSE) is evaluated (on a logarithmic scale) by averaging the squared $\check{\nu}_n^{[q]}$ over the $N = 500$ replications.

The resulting bias and MSE are displayed on Figures 2–5 as functions of k_n. In terms of bias, it appears that $\check{Q}_n^{[1]}(\beta_n)$ show pretty good results with a small bias over a large range of k_n values for Gamma, Weibull, Gaussian, Lognormal, super heavy-tail and finite endpoint distributions. The bias behavior of $\check{Q}_n^{[1]}(\beta_n)$ is less satisfying on Burr and Pareto-like distributions ($\theta = 1$ in both cases) where $\check{Q}_n^{[4]}(\beta_n)$ is the best in terms of bias stability. From the MSE point of view, $\check{Q}_n^{[1]}(\beta_n)$ achieves better performances than $\check{Q}_n^{[4]}(\beta_n)$ on almost all distributions except the Pareto-like where the results are similar and the Burr distribution where $\check{Q}_n^{[4]}(\beta_n)$ is better than $\check{Q}_n^{[1]}(\beta_n)$. Similar behaviors can be observed on Figures 6–9 where the estimators $\hat{\theta}_n^{(M)}$ and $\hat{\theta}_{k_n,n}^{[4]}$ are compared.

Let us also note that assuming $\theta = 0$ improves the results only on the strict Weibull distribution, the results of $\check{Q}_n^{[3]}(\beta_n)$ being disappointing for other Weibull tail-distributions such as Gaussian or Gamma. Similarly, assuming $\theta > 0$ improves the results only on the Gamma distribution, the results of $\check{Q}_n^{[2]}(\beta_n)$ are not convincing on other distributions. This phenomenon indicates that $\hat{\theta}_{n,-}^{(M)}$ is useful even in case where $\theta > 0$, since it may temper the positive bias associated with $\hat{\theta}_{n,+}^{(M)}$.

5 Illustration on real data

In this section, the extreme quantile estimators $\check{Q}_n^{[1]}(\beta_n)$ and $\check{Q}_n^{[4]}(\beta_n)$ are compared on the average daily river flows (in m^3/s) of the Rhône river (France). The dataset covers the period 1915–2013, and for stationarity reasons, only the winter and spring seasons were considered (from December, 1st to May, 31st), leading to $n = 18043$ measures. We focus on the extreme quantile $Q(\beta_n)$ with $\beta_n = 5.5 \times 10^{-6}$ which is exceeded with a frequency of 10^{-3} per year. Figure 11 displays the index estimates $\hat{\theta}_n^{(M)}$ and $\hat{\theta}_{k_n,n}^{[4]}$ as well as the estimates $\check{Q}_n^{[1]}(\beta_n)$ and $\check{Q}_n^{[4]}(\beta_n)$ of the extreme quantile together with their corresponding 95% asymptotic confidence intervals.

In both cases, the index estimates seem fairly stable as a function of k_n, suggesting a positive value for $\theta \in [0.3, 0.4]$ associated with a log-Weibull tail-distribution. This hypothesis is confirmed by the quantile-quantile plot displayed on Figure 10. This plot is inspired from approximation (10) which suggests that the points

$$\left(\ln\left(1 + \frac{\ln(i/k_n)}{\ln(k_n/n)}\right), \ln_2 X_{n-i+1,n} - \ln_2 X_{n-k_n+1,n}\right),\ i = 1, \ldots, k_n - 1$$

should be approximately located on a line of slope $\hat{\theta}_{n,+}^{(M)}$. Following hydrologists advice, $k_n = 252$

was selected, corresponding to a flow of $2400m^3/s$. The very good fit can be interpreted as an empirical validation of the log-Weibull tail-distribution assumption.

The behavior of extreme quantile estimates $\check{Q}_n^{[1]}(\beta_n)$ and $\check{Q}_n^{[4]}(\beta_n)$ are also similar, $\check{Q}_n^{[1]}(\beta_n)$ being more stable with respect to k_n than $\check{Q}_n^{[4]}(\beta_n)$. The first estimator $\check{Q}_n^{[1]}(\beta_n)$ points towards a constant value $Q(\beta_n) \approx 10,000m^3/s$ while the second one $\check{Q}_n^{[4]}(\beta_n)$ exhibits a trend from 8000 to $12{,}000m^3/s$ as k_n vary from 100 to 2000. At the opposite, the widths of the 95% asymptotic confidence intervals associated with both estimators are significantly different. Indeed, the interval associated with $\check{Q}_n^{[1]}(\beta_n)$ is 10 times narrower than the one associated with $\check{Q}_n^{[4]}(\beta_n)$. This result is in accordance with (18) since here $\tau \simeq 1.24$ yields $\Lambda_\theta(\tau) \simeq 0.1$ for a large range of θ values, see Figure 1.

References

[1] Abramowitz, M. and Stegun, I.A. (1965). *Handbook of mathematical functions with formulas, graphs and mathematical tables*, Dover Book on Advanced Mathematics, New York.

[2] Albert, C., Dutfoy, A. and Girard, S. (2018). Asymptotic behavior of the extrapolation error associated with the estimation of extreme quantiles, `https://hal.inria.fr/hal-01692544`.

[3] Alves, I., de Haan, L. and Neves, C. (2009). A test procedure for detecting super-heavy tails, *Journal of Statistical Planning and Inference*, **139**(2), 213–227.

[4] Beirlant, J., Broniatowski, M., Teugels, J. and Vynckier, P. (1995). The mean residual life function at great age: Applications to tail estimation, *Journal of Statistical Planning and Inference*, **45**(1-2), 21–48.

[5] Bingham, N.H., Goldie, C.M. and Teugels, J.L. (1987). *Regular Variation*, Cambridge University Press.

[6] Coles, S., Pericchi, L.R., and Sisson, S. (2003). A fully probabilistic approach to extreme rainfall modeling. *Journal of Hydrology*, **273**(1-4), 35–50.

[7] Dekkers, A., Einmhal, J. and de Haan, L. (1989). A moment estimator for the index of an extreme-value distribution, *The Annals of Statistics*, **17**(4), 1833–1855.

[8] El Methni, J., Gardes, L., Girard, S. and Guillou, A. (2012). Estimation of extreme quantiles from heavy and light tailed distributions, *Journal of Statistical Planning and Inference*, **142**(10), 2735–2747.

[9] Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*, Springer.

[10] Embrechts, P. (2000). *Extremes and integrated risk management*, Risk Books.

[11] Gardes, L. and Girard, S. (2008). Estimation of the Weibull tail-coefficient with linear combination of upper order statistics, *Journal of Statistical Planning and Inference*, **138**(5), 1416–1427.

[12] Gardes, L., Girard, S. and Guillou, A. (2011). Weibull tail-distributions revisited: a new look at some tail estimators, *Journal of Statistical Planning and Inference*, **141**(1), 429–444.

[13] Gardes, L. and Girard, S. (2006). Comparison of Weibull tail-coefficients estimators, *REVSTAT - Statistical Journal*, **4**, 163–188.

[14] Goegebeur, Y., Beirlant, J. and De Wet, T. (2010). Generalized kernel estimators for the Weibull-tail coefficient, *Communications in Statistics-Theory and Methods*, **39**(20), 3695–3716.

[15] de Haan, L., and Ferreira, A. (2006). *Extreme Value Theory: An introduction*, Springer Series in Operations Research and Financial Engineering, Springer.

[16] Jagger, T.H. and Elsner, J.B. (2006). Climatology models for extreme hurricane winds near the United States. *Journal of Climate*, **19**(13), 3220–3236.

[17] Katz, R W., Parlange, M.B. and Naveau, P. (2002). Statistics of extremes in hydrology. *Advances in water resources*, **25**(8-12), 1287–1304.

[18] Muir, L.R. and El-Shaarawi, A.H. (1986). On the calculation of extreme wave heights: a review. *Ocean Engineering*, **13**(1), 93–118.

[19] McNeil, A.J., Frey, R. and Embrechts, P. (2005). *Quantitative risk management: concepts, techniques, and tools*, Princeton university press.

[20] Smirnov, N.V. (1949). Limit distributions for the terms of a variational series, *Trudy Matematicheskogo Instituta im. V.A. Steklova*, **25**, 3–60.

[21] de Valk, C. (2016). Approximation of high quantiles from intermediate quantiles, *Extremes*, **19**(4), 661–686.

[22] de Valk, C. (2016). Approximation and estimation of very small probabilities of multivariate extreme events, *Extremes*, **19**(4), 686–717.

[23] de Valk, C., and Cai, J.-J. (2018). A high quantile estimator based on the log-generalized Weibull tail limit, *Econometrics and Statistics*, **6**, 107–128.

Acknowledgments

The authors would like to thank Cees de Valk for fruitful discussions and the referees for their valuable suggestions, which have significantly improved the paper.

6 Appendix: Proofs

Some preliminary lemmas are first provided in Paragraph 6.1, their proofs being postponed to Paragraph 6.3. Proofs of main results are given in Paragraph 6.2.

6.1 Preliminary lemmas

We first give a general tool for establishing the convergence in distribution of random vectors.

Lemma 2 *For $p \in \mathbb{N} \setminus \{0\}$ and $n \in \mathbb{N}$, let $W_n := (W_{n,1}, \ldots, W_{n,p})^\top$ and $W := (W_1, \ldots, W_p)^\top$ be two random vectors in $\mathbb{R}^p$. If there exist a sequence $\sigma_n \to 0$ and $\lambda := (\lambda_1, \ldots, \lambda_p)^\top \in \mathbb{R}^p$ such that $\sigma_n^{-1}(W_n - \lambda) \xrightarrow{d} W$ then, for all $q \in \mathbb{N} \setminus \{0\}$ and all continuously differentiable functions $\varphi_1, \ldots, \varphi_q$ from $\mathbb{R}^p$ to $\mathbb{R}$,*

$$\sigma_n^{-1}\left((\varphi_1(W_n), \ldots, \varphi_q(W_n))^\top - (\varphi_1(\lambda), \ldots, \varphi_q(\lambda))^\top\right) \xrightarrow{d} \left(W^\top \nabla\varphi_1(\lambda), \ldots, W^\top \nabla\varphi_q(\lambda)\right),$$

where, for all $i \in \{1, \ldots, q\}$, $\nabla\varphi_i(\lambda)$ is the gradient of φ_i evaluated at point λ.

The following lemma is the cornerstone for establishing the asymptotic normality of the quantile estimator in the intermediate case.

Lemma 3 *Let $Z_1, \ldots, Z_n$ be n independent copies of a random variable Z. Denote by S_Z the survival function of Z and by $Q_Z = S_Z^{\leftarrow}$ the associated quantile function. Assume Q_Z is differentiable and that $-Q_Z'(1/\cdot)$ is regularly varying. Then, for all sequence (α_n) such that $\alpha_n \to 0$ and $n\alpha_n \to \infty$,*

$$\frac{n^{1/2}}{\alpha_n^{1/2} Q_Z'(\alpha_n)} \left(Z_{n-\lfloor n\alpha_n \rfloor, n} - Q_Z(\alpha_n)\right) \xrightarrow{d} \mathcal{N}(0,1).$$

An elementary result on ordered statistics from standard uniform random variables is provided below.

Lemma 4 *Let $U_1, \ldots, U_n$ be independent standard uniform variables. For all intermediate sequence (k_n), i.e. such that $k_n \to \infty$ and $k_n/n \to 0$, one has*

(i) $U_{k_n+1,n} \xrightarrow{\mathbb{P}} 0$.

(ii) Let $\{F_1, \ldots, F_{k_n}\}$ and $\{E_1, \ldots, E_n\}$ be two independent samples of independent standard exponential random variables. Then,

$$\left\{\frac{\ln(U_{i+1,n}/U_{k_n+1,n})}{\ln(U_{k_n+1,n})},\ i = 0, \ldots, k_n - 1\right\} \overset{d}{=} \left\{\frac{F_{k_n-i,k_n}}{E_{n-k_n,n}},\ i = 0, \ldots, k_n - 1\right\},$$

where $\{F_1, \ldots, F_{k_n}\}$ are independent from $E_{n-k_n,n}$.

Let us introduce some additional notations. For $J \in \mathbb{N} \setminus \{0\}$, $\zeta := (\zeta_1, \ldots, \zeta_J)^\top \in (-\infty, 1)^J$ and $t > 0$, consider the functions

$$S_n(\zeta) := \frac{1}{k_n} \sum_{i=0}^{k_n-1} \prod_{j=1}^{J} L_{\zeta_j}\left(\frac{\ln U_{i+1,n}}{\ln U_{k_n+1,n}}\right) \text{ and } \mu(t,\zeta) := \int_0^1 \prod_{j=1}^{J} L_{\zeta_j}\left(1 - \frac{\ln s}{t}\right) ds$$

and remark that, for $J = 1$ and $\zeta < 1$, $\mu(t,\zeta) = \mu_1(t,\zeta)$ and, for $J = 2$, $\mu(t,(\zeta,\zeta)) = \mu_2(t,\zeta)$. Let us also recall that, from Section 2, $\Psi_t(\zeta) = \mu_1^2(t,\zeta)/\mu_2(t,\zeta)$ for all $t > 0$ and $\zeta < 1$. The next result is of analytical nature. It provides first-order asymptotic expansions as $t \to \infty$ for functions $\mu(t,\zeta)$ and $\Psi_t(\zeta)$ locally uniformly on ζ.

Lemma 5 *(i) Let $J \in \mathbb{N} \setminus \{0\}$. For all hyper-rectangle $\mathcal{R}_J \subset (-\infty, 1)^J$, one has*

$$\lim_{t\to\infty} \sup_{\zeta\in\mathcal{R}_J} \left| t^J \mu(t,\zeta) - J! \right| = 0.$$

(ii) Denoting by Ψ_t' the first derivative of Ψ_t, one has, for all closed interval $I \subset (-\infty, 1)$,

$$\lim_{t\to\infty} \sup_{\zeta\in I} \left| t\Psi_t'(\zeta) + \frac{1}{2} \right| = 0.$$

As a consequence of Lemma 5(ii), the function Ψ_t is decreasing at least for t large enough. Lemma 6 below states Law of Large Numbers type results dedicated to particular triangular arrays of random variables.

Lemma 6 *Let (t_m) be a sequence such that $\log(m)/t_m \to 0$ as $m \to \infty$ and let $F_1, \ldots, F_m$ be independent copies of a standard exponential random variable.*
(i) For all $\delta > 0$ and $\zeta \in (-\infty, 1)^J$ with $J \in \mathbb{N} \setminus \{0\}$, one has

$$\frac{t_m^{J-1}}{m} \sum_{i=1}^{m} F_i \left(1 + \frac{F_i}{t_m}\right)^{\zeta_1 - 1} \prod_{j=2}^{J} L_{\zeta_j}\left(1 + \frac{F_i}{\delta t_m}\right) \xrightarrow{\mathbb{P}} J!/\delta^{J-1}.$$

(ii) For all $\delta > 0$, $(\xi_1, \ldots, \xi_4) \in (-\infty, 1)^4$ with $\xi_3 > \xi_4$ and $J_i \in \mathbb{N}$, $i \in \{1, 2, 3\}$, one has for $J = J_1 + J_2 + 2J_3$ that

$$\frac{t_m^J}{m} \sum_{i=1}^{m} L_{\xi_1}^{J_1}\left(1 + \frac{F_i}{\delta t_m}\right) L_{\xi_2}^{J_2}\left(1 + \frac{F_i}{\delta t_m}\right) \left[L_{\xi_3}\left(1 + \frac{F_i}{\delta t_m}\right) - L_{\xi_4}\left(1 + \frac{F_i}{\delta t_m}\right)\right]^{J_3} \xrightarrow{\mathbb{P}} \frac{J!}{\delta^J} \left(\frac{\xi_3 - \xi_4}{2}\right)^{J_3}.$$

Finally, Lemmas 7 and 8 are the key tools for establishing the joint asymptotic normality of the random pair $(M_n^{(1)}, M_n^{(2)})$.

Lemma 7 *Let (k_n) be an intermediate sequence such that $k_n \to \infty$ and $\ln(k_n)/\ln(n) \to 0$ as $n \to \infty$. For $J_1, J_2 \in \mathbb{N} \setminus \{0\}$ and for all $\zeta^{(1)} \in (-\infty, 1)^{J_1}$, $\zeta^{(2)} \in (-\infty, 1)^{J_2}$, the random vector*

$$k_n^{1/2} \left\{ \frac{S_n(\zeta^{(1)})}{\mu(\ln(n/k_n), \zeta^{(1)})} - 1, \frac{S_n(\zeta^{(2)})}{\mu(\ln(n/k_n), \zeta^{(2)})} - 1 \right\}$$

converges in distribution to a centered Gaussian random vector with covariance matrix

$$\begin{pmatrix} (2J_1)!/(J_1!)^2 - 1 & (J_1 + J_2)!/(J_1!J_2!) - 1 \\ (J_1 + J_2)!/(J_1!J_2!) - 1 & (2J_2)!/(J_2!)^2 - 1 \end{pmatrix}.$$

Lemma 8 *Let (k_n) be an intermediate sequence such that $k_n \to \infty$ and $\ln(k_n)/\ln(n) \to 0$ as $n \to \infty$. For all $(\xi_1, \ldots, \xi_4) \in (-\infty, 1)^4$ with $\xi_3 > \xi_4$ and $J_i \in \mathbb{N}$, $i \in \{1, 2, 3\}$, one has for $J = J_1 + J_2 + 2J_3$ that*

$$\frac{[\ln(n/k_n)]^J}{k_n} \sum_{i=0}^{k_n - 1} L_{\xi_1}^{J_1}\left(\frac{\ln U_{i+1,n}}{\ln U_{k_n+1,n}}\right) L_{\xi_2}^{J_2}\left(\frac{\ln U_{i+1,n}}{\ln U_{k_n+1,n}}\right) \left[L_{\xi_3}\left(\frac{\ln U_{i+1,n}}{\ln U_{k_n+1,n}}\right) - L_{\xi_4}\left(\frac{\ln U_{i+1,n}}{\ln U_{k_n+1,n}}\right)\right]^{J_3}$$

converges in probability to $J!\left(\frac{\xi_3 - \xi_4}{2}\right)^{J_3}$.

6.2 Proofs of main results

Proof of Theorem 1 – Let $\{Z_i := \ln(X_i),\ i = 1, \ldots, n\}$. These random variables are independent with common quantile function $Q_Z(u) = V(\ln(1/u))$, $u \in (0,1)$. Under **(A1)**, Q_Z is differentiable with first derivative verifying $-Q_Z'(1/x) = xV'(\ln(x)) \in \mathcal{RV}(1)$. One can thus apply Lemma 3 to obtain

$$\frac{(n\alpha_n)^{1/2}}{V'(\ln(\alpha_n^{-1}))}\left(Z_{n-\lfloor n\alpha_n\rfloor,n} - Q_Z(\alpha_n)\right) \xrightarrow{d} \mathcal{N}(0,1).$$

Now, since $a(x) \sim xV'(x)$ as $x \to \infty$ in view of [15, Corollary 1.1.10], it follows that

$$\frac{(n\alpha_n)^{1/2}\ln(\alpha_n^{-1})}{a(\ln(\alpha_n^{-1}))}\left(Z_{n-\lfloor n\alpha_n\rfloor,n} - Q_Z(\alpha_n)\right) \xrightarrow{d} \mathcal{N}(0,1).$$

The result is then proved by remarking that $Z_{n-\lfloor n\alpha_n\rfloor,n} = \ln(X_{n-\lfloor n\alpha_n\rfloor,n})$ and $Q_Z = \ln Q$. ■

Proof of Proposition 1 – Let us first show that

$$\frac{\hat{a}_n(\ln(n/k_n))}{a(\ln(n/k_n))} \xrightarrow{\mathbb{P}} 1.$$

In view of

$$\sigma_n^{-1}\frac{L_\theta(d_n)}{H_{\theta,0}(d_n)}\left(\frac{\hat{a}_n(\ln(n/k_n))}{a(\ln(n/k_n))} - 1\right) \xrightarrow{d} \Lambda,$$

it is sufficient to prove that $\sigma_n^{-1}L_\theta(d_n)/H_{\theta,0}(d_n) \to \infty$ as $n \to \infty$.
Let us first assume that $d_n \to 1$ as $n \to \infty$. Since $L_\theta(1+u) \sim u$ and $H_{\theta,0}(1+u) \sim u^2/2$ as $u \to 0$, $L_\theta(d_n)/H_{\theta,0}(d_n) \sim 2/(d_n - 1)$ and $\sigma_n^{-1}L_\theta(d_n)/H_{\theta,0}(d_n) \to \infty$ as $n \to \infty$. Second, if $d_n \to d \in (1,\infty)$ then $L_\theta(d_n)/H_{\theta,0}(d_n) \to L_\theta(d)/H_{\theta,0}(d) > 0$ and the result is proved. Finally, if $d_n \to \infty$, remarking that, as $t \to \infty$,

$$\frac{L_\theta(t)}{H_{\theta,0}(t)} \sim \begin{cases} 1/\ln(t) & \text{if } \theta > 0, \\ 2/\ln(t) & \text{if } \theta = 0, \\ -\theta & \text{if } \theta < 0 \end{cases} \tag{20}$$

implies $\sigma_n^{-1}L_\theta(d_n)/H_{\theta,0}(d_n) \to \infty$ by assumption.

Let us now prove the second part of Proposition 1. The following equality holds:

$$H_{\hat{\theta}_n,0}(d_n) - H_{\theta,0}(d_n) = (\hat{\theta}_n - \theta)\int_1^{d_n} s^{\theta-1}\ln^2(s)\frac{\exp((\hat{\theta}_n - \theta)\ln(s)) - 1}{(\hat{\theta}_n - \theta)\ln(s)}ds. \tag{21}$$

Since for all $s \in (1, d_n)$, $|(\hat{\theta}_n - \theta)\ln(s)| \le |\hat{\theta}_n - \theta|\ln(d_n) = O_{\mathbb{P}}(\sigma_n \ln d_n) = o_{\mathbb{P}}(1)$ by assumption, it is easy to check that

$$H_{\hat{\theta}_n,0}(d_n) - H_{\theta,0}(d_n) = (\hat{\theta}_n - \theta)\int_1^{d_n} s^{\theta-1}\ln^2(s)ds(1 + o_{\mathbb{P}}(1)),$$

or equivalently,

$$\frac{H_{\hat{\theta}_n,0}(d_n)}{H_{\theta,0}(d_n)} - 1 = \sigma_n^{-1}(\hat{\theta}_n - \theta) \times \frac{\sigma_n}{H_{\theta,0}(d_n)}\int_1^{d_n} s^{\theta-1}\ln^2(s)ds(1 + o_{\mathbb{P}}(1)).$$

The three situations $d_n \to 1$, $d_n \to d > 1$ and $d_n \to \infty$ are again considered separately. First, since

$$\int_1^{1+u} s^{\theta-1} \ln^2(s) ds \sim \frac{u^3}{3} \text{ and } H_{\theta,0}(u) \sim \frac{u^2}{2},$$

as $u \to 0$, one has for $d_n \to 1$ that

$$\frac{H_{\hat{\theta}_n,0}(d_n)}{H_{\theta,0}(d_n)} - 1 \sim \sigma_n^{-1}(\hat{\theta}_n - \theta) \times \frac{2}{3}\sigma_n(d_n - 1) \xrightarrow{\mathbb{P}} 0.$$

The case $d_n \to d$ is straightforward. Finally, when $d_n \to \infty$,

$$\frac{1}{H_{\theta,0}(d_n)} \int_1^{d_n} s^{\theta-1} \ln^2(s) ds \sim \begin{cases} \ln(d_n) & \text{if } \theta > 0, \\ 3\ln(d_n)/2 & \text{if } \theta = 0, \\ -2/\theta & \text{if } \theta < 0. \end{cases}$$

Collecting conditions $\sigma_n \ln(d_n) \to 0$ and $\sigma_n^{-1}(\hat{\theta}_n - \theta) \xrightarrow{d} \Theta$ concludes the proof. ■

Proof of Theorem 2 – Let us start with the expansion:

$$\frac{\sigma_n^{-1}}{a(\ln(n/k_n))H_{\theta,0}(d_n)} \ln \frac{\check{Q}_n(\beta_n)}{Q(\beta_n)} = T_{1,n} + T_{2,n} + T_{3,n} + T_{4,n},$$

with

$$\begin{aligned}
T_{1,n} &= \frac{\sigma_n^{-1}}{a(\ln(n/k_n))H_{\theta,0}(d_n)} \left(\ln X_{n-k_n,n} - \ln Q(k_n/n)\right), \\
T_{2,n} &= \frac{\hat{a}_n(\ln(n/k_n))}{a(\ln(n/k_n))H_{\theta,0}(d_n)} \sigma_n^{-1} \left(L_{\hat{\theta}_n}(d_n) - L_\theta(d_n)\right), \\
T_{3,n} &= \frac{L_\theta(d_n)}{H_{\theta,0}(d_n)} \sigma_n^{-1} \left(\frac{\hat{a}_n(\ln(n/k_n))}{a(\ln(n/k_n))} - 1\right), \\
T_{4,n} &= \frac{\sigma_n^{-1}}{H_{\theta,0}(d_n)} \left(\frac{\ln Q(k_n/n) - \ln Q(\beta_n)}{a(\ln(n/k_n))} + L_\theta(d_n)\right).
\end{aligned}$$

Clearly, under **(A3)**, $T_{1,n} \xrightarrow{d} \Omega$ and $T_{3,n} \xrightarrow{d} \Lambda$. Next, remark that Proposition 1 entails that the asymptotic distribution of $T_{2,n}$ is the same as the one of

$$\frac{\sigma_n^{-1}}{H_{\theta,0}(d_n)} \left(L_{\hat{\theta}_n}(d_n) - L_\theta(d_n)\right).$$

Furthermore, similarly to (21) in the proof of Proposition 1, one can show that $L_{\hat{\theta}_n}(d_n) - L_\theta(d_n) = (\hat{\theta}_n - \theta)H_{\theta,0}(d_n)(1 + o_{\mathbb{P}}(1))$. As a consequence, $T_{2,n} \xrightarrow{d} \Theta$. Finally, since $\ln Q(\alpha) = V(\ln(1/\alpha))$, it follows that

$$H_{\theta,0}(d_n)\sigma_n T_{4,n} = \frac{V(\ln(n/k_n)) - V(\ln(\beta_n))}{a(\ln(n/k_n))} + L_\theta(d_n).$$

Let us consider separately the three cases $d_n \to 1$, $d_n \to d > 1$ and $d_n \to \infty$.
First, if $d_n \to 1$, the second order condition **(A2)** entails $H_{\theta,0}(d_n)\sigma_n T_{4,n} \sim H_{\theta,\rho}(d_n)\tilde{A}(\ln(n/k_n))$. Since for all $\rho \leq 0$, $H_{\theta,\rho}(1+u) \sim H_{\theta,0}(1+u) \sim u^2/2$ as $u \to 0$, it follows that $T_{4,n} \sim \sigma_n^{-1}\tilde{A}(\ln(n/k_n)) = o(1)$ by assumption.

Next, if $d_n \to d > 1$, conditions **(A2)** and $\sigma_n^{-1}\tilde{A}(\ln(n/k_n)) \to 0$ imply that $T_{4,n} \to 0$ as $n \to \infty$. Finally, when $d_n \to \infty$, [15, Lemma 4.3.5] entails that

$$T_{4,n} = \mathcal{O}\left\{\frac{L_\theta(d_n)}{H_{\theta,0}(d_n)}\sigma_n^{-1}\tilde{A}(\ln(n/k_n))\right\} = o(1),$$

using (20). To conclude, if $d_n \to d \in [1,\infty]$, $T_{4,n} \to 0$ as $n \to \infty$ and the result is proved. ■

Proof of Theorem 3 – For $i = 1,\ldots,n$, let $U_i := S(X_i)$ so that $\{U_1,\ldots,U_n\}$ is a set of independent standard uniform random variables. Let $\delta > 0$ and $r(\cdot) := a(\cdot)/V(\cdot)$. For $s \in (\alpha^\delta, 1)$, let us plug $x := \ln(1/\alpha)$ and $t := 1 + \ln s/\ln\alpha$ in (12). Consequently, as $\alpha \to 0$,

$$r^{-1}(\ln(1/\alpha))\ln\frac{\ln Q(s\alpha)}{\ln Q(\alpha)} = L_{\theta_-}\left(1+\frac{\ln s}{\ln\alpha}\right) + A(\ln(1/\alpha))H_{\theta_-,\rho'}\left(1+\frac{\ln s}{\ln\alpha}\right) + o[A(\ln 1/\alpha)], \quad (22)$$

uniformly in $s \in (\alpha^\delta, 1)$. As a consequence of Lemma 4(i, ii), one may apply (22) with α replaced by $U_{k_n+1,n}$ and s replaced by $U_{i+1,n}/U_{k_n+1,n}$ to get, for n large enough,

$$\begin{aligned} r^{-1}(\ln U_{k_n+1,n}^{-1})\frac{M_n^{(1)}}{\mu_1(\ln(n/k_n),\theta_-)} - 1 &= \frac{S_n(\theta_-)}{\mu_1(\ln(n/k_n),\theta_-)} - 1 \\ &+ \frac{A(\ln U_{k_n+1,n}^{-1})}{k_n\mu_1(\ln(n/k_n),\theta_-)}\sum_{i=0}^{k_n-1} H_{\theta_-,\rho'}\left(\frac{\ln U_{i+1,n}}{\ln U_{k_n+1,n}}\right) \\ &+ o_{\mathbb{P}}\left(\frac{A(\ln U_{k_n+1,n}^{-1})}{\mu_1(\ln(n/k_n),\theta_-)}\right). \end{aligned}$$

Similarly, we have

$$\begin{aligned} & r^{-2}(\ln U_{k_n+1,n}^{-1})\frac{M_n^{(2)}}{\mu_2(\ln(n/k_n),\theta_-)} - 1 = \frac{S_n((\theta_-,\theta_-))}{\mu_2(\ln(n/k_n),\theta_-)} - 1 \\ + & \frac{2A(\ln U_{k_n+1,n}^{-1})}{k_n\mu_2(\ln(n/k_n),\theta_-)}\sum_{i=0}^{k_n-1} L_{\theta_-}\left(\frac{\ln U_{i+1,n}}{\ln U_{k_n+1,n}}\right)H_{\theta_-,\rho'}\left(\frac{\ln U_{i+1,n}}{\ln U_{k_n+1,n}}\right) \\ + & \frac{A^2(\ln U_{k_n+1,n}^{-1})}{k_n\mu_2(\ln(n/k_n),\theta_-)}\sum_{i=0}^{k_n-1} H_{\theta_-,\rho'}^2\left(\frac{\ln U_{i+1,n}}{\ln U_{k_n+1,n}}\right) + o_{\mathbb{P}}\left(\frac{A^2(\ln U_{k_n+1,n}^{-1})}{\mu_2(\ln(n/k_n),\theta_-)}\right) \\ + & \frac{S_n(\theta_-)}{\mu_2(\ln(n/k_n),\theta_-)}o_{\mathbb{P}}[A(\ln U_{k_n+1,n}^{-1})] + \frac{o_{\mathbb{P}}[A^2(\ln U_{k_n+1,n}^{-1})]}{k_n\mu_2(\ln(n/k_n),\theta_-)}\sum_{i=0}^{k_n-1} H_{\theta_-,\rho'}\left(\frac{\ln U_{i+1,n}}{\ln U_{k_n+1,n}}\right). \end{aligned}$$

From Rényi's representation, $\ln(1/U_{k_n+1,n})/\ln(n/k_n) \xrightarrow{\mathbb{P}} 1$ and since $|A|$ is regularly varying, it follows that

$$\left|\frac{A(\ln(1/U_{k_n+1,n}))}{A(\ln(n/k_n))}\right| \xrightarrow{\mathbb{P}} 1 \quad (23)$$

as $n \to \infty$. Now, since for all $t > 0$,

$$H_{\theta_-,\rho'}(t) = \begin{cases} 1/\rho'(L_{\theta_-+\rho'}(t) - L_{\theta_-}(t)) & \text{if } \rho' \neq 0, \\ L_{\theta_-}(t)L_0(t) - \theta_-^{-1}(L_{\theta_-}(t) - L_0(t)) & \text{if } \rho' = 0 \text{ and } \theta_- \neq 0, \\ L_0^2(t)/2 & \text{if } \rho' = \theta_- = 0, \end{cases}$$

Lemma 5(i), Lemma 7, Lemma 8 and condition $k_n A^2(\ln(n/k_n))/\ln^2(n/k_n) \to 0$ yield

$$r^{-1}(\ln U^{-1}_{k_n+1,n})\frac{M_n^{(1)}}{\mu_1(\ln(n/k_n),\theta_-)} - 1 = \frac{S_n(\theta_-)}{\mu_1(\ln(n/k_n),\theta_-)} - 1 + o_{\mathbb{P}}\left(k_n^{-1/2}\right),$$

and

$$r^{-2}(\ln U^{-1}_{k_n+1,n})\frac{M_n^{(2)}}{\mu_2(\ln(n/k_n),\theta_-)} - 1 = \frac{S_n((\theta_-,\theta_-))}{\mu_2(\ln(n/k_n),\theta_-)} - 1 + o_{\mathbb{P}}\left(k_n^{-1/2}\right).$$

Note that Lemma 7 can be applied since $k_n A^2(\ln(n/k_n))/\ln^2(n/k_n) \to 0$ implies $\ln(k_n)/\ln(n) \to 0$, see [13, Lemma 1]. As a consequence, the random vector

$$k_n^{1/2}\left(r^{-1}(\ln U^{-1}_{k_n+1,n})\frac{M_n^{(1)}}{\mu(\ln(n/k_n),\theta_-)} - 1, r^{-2}(\ln U^{-1}_{k_n+1,n})\frac{M_n^{(2)}}{\mu_2(\ln(n/k_n),\theta_-)} - 1\right) \tag{24}$$

converges in distribution to a centered Gaussian random vector (P_1, P_2) with covariance matrix

$$\begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix}.$$

Let us investigate the asymptotic distribution of $k_n^{1/2}(\hat{\theta}^{(M)}_{n,+} - \theta_+)$ where we recall that

$$\hat{\theta}^{(M)}_{n,+} = \frac{M_n^{(1)}}{\mu_1(\ln(n/k_n),0)},$$

see (5). From [15, Eq. 3.5.13], $r(\cdot)$ is regularly varying and thus $\ln(1/U_{k_n+1,n})/\ln(n/k_n) \xrightarrow{\mathbb{P}} 1$ implies $r(\ln(1/U_{k_n+1,n}))/r(\ln(n/k_n)) \xrightarrow{\mathbb{P}} 1$. Since $\mu_1(\ln(n/k_n),\theta_-) = \mu_1(\ln(n/k_n),0)$ for $\theta > 0$ and $\mu_1(\ln(n/k_n),\theta_-) \sim \mu_1(\ln(n/k_n),0)$ for $\theta \leq 0$ from Lemma 5(i), convergence in distribution (24) yields in both cases

$$k_n^{1/2}\left(\hat{\theta}^{(M)}_{n,+} - \theta_+\right) = r(\ln(n/k_n))P_{1,n} + k_n^{1/2}\left\{r(\ln(1/U_{k_n+1,n})) - \theta_+\right\}(1+o(1)),$$

where $P_{1,n} \xrightarrow{d} P_1$. Now, [15, Eq. B.3.46] ensures that $(r(x)-\theta_+)/A(x) \to \lambda \in \mathbb{R}$ as $x \to \infty$ and thus, taking into account of (23),

$$k_n^{1/2}\left(\hat{\theta}^{(M)}_{n,+} - \theta_+\right) = r(\ln(n/k_n))P_{1,n} + \mathcal{O}_{\mathbb{P}}(k_n^{1/2}A(\ln(n/k_n)) = \theta_+ P_{1,n} + \mathcal{O}_{\mathbb{P}}(k_n^{1/2}A(\ln(n/k_n))), \tag{25}$$

since $r(\ln(n/k_n)) \to \theta_+$ as $n \to \infty$. Now, using convergence in distribution (24) and a Taylor expansion yield

$$\frac{1}{\Psi_{\ln(n/k_n)}(\theta_-)}\frac{(M_n^{(1)})^2}{M_n^{(2)}} = 1 + k_n^{-1/2}(2P_{1,n} - P_{2,n}) + o_{\mathbb{P}}(k_n^{-1/2}),$$

where $(P_{1,n}, P_{2,n}) \xrightarrow{d} (P_1, P_2)$. From Lemma 5(i), $\Psi_{\ln(n/k_n)}(\theta_-) \to 1/2$ as $n \to \infty$, and thus

$$2k_n^{1/2}\left(\frac{(M_n^{(1)})^2}{M_n^{(2)}} - \Psi_{\ln(n/k_n)}(\theta_-)\right) \xrightarrow{d} 2P_1 - P_2, \tag{26}$$

where it is easily seen that $2P_1 - P_2 \sim \mathcal{N}(0,1)$. Now let $\sigma_n := k_n^{-1/2}\ln(n/k_n) \to 0$. For all $z \in \mathbb{R}$ and n large enough,

$$\mathbb{P}\left(\sigma_n^{-1}\left(\hat{\theta}^{(M)}_{n,-} - \theta_-\right) \leq z\right) = \mathbb{P}\left(\frac{(M_n^{(1)})^2}{M_n^{(2)}} \geq \Psi_{\ln(n/k_n)}\left(\theta_- + \sigma_n z\right)\right),$$

since for n large enough, $\Psi_{\ln(n/k_n)}$ is decreasing. Hence,

$$\mathbb{P}\left(\sigma_n^{-1}\left(\hat{\theta}_{n,-}^{(M)}-\theta_-\right)\leq z\right)=\mathbb{P}\left(2k_n^{1/2}\left(\frac{(M_n^{(1)})^2}{M_n^{(2)}}-\Psi_{\ln(n/k_n)}(\theta_-)\right)\geq z_{n,k_n}\right),$$

with $z_{n,k_n}:=2k_n^{1/2}(\Psi_{\ln(n/k_n)}(\theta_-+\sigma_n z)-\Psi_{\ln(n/k_n)}(\theta_-))$. The mean-value theorem entails that

$$z_{n,k_n}=2k_n^{1/2}\sigma_n z\Psi'_{\ln(n/k_n)}(\theta_-+\tau_n\sigma_n z),$$

where $\tau_n\in(0,1)$. We thus have that $z_{n,k_n}\to -z$ as $n\to\infty$ from Lemma 5(ii) and replacing σ_n by its expression. Taking into account of convergence (26) leads to

$$\frac{k_n^{1/2}}{\ln(n/k_n)}\left(\hat{\theta}_{n,-}^{(M)}-\theta_-\right)\xrightarrow{d}2P_1-P_2\sim\mathcal{N}(0,1). \tag{27}$$

Collecting (25) and (27) concludes the proof. ■

Proof of Theorem 4 – Keeping in mind the notations introduced in the proof of Theorem 3, the following expansion holds

$$\frac{\hat{a}_n^{(M)}(\ln(n/k_n))}{a(\ln(n/k_n))}=\mathcal{F}_{1,n}\times\mathcal{F}_{2,n}\times\mathcal{F}_{3,n},$$

with

$$\mathcal{F}_{1,n}:=\frac{r^{-1}(\ln(1/U_{k_n+1,n}))M_n^{(1)}}{\mu_1(\ln(n/k_n),\theta_-)},\ \mathcal{F}_{2,n}:=\frac{a(\ln(1/U_{k_n+1,n}))}{a(\ln(n/k_n))}\text{ and }\mathcal{F}_{3,n}:=\frac{\mu_1(\ln(n/k_n),\theta_-)}{\mu_1(\ln(n/k_n),\hat{\theta}_{n,-}^{(M)})}.$$

First, (24) entails that

$$k_n^{1/2}(\mathcal{F}_{1,n}-1)\xrightarrow{d}P_1. \tag{28}$$

Let us now consider $\mathcal{F}_{2,n}$. From [15, Theorem 2.3.6 and Corollary 2.3.5], there exist a function a_0 with, as $t\to\infty$

$$\frac{a_0(t)}{a(t)}=1+\mathcal{O}(\tilde{A}(t))$$

and a function A_0 with $A_0(t)=\mathcal{O}(\tilde{A}(t))$ as $t\to\infty$ such that, for all $\varepsilon>0$, $\delta>0$ and n large enough,

$$A_0^{-1}(\ln(n/k_n))\left(\frac{a_0(\ln(1/U_{k_n+1,n}))}{a_0(\ln(n/k_n))}-\left(\frac{\ln(1/U_{k_n+1,n})}{\ln(n/k_n)}\right)^{\theta}\right)=\left(\frac{\ln(1/U_{k_n+1,n})}{\ln(n/k_n)}\right)^{\theta}L_\rho\left(\frac{\ln(1/U_{k_n+1,n})}{\ln(n/k_n)}\right)+R_n,$$

where

$$|R_n|\leq\varepsilon\max\left\{\left(\frac{\ln(1/U_{k_n+1,n})}{\ln(n/k_n)}\right)^{\theta+\rho+\delta},\left(\frac{\ln(1/U_{k_n+1,n})}{\ln(n/k_n)}\right)^{\theta+\rho-\delta}\right\}.$$

Hence, since $|\tilde{A}|$ is a regularly varying function and $nU_{k_n+1,n}/k_n\xrightarrow{\mathbb{P}}1$,

$$\begin{aligned}\mathcal{F}_{2,n}&=\left\{\left(\frac{\ln(1/U_{k_n+1,n})}{\ln(n/k_n)}\right)^{\theta}+\mathcal{O}\{\tilde{A}(\ln(n/k_n))\}\left(\frac{\ln(1/U_{k_n+1,n})}{\ln(n/k_n)}\right)^{\theta}L_\rho\left(\frac{\ln(1/U_{k_n+1,n})}{\ln(n/k_n)}\right)\right.\\&+\left.\mathcal{O}\{\tilde{A}(\ln(n/k_n))R_n\}\right\}\{1+\mathcal{O}\{\tilde{A}(\ln(n/k_n))\}\}.\end{aligned} \tag{29}$$

Let us now consider the expansion

$$k_n^{1/2}(\mathcal{F}_{2,n}-1) = k_n^{1/2}\left(\mathcal{F}_{2,n} - \left(\frac{\ln(1/U_{k_n+1,n})}{\ln(n/k_n)}\right)^\theta\right) + k_n^{1/2}\left(\left(\frac{\ln(1/U_{k_n+1,n})}{\ln(n/k_n)}\right)^\theta - 1\right) =: T_{1,n}+T_{2,n}.$$

Since $\ln(U_{k_n+1,n})/\ln(k_n/n) \xrightarrow{\mathbb{P}} 1$, (29) entails that

$$T_{1,n} = \mathcal{O}\{k_n^{1/2}\tilde{A}(\ln(n/k_n))\} = o_{\mathbb{P}}(1),$$

by assumption. Next, Lemma 3 yields $\xi_n := k_n^{1/2}(\ln(1/U_{k_n+1,n}) - \ln(n/k_n)) \xrightarrow{d} \mathcal{N}(0,1)$ and thus

$$T_{2,n} = k_n^{1/2}\left(\left(1 + \frac{\xi_n}{k_n^{1/2}\ln(n/k_n)}\right)^\theta - 1\right) = o_{\mathbb{P}}(1).$$

To sum up, we have shown that

$$k_n^{1/2}(\mathcal{F}_{2,n} - 1) \xrightarrow{\mathbb{P}} 0. \tag{30}$$

Let us finally focus on $\mathcal{F}_{3,n}$. The mean-value theorem entails that

$$\mu_1\left(\ln(n/k_n), \hat{\theta}_{n,-}^{(M)}\right) - \mu_1\left(\ln(n/k_n), \theta_-\right) = (\hat{\theta}_{n,-}^{(M)} - \theta_-)\dot{\mu}\left(\ln(n/k_n), \theta_{n,-}^*\right),$$

where $\theta_{n,-}^* = \theta_- + \tau(\hat{\theta}_{n,-}^{(M)} - \theta_-)$ for some random value $\tau \in (0,1)$ and

$$\dot{\mu}(t,x) = \frac{\partial}{\partial x}\mu(t,x).$$

It has been shown in the proof of Lemma 5(ii) that $\mathcal{I}_1(t,x) = t^2\dot{\mu}(t,x) \to 1$ as $t \to \infty$, uniformly on all closed intervals included in $(-\infty, 1)$. Hence, under the assumptions of Theorem 4, $\theta_{n,-}^* \xrightarrow{\mathbb{P}} \theta_-$ from Theorem 3 and

$$\begin{aligned} & k_n^{1/2}\ln(n/k_n)\left\{\mu_1\left(\ln(n/k_n), \hat{\theta}_{n,-}^{(M)}\right) - \mu_1\left(\ln(n/k_n), \theta_-\right)\right\} \\ = \quad & (\ln(n/k_n))^2\dot{\mu}\left(\ln(n/k_n), \theta_{n,-}^*\right)\frac{k_n^{1/2}}{\ln(n/k_n)}\left(\hat{\theta}_{n,-}^{(M)} - \theta_-\right) \xrightarrow{d} 2P_1 - P2 \end{aligned}$$

from (27). Lemma 5(i) yields

$$k_n^{1/2}(\mathcal{F}_{3,n} - 1) \xrightarrow{d} P_2 - 2P_1. \tag{31}$$

Collecting (28), (30) and (31), Lemma 2 leads to

$$k_n^{1/2}\left(\frac{\hat{a}_n^{(M)}(\ln(n/k_n))}{a(\ln(n/k_n))} - 1\right) \xrightarrow{d} P_2 - P_1 \sim \mathcal{N}(0,2),$$

which is the expected result. ■

Proof of Corollary 1 – It is sufficient to show that condition **(A3)** is satisfied by the estimators $\hat{\theta}_n^{(M)}$ and $\hat{a}_n^{(M)}(\ln(n/k_n))$ with $\sigma_n := k_n^{-1/2}\ln(n/k_n)$ and $(\Omega, \Theta, \Lambda) = (0, \Theta, 0)$ where Θ follows a standard Gaussian distribution. First, from Theorem 1,

$$\frac{k_n^{1/2}}{\ln(n/k_n)}\frac{\ln(X_{n-k_n,n}) - \ln Q(k_n/n)}{a(\ln(n/k_n))H_{\theta,0}(d_n)} = \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\ln^2(n/k_n)H_{\theta,0}(d_n)}\right).$$

Clearly, when $d_n \to d \in (1,\infty]$, $\ln^2(n/k_n)H_{\theta,0}(d_n) \to \infty$. When $d_n \to 1$,

$$\ln^2\left(\frac{n}{k_n}\right) H_{\theta,0}(d_n) \sim \frac{1}{2}\ln^2\left(\frac{n}{k_n}\right)(d_n-1)^2 = \ln^2\left(\frac{k_n}{n\beta_n}\right) \to \infty, \tag{32}$$

since $n\beta_n \to c \geq 0$. As a consequence, if $d_n \to d \in [1,\infty]$,

$$\sigma_n^{-1}\frac{\ln(X_{n-k_n,n}) - \ln Q(k_n/n)}{a(\ln(n/k_n))H_{\theta,0}(d_n)} \xrightarrow{\mathbb{P}} 0. \tag{33}$$

Next, Theorem 3 entails that

$$\sigma_n^{-1}\left(\hat{\theta}_n^{(M)} - \theta\right) \xrightarrow{d} \mathcal{N}(0,1). \tag{34}$$

Finally, from Theorem 4,

$$\frac{k_n^{1/2}}{\ln(n/k_n)}\frac{L_\theta(d_n)}{H_{\theta,0}(d_n)}\left(\frac{\hat{a}_n^{(M)}(\ln(n/k_n))}{a(\ln(n/k_n))} - 1\right) = \mathcal{O}_{\mathbb{P}}\left(\frac{L_\theta(d_n)}{\ln(n/k_n)H_{\theta,0}(d_n)}\right).$$

If $d_n \to 1$, since $H_{\theta,0}(d_n)/L_\theta(d_n) \sim (d_n-1)/2$,

$$\frac{L_\theta(d_n)}{\ln(n/k_n)H_{\theta,0}(d_n)} \to 0, \tag{35}$$

as shown in (32). When $d_n \to d > 1$, it is clear that (35) holds. Finally, when $d_n \to \infty$, (20) entails that $L_\theta(d_n)/H_{\theta,0}(d_n) \to -\theta_-$ and thus (35) also holds. To sum up, when $d_n \to d \in [1,\infty]$,

$$\sigma_n^{-1}\frac{L_\theta(d_n)}{H_{\theta,0}(d_n)}\left(\frac{\hat{a}_n^{(M)}(\ln(n/k_n))}{a(\ln(n/k_n))} - 1\right) \xrightarrow{\mathbb{P}} 0, \tag{36}$$

and the conclusion follows from (33), (34) and (36). ■

6.3 Proofs of auxiliary results

Proof of Lemma 2 – Let $\varphi : \mathbb{R}^p \mapsto \mathbb{R}$ be a continuously differentiable function. It suffices to show that

$$\sigma_n^{-1}(\varphi(W_n) - \varphi(\lambda)) \xrightarrow{d} W^\top \nabla\varphi(\lambda).$$

Conclusion of the proof will be then straightforward by applying the Cramér-Wold device. The multivariate version of the mean-value theorem leads to

$$\sigma_n^{-1}(\varphi(W_n) - \varphi(\lambda)) = \sigma_n^{-1}(W_n - \lambda)^\top \nabla\varphi(\lambda_n^*),$$

where $\lambda_n^* := (\lambda_{n,1}^*, \ldots, \lambda_{n,p}^*)^\top$ with for all $i \in \{1,\ldots,p\}$, $\lambda_{n,i}^* = \lambda_i + \tau_i(W_{n,i} - \lambda_i)$ where $\tau_i \in (0,1)$. By assumption, $\lambda_{n,i}^* \xrightarrow{\mathbb{P}} \lambda_i$ and the continuous mapping theorem entails that $\nabla\varphi(\lambda_n^*) \xrightarrow{\mathbb{P}} \nabla\varphi(\lambda)$ and the proof is completed. ■

Proof of Lemma 3 – We start with a result due to Smirnov [20] and that can be found for instance in [15, Lemma 2.2.3]. Let (α_n) be a sequence such that $\alpha_n \to 0$ and $n\alpha_n \to \infty$. If $U_1, \ldots, U_n$ are independent random variables from a standard uniform distribution,

$$\frac{n^{1/2}}{\alpha_n^{1/2}}\left(U_{\lfloor n\alpha_n \rfloor + 1,n} - \alpha_n\right) \xrightarrow{d} \mathcal{N}(0,1). \tag{37}$$

Since $Z_{n-\lfloor n\alpha_n\rfloor,n} \stackrel{d}{=} Q_Z(U_{\lfloor n\alpha_n\rfloor+1,n})$, our aim is to show that

$$\frac{n^{1/2}}{\alpha_n^{1/2}Q'_Z(\alpha_n)}\left(Q_Z(U_{\lfloor n\alpha_n\rfloor+1,n}) - Q_Z(\alpha_n)\right) \xrightarrow{d} \mathcal{N}(0,1).$$

Since Q_Z is a differentiable function, the mean-value theorem leads to

$$Q_Z(U_{\lfloor n\alpha_n\rfloor+1,n}) - Q_Z(\alpha_n) = \left(U_{\lfloor n\alpha_n\rfloor+1,n} - \alpha_n\right) Q'_Z(\alpha_n^*),$$

where $\alpha_n^* := \alpha_n + \tau(U_{\lfloor n\alpha_n\rfloor+1,n} - \alpha_n)$ with $\tau \in (0,1)$. From (37) and since $n\alpha_n \to \infty$,

$$\frac{U_{\lfloor n\alpha_n\rfloor+1,n} - \alpha_n}{\alpha_n} \xrightarrow{\mathbb{P}} 0,$$

and thus $\alpha_n^* = \alpha_n(1 + o_{\mathbb{P}}(1))$. Since $-Q'_Z(1/\cdot)$ is regularly varying,

$$\frac{n^{1/2}}{\alpha_n^{1/2}Q'_Z(\alpha_n)}\left(Q_Z(U_{\lfloor n\alpha_n\rfloor+1,n}) - Q_Z(\alpha_n)\right) = \frac{n^{1/2}}{\alpha_n^{1/2}}\left(U_{\lfloor n\alpha_n\rfloor+1,n} - \alpha_n\right)(1 + o_{\mathbb{P}}(1)) \xrightarrow{d} \mathcal{N}(0,1),$$

and the proof is completed. ■

Proof of Lemma 4 – (i) The proof is based on Rényi's representation of standard uniform ordered statistics:

$$U_{k_n+1,n} \stackrel{d}{=} \frac{T_{k_n+1}}{T_{n+1}},$$

where for $j \in \mathbb{N}\setminus\{0\}$, T_j is the sum of j independent standard exponential random variables. The law of large numbers shows that $U_{k_n+1,n} \stackrel{\mathbb{P}}{\sim} k_n/n$ and the conclusion follows.
(ii) Remarking that

$$\left\{\frac{\ln(U_{i+1,n}/U_{k_n+1,n})}{\ln(U_{k_n+1,n})},\ i = 0,\dots,k_n-1\right\} \stackrel{d}{=} \left\{\frac{E_{n-i,n} - E_{n-k_n,n}}{E_{n-k_n,n}},\ i = 0,\dots,k_n-1\right\},$$

the result is then a consequence of the following Rényi's representation:

$$\{E_{j,n},\ j = 1,\dots,n\} \stackrel{d}{=} \left\{\sum_{r=1}^{j}\frac{F_r}{n-r+1},\ j = 1,\dots,n\right\}.$$

(iii) It is clear that

$$0 \le \max_{i\in\{0,\dots,k_n-1\}}\frac{F_{k_n-i,k_n}}{E_{n-k_n,n}} \le \frac{F_{k_n,k_n}}{E_{n-k_n,n}}.$$

Using the facts that $k_n^{1/2}(E_{n-k_n,n} - \ln(n/k_n)) \xrightarrow{d} \mathcal{N}(0,1)$ and that $F_{k_n,k_n} - \ln k_n$ converges in distribution to a Gumbel random variable entails

$$\frac{F_{k_n,k_n}}{E_{n-k_n,n}} \stackrel{\mathbb{P}}{\sim} \frac{\ln k_n}{\ln(n/k_n)} \to 0,$$

since $\log(k_n)/\log(n) \to 0$ as $n \to \infty$ and the conclusion follows. ■

Proof of Lemma 5 – (i) For $j = 1,\dots,J$, let

$$R_j(t,s) := tL_{\zeta_j}\left(1 + \frac{\ln(1/s)}{t}\right) - \ln(1/s).$$

Since $2|R_j(t,s)| \leq (1-\zeta_j)\ln^2(1/s)/t$, denoting by $\underline{\zeta} := \min\{\zeta_1, \ldots, \zeta_J\}$, one has

$$-\frac{1-\underline{\zeta}}{2}\frac{\ln^2(1/s)}{t} \leq R_j(t,s) \leq \frac{1-\underline{\zeta}}{2}\frac{\ln^2(1/s)}{t}.$$

Hence

$$\int_0^1 \left(\ln(1/s) - \frac{1-\underline{\zeta}}{2}\frac{\ln^2(1/s)}{t}\right)^J ds \leq t^J\mu(t,\zeta) \leq \int_0^1 \left(\ln(1/s) + \frac{1-\underline{\zeta}}{2}\frac{\ln^2(1/s)}{t}\right)^J ds.$$

As a consequence,

$$t^J\mu(t,\zeta) - J! \geq \sum_{j=0}^{J-1}(-1)^{J-j}(2J-j)!C_J^j\left(\frac{1-\underline{\zeta}}{2}\right)^{J-j}\frac{1}{t^{J-j}} \to 0,$$

uniformly for any hyper-rectangle included in $(-\infty,1)^J$. Similarly,

$$t^J\mu(t,\zeta) - J! \leq \sum_{j=0}^{J-1}(2J-j)!C_J^j\left(\frac{1-\underline{\zeta}}{2}\right)^{J-j}\frac{1}{t^{J-j}} \to 0,$$

uniformly locally and the proof is completed.
(ii) It is easily seen that

$$t(t^2\mu_2(t,x))^2\Psi_t'(x) \quad = \quad 2(t\mu_1(t,x))(t^2\mu_2(t,x))\mathcal{I}_1(t,x) - (t\mu_1(t,x))^2\mathcal{I}_2(t,x),$$

with

$$\begin{aligned}\mathcal{I}_1(t,x) &= t^2\dot{\mu_1}(t,x) = \int_0^1 t^2\dot{L}_x\left(1+\frac{\ln(1/s)}{t}\right)ds,\\ \mathcal{I}_2(t,x) &= t^2\dot{\mu_2}(t,x) = 2\int_0^1 t^2\dot{L}_x\left(1+\frac{\ln(1/s)}{t}\right)tL_x\left(1+\frac{\ln(1/s)}{t}\right)ds,\end{aligned}$$

and where the following notations have been introduced

$$\dot{L}_x(u) := \frac{\partial}{\partial x}L_x(u) \text{ and } \dot{\mu}_b(t,x) := \frac{\partial}{\partial x}\mu_b(t,x),\ b \in \{1,2\}.$$

The first step consists in studying the quantities $L_x(1+u)$ and $\dot{L}_x(1+u)$ for $u \geq 0$ and $x < 1$. A Taylor expansion leads to

$$L_x(1+u) = u + \frac{x-1}{2}u^2 + R_x(u), \tag{38}$$

where

$$0 \leq R_x(u) \leq \frac{(x-1)(x-2)}{6}u^3. \tag{39}$$

Next, an integration by part entails

$$\dot{L}_x(1+u) = \frac{1}{x}\left(\ln(1+u)(1+u)^x - L_x(1+u)\right). \tag{40}$$

Let us note that when $x = 0$,

$$\dot{L}_0(1+u) = \lim_{x\to 0}\dot{L}_x(1+u) = \frac{1}{2}\ln^2(1+u).$$

Using (38), (40) and remarking that $R_x(u) = \ln(1+u) - u + u^2/2$ yield

$$\dot{L}_x(1+u) = \frac{u^2}{2} + \bar{R}_x(u), \tag{41}$$

where

$$\begin{aligned} \bar{R}_x(u) &= \frac{x-2}{2}u^3 - \frac{x-1}{4}u^4 + u(R_x(u)+R_0(u)) - \frac{1}{2}R_x(u)u^2 + \frac{R_0(u)-R_x(u)}{x} \\ &+ \frac{x-1}{2}R_0(u)u^2 + R_x(u)R_0(u). \end{aligned} \tag{42}$$

Taking account of

$$\lim_{x\to 0}\frac{R_0(u)-R_x(u)}{x} = -\dot{K}_0(1+u) + \frac{u^2}{2} = -\frac{1}{2}\ln^2(1+u) + \frac{u^2}{2},$$

it follows that

$$\bar{R}_0(u) = \lim_{x\to 0}\bar{R}_x(u) = \dot{L}_0(1+u) - \frac{u^2}{2} = \frac{u^4}{8} + \frac{R_0^2(u)}{2} - \frac{u^3}{2} + uR_0(u) - \frac{u^2R_0(u)}{2}.$$

The second step is to focus on the integral $\mathcal{I}_1(t,x)$. From (41), it can be rewritten as

$$\mathcal{I}_1(t,x) = 1 + \int_0^1 t^2\bar{R}_x\left(\frac{\ln(1/s)}{t}\right)ds.$$

Now, using (42),

$$\begin{aligned} \int_0^1 t^2\bar{R}_x\left(\frac{\ln(1/s)}{t}\right)ds &= \frac{3(x-2)}{t} - \frac{6(x-1)}{t^2} + \int_0^1 \ln(1/s)t\left(R_x\left(\frac{\ln(1/s)}{t}\right) + R_0\left(\frac{\ln(1/s)}{t}\right)\right)ds \\ &- \frac{1}{2}\int_0^1 \ln^2(1/s)R_x\left(\frac{\ln(1/s)}{t}\right)ds + \frac{1}{x}\int_0^1 t^2\left(R_0\left(\frac{\ln(1/s)}{t}\right) - R_x\left(\frac{\ln(1/s)}{t}\right)\right)ds \\ &+ \frac{x-1}{2}\int_0^1 \ln^2(1/s)R_0\left(\frac{\ln(1/s)}{t}\right)ds + \int_0^1 t^2R_x\left(\frac{\ln(1/s)}{t}\right)R_0\left(\frac{\ln(1/s)}{t}\right)ds. \end{aligned}$$

It is clear that the first two terms converge to 0 as $t\to\infty$ uniformly on $x\in I$. Considering the third term, one has by (39) that

$$0 \le \int_0^1 \ln(1/s)t\left(R_x\left(\frac{\ln(1/s)}{t}\right) + R_0\left(\frac{\ln(1/s)}{t}\right)\right)ds \le \frac{4(x-1)(x-2)}{t^2} + \frac{8}{t^2}.$$

As a consequence, the third term also converges to 0 as $t\to\infty$ uniformly on $x\in I$. A similar proof can be done for the fifth, sixth and seventh terms. Considering the fourth term, let us remark that the function $x\to x^{-1}(R_0(u)-R_x(u))$ is decreasing for all $u>0$. Thus, for all $x\in I =: [x_1,x_2]$,

$$\mathcal{J}_{x_2}(t) \le \frac{1}{x}\int_0^1 t^2\left(R_0\left(\frac{\ln(1/s)}{t}\right) - R_x\left(\frac{\ln(1/s)}{t}\right)\right)ds \le \mathcal{J}_{x_1}(t)$$

where we have introduced

$$\mathcal{J}_x(t) := \frac{1}{x}\int_0^1 t^2\left(R_0\left(\frac{\ln(1/s)}{t}\right) - R_x\left(\frac{\ln(1/s)}{t}\right)\right)ds$$

for $x\neq 0$ and

$$\mathcal{J}_0(t) := \lim_{x\to 0}\mathcal{J}_x(t) = -\frac{1}{2}\int_0^1 t^2\ln^2\left(1+\frac{\ln(1/s)}{t}\right)ds + 1.$$

From (39), one can show that $\mathcal{J}_x(t) \to 0$ as $t \to \infty$ uniformly on $x \in I$. Finally $\mathcal{I}_1(t,x) \to 1$ as $t \to \infty$, uniformly on $x \in I$. A similar proof can be done to show that $\mathcal{I}_2(t,x) \to 6$ as $t \to \infty$, uniformly on $x \in I$. Moreover, Lemma 5(i) implies that $t(t^2\mu_2(t,x))^2\Psi_t'(x) \to -2$ and that $(t^2\mu_2(t,x))^2 \to 4$ as $t \to \infty$ uniformly on $x \in I$. The result is thus proved: $\Psi_t'(x) \to -1/2$ as $t \to \infty$ uniformly on $x \in I$. ■

Proof of Lemma 6 – (i) For $i = 1, \ldots, m$, let

$$Y_{m,i} := t_m^{J-1} F_i \left(1 + \frac{F_i}{t_m}\right)^{\zeta_1 - 1} \prod_{j=2}^{J} L_{\zeta_j}\left(1 + \frac{F_i}{\delta t_m}\right).$$

Let $\underline{\zeta} := \min\{\xi_1, \xi_2\}$. The following inequalities hold for all $i = 1, \ldots, m$:

$$\frac{F_i}{\delta t_m} + \frac{\underline{\zeta} - 1}{2}\frac{F_i^2}{\delta^2 t_m^2} \leq \frac{F_i}{\delta t_m} + \frac{\zeta_j - 1}{2}\frac{F_i^2}{\delta^2 t_m^2} \leq L_{\zeta_j}\left(1 + \frac{F_i}{\delta t_m}\right) \leq \frac{F_i}{\delta t_m}.$$

Since $\max\{F_1, \ldots, F_m\} - \ln(m)$ converges in distribution to a Gumbel random variable,

$$0 \leq \max_{1 \leq i \leq m} \frac{F_i}{t_m} = \frac{\ln(m)}{t_m}\left(1 + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\ln(m)}\right)\right) = o_{\mathbb{P}}(1), \tag{43}$$

by assumption. It follows that uniformly in $i \in \{1, \ldots, m\}$,

$$Y_{m,i} = \delta^{1-J} F_i^J \left(1 + o_{\mathbb{P}}(1)\right).$$

The law of large numbers entails that

$$\frac{1}{m}\sum_{i=1}^{m} F_i^J \xrightarrow{\mathbb{P}} \mathbb{E}(F_1^J) = J!,$$

and the conclusion follows.
(ii) Since for all $\xi < 1$ and $u > 0$,

$$L_\xi(1+u) = u - \frac{\xi - 1}{2}u^2 + R_\xi(u),$$

with $0 \leq R_\xi(u) \leq (\xi - 1)(\xi - 2)u^3/6$ and taking into account of (43), it follows that uniformly in $i \in \{1, \ldots, m\}$

$$L_{\xi_3}\left(1 + \frac{F_i}{\delta t_m}\right) - L_{\xi_4}\left(1 + \frac{F_i}{\delta t_m}\right) = \frac{\xi_3 - \xi_4}{2}\left(\frac{F_i}{\delta t_m}\right)^2 (1 + o_{\mathbb{P}}(1)).$$

The rest of the proof follows the same lines as the one of (i) and is thus omitted. ■

Proof of Lemma 7 – Using the Cramér-Wold device, it suffices to obtain the asymptotic distribution of

$$T_n := k_n^{1/2}\left\{\beta_1 \frac{S_n(\zeta^{(1)})}{\mu(\ln(n/k_n), \zeta^{(1)})} + \beta_2 \frac{S_n(\zeta^{(2)})}{\mu(\ln(n/k_n), \zeta^{(2)})} - (\beta_1 + \beta_2)\right\},$$

where $(\beta_1, \beta_2) \in \mathbb{R}^2$. Let us introduce the random processes indexed by $t > 0$

$$W_{i,n}(t) := \frac{\beta_1}{\mu(\ln(n/k_n), \zeta^{(1)})} \prod_{j=1}^{J_1} L_{\zeta_j^{(1)}}\left(1 + \frac{F_i}{t}\right) + \frac{\beta_2}{\mu(\ln(n/k_n), \zeta^{(2)})} \prod_{j=1}^{J_2} L_{\zeta_j^{(2)}}\left(1 + \frac{F_i}{t}\right),$$

for $i = 1, \ldots, k_n$ and where $F_1, \ldots, F_{k_n}$ are independent standard exponential random variables. Lemma 4(ii) yields

$$T_n \stackrel{d}{=} k_n^{-1/2} \sum_{i=1}^{k_n} \left\{ W_{i,n}(E_{n-k_n,n}) - \mathbb{E}\left(W_{i,n}(\ln(n/k_n))\right) \right\},$$

where $E_{n-k_n,n}$ is the $(n-k_n)$th ordered statistic associated to a sample $E_1, \ldots, E_n$ of standard exponential random values independent of $F_1, \ldots, F_{k_n}$. Let us consider the following expansion $T_n =: T_{n,1} + T_{n,2}$ with

$$T_{n,1} := k_n^{-1/2} \sum_{i=1}^{k_n} \bar{W}_{i,n}(\ln(n/k_n))$$

where $\bar{W}_{i,n}(\ln(n/k_n)) := W_{i,n}(\ln(n/k_n)) - \mathbb{E}\left(W_{i,n}(\ln(n/k_n))\right)$ and

$$T_{n,2} := k_n^{-1/2} \sum_{i=1}^{k_n} \left\{ W_{i,n}(E_{n-k_n,n}) - W_{i,n}(\ln(n/k_n)) \right\}.$$

The asymptotic normality of the random term $T_{n,1}$ is obtained by Lyapunov's theorem. Let us observe that

$$s_n^2 := \mathrm{Var}\left(\sum_{i=1}^{k_n} \bar{W}_{i,n}(\ln(n/k_n)) \right) = k_n \left\{ \mathbb{E}\left(W_{1,n}^2(\ln(n/k_n))\right) - (\beta_1+\beta_2)^2 \right\}.$$

Straightforward calculations then lead to

$$\begin{aligned} \mathbb{E}\left(W_{1,n}^2(\ln(n/k_n))\right) &= \beta_1^2 \frac{\mu(\ln(n/k_n), (\zeta^{(1)}, \zeta^{(1)}))}{\mu^2(\ln(n/k_n), \zeta^{(1)})} + \beta_2^2 \frac{\mu(\ln(n/k_n), (\zeta^{(2)}, \zeta^{(2)}))}{\mu^2(\ln(n/k_n), \zeta^{(2)})} \\ &+ 2\beta_1\beta_2 \frac{\mu(\ln(n/k_n), (\zeta^{(1)}, \zeta^{(2)}))}{\mu(\ln(n/k_n), \zeta^{(1)})\mu(\ln(n/k_n), \zeta^{(2)})}. \end{aligned}$$

As a direct consequence of Lemma 5(i), one has

$$\lim_{n\to\infty} \mathbb{E}\left(W_{1,n}^2(\ln(n/k_n))\right) = \frac{(2J_1)!}{(J_1!)^2}\beta_1^2 + \frac{(2J_2)!}{(J_2!)^2}\beta_2^2 + 2\frac{(J_1+J_2)!}{J_1!J_2!}\beta_1\beta_2, \tag{44}$$

and thus $s_n^2 \sim c(\beta_1,\beta_2)k_n$ as $n \to \infty$ where the constant $c(\beta_1,\beta_2)$ is given by

$$c(\beta_1,\beta_2) := ((2J_1)!/(J_1!)^2 - 1)\beta_1^2 + ((2J_2)!/(J_2!)^2 - 1)\beta_2^2 + 2((J_1+J_2)!/(J_1!J_2!) - 1)\beta_1\beta_2.$$

Let us now check Lyapunov's condition *i.e.* that

$$\frac{1}{k_n^2} \sum_{i=1}^{k_n} \mathbb{E}\left(\bar{W}_{i,n}^4(\ln(n/k_n))\right) = \frac{1}{k_n} \mathbb{E}\left(\bar{W}_{1,n}^4(\ln(n/k_n))\right) \to 0, \tag{45}$$

as $n \to \infty$. By similar arguments as the ones leading to (44), one can show that

$$\mathbb{E}\left(\bar{W}_{1,n}^4(\ln(n/k_n))\right) = \sum_{l=1}^{4} (-1)^l C_4^l \mathbb{E}\left(W_{1,n}^l(\ln(n/k_n))\right) \mathbb{E}^{4-l}\left(W_{1,n}(\ln(n/k_n))\right)$$

converges to a constant as $n \to \infty$ and thus (45) holds. As a conclusion

$$T_{n,1} \xrightarrow{d} \mathcal{N}\left(0, c(\beta_1,\beta_2)\right). \tag{46}$$

It remains to prove that $T_{n,2} \xrightarrow{\mathbb{P}} 0$. For $i = 1, \ldots, k_n$, let $\dot{W}_{i,n}(\cdot)$ be the first derivative of the random function $W_{i,n}(\cdot)$. The mean-value theorem entails that

$$W_{i,n}(E_{n-k_n,n}) - W_{i,n}(\ln(n/k_n)) = \dot{W}_{i,n}(E^*_{n,i})(E_{n-k_n,n} - \ln(n/k_n)),$$

where for $i = 1, \ldots, k_n$, $E^*_{n,i} = \ln(n/k_n) + \Theta_{n,i}(E_{n-k_n,n} - \ln(n/k_n))$ with $\Theta_{n,i}$ a random variable in $(0,1)$. Recalling that $k_n^{1/2}(E_{n-k_n,n} - \ln(n/k_n)) \xrightarrow{d} \mathcal{N}(0,1)$, in order to show that $T_{n,2} \xrightarrow{\mathbb{P}} 0$, it suffices to prove that

$$\frac{1}{k_n}\sum_{i=1}^{k_n} \dot{W}_{i,n}(E^*_{n,i}) \xrightarrow{\mathbb{P}} 0. \tag{47}$$

First, simple calculations show that for all $t > 0$,

$$\begin{aligned} -\frac{1}{k_n}\sum_{i=1}^{k_n} \dot{W}_{i,n}(t) &= \frac{\beta_1}{t^2\mu(\ln(n/k_n),\zeta^{(1)})}\sum_{l=1}^{J_1}\left(\frac{1}{k_n}\sum_{i=1}^{k_n} F_i\left(1+\frac{F_i}{t}\right)^{\zeta_l^{(1)}-1}\prod_{j\neq l} L_{\zeta_j^{(1)}}\left(1+\frac{F_i}{t}\right)\right) \\ &+ \frac{\beta_2}{t^2\mu(\ln(n/k_n),\zeta^{(2)})}\sum_{l=1}^{J_2}\left(\frac{1}{k_n}\sum_{i=1}^{k_n} F_i\left(1+\frac{F_i}{t}\right)^{\zeta_l^{(2)}-1}\prod_{j\neq l} L_{\zeta_j^{(2)}}\left(1+\frac{F_i}{t}\right)\right). \end{aligned}$$

Hence, we have to deal with random terms proportional to $T_{3,n}(t,t)$ where

$$T_{3,n}(t_1,t_2) := \frac{1}{t_2^2\mu(\ln(n/k_n),\zeta)}\frac{1}{k_n}\sum_{i=1}^{k_n} F_i\left(1+\frac{F_i}{t_1}\right)^{\zeta_1-1}\prod_{j=2}^{J} L_{\zeta_j}\left(1+\frac{F_i}{t_2}\right)$$

with $J \in \mathbb{N}\setminus\{0\}$ and $\zeta \in \mathbb{R}^J$. To prove (47) let us check that

$$\bar{T}_{3,n} := \frac{1}{(E^*_{n,i})^2\mu(\ln(n/k_n),\zeta)}\frac{1}{k_n}\sum_{i=1}^{k_n} F_i\left(1+\frac{F_i}{E^*_{n,i}}\right)^{\zeta_1-1}\prod_{j=2}^{J} L_{\zeta_j}\left(1+\frac{F_i}{E^*_{n,i}}\right) \xrightarrow{\mathbb{P}} 0. \tag{48}$$

For all $\eta > 0$, let $0 < \varepsilon < \min\{\eta_+, \eta_-\}$ where $\eta_+ = 1 - ((1+\eta/2)/(1+\eta))^{1/2}$ and $\eta_- = ((1-\eta/2)/(1-\eta))^{1/2} - 1$. Let us also introduce the Borel set

$$A_{n,\varepsilon} = \left\{\left|\frac{E_{n-k_n,n} - \ln(n/k_n)}{\ln(n/k_n)}\right| \leq \varepsilon\right\}. \tag{49}$$

For all $\eta > 0$, remark that

$$\mathbb{P}\left(|\bar{T}_{n,3}| > \eta\right) \leq \mathbb{P}(A^{\mathrm{C}}_{n,\varepsilon}) + \mathbb{P}\left(\{|\bar{T}_{n,3}| > \eta\} \cap A_{n,\varepsilon}\right) \tag{50}$$

and recall that, for $i = 1, \ldots, k_n$, $E^*_{n,i} = \ln(n/k_n) + \Theta_{i,n}(E_{n-k_n,n} - \ln(n/k_n))$. Since $\Theta_{n,i} \in (0,1)$ it is clear that under $A_{n,\varepsilon}$, one has

$$(1-\varepsilon)\ln(n/k_n) \leq E^*_{n,i} \leq (1+\varepsilon)\ln(n/k_n)$$

for all $i = 1, \ldots, k_n$. Hence

$$T_{n,3}\left((1-\varepsilon)\ln(n/k_n), (1+\varepsilon)\ln(n/k_n)\right) \leq \bar{T}_{n,3} \leq T_{n,3}\left((1+\varepsilon)\ln(n/k_n), (1-\varepsilon)\ln(n/k_n)\right),$$

and thus

$$\begin{aligned} \mathbb{P}\left(\{|\bar{T}_{n,3}| > \eta\} \cap A_{n,\varepsilon}\right) &\leq \mathbb{P}\left(T_{n,3}\left((1+\varepsilon)\ln(n/k_n), (1-\varepsilon)\ln(n/k_n)\right) > \eta\right) \\ &+ \mathbb{P}\left(T_{n,3}\left((1-\varepsilon)\ln(n/k_n), (1+\varepsilon)\ln(n/k_n)\right) < -\eta\right). \end{aligned}$$

Applying Lemma 6(i) and the fact that from Lemma 5(i),

$$\frac{\ln^{1-J}(n/k_n)}{\ln^2(n/k_n)\mu(\ln(n/k_n),\zeta)} \sim \frac{1}{J!}\frac{1}{\ln(n/k_n)} \to 0,$$

it follows that $T_{n,3}\left((1\pm\varepsilon)\ln(n/k_n),(1\mp\varepsilon)\ln(n/k_n)\right) \xrightarrow{\mathbb{P}} 0$. As a consequence,

$$\mathbb{P}\left(\{|\bar{T}_{n,3}|>\eta\}\cap A_{n,\varepsilon}\right)\to 0 \tag{51}$$

as $n\to\infty$. Furthermore, since $(E_{n-k_n,n}-\ln(n/k_n))/\ln(n/k_n) = \mathcal{O}_{\mathbb{P}}(k_n^{-1/2}\ln^{-1}(n/k_n)) = o_{\mathbb{P}}(1)$, one has that $\mathbb{P}(A_{n,\varepsilon}) \xrightarrow{\mathbb{P}} 1$ as $n\to\infty$. Collecting this last result, (50) and (51) implies (48) and thus (47) and the conclusion follows. ■

Proof of Lemma 8 – Let $E_1,\ldots,E_n,F_1,\ldots,F_n$ be a sample of $2n$ independent standard exponential random variables and let $E_{n-k_n,n}$ be the $(n-k_n)$th ordered statistic associated with the sample $E_1,\ldots,E_n$. Let us also introduce the random variable defined for all $t>0$ by

$$W_n(t) := \frac{(\ln(n/k_n))^J}{k_n}\sum_{i=1}^{k_n} L_{\xi_1}^{J_1}\left(1+\frac{F_i}{t}\right)L_{\xi_2}^{J_2}\left(1+\frac{F_i}{t}\right)\left(L_{\xi_3}\left(1+\frac{F_i}{t}\right)-L_{\xi_4}\left(1+\frac{F_i}{t}\right)\right)^{J_3}.$$

According to Lemma 4(ii), we have to prove that $W_n(E_{n-k_n,n}) \xrightarrow{\mathbb{P}} K := J!((\xi_3-\xi_4)/2)^{J_3}$. For all $\varepsilon>0$, let us consider the Borel set

$$\mathcal{A}_{n,\varepsilon} = \left\{\left|\frac{E_{n-k_n,n}}{\ln(n/k_n)}-1\right|\le\varepsilon\right\},$$

introduced in (49). Remarking that the function $x\mapsto L_{\xi_1}^{J_1}(x)L_{\xi_2}^{J_2}(x)(L_{\xi_3}(x)-L_{\xi_4}(x))^{J_3}$ is increasing on $(1,\infty)$ leads to

$$\begin{aligned}\mathbb{P}\left\{|W_n(E_{n-k_n,n})-K|>\varepsilon\right\} &\le \mathbb{P}\left\{W_n((1+\varepsilon)\ln(n/k_n))>K+\varepsilon\right\}\\ &+ \mathbb{P}\left\{W_n((1-\varepsilon)\ln(n/k_n))<K-\varepsilon\right\}+1-\mathbb{P}(\mathcal{A}_{n,\varepsilon}).\end{aligned} \tag{52}$$

Using the inequality $K+\varepsilon-K/(1+\varepsilon)^J\ge\varepsilon$ yields

$$\mathbb{P}\left\{W_n((1+\varepsilon)\ln(n/k_n))>K+\varepsilon\right\}\le\mathbb{P}\left\{W_n((1+\varepsilon)\ln(n/k_n))-\frac{K}{(1+\varepsilon)^J}>\varepsilon\right\}.$$

Since $\ln(k_n)/\ln(n)\to 0$, one can apply Lemma 6(ii) with $m=k_n$ and $t_m=\ln(n/k_n)$ to obtain

$$\lim_{n\to\infty}\mathbb{P}\left\{W_n((1+\varepsilon)\ln(n/k_n))>K+\varepsilon\right\}=0. \tag{53}$$

In the same way, one has

$$\lim_{n\to\infty}\mathbb{P}\left\{W_n((1-\varepsilon)\ln(n/k_n))<K-\varepsilon\right\}=0. \tag{54}$$

Finally, since $E_{n-k_n,n}/\ln(n/k_n)\xrightarrow{\mathbb{P}}1$, we conclude the proof by collecting (52) to (54). ■

	$\bar{F}(x)$	θ	ρ	ρ'	ρ''
$\boldsymbol{\theta = 0}$					
Gamma $(a > 0, s > 0)$	$\frac{1}{s^a \Gamma(a)} \int_x^\infty t^{a-1} e^{-t/s} dt$ $x \geq 0$	0	-1	0	-1
Weibull $(k \neq 1, \lambda > 0)$	$e^{-(x/\lambda)^k}$ $x \geq 0$	0	$-\infty$	0	-1
Gaussian $(\mu \in \mathbb{R}, \sigma > 0)$	$\frac{1}{\sigma\sqrt{2\pi}} \int_x^\infty \exp\left(-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right) dt$ $x \in \mathbb{R}$	0	-1	0	-1
$\boldsymbol{\theta > 0}$					
Lognormal $(\mu \in \mathbb{R}, \sigma > 0)$	$\frac{1}{\sigma\sqrt{2\pi}} \int_x^\infty \frac{1}{t} \exp\left(-\frac{(\ln t - \mu)^2}{2\sigma^2}\right) dt$ $x \geq 0$	1/2	-1	-1	-1
Burr $(\lambda > 0, c > 0, k > 0)$	$\left(1 + \left(\frac{x}{\lambda}\right)^c\right)^{-k}$ $x \geq 0$	1	$-\infty$	-1	-2
Pareto-like	$1/U^{\leftarrow}(x)$, $U(x) = x\left(1 + 2\ln^2(x)\right)$	1	-1	-1	-1
Super heavy-tail	$e^{-\ln^{1/2}(x)}$ $x \geq 1$	2	$-\infty$	$-\infty$	$-\infty$
$\boldsymbol{\theta < 0}$					
Finite endpoint $(x^* > 0)$	$\exp\left(-\frac{1}{\ln x^* - \ln x}\right)$ $x \in (0, x^*)$	-1	$-\infty$	-1	-2

Table 1: Examples of distributions verifying **(A1)** and **(A2)** with associated values of θ, ρ, ρ' and $\rho'' := \max(\rho, \rho' - 1)$, see (13).

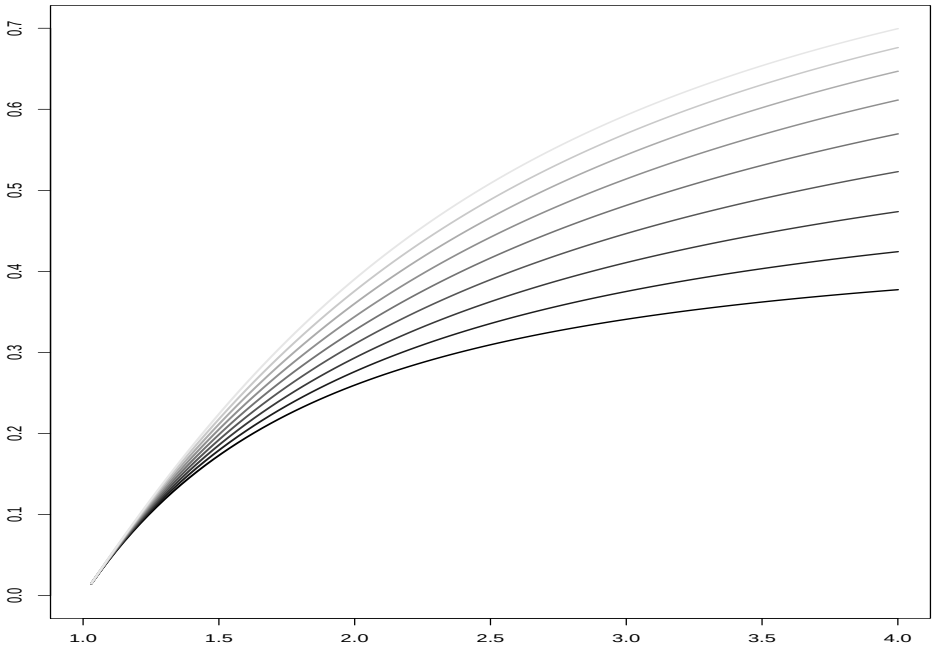

Figure 1: Ratio $\Lambda_\theta(\tau)$ between the asymptotic standard deviations σ_n and σ'_n (see equation (18)) as a function of $\tau \geq 1$ for $\theta \in \{-2, -1.5, \ldots, 2\}$. Dark lines are associated with small values of θ.

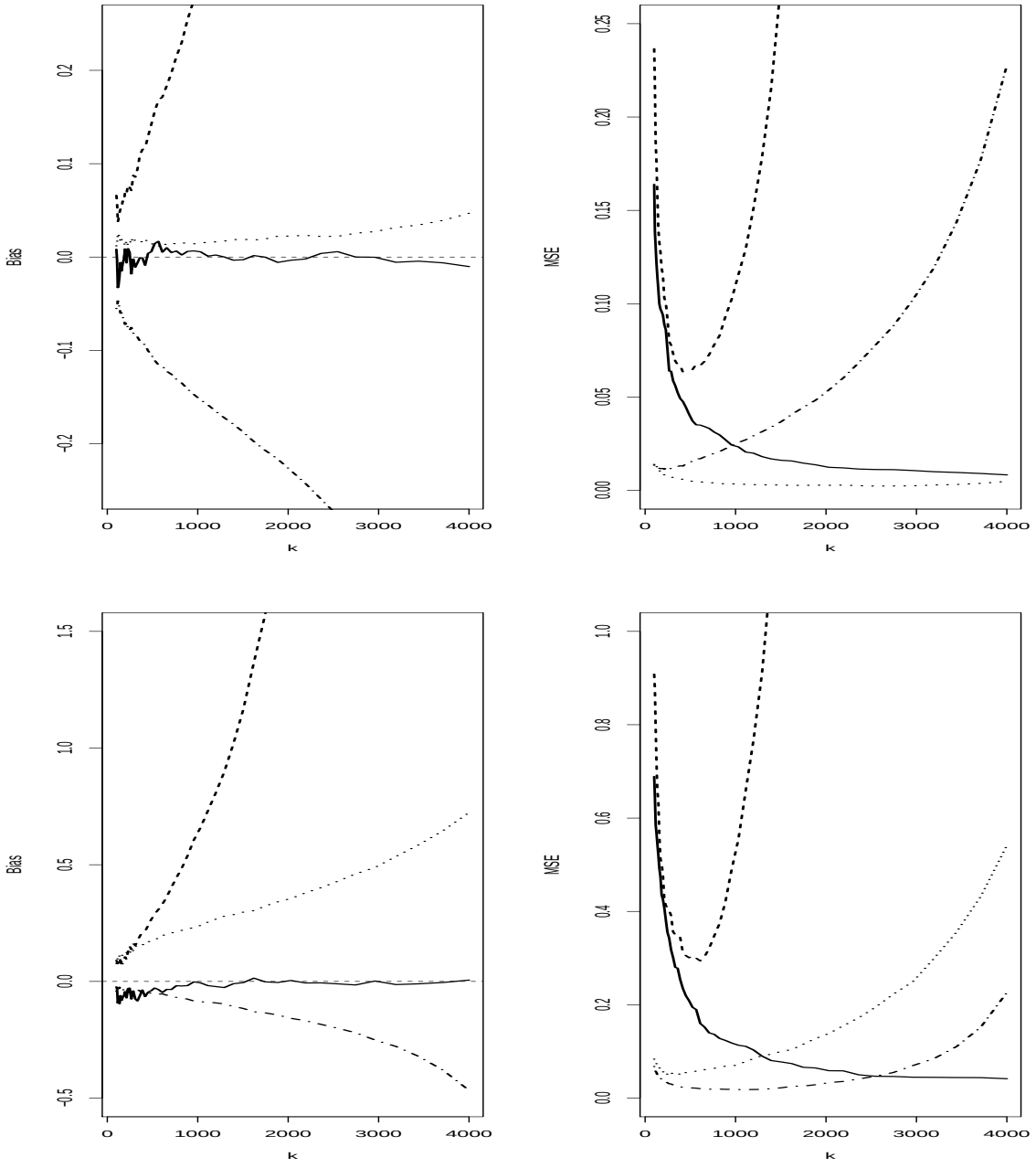

Figure 2: Results on simulated data. Bias (left) and MSE (right) associated with $\check{Q}_n^{[1]}(\beta_n)$ (solid line), $\check{Q}_n^{[2]}(\beta_n)$ (dotted line), $\check{Q}_n^{[3]}(\beta_n)$ (dash-dotted line) and $\check{Q}_n^{[4]}(\beta_n)$ (dashed line) as functions of k_n for $\beta_n = n^{-2}$ and $n = 5000$. Top: Gamma, bottom: Weibull.

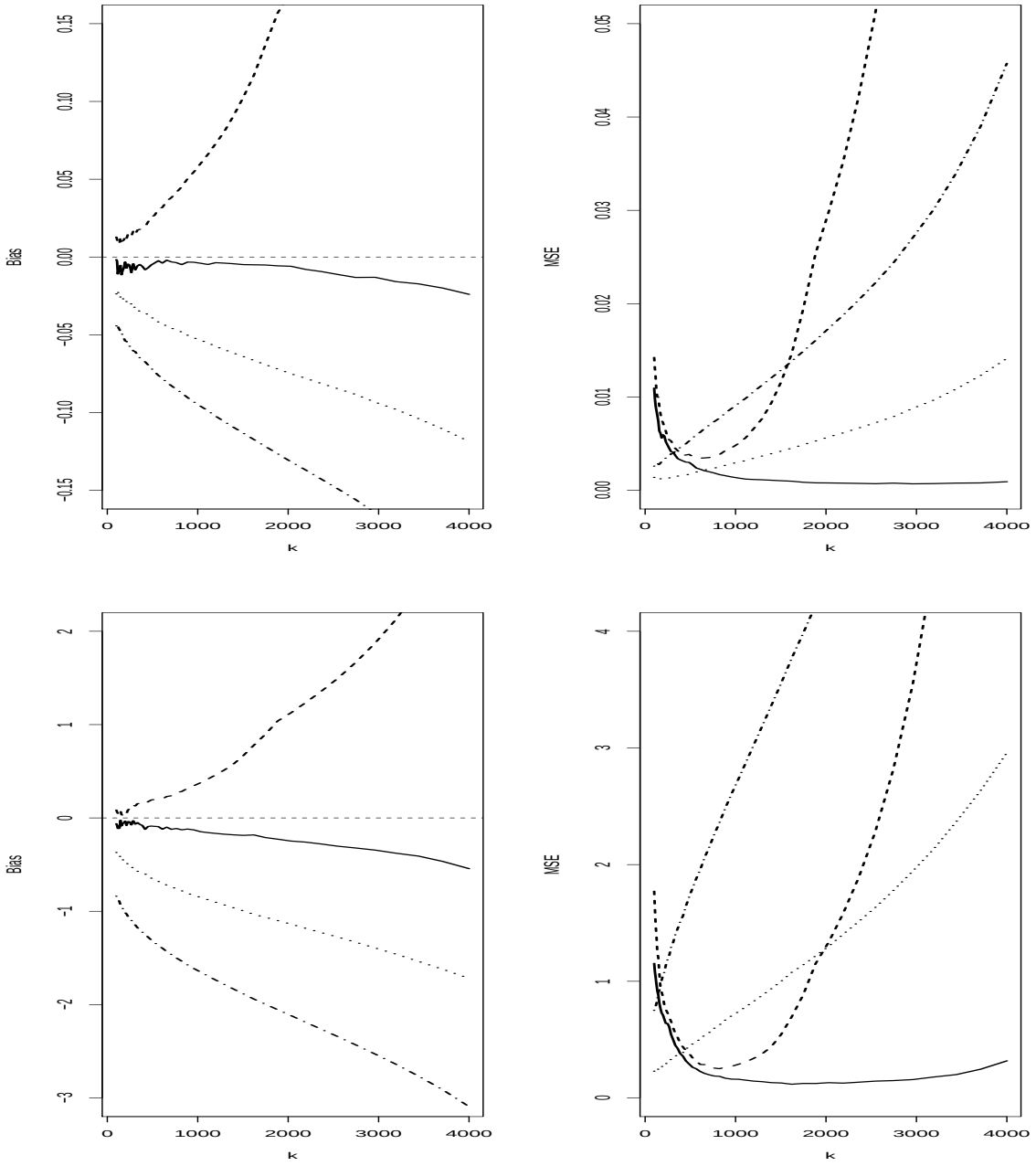

Figure 3: Results on simulated data. Bias (left) and MSE (right) associated with $\check{Q}_n^{[1]}(\beta_n)$ (solid line), $\check{Q}_n^{[2]}(\beta_n)$ (dotted line), $\check{Q}_n^{[3]}(\beta_n)$ (dash-dotted line) and $\check{Q}_n^{[4]}(\beta_n)$ (dashed line) as functions of k_n for $\beta_n = n^{-2}$ and $n = 5000$. Top: Gaussian, bottom: Lognormal.

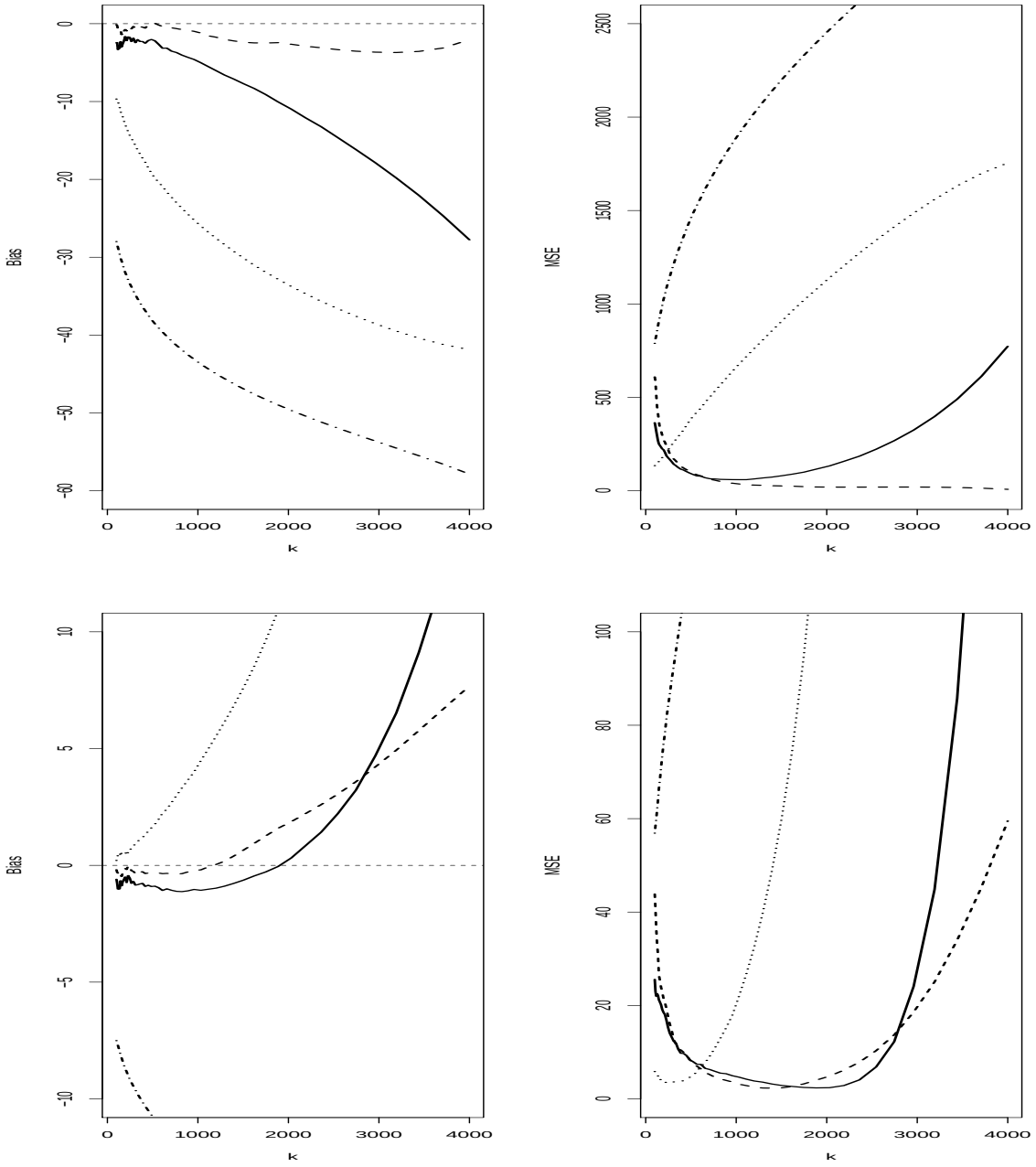

Figure 4: Results on simulated data. Bias (left) and MSE (right) associated with $\check{Q}_n^{[1]}(\beta_n)$ (solid line), $\check{Q}_n^{[2]}(\beta_n)$ (dotted line), $\check{Q}_n^{[3]}(\beta_n)$ (dash-dotted line) and $\check{Q}_n^{[4]}(\beta_n)$ (dashed line) as functions of k_n for $\beta_n = n^{-2}$ and $n = 5000$. Top: Burr, bottom: Pareto-like.

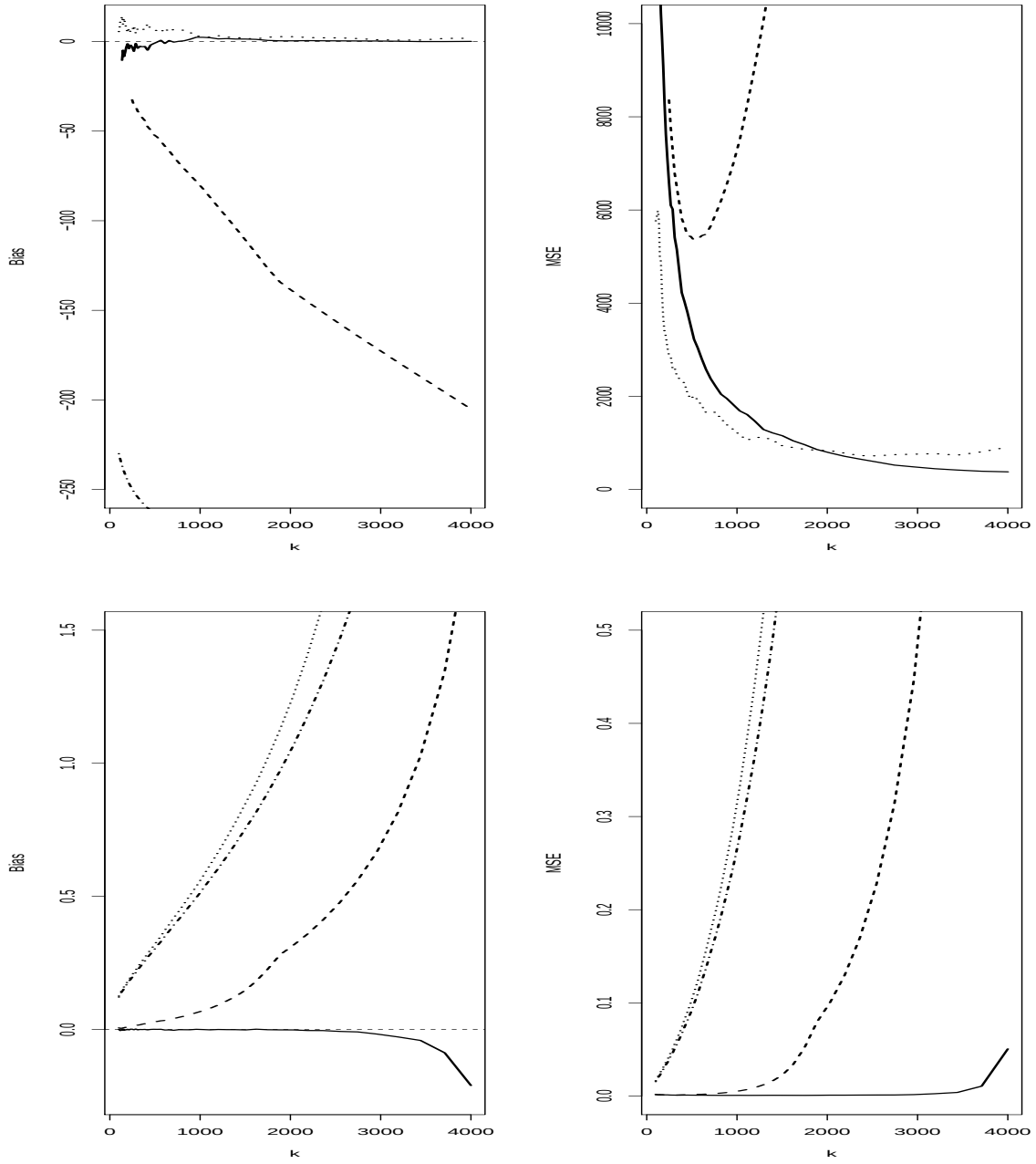

Figure 5: Results on simulated data. Bias (left) and MSE (right) associated with $\check{Q}_n^{[1]}(\beta_n)$ (solid line), $\check{Q}_n^{[2]}(\beta_n)$ (dotted line), $\check{Q}_n^{[3]}(\beta_n)$ (dash-dotted line) and $\check{Q}_n^{[4]}(\beta_n)$ (dashed line) as functions of k_n for $\beta_n = n^{-2}$ and $n = 5000$. Top: super heavy-tail, bottom: finite endpoint.

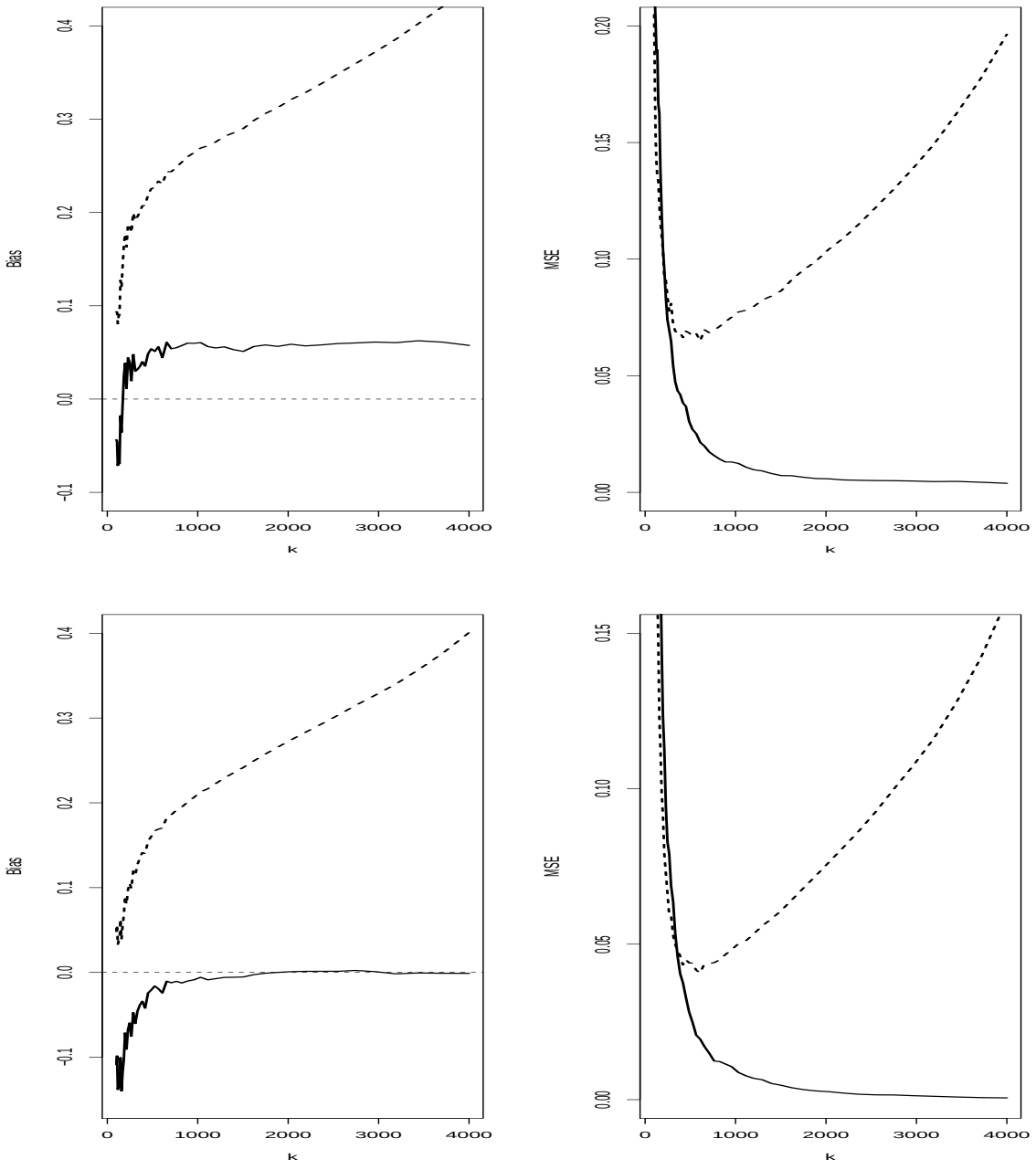

Figure 6: Results on simulated data. Bias (left) and MSE (right) associated with $\hat{\theta}_n^{(M)}$ (solid line) and $\hat{\theta}_{k_n,n}^{[4]}$ (dashed line) as functions of k_n for $n = 5000$. Top: Gamma, bottom: Weibull.

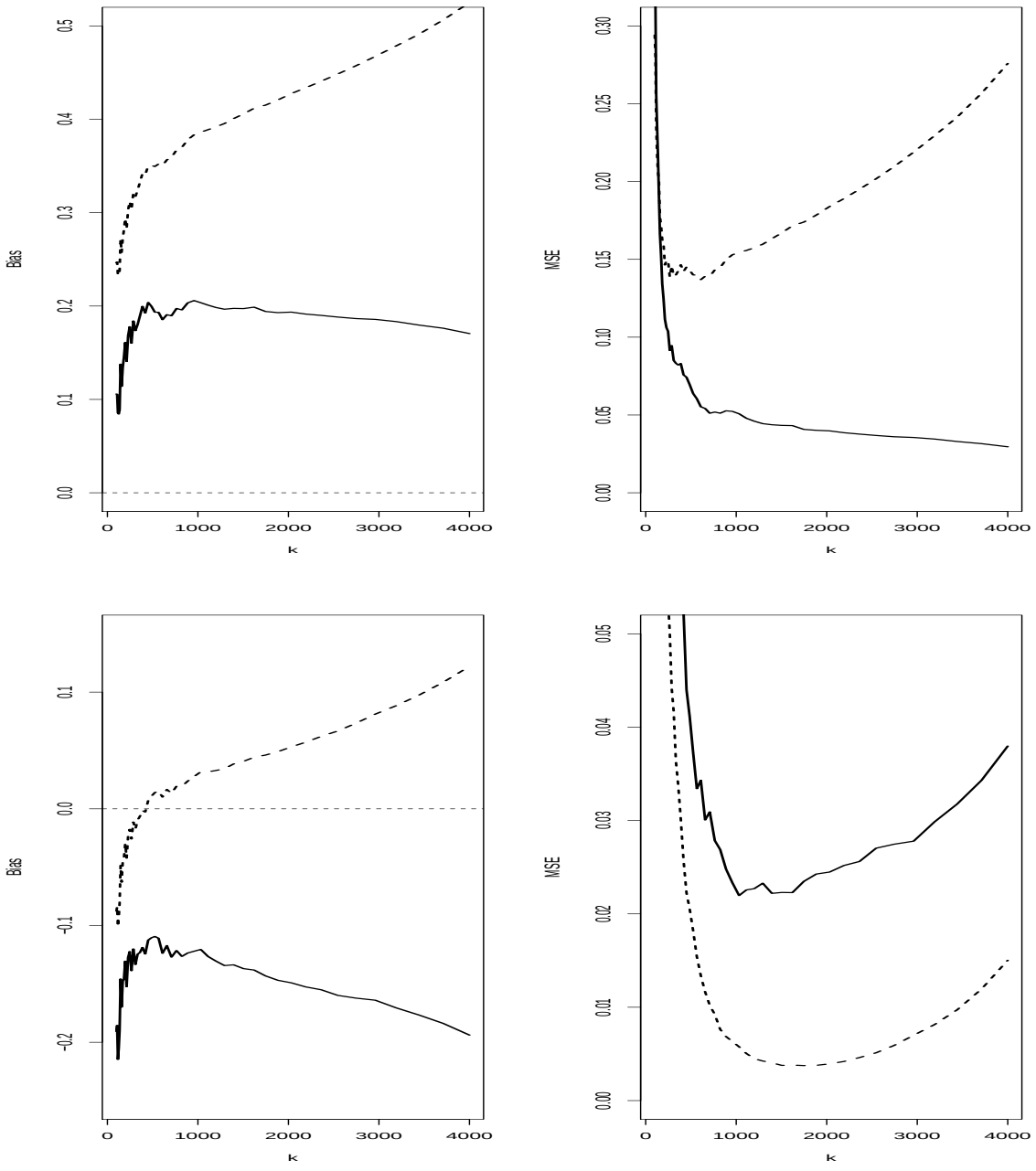

Figure 7: Results on simulated data. Bias (left) and MSE (right) associated with $\hat{\theta}_n^{(M)}$ (solid line) and $\hat{\theta}_{k_n,n}^{[4]}$ (dashed line) as functions of k_n for $n = 5000$. Top: Gaussian, bottom: Lognormal.

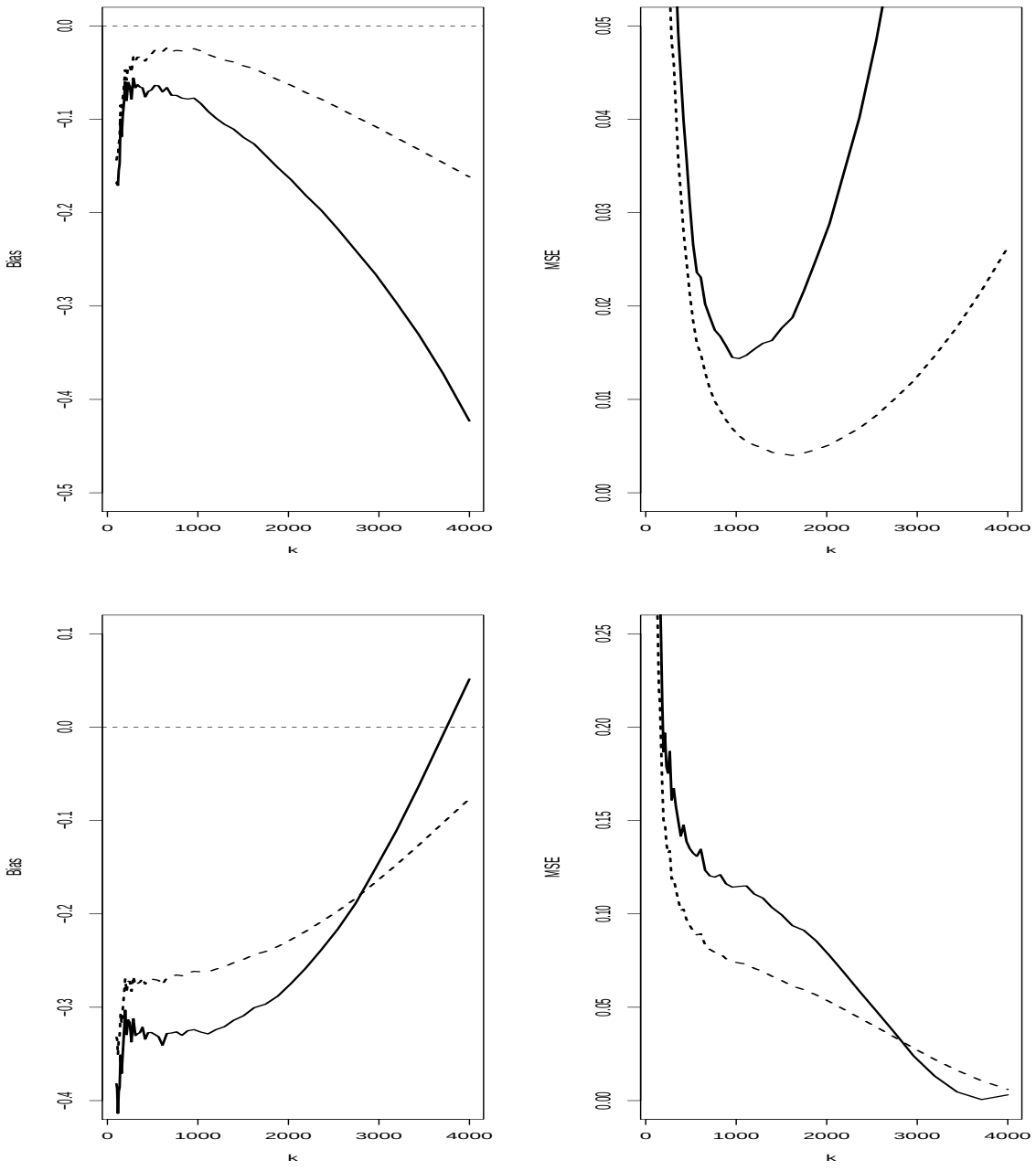

Figure 8: Results on simulated data. Bias (left) and MSE (right) associated with $\hat{\theta}_n^{(M)}$ (solid line) and $\hat{\theta}_{k_n,n}^{[4]}$ (dashed line) as functions of k_n for $n = 5000$. Top: Burr, bottom: Pareto-like.

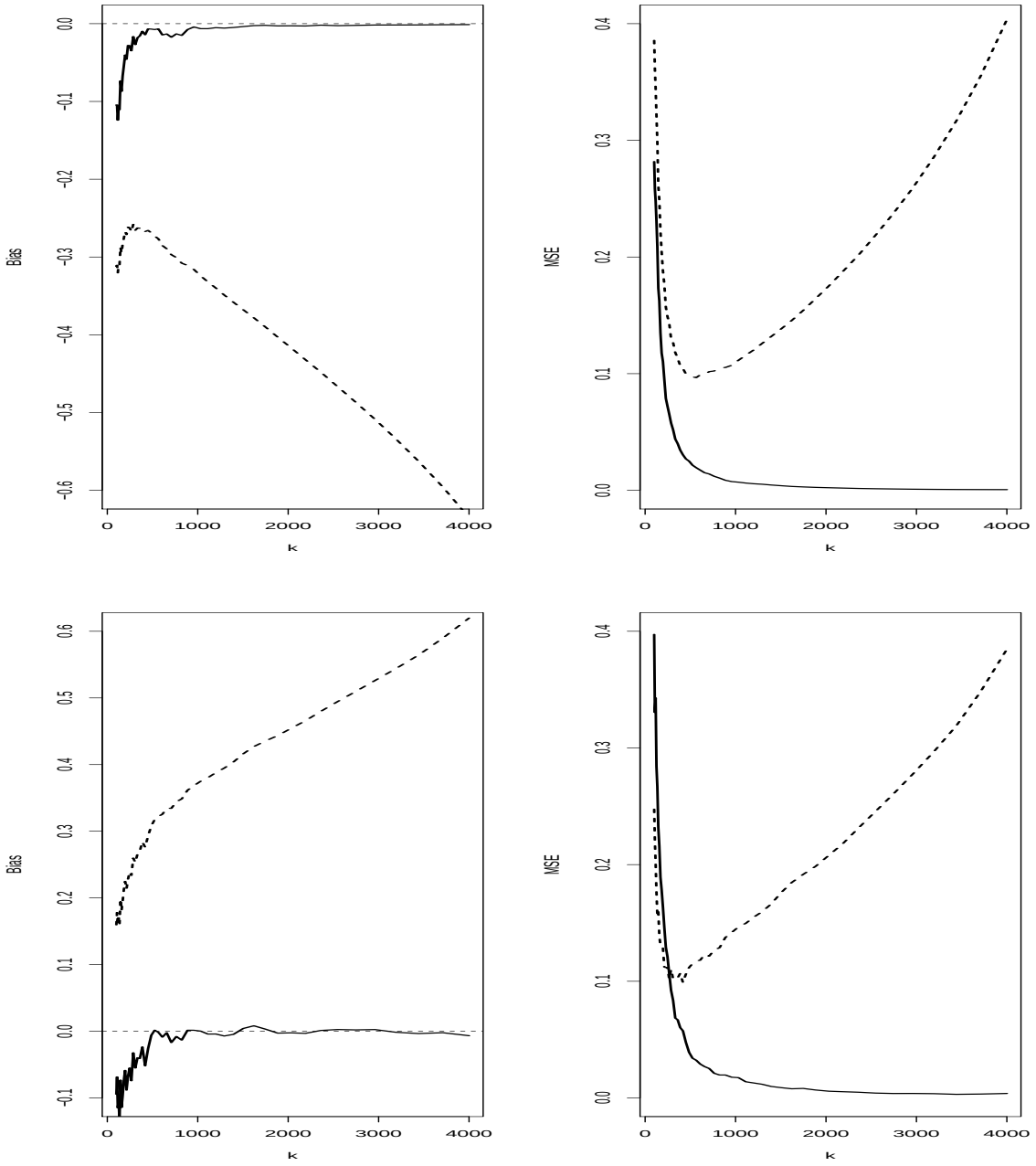

Figure 9: Results on simulated data. Bias (left) and MSE (right) associated with $\hat{\theta}_n^{(M)}$ (solid line) and $\hat{\theta}_{k_n,n}^{[4]}$ (dashed line) as functions of k_n for $n = 5000$. Top: super heavy-tail, bottom: finite endpoint.

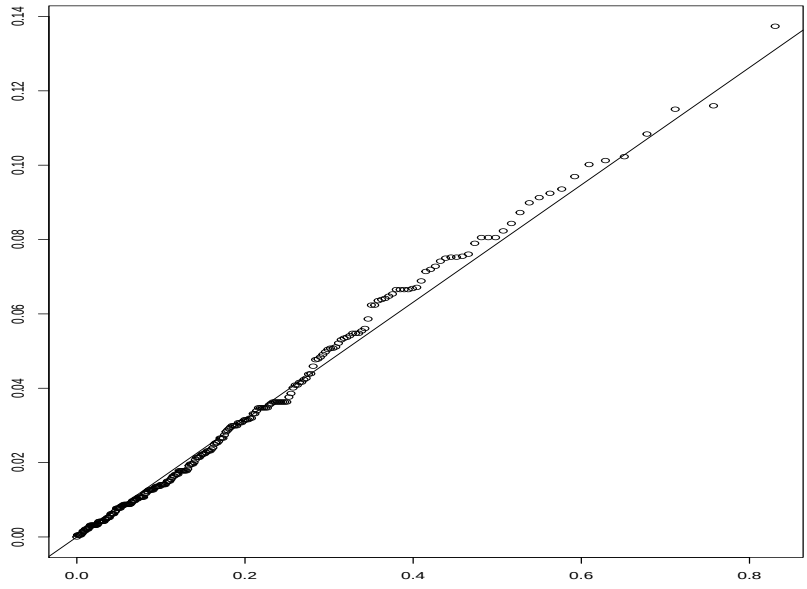

Figure 10: Results on Rhône data. Line of slope $\hat{\theta}_{n,+}^{(M)}$ superimposed to the quantile-quantile plot obtained with $k_n = 252$ (see text for details).

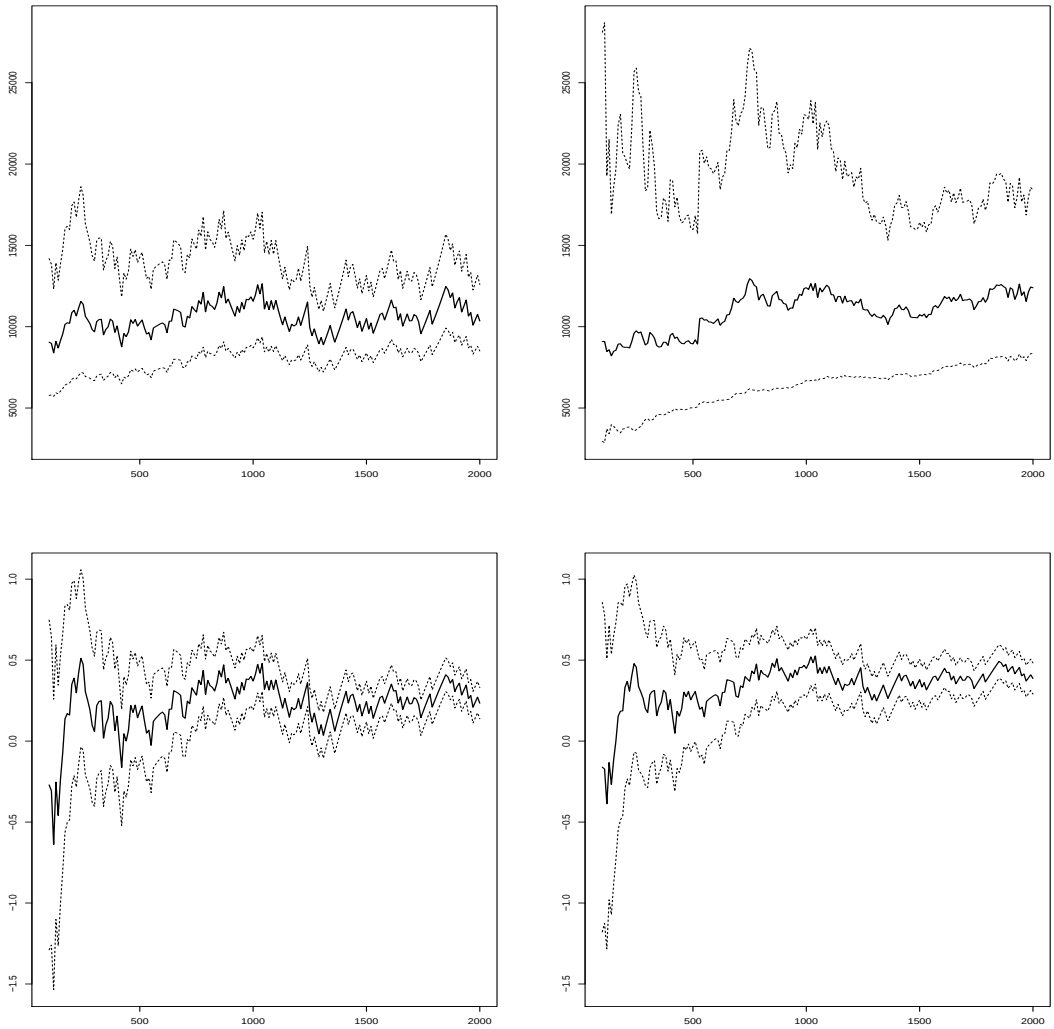

Figure 11: Results on Rhône data. Estimates $\check{Q}_n^{[1]}(\beta_n)$ (top left) and $\check{Q}_n^{[4]}(\beta_n)$ (top right) of $Q(\beta_n)$ with ($\beta_n = 5.5\ 10^{-6}$) and their corresponding index estimates $\hat{\theta}_n^{(M)}$ (bottom left) and $\hat{\theta}_{k_n,n}^{[4]}$ (bottom right) as functions of $k \in \{100, \ldots, 2000\}$. The 95% asymptotic confidence intervals are depicted by dotted lines.

3.3 Perspectives

Dans ce chapitre, nous avons souhaité contribuer à la théorie des valeurs extrêmes, et en particulier contribuer à la popularisation du modèle des lois à queue de type log-Weibull généralisé introduit par DE VALK [2016b], en proposant un estimateur alternatif des quantiles extrêmes. Nous avons ainsi pu montrer que, dans certaines situations, l'estimateur proposé était plus efficace que celui de DE VALK et CAI [2018]. Plusieurs perspectives sont envisageables afin d'améliorer les performances de l'estimateur ainsi proposé.

3.3.1 Lever la contrainte sur ρ

Au vu de la remarque associée au Corollaire 1 (voir partie "Parameters estimation" du Paragraphe 3.2), plusieurs restrictions apparaissent sur le paramètre ρ, et ce en fonction de la vitesse de convergence de l'ordre β_n. Ainsi, si $\beta_n = c/n$, $c \in (0,1)$, alors les lois de type "super heavy tail" avec $\theta > 2$ ne vérifient plus les conditions du Corollaire en question. Dans le cas où $\beta_n = n^{-\tau}$, c'est l'ensemble des lois "super heavy tail" qui ne vérifient plus les conditions, mais aussi certaines lois du domaine d'attraction de Fréchet, telle que la "Pareto-like" décrite Table 1 de l'article. C'est le cas également de certaines lois de type"log-Weibull tail". Pour terminer, dans le cas $\beta_n = \exp(-cn)$, $c > 0$, seules les lois du domaine d'attraction de Gumbel ayant un point terminal fini vérifient les conditions.

Une des perspectives de travail serait donc de lever la contrainte sur ρ. Pour ce faire, on peut commencer par remarquer (cf Corollaire 1 toujours) que ces restrictions dans la vitesse de convergence du terme

$$\frac{k_n^{1/2}/\ln(n/k_n)}{a[\ln(n/k_n)]H_{\theta,0}(d_n)} \ln\left(\frac{\check{Q}_n^{(M)}(\beta_n)}{Q(\beta_n)}\right)$$

proviennent du numérateur $k_n^{1/2}/\ln(n/k_n)$. Or ce dernier est issu de l'estimateur de θ, voir Théorème 3 dans l'article. Une idée serait d'estimer le paramètre θ_- à l'aide d'une suite intermédiaire k_n et de le combiner avec des estimateurs de θ_+ et de a utilisant une autre suite intermédiaire l_n telle que $1 < l_n < k_n$ pour tout n, à l'image de ce qui est fait dans DE VALK et CAI [2018]. Finalement, l'extrapolation du quantile se ferait également à partir du point l_n, permettant de supprimer les précédentes restrictions et d'obtenir un résultat théorique ayant une large portée d'application en pratique.

Evidemment, cela poserait le choix de la suite l_n. Dans DE VALK et CAI [2018], ce choix est effectué à l'aide d'un facteur σ posé égal à un dans les simulations, voir DE VALK et CAI [2018, Proposition 2] pour plus de précisions à ce sujet. En pratique, dans le cas où n n'est pas très grand, il est toujours possible de prendre une plus grande valeur du facteur σ de telle sorte que l_n reste raisonnablement grand.

3.3.2 Choix du nombre k_n de statistiques d'ordre considérées

L'estimateur des quantiles extrêmes proposé dans ce chapitre dépend de k_n. Ce paramètre représente le nombre de statistiques d'ordre que l'on souhaite utiliser pour estimer le quantile en question. Plus ce nombre est grand, plus l'estimation est aisée, menant à des intervalles de confiance assez étroits. D'un autre côté, plus ce nombre est grand, plus l'estimation est biaisée, puisque se basant sur des statistiques d'ordre issues du ventre, et non plus de la queue de distribution. Il s'agit là d'un compromis biais-variance. Trouver un équilibre pour le choix de k_n (et/ou de l_n, voir ci-dessus) demeure un choix délicat en pratique.

Dans la littérature, de nombreux articles cherchent à déterminer le nombre optimal de statistiques d'ordre servant à l'inférence d'un paramètre. C'est particulièrement le cas pour l'estimateur de l'indice des valeurs extrêmes de Hill, voir Paragraphe 1.3.4. On peut citer par exemple DREES et KAUFMANN [1998], DRAISMA et collab. [1999], BEIRLANT et collab. [1999], MATTHYS et BEIRLANT [2000] ou encore DE HAAN et FERREIRA [2007], pages 77-82.

Un des perspectives de ce chapitre serait de trouver le nombre optimal k_n permettant d'estimer au mieux le paramètre θ, qui correspond au paramètre de forme dans le modèle introduit par DE VALK [2016b]. Un premier point serait de considérer l'estimateur de θ_+, très similaire à l'indice des valeurs extrêmes, et d'adapter les méthodes proposées dans les articles ci-dessus à nos hypothèses de modèle.

3.3.3 Tests d'hypothèses

Pour terminer, une perspective plus générale serait de construire des tests d'hypothèses pour le modèle des lois à queue de type log-Weibull généralisé, basé sur le paramètre θ.

Dans la littérature, il existe de nombreux articles qui traitent de la construction de tests permettant de juger de l'appartenance d'une loi à un domaine d'attraction. Des résumés de ces différentes méthodes sont proposés par FRAGA ALVES et GOMES [1996] et NEVES et FRAGA ALVES [2008]. Parmi ces méthodes, on distingue deux types d'approches :

— Les approches paramétriques dont la principale hypothèse concerne l'existence d'une classe de modèles permettant de décrire le processus ayant généré les données. Ces approches se divisent en trois sous-catégories, la première s'intéressant au maximum et les autres aux excès ou à un mélange des deux. Pour des exemples de mise en application de ces approches, citons GOMES [1982], HOSKING [1984], VAN MONTFORT et WITTER [1985] ou encore MAROHN [1998].

— Les approches semi-paramétriques qui consistent à supposer que la loi sous-jacente appartient à un domaine d'attraction. Ces dernières s'intéressent par exemple à des tests du type :

$$H_0 : F \in DA(G_0) \quad vs \quad H_1 : F \in DA(G_\gamma)_{\gamma \neq 0}.$$

Citons par exemple CASTILLO et collab. [1989], HASOFER et WANG [1992], DIEBOLT et collab. [2003] ou encore NEVES et collab. [2006].

Dans notre cas, l'idée serait de construire un test basé sur une approche semi-paramétrique, où l'on suppose que la loi en question vérifie le modèle de DE VALK [2016b]. Ce test pourrait alors nous permettre de mieux caractériser les queues de distribution des lois vérifiant le modèle. A terme, il nous permettrait de pouvoir distinguer les lois à queue super lourde, vérifiant $\theta > 1$, des lois à queue lourde du domaine d'attraction de Fréchet, vérifiant $\theta = 1$, des lois de type Weibull tail, vérifiant $\theta = 0$, ou encore des lois à point terminal fini vérifiant $\theta < 0$.

3.3.4 Estimateur de petites probabilités d'évènements extrêmes multivariés

La proposition d'un estimateur de petites probabilités d'évènements extrêmes multivariés sous le modèle de DE VALK [2016b] est aussi une des futures pistes envisagées. Il faudrait alors le comparer à l'estimateur proposé par DE VALK [2016a]. Cela nous permettrait de rendre compte de la dépendance existant entre les évènements. A ce sujet, citons FALK et MICHEL [2006], BACRO et collab. [2010] ou encore WADSWORTH et collab. [2017]; WADSWORTH et TAWN [2012].

Chapitre 4

Estimation des limites d'extrapolation sur des données environnementales

Sommaire

Résumé

Nous développons dans ce chapitre des outils permettant l'estimation des limites d'extrapolation à partir de jeux de données réelles. Ces estimations nécessitent de savoir le domaine d'attraction de la loi sous-jacente aux données. Les domaines d'attraction de Gumbel et de Fréchet sont traités. La Partie 4.1 débute ainsi avec la proposition d'un estimateur de l'erreur d'extrapolation relative dédié au domaine d'attraction de Gumbel. Celui-ci se base sur les équivalents obtenus Chapitre 2 combinés aux estimateurs des paramètres introduits Chapitre 3. Nous montrons sur simulations comment cet estimateur se comporte en pratique, pour différentes lois du domaine d'attraction de Gumbel. La Partie 4.2 illustre l'utilisation de cet estimateur sur deux jeux de données, le premier constitué de mesures journalières du débit du Rhône et le second correspondant à des mesures faites à Reims de vitesses instantanées de vent. Etant donné une erreur maximale admissible, nous sommes alors capables d'estimer jusqu'où il est possible d'extrapoler. Les résultats indiquent que l'extrapolation est beaucoup plus limitée pour le premier jeu de données que pour le second. La Partie 4.3 s'intéresse à un estimateur de l'erreur d'extrapolation relative dédié au domaine d'attraction de Fréchet. Sa construction y est détaillée et son comportement pratique est étudié sur simulations. Cet estimateur est alors illustré Partie 4.4 sur un troisième jeu de données, constitué de cumuls journaliers de précipitations enregistrés à Valleraugue dans les Cévennes. Enfin, la Partie 4.5 propose des perspectives de travail.

4.1 Un estimateur de l'erreur d'extrapolation dédié au domaine d'attraction de Gumbel

Nous proposons dans cette partie un estimateur de l'erreur d'extrapolation relative associée à l'approximation Exponential Tail, dédiée au domaine d'attraction de Gumbel. C'est cet estimateur qui nous servira à estimer les limites d'extrapolation dans des cas pratiques où la loi F sous-jacente appartient au domaine d'attraction de Gumbel. Pour plus de détails sur l'approximation Exponential Tail, voir le Paragraphe 1.3.3.

Rappelons que l'erreur d'extrapolation relative est définie par (cf équation (2.3) du Chapitre 2) :

$$\epsilon_{ext}(p_n) := \frac{q(p_n) - \tilde{q}(p_n)}{q(p_n)}.$$

En particulier, dans le cas de l'approximation Exponential Tail, cette erreur se réécrit comme suit (voir Paragraphe 2.1.1) :

$$\epsilon_{\mathrm{ET}}(p_n;\alpha_n) = \frac{q(p_n) - q(\alpha_n) - \sigma_n \log\left(\frac{\alpha_n}{p_n}\right)}{q(p_n)}. \tag{4.1}$$

Notre objectif est de proposer un estimateur $\hat{\epsilon}_{\mathrm{ET}}(p_n;\alpha_n)$ de (4.1).

4.1.1 Une approximation générale de l'erreur d'extrapolation

L'étude de (4.1) est l'objet du Chapitre 2, Paragraphe 2.2. Rappelons que dans ce dernier paragraphe, $\varphi(\cdot) = \mathrm{H}^{-1}(\cdot)$ (voir Paragraphe "Application to the ET approximation") avec $\mathrm{H}(\cdot) = -\log(1-\mathrm{F}(\cdot))$ la fonction de hasard cumulé, de telle sorte que les fonctions K_1 et K_2 définies Paragraphe "Theoretical framework" s'écrivent :

$$\mathrm{K}_1(s) := \frac{s\varphi'(s)}{\varphi(s)} = s\frac{(\mathrm{H}^{-1})'(s)}{\mathrm{H}^{-1}(s)}$$

et

$$\mathrm{K}_2(s) := \frac{s^2\varphi''(s)}{\varphi(s)} = s^2\frac{(\mathrm{H}^{-1})''(s)}{\mathrm{H}^{-1}(s)}.$$

Enfin, les limites des fonctions K_1 et K_2 sont respectivement notées ℓ_1 et ℓ_2 quand elles existent. Moyennant ces notations, nous avons obtenu des équivalents simples de l'erreur d'extrapolation, voir le Théorème 2 du Chapitre 2.

Une première idée consiste ainsi à s'appuyer sur ces équivalents pour proposer un estimateur de l'erreur d'extrapolation. De plus, ces équivalents sont, au vu des Figures 1-4 associées (voir Chapitre 2), particulièrement efficaces, même pour de petites tailles d'échantillon. Cependant, ils requièrent de connaître dans quel sous-domaine du domaine d'attraction de Gumbel la loi des données F appartient. Or ce n'est pas le cas en pratique, nous obligeant à trouver un autre moyen de proposer un estimateur de l'erreur relative d'extrapolation.

Pour ce faire, une autre idée est de partir de l'approximation générale donnée Lemme 3-(i) du Chapitre 2. Dans ce Lemme, et sous les hypothèses de modèle **(A1)**-**(A4)**, nous montrons que l'erreur d'extrapolation relative peut s'écrire

$$\epsilon_{\mathrm{ET}}(p_n;\alpha_n) = \Delta(n) = \delta^2(n)\int_0^1 \frac{\mathrm{K}_2(y(n)(1-\delta(n)u))}{(1-\delta(n)u)^2}\exp\left(\mathrm{K}_1(y(n))\mathrm{L}_{\theta_1}(1-\delta(n)u)(1+o(1))\right)u\,du,$$

avec (cf Page 54)

$$\begin{aligned}\delta(n) &:= (y(n)-x(n))/y(n),\\ x(n) &:= \log(1/\alpha_n),\\ y(n) &:= \log(1/p_n),\\ \mathrm{L}_\theta(t) &:= \int_1^t u^{\theta-1}\,du\end{aligned}$$

et α_n et p_n qui représentent respectivement des ordres intermédiaire et extrême (voir Paragraphe 1.3).

Ce résultat est plus général que les équivalents obtenus Théorème 2 du Chapitre 2 puisqu'il ne suppose pas d'hypothèse sur l'appartenance à un sous-domaine du domaine d'attraction de Gumbel.

Supposons par la suite, sans perdre beaucoup plus de généralité, qu'il existe $\theta_2 \in \mathbb{R}$ tel que $|\mathrm{K}_2| \in \mathrm{RV}_{\theta_2}$. Nous avons vu Page 56 que cette hypothèse est automatiquement vérifiée dès lors que $\ell_1 \neq 1$. Il vient alors :

$$\epsilon_{\mathrm{ET}}(p_n;\alpha_n) = \delta^2(n)\mathrm{K}_2(y(n))\int_{1-\delta(n)}^1 (1-\delta(n)u)^{\theta_2-2}\exp\left(\mathrm{K}_1(y(n))\mathrm{L}_{\theta_1}(1-\delta(n)u)(1+o(1))\right)u\,du.$$

Cette équation nous suggère ainsi qu'il est possible d'estimer l'erreur d'extrapolation relative si l'on parvient toutefois à proposer des estimateurs appropriés des quantités θ_1, θ_2, K_1 et K_2 au point $y(n)$.

L'estimation des quantités θ_1, θ_2 n'est pas très difficile en soi. L'estimation des fonctions K_1 et K_2 au point $y(n)$ est quant à elle bien plus problématique, dans le sens où $y(n)$ représente un quantile situé en dehors de l'échantillon (cf Paragraphe 1.3.1), ramenant le problème de l'estimation d'un quantile extrême à l'estimation des fonctions K_1 et K_2 au point $y(n)$.

Pour pallier ce problème, l'idée est d'utiliser les propriétés de conservation des équivalents des fonctions à variation régulière (cf Proposition 4), tout en se rappelant que $y(n) = x(n)/(1-\delta(n))$:

$$\varepsilon_{\mathrm{ET}}(p_n;\alpha_n) \approx \delta^2(n)K_2(x(n)) \int_0^1 \frac{(1-\delta(n)u)^{\theta_2-2}}{(1-\delta(n))^{\theta_2}} \exp\left(K_1(x(n))\frac{L_{\theta_1}(1-\delta(n)u)}{(1-\delta(n))^{\theta_1}}\right) u\,du.$$

Cette équation nous suggère maintenant d'estimer K_1 et K_2 non plus au point $y(n)$, mais au point $x(n)$, l'estimation en ce dernier point étant bien plus aisée puisque $x(n)$ représente un quantile intermédiaire, se situant dans la partie "haute" de l'échantillon. En particulier, si l'on pose $\alpha_n = k_n/n$ de telle façon que $x_n := \log(1/\alpha_n) = \log(n/k_n)$, il vient :

$$\varepsilon_{\mathrm{ET}}(p_n;\alpha_n) \approx \delta^2(n)K_2[\log(n/k_n)] \int_0^1 \frac{(1-\delta(n)u)^{\theta_2-2}}{(1-\delta(n))^{\theta_2}} \exp\left(K_1[\log(n/k_n)]\frac{L_{\theta_1}(1-\delta(n)u)}{(1-\delta(n))^{\theta_1}}\right) u\,du := \tilde{\varepsilon}_{\mathrm{ET}}(p_n;\alpha_n). \quad (4.2)$$

On propose finalement d'estimer $\varepsilon_{\mathrm{ET}}(p_n;\alpha_n)$ par

$$\hat{\varepsilon}_{\mathrm{ET}}(p_n;\alpha_n) := \delta^2(n)\hat{K}_2[\log(n/k_n)] \int_0^1 \frac{(1-\delta(n)u)^{\hat{\theta}_2-2}}{(1-\delta(n))^{\hat{\theta}_2}} \exp\left(\hat{K}_1[\log(n/k_n)]\frac{L_{\hat{\theta}_1}(1-\delta(n)u)}{(1-\delta(n))^{\hat{\theta}_1}}\right) u\,du, \quad (4.3)$$

où $\hat{\theta}_1, \hat{\theta}_2, \hat{K}_1[\log(n/k_n)]$ et $\hat{K}_2[\log(n/k_n)]$ sont des estimateurs appropriés de $\theta_1, \theta_2, K_1[\log(n/k_n)]$ et $K_2[\log(n/k_n)]$. Le paragraphe suivant traite de l'estimation de ces quantités.

4.1.2 Utilisation des estimateurs des paramètres du modèle de queue de type log-Weibull généralisé proposés Chapitre 3

Les estimateurs que nous proposons des quantités θ_1, θ_2, $K_1[\log(n/k_n)]$ et $K_2[\log(n/k_n)]$ se basent sur les estimateurs des paramètres du modèle de queue de type log-Weibull généralisé proposés Paragraphe "Inférence" du Chapitre 3. Ci-dessous, nous faisons le lien entre ces derniers et les quantités en question.

Pour ce faire, rappelons que l'étude de l'erreur d'extrapolation associée à l'approximation ET s'appuie sur l'hypothèse que la fonction K_1 est à variation régulière d'indice $\theta_1 \leq 1$ (cf Condition **(A3)**). Or, dans le cas ET,

$$K_1(s) = s\frac{(H^{-1})'(s)}{H^{-1}(s)},$$

voir (4.1.1). Par conséquent,

$$\begin{aligned} & K_1 \in \mathrm{RV}_{\theta_1} \\ \Longleftrightarrow\ & (H^{-1})'/H^{-1} \in \mathrm{RV}_{\theta_1-1} \\ \Longleftrightarrow\ & (\log H^{-1})' \in \mathrm{RV}_{\theta_1-1} \\ \Longrightarrow\ & \log H^{-1} \in \mathrm{ERV}_{\theta_1}. \end{aligned}$$

d'après DE HAAN et FERREIRA [2007, Corollaire 1.1.10]. Or, par définition de la fonction de taux de hasard cumulé H,

$$\log H^{-1}(\cdot) = \log \overline{F}^{-1}(\exp(\cdot)) = \log q(1/\exp(\cdot)) = \log U(\exp(\cdot)) = V(\cdot).$$

Autrement dit, on a supposé Paragraphe "Application to the ET approximation" du Chapitre 2 que V était à variation régulière étendue d'indice θ_1, tout comme dans le Chapitre 3 où on s'est également placé dans le cadre du modèle de queue de type log-Weibull généralisé (voir Paragraphe "Tail model"). Se faisant, il est alors possible de remarquer que θ_1 joue le rôle de θ dans le Chapitre 3. De la même façon, la fonction K_1 joue le rôle de la fonction a dans ce dernier. Or, dans le Chapitre 3, nous avons proposé Paragraphe "Inférence" des estimateurs de ces dernières quantités.

Par conséquent, on propose dans ce chapitre d'estimer $K_1[\log(n/k_n)]$ par :

$$\hat{K}_1[\log(n/k_n)] := \frac{\log X_{n-k_n,n}}{\mu_1[\log(n/k_n), \hat{\theta}^{(\mathrm{M})}_{n,-}]} M_n^{(1)},$$

qui n'est rien d'autre que l'estimateur proposé de $a(\log(n/k_n))$ dans le Chapitre 3 (cf équation (8)), où, pour $j \in \{1,2\}$, $M_n^{(j)}$ est défini par

$$M_n^{(j)} := \frac{1}{k_n} \sum_{i=0}^{k_n-1} \left[\log_2(X_{n-i,n}) - \log_2(X_{n-k_n,n})\right]^j,$$

avec $\log_2 := \log\log$, $\mu_b(\cdot,\cdot)$ est défini par

$$\mu_b(t,\zeta) := \int_0^1 \left[L_\zeta\left(1+\frac{\log(1/s)}{t}\right)\right]^b ds$$

et

$$\Psi_t(\zeta) := \frac{\mu_1^2(t,\zeta)}{\mu_2(t,\zeta)},$$

pour tout $t>0$, $b \in \mathbb{N}\setminus\{0\}$ avec $\zeta<1$. Pour ce qui est de θ_1, nous proposons l'estimateur suivant (cf équation (5)-(7)) :

$$\hat{\theta}_1 := \hat{\theta}_{1,+} + \hat{\theta}_{1,-}, \tag{4.4}$$

avec

$$\hat{\theta}_{1,+} := \frac{M_n^{(1)}}{\mu_1[\ln(n/k_n),0]}$$

et

$$\hat{\theta}_{1,-} := \Psi^{\leftarrow}_{\ln(n/k_n)}\left(\frac{[M_n^{(1)}]^2}{M_n^{(2)}}\right).$$

Ayant proposé des estimateurs des quantités θ_1 et $K_1[\log(n/k_n)]$, il reste finalement à proposer des estimateurs de θ_2 et $K_2[\log(n/k_n)]$. Leur construction se base sur la preuve du Lemme 2 du Chapitre 2, qui donne une relation entre K_2 et K_1 : pour tout $s \in \mathbb{R}$,

$$K_2(s) = K_1^2(s) + K_1(s)\left(\frac{sK_1'(s)}{K_1(s)} - 1\right).$$

En se rappelant alors que K_1' est ultimement monotone (cf Hypothèse (A4)) et en utilisant le Corollaire 1, on propose d'estimer $K_2[\log(n/k_n)]$ par :

$$\hat{K}_2[\log(n/k_n)] := \hat{K}_1^2[\log(n/k_n)] + \hat{K}_1[\log(n/k_n)](\hat{\theta}_1 - 1).$$

Par ailleurs, une conséquence du Lemme 2 est que $K_1 \in RV_{\theta_1}$, $\ell_1 \neq 1$ implique $K_2 \in RV_{\theta_2}$ avec

$$\theta_2 = \begin{cases} 2\theta_1 & \theta_1 > 0 \\ \theta_1 & \theta_1 < 0 \\ 0 & \theta_1 = 0 \text{ et } \ell_1 \neq 1 \end{cases}$$

nous suggèrant d'estimer θ_2 comme suit :

$$\hat{\theta}_2 := 2\hat{\theta}_1 \mathbb{1}\{\hat{\theta}_1 > 0\} + \hat{\theta}_1 \mathbb{1}\{\hat{\theta}_1 < 0\}.$$

Notons que cet estimateur pourra ne pas être consistant si $\ell_1 = 1$ (et donc $\theta_1 = 0$, voir la remarque sous la Définition 1). L'estimation de l'erreur d'extrapolation relative est ensuite obtenue en faisant du plug-in des estimateurs précédents de θ_1, θ_2, $K_1[\log(n/k_n)]$ et $K_2[\log(n/k_n)]$ dans l'équation (4.2).

4.1.3 Illustration sur données simulées

Nous proposons d'évaluer les performances de l'estimateur de l'erreur d'extrapolation relative donné équation (4.3) sur des données simulées. Pour ce faire, le comportement de $\hat{\varepsilon}_{ET}(p_n;\alpha_n)$ est étudié comme une fonction de n (de 100 à 10 000) à partir de N = 500 échantillons aléatoires simulés, et ce dans le cas où

$$p_n = n^{-4/3}, \alpha_n = n^{-1/3} \text{ et } \delta_n = 3/4. \tag{4.5}$$

Sur chacun des échantillons, l'estimateur en question est calculé, puis l'espérance de l'estimateur est estimée en moyennant sur les 500 réplications. Un intervalle de confiance empirique à 90% est alors obtenu à

partir des quantiles à 5 et 95% de l'échantillon. L'espérance est alors comparée avec la vraie erreur d'extrapolation relative

$$\varepsilon_{\mathrm{ET}}(p_n;\alpha_n) := (q(p_n) - \tilde{q}_{\mathrm{ET}}(p_n;\alpha_n))/q(p_n)$$

ainsi qu'avec l'équivalent théorique donné équation (4.2). La procédure est répétée pour des lois Exponentielle, Reflected Gumbel, Gamma, Weibull, Normale, Lognormale, Point Terminal Fini et Pareto. Ces lois sont décrites Table 4.1. Les résultats sont présentés Figures 4.1 et 4.2.

Avant de commenter directement ces Figures, il est intéressant de remarquer que, clairement, toutes les lois utilisées vérifient K_1 et $|K_2|$ sont à variation régulière. De plus, les choix de suites opérés (cf équation (4.5)) impliquent $\delta_\infty = 3/4$ de telle façon que le Théorème 2 (i,ii,iii)-(b) peut-être appliqué et l'erreur $\varepsilon_{\mathrm{ET}}(p_n;\alpha_n)$ tend vers zéro uniquement pour les lois appartenant à MDA_1(Gumbel).

En effet, comme prédit par le précédent Théorème, il apparaît sur les Figures 4.1 et 4.2 que l'erreur d'extrapolation relative tend vers zéro uniquement pour les lois Exponentielle, Reflected Gumbel, Gamma et Point Terminal qui sont les lois qui vérifient F $\in \mathrm{MDA}_1$(Gumbel), au vu de la Table 1 du Chapitre 2. Les autres lois considérées ne vérifient pas F $\in \mathrm{MDA}_1$(Gumbel). Elles présentent une erreur qui tend vers une constante non nulle, ceci étant en accord avec le Théorème 2 (i,ii,iii)-(b) du Chapitre 2.

Ayant vérifié que les Figures 4.1 et 4.2 étaient en accord avec la théorie développée Chapitre 2, nous nous attachons maintenant à commenter le comportement de l'équivalent de l'erreur d"extrapolation. A ce titre, les Figures 4.1 et 4.2 nous permettent de constater que la courbe de l'équivalent donné équation (4.2) (en bleu) est très proche de la vraie erreur d'extrapolation relative (en noir), et ce même pour des petites tailles d'échantillons (de l'ordre de 100). C'est le cas pour la plupart des lois exposées. La plus grande différence entre ces deux courbes s'observe dans le cas de la loi Lognormale, où la convergence de l'équivalent est très lente.

Enfin, l'estimateur proposé de l'erreur (en rouge) semble particulièrement efficace, dans le sens où il est à la fois très proche des courbes bleues et noires, et ce aussi bien pour des lois appartenant à MDA_1(Gumbel) (Exponentielle, Reflected Gumbel), que pour des lois appartenant à MDA_3(Gumbel) (Lognormale Pareto).

Bien que l'on ne dispose pas de garantie de convergence de cet estimateur - cela constitue une des perspectives de cette thèse -, ces résultats sont rassurants quant à la portée d'applicabilité de (4.3) en pratique. L'application de cet estimateur sur des cas réels est l'objet du paragraphe suivant.

TABLEAU 4.1 – Domaine d'attraction, support, H, K_1, K_2, θ_1, θ_2, ℓ_1, ℓ_2 pour plusieurs lois de la Table 1.1.

Lois	Domaine d'attraction	Support de la loi	H(x)	$K_1(x)$	$K_2(x)$	θ_1	θ_2	ℓ_1	ℓ_2
Exponentielle	Gumbel	$[0,+\infty[$	$= x$	$= 1$	$= 0$	0	0	1	0
Gamma(μ), $\mu \neq 1$	Gumbel	$[0,+\infty[$	$= x + (1-\mu)\log x + \log\Gamma(\mu) + o(1)$	~ 1	$\sim \dfrac{1-\mu}{x}$	0	-1	1	0
Logistique	Gumbel	$\mathbb{R}$	$= x + \log(1+e^{-x})$	~ 1	$\sim -xe^{-x}$	0	X	1	0
Reflected Gumbel	Gumbel	$\mathbb{R}$	$= e^x$	$= \dfrac{1}{\log x}$	$= -\dfrac{1}{\log x}$	0	0	0	0
Point terminal fini	Gumbel	$[0,1[$	$= \dfrac{x}{1-x}$	$= \dfrac{1}{x+1}$	$= -\dfrac{2x}{(x+1)^2}$	-1	-1	0	0
Point terminal fini 2	Gumbel	$[0,1[$	$= \dfrac{1}{\log x^* - \log x}$	$= \dfrac{1}{x}$	$= -\dfrac{2}{x} + \dfrac{1}{x^2}$	-1	-1	0	0
Weibull(β), $\beta \neq 1$	Gumbel	$[0,+\infty[$	$= x^\beta$	$= \dfrac{1}{\beta}$	$= \dfrac{1}{\beta}\left(\dfrac{1}{\beta} - 1\right)$	0	0	$\dfrac{1}{\beta}$	$\dfrac{1}{\beta}\left(\dfrac{1}{\beta} - 1\right)$
Normale	Gumbel	$\mathbb{R}$	$= x^2/2 + \log x + (2\pi)^{1/2} + o(1)$	$\sim \dfrac{1}{2}$	$\sim -\dfrac{1}{4}$	0	0	$\dfrac{1}{2}$	$-\dfrac{1}{4}$
Lognormale	Gumbel	$]0,+\infty[$	$= (\log x)^2/2 + \log\log x + \log(2\pi)^{1/2} + o(1)$	$\sim \sqrt{\dfrac{x}{2}}$	$\sim \dfrac{x}{2}$	1/2	1	$+\infty$	$+\infty$
Pareto(γ), $\gamma > 0$	Fréchet	$]1,+\infty[$	$= \dfrac{1}{\gamma}\log x$	$= \gamma x$	$= \gamma^2 x^2$	1	2	$+\infty$	$+\infty$

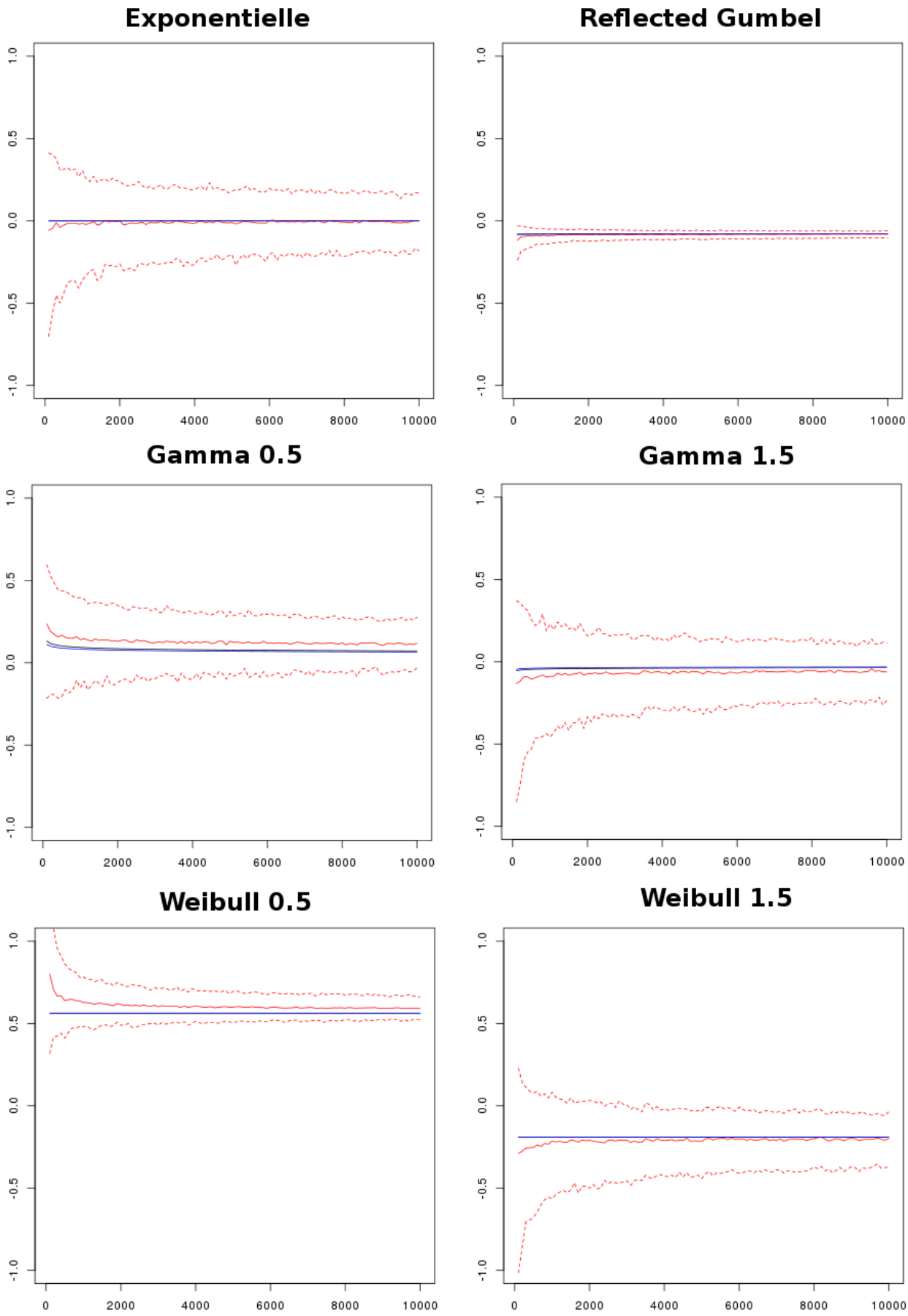

FIGURE 4.1 – L'erreur d'extrapolation relative comme une fonction de n. En noir, $\varepsilon_{\mathrm{ET}}(p_n;\alpha_n)$. En bleu, $\tilde{\varepsilon}_{\mathrm{ET}}(p_n;\alpha_n)$. En rouge, $\hat{\varepsilon}_{\mathrm{ET}}(p_n;\alpha_n)$ ainsi qu'un intervalle de confiance empirique à 90%.

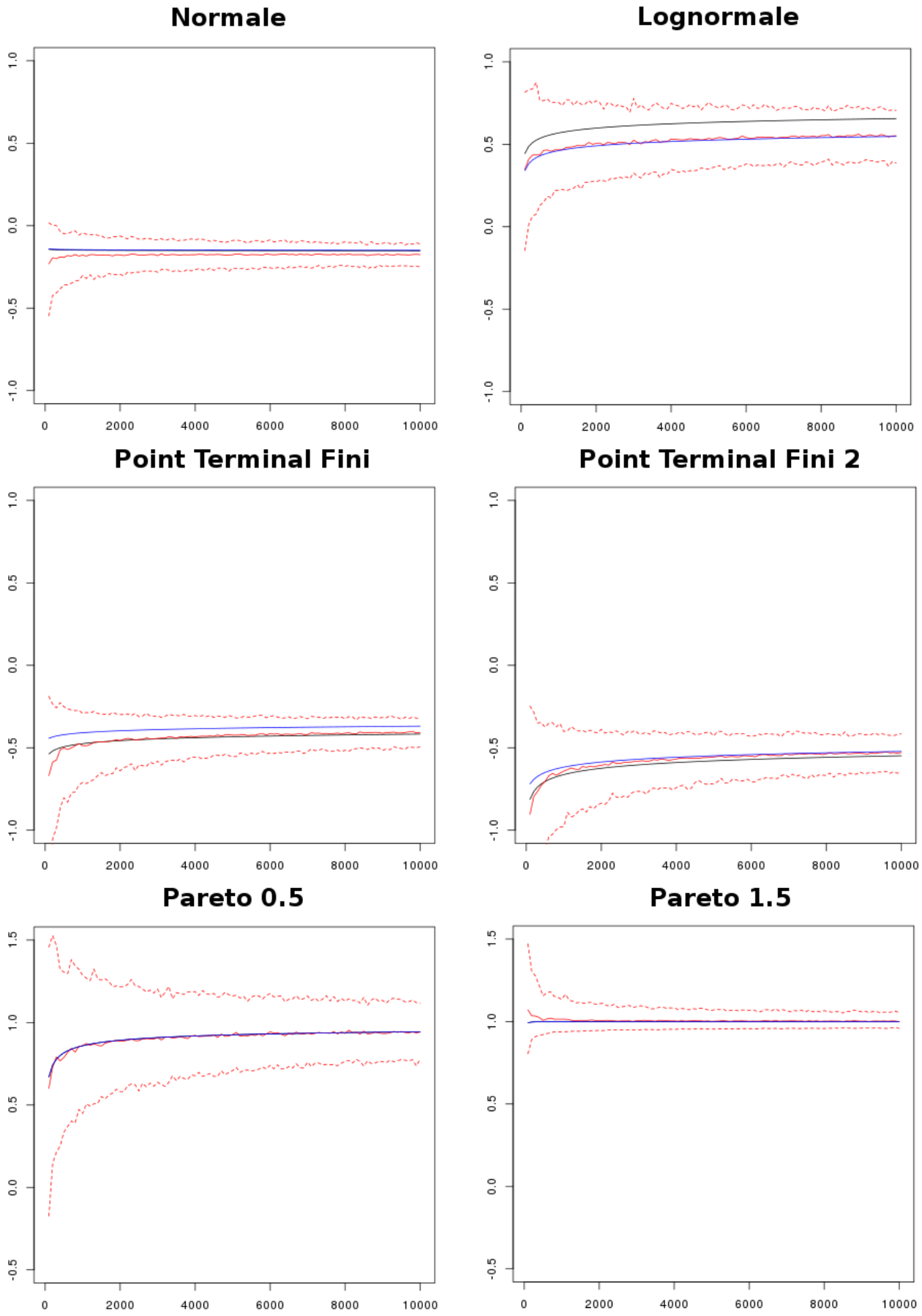

FIGURE 4.2 – L'erreur d'extrapolation relative comme une fonction de n. En noir, $\varepsilon_{\mathrm{ET}}(p_n;\alpha_n)$. En bleu, $\tilde{\varepsilon}_{\mathrm{ET}}(p_n;\alpha_n)$. En rouge, $\hat{\varepsilon}_{\mathrm{ET}}(p_n;\alpha_n)$ ainsi qu'un intervalle de confiance empirique à 90%.

4.2 Applications à des séries de mesures de variables environnementales

Nous mettons en pratique l'estimateur de l'erreur d'extrapolation relative présenté ci-dessus sur deux jeux de données réelles. Le logiciel utilisé est R, librairie evd. L'objectif est de répondre aux deux questions suivantes : étant donné une série temporelle de mesures et une erreur ϵ^*, jusqu'où est-il possible d'extrapoler ? Autrement dit, quelle est la période de retour maximale T^* (voir le Paragraphe 1.3.5) au delà de laquelle l'estimation du niveau de retour associé à cette période T^* induit une erreur d'extrapolation relative plus grande que ϵ^* ? La deuxième question est la question duale à la première : Quelle erreur d'approximation relative commet-on en estimant un niveau de retour de période de retour T_0 ?

A cette fin, nous commençons par étudier le premier jeu de données, constitué de mesures journalières du débit moyen du Rhône. Nous montrons que la loi sous-jacente est très proche d'une loi Lognormale et qu'il est donc sensé d'utiliser les résultats théoriques de l'erreur d'approximation relative dans le cas $F \in DA(Gumbel)$. Suite à cela, nous proposons des estimations des paramètres du modèle basées sur le jeu de données en question. Utilisant le lien entre quantile et niveau de retour, nous montrons que l'extrapolation est particulièrement limitée dans ce cas.

Puis un deuxième jeu de données, constitué de mesures de vitesses maximales journalières de vents, est introduit. Suivant la même procédure, nous concluons que les limites d'extrapolation sont bien moins restrictives dans ce cas.

4.2.1 Application à des données de débits du Rhône

Le jeu de données étudié est celui brièvement présenté Paragraphe "Illustration on real data" du Chapitre 3. Il consiste en des mesures faites à Malause (série CASTOR Q5382 DI1 J1) du débit moyen journalier du Rhône de 1915 à 1966, complétées par des mesures faites à Lamagistère (série CASTOR Q1396 DI1 J1) de 1967 à 2013, pour un total de 36160 mesures. Les données manquantes ont été reconstituées à partir de la série de la Banque Hydro O6140010.

La Figure 4.7 représente un Boxplot des données en fonction des différents mois de l'année (de Janvier à Décembre). Cette Figure permet de mettre en évidence une saisonnalité des données due à la succession des périodes sèches et pluvieuses. Ainsi, le débit du Rhône est en moyenne plus élevé en Mars qu'en Août. Cette non-stationnarité peut poser problème dans le cadre d'une étude statistique basée sur l'hypothèse de variables aléatoires (les débits) indépendantes et identiquement distribuées. Pour pallier ce problème, il est courant de considérer uniquement la saison pluvieuse, c'est à dire les mois de Décembre à Mai, ramenant le nombre de mesures à $n = 18043$. C'est également ce qui est préconisé par les études EDF sur ce jeu de données.

Un autre facteur à prendre en compte quand on a affaire à des données réelles est celui de la dépendance. Encore une fois, les modèles développés dans les chapitres précédents se basent sur une hypothèse d'indépendance des variables aléatoires sous-jacentes. Cependant, au vu de la Figure 4.8, qui représente un graphe d'auto-corrélation partielle de la série des débits en fonction des jours d'espacement des mesures, il semblerait que les débits présentent une dépendance à trois jours. Il convient par la suite de prendre en compte cette dépendance. Pour ce faire, il est courant de généraliser l'approche POT à des clusters de dépassements, au lieu de considérer tous les excès au delà d'un seuil u (cf Paragraphe 1.3.5). Nous commençons ainsi par ajuster une loi de Pareto Généralisée (GPD) à la série de débits décrite ci-dessus, en considérant un seuil $u = 2400 m^3/s$ et des clusters de dépassements de paramètre $r = 3$ jours, un cluster de dépassements étant défini comme un bloc ayant au moins un dépassement et dans lequel deux dépassements consécutifs sont toujours distants de moins de r jours. Le choix du paramètre $r = 3$ est motivé par la Figure 4.8. Le choix du seuil fait quant à lui suite à des études antérieures EDF, basées sur la modélisation des excès par un processus de Poisson approprié. Les résultats numériques liés à l'application de la méthode POT sont représentés Figure 4.3 et la qualité d'ajustement est illustrée Figures 4.9 et 4.10. Nous commençons par brièvement commenter ces dernières.

La Figure 4.9 propose une comparaison de la densité théorique de la GPD ajustée (ligne pleine) et d'un estimateur à noyau de la densité des données (en pointillés), en fonction du débit. Les deux courbes étant presque superposées, cette figure témoigne de la très grande qualité d'ajustement de la GPD aux excès, et ce même pour les valeurs les plus extrêmes du jeu de données. Cette très grande qualité d'ajustement est confirmée par la Figure 4.10 qui oppose les quantiles théoriques de la loi GPD ajustée (abcisse) aux quantiles empiriques du jeu de données (à savoir les excès, en ordonnées). Encore une fois, même si le point de vue est différent, l'ajustement est excellent, même pour les valeurs les plus extrêmes du jeu de données.

Nous nous intéressons maintenant aux résultats numériques liés à l'ajustement de la GPD aux excès, compilés Figure 4.3. Cette figure nous indique tout d'abord qu'il y a 251 excès au dessus du seuil

```
Call: fpot(x = X$Debit, threshold = 2400, cmax = TRUE, r = 3)
Deviance: 1871.069

Threshold: 2400
Number Above: 251
Proportion Above: 0.0139

Clustering Interval: 3
Number of Clusters: 120
Extremal Index: 0.4781

Estimates
    scale      shape
8.944e+02  5.332e-03

Standard Errors
    scale      shape
119.95918    0.09727
```

FIGURE 4.3 – Données du Rhône : Ajustement d'une loi de Pareto Généralisée aux excès. Estimation des paramètres par maximum de vraisemblance.

$u = 2400 m^3/s$, pour seulement 120 clusters de dépassements, signifiant que ces derniers ont une taille moyenne de deux jours. Ceci est cohérent avec l'estimation de l'indice extrêmal (environ 1/2) dont l'inverse rappelons le peut-être interprété comme la taille moyenne des clusters de dépassements (cf Paragraphe 1.3.5). La Figure 4.3 donne également l'estimation des paramètres de la GPD obtenus par maximum de vraisemblance (voir Paragraphe 1.3.3). En particulier, le paramètre de forme γ (ou indice des valeurs extrêmes), est estimé à $5e-03$, soit une valeur très proche de zéro. Cette estimation de l'indice des valeurs extrêmes et l'écart-type associé, d'environ 0.1, suggère qu'on a affaire à une loi des débits qui appartient au domaine d'attraction de Gumbel (cf Paragraphe 1.2.6).

Les Figures 4.12 (haut) et 4.11 s'attachent à confirmer cette intuition. La Figure 4.12 (haut) représente la valeur du paramètre de forme de la GPD (estimé par maximum de vraisemblance) en fonction du seuil lorsqu'une dépendance à trois jours est prise en compte. Cette figure nous permet d'observer que la fonction constante égale à zéro en rouge recoupe tous les intervalles de confiance à 95% associés à l'estimation de l'indice des valeurs extrêmes, et ce pour une vaste plage de valeurs du seuil (de 1000 à 4000 m^3/s), suggérant à nouveau que la loi des débits appartient au domaine d'attraction de Gumbel.

La Figure 4.11 juxtapose un graphe quantile-quantile permettant de tester l'adéquation des débits à une loi Lognormale (en haut) et un graphe de densité représentant la superposition de la densité d'une loi Lognormale à l'histogramme des données (en bas). L'excellente qualité d'ajustement nous renseigne sur le fait que les débits en question semblent issus d'une loi appartenant à $\text{MDA}_3(\text{Gumbel})$ (cf Page 56), très proche d'une loi Lognormale, loi qui rappelons le appartient au domaine d'attraction de Gumbel et dont la fonction K_1 associée est à variation régulière, avec $\theta_1 = 1/2$ (voir Table 4.1).

Ainsi, il semble naturel de supposer que la loi des débits appartient au domaine d'attraction de Gumbel et que sa fonction K_1 associée est à variation régulière. Au vu des remarques précédentes, il fait sens de vouloir estimer l'erreur d'extrapolation relative en utilisant l'estimateur dédié au domaine d'attraction de Gumbel donné équation (4.3). C'est l'objet des paragraphes qui suivent.

Estimation de θ_1

Nous commençons par nous intéresser à l'estimation de θ_1, qui nous renseigne sur le type de sous-domaine d'attraction auquel appartient la loi.

La Figure 4.12 (bas) représente les estimations obtenues de θ_1 par (4.4) (en noir) et par l'estimateur proposé par DE VALK et CAI [2018] (en rouge) en fonction du seuil en m^3/s. Elle nous permet de voir que ces deux estimateurs présentent des performances très similaires, les courbes étant quasiment superposées, l'estimateur proposé par DE VALK et CAI [2018] estimant des valeurs de θ_1 très légèrement supérieures aux notres. Ensuite, cette figure n'est pas sans rappeler la Figure 4.12 (haut) pour l'indice des valeurs extrêmes γ, d'où notre volonté de juxtaposer les deux. En effet, on retrouve dans la Figure 4.12 (bas) un comportement

similaire à l'estimation par maximum de vraisemblance du paramètre γ, et ce quel que soit l'estimateur proposé : L'estimation de θ_1 est stable autour de $1/2$ pour une plage de débits allant de 1000 à $2500 m^3/s$, avant de chuter brutalement pour des seuils supérieurs à $2600 m^3/s$. Ce décrochage se retrouve également dans la Figure 4.12 (haut) pour γ, bien que celui-ci ait lieu pour une plus grande valeur du seuil ($\approx 3100 m^3/s$). Par ailleurs, si l'on revient à l'estimation de θ_1, l'intervalle de confiance associé est de plus en plus large avec le seuil. Cela est cohérent avec la Figure 4.12 (haut) et avec le fait que, plus le seuil est haut, moins il y a d'excès disponibles pour estimer les paramètres en question.

D'autre part, il est intéressant de voir que la valeur obtenue pour $u = 2400 m^3/s$ se situe dans la "zone de stabilité" de $\hat{\theta}_1$. Pour un tel seuil, θ_1 est estimé à environ $1/3$. Cette estimation légèrement positive de θ_1 nous conforte dans le fait que la distribution des débits est proche d'une loi lognormale (qui, rappelons le, presénte un θ_1 égal à $1/2$, voir Table 4.1).

Par la suite nous nous focalisons sur l'estimation de l'erreur d'extrapolation relative.

Estimation de l'erreur d'extrapolation

Nous nous proposons d'estimer l'erreur d'extrapolation associée à un niveau de retour à T = 1000 années. La première étape consiste à écrire l'erreur d'extrapolation comme une fonction de T dans (4.3), et non plus comme une fonction de $\delta_n := 1 - \log(1/\alpha_n)/\log(1/p_n)$.

Dans le cas de données présentant de la dépendance, rappelons que le Paragraphe 1.3.5 nous donne une relation entre p_n et T :

$$p_n = 1 - \left(1 - \frac{1}{\mathrm{T}}\right)^{1/(n\omega)} \approx \frac{1}{m\omega\mathrm{T}},$$

où m représente la taille des blocs et ω est l'indice extrémal. Rappelons en effet que la période de retour est définie à travers la notion de niveau de retour, ce dernier étant le quantile extrême d'ordre $1-q$ de la loi du maximum $\mathrm{X}_{m,m}$ sur une période donnée. Par conséquent ici, $m = 182.5$ jours (on considère six mois de l'année).

En remplaçant alors $\delta(n)$ par son expression dans (4.3), puis p_n par son équivalent en T, il vient :

$$\begin{aligned}\hat{\varepsilon}_{\mathrm{ET}}(p_n;\alpha_n) &= \left(1 - \frac{\log(n/k_n)}{\log(m\hat{\omega}\mathrm{T})}\right)^2 \left(\frac{\log(n/k_n)}{\log(m\hat{\omega}\mathrm{T})}\right)^{-\hat{\theta}_2} \hat{\mathrm{K}}_2[\log(n/k_n)] \int_0^1 \left[\left(1 - \left(1 - \frac{\log(n/k_n)}{\log(m\hat{\omega}\mathrm{T})}\right)u\right)^{\hat{\theta}_2 - 2}\right] \\ &\times \exp\left(\hat{\mathrm{K}}_1[\log(n/k_n)] \frac{\mathrm{L}_{\hat{\theta}_1}\left(1 - \left(1 - \frac{\log(n/k_n)}{\log(m\hat{\omega}\mathrm{T})}\right)u\right)}{\left(\frac{\log(n/k_n)}{\log(m\hat{\omega}\mathrm{T})}\right)^{\hat{\theta}_1}}\right) u\,du, \end{aligned} \tag{4.6}$$

qui constitue une fonction de T.

La Figure 4.13 représente l'estimation de l'erreur d'extrapolation relative en fonction du seuil ($k_n = \sum_{i=1}^n \mathbb{1}\{\mathrm{X}_i \geq seuil\}$), lorsque T = 1000 ans. A l'image de l'estimation de θ_1, l'estimation de l'erreur est particulièrement stable pour une plage de débits allant de 1000 à $2600 m^3/s$. La légère décroissance vers zéro que l'on peut observer s'explique par le fait que l'approximation GPD devient de plus en plus précise avec l'augmentation du seuil, le résultat théorique étant un résultat asymptotique. Par ailleurs, là aussi il est possible d'observer un décrochage de l'estimation due au manque de données au dessus d'un seuil de $2600 m^3/s$. Enfin, pour une valeur du seuil de $2400 m^3/s$, l'erreur est ponctuellement estimée à environ 22%.

En comparaison, l'estimation du niveau de retour à 1000 ans que nous avions obtenus au Chapitre 3 (cf Page 100), basée sur un estimateur des moments et une hypothèse d'indépendance, était d'environ $10\,000 m^3/s$. Une erreur de 22% d'une telle quantité n'est clairement pas négligeable. Le résultat est toutefois à relativiser au vu du Théorème 2 du Chapitre 2. En effet, il semblerait que la loi des débits $\mathrm{F} \in \mathrm{MDA}_3(\mathrm{Gumbel})$, ce qui constitue le pire cas possible en terme d'erreur d'extrapolation. En particulier, pour d'autres aléas climatiques, on peut s'attendre à obtenir une estimation de l'erreur d'extrapolation plus faible que celle obtenue ici.

Enfin pour terminer, il est intéressant de répondre à la question duale qui est de savoir, moyennant une erreur ϵ^* donnée, quelle est la période de retour jusqu'où il est possible d'extrapoler. La Figure 4.14 représente $\hat{\epsilon}_{app_n}$ qui varie cette fois-ci comme une fonction de la période de retour T, où le seuil a été fixé à $2400 m^3/s$. Avant de commenter cette figure, précisons qu'il convient de l'étudier en comparaison avec la Figure 4.13. En effet, il est possible de remarquer que pour T = 1000 ans, on retrouve bien une erreur d'estimation d'environ 22%, comme c'était le cas sur la Figure 4.13 pour un seuil de $2400 m^3/s$. De plus, le seuil ayant été fixé sur la Figure 4.14, l'apparence lisse de la courbe s'explique par le fait que, pour chaque période de retour T, on considère toujours les mêmes excès, à savoir les débits supérieurs à $2400 m^3/s$.

Ceci ayant été souligné, nous nous intéressons par la suite au potentiel pratique de la Figure 4.14. Cette Figure permet tout d'abord de rendre compte du fait que l'erreur d'extrapolation est une fonction croissante non linéaire de la période de retour. En particulier, l'estimation faite de l'erreur d'extrapolation est ici concave, de telle sorte qu'un accroissement de la période de retour T a plus d'impact sur l'erreur d'extrapolation pour des valeurs faibles de T qu'il n'en a pour de grandes valeurs. D'autre part, cette Figure nous permet de répondre à la question posée : étant donnée une erreur maximale admissible ϵ^*, il est possible de savoir jusqu'où il est possible d'extrapoler. Bien qu'au vu de l'équation (4.3) il semble compliqué d'obtenir une fonction inverse explicite, la Figure 4.14 nous présente une inverse numérique monotone qui nous permet de répondre au problème posé. Par exemple, pour $\epsilon^* = 20\%$, ce graphe nous indique qu'il n'est pas possible d'extrapoler au delà de 500 ans, sous peine de dépasser ladite erreur.

Enfin il est important de noter que l'estimation faite de l'erreur d'extrapolation dépend grandement du seuil choisi. Ainsi, quand le choix d'un seuil $u = 2400 m^3/s$ mène à une erreur estimée de 22% pour une extrapolation à 1000 ans, le choix d'un autre seuil, par exemple $u = 2000 m^3/s$ mène à une erreur estimée de 11% environ pour la même période de retour. Remarquons que dans ce dernier cas, non seulement l'erreur estimée est plus faible, mais le gain est double puisque, le seuil étant plus faible, plus de données sont disponibles pour quantifier l'erreur d'estimation. Choisir un seuil optimal est ainsi une des perspectives de ce travail. Cela est discuté dans la partie Perspectives, voir Paragraphe 4.5.

4.2.2 Application à des mesures de vitesses instantanées de vents

Le jeu de données que nous étudions ici consiste en 22 218 mesures de vitesses maximales journalières en mètre par seconde de vents relevées à Reims par Météo-France du 01/01/1949 au 30/04/2011. Pour des raisons de stationnarité, le capteur servant à faire les mesures ayant été changé en 1981, nous considérons par la suite uniquement les mesures faites après 1981. Enfin, pour des raisons de saisonnalité, seuls les mois d'Octobre à Mars sont retenus, menant à $n = 5371$ mesures.

Nous commençons par ajuster ajuster une loi de Pareto Généralisée aux plus grandes valeurs de la série des vitesses de vents ainsi obtenue. Pour ce qui est du seuil, nous choisissons $u = 26 m/s$. Ce choix est justifié par la suite. La Figure 4.4 résume les résultats numériques liés à l'ajustement des excès au-delà du seuil précédent par une GPD. Elle nous permet tout d'abord de dénombrer 43 excès au dessus du seuil $u = 26 m/s$, soit moins de 1% du jeu de données. Ensuite, l'indice des valeurs extrêmes est estimé à 0.07, soit une valeur très proche de zéro. L'écart-type associé à ce dernier ne permet pas d'exclure l'appartenance de la loi sous-jacente F au domaine d'attraction de Gumbel.

```
Call: fpot(x = X$Vent, threshold = 26)
Deviance: 182.0964

Threshold: 26
Number Above: 43
Proportion Above: 0.008

Estimates
  scale    shape
2.84834  0.07066

Standard Errors
 scale   shape
0.6779  0.1830
```

FIGURE 4.4 – Mesures de vents : Ajustement d'une loi de Pareto Généralisée aux excès. Estimation des paramètres par maximum de vraisemblance.

Les Figures 4.15 et 4.16 représentent respectivement un graphe de densité et un graphe des quantiles. Sur la première, nous avons représenté une estimation à noyau de la densité (en pointillés) à laquelle nous avons superposé la densité de la loi GPD de paramètres donnés Figure 4.4 (en trait plein). La proximité de ces deux courbes nous permet de juger de la très bonne qualité de l'ajustement fait, et ce même pour les plus grandes valeurs du jeu de données. Cette qualité d'ajustement est confirmée par la Figure 4.16, sur laquelle nous avons confronté les observations ordonnées (verticalement) aux quantiles théoriques de cette même loi GPD (horizontalement). Ici, la qualité de l'ajustement est mesuré par la proximité des points à la

bissectrice, en trait plein sur le graphe.

Avant de s'intéresser à l'estimation des paramètres du modèle, nous souhaitons vérifier graphiquement que nous sommes bien confrontés à une loi du domaine d'attraction de Gumbel, afin de justifier l'utilisation de l'approximation ET. La Figure 4.18 (haut) représente l'estimation de l'indice des valeurs extrêmes par maximum de vraisemblance en fonction du seuil ainsi qu'un intervalle de confiance à 95%. Sur cette figure, on s'aperçoit que la ligne rouge correspondant à $\gamma = 0$ recoupe tous les intervalles de confiance associés aux différentes estimations de l'indice des valeurs extrêmes, et ce pour des seuils allant de 19 à $28 m/s$. On ne peut ainsi pas exclure que la loi des données appartient au domaine d'attraction de Gumbel.

La Figure 4.17 va plus loin encore. Sur cette dernière, on a juxtaposé un graphe des quantiles (en haut) superposant quantiles empiriques et quantiles théoriques d'une loi Gamma de paramètres appropriés avec un graphe de densité (en bas) qui superpose un histogramme des données à la densité théorique d'une loi Gamma de mêmes paramètres. Au vu de ces deux graphes, on s'aperçoit que les vitesses de vent sont presque parfaitement issues d'une loi Gamma, et donc la loi sous-jacente appartient au domaine d'attraction de Gumbel. De plus, la loi gamma appartient au modèle des lois à queue de type Weibull généralisées, justifiant ainsi l'utilisation des estimateurs présentés au Paragraphe 4.1. L'estimation d'un de ces estimateurs, θ_1, est l'objet du paragraphe qui suit.

Estimation de θ_1

Là encore, nous commençons par estimer θ_1. Au vu de l'ajustement d'une loi Gamma aux données (voir Figure 4.17), on s'attend à obtenir une estimation de θ_1 proche de zéro (cf Table 4.1).

La Figure 4.18 (bas) représente les estimations obtenues de θ_1 par (4.4) (en noir) et par l'estimateur proposé par DE VALK et CAI [2018] (en rouge) en fonction du seuil en m/s. Sur cette dernière, on distingue que les deux estimateurs précédents proposent des estimations équivalentes de θ_1, l'estimateur de DE VALK et CAI [2018] estimant des valeurs très légèrement plus élevées. De manière absolue maintenant, on peut voir que l'estimation de θ_1 est stable autour d'environ 1/2 pour une plage de vents allant de 22 à $26 m/s$. En particulier, pour un seuil de $26 m/s$, $\hat{\theta}_1 \simeq 0.4$, ce qui représente une valeur légèrement plus grande que celle à laquelle on pouvait s'attendre. Cependant, au vu de l'intervalle de confiance, la valeur attendue de zéro n'est pas à exclure. Enfin, ce graphe est très semblable à la Figure 4.18 (haut) pour l'indice des valeurs extrêmes. Notons qu'on observe les mêmes décrochages des estimateurs pour des valeurs trop élevées du seuil. C'est d'ailleurs ces considérations qui ont motivé le choix du seuil : $26 m/s$ correspond au plus grand seuil appartenant à la zone de stabilité.

Par la suite nous nous focalisons sur l'estimation de l'erreur d'extrapolation relative.

Estimation de l'erreur d'extrapolation

Nous nous proposons d'estimer l'erreur d'extrapolation associée à un niveau de retour à T = 1000 années, à l'aide de l'équation (4.6).

La Figure 4.19 représente deux estimations de l'erreur d'extrapolation relative en fonction du seuil, lorsque $p \approx 5.48 \times 10^{-6}$. La courbe en noir correspond à l'estimateur donné équation (4.3) alors que la courbe rouge correspond à de même estimateur dans lequel θ_2 est estimé à -1. Ce deuxième choix vient du fait que, dans le cas d'une loi Gamma, l'estimateur de θ_2 donné au Paragraphe 4.1.2 n'est pas consistant (nous y revenons dans les perspectives). On remarque sur cette figure que l'estimation de l'erreur d'extrapolation semble légèrement négative, en comparaison avec l'estimation faite dans le cas des données de débits du Rhône (voir Figure 4.13), signifiant que $\tilde{q}_{\mathrm{ET}}(p_n;\alpha_n)$ est légèrement plus grand que le quantile $q(p_n)$. En particulier, pour un seuil de $26 m/s$, l'estimation de l'erreur d'extrapolation est de -0.005. Cela est en accord avec la théorie développée au Chapitre 2, selon laquelle les limites d'extrapolation sont plus fortes dans le cas de données issues d'une loi Lognormale (appartenant à MDA_3(Gumbel)) que dans le cas de données issues d'une loi Gamma (appartenant à MDA_1(Gumbel)).

La Figure 4.20 confirme cette conclusion. Cette dernière représente $\hat{\epsilon}_{app_n}$ qui varie cette fois-ci comme une fonction de la période de retour T, où le seuil a été fixé à $26 m/s$, nous permettant ainsi de répondre à la question de savoir, moyennant une erreur ϵ^* donnée, quelle est la période de retour jusqu'où il est possible d'extrapoler. Evidemment, pour T = 1000 ans, on retrouve le fait que l'erreur d'estimation est d'environ de -0.005, comme c'était le cas sur la Figure 4.19 pour un seuil de $26 m/s$. Rappelons également que, sur ce graphe, l'estimation faite de l'erreur d'extrapolation dépend du seuil choisi, bien qu'une certaine stabilité ait été observée sur la Figure 4.19. Ainsi, deux choix de seuil ne mèneront pas exactement à la même figure. De plus, soulignons une fois de plus l'absence d'intervalles de confiance. Cependant, cette figure nous permet d'apprécier globalement les limites d'extrapolation associées à ce jeu de données. Ici, elle nous permet de voir que l'erreur d'extrapolation est très faible, et ce même pour des extrapolation à 10 000 ans, suggérant ainsi que le seul facteur limitant en pratique consiste en l'erreur d'estimation.

4.3 Un estimateur de l'erreur d'extrapolation dédié au domaine d'attraction de Fréchet

Dans la Partie 4.1, nous nous sommes intéressés à l'erreur d'extrapolation relative associée à l'approximation ET, $\varepsilon_{ET}(p_n;\alpha_n)$, et par conséquent dédiée au domaine d'attraction de Gumbel. Nous avons ainsi pu, grâce à l'étude théorique faite Chapitre 2, proposer une estimation de cette erreur. Dans cette partie, l'idée est de proposer un estimateur de $\varepsilon_W(p_n;\alpha_n)$, l'erreur d'extrapolation relative associée à l'approximation Weissman.

4.3.1 Utilisation d'un équivalent général de l'erreur d'extrapolation

Rappelons que l'erreur d'extrapolation relative associée à l'approximation Weissman est définie par (cf équation (10) du Chapitre 2) :

$$\varepsilon_W(p_n;\alpha_n) := (q(p_n) - \tilde{q}_W(p_n;\alpha_n))/q(p_n) \tag{4.7}$$

avec

$$\tilde{q}_W(p_n;\alpha_n) := q(\alpha_n)\left(\frac{\alpha_n}{p_n}\right)^{\gamma}.$$

Notre objectif est de proposer un estimateur $\hat{\varepsilon}_W(p_n;\alpha_n)$ de (4.7). Là aussi, l'idée est de se baser sur un équivalent de $\varepsilon_W(p_n;\alpha_n)$. Celui-ci est donné par le Théorème 3 du Chapitre 2. Ce dernier nous indique que, lorsque $n \to \infty$,

$$\varepsilon_W(p_n;\alpha_n) \sim -\frac{\delta_\infty}{1-\delta_\infty}\log(1/\alpha_n)\eta(1/\alpha_n) := \tilde{\varepsilon}_W(p_n;\alpha_n). \tag{4.8}$$

En posant alors $\alpha_n = k_n/n$, il vient :

$$\varepsilon_W(p_n;\alpha_n) \sim -\frac{\delta_\infty}{1-\delta_\infty}\log(n/k_n)\eta(n/k_n).$$

Il est donc nécessaire d'estimer la fonction auxliaire η au point n/k_n. Soient $X_1,...,X_n$ une suite de variables aléatoires iid de fonction de répartition F $\in$DA(Fréchet) et $X_{1,n},...,X_{n,n}$ leurs statistiques d'ordre associées. Considérons les statistiques suivantes :

$$Z_j := j\left(\log X_{n-j+1,n} - \log X_{n-j,n}\right),$$

avec $1 \le j \le k_n < n$. Beirlant et collab. [2002] proposent d'estimer $\eta(n/k_n)$ par :

$$\hat{\eta}(n/k_n) := \frac{(1-\hat{\rho})^2(1-2\hat{\rho})}{\hat{\rho}^2}\frac{1}{k_n}\sum_{j=1}^{k_n}\left(\left(\frac{j}{k_n+1}\right)^{-\hat{\rho}} - \frac{1}{1-\hat{\rho}}\right)Z_j, \tag{4.9}$$

où $\hat{\rho}$ est un estimateur approprié du paramètre du second-ordre ρ (cf Paragraphe 1.2.3). Gomes et collab. [2002] proposent l'estimateur suivant :

$$\hat{\rho}_G := -\frac{2(3s_n-2)+\sqrt{3s_n-2}}{3-4s_n}$$

à condition que $2/3 \le s_n < 3/4$, où

$$\begin{aligned} s_n &:= \frac{9\Gamma^2(2)}{2\Gamma(4)}\frac{r_n^{(4)}}{\left[r_n^{(3)}\right]^2}, \\ r_n^{(\alpha)} &:= \frac{m_n^{(\alpha)} - \Gamma(\alpha+1)\left[m_n^{(1)}\right]^{\alpha}}{m_n^{(2)} - 2\left[m_n^{(1)}\right]^2}, \\ m_n^{(\alpha)} &:= \frac{1}{k_n}\sum_{i=1}^{k_n}\left[\log X_{n-i+1,n} - \log X_{n-k_n,n}\right]^{\alpha}, \end{aligned}$$

avec $\alpha > 0$, Γ la fonction gamma et $m_n^{(\alpha)}$ les statistiques appelées "moments" dans l'estimateur des quantiles extrêmes proposé par Dekkers et collab. [1989] (voir Paragraphe 1.3.3). Nous proposons finalement d'estimer ρ par

$$\hat{\rho} := \begin{cases} -\dfrac{2(3s_n-2)+\sqrt{3s_n-2}}{3-4s_n} & \text{si} \quad 2/3 \le s_n < 3/4 \\ -1 & \text{sinon.} \end{cases} \tag{4.10}$$

Finalement, nous estimons $\varepsilon_W(p_n;\alpha_n)$ en remplaçant $\eta(1/\alpha_n)$ dans (4.8) par (4.9) :

$$\hat{\varepsilon}_W(p_n;\alpha_n) := -\frac{\delta_\infty}{1-\delta_\infty}\log(n/k_n)\frac{(1-\hat{\rho})^2(1-2\hat{\rho})}{\hat{\rho}^2}\frac{1}{k_n}\sum_{j=1}^{k_n}\left(\left(\frac{j}{k_n+1}\right)^{-\hat{\rho}}-\frac{1}{1-\hat{\rho}}\right)Z_j, \tag{4.11}$$

avec $\hat{\rho}$ donné par (4.10).

Le comportement asymptotique de $\hat{\eta}(n/k_n)$ est étudié dans BEIRLANT et collab. [2002] alors que celui de $\hat{\rho}_G$ est étudié dans GOMES et collab. [2002]. D'autres choix d'estimateurs de ρ sont possibles, voir Paragraphe 1.2.3.

4.3.2 Illustration sur données simulées

Dans ce paragraphe, les performances de l'estimateur (4.11) sont évaluées sur des données simulées. Pour ce faire, le comportement de $\hat{\varepsilon}_W(p_n;\alpha_n)$ est étudié comme une fonction de n (de 100 à 10 000) à partir de N = 500 échantillons aléatoires simulés, et ce dans le cas où

$$p_n = n^{-4/3}, \alpha_n = n^{-1/3} \text{ et } \delta_n = 3/4. \tag{4.12}$$

Sur chacun des échantillons, l'estimateur en question est calculé, puis l'espérance de l'estimateur est estimée en moyennant sur les 500 réplications. Un intervalle de confiance empirique à 90% est alors obtenu à partir des quantiles à 5 et 95% de l'échantillon. L'espérance est alors comparée avec la vraie erreur d'extrapolation relative, (4.7), ainsi qu'avec l'équivalent théorique donné équation (4.8). La procédure est répétée pour des lois Burr, GPD et log-Logistique. Ces lois sont décrites Table 4.2. Les résultats sont présentés Figure 4.5.

La Figure 4.5 nous permet d'observer dans un premier temps que toutes les erreurs d'extrapolation relatives tendent vers zéro quand n tend vers l'infini, et ce quelle que soit la loi considérée. Cela est corroboré par le Théorème 3 (i) du Chapitre 2. En effet, au vu des suites considérées (cf (4.12)), ce dernier nous indique que, pour toute loi du domaine d'attraction de Fréchet, $\varepsilon_W(p_n;\alpha_n) \to 0$ lorsque n tend vers l'infini.

Dans un deuxième temps, la Figure 4.5 est également intéressante pour la comparaison qu'elle offre entre l'équivalent donné équation (4.8) et la vraie erreur d'extrapolation relative $\varepsilon_W(p_n;\alpha_n)$. On peut observer que les courbes associées à ces quantités (respectivement en bleu et en noir) sont très proches, parfois même presque superposées, et ce même pour des tailles d'échantillon modestes (de l'ordre de 100).

Enfin, l'estimateur proposé de l'erreur (cf (4.11)) semble particulièrement efficace en terme de biais, dans la mesure où la courbe associée (en rouge) est très proche des courbes bleue et noire. Cependant, sa variance est particulièrement grande dans certaines situations. Pour s'en apercevoir, il suffit de regarder l'intervalle de confiance associé (en pointillés rouge) ainsi que l'échelle en ordonnée du graphe. Il est par exemple possible de constater que l'intervalle de confiance associé à la loi GPD est particulièrement large, même quand n est grand. D'un autre côté, l'intervalle de confiance associé à la loi log-Logistique ($\alpha = 3$, $\beta = 10$) est relativement étroit. De manière générale, il semble que la largeur de l'intervalle de confiance soit liée à la lourdeur de la queue. Ainsi, pour une loi GPD ($\mu = 5$, $\sigma = 1$, $\gamma = 2$), $\gamma = 2$ alors que pour une loi log-Logistique ($\alpha = 3$, $\beta = 10$), $\gamma = 0.1$. Notons que pour la loir de Burr ($k = 1/2$, $c = 2$), $\gamma = 1$ et $\gamma = 1/2$ pour la loi log-Logistique ($\alpha = 3$, $\beta = 2$).

L'application pratique de l'estimateur proposé est l'objet du paragraphe suivant.

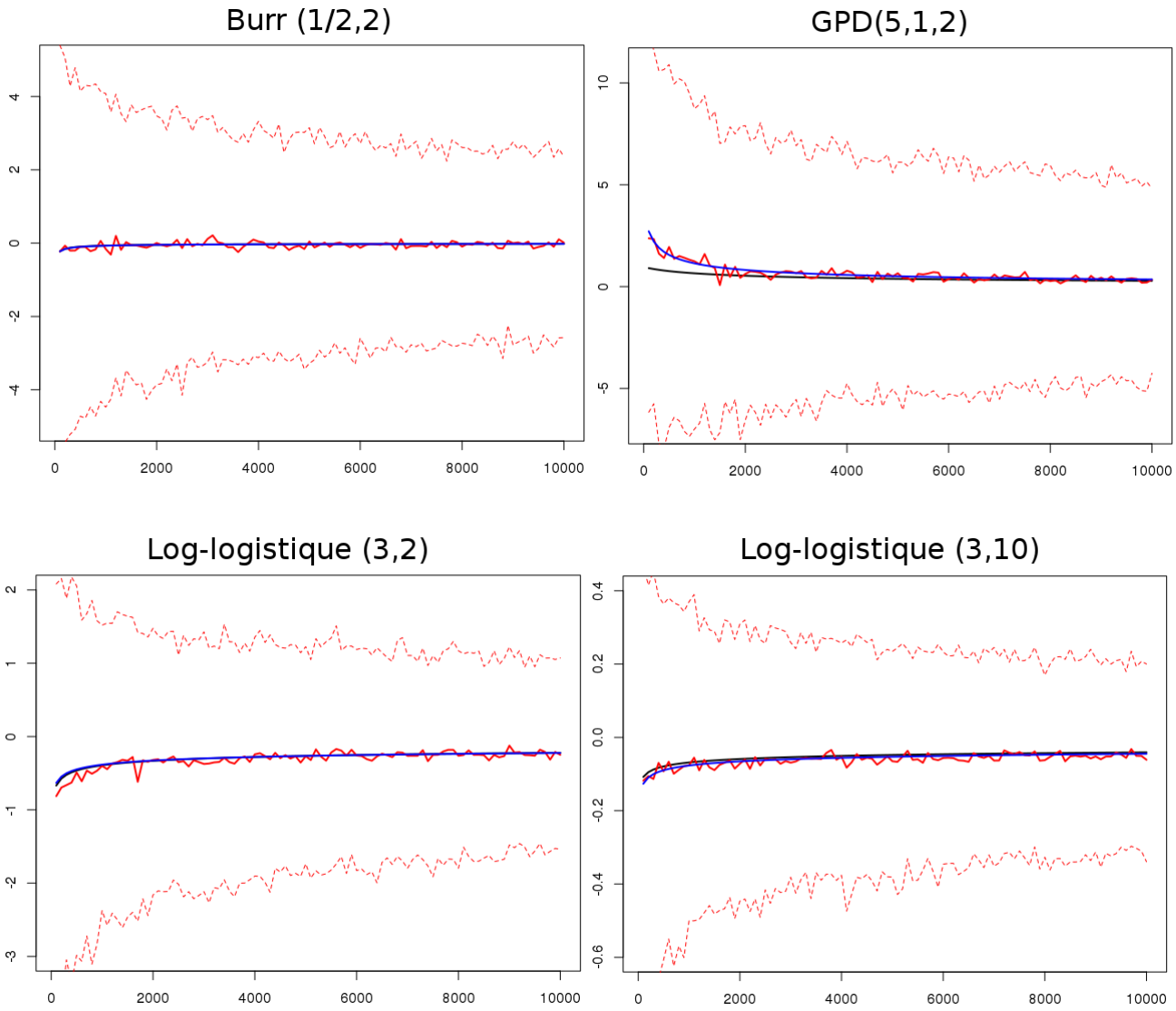

FIGURE 4.5 – L'erreur d'extrapolation relative comme une fonction de n. En noir, $\varepsilon_W(p_n;\alpha_n)$ (cf (4.7)). En bleu, $\tilde{\varepsilon}_W(p_n;\alpha_n)$ (cf (4.8)). En rouge, $\hat{\varepsilon}_W(p_n;\alpha_n)$ ainsi qu'un intervalle de confiance empirique à 90% (cf (4.11)). Voir la Table 4.2 pour plus de détails sur les lois considérées.

Lois	$U(x)$	$\eta(x)$	γ	ρ
Burr $(k>0, c>0)$	$(x^{1/k}-1)^{1/c}$ $x>1$	$\frac{1}{kc(x^{1/k}-1)}$	$1/(kc)$	$-1/k$
GPD $(\mu\in\mathbb{R}, \sigma>0, \gamma>0)$	$\mu+\frac{\sigma}{\gamma}(x^\gamma-1)$ $x>1$	$\gamma\frac{\sigma-\mu\gamma}{\sigma(x^\gamma-1)+\mu\gamma}$	γ	$-\gamma$
log-Logistique $(\alpha>0, \beta>0)$	$\alpha(x-1)^{1/\beta}$ $x>1$	$\frac{1}{\beta(x-1)}$	$1/\beta$	-1

TABLEAU 4.2 – Exemples de lois appartenant au domaine d'attraction de Fréchet et valeurs associées de γ et ρ.

4.4 Application à un cas réel de mesures de variables environnementales

Dans cette partie, le comportement de l'estimateur de l'erreur d'extrapolation donné par l'équation (4.11) est illustré sur des données de précipitations journalières (en mm) mesurées dans la station de Valleraugue (Cévennes, France) par Météo-France. Le jeu de données utilisé consiste en 20220 mesures faites entre 1958 et 2000. On décompte 5597 valeurs manquantes, 11768 valeurs nulles, et 2855 valeurs non nulles, comprises entre 0 et 51 mm.

Après s'être fixé un seuil égal à $u = 12mm$, nous ajustons une loi GPD aux 84 valeurs supérieures à ce seuil. Les résultats numériques sont présentés Figure 4.6, nous permettant de remarquer que l'indice des valeurs extrêmes est estimé à environ 0.29 ± 0.14. Cela nous laisse supposer que la loi sous-jacente est à queue lourde. Ceci est cohérent avec les précédentes études sur ce jeu de données, voir METHNI et collab. [2014] par exemple et avec la Figure 4.25 qui représente l'estimation de l'indice des valeurs extrêmes par maximum de vraisemblance en fonction du seuil.

```
Call: fpot(x = Z, threshold = 12)
Deviance: 444.7201

Threshold: 12
Number Above: 84
Proportion Above: 0.0057

Estimates
 scale   shape
3.8973  0.2868

Standard Errors
 scale   shape
0.6868  0.1419
```

FIGURE 4.6 – Cumuls de précipitations : Ajustement d'une loi de Pareto Généralisée aux excès. Estimation des paramètres par maximum de vraisemblance.

L'ajustement graphique, donné sous forme de diagramme quantile-quantile sur la Figure 4.21 et de graphe de densité sur la Figure 4.22 est moins satisfaisant que ceux présentés dans le cas des débits du Rhône ou encore des vitesses de vent à Reims. Ainsi, sur la Figure 4.21, l'estimateur à noyau de la densité fluctue beaucoup plus pour ce jeu de données et est bien moins lisse que la densité de la GPD qui lui est superposée. De même, les intervalles de confiance sont particulièrement larges pour les plus grandes valeurs du jeu de données sur la Figure 4.22. Cela est dû aux petits nombres d'excès au-dessus du seuil (seulement 84), lui même dû à la lourdeur de la queue de la loi sous-jacente, qui fait qu'on observe beaucoup d'évènements fréquents et peu d'évènements de grandes ampleurs. C'est une des caractéristiques de la région, les grandes valeurs du jeu de données correspondant à des mesures faites durant ce que l'on appelle les épisodes cévenols.

La Figure 4.23 représente deux estimations de l'erreur d'extrapolation associée à l'approximation Weissman en fonction du seuil pour T = 1000 ans. Le premier estimateur considéré, en noir, correspond à l'équation (4.11), avec $\hat{\rho}$ décrit équation (4.10). Le deuxième estimateur, en rouge, correspond à cette même équation dans laquelle on a posé $\hat{\rho} = -1$. Cette figure nous permet de remarquer la haute instabilité de l'estimateur de ρ donné équation (4.10), qui se caractérise par les écarts que produit la courbe noire. La courbe rouge présente quant à elle un comportement un peu plus acceptable : elle semble croître pour des seuils allant de 5 à 11 mm avant de se stabiliser pour des seuils supérieurs à cette dernière valeur. Malgré tout, cette stabilisation reste toute relative, les valeurs de la courbe rouge pouvant varier entre -1 et 0.3. De plus, ces variations peuvent être très abruptes, et ce même pour des seuils dont l'écart est minime, limitant la portée d'exploitation de ces résultats. Ainsi, le choix d'un seuil de 12 mm mène à une estimation de l'erreur d'extrapolation d'environ 0.11 alors qu'un choix de 13 mm donne une valeur de -0.38.

Cet effet se répercute sur la Figure 4.24, qui représente l'estimateur de l'erreur d'extrapolation décrit équation (4.11) (dans lequel on a posé $\hat{\rho} = -1$) en fonction de la période de retour T pour trois seuils différents : $u = 12mm$ (en noir), $u = 12.2mm$ (en bleu) et $u = 11.8mm$ (en rouge). Notons que, pour une période de retour fixée T = 1000 ans sur la Figure 4.24, on retrouve les estimations de l'erreur d'extrapolation ob-

tenues quand on fixe le seuil respectivement à $u = 12mm$, $u = 12.2mm$ et $u = 11.8mm$ sur la Figure 4.23, à savoir 0.11, -0.03 et -0.10. La Figure 4.24 nous permet de constater qu'il est très délicat de quantifier les limites d'extrapolation avec les estimateurs développés, qui manquent de robustesse. Ainsi, pour un seuil $u = 12mm$, la courbe noire suggère une erreur d'extrapolation positive, qui croît légèrement avec la période de retour quand des choix de seuil $u = 12.2mm$ et $u = 11.8mm$ mènent à la conclusion que l'erreur est négative et décroît avec la période de retour. Dans le cas des extrapolations basées sur des lois du domaine d'attraction de Fréchet, il semblerait donc que l'erreur d'extrapolation soit moins un problème que l'estimation des paramètres, dont la volatilité est particulièrement importante.

4.5 Perspectives

Dans ce chapitre, nous avons proposé des outils mathématiques permettant de quantifier l'erreur d'extrapoaltion faite à partir d'un jeu de données. Nous avons ainsi pu montrer que des limites d'extrapolation existaient bel et bien, et qu'elles étaient plus ou moins restrictives selon l'aléa climatique considéré. Ces travaux ouvrent la voie à plusieurs perspectives. Ces dernières sont résumées dans les paragraphes qui suivent.

4.5.1 Autres sources d'erreur

Rappelons que l'erreur globale $\epsilon(p_n)$ se décompose comme la somme de deux termes :

$$\epsilon(p_n) \quad = \quad \epsilon_{est}(p_n) + \epsilon_{ext}(p_n),$$

le premier terme correspondant à l'erreur d'estimation des paramètres et le deuxième à l'erreur d'approximation du vraie quantile.

Une première perspective serait de proposer un outil permettant de quantifier l'erreur globale. Pour EDF, cela permettrait de juger de l'erreur totale faite sur l'estimation d'un niveau de retour. A l'heure actuelle, il existe des outils permettant uniquement de quantifier l'erreur d'estimation : ces outils consistent en des intervalles de confiance sur les niveaux de retour (cf Paragraphe 2.1). Au vu de la décomposition ci-dessus - et ayant déjà quantifiée l'erreur d'approximation - la construction d'un outil permettant de quantifier l'erreur globale passerait par l'étude de l'erreur d'estimation, qui est une des perspectives dejà évoquées Paragraphe 2.3. Moyennant une étude de l'erreur d'estimation, il faudrait alors prendre en compte le fait que cette dernière est de nature différente de l'erreur d'approximation. En effet, l'erreur d'estimation étant aléatoire, il faudrait la condidérer en espérance, pour pouvoir la comparer à l'erreur d'approximation. On pourrait alors voir si ces deux erreurs s'ajoutent ou se compensent, et quantifier leur somme.

4.5.2 Choix optimal du seuil

La perspective que nous développons ici est également liée à la décomposition de l'erreur globale donnée ci-dessus. Cette perspective consiste à proposer une méthode pour choisir de manière optimale le seuil u (ou le nombre de statistiques d'ordre k_n).

En effet, dans les approximations utilisées ci-dessus, issues de l'approximation GPD, l'erreur d'approximation quantifie à quel point les excès ont convergé vers leur loi limite, voir Theorème 3. Or ce dernier Théorème est valide lorsque le seuil u est grand. Ainsi, plus ce seuil est grand, meilleure est la qualité d'approximation. En contrepartie, plus le seuil est grand et moins on dispose d'excès pour estimer les paramètres du modèle, menant à une erreur d'estimation plus élevée. C'est un problème de compromis biais-variance.

La Figure 4.13 du Paragraphe 4.2 pourrait en partie nous permettre de répondre graphiquement à cette question. Rappelons que cette figure représente l'estimation de l'erreur d'extrapolation relative (cf (4.3)) en fonction du seuil, lorsque la période de retour T est fixée, pour le jeu de données des débits du Rhône. Ainsi, pour un seuil de $2400m^3/s$, nous estimions une erreur d'environ 20%. Si l'on considère plutôt un seuil de $2000m^3/s$ sur cette figure, on se rend compte que ce seuil correspond à un minimum local. Suivant ce seuil, l'erreur est estimée à environ 10%, soit une valeur plus faible que pour un seuil de $2400m^3/s$. Or ce dernier est plus grand que $2000m^3/s$. Par conséquent, l'estimation des paramètres sera plus difficile avec un seuil de $2400m^3/s$ qu'avec un seuil de $2000m^3/s$, puisque le nombre d'excès pour mener l'estimation en question sera plus faible. Ainsi, il semblerait que le choix d'un seuil de $2000m^3/s$ soit doublement gagnant pour ce jeu de données. D'une part, l'erreur d'extrapolation semble être modérée (d'environ 10%). D'autre part, l'erreur d'estimation sera également plus faible, puisque disposant d'un nombre d'excès plus élevé.

Proposer un outil mathématique serait donc des plus intéressants pour répondre à cette question. Cela consisterait en un outil d'optimisation permettant de minimiser en k_n l'erreur globale, c'est à dire de la

somme des deux erreurs. Avoir un résultat sur le comportement asymptotique de l'erreur d'estimation serait nécessaire pour mener à bien ce projet, nous renvoyant là aussi aux perspectives évoquées Paragraphe 2.3.

4.5.3 Proposer des estimateurs alternatifs...

du paramètre θ_2

Rappelons que l'estimateur de l'erreur d'extrapolation associée à l'approximation ET fait appel à un estimateur de θ_2, voir (4.3). Au Paragraphe 4.1.2, nous proposions d'estimer θ_2 par

$$\hat{\theta}_2 := 2\hat{\theta}_1 \mathbb{1}\{\hat{\theta}_1 > 0\} + \hat{\theta}_1 \mathbb{1}\{\hat{\theta}_1 < 0\}.$$

Ce choix était motivé par le Lemme 2 du Chapitre 2, qui établit un lien entre θ_1 et θ_2 :

$$\theta_2 = \begin{vmatrix} 2\theta_1 & \theta_1 > 0 \\ \theta_1 & \theta_1 < 0 \\ 0 & \theta_1 = 0 \text{ et } \ell_1 \neq 1. \end{vmatrix}$$

Cependant, dans le cas où $\ell_1 = 1$ ($\theta_1 = 0$), l'estimateur de θ_2 en question peut ne pas être consistant. C'est par exemple le cas de la loi Gamma, pour laquelle $\theta_1 = 0$, $\ell_1 = 1$ et K_2 est à variation régulière d'indice $\theta_2 = -1$. Dans ce cas précis, il est aisé de voir que $\hat{\theta}_2$ est un estimateur biaisé du vrai θ_2. Une des perspectives de travail concernant ce chapitre est de proposer un estimateur consistant de θ_2.

Une idée serait de prendre exemple sur l'estimateur de θ proposé par DE VALK et CAI [2018]. Pour proposer cet estimateur, DE VALK et CAI [2018] se basent sur le fait que $V \in ERV$ (voir Paragraphe 1.3.4). Plus précisémment, il existe une fonction positive a telle que, pour tout $t > 0$,

$$\lim_{x\to\infty} \frac{V(tx) - V(x)}{a(x)} = \int_1^t u^{\theta-1} du =: L_\theta(t). \tag{4.13}$$

D'après, DE HAAN et FERREIRA [2007, Corollaire 1.1.10], une condition suffisante est V est dérivable de dérivée V' vérifiant

$$\lim_{x\to\infty} \frac{V'(tx)}{V'(x)} = t^{\theta-1}. \tag{4.14}$$

De plus, sous cette condition, un choix possible pour la fonction a est $a(x) = xV'(x)$. Moyennant ce choix, a est à variation régulière d'indice θ. DE VALK et CAI [2018] tire profit de ce résultat pour proposer un estimateur de θ basé sur l'accroissement des logarithmes de la fonction a à partir du point k_n (voir équation (1.73)).

Pour proposer un estimateur de θ_2, l'idée serait de renforcer les hyptohèses précédentes en supposant que V est à variation régulière étendue d'ordre deux (cf Définition 5), c'est à dire qu'il existe une fonction A avec $A(x) \to 0$ lorsque $x \to \infty$ et $\rho < 0$ tels que,

$$\lim_{x\to\infty} \frac{1}{A(x)} \left[\frac{V(tx) - V(x)}{a(x)} - L_\theta(t) \right] = H_{\theta,\rho}(t) := \int_1^t s^{\theta-1} L_\rho(s) ds.$$

converge uniformément sur tout compact pour $t > 0$.

La clé serait alors de trouver un lien entre A et K_2 dans l'équation précédente, de la même façon qu'il existait une analogie entre les fonctions a et K_1 et entre les indices θ et θ_1, voir Paragraphe 4.1.2. L'idée serait ensuite d'utiliser le fait que A est à variation régulière d'indice ρ, voir Paragraphe 1.2.3. Si nous parvenions à établir un tel lien, alors θ_2 pourrait s'obtenir en fonction de ρ, qui lui s'estimerait en se basant sur l'accroissement des logarithmes de la fonction A par exemple, ou alors en utilisant les divers estimateurs de ρ proposés dans la littérature, voir Paragraphes 1.2.3 et 4.3. Notons que ces derniers sont proposés sous l'hypothèse que la fonction quantile de queue U est à variation régulière étendue d'ordre deux et non V. Il faudrait donc les adapter à nos hypothèses de modèle.

du paramètre du second-ordre ρ

Dans le cas de l'approximation Weissman, nous avons eu à estimer le paramètre du second-ordre ρ, voir partie 4.3. Ce paramètre est reconnu dans la littérature comme difficile à estimer, et est bien souvent estimé par une valeur ponctuelle, de -1, voir par exemple BEIRLANT et collab. [2002]. L'estimateur que nous avons utilisé (cf équation 4.10) s'est basé sur un cas particulier de l'estimateur proposé dans GOMES et collab. [2002]. Cependant, dans le cas d'application des données de pluie, cet estimateur s'est révélé particulièrement instable, voir Figure 4.23.

Utiliser un estimateur plus robuste de ρ est ainsi une des perspectives de ce chapitre. La première idée serait de considérer l'estimateur général de ρ proposé dans GOMES et collab. [2002], et non plus le cas particulier où $\alpha = 2$, qui mène à une forme explicite de ce dernier.

Tester d'autres estimateurs de la littérature est aussi une possibilité envisagée, nous renvoyons encore une fois le lecteur aux Paragraphes 1.2.3 et 4.3 pour plus de détails à ce sujet.

4.5.4 Intervalles de confiance asymptotiques des estimateurs de l'erreur d'extrapolation

Obtenir des garanties asymptotiques pour les estimateurs (4.3) et (4.11) est une des perspectives de ce chapitre.

Approximation Weissman Dans le cas de l'erreur d'extrapolation associée à Weissman, cela consisterait à obtenir un résultat de normalité asymptotique pour (4.9). Pour ce faire, on pourrait se baser sur BEIRLANT et collab. [2002, Théorème 3.2]. Une représentation asymptotique de $\hat{\eta}(n/k_n)$ y est donnée. Elle fait cependant intervenir ρ, qui est inconnu, la rendant inutilisable en pratique. L'étude du comportement asymptotique de $\hat{\rho}$ est au préalable nécessaire. Cette dernière pourrait se baser sur la normalité asymptotique de $\hat{\rho}_G$, prouvée dans GOMES et collab. [2002, Théorème 2.2].

Approximation ET Contrairement à l'estimateur proposé dans le cas de l'approximation Weissman, celui associé à l'approximation ET dépend de quatre paramètres : θ_1, θ_2, $K_1(\log n/k_n)$ et $K_2(\log n/k_n)$. Proposer un résultat de normalité asymptotique pour l'estimateur de l'erreur d'extrapolation ET consisterait donc dans un premier temps à établir la loi jointe de $\hat{\theta}_1$, $\hat{K}_1(\log n/k_n)$, $\hat{\theta}_2$ et $\hat{K}_2(\log n/k_n)$. Dans un deuxième temps, il s'agirait d'adapter une méthode delta dans un cadre multivarié, afin d'en déduire la loi de l'estimateur de l'erreur d'extrapolation, qui est une fonction des ces quatre quantités.

Pour ce qui est de la première étape, rappelons que $\hat{\theta}_2$ et $\hat{K}_2[\log(n/k_n)]$ sont liés à $\hat{\theta}_1$ et $\hat{K}_1(\log n/k_n)$. En effet,

$$\hat{\theta}_2 := 2\hat{\theta}_1 \mathbb{1}\{\hat{\theta}_1 > 0\} + \hat{\theta}_1 \mathbb{1}\{\hat{\theta}_1 < 0\}$$

et

$$\hat{K}_2[\log(n/k_n)] := \hat{K}_1^2[\log(n/k_n)] + \hat{K}_1[\log(n/k_n)](\hat{\theta}_1 - 1),$$

de sorte que seule la loi jointe entre $\hat{\theta}_1$ et $\hat{K}_1(\log n/k_n)$ serait nécessaire. Or cette loi jointe est l'objet de la condition **(A3)** du Chapitre 3 (voir les Théorèmes 3 et 4 pour ce qui est des lois marginales).

4.5.5 Utilisation d'un modèle paramétrique

Pour terminer, il serait intéressant de quantifier l'erreur totale faite en utilisant des modèles paramétriques. Dans le but de modéliser au mieux la loi du hasard régissant un phénomène, rappelons que ces derniers consistent à utiliser une famille de lois statistiques paramétriques qui ont tendance à bien s'ajuster à la queue de distribution empirique des données rencontrées dans un domaine d'application précis, voir EVIN et collab. [2016]. Dans le cas des débits du Rhône, l'idée serait alors d'exploiter l'hypothèse de lognormalité des données et dans le cas des vitesses de vent, le fait que les mesures suivent une loi Gamma.

Pour quantifier l'erreur totale faite, une première étape consisterait à effectuer un test d'adéquation de la loi en question aux données. Si ce dernier n'est pas rejeté, il s'agirait alors de montrer que

$$\hat{q}_{\text{EVT}} - q = (\hat{q}_{\text{EVT}} - \hat{q}_{param}) + (\hat{q}_{param} - q) \simeq \hat{q}_{\text{EVT}} - \hat{q}_{param},$$

avec $\hat{q}_{\text{EVT}}$ un estimateur des quantiles extrêmes par la théorie des valeurs extrêmes et $\hat{q}_{param}$ l'estimateur obtenu directement à partir de la fonction quantile théorique de la loi ajustée.

Pour démontrer cette relation, une idée serait de montrer dans un premier temps que

$$\hat{q}_{param} - q = \mathrm{O}\left(\frac{1}{\sqrt{n}}\right),$$

en utilisant le fait que l'estimateur paramétrique est basé sur tout l'échantillon (à savoir n statistiques d'ordre). Dans un deuxième temps, il faudrait montrer que

$$\hat{q}_{\text{EVT}} - \hat{q}_{param} = \mathrm{O}\left(\frac{1}{\sqrt{k}}\right),$$

en utilisant cette fois-ci le fait que l'estimateur $\hat{q}_{\text{EVT}}$ est basé sur les k plus grandes statistiques d'ordre. Il suffirait alors de remarque que $k << n$ pour conclure le résultat voulu.

Moyennant ce résultat, à savoir $\hat{q}_{\text{EVT}} - q \simeq \hat{q}_{\text{EVT}} - \hat{q}_{param}$, étudier l'erreur totale reviendrait alors à comparer $\hat{q}_{\text{EVT}}$ à $\hat{q}_{param}$.

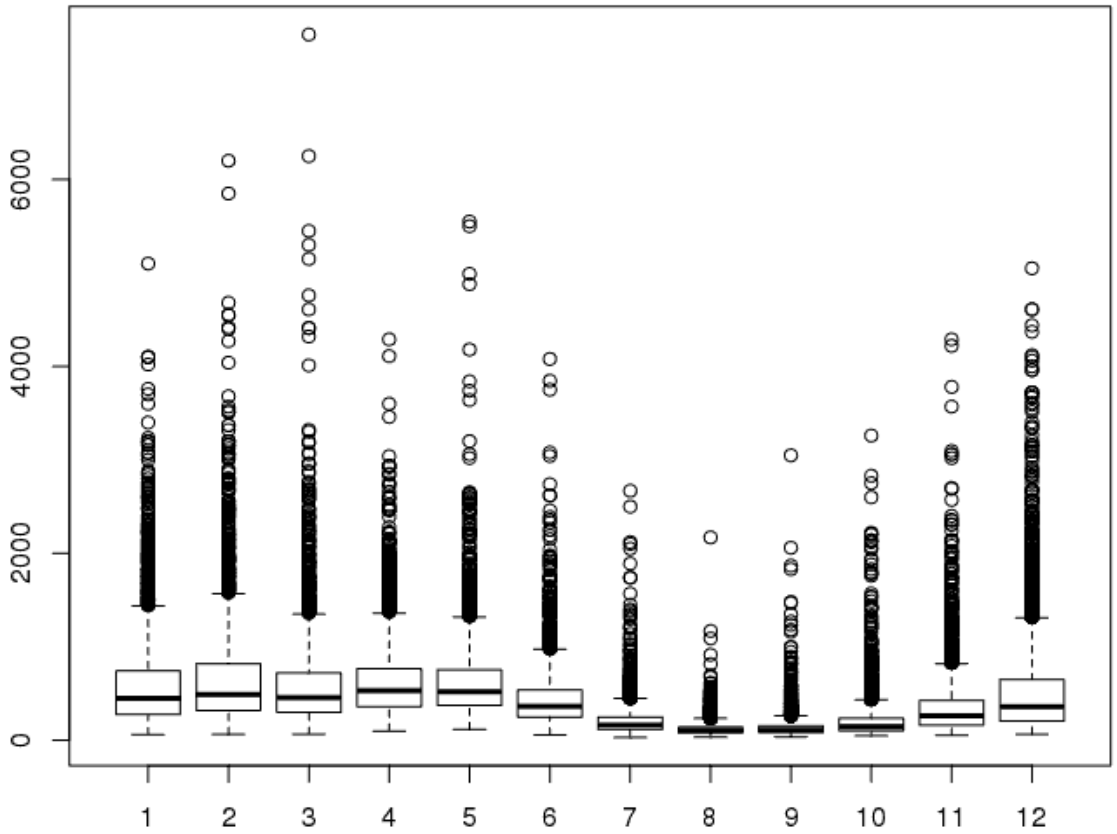

FIGURE 4.7 – Données du Rhône : Boxplot en fonction des différents mois de l'année (de Janvier à Décembre).

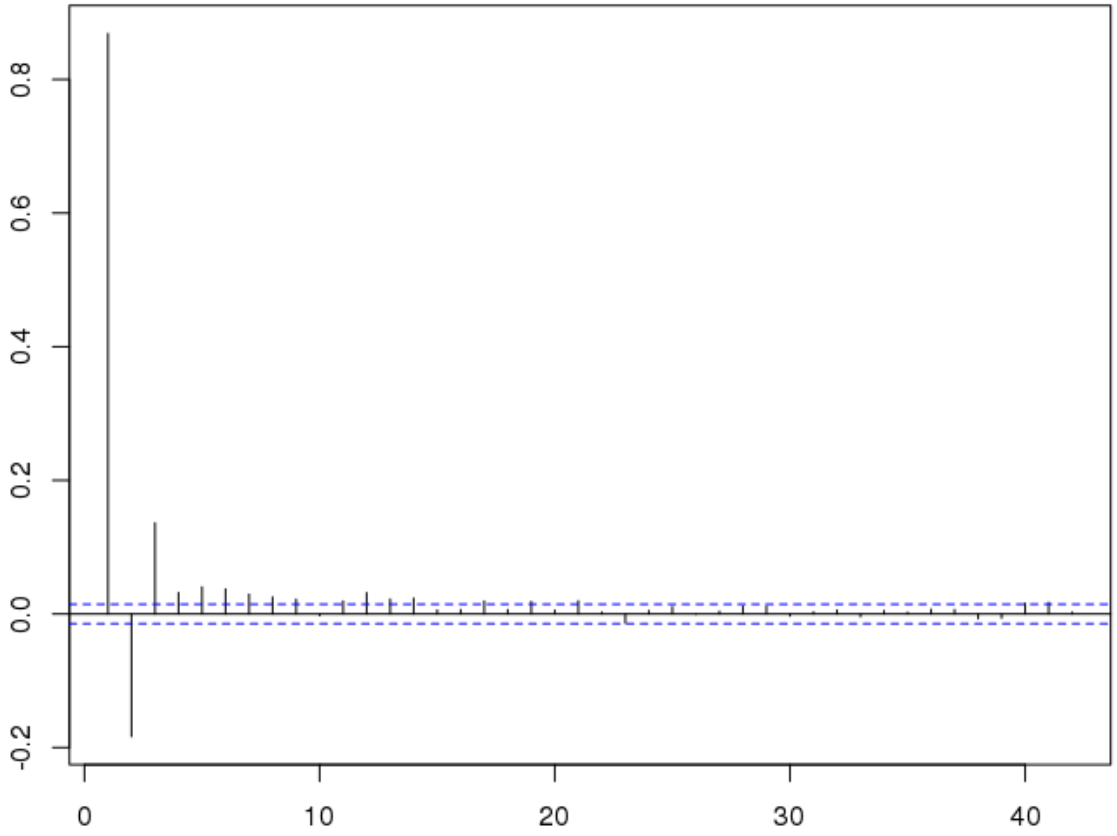

FIGURE 4.8 – Données du Rhône : Graphe d'auto-corrélation partielle représentant l'auto-corrélation partielle de la série (en ordonnées) en fonction du nombre de jours d'espacement entre les mesures (en abcisse). Intervalle de confiance basé sur la série décorrélée (en pointillés bleus).

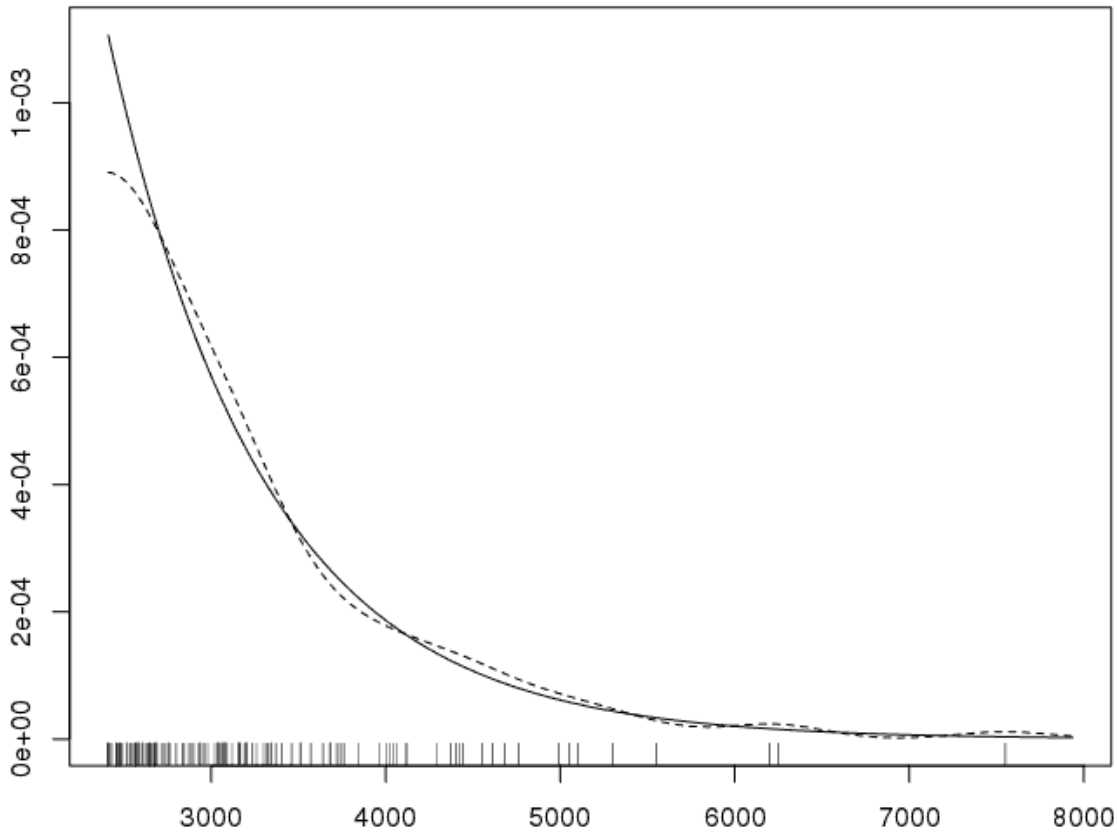

FIGURE 4.9 – Données du Rhône : Graphe de densité superposant une estimation à noyau de la densité des données (en pointillés) à la densité d'une loi GPD de paramètres donnés Figure 4.3. Les traits en abcisse représentent les observations.

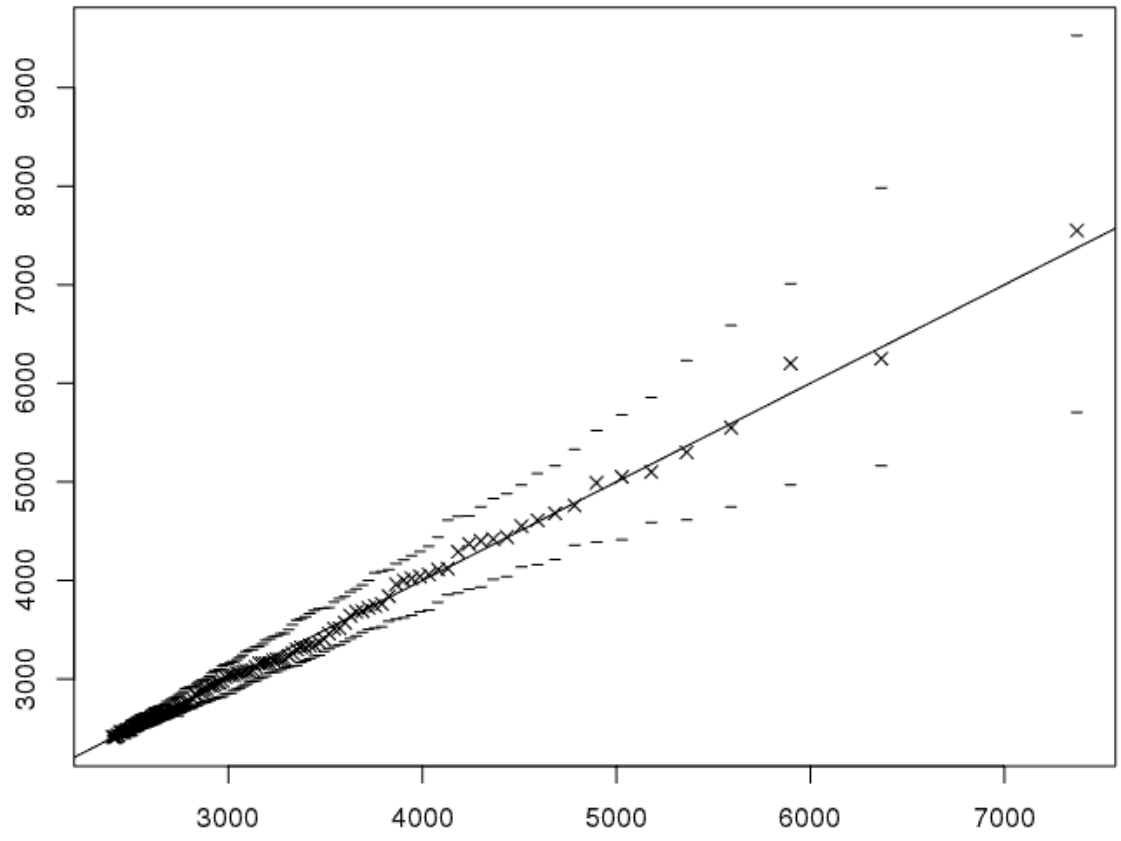

FIGURE 4.10 – Données du Rhône : Diagramme Quantile-Quantile. En ordonnées, les quantiles empiriques. En abcisse, les quantiles théoriques d'une loi GPD de paramètres donnés Figure 4.3.

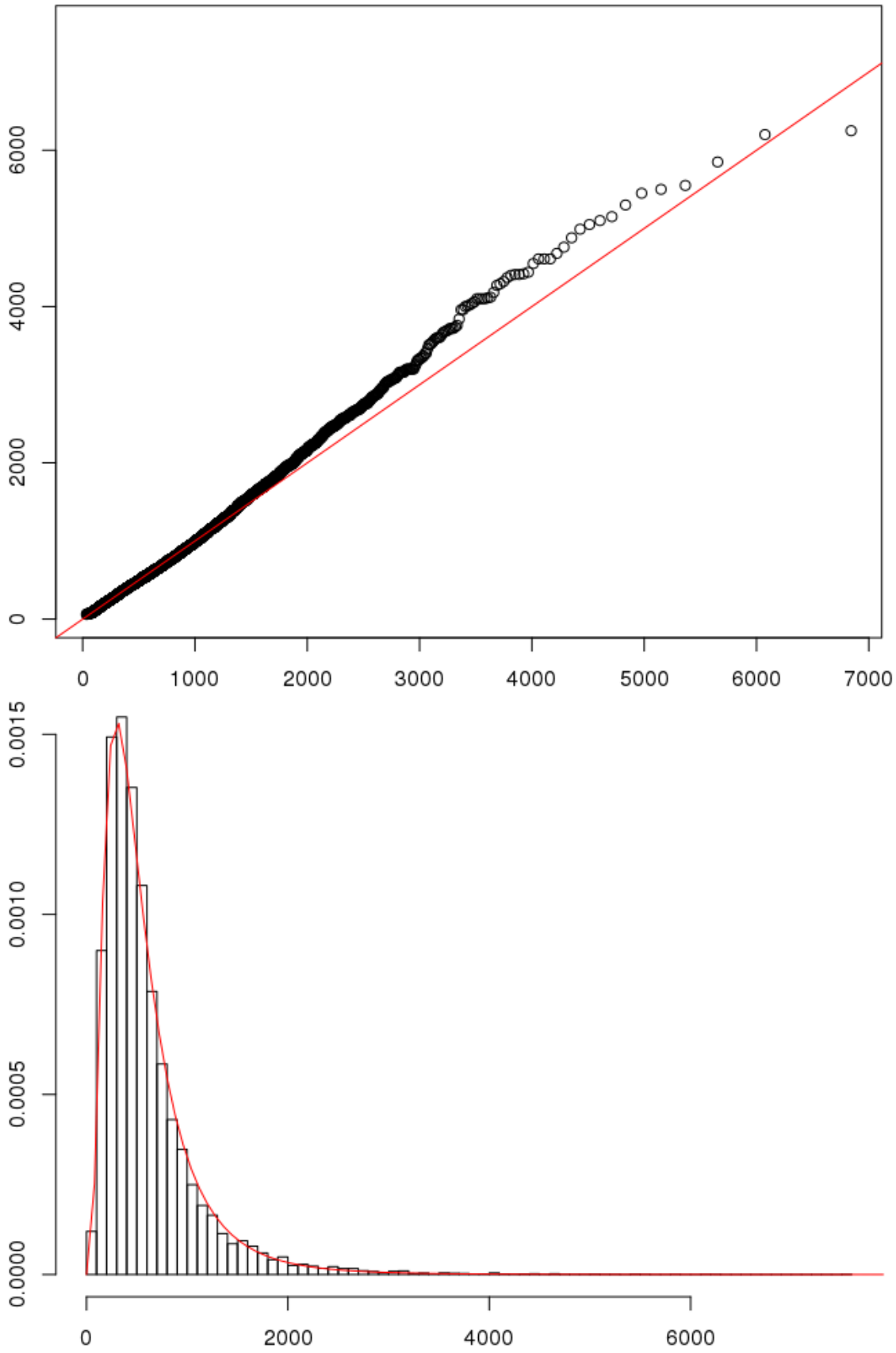

FIGURE 4.11 – Données du Rhône : Diagramme Quantile-Quantile (en haut) mettant en relation les quantiles empiriques (en ordonnées) et les quantiles théoriques d'une loi Lognormale de paramètres appropriés (en abcisse). Graphe de densité (en bas) superposant un histogramme des données à la densité théorique d'une loi Lognormale de mêmes paramètres.

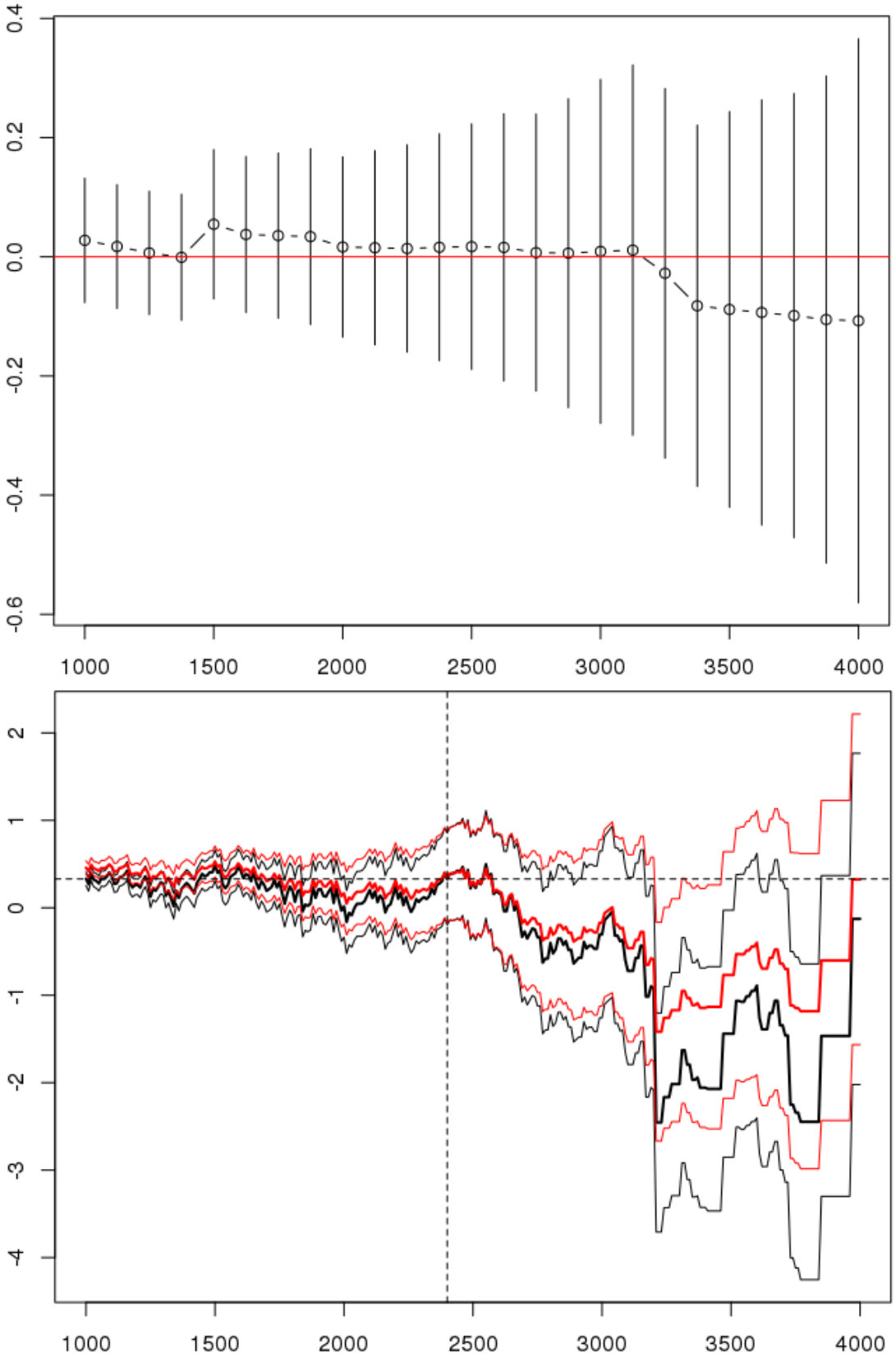

FIGURE 4.12 – Données du Rhône : Estimation de l'indice des valeurs extrêmes par maximum de vraisemblance en fonction du seuil ainsi qu'un intervalle de confiance à 95% (en haut). Estimation de θ_1 par l'estimateur proposé Chapitre 3 (en noir) et l'estimateur proposé par DE VALK et CAI [2018] (en rouge) en fonction du seuil en m^3/s ainsi que des intervalles de confiance à 95% (en bas).

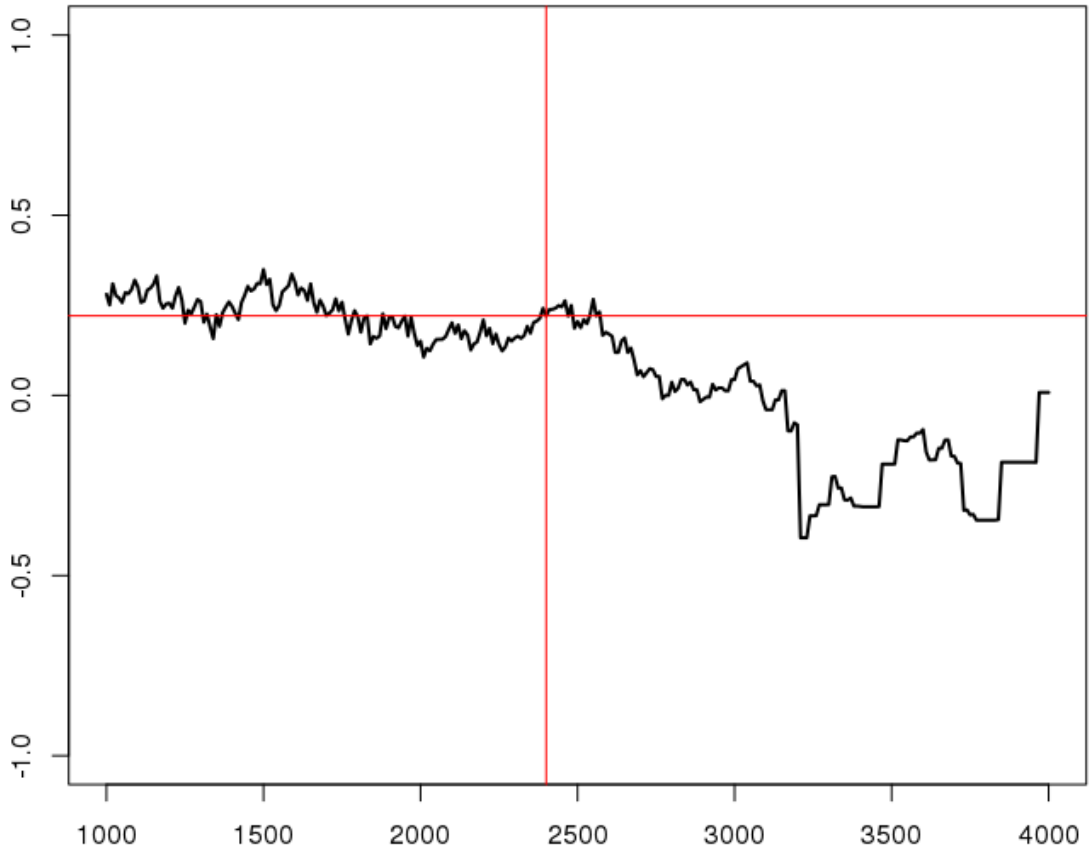

FIGURE 4.13 – Données du Rhône : Estimation de $\tilde{\varepsilon}_{\mathrm{ET}}(p_n;\alpha_n)$ en fonction du seuil en m^3/s pour T = 1000 ans.

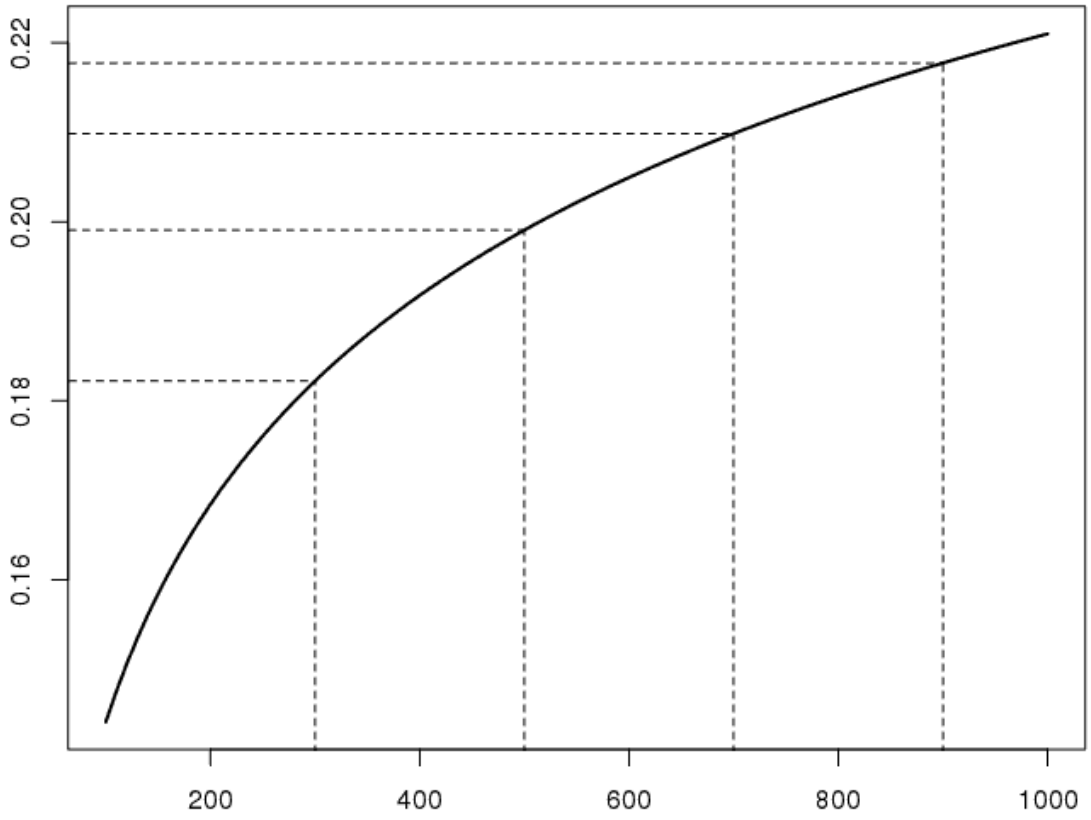

FIGURE 4.14 – Données du Rhône : Estimation de $\tilde{\varepsilon}_{\mathrm{ET}}(p_n;\alpha_n)$ en fonction de la période de retour T, pour $u = 2400 m^3/s$.

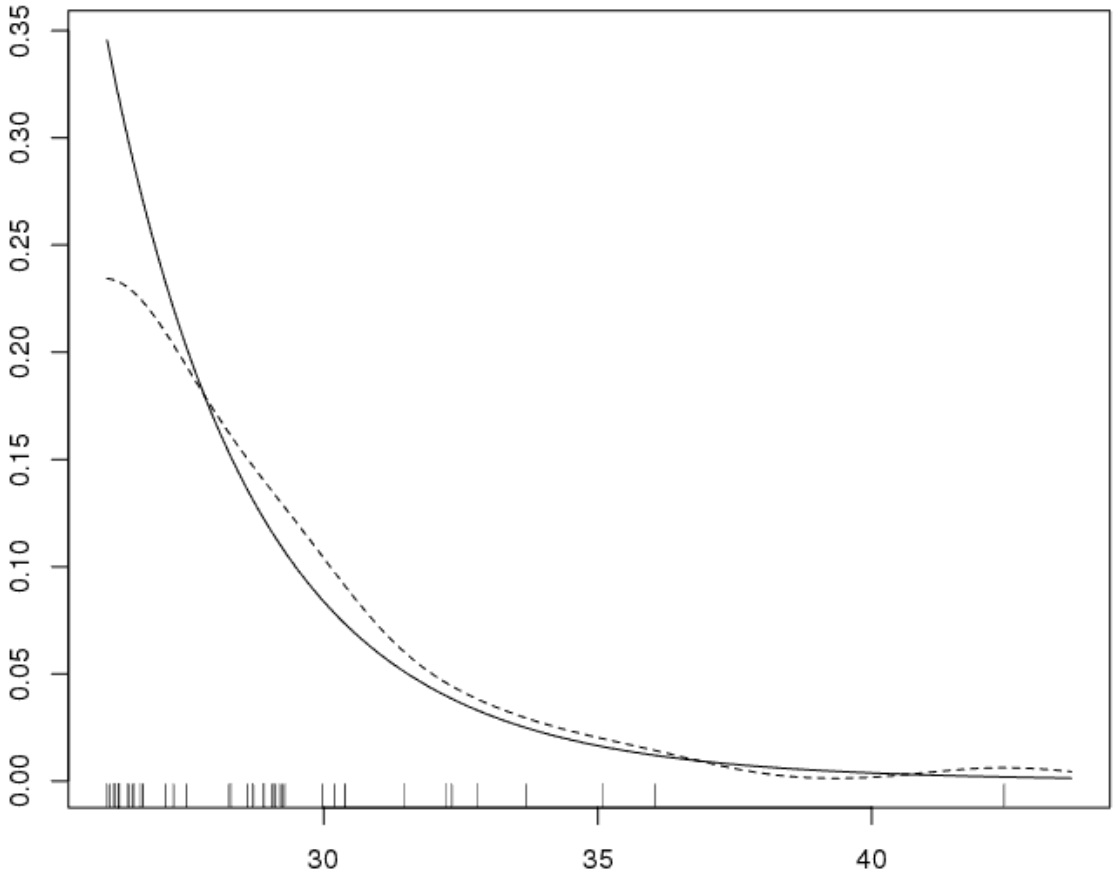

FIGURE 4.15 – Mesures de vents : Graphe de densité superposant une estimation à noyau de la densité des données (en pointillés) à la densité d'une loi GPD de paramètres donnés Figure 4.4. Les traits en abcisse représentent les observations.

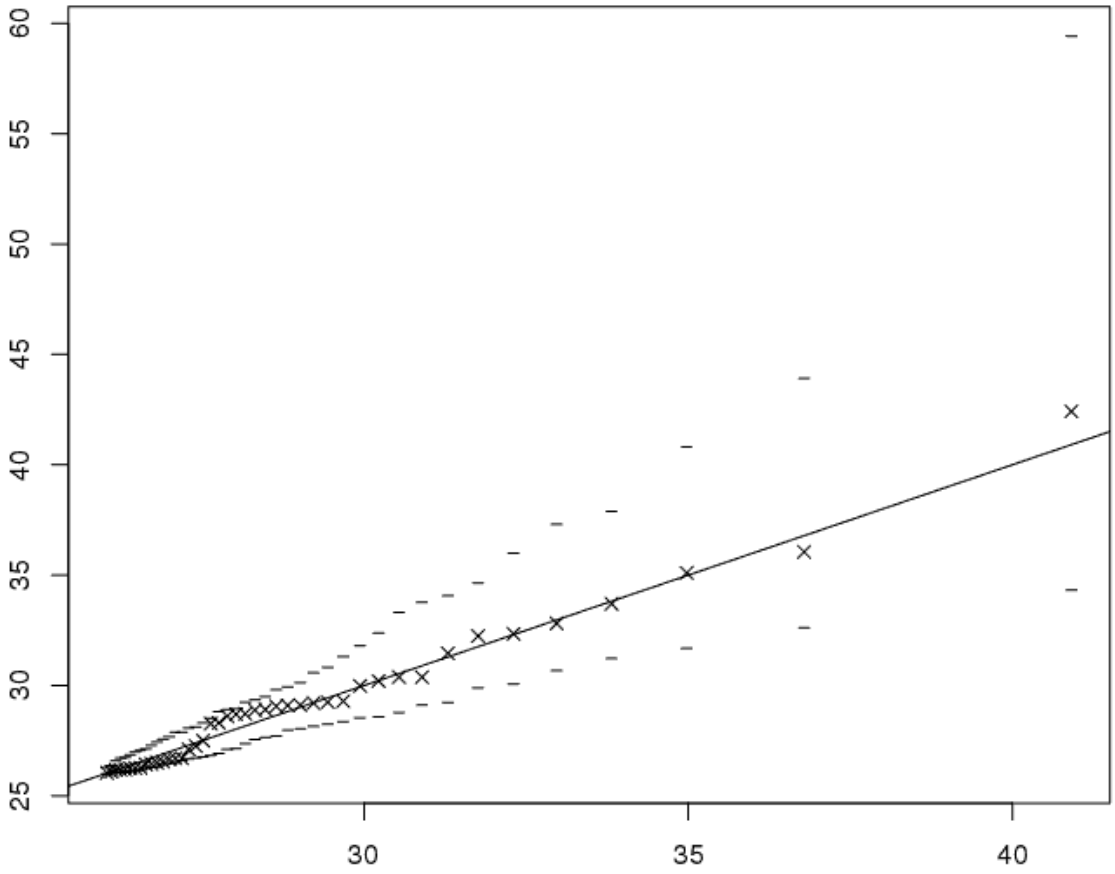

FIGURE 4.16 – Mesures de vents : Diagramme Quantile-Quantile. En ordonnées, les quantiles empiriques. En abcisse, les quantiles théoriques d'une loi GPD de paramètres donnés Figure 4.4.

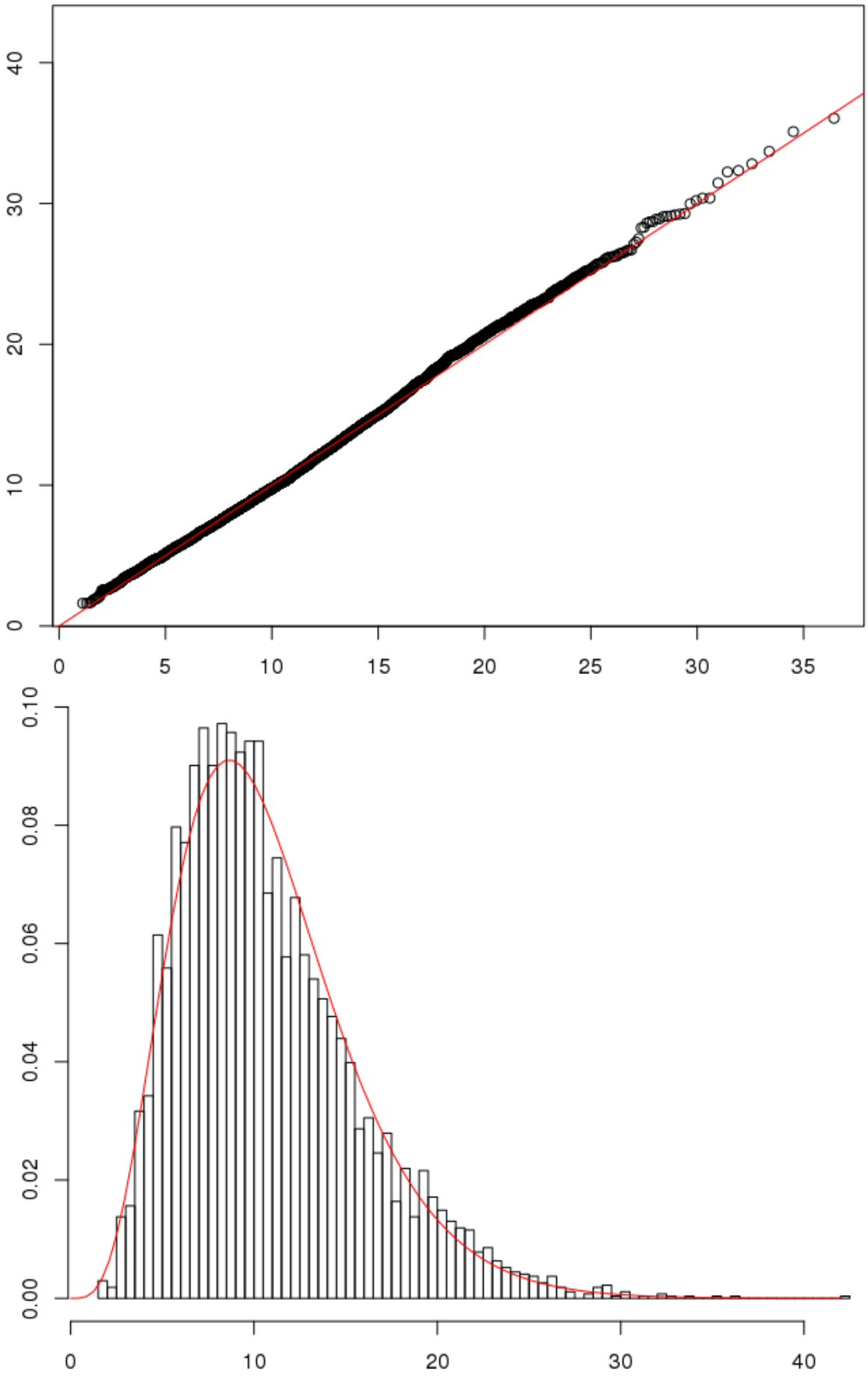

FIGURE 4.17 – Mesures de vents : Diagramme Quantile-Quantile (en haut) mettant en relation les quantiles empiriques (en ordonnées) et les quantiles théoriques d'une loi Gamma de paramètres appropriés (en abcisse). Graphe de densité (en bas) superposant un histogramme des données à la densité théorique d'une loi Gamma de mêmes paramètres.

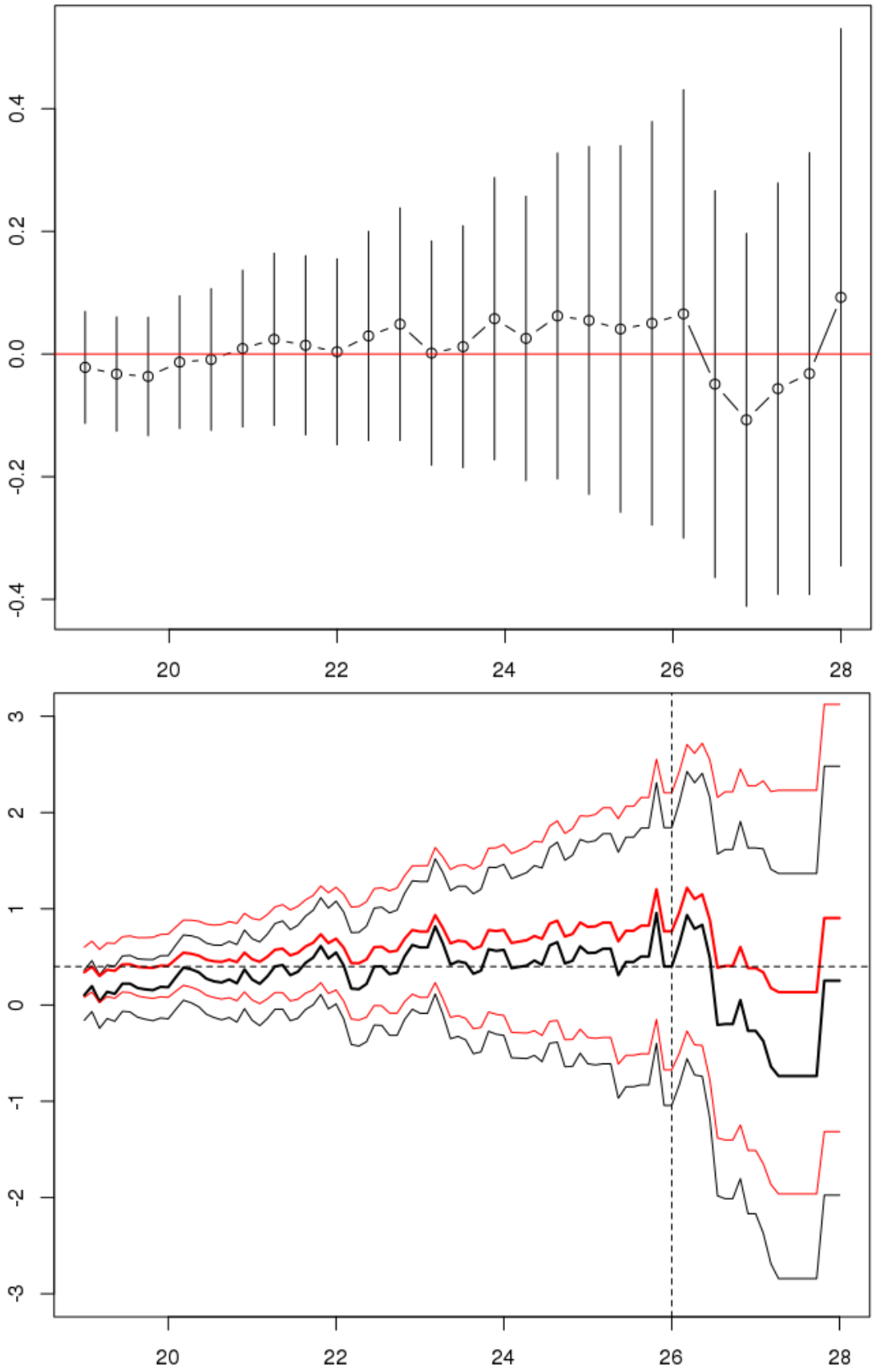

FIGURE 4.18 – Mesures de vents : Estimation de l'indice des valeurs extrêmes par maximum de vraisemblance en fonction du seuil ainsi qu'un intervalle de confiance à 95% (en haut). Estimation de θ_1 par l'estimateur proposé Chapitre 3 (en noir) et l'estimateur proposé par de Valk et Cai (en rouge) en fonction du seuil en m^3/s ainsi que des intervalles de confiance à 95% (en bas).

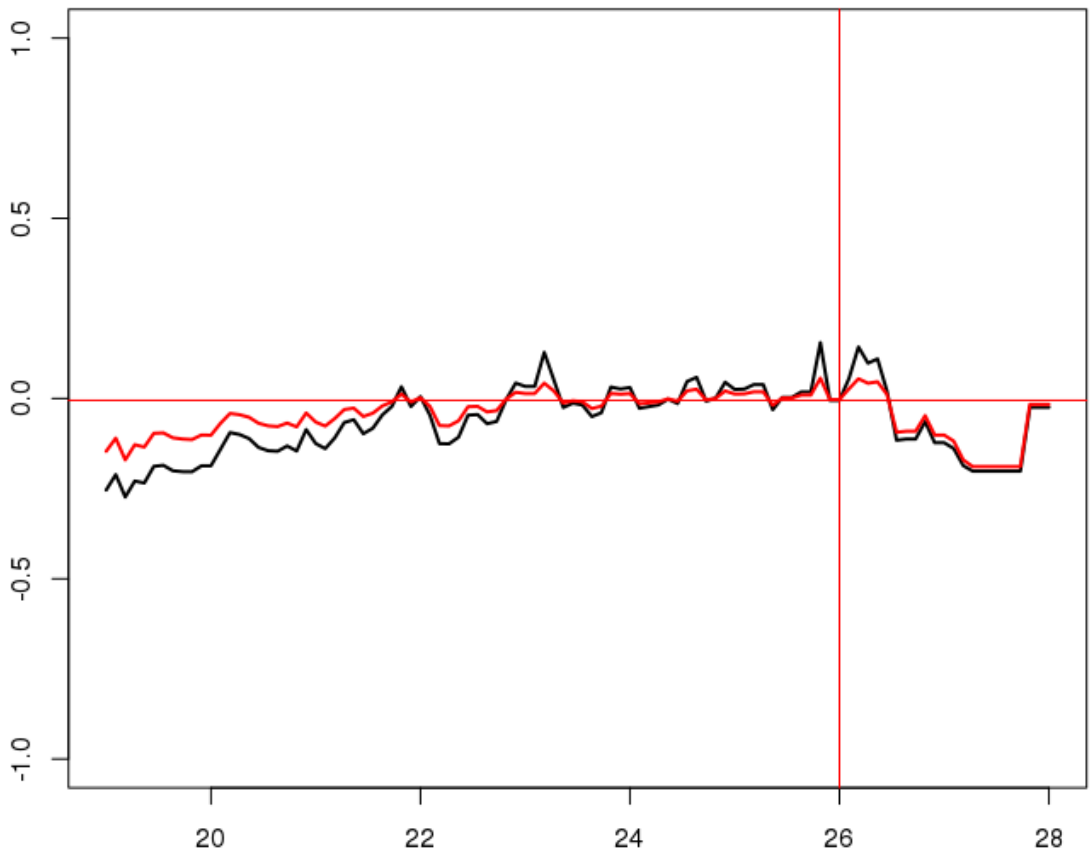

FIGURE 4.19 – Mesures de vents : Estimations de $\tilde{\varepsilon}_{\mathrm{ET}}(p_n;\alpha_n)$ en fonction du seuil en m^3/s pour T = 1000 ans. En noir, l'estimateur donné par l'équation (4.3). En rouge, ce même estimateur dans lequel on a posé $\hat{\theta}_2 = -1$.

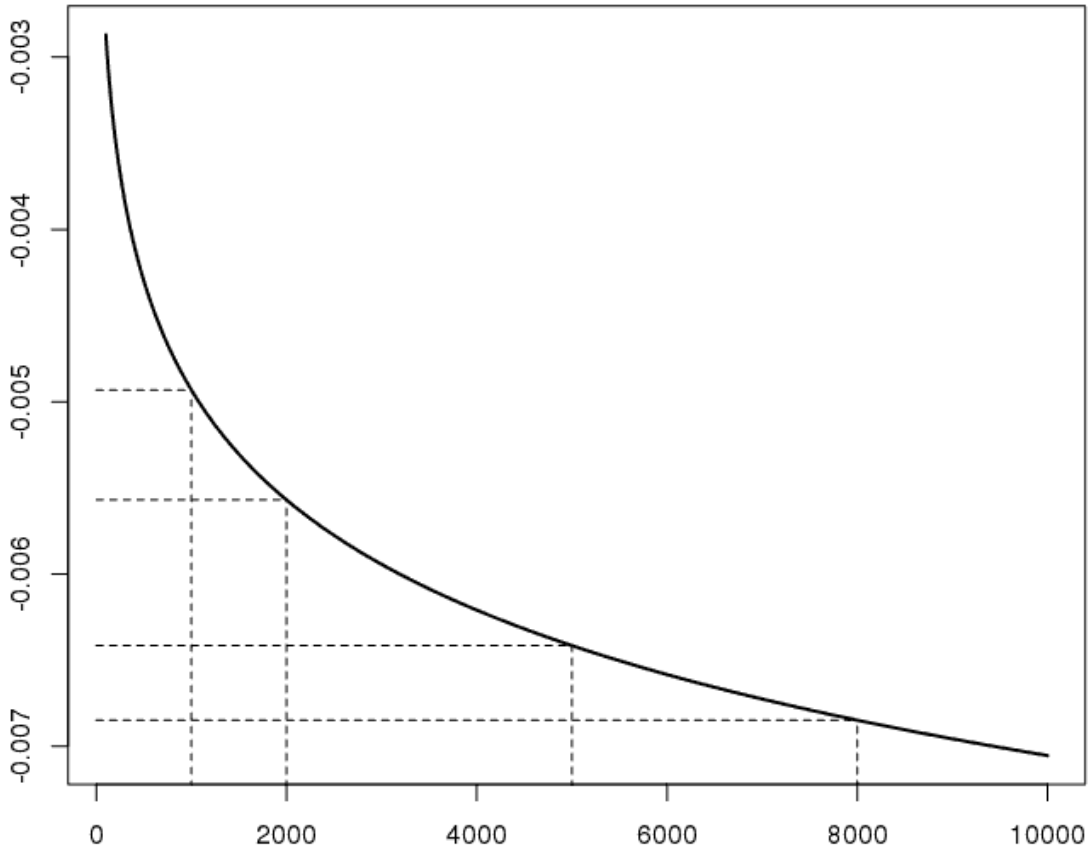

FIGURE 4.20 – Mesures de vents : Estimateur de l'erreur d'extrapolation (voir équation (4.3)) en fonction de la période de retour T, pour $u = 2400 m^3/s$.

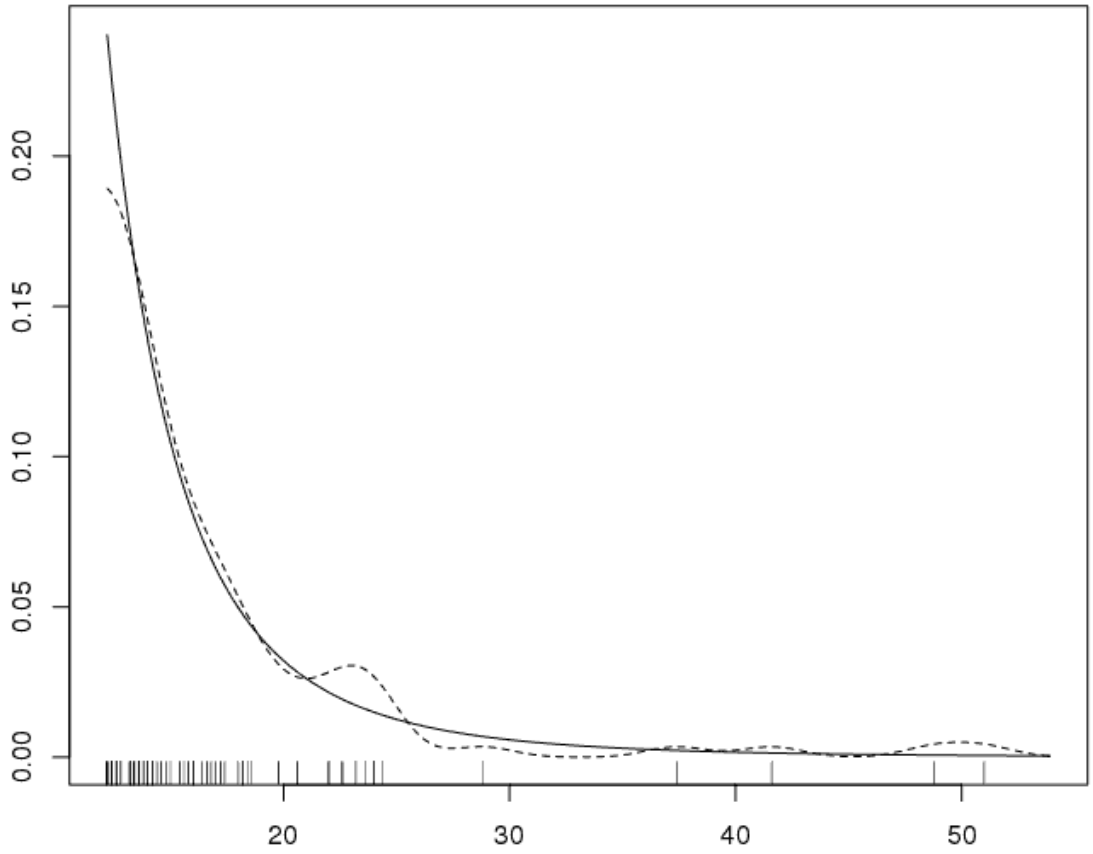

FIGURE 4.21 – Cumuls de précipitations : Graphe de densité superposant une estimation à noyau de la densité des données (en pointillés) à la densité d'une loi GPD de paramètres donnés Figure 4.6. Les traits en abcisse représentent les observations.

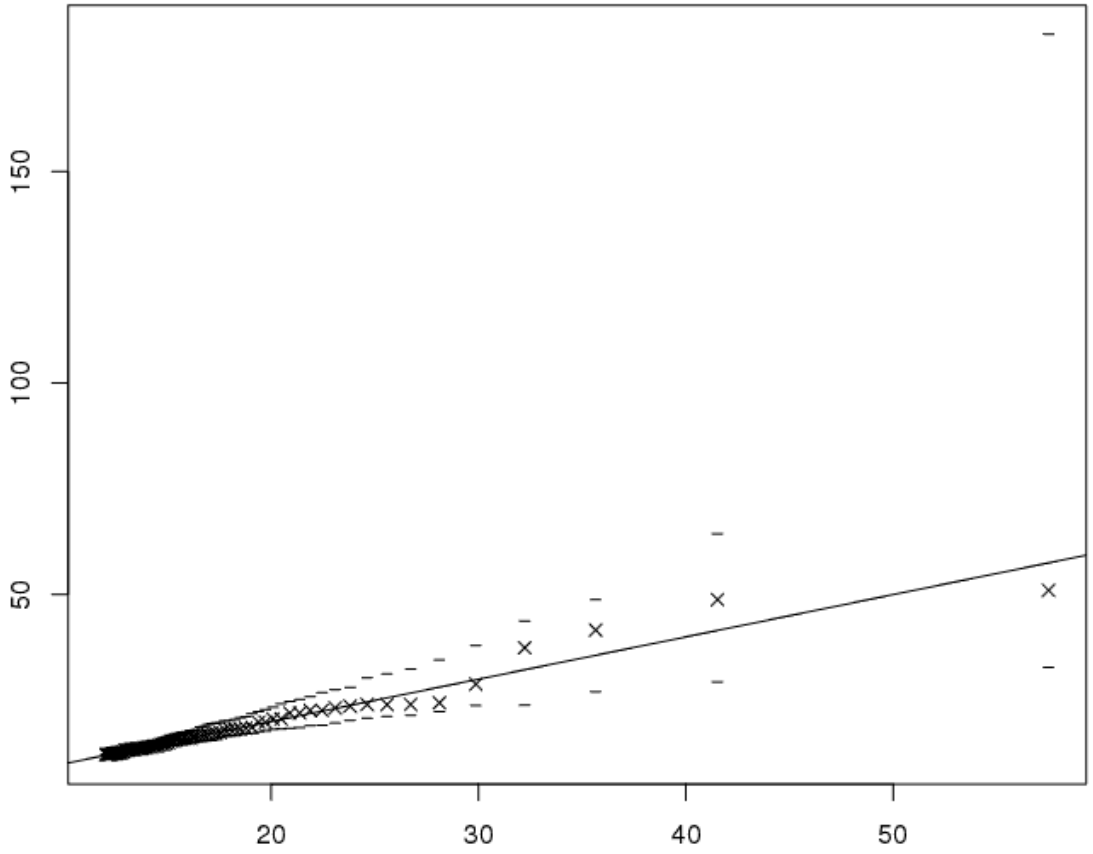

FIGURE 4.22 – Cumuls de précipitations : Diagramme Quantile-Quantile. En ordonnées, les quantiles empiriques. En abcisse, les quantiles théoriques d'une loi GPD de paramètres donnés Figure 4.6.

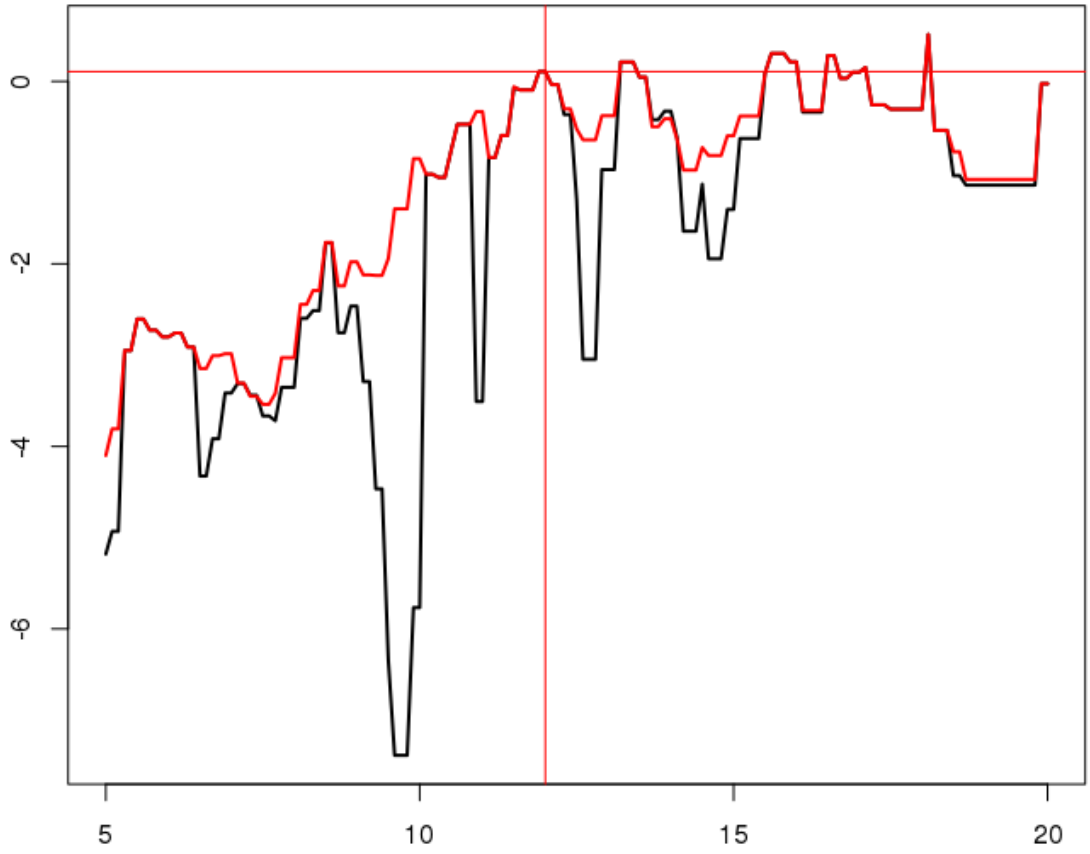

FIGURE 4.23 – Cumuls de précipitations : Estimation de l'erreur d'extrapolation associée à l'approximation Weissman en fonction du seuil en mm pour T = 1000 ans. En noir, $\hat{\varepsilon}_W(p_n;\alpha_n)$ donné par l'équation (4.11). En rouge, ce même estimateur dans lequel on a posé $\hat{\rho} = -1$.

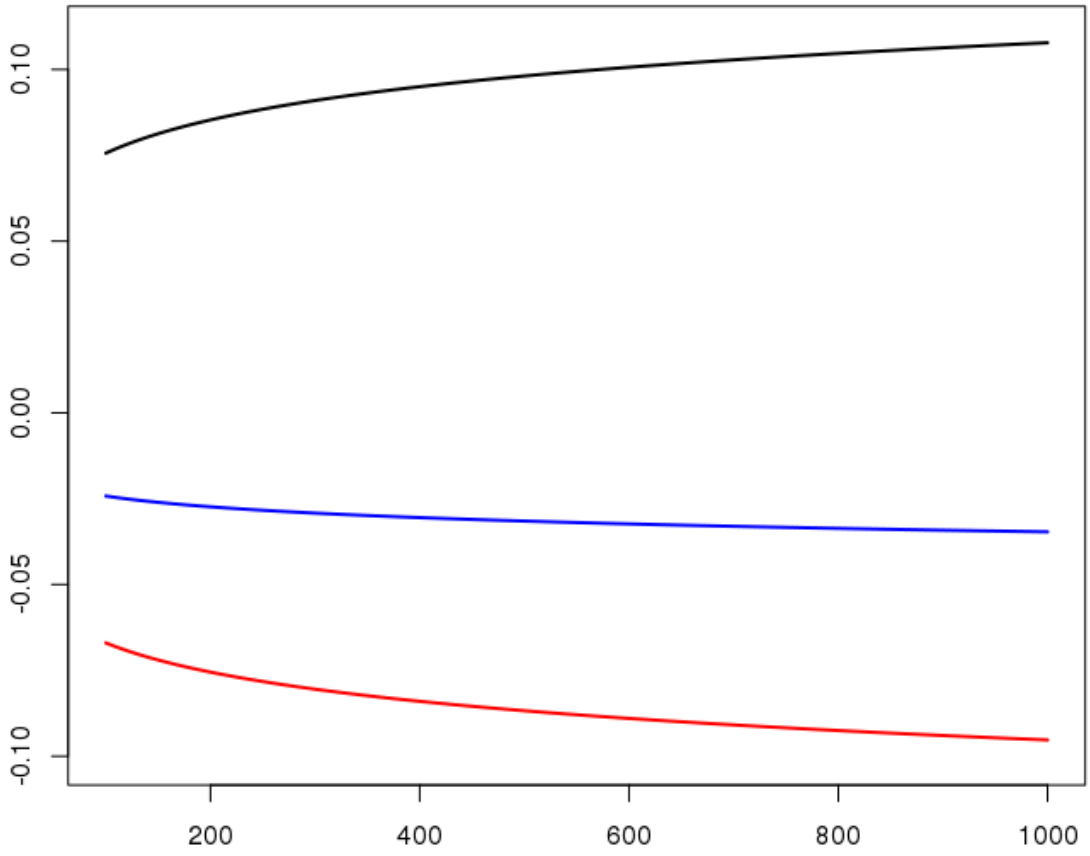

FIGURE 4.24 – Cumuls de précipitations : Estimation de $\tilde{\varepsilon}_W(p_n;\alpha_n)$ (voir équation (4.8)) via l'estimateur décrit équation (4.11) dans lequel on a posé $\hat{\rho} = -1$, et ce en fonction de la période de retour T, pour u = 12mm (en noir), u = 12.2mm (en bleu) et u = 11.8mm (en rouge).

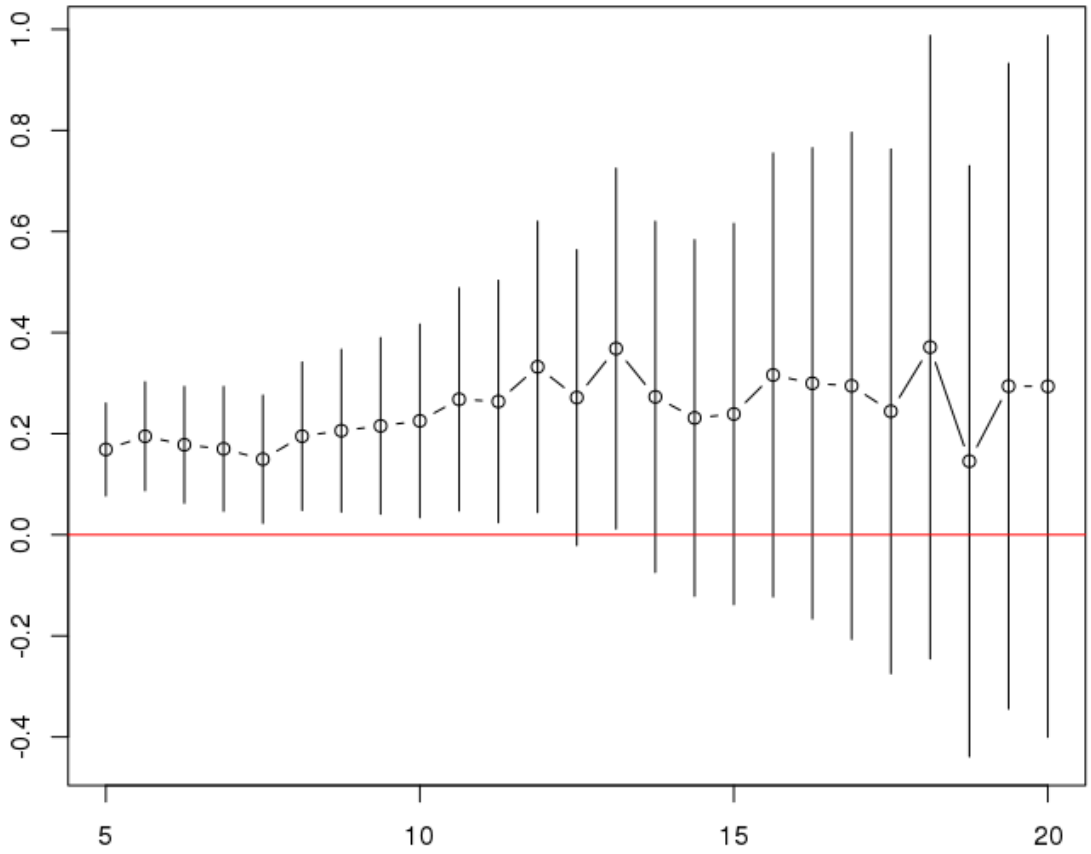

FIGURE 4.25 – Cumuls de précipitations : Estimation de l'indice des valeurs extrêmes par maximum de vraisemblance en fonction du seuil ainsi qu'un intervalle de confiance à 95%.

Conclusion et perspectives

L'objectif de cette thèse était double :

1. Tout d'abord, le but était de répondre aux problématiques d'EDF, à savoir la quantification des limites d'extrapolation. Pour ce faire, nous nous sommes dans un premier temps attelés à développer un pan théorique concernant l'étude du comportement asymptotique de ce que nous avons appelé l'"erreur d'extrapolation". Ce pan théorique est une des contributions majeures de cette thèse et a abouti à un article soumis pour publication, ALBERT et collab. [2018b]. Puis, dans un second temps, nous avons proposé des estimateurs dédiés au modèle sous lequel nous nous sommes placés pour étudier l'erreur. En combinant ces travaux, nous avons alors proposé des estimateurs de l'erreur d'extrapolation, et ce dans plusieurs situations (F $\in$DA(Gumbel) et F $\in$DA(Fréchet)). Ces outils statistiques nous ont finalement permis de quantifier en pratique les limites d'extrapolation à partir d'un jeu de données réelles, constituant ainsi un nouvel outil dans la prévention des risques. Nous avons ainsi pu montrer, sur trois séries de mesures de variables environnementales, que des limites d'extrapolation existaient bel et bien, et qu'elles étaient plus ou moins restrictives en fonction de l'aléa climatique considéré.
2. Ensuite, le deuxième objectif de cette thèse était de contribuer à la théorie des valeurs extrêmes, en proposant de nouveaux estimateurs basés sur la littérature récente. Dans cette optique, nous avons fait la proposition d'un nouvel estimateur des quantiles extrêmes, en se basant sur un modèle très prometteur introduit et discuté par Cees de Valk dans une série d'articles, DE VALK [2016a,b]; DE VALK et CAI [2018]. Nous en avons alors prouvé les propriétés de normalité asymptotique et avons illustré son comportement pratique sur données réelles et simulées, tout en le comparant à l'estimateur proposé par Cees de Valk lui-même en 2018, voir DE VALK et CAI [2018]. L'estimateur des quantiles extrêmes proposé a finalement fait l'objet d'un article soumis pour publication, ALBERT et collab. [2018a].

Cette thèse offre par ailleurs de nombreuses perspectives de recherche, aussi bien théoriques que pratiques. Ces perspectives sont respectivement discutées en détails dans les paragraphes 2.3, 3.3 et 4.5 des Chapitres 2, 3 et 4. Nous en résumons ici les principales.

- Du côté pratique tout d'abord, une des priorités serait d'obtenir des résultats de normalité asymptotique des estimateurs de l'erreur d'extrapolation dans le cas ET et Weissman. Cela nous permettrait de faire figurer des intervalles de confiance asymptotique sur les graphes 4.13, 4.19 et 4.23 et ainsi faciliter les prises de décision.

  Proposer un estimateur consistant de θ_2 et tester d'autres estimateurs de ρ est également une des perspectives qui pourrait faciliter les prises de décision, ces paramètres servant à l'élaboration des estimateurs de l'erreur d'extrapolation. Pour proposer un estimateur alternatif de θ_2, l'idée serait de supposer que V est à variation régulière étendue d'ordre deux et de s'inspirer de l'estimateur de θ introduit par DE VALK et CAI [2018]. Pour ce qui est de ρ, d'autres estimateurs ont été proposés dans la littérature. Citons GOMES et collab. [2002], BEIRLANT et collab. [2002] ou encore E. DEME [2013].

  Enfin, toujours dans l'idée de faciliter les prises de décisions, une autre perspective serait de proposer une méthode permettant de choisir le seuil de manière automatique. En effet, les estimateurs de l'erreur d'extrapolation que nous avons développés dépendent de k_n le nombre de statistiques d'ordre que l'on considère dans l'échantillon, et donc du seuil. Proposer une méthode permettant un choix pratique du seuil permettrait d'extrapoler plus loin dans les queues de distribution. Pour ce faire, on pourrait s'inspirer du nombre optimal de statistiques d'ordre utilisées par l'estimateur de Hill pour l'indice des valeurs extrêmes. A ce sujet, citons DREES et KAUFMANN [1998], DRAISMA et collab. [1999], BEIRLANT et collab. [1999], MATTHYS et BEIRLANT [2000] ou encore DE HAAN et FERREIRA [2007], pages 77-82. Développer des outils graphiques comme la Figure 4.13 est aussi une piste envisagée pour répondre à cette problématique.
- Du côté théorique ensuite, où une première contribution serait d'étudier l'erreur d'extrapolation associée à d'autres approximations quantiles. En particulier, il serait intéressant de généraliser nos tra-

vaux à l'approximation GPD en général, donnée par (1.37). Il s'agirait alors d'adapter des études telles que celles proposées par DE HAAN et RESNICK [1996] et DE HAAN et FERREIRA [2007, Théorème 4.3.1] à nos hypothèses de modèle. Cela nous permettrait également de traiter tous les domaines d'attraction sous un même formalisme, y compris le domaine d'attraction de Weibull.

En ce qui concerne le Chapitre 2, on pourra par exemple s'intéresser à l'erreur d'estimation aléatoire, pendant de l'erreur d'extrapolation. Dans le cas de l'approximation Weissman, le comportement asymptotique de l'erreur d'estimation pourrait se déduire DE HAAN et FERREIRA [2007, Théorème 4.3.8]. Dans le cas de l'approximation ET, cela nécessiterait l'adaptation de résultats déjà existants dans la littérature : citons DE HAAN et ROOTZÉN [1993, Proposition 1] ou encore DIEBOLT et GIRARD [2003, Théorème 1].

Pour ce qui est de perspectives à plus long terme, nous pourrions également envisager de construire des tests d'hypothèses pour le modèle des lois à queue de type log-Weibull généralisé, basé sur le paramètre θ. Ces tests pourraient alors nous permettre de mieux caractériser les queues des lois vérifiant le modèle. A terme, il nous permettrait de pouvoir distinguer les lois à queue super lourde, vérifiant $\theta > 1$, des lois à queue lourde du domaine d'attraction de Fréchet, vérifiant $\theta = 1$, des lois de type Weibull tail, vérifiant $\theta = 0$, ou encore des lois à point terminal fini vérifiant $\theta < 0$. Citons FRAGA ALVES et GOMES [1996] et NEVES et FRAGA ALVES [2008] pour un résumé des différents tests permettant de juger de l'appartenance d'une loi à un domaine d'attraction.

Enfin, proposer un estimateur d'une faible probabilité d'occurence d'un évènement extrême multivarié dans le cadre du modèle introduit par DE VALK [2016b] constituerait une perspective particulièrement ambitieuse. Il conviendrait de prendre en compte la dépendance existant entre les évènements, voir BACRO et collab. [2010], FALK et MICHEL [2006] ou encore WADSWORTH et collab. [2017]; WADSWORTH et TAWN [2012] à ce sujet. Pour terminer, il faudrait alors comparer l'estimateur obtenu avec celui proposé par DE VALK [2016a].

Bibliographie

AARSSEN, K. et L. DE HAAN. 1994, «On the maximal life span of humans», *Mathematical Population Studies*, vol. 4, nº 4, p. 259–281. 21

ALBERT, C., A. DUTFOY, L. GARDES et S. GIRARD. 2018a, «An extreme quantile estimator for the log-generalized Weibull-tail model», URL https://hal.inria.fr/hal-01783929/. 86, 87, 167

ALBERT, C., A. DUTFOY et S. GIRARD. 2018b, «Asymptotic behavior of the extrapolation error associated with the estimation of extreme quantiles», URL https://hal.archives-ouvertes.fr/hal-01692544/. 44, 49, 167

ANDERSON, P. L. et M. M. MEERSCHAERT. 1998, «Modeling river flows with heavy tails», *Water Resources Research*, vol. 34, nº 9, p. 2271–2280. 21

ARNAUD, P., P. CANTET et Y. AUBERT. 2016, «Relevance of an at-site flood frequency analysis method for extreme events based on stochastic simulation of hourly rainfall», *Hydrological Sciences Journal*, vol. 61, nº 1, p. 36–49. 5

ARNAUD, P., P. CANTET et J. ODRY. 2017, «Uncertainties of flood frequency estimation approaches based on continuous simulation using data resampling», *Journal of Hydrology*, vol. 554, p. 360–369. 5

ARNAUD, P. et J. LAVABRE. 1999, «Nouvelle approche de la prédétermination des pluies extrêmes», *Comptes Rendus de l'Académie des Sciences-Series IIA-Earth and Planetary Science*, vol. 328, nº 9, p. 615–620. 5

BACRO, J.-N., L. BEL et C. LANTUÉJOUL. 2010, «Testing the independence of maxima : from bivariate vectors to spatial extreme fields», *Extremes*, vol. 13, nº 2, p. 155–175. 131, 168

BECHLER, A., L. BEL et M. VRAC. 2015, «Conditional simulations of the extremal t process : application to fields of extreme precipitation», *Spatial statistics*, vol. 12, p. 109–127. 6

BEIRLANT, J., M. BRONIATOWSKI, J. L. TEUGELS et P. VYNCKIER. 1995, «The mean residual life function at great age : Applications to tail estimation», *Journal of Statistical Planning and Inference*, vol. 45, nº 1-2, p. 21–48. 23

BEIRLANT, J., G. DIERCKX, Y. GOEGEBEUR et G. MATTHYS. 1999, «Tail index estimation and an exponential regression model», *Extremes*, vol. 2, nº 2, p. 177–200. 130, 167

BEIRLANT, J., G. DIERCKX, A. GUILLOU et C. STAĂRICAĂ. 2002, «On exponential representations of log-spacings of extreme order statistics», *Extremes*, vol. 5, nº 2, p. 157–180. 19, 147, 148, 152, 153, 167

BEIRLANT, J., Y. GOEGEBEUR, J. SEGERS et J. L. TEUGELS. 2006, *Statistics of extremes : theory and applications*, John Wiley & Sons. 6, 10, 34

BEIRLANT, J., J.-P. RAOULT et R. WORMS. 2003, «On the relative approximation error of the generalized Pareto approximation for a high quantile», *Extremes*, vol. 13, p. 335–360. 45

BEIRLANT, J. et J. L. TEUGELS. 1992, «Modeling large claims in non-life insurance», *Insurance : Mathematics and Economics*, vol. 11, nº 1, p. 17–29. 6, 23

BEL, L., J.-N. BACRO et C. LANTUÉJOUL. 2008, «Assessing extremal dependence of environmental spatial fields», *Environmetrics : The official journal of the International Environmetrics Society*, vol. 19, nº 2, p. 163–182. 6

BERRED, M. 1991, «Record values and the estimation of the Weibull tail-coefficient», *Comptes-Rendus de l'Académie des Sciences*, vol. 312, nº 12, p. 943–946. 23

BINGHAM, N., C. GOLDIE et J. TEUGELS. 1987, *Regular Variation, Encyclopedia of Mathematics and its application*, vol. 27, Cambridge University Press. 16, 17, 18

BREIMAN, L., C. STONE et C. KOOPERBERG. 1990, «Robust confidence bounds for extreme upper quantiles», *Journal of Statistical Computation and Simulation*, vol. 37, p. 127–149. 31

BRONIATOWSKI, M. 1993, «On the estimation of the Weibull tail coefficient», *Journal of Statistical Planning and Inference*, vol. 35, nº 3, p. 349–365. 23

CASTILLO, E., J. GALAMBOS et J. M. SARABIA. 1989, «The selection of the domain of attraction of an extreme value distribution from a set of data», dans *Extreme Value Theory*, vol. 51, Springer, p. 181–190. 131

CERESETTI, D., E. URSU, J. CARREAU, S. ANQUETIN, J.-D. CREUTIN, L. GARDES, S. GIRARD et G. MOLINIE. 2012, «Evaluation of classical spatial-analysis schemes of extreme rainfall», *Natural Hazards and Earth System Sciences*, vol. 12, p. 3229–3240. 6

COLES, S., J. BAWA, L. TRENNER et P. DORAZIO. 2001, *An introduction to statistical modeling of extreme values*, vol. 208, Springer. 6, 10, 40, 45

COLES, S., L. R. PERICCHI et S. SISSON. 2003, «A fully probabilistic approach to extreme rainfall modeling», *Journal of Hydrology*, vol. 273, nº 1-4, p. 35–50. 6, 24

COLES, S. G. et J. A. TAWN. 1996, «A Bayesian analysis of extreme rainfall data», *Applied statistics*, vol. 45, nº 4, p. 463–478. 6

DEGEN, M., P. EMBRECHTS et D. D. LAMBRIGGER. 2007, «The quantitative modeling of operational risk : between g-and-h and EVT», *ASTIN Bulletin : The Journal of the IAA*, vol. 37, nº 2, p. 265–291. 6

DEHEUVELS, P., E. HAEUSLER et D. M. MASON. 1988, «Almost sure convergence of the Hill estimator», dans *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 104, Cambridge University Press, p. 371–381. 34

DEKKERS, A. L., J. H. EINMAHL et L. DE HAAN. 1989, «A moment estimator for the index of an extreme-value distribution», *The Annals of Statistics*, vol. 17, nº 4, p. 1833–1855. 32, 33, 87, 147

DIEBOLT, J., L. GARDES, S. GIRARD et A. GUILLOU. 2008a, «Bias-reduced estimators of the Weibull tail-coefficient», *Test*, vol. 17, nº 2, p. 311–331. 23

DIEBOLT, J., M. GARRIDO et S. GIRARD. 2003, «Asymptotic normality of the ET method for extreme quantile estimation. application to the ET test», *Comptes-Rendus de l'Académie des Sciences*, vol. 337, p. 213–218. 131

DIEBOLT, J. et S. GIRARD. 2003, «A note on the asymptotic normality of the ET method for extreme quantile estimation», *Statistics & Probability Letters*, vol. 62, nº 4, p. 397–405. 45, 80, 168

DIEBOLT, J., A. GUILLOU, P. NAVEAU et P. RIBEREAU. 2008b, «Improving probability-weighted moment methods for the generalized extreme value distribution», *REVSTAT-Statistical Journal*, vol. 6, nº 1, p. 33–50. 29

DIEBOLT, J., A. GUILLOU et I. RACHED. 2004, «A new look at probability-weighted moments estimators», *Comptes-Rendus de l'Académie des Sciences*, vol. 338, nº 8, p. 629–634. 32

DIEBOLT, J., A. GUILLOU et I. RACHED. 2007, «Approximation of the distribution of excesses through a generalized probability-weighted moments method», *Journal of Statistical Planning and Inference*, vol. 137, nº 3, p. 841–857. 32

DITLEVSEN, O. 1994, «Distribution arbitrariness in structural reliability», *Structural Safety and Reliability*, p. 1241–1247. 6

DOMBRY, C. 2015, «Existence and consistency of the maximum likelihood estimators for the extreme value index within the block maxima framework», *Bernoulli*, vol. 21, nº 1, p. 420–436. 28

DRAISMA, G., L. DE HAAN, L. PENG et T. T. PEREIRA. 1999, «A bootstrap-based method to achieve optimality in estimating the extreme-value index», *Extremes*, vol. 2, nº 4, p. 367–404. 130, 167

DREES, H., A. FERREIRA et L. DE HAAN. 2004, «On maximum likelihood estimation of the extreme value index», *Annals of Applied Probability*, vol. 14, nº 3, p. 1179–1201. 31

DREES, H. et E. KAUFMANN. 1998, «Selecting the optimal sample fraction in univariate extreme value estimation», *Stochastic Processes and their Applications*, vol. 75, nº 2, p. 149–172. 130, 167

DUTFOY, A., S. PAREY et N. ROCHE. 2014, «Multivariate extreme value theory-A tutorial with applications to hydrology and meteorology», *Dependence Modeling*, vol. 2, nº 1, p. 30–48. 6

E. DEME, L. G. . S. G. 2013, «On the estimation of the second order parameter for heavy-tailed distributions», *REVSTAT-Statistical Journal*, vol. 11, nº 3, p. 277–299. 19, 167

EINMAHL, J. H. et J. R. MAGNUS. 2008, «Records in athletics through extreme-value theory», *Journal of the American Statistical Association*, vol. 103, nº 484, p. 1382–1391. 21

EL METHNI, J., L. GARDES, S. GIRARD et A. GUILLOU. 2012, «Estimation of extreme quantiles from heavy and light tailed distributions», *Journal of Statistical Planning and Inference*, vol. 142, nº 10, p. 2735–2747. 6, 21, 23, 35, 36, 37

EMBRECHTS, P. 2000, *Extremes and integrated risk management*, Risk Books. 24

EMBRECHTS, P., C. KLÜPPELBERG et T. MIKOSCH. 2013, *Modelling extremal events : for insurance and finance*, vol. 33, Springer Science & Business Media. 6, 12, 87

EMBRECHTS, P., S. I. RESNICK et G. SAMORODNITSKY. 1999, «Extreme value theory as a risk management tool», *North American Actuarial Journal*, vol. 3, nº 2, p. 30–41. 6

EVIN, G., J. BLANCHET, E. PAQUET, F. GARAVAGLIA et D. PENOT. 2016, «A regional model for extreme rainfall based on weather patterns subsampling», *Journal of Hydrology*, vol. 541, p. 1185–1198. 5, 153

FALK, M. 1995, «Some best parameter estimates for distributions with finite endpoint», *Statistics : A Journal of Theoretical and Applied Statistics*, vol. 27, nº 1-2, p. 115–125. 21

FALK, M. et R. MICHEL. 2006, «Testing for tail independence in extreme value models», *Annals of the Institute of Statistical Mathematics*, vol. 58, nº 2, p. 261–290. 131, 168

FERREIRA, A. et L. D. HAAN. 2015, «On the block maxima method in extreme value theory : PWM estimators», *The Annals of Statistics*, vol. 43, nº 1, p. 276–298. 29

FISHER, R. A. et L. H. C. TIPPETT. 1928, «Limiting forms of the frequency distribution of the largest or smallest member of a sample», dans *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 24, Cambridge University Press, p. 180–190. 6, 11

FRAGA ALVES, M. et M. I. GOMES. 1996, «Statistical choice of extreme value domains of attraction—a comparative analysis», *Communications in Statistics-Theory and Methods*, vol. 25, nº 4, p. 789–811. 130, 168

FRÉCHET, M. 1927, «Sur la loi de probabilité de l'écart maximum», dans *Annales de la société Polonaise de Mathématique*, vol. 6, p. 93–116. 12

GALAMBOS, J. 1977, «The asymptotic theory of extreme order statistics», dans *The Theory and Applications of Reliability with Emphasis on Bayesian and Nonparametric Methods*, Elsevier, p. 151–164. 22

GARDES, L. et S. GIRARD. 2005, «Estimating extreme quantiles of Weibull tail distributions», *Communications in Statistics—Theory and Methods*, vol. 34, nº 5, p. 1065–1080. 23, 35

GARDES, L. et S. GIRARD. 2006, «Comparison of Weibull tail-coefficient estimators», *REVSTAT - Statistical Journal*, vol. 4, nº 2, p. 163–188. 23

GARDES, L. et S. GIRARD. 2010, «Conditional extremes from heavy-tailed distributions : An application to the estimation of extreme rainfall return levels», *Extremes*, vol. 13, nº 2, p. 177–204. 6, 21

GARDES, L. et S. GIRARD. 2013, «Estimation de quantiles extrêmes pour les lois à queue de type Weibull : une synthèse bibliographique», *Journal de la Société Française de Statistique*, vol. 154, nº 2, p. 98–118. 23

GARDES, L., S. GIRARD et A. GUILLOU. 2011, «Weibull tail-distributions revisited : a new look at some tail estimators», *Journal of Statistical Planning and Inference*, vol. 141, nº 1, p. 429–444. 23, 36, 37, 38

GAUME, J., N. ECKERT, G. CHAMBON, M. NAAIM et L. BEL. 2013, «Mapping extreme snowfalls in the French Alps using max-stable processes», *Water Resources Research*, vol. 49, nº 2, p. 1079–1098. 6

GIRARD, S. 2004, «A Hill type estimator of the Weibull tail-coefficient», *Communications in Statistics-Theory and Methods*, vol. 33, nº 2, p. 205–234. 23, 35

GIRARD, S. et J. DIEBOLT. 1999, «Consistency of the ET method and smooth variations», *Comptes-Rendus de l'Académie des Sciences*, vol. 329, nº 9, p. 821–826. 45

GIRARD, S., A. GUILLOU et G. STUPFLER. 2012, «Estimating an endpoint with high order moments in the Weibull domain of attraction», *Statistics & Probability Letters*, vol. 82, nº 12, p. 2136–2144. 21

GNEDENKO, B. 1943, «Sur la distribution limite du terme maximum d'une serie aléatoire», *Annals of Mathematics*, vol. 44, nº 3, p. 423–453. 6, 11, 12, 21

GOEGEBEUR, Y., J. BEIRLANT et T. DE WET. 2010, «Generalized kernel estimators for the Weibull-tail coefficient», *Communications in Statistics-Theory and Methods*, vol. 39, nº 20, p. 3695–3716. 23

GOMES, M. I. 1982, «A note on statistical choice of extremal models», dans *Actas de las IX Jornadas hispano-lusas : Salamanca 12-16 abril 1982*, vol. 2, Universidad de Salamanca, p. 653–656. 131

GOMES, M. I. et A. GUILLOU. 2015, «Extreme value theory and statistics of univariate extremes : a review», *International Statistical Review*, vol. 83, nº 2, p. 263–292. 6, 10

GOMES, M. I., L. DE HAAN et L. PENG. 2002, «Semi-parametric estimation of the second order parameter in statistics of extremes», *Extremes*, vol. 5, nº 4, p. 387–414. 19, 147, 148, 152, 153, 167

GOMES, M. I. et D. PESTANA. 2007, «A sturdy reduced-bias extreme quantile (VaR) estimator», *Journal of the American Statistical Association*, vol. 102, nº 477, p. 280–292. 45

GREENWOOD, J. A., J. M. LANDWEHR, N. C. MATALAS et J. R. WALLIS. 1979, «Probability weighted moments : definition and relation to parameters of several distributions expressable in inverse form», *Water Resources Research*, vol. 15, nº 5, p. 1049–1054. 29

GUMBEL, E. J. 1941, «The return period of flood flows», *The Annals of Mathematical Statistics*, vol. 12, nº 2, p. 163–190. 23

GUMBEL, E. J. 1954, «Statistical theory of extreme values and some practical applications», *NBS Applied Mathematics Series*, vol. 33. 6, 23

GUMBEL, E. J. 1958, «Statistics of extremes», *Columbia Univ. press, New York*, vol. 247. 12, 23, 27

DE HAAN, L. 1990, «Fighting the arch–enemy with mathematics», *Statistica Neerlandica*, vol. 44, nº 2, p. 45–68. 6, 23

DE HAAN, L. et A. FERREIRA. 2007, *Extreme value theory : an introduction*, Springer Science & Business Media. 6, 10, 19, 20, 21, 22, 23, 25, 33, 34, 80, 82, 87, 130, 136, 152, 167, 168

DE HAAN, L. et S. RESNICK. 1996, «Second-order regular variation and rates of convergence in extreme-value theory», *The Annals of Probability*, p. 97–124. 82, 168

DE HAAN, L. et S. RESNICK. 1998, «On asymptotic normality of the Hill estimator», *Stochastic Models*, vol. 14, nº 4, p. 849–866. 34

DE HAAN, L. et H. ROOTZÉN. 1993, «On the estimation of high quantiles», *Journal of Statistical Planning and Inference*, vol. 35, nº 1, p. 1–13. 45, 80, 168

HAEUSLER, E. et J. L. TEUGELS. 1985, «On asymptotic normality of Hill's estimator for the exponent of regular variation», *The Annals of Statistics*, vol. 13, nº 2, p. 743–756. 34

HALL, P. 1982, «On estimating the endpoint of a distribution», *The Annals of Statistics*, vol. 10, nº 2, p. 556–568. 21

HASOFER, A. M. et Z. WANG. 1992, «A test for extreme value domain of attraction», *Journal of the American Statistical Association*, vol. 87, nº 417, p. 171–177. 131

HILL, B. 1975, «A simple general approach to inference about the tail of a distribution», *The Annals of Statistics*, vol. 3, nº 5, p. 1163–1174. 32, 34

HOSKING, J. R. 1984, «Testing whether the shape parameter is zero in the generalized extreme-value distribution», *Biometrika*, vol. 71, nº 2, p. 367–374. 131

HOSKING, J. R. et J. R. WALLIS. 1987, «Parameter and quantile estimation for the generalized Pareto distribution», *Technometrics*, vol. 29, nº 3, p. 339–349. 31, 32

HOSKING, J. R., J. R. WALLIS et E. F. WOOD. 1985, «Estimation of the generalized extreme-value distribution by the method of probability-weighted moments», *Technometrics*, vol. 27, nº 3, p. 251–261. 28, 29

JAGGER, T. H. et J. B. ELSNER. 2006, «Climatology models for extreme hurricane winds near the United States», *Journal of Climate*, vol. 19, nº 13, p. 3220–3236. 24

KATZ, R. W., M. B. PARLANGE et P. NAVEAU. 2002, «Statistics of extremes in hydrology», *Advances in Water Resources*, vol. 25, nº 8-12, p. 1287–1304. 6, 24, 29

LEADBETTER, M. R. 1983, «Extremes and local dependence in stationary sequences», *Probability Theory and Related Fields*, vol. 65, nº 2, p. 291–306. 40, 41

MACLEOD, A. J. 1989, «A remark on algorithm AS 215 : Maximum-likelihood estimation of the parameters of the generalized extreme-value distribution», *Applied Statistics*, vol. 38, nº 1, p. 198–199. 28

MAROHN, F. 1998, «Testing the Gumbel hypothesis via the POT-method», *Extremes*, vol. 1, nº 2, p. 191–213. 131

MASON, D. M. 1982, «Laws of large numbers for sums of extreme values», *The Annals of Probability*, vol. 10, nº 3, p. 754–764. 34

MATTHYS, G. et J. BEIRLANT. 2000, «Adaptive threshold selection in tail index estimation», *Extremes and Integrated Risk Management*, p. 37–49. 130, 167

MCNEIL, A. J., R. FREY, P. EMBRECHTS et collab.. 2005, *Quantitative risk management : Concepts, techniques and tools*, vol. 3, Princeton university press Princeton. 24

METHNI, J. E., L. GARDES et S. GIRARD. 2014, «Non-parametric estimation of extreme risk measures from conditional heavy-tailed distributions», *Scandinavian Journal of Statistics*, vol. 41, nº 4, p. 988–1012. 6, 21, 150

MIGNOLA, G. et R. UGOCCIONI. 2005, «Tests of extreme value theory», *Operational Risk*, vol. 6, nº 10, p. 32–35. 6

MUIR, L. R. et A. EL-SHAARAWI. 1986, «On the calculation of extreme wave heights : a review», *Ocean Engineering*, vol. 13, nº 1, p. 93–118. 24

NEVES, C. et M. I. FRAGA ALVES. 2008, «Testing extreme value conditions—an overview and recent approaches», *REVSTAT - Statistical Journal*, vol. 6, nº 1, p. 83–100. 130, 168

NEVES, C., J. PICEK et M. F. ALVES. 2006, «The contribution of the maximum to the sum of excesses for testing max-domains of attraction», *Journal of Statistical Planning and Inference*, vol. 136, nº 4, p. 1281–1301. 131

PICKANDS, J. 1975, «Statistical inference using extreme order statistics», *The Annals of Statistics*, vol. 3, p. 119–131. 6, 11, 14

PRESCOTT, P. et A. WALDEN. 1980, «Maximum likelihood estimation of the parameters of the generalized extreme-value distribution», *Biometrika*, vol. 67, nº 3, p. 723–724. 28

PRESCOTT, P. et A. WALDEN. 1983, «Maximum likelihood estimation of the parameters of the three-parameter generalized extreme-value distribution from censored samples», *Journal of Statistical Computation and Simulation*, vol. 16, nº 3-4, p. 241–250. 28

RESNICK, S. 1987, *Extreme values, regular variation, and point processes*, Springer, New York. 6, 10

ROOTZÉN, H. et N. TAJVIDI. 2001, «Can losses caused by wind storms be predicted from meteorological observations?», *Scandinavian Actuarial Journal*, vol. 2001, nº 2, p. 162–175. 6

SMITH, R. L. 1985, «Maximum likelihood estimation in a class of nonregular cases», *Biometrika*, vol. 72, nº 1, p. 67–90. 28

DE VALK, C. 2016a, «Approximation and estimation of very small probabilities of multivariate extreme events», *Extremes*, vol. 19, nº 4, p. 687–717. 87, 131, 167, 168

DE VALK, C. 2016b, «Approximation of high quantiles from intermediate quantiles», *Extremes*, vol. 19, nº 4, p. 661–686. 7, 35, 37, 46, 86, 87, 130, 131, 167, 168

DE VALK, C. et J.-J. CAI. 2018, «A high quantile estimator based on the log-generalized Weibull tail limit», *Econometrics and Statistics*, vol. 6, p. 107–128. 7, 35, 37, 38, 86, 87, 130, 143, 146, 152, 157, 167

VAN DER VAART, A. W. et J. A. WELLNER. 1996, «Weak convergence», dans *Weak convergence and empirical processes*, Springer, p. 16–28. 24

VAN MONTFORT, M. et J. WITTER. 1985, «Testing exponentiality against generalised Pareto distribution», *Journal of Hydrology*, vol. 78, nº 3-4, p. 305–315. 131

WADSWORTH, J., J. A. TAWN, A. DAVISON et D. M. ELTON. 2017, «Modelling across extremal dependence classes», *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, vol. 79, nº 1, p. 149–175. 131, 168

WADSWORTH, J. L. et J. A. TAWN. 2012, «Dependence modelling for spatial extremes», *Biometrika*, vol. 99, nº 2, p. 253–272. 131, 168

WEIBULL, W. 1951, «Wide applicability», *Journal of Applied Mechanics*, vol. 103, nº 730, p. 293–297. 12

WEISSMAN, I. 1978, «Estimation of parameters and large quantiles based on the k largest observations», *Journal of the American Statistical Association*, vol. 73, nº 364, p. 812–815. 33, 34

WILLEMS, P. et A. GUILLOU. 2006, «Application de la théorie des valeurs extrêmes en hydrologie», *Revue de statistique appliquée*, vol. 54, nº 2, p. 5–31. 6

ZHOU, C. 2009, «Existence and consistency of the maximum likelihood estimator for the extreme value index», *Journal of Multivariate Analysis*, vol. 100, nº 4, p. 794–815. 28, 31

ZHOU, C. 2010, «The extent of the maximum likelihood estimator for the extreme value index», *Journal of Multivariate Analysis*, vol. 101, nº 4, p. 971–983. 28