République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ MOHAMED KHIDER, BISKRA

FACULTÉ des SCIENCES EXACTES et des SCIENCES de la NATURE et de la VIE

DÉPARTEMENT DE MATHÉMATIQUES

Mémoire présenté en vue de l'obtention du Diplôme **:**

MASTER en Mathématiques

Option : **Statistique**

Par

BOUACIDA Hocine

Titre :

Choix du nombre optimal de statistiques d'ordre extrêmes

Membres du Comité d'Examen :

Dr.	BENAMEUR Sana	UMKB	Encadreuse
Dr.	YAHIA Djabrane	UMKB	Co-Encadreur
Pr.	MERAGHNI Djamel	UMKB	Président
Dr.	KHEIREDDINE Souraya	UMKB	Examinatrice

Juin 2018

Dedicace

Je dédie ce humble travail

A mes chers parents pour leur soutien, leur patience,

leur encouragement durant mon parcours scolaire.

A mes frères et mes sœurs ainsi à toute ma famille.

A tous mes amis,

et à l'ensemble des étudiants de la promotion master

LMD/MI de l'année 2017-2018

REMERCIEMENTS

Je tiens tout d'abord à remercier Allah le tout puissant et miséricordieux, qui m'a donné la force et la patience d'accomplir ce travail.

En second lieu, je tiens à remercier mes encadreurs Madame ***Benameur Sana*** *et Monsieur* ***Yahia Djabrane****, pour leurs précieux conseils et leurs aide durant toute la période du travail.*

Je suis conscient de l'honneur que nous a fait Monsieur ***Meraghni Djamel*** *en étant président du jury et Madame* ***Kheireddine Souraya*** *d'avoir accepté d'examiner ce travail.*

Mes plus vifs remerciements vont à Monsieur ***Meraghni Djamel****, Monsieur* ***Necir Abdelhakim*** *et Monsieur* ***Hafayed Mokhtar*** *pour toute l'aide qu'ils m'ont apporté.*

Mes remerciements s'étendent également à tous mes enseignants du département de mathématiques à l'université de Biskra durant les années d'étude, ma famille et mes amis pour leurs aides et leurs encouragements.

Enfin, je tiens à remercier toutes les personnes qui ont participé de près ou de loin à la réalisation de ce travail.

Table des matières

Remerciements ii

Table des matières iii

Liste des figures v

Introduction 1

1 Introduction à la théorie des valeurs extrêmes 3

1.1 Définitions et caractéristiques de bases 3

1.1.1 Loi des grands nombres 3

1.1.2 Théorème central limite 4

1.2 Statistiques d'ordre 6

1.2.1 Distribution de la k^{ième}statistique d'ordre 7

1.2.2 Distribution jointe de deux statistiques d'ordre 7

1.2.3 Densité jointe de n statistique d'ordre 8

1.3 Distribution des valeurs extrêmes 9

1.3.1 Loi max stable 10

1.3.2 Distribution des valeurs extrêmes généralisé (GEV) 10

1.3.3 Domaine d'attraction 12

1.3.4 Distribution conditionnelle des excès 18

1.3.5 Distribution de Pareto généralisé (GPD) 19

1.4 Estimation de l'indice des valeurs extrêmes γ 21
1.4.1 Estimateur de Pickands 21
1.4.2 Estimateur de Hill 22

2 Détermination du nombre de statistiques d'ordre extrêmes 24

2.1 Méthode Graphique 24
2.1.1 Méthode Sum-Plot 25
2.2 Minimisation de l'erreur quadratique moyenne asymptotique 26
2.2.1 Résultat pour l'estimateur de Hill 26
2.3 Procédures adaptatives 27
2.3.1 Approche de Hall et Welsh 27
2.3.2 Approche de Bootstrap 28
2.3.3 Approche Séquentielle 30
2.3.4 Approche de Cheng et Peng 31

Conclusion 34

Bibliographie 35

Annexe B : Abréviations et Notations 37

Table des figures

1.1 Distributions et densités standard des valeurs extrêmes 12

1.2 Distribution et densité standard de GPD 21

1.3 Estimateur de Pickands, avec un intervalle de confiance de niveau 95%, pour l'EVI de la distribution uniforme standard ($\gamma = -1$) basé sur 100 échantillons de 3000 observations. 22

1.4 Estimateur de Hill,avec un intervalle de confiance de niveau 95%, pour l'EVI de la distribution Pareto standard ($\gamma = 1$) basé sur 100 échantillons de 3000 observations. 23

2.1 Méthode Graphique du choix de k. 24

Introduction

La théorie des valeurs extrêmes (TVE) est une vaste théorie dont le but est d'étudier les événements rares (les événements dont la probabilité d'apparition est faible). L'objectif de cette théorie est la connaissance de la loi asymptotique des extrêmes d'un échantillon, nommée loi des valeurs extrêmes. La loi des valeurs extrêmes dépende d'un paramètre réel inconnu γ appelé l'indice de queue ou l'indice des valeurs extrêmes (IVE), cet indice est un élément fondamental permettant de contrôler la lourdeur de la queue de distribution étudiée.

L'estimation de l'indice des valeurs extrêmes d'une distribution à queue lourde dépend fortement du choix du nombre de statistiques d'ordre extrêmes à utiliser dans cette estimation. Il est bien connu que la façon de choisir la valeur du nombre de statistiques d'ordre extrêmes que l'on note k est toujours un problème difficile même si la forme d'estimation a été déterminée. En effet, sélectionner une bonne valeur de k est une tâche sensible. Lorsque k est petit la variance de l'estimateur est grande et l'utilisation de grande valeur de k introduit un grand biais dans l'estimation. L'équilibrage de ces composants (la variance et le biais) est important dans les applications.

Dans ce mémoire, nous nous intéressons à répondre à la question : comment choisir le nombre de statistiques d'ordre extrêmes impliqué dans le calcul de l'estimateur de L'IVE. Pour cela, l'organisation de ce travail est la suivante :

Chapitre 1 : Introduction à la théorie des valeurs extrêmes : dans ce chapitre nous rappelons la définition des statistiques d'ordre, le comportement asymptotique du maximum

d'un échantillon. On donne ensuite des résultats décrivant les limites possibles de la loi du maximum d'un échantillon. Pour cela on a introduit deux théorèmes essentiels à la compréhension de la TVE : théorème de Fisher-Tippet et celui de Balkema-de Haan-Pickands. On s'intéresse à l'estimation semi-paramétrique de IVE.

Chapitre 2 : Détermination du nombre de statistiques d'ordre extrêmes : Le problème du choix du nombre de statistiques d'ordre extrêmes a reçu beaucoup d'attention dans la littérature. Nous abordons dans ce chapitre certaines des méthodes proposées pour équilibrer entre deux vices : Biais et variance, afin d'obtenir un nombre optimal de statistiques d'ordre qui localise où la queue de distribution (vraiment) commence.

Chapitre 1

Introduction à la théorie des valeurs extrêmes

La théorie des valeurs extrêmes vient en complément de la statistique classique où il est généralement question d'étudier le comportement de variables aléatoires autour de leurs moyennes, il s'agit dans cette théorie de caractériser le comportement des queues de distribution à l'aide de modèles permettant un bon ajustement au-delà du maximum de l'échantillon. Nous introduisons tout d'abord quelques caractéristiques de bases sur les théorèmes limite et les statistiques d'ordre, puis nous présentons les notions essentielles dans la théorie des valeurs extrêmes (TVE) telles que les différentes distributions des valeurs extrêmes, domaine d'attraction et estimation semi-paramétrique de l'indice des valeurs extrêmes.

1.1 Définitions et caractéristiques de bases

1.1.1 Loi des grands nombres

Soit $(X_i)_{i\geq 1}$ une suite des variables aléatoires (va's) réelles indépendantes et identiquement distribuées (i.i.d) intégrables d'espérance commun $\mathbb{E}(X)$, les lois des grands nombres

portent sur le comportement de la moyenne empirique $\bar{X}_n$ quand $n \longrightarrow \infty$ où

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\text{Loi faible des grands nombres : } \bar{X}_n \xrightarrow{P} \mathbb{E}(X), \text{ quand } n \longrightarrow \infty. \tag{1.1}$$

$$\text{Loi forte des grands nombres : } \bar{X}_n \xrightarrow{p.s} \mathbb{E}(X), \text{ quand } n \longrightarrow \infty. \tag{1.2}$$

1.1.2 Théorème central limite

Le résultat suivant explique la place fondamentale des variables aléatoire gaussienne en probabilité et en statistique.

Théorème 1.1.1 (Théorème Central limite) *Soit $(X_i)_{i\geq 1}$ une suite de va's réelles i.i.d, telle que $\mathbb{E}(X_1^2) < \infty$, et $\bar{X}_n$ sa moyen empirique alors*

$$\frac{\sqrt{n}(\bar{X}_n - \mathbb{E}(X_1))}{\sqrt{var(X_1)}} \xrightarrow{d} \mathcal{N}(0,1), \text{ quand } n \longrightarrow \infty. \tag{1.3}$$

On peut aussi en déduire que la loi de $\sum_{i=1}^{n} X_i = n\bar{X}_n$ est proche de $\mathcal{N}(n\mathbb{E}(X_1), n\sigma^2)$, où $\sigma^2 = var(X_1)$.

Preuve. Voir Jean-François Delmas [8]. ■

Définition 1.1.1 (Distribution empirique) *la fonction de répartition empirique F_n*

associée a une échantillon $(X_1, \cdots, X_n)$ *est :*

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(X_i \leq x) \tag{1.4}$$
$$= \begin{cases} 0 & si \quad x < X_{1.n} \\ \frac{k}{n} & si \quad X_{k,n} \leq x < X_{k+1,n} \quad , \ pour\ 1 \leq k \leq n, \\ 1 & si \quad x \geq X_{n,n} \end{cases}$$

voir la page 6

Définition 1.1.2 (Inverse généralisé) *On appelle L'inverse généralisé de* h *, l'application notée* $h^{\leftarrow}$ *telle que :*

$$h^{\leftarrow}(t) = \inf\{x \in \mathbb{R},\ h(x) \geq t\}. \tag{1.5}$$

Remarque 1.1.1 *La fonction de quantile de* F *c'est I'inverse généralisé et notée* $Q(t)$ *telle que*

$$Q(t) = F^{\leftarrow}(t) = \inf\{x \in \mathbb{R},\ F(x) \geq t\},\ 0 < t < 1. \tag{1.6}$$

Définition 1.1.3 (fonction de quantile empirique) *On appelle fonction de quantile empirique de distribution* F *et notée* $Q_n(t)$ *telle que :*

$$Q_n(t) = F_n^{\leftarrow}(t) = \inf\{x \in \mathbb{R}, F_n(x) \geq t\},\ 0 < t < 1 \tag{1.7}$$
$$= X_{i,n},\ \frac{i-1}{n} < t \leq \frac{i}{n},$$

où $F(x) = P(X \leq x)$.

Définition 1.1.4 (fonction de queue) *On appelle fonction de queue ou (de survie)* $\bar{F}$*, la fonction définie par*

$$\bar{F}(x) = 1 - F(x). \tag{1.8}$$

Remarque 1.1.2 *Pour plus de détails sur ces fonctions, voir Embrech et al. [10], ou bien Beirlant et al. [3].*

1.2 Statistiques d'ordre

Soit $X_1, \cdots, X_n$, n va i.i.d de densité de probabilité f et de fonction de distribution F. On appelle statistique d'ordre notées $X_{1,n}, \cdots, X_{n,n}$ les va's ordonnées comme suit

$$X_{1,n} \leq X_{2,n} \leq \cdots \leq X_{n,n},$$

telle que :

$$X_{1,n} = min(X_1, \cdots, X_n), \qquad X_{n,n} = max(X_1, \cdots, X_n).$$

La fonction de répartition de la statistique d'ordre du maximum $X_{n,n}$ est :

$$F_{X_{n,n}}(x) = P(X_{n,n} \leq x) = [F(x)]^n, \tag{1.9}$$

et sa densité :

$$f_{X_{n,n}}(x) = nf(x)F(x)^{n-1}, \text{ pour } x \in \mathbb{R}. \tag{1.10}$$

La fonction de répartition de la statistique d'ordre du minimum $X_{1,n}$ est :

$$F_{X_{1,n}}(x) = P(X_{1,n} \leq x) = 1 - [1 - F(x)]^n,$$

de même, sa densité est :

$$f_{X_{1,n}}(x) = nf(x)[1 - F(x)]^{n-1}, \text{ pour } x \in \mathbb{R}. \tag{1.11}$$

1.2.1 Distribution de la k$^{\text{ième}}$statistique d'ordre

La fonction de répartition de la k$^{\text{ième}}$statistique d'ordre est :

$$\begin{aligned} F_{X_{k,n}}(x) &= P(X_{k,n} \leq x) \\ &= P\{\text{au moins } k \text{ des } X_r \text{ sont inférieur à } x\} \\ &= \sum_{r=k}^{n} P\{\text{exactement } r \text{ de } X_1, \cdots, X_n \text{ sont inférieur à } x\} \\ &= \sum_{r=k}^{n} C_n^r [F(x)]^r [1-F(x)]^{n-r}, \ x \in \mathbb{R}, \end{aligned} \tag{1.12}$$

où

$$C_n^r = \frac{n!}{r!(n-r)!}.$$

Sa fonction densité est :

$$f_{X_{k,n}}(x) = \frac{n!}{(k-1)!(n-k)!} [F(x)]^{k-1} [1-F(x)]^{n-k} f(x), \ x \in \mathbb{R}. \tag{1.13}$$

1.2.2 Distribution jointe de deux statistiques d'ordre

Soit $X_{i,n}$ et $X_{j,n}$ deux statistiques d'ordre si $x \geq y$:

$$\begin{aligned} F_{(X_{i,n},\, X_{j,n})}(x,y) &= P(X_{i,n} \leq x, X_{j,n} \leq y) \\ &= P(X_{j,n} \leq y) \\ &= F_{X_{j,n}}(y), \end{aligned} \tag{1.14}$$

et si $x < y$:

$$\begin{aligned} F_{(X_{i,n},\, X_{j,n})}(x,y) &= P(X_{i,n} \leq x,\ X_{j,n} \leq y) \\ &= P\{\text{au moins } i \text{ de } X_1, \cdots, X_n \text{ sont inférieure à } x \text{ et} \\ &\quad \text{au moins } j \text{ de } X_1, \cdots, X_n \text{ sont inférieure à } y\} \\ &= \sum_{r=j}^{n} \sum_{s=i}^{r} P\{\text{exactement } s \text{ de } X_1, \cdots, X_n \text{ sont inférieure à } x \text{ et} \\ &\quad \text{exactement } r \text{ de } X_1, \cdots, X_n \text{ sont inférieure à } y\} \\ &= \sum_{r=j}^{n} \sum_{s=i}^{r} \frac{n!}{s!(r-s)!(n-r)!} [F(x)]^s [F(y) - F(x)]^{r-s} [1 - F(y)]^{n-r}, \end{aligned} \tag{1.15}$$

$$-\infty < x < y < +\infty,$$

et sa densité

$$f_{(X_{i,n},\, X_{j,n})}(x,y) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} [F(x)]^{i-1} f(x) [F(y) - F(x)]^{j-i-1} f(y) [1 - F(y)]^{n-j}, \tag{1.16}$$

$$-\infty < x < y < +\infty.$$

1.2.3 Densité jointe de n statistique d'ordre

Soit $X_1, \cdots, X_n$, n va i.i.d de densité de probabilité f, alors la densité de la statistique d'ordre $(X_{1,n} \leq X_{2,n} \leq \cdots \leq X_{n,n})$ est :

$$f_{(X_{i,n}, \cdots, X_{n,n})}(x_{1,\cdots,}x_n) = n! \prod_{i=1}^{n} f_{(x_i)},\ -\infty < x_1 < \cdots < x_n < +\infty. \tag{1.17}$$

Preuve. Les démonstrations détaillées sont données dans Arnold [1]. ■

1.3 Distribution des valeurs extrêmes

Théorème 1.3.1 (Fisher et Tippet(1928), Gnedenko(1943)) *Soit $(X_1, \cdots, X_n)$ est une suite des va's i.i.d si il existe deux suites normalisant $b_n \in \mathbb{R},\ a_n > 0$ alors :*

$$\lim_{n\to\infty} P[(\frac{M_n - b_n}{a_n} \leq x)] = \lim_{n\to\infty} F^n(a_n x + b_n) \longrightarrow G(x), \tag{1.18}$$

telle que G fonction de distribution non dégénérée et doit être qu' une de ces trois distribution suivantes :

1)Loi de Fréchet :

$$\Phi_\alpha(x) = \begin{cases} 0 & si \quad x \leq 0 \\ \exp(-x^{-\alpha}) & si \quad x > 0 \end{cases},\ \alpha > 0. \tag{1.19}$$

2)Loi de Weibull :

$$\Psi_\alpha(x) = \begin{cases} 1 & si \quad x > 0 \\ \exp(-(-x)^{\alpha}) & si \quad x \leq 0 \end{cases},\ \alpha > 0. \tag{1.20}$$

3)Loi de Gumbel :

$$\Lambda(x) = \exp(-e^{-x}),\ x \in \mathbb{R}. \tag{1.21}$$

Les suites $(a_n),\ (b_n)$ ne sont pas unique et les trois distribution $\Phi_\alpha,\ \Psi_\alpha,\ \Lambda$ sont les distributions des valeurs extrêmes standard.

Remarque 1.3.1 *Pour plus de détaille on pourra se référer à Embrech et al* (1997) *[10], Resnick* (1987) *[17].*

1.3.1 Loi max stable

Définition 1.3.1 *On dit que F est max stable si pour tout $n \geq 1$, $X_1, \cdots, X_n$ des va's i.i.d de loi F, $M_n = \max(X_1, \cdots, X_n)$, il existe suite $a_n > 0,\ b_n \in \mathbb{R}$ telle que :*

$$a_n^{-1}(M_n - b_n) \sim F.$$

Remarque 1.3.2 *Si $X_1, \cdots, X_n$ des va's i.i.d de loi Fréchet Φ_α il existe $a_n = n^{\frac{-1}{\alpha}},\ b_n = 0$ alors :*

$$n^{\frac{-1}{\alpha}} M_n \stackrel{d}{=} X.$$

Si $X_1, \cdots, X_n$ des va's i.i.d de loi Weibull Ψ_α il existe $a_n = n^{\frac{1}{\alpha}},\ b_n = 0$ alors :

$$n^{\frac{1}{\alpha}} M_n \stackrel{d}{=} X.$$

Si $X_1, \cdots, X_n$ des va's i.i.d de loi Gumbel Λ il existe $a_n = 1,\ b_n = \ln n$ alors :

$$M_n - \ln n \stackrel{d}{=} X.$$

Remarque 1.3.3 *D'après Embrech et al* (1997) *[10], les va's dans les trois domaines d'attraction de Fréchet, Gumbel, Weibull sont liées par la relation suivante pour $X > 0$*

$$X \sim \Phi_\alpha \Longleftrightarrow -X^{-1} \sim \Psi_\alpha \Longleftrightarrow \ln X^\alpha \sim \Lambda. \tag{1.22}$$

1.3.2 Distribution des valeurs extrêmes généralisé (GEV)

De représentation de Jenkinson-Von Mises, on peut ressembler les trois familles de lois en une seule famille paramétrique $G\gamma$ donnée par :

$$G\gamma(x) = \begin{cases} \exp\{-(1+\gamma x)^{-\frac{1}{\gamma}}\} & si \quad \gamma \neq 0,\ 1+\gamma x > 0 \\ \exp\{-\exp(-x)\} & si \qquad\qquad \gamma = 0 \end{cases}, \tag{1.23}$$

$\gamma \in \mathbb{R}$ c'est l'indice de queue ou l'indice des valeurs extrêmes, ce indice est un élément fondamental permettant de contrôler la lourdeur de la queue de la distribution. La fonction $G_\gamma$est appelée loi des valeurs extrêmes généralisé, voir Embrech et al (1997) [10].

Remarque 1.3.4 *Si $\gamma > 0$ on dit que F appartient au domaine d'attraction de Fréchet, ce domaine contient les fonctions de répartition à queue lourdes par exemple : Loi de Cauchy, Loi Student.*

Si $\gamma < 0$: on dit que F appartient au domaine d'attraction de Weibull, ce domaine contient la majorité des fonctions de répartition dont le point terminal est fini par exemple : Loi Uniform, Loi Bêta.

Si $\gamma = 0$: on dit que F appartient au domaine d'attraction de Gumbel, ce domaine contient les fonctions de répartition a décroissante exponentielle par exemple : Loi Normal, Loi Exponentielle, Loi Gamma.

Remarque 1.3.5 *pour GEV nous avons les correspondances suivantes :*

Fréchet : $\gamma = \alpha^{-1} > 0$.

Weibull : $\gamma = -\alpha^{-1} < 0$.

Gumbull : $\gamma = 0$.

Pour x non centrée et non réduite on écrit :

$$G_{\mu,\delta,\gamma}(x) = \begin{cases} \exp\{-(1+\gamma(\frac{x-\mu}{\delta}))^{-\frac{1}{\gamma}}\} & si \quad \gamma \neq 0,\ 1+\gamma(\frac{x-\mu}{\delta}) > 0 \\ \exp\{-\exp(-(\frac{x-\mu}{\delta}))\} & si \qquad \gamma = 0 \end{cases}, \tag{1.24}$$

avec $\mu \in \mathbb{R}$ le paramètre de position et $\delta > 0$ le paramètre d'échelle. La densité de $G_{\mu,\delta,\gamma}$ est $g_{\mu,\delta,\gamma}$ telle que :

$$g_{\mu,\delta,\gamma}(x) = \begin{cases} \frac{1}{\delta}(1+\gamma(\frac{x-\mu}{\delta}))^{-\frac{(1+\gamma)}{\gamma}} \exp\{-(1+\gamma(\frac{x-\mu}{\delta}))^{-\frac{1}{\gamma}} & si \quad \gamma \neq 0,\ 1+\gamma(\frac{x-\mu}{\delta}) > 0 \\ \frac{1}{\delta}\exp\{(-(\frac{x-\mu}{\delta}) - \exp(-(\frac{x-\mu}{\delta}))\} & si \qquad \gamma = 0 \end{cases}. \tag{1.25}$$

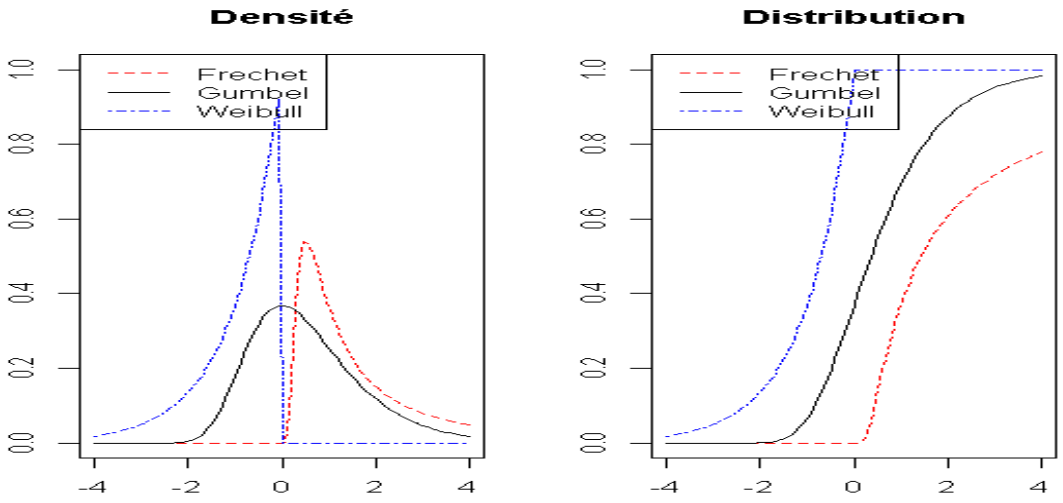


FIG. 1.1 – Distributions et densités standard des valeurs extrêmes

1.3.3 Domaine d'attraction

On appelle point terminal de la fonction de répartition F le point noté x_F telle que :

$$x_F = \sup\{x \in \mathbb{R} : \ F(x) < 1\}. \tag{1.26}$$

et on note aussi la fonction quantile de queue par

$$U(t) = F^{\leftarrow}(1 - \frac{1}{t}), \ t > 1. \tag{1.27}$$

Fonction à variation régulière

On dit que F est une fonction à variation régulière à l'infini si F est positive à l'infini et

$$\lim_{t\to\infty} \frac{F(tx)}{F(t)} = x^{\alpha}, \ x > 0, \tag{1.28}$$

et notée $F \in RV_{\alpha}, \ \alpha \in \mathbb{R}$, le nombre α appelé l'indice de variation régulière.

Remarque 1.3.6 *Soit L une fonction satisfait* 1.28 *avec* $\alpha = 0$ *appelé fonction a variation lente, (voir de Haan et Ferreira,* 2006*).*

Représentation de Karamata Si l est une à variation lente il existe des fonctions mesurables $c : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$ et $\varepsilon : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$ avec $\lim_{t\to\infty} c(t) = c_0 > 0$ et $\lim_{s\to\infty} \varepsilon(s) = 0$, pour $t > 0$

$$l(t) = c(t) \exp(\int_1^t \frac{\varepsilon(s)}{s} ds). \tag{1.29}$$

Remarque 1.3.7 *Soit g est une fonction à variation régulière a l'infini*($g \in RV_\alpha$) *si et seulement si pour tout* $t > 0$

$$g(t) = c(t) \exp(\int_1^t \frac{\varepsilon(s)}{s} ds),$$

où $\lim_{t\to\infty} c(t) = c_0 > 0,\ \lim_{s\to\infty} \varepsilon(s) = \alpha.$

i) Si la fonction c est constante, on dit que la fonction l est normalisé et dérivable telle que sa dérivée ζ *est pour* $t > 0$

$$\zeta(t) = \frac{\varepsilon(t) l(t)}{t}.$$

ii) Si $f_1 \in RV_{\alpha_1},\ f_2 \in RV_{\alpha_2}$ *alors :*

$$f_1 f_2 \in RV_{\max(\alpha_1,\alpha_2)},$$

iii) Si de plus $\lim_{t\to\infty} f_2(t) = \infty$ *alors :*

$$f_1 \circ f_2 \in RV_{\alpha_1\alpha_2},$$

iv) Si $f \in RV_\alpha$, $\alpha > 0$ *à l'infini alors*

$$f^{-1} \in RV_{\frac{1}{\alpha}}$$

Proposition 1.3.1 (Potter 1942) *Suppose que $f \in RV_\alpha$ si $\delta_1, \delta_2 > 0$ sont arbitraires il existe $t_0 = t_0(\delta_1, \delta_2)$ tell que pour $t \geq t_0$, $tx \geq t_0$*

$$(1-\delta_1)x^\alpha \min(x^{-\delta_2}, x^{\delta_2}) < \frac{f(tx)}{f(t)} < (1+\delta_1)x^\alpha \max(x^{-\delta_2}, x^{\delta_2}). \tag{1.30}$$

Proposition 1.3.2 (Dress 1998) *Si $f \in RV_\alpha$ pour tout $\varepsilon,\ \delta > 0$ il existe $t_0 = t_0(\varepsilon, \delta)$ telle que pour t, $tx \geq t_0$*

$$\left| \frac{f(tx)}{f(t)} - x^\alpha \right| \leq \varepsilon \max(x^{\alpha+\delta}, x^{\alpha-\delta}). \tag{1.31}$$

Proposition 1.3.3 (de Haan et Ferreira (2006)**)** *Soit $f \in RV_\alpha$, $\alpha \in \mathbb{R}$ à l'infini alors il existe une fonction à variation lente l à l'infini telle que*

$$f(x) = x^\alpha l(x),\ x > 0.$$

Domaine d'attraction de Fréchet :

On dit que F est appartient au domaine d'attraction de Fréchet notée $F \in D(\Phi_\gamma),\ \gamma > 0$ si et seulement si x_F le point terminal $x_F = +\infty$ et $1 - F(x)$ est une fonction à variation régulière d'indice $-\frac{1}{\gamma}$. Et pour les suites on choisit :

$$a_n = F^{\leftarrow}(1 - \frac{1}{n}) = U(n),\ b_n = 0.$$

Si $F \in D(\Phi_\gamma)$ alors (voir Embrech et al (1997) [10]) :

$$a_n^{-1} M_n \xrightarrow{d} \Phi_\gamma.$$

Remarque 1.3.8 (Condition de Von-Mises) *Supposons que F est absolument conti-*

nue avec densité $f(x) > 0,\ x \in]x_1, +\infty[,$ *si pour* $\gamma > 0$

$$\lim_{x \longrightarrow \infty} \frac{xf(x)}{1 - F(x)} = \gamma, \tag{1.32}$$

alors (Resnick (1987) *[17].) :* $F \in D(\Phi_\gamma)$.

Domaine d'attraction de Weibull

Théorème 1.3.2 *On dit que* F *est appartient au domaine d'attraction de Weibull noté :*$F \in D(\psi_\gamma),\ \gamma < 0$ *si et seulement si* $x_F < +\infty$ *et* $\bar{F}(x_F - x^{-1}) = x^{\frac{1}{\gamma}} L(x)$. *On défini*

$$F^*(x) = F(x_F - x^{-1}) \text{ si } x > 0,$$

alors $\bar{F}^* \in RV_{\frac{1}{\gamma}}$,*et pour les suites on choisit*

$$a_n = x_F - F^{\longleftarrow}(1 - \frac{1}{n}),\ b_n = x_F.$$

Si $F \in D(\psi_\gamma)$ *alors*

$$a_n^{-1}(M_n - b_n) \xrightarrow{d} \psi_\gamma.$$

Preuve. Voir Embrech et al (1997) [10]. ■

Remarque 1.3.9 (Condition de Von-Mises) *Si* F *est absolument continue avec densité* $f(x) > 0$, *pour* $x \in]x_1, x_F[$ *et* $f(x) = 0$ *pour* $x > x_F$, *pour* $\gamma > 0$

$$\lim_{x \longrightarrow x_F} \frac{(x_F - x)f(x)}{1 - F(x)} = \gamma, \tag{1.33}$$

alors (Resnick (1987) *[17]) :* $F \in D(\psi_\gamma)$

Domaine d'attraction de Gumbell :

Définition 1.3.2 (fonction de von-mises) *F est une fonction de von-mises si il existe $z < x_F$ (x_F le point terminal) telle que*

$$1 - F(x) = c \exp\{-\int_z^x \frac{1}{\alpha(t)} dt\},\ z < x < x_F, \tag{1.34}$$

avec $c > 0$ et α est une fonction positive absolument continue de densité α' vérifiant $\lim_{x \longrightarrow x_F} \alpha'(x) = 0$

Théorème 1.3.3 *On dit que F est appartient au domaine d'attraction de Gumbel noté $F \in D(\Lambda)$ si et seulement si il existe une fonction de von-mises F^* telle que pour $z < x < x_F$:*

$$1 - F(x) = c(x)[1 - F^*(x)] = c(x) \exp\{-\int_z^x \frac{1}{\alpha(t)} dt\}, \tag{1.35}$$

où $\lim_{x \longrightarrow x_F} c(x) = c > 0$, et pour les suites on choisit :

$$b_n = F^{\longleftarrow}(1 - \frac{1}{n}),\ a_n = \alpha(b_n).$$

Si $F \in D(\Lambda)$ alors

$$a_n^{-1}(M_n - b_n) \xrightarrow{d} \Lambda.$$

Preuve. Resnick(1987) [17]. ■

Remarque 1.3.10 (Condition de Von-Mises) *Suppose que F admet une deuxième dérivée négative $f'(x) < 0$ pour tout $x \in]x_1, x_F[$ et pour tout $x \geq x_F \quad f(x) = 0$ et*

$$\lim_{x \longrightarrow x_F} \frac{f'(x)(1 - F(x))}{(f(x))^2} = -1, \tag{1.36}$$

alors (Resnick (1987) [17]) : $F \in D(\Lambda)$.

Proposition 1.3.4 *$F \in D(G_\gamma)$ si et seulement si $n\ \bar{F}(a_n x + b_n) \underset{n\to\infty}{\longrightarrow} -\log G_\gamma(x)$ pour une certaine suite $a_n > 0$, $b_n \in \mathbb{R}$ on a alors :*

$$a_n^{-1}(M_n - b_n) \xrightarrow{d} G_\gamma(x).$$

Théorème 1.3.4 (caractérisation de $D(G_\gamma)$) *Les affirmation suivantes sont équivalentes :*

a) $F \in D(G_\gamma)$.

b) il existe a une fonction mesurable positive telle que pour $1 + \gamma x > 0$

$$\lim_{u\longrightarrow x_F} \frac{\bar{F}(u + xa(u))}{\bar{F}(u)} = \begin{cases} (1+\gamma x)^{-\frac{1}{\gamma}} & si \quad \gamma \neq 0 \\ \exp(-x) & si \quad \gamma = 0 \end{cases}, \tag{1.37}$$

c) pour $x, y > 0$, $y \neq 1$

$$\lim_{u\longrightarrow\infty} \frac{U(sx) - U(s)}{U(sy) - U(s)} = \begin{cases} \frac{x^\gamma - 1}{y^\gamma - 1} & si \quad \gamma \neq 0 \\ \frac{\ln x}{\ln y} & si \quad \gamma = 0 \end{cases}. \tag{1.38}$$

Preuve. Embrech et al (1997) [10]. ■

Définition 1.3.3 (condition de première ordre) *Pour $x > 0$, la fonction de queue de quantile U à variation régulière a l'infini avec indice γ admet condition de première ordre si :*

$$\lim_{t\longrightarrow\infty} \frac{U(tx)}{U(t)} = x^\gamma \Longleftrightarrow \lim_{t\longrightarrow\infty} \frac{\bar{F}(tx)}{\bar{F}(t)} = x^{-\frac{1}{\gamma}}, \tag{1.39}$$

tell que $\bar{F}$ est une fonction à variation régulière à l'infini avec indice $-\frac{1}{\gamma}$, $\bar{F} \in RV_{\frac{-1}{\gamma}}$ et $U \in RV_\gamma$.

Définition 1.3.4 (condition de seconde ordre) *Une fonction U admet une condition de seconde ordre a l'infini s'il existe un paramètre $\rho \leq 0$ et une fonction A_1 avec $\lim_{t\longrightarrow\infty} A_1(t) =$*

0, $\forall x > 0$

$$\lim_{t\longrightarrow\infty} \frac{\frac{U(tx)}{U(t)} - x^{\gamma}}{A_1(t)} = x^{\gamma}\frac{x^{\rho}-1}{\rho}, \tag{1.40}$$

avec $|A_1| \in RV_\rho,\ \gamma > 0$ *et pour la fonction* $\bar{F}$ *admet une condition de seconde ordre a l'infini s'il existe un paramètre* $\rho \leq 0$ *et une fonction* A_2 *avec* $\lim_{t\longrightarrow\infty} A_2(t) = 0,\ \forall x > 0$

$$\lim_{t\longrightarrow\infty} \frac{\frac{\bar{F}(tx)}{\bar{F}(t)} - x^{\frac{-1}{\gamma}}}{A_2(t)} = x^{\frac{-1}{\gamma}}\frac{x^{\frac{\rho}{\gamma}}-1}{\rho\gamma}, \tag{1.41}$$

où $A_2(t) = A_1(1/1 - F(t)),\ |A_2| \in RV_{\frac{\rho}{\gamma}}$

Remarque 1.3.11 (class de Hall) *La classe de Hall est composée de l'ensemble des distributions* F *telle que*

$$F(x) = 1 - cx^{\frac{-1}{\gamma}}(1 + dx^{\frac{\rho}{\gamma}} + 0(x^{\frac{\rho}{\gamma}})),\ \textit{quand}\ x \longrightarrow \infty, \tag{1.42}$$

où $\gamma > 0,\ \rho \leq 0,\ c > 0$ *et* $d \in \mathbb{R}^*$. *Il s'atgit d'une sous classe de distributions à queues lourdes, la relation (1.42) peut être reformulée en termes de fonction quantile de queue* U *comme suit :*

$$U(t) = c^{\gamma}t^{\gamma}(1 + \gamma dc^{\rho}t^{\rho} + 0(t^{\rho})),\ \textit{quand}\ t \longrightarrow \infty. \tag{1.43}$$

1.3.4 Distribution conditionnelle des excès

Plutôt que de considère le maximum d'un échantillon $X_1, \cdots, X_n$, on étudie les valeur dépassent un seuil donnée , l'excès Y de la variable X au- dessus du seuil u est définit par : $X - u$ quand $X \geq u$

Définition 1.3.5 (la fonction distribution des excès) *Soit* X *un variable aléatoire de fonction répartition* F *et de points terminal* x_F, *pour tout* $u < x_F$ *la fonction distribution des excès définit par :*

$$F_u(y) = P(X - u \leq y/X > u) = \frac{F(u+y) - F(u)}{1 - F(u)},\ y > 0. \tag{1.44}$$

Définition 1.3.6 (Embrech et al (1997) **[10], Coles [6])** *On appelle fonction des excès en moyen "Mean excess function" la fonction $e(u)$ définit par :*

$$e(u) = \mathbb{E}[X - u/X > u], \text{ pour } u \geq 0. \tag{1.45}$$

1.3.5 Distribution de Pareto généralisé (GPD)

La loi de Pareto généralisé noté GPD :

$$H_\gamma(x) = \begin{cases} 1 - (1+\gamma x)^{\frac{-1}{\gamma}} & \text{si} \quad \gamma \neq 0 \\ 1 - \exp(-x) & \text{si} \quad \gamma = 0 \end{cases} \tag{1.46}$$

$$\text{avec :} \quad \begin{array}{rl} x \geq 0 & \text{si} \quad \gamma \geq 0 \\ 0 \leq x \leq -\frac{1}{\gamma} & \text{si} \quad \gamma < 0 \end{array}$$

H_γ est appelé loi Pareto Généralisé standard.

La fonction général de GPD noté $H_{\gamma,v,\beta}(x) = H_\gamma(\frac{x-v}{\beta})$, On remplacent x par $\frac{x-v}{\beta}$ dans (1.47), telle que $v \in \mathbb{R}$ appelé le paramètre de localisation $\beta > 0$ appelé paramètre d'échelle. Le cas où le paramètre de localisation est nul ($v = 0$) et ($\beta > 0$) est important dans l'analyse statistique des évènement extrême, cette famille noté par $H_{\alpha,\beta}(x)$ telle que :

$$H_{\gamma,\beta}(x) = \begin{cases} 1 - (1+\gamma\frac{x}{\beta})^{\frac{-1}{\gamma}} & \text{si} \quad \gamma \neq 0 \\ 1 - e^{-\frac{x}{\beta}} & \text{si} \quad \gamma = 0 \end{cases} \tag{1.47}$$

$$\text{avec :} \quad \begin{array}{rl} x \geq 0 & \text{si} \quad \gamma \geq 0 \\ 0 \leq x \leq -\frac{\beta}{\gamma} & \text{si} \quad \gamma < 0 \end{array}$$

Sa densité est $h_{\gamma,\beta}$ où

$$h_{\gamma,\beta}(x) = \begin{cases} \frac{1}{\beta}(1+\gamma\frac{x}{\beta})^{\frac{-1}{\gamma}-1} & \text{si} \quad \gamma \neq 0 \\ \frac{1}{\beta}e^{-\frac{x}{\beta}} & \text{si} \quad \gamma = 0 \end{cases} \tag{1.48}$$

Remarque 1.3.12 *La relation entre le loi Pareto généralisé standard H_γ et la loi des valeurs extrême généralisé standard G_γ est :*

$$H_\gamma(x) = 1 + \log G_\gamma(x), \ si \log G_\gamma(x) > -1. \tag{1.49}$$

Suppose $x_i \in \begin{cases} [0,\infty[& si \quad \gamma \geq 0 \\ [0,\frac{-\beta}{\gamma}[& si \quad \gamma < 0 \end{cases}$ $i = 1,2$ *alors :*

$$\frac{\bar{H}_{\gamma,\beta}(x_1+x_2)}{\bar{H}_{\gamma,\beta}(x_1)} = \bar{H}_{\gamma,\beta+\gamma x_1}(x_2) \tag{1.50}$$

Suppose X a GPD avec paramètre $\gamma < 1$ et β, et pour $u < x_F$

$$e(u) = \mathbb{E}(X-u/X>u) = \frac{\beta+\gamma u}{1-\gamma}, \ \beta+\gamma u > 0 \tag{1.51}$$

Théorème 1.3.5 (Pickands-Balkema-de Haan) *Une fonction de répartition F appartient au domaine d'attraction de G_γ si et seulement si il existe une fonction positive $\beta(u)$ tell que :*

$$\lim_{u \longrightarrow x_F} \sup_{0<y<x_F-u} \left| F_u(y) - H_{\gamma,\beta(u)}(y) \right| = 0, \tag{1.52}$$

où $F_u(y)$: la fonction distribution des excès, $H_{\gamma,\beta(u)}$:le GPD et x_F le point terminal, voir Reiss et M.Thomas [16].

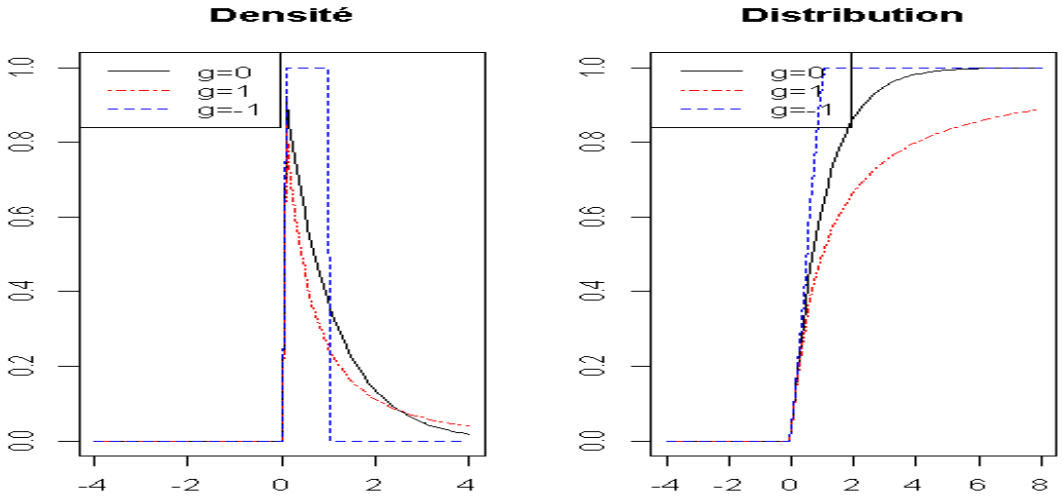


FIG. 1.2 – Distribution et densité standard de GPD

1.4 Estimation de l'indice des valeurs extrêmes γ

On définit deux estimateurs semi paramétriques différentes basée sur la statistique d'ordre $X_{1,n} \leq \cdots \leq X_{i,n}$,on considérant les k valeurs les plus grandes, $k \to \infty$ lorsque $n \to \infty$ et $\frac{k}{n} \to 0$.

1.4.1 Estimateur de Pickands

$F \in D(G\gamma),\ \gamma \in \mathbb{R}$ on définit l'estimateur de Pickands par :

$$\hat{\gamma}_{k,n}^{(P)} = \frac{1}{\ln 2} \ln\left(\frac{X_{n-k,n} - X_{n-2k,n}}{X_{n-2k,n} - X_{n-4k,n}}\right). \tag{1.53}$$

Théorème 1.4.1 *Soit $X_1, \cdots, X_n$ n va iid,$F \in D(G\gamma), \gamma \in \mathbb{R}$ la condition du second ordre est vérifiée, soit $k = k_n,\ n \geq 1$ une suite des entiers telle que $1 < k < n,\ k \to \infty$ et $\frac{k}{n} \to 0$ lorsque $n \to \infty$*

Consistant faible :

$$\hat{\gamma}_{k,n}^{(p)} \xrightarrow{P} \gamma,\ \ quand\ n \longrightarrow \infty. \tag{1.54}$$

Consistant fort : si de plus $\frac{k}{\ln\ln n} \longrightarrow \infty$ *quand* $n \longrightarrow \infty$ *alors*

$$\hat{\gamma}_{k,n}^{(p)} \xrightarrow{p.s} \gamma, \ \ quand\ n \longrightarrow \infty. \tag{1.55}$$

Normalité asymptotique :

$$\sqrt{k}(\hat{\gamma}_{k,n}^{(p)} - \gamma) \xrightarrow{d} \mathcal{N}(0.\delta(\gamma)), \ \ quand \ \ n \longrightarrow \infty, \tag{1.56}$$

où $\delta(\gamma) = \frac{\gamma^2(2^{2\gamma+1}+1)}{(2(2^\gamma-1)\ln 2)^2}$.

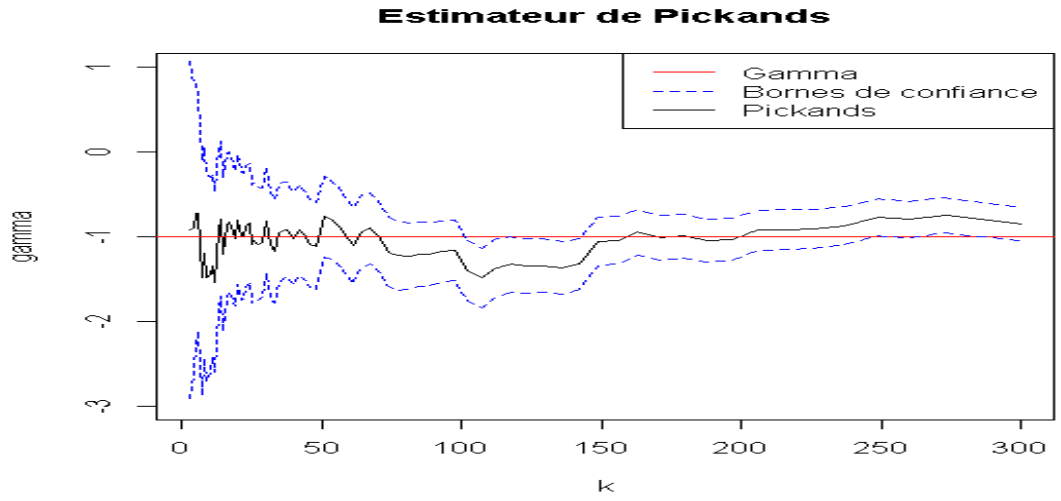


FIG. 1.3 – Estimateur de Pickands, avec un intervalle de confiance de niveau 95%, pour l'EVI de la distribution uniforme standard ($\gamma = -1$) basé sur 100 échantillons de 3000 observations.

1.4.2 Estimateur de Hill

Pour $F \in D(G\gamma),\ \gamma > 0$, on définit l'estimateur de Hill par :

$$\hat{\gamma}_{k,n}^{(H)} = \frac{1}{k}\sum_{j=1}^{k} \ln X_{n-j+1,n} - \ln X_{n-k,n} \tag{1.57}$$

Théorème 1.4.2 *Soit $X_1, \cdots, X_n$ n va iid, de loi F telle que $\bar{F}(x) = x^{-\frac{1}{\gamma}} L(x)$, $x > 0$, $\gamma > 0$ et L une fonction a variation lente :*

i) Consistant faible : si $k \longrightarrow \infty$ et $\frac{k}{n} \longrightarrow 0$ et $n \longrightarrow \infty$ alors

$$\hat{\gamma}_{k,n}^{(H)} \xrightarrow{P} \gamma, \ \ quand\ n \longrightarrow \infty \tag{1.58}$$

ii) Consistant fort : si de plus $\frac{k}{\ln \ln n} \longrightarrow \infty$ quand $n \longrightarrow \infty$ alors

$$\hat{\gamma}_{k,n}^{(H)} \xrightarrow{p.s} \gamma, \ \ quand\ n \longrightarrow \infty \tag{1.59}$$

iii) Normalité asymptotique : supposons que F satisfaisant (1.41)*, si $\sqrt{k} A(\frac{n}{k}) \longrightarrow \lambda$ quand $n \longrightarrow \infty$ alors*

$$\sqrt{k}(\hat{\gamma}_{k,n}^{(H)} - \gamma) \xrightarrow{d} \mathcal{N}(\frac{\lambda}{1-\rho}.\gamma^2), \ \ quand\ n \longrightarrow \infty \tag{1.60}$$

Preuve. de Haan et Ferreira (2006) [11]. ■

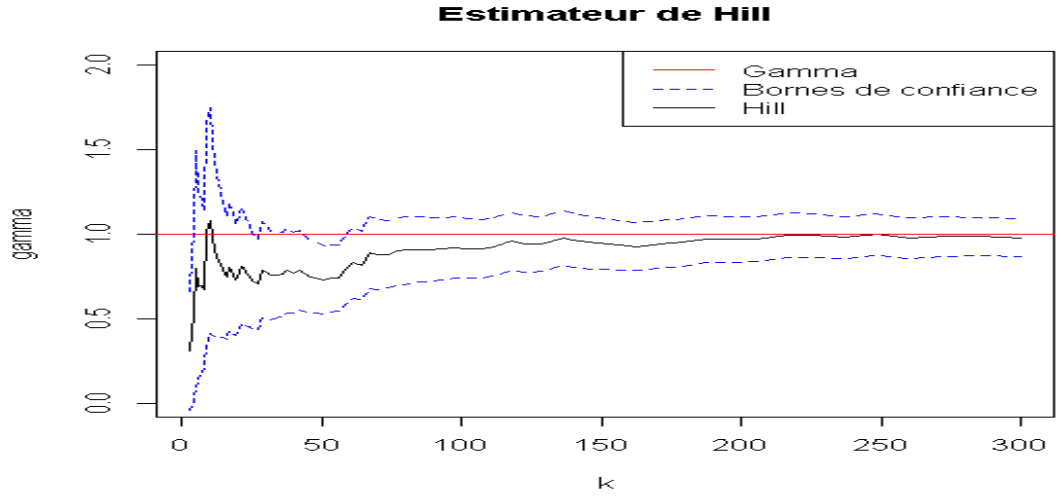


FIG. 1.4 – Estimateur de Hill,avec un intervalle de confiance de niveau 95%, pour l'EVI de la distribution Pareto standard ($\gamma = 1$) basé sur 100 échantillons de 3000 observations.

Chapitre 2

Détermination du nombre de statistiques d'ordre extrêmes

2.1 Méthode Graphique

Cette méthode consiste a tracer le graphe$\{(k, \hat{\gamma}_{k,n}) : 1 < k < n\}$ dans le but de trouver une valeur optimal de k noté k_{opt} est choisi dans la première région ou l'estimateur $\hat{\gamma}_{k,n}$ devient stable.

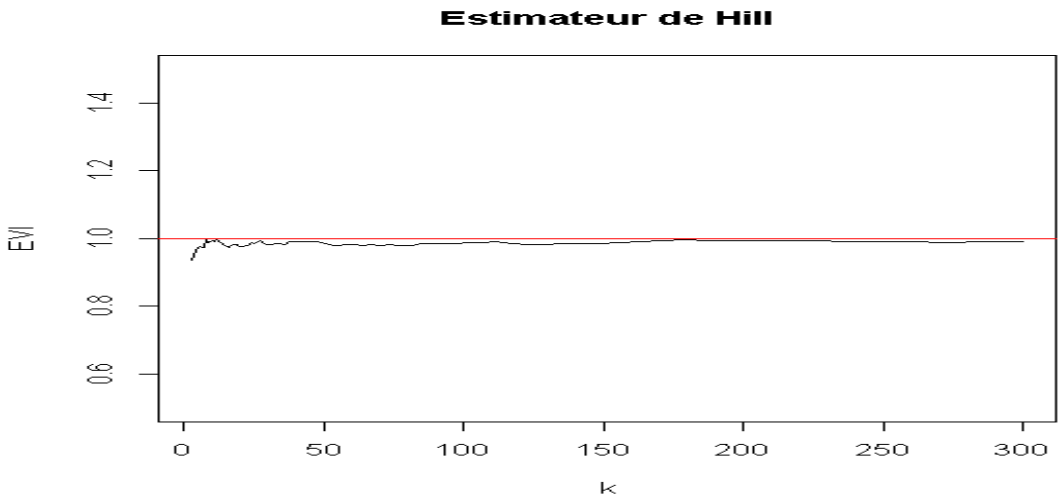


FIG. 2.1 – Méthode Graphique du choix de k.

2.1.1 Méthode Sum-Plot

Sousa et Michaildis [18] présentent une méthode qui basée sur le graphe du couple (k, S_k), avec S_k une variable aléatoire définit par

$$S_k = \sum_{i=1}^{k} iVi = \sum_{i=1}^{k} i \log \frac{X_{i,n}}{X_{i+1,n}}, \quad pour\ k = 1, \cdots, n \tag{2.1}$$

où $X_{i,n}$ la statistique d'ordre de $(X_1, \cdots, X_n)$ de distribution F tel que

$$1 - F(x) = x^{-\frac{1}{\gamma}} l(x), \tag{2.2}$$

avec l est une fonction a variation lente, on choisit le nombre k, pour toute $x > X_{k+1,n}$, le S_k est suit dans ce cas la distribution de gamma de paramètre $(k, \frac{1}{\gamma})$, on traçons S_k contre k, on s'attend à ce que le graphique obtenue doit être linéaire dans la région correspondant de la partie supérieur des statistiques d'ordre. Le fait d'être estimé par le coefficient de régression de la pente du modèle de régression linéaire simple ;

$$S_i = \beta_0 + \beta_1 i + \varepsilon_i,\ i = 1, \cdots, k. \tag{2.3}$$

Par l'estimation des moindres carrés généralisés

$$\hat{\gamma} = \hat{\beta}_1 = \frac{k}{k-1} \hat{\gamma}^H_{k,n} - \frac{k}{k-1} \log X_{1,n}, \tag{2.4}$$

où $\hat{\gamma}^H_{k,n}$: l'estimateur de Hill. En conclusion, si $\beta_0 = 0$, alors $\hat{\gamma} = \hat{\gamma}^H_{k,n}$.

2.2 Minimisation de l'erreur quadratique moyenne asymptotique

Dans cette section nous utilisons l'approche ci-dessus sur la base d'un modèle linéaire généralisé de formulation, dans le but de déterminer la fraction de l'échantillon extrême, peut être utilisé lors de l'estimation de γ, l'objectif est d'estimer l'erreur quadratique moyenne asymptotique ($AMSE$) des estimateurs. Alors, k sera déterminé en minimisant ces expressions d'erreur :

$$AMSE(\hat{\gamma}_n) = E_\infty[(\hat{\gamma}_n - \gamma)^2] = Var(\hat{\gamma}_n) + Biais^2(\hat{\gamma}_n). \tag{2.5}$$

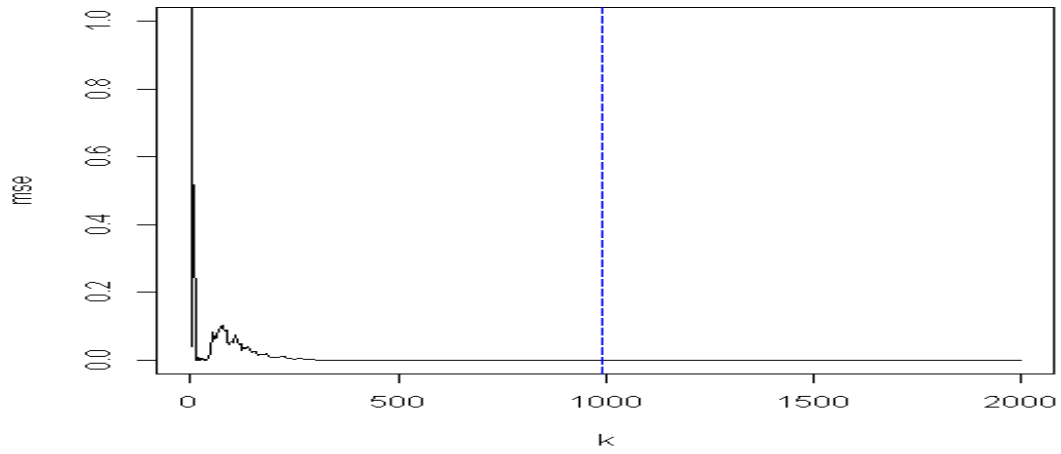
où E_∞ est l'espérance mathématique suivant la distribution asymptotique. Le k_{opt} est :

$$k_{opt} = \arg\min_k AMSE(\hat{\gamma}_n). \tag{2.6}$$

2.2.1 Résultat pour l'estimateur de Hill

Concernant l'estimateur de Hill pour les fonction appartenant au domaine d 'attraction maximale de Fréchet, $\gamma > 0$ ($\rho = \max(-1, -\gamma)$), de Haan et Peng (1998) ont proposé un nombre optimale qui minimise l'erreur moyenne quadratique de l'estimateur de Hill (voir Neves et Fraga Alves [15]) :

$$k_{opt} \sim \begin{cases} \left\{\frac{(1+\gamma)^2}{2\gamma}\right\}^{\frac{1}{(2\gamma+1)}} n^{\frac{2\gamma}{(2\gamma+1)}} & si \quad 0 < \gamma < 1 \\ 2\left\{\frac{n}{3}\right\}^{\frac{2}{3}} & si \quad \gamma = 1 \\ 2n^{\frac{2}{3}} & si \quad \gamma > 1 \end{cases} \tag{2.7}$$



2.3 Procédures adaptatives

La procédure du choix optimale de k est souvent difficile du fait que celui-ci ne dépend pas exclusivement de la taille de l'échantillon et de l'indice des valeurs extrêmes γ, mais il dépend d'autre paramètres inconnues (le paramètre de seconde ordre ρ etc...) intervenant dans la fonction de distribution F. Il existe plusieurs algorithmes de procédures adaptatives pour trouver un estimateur $\hat{k}_{opt}$de k_{opt} :

$$\frac{\hat{k}_{opt}}{k_{opt}} \xrightarrow{P} 1, \quad \text{quand } n \longrightarrow \infty.$$

2.3.1 Approche de Hall et Welsh

Hall et Welsh [12] prouvent que l'erreur quadratique moyenne asymptotique de l'estimateur de Hill est minimisée par

$$k_{opt} \sim \left(\frac{c^2(1+\rho)^2}{2d^2\rho^3}\right)^{\frac{1}{(2\rho+1)}} n \frac{2\rho}{(2\rho+1)}. \tag{2.8}$$

Si la fonction de distribution satisfait la classe de Hall (1.42) avec les paramètres $\rho > 0$, $c > 0$ et $d \neq 0$ sont inconnus, ce résultat ne peuvent pas être utilisé directement

pour déterminer le nombre optimal de statistique d'ordre, la procédure de Hall et Welsh construit un estimateur consistant pour k_{opt} :

$$\hat{k}_{opt} = [\hat{\lambda}_0 n^{\frac{2\hat{\rho}}{(2\hat{\rho}+1)}}], \tag{2.9}$$

où

$$\hat{\rho} = \left| \log \left| \frac{(\hat{\gamma}_n^H(t_1))^{-1} - (\hat{\gamma}_n^H(s))^{-1}}{(\hat{\gamma}_n^H(t_2))^{-1} - (\hat{\gamma}_n^H(s))^{-1}} \right| / \log \frac{t_1}{t_2} \right|, \tag{2.10}$$

et

$$\hat{\lambda}_0 = \left| (2\hat{\rho})^{-1/2} (\frac{n}{t_1})^{\hat{\rho}} \frac{(\hat{\gamma}_n^H(t_1))^{-1} - (\hat{\gamma}_n^H(s))^{-1}}{\hat{\gamma}_n^H(s)} \right|^{\frac{2}{(2\hat{\rho}+1)}}. \tag{2.11}$$

Ce estimateur consistant de k_{opt} dans le sens quadratique :

$$\frac{\hat{k}_{opt}}{k_{opt}} \xrightarrow{P} 1 \quad si\ t_i = [n^{t_i}],\ i = 1, 2 \quad et \quad s = [n^6]$$

pour certains

$$0 < 2\rho(1 - \tau_1) < \delta < 2\rho(2\rho + 1) < \tau_1 < \tau_2 < 1.$$

2.3.2 Approche de Bootstrap

La méthode de Bootstrap proposée initialement par Efron (1979), le principe de cette méthode est rééchantillonner un grand nombre de fois l'échantillon initial. Cette procédure exige que le paramètre ρ est connue et la taille n de l'échantillon est très grande, cet approche détermine la fraction de l'échantillon qui minimise l'erreur quadratique.

Bootstrap proposé par Hall

Hall [13] propose la méthode du Bootstrap dans l'estimation de l'indice de queue pour le nombre de statistique d'ordre extrême, cette méthode estime $AMSE$.

Nous avons quand $k \longrightarrow \infty,\ \frac{k}{n} \longrightarrow 0$:

$$\sqrt{k_{opt}}(\hat{\gamma}_n^H(k_{opt}) - \gamma) \xrightarrow{d} N(b, \gamma^2). \tag{2.12}$$

Soit F une fonction de distribution satisfait (2.2) et (1.41) avec ρ et γ sont connus, le but est de déterminer une estimateur $\hat{k}_{opt}$ tel que

$$\sqrt{\hat{k}_{opt}}(\hat{\gamma}_n^H(\hat{k}_{opt}) - \gamma) \xrightarrow{d} N(b, \gamma^2). \tag{2.13}$$

Pour ceci c'est suffisant pour prouver

$$\frac{\hat{k}_{opt}}{k_{opt}} \xrightarrow{P} 1. \tag{2.14}$$

Hall utilise des échantillons dont la taille n_1 est d'un ordre plus petit que la taille de l'échantillon initiale on retire un sous -échantillon $\chi_{n_1}^* = \{X_1^*, \cdots, X_{n_1}^*\}$ du l'échantillon total $\chi_n = \{X_1, \cdots, X_n\}$ $(n_1 << n)$ et $X_{1,n_1}^* \leq X_{2,n_1}^* \leq \cdots \leq X_{n_1,n_1}^*$ les statistiques d'ordres de $\chi_{n_1}^*$. Nous avons défini :

$$\gamma_{n_1}^*(k_1) = \frac{1}{k_1} \sum_{i=1}^{k_1} \log X_{n_1-i+1,n_1}^* - \log X_{n_1-k_1,n_1}^*. \tag{2.15}$$

Le Bootsrap estime $AMSE\ (n_1, k_1)$:

$$\widehat{MSE}(n_1, k_1) = E[(\gamma_{n_1}^*(k_1) - \gamma_{n_1}(k_1))^2 \mid \chi_n]. \tag{2.16}$$

Nous déterminons k et k_1 en minimisons $\widehat{MSE}(n_1, k_1)$, Supposons que nous savons que l'asymptotique k est $k = cn^\alpha$ avec $0 < \alpha < 1$ et c est inconnue et la relation $k = k_1(\frac{n}{n_1})^\alpha$.

Bootstrap proposé par Danielsson et al

Danielsson et al [7] proposent une méthode de remplacé $\hat{\gamma}_n^H(k)$ par $M_n(k)$ où :

$$M_n(k) = \frac{1}{k}\sum_{i=1}^{k}(\log X_{i,n} - \log X_{k+1,n})^2. \tag{2.17}$$

Nous avons prouvé $M_n(k)/2\gamma_n(k)$ converge en probabilité vers γ quand $k \longrightarrow \infty$, il a aussi équilibrer la variance et le biais, il peut être démontrer que les statistiques $M_n(k)/(2\gamma_n(k)-\gamma_n(k))$ et $\gamma_n(k)$ en la même propriété asymptotique que leur moyenne asymptotique est 0 , il peut recevoir la même valeur de k pour minimisé $AMSE$ et $E_\infty(M_n(k) - 2(\gamma_n(k))^2)^2$. Nous sélectionnons les statistiques pour détermine $k_1$en minimisons $Q(n_1, k_1)$:

$$Q(n_1, k_1) = E[(M_n^*(k_1) - 2(\gamma_n^*(k_1))^2)^2 \diagdown \chi_n]. \tag{2.18}$$

La procédure de calcul $\gamma_n^H(\hat{k})$ est :

Etape 1 choisir $n_1$et k_1, $(n_1 = O(n^{1-\varepsilon}),\ 0 \leq \varepsilon \leq 1)$ et minimiser $Q(n_1, k_1)$ pour déterminer k_1

Etape 2 nous supposons que $n_2 = \frac{n_1^2}{n}$ et répéter Etape 1 pour déterminer k_2

Etape 3 nous pouvons calculer $\hat{k}$:

$$\hat{k} = \frac{k_1^2}{k_2}\left(\frac{(\log k_1)^2}{(2\log n_1 - \log k_1)^2}\right)^{(\log n_1 - \log k_1)/\log n_1}. \tag{2.19}$$

2.3.3 Approche Séquentielle

Cette procédure séquentielle est présenté par Dress et Kaufmann [9] pour construire une estimation consistant du k_{opt} sans n'importe quelle connaissance sur la fonction de distribution de la classe de Hall (1.42). Leur estimateur se fonde sur le fait que la fluctuation aléatoire maximum $i^{\frac{1}{2}}(\hat{\gamma}_{i,n} - \gamma)$ avec $2 \leq i \leq k_n$, est de l'ordre $(\log\log n)^{\frac{1}{2}}$ pour tous les séquences intermédiaires k_n, cet estimateur est basé sur le temps d'arrêt pour des séquences

de l'estimateur de Hill :

$$\tilde{k}(r_n) = \min\{k \in \{2, ..., n\} : \max_{2 \leq i \leq k_n} i^{\frac{1}{2}} \left|\hat{\gamma}_{i,n}^H - \gamma_{k,n}^H\right| > r_n\}, \tag{2.20}$$

où r_n les seuils constituent une séquence qui est d'ordre supérieur que $(\log\log n)^{\frac{1}{2}}$ et d'ordre inferieur que $n^{\frac{1}{2}}$.

Le paramètre de seconde d'ordre ρ est remplacé par une valeur fixée ρ_0 dans cette méthode, pour fixer $\rho_0 = -1$, la méthode de Dress et Kaufmann est résumé dans les trois étapes suivantes :

Etape 1 : soit $r_n = 2.5 * \hat{\gamma}_n^H * n^{0.25}$ avec $\hat{\gamma}_n^H = \hat{\gamma}_{2\sqrt{n},n}^H$.

Etape 2 : obtenir $\tilde{k}(r_n)$ si la condition $\max_{2 \leq i \leq k_n} i^{\frac{1}{2}} \left|\hat{\gamma}_{i,n}^H - \gamma_{k,n}^H\right| > r_n$ est satisfait pour certains k, puis passer à l'étape 3 , sinon assigner $0.9 * r_n$ à r_n et répéter l'étape 2.

Etape 3 : pour $\varepsilon \in (0.1)$(en particulier $\varepsilon = 0.7$) déterminer

$$\hat{k}_{opt} = [\frac{1}{3}(2(\hat{\gamma}_n^H)^2)^{\frac{1}{3}}(\frac{\tilde{k}(r_n^{\varepsilon})}{(\tilde{k}(r_n))^{\varepsilon}})^{\frac{1}{1-\varepsilon}}]. \tag{2.21}$$

où $[a]$ désigne le plus grand entier inferieur ou égale à a.

2.3.4 Approche de Cheng et Peng

L'estimateur de Hill c'est l'estimateur le plus connue. Considérons les intervalles de confiance basés sur l'approximation asymptotique normal de l'estimateur de Hill.

Proposition 2.3.1 *Supposons que (1.42) est satisfait et* $k \longrightarrow \infty$, $\frac{k}{n} \longrightarrow 0$, *alors*

$$\sqrt{k}(\hat{\gamma}_n^H - \gamma) \xrightarrow{d} \mathcal{N}(0, \gamma^2), \tag{2.22}$$

si et seulement si $k = 0(n^{2\rho/(1+2\rho)})$. *Puis avec* $0 < \alpha < 1$ *les intervalles de confiance au*

niveau $(1-\alpha)$ *pour l'indice* γ *baseé sur l'approximation normal (2.22) sont :*

$$I_1(\alpha) = (0, \hat{\gamma}_n^H + x_\alpha \frac{\hat{\gamma}_n^H}{\sqrt{k}}), \tag{2.23}$$

et

$$I_2(\alpha) = (\hat{\gamma}_n^H - x_{\frac{\alpha}{2}} \frac{\hat{\gamma}_n^H}{\sqrt{k}}, \hat{\gamma}_n^H + x_{\frac{\alpha}{2}} \frac{\hat{\gamma}_n^H}{\sqrt{k}}) \tag{2.24}$$

avec x_θ *est le quantile d'ordre* $1-\theta$ *de loi Normal standard* $0 \leq \theta \leq 1$.

Nous étudions la précesion de couverture pour les intervalles de confiance $I_1(\alpha)$ et $I_2(\alpha)$.

Théorème 2.3.1 *Supposons* 1.42 *satisfait et* $k \longrightarrow \infty$, $\frac{k}{n} \longrightarrow 0$ *alors :*

$$P(\gamma \in I_1(\alpha)) = 1 - \alpha - \phi(x_\alpha) \left\{ \frac{1+2x_\alpha^2}{3\sqrt{k}} - \frac{\rho d c^\rho}{(1-\rho)} \sqrt{k} (\frac{n}{k})^\rho \right\} + 0(\frac{1}{\sqrt{k}} + \sqrt{k}(\frac{n}{k})^\rho), \tag{2.25}$$

et

$$P(\gamma \in I_2(\alpha)) = 1 - \alpha + 0(\frac{1}{\sqrt{k}} + \sqrt{k}(\frac{n}{k})^\rho), \tag{2.26}$$

telle que ϕ *est une fonction densité de loi Normal standard ou la valeur optimal de* k *qui minimise la valeur absolue de l'erreur de couverture pour* $I_1(\alpha)$ *est :*

$$k_{opt} = \begin{cases} \left\{ \frac{(1+2x_\alpha^2)(1-\rho)c^{-\rho}}{-3d\rho(1-2\rho)} \right\}^{\frac{1}{1-\rho}} n^{-\rho\frac{1}{(1-\rho)}} & si \quad d > 0 \\ \left\{ \frac{(1+2x_\alpha^2)(1-\rho)c^{-\rho}}{-3d\rho} \right\}^{\frac{1}{1-\rho}} n^{-\rho\frac{1}{(1-\rho)}} & si \quad d < 0 \end{cases} \tag{2.27}$$

Qui satisfait automatiquement la condition $k = 0(n^{-2\rho/(1-2\rho)})$ *dans la proposition précèdent, en outre la précesion de la couverture optimale pour* $I_1(\alpha)$ *est :*

$$P(\gamma \in I_1(\alpha)) = \begin{cases} 1 - \alpha - 2 \left\{ \frac{(1-\rho)(1+2x_\alpha^2)}{3(1-2\rho)} \right\}^{\frac{(1-2\rho)}{2(1-\rho)}} \{-d\rho c^\rho\}^{\frac{1}{(2(1-\rho))}} & \\ \times \phi(x_\alpha) n^{\frac{\rho}{(2(1-\rho))}} (1 + 0(1)) & si \quad d > 0 \\ 1 - \alpha + 0(n^{\frac{\rho}{(2(1-\rho))}} & si \quad d < 0 \end{cases} \tag{2.28}$$

Remarque 2.3.1 *De (2.27) et (2.28) on remarque que l'ordre de la précesion du couverture pour $I_1(\alpha)$ dépend du signe de la variation régulière du second ordre .*

Théorème 2.3.2 *Comme la fraction de l'échantillons optimal en termes de probabilité de couverture dépend de quelque quantité inconnue, Cheng et Peng [5] proposent un estimateur Plug-in, en se concentrant sur l'intervalle de confiance $I_1(\alpha)$*

$$\hat{k}_{opt} = \begin{cases} \left\{ \frac{(1+2x_\alpha^2)}{3\hat{\delta}_n(1+2\hat{\rho}_n)} \right\}^{\frac{1}{(1+\hat{\rho}_n)}} n^{\hat{\rho}_n \frac{1}{(1+\hat{\rho}_n)}} & si \quad \hat{\delta}_n > 0 \\ \left\{ \frac{(1+2x_\alpha^2)}{-3\hat{\delta}_n} \right\}^{\frac{1}{(1+\hat{\rho}_n)}} n^{\hat{\rho}_n \frac{1}{(1+\hat{\rho}_n)}} & si \quad \hat{\delta}_n < 0 \end{cases}, \tag{2.29}$$

où

$$\hat{\rho}_n = -(\log 2)^{-1} \log \left| \frac{M_n^{(2)}(\frac{n}{(2\sqrt{\log n})}) - 2[M_n^{(1)}(\frac{n}{(2\sqrt{\log n})})]^2}{M_n^{(2)}(\frac{n}{(\sqrt{\log n})}) - 2[M_n^{(1)}(\frac{n}{(\sqrt{\log n})})]^2} \right|, \tag{2.30}$$

et

$$\hat{\delta}_n = \left\{ \frac{M_n^{(2)}(\frac{n}{(\sqrt{\log n})}) - 2[M_n^{(1)}(\frac{n}{(\sqrt{\log n})})]^2}{2\hat{\rho}_n[M_n^{(1)}(\frac{n}{(\sqrt{\log n})})]^2} \right\} (1+\hat{\rho}_n)(\sqrt{\log n})^{\hat{\rho}_n}, \tag{2.31}$$

avec

$$M_n^{(j)}(k) = \frac{1}{k} \sum_{i=1}^{k} (\log X_{n-i+1,n} - \log X_{n-k,n})^j, \ j = 1, 2. \tag{2.32}$$

Conclusion

Les résultats asymptotiques concernant les estimateurs de l'indice des valeurs extrêmes sont obtenus lorsque $k \longrightarrow \infty \quad et \quad \frac{k}{n} \longrightarrow 0,$ quand $n \longrightarrow \infty$La difficulté en pratique consiste à choisir le nombre d'extrêmes k utilisé dans l'estimation. Le choix de k est d'une grande importance en pratique, lorsque k est petit la variance de l'estimateur est grande et l'utilisation de grande valeur de k introduit un grand biais dans l'estimation, il s'agit donc d'effectuer un compromis entre biais et variance.

Dans ce mémoire nous avons donné un aperçu général sur les méthodes et approches existant et traitant le problème de choix du nombre optimal d'extrême à utiliser dans l'estimation de l'indice de queue. L'une des méthodes les plus utilisée est celle de Reiss et Thomas [15] dont il s'agit d'une manière automatique de choisir k_{opt} en minimisant

$$\frac{1}{k}\sum_{i\leq k} i^{\beta}\left|\hat{\gamma}_{i,n} - med(\hat{\gamma}_{1,n},...,\hat{\gamma}_{k,n})\right|,\ 0 \leq \beta \leq \frac{1}{2}.$$

Reiss et Thomas ont suggéré de minimiser la modification de précèdent ;

$$\frac{1}{k-1}\sum_{i<k} i^{\beta}(\hat{\gamma}_{i,n} - \hat{\gamma}_{k,n})^2,\ 0 \leq \beta \leq \frac{1}{2}.$$

Cette méthode fera l'objet d'un travail futur dont on suggère une étude de comparaison des différentes approches existant avec des applications sur des données simulées ou réelles.

Bibliographie

[1] Arnold, B.C., Balakrishan, N.and Nagaraja, H.N. (1992). A First Course in Order Statistics. Wiley, New York.

[2] Bateka, S. (2010). Détermination du nombre de statistiques d'ordre extrêmes. Memoire de magister, Université de Biskra.

[3] Beirlant, J., Goegebeur, Y., Segers, J. and Teugels, J. (2004). Statistics of Extremes - Theory and Application. Wiley.

[4] Benameur, S. (2010). Sur l'estimation de l'indice des valeurs extrêmes. Memoire de magister, Université de Biskra

[5] Cheng, S. et Peng, L. (2001). Confidence Intervals for the Tail Index. Bernoulli 7, 751-760.

[6] Coles, S. An Introduction to Statistical Modeling of Extreme Values. Springer.

[7] Danielsson, J., de Haan, L., Peng, L. et de Vries, C.G. (2001). Using a Bootstrap Method to Choose the Sample Fraction in Tail Index Estimation. Journal of Multivariate Analysis 76, 226-248.

[8] Delmas, J, F. (2013) . Introduction au calcul des probabilités et à la statistique. Les Presses de l'ENSTA.

[9] Dress, H. et Kaufmann, E. (1998). Selection of the Optimal Sample Fraction in Univariate Extreme Value Estimation. Stochastic Processes and their Application 75, 149-195.

[10] Embrechts, P., Klüpperlberg, C.and Mikosch,T. (1997). Modelling Extremal Event for Insurance and Finance. Springer, Berlin.

[11] de Haan, L et Ferreira, A. (2006). Extreme Value Theory : an introduction. Springer.

[12] Hall, P. et Welsh, A.H. (1985). Adaptive Estimates of Parameters of Regular Variation. Annals of Statistics 13, 331-341.

[13] Hall, P. (1982). Using the Bootstrap to Estimates Mean Squared Error and Select Smoothing Parameter in Nonparametric Problems. Journal of Multivariate Analysis 32, 177-203.

[14] Meraghni, D. (2008). Modelling distribution tails. Thèse de doctorat, Université de Biskra.

[15] Neves, C.et Fraga Alves, M.I. (2004). Reiss and Thomas' Automatic Selection of the Number of Extremes. Computational Statistics and Data Analysis 47, 689-704.

[16] Reiss, R.D. et Thomas, M. (1997). Statistical Analysis of Extreme Values with Application to Insurance, Finance, Hydrology and Other Fields. Birkhäuser, Basel.

[17] Resnick, S.I. (1987). Extreme Values, Regular Variation, and Point Processes. Springer, New York.

[18] Sousa, B. et Michailidis, G. (2004). A Diagnostic Plot for Estimating the Tail Index of a Distribution.Journal of Computational and Graphical Statistics, 13, 1-22.

Annexe : Abréviations et Notations

Les différentes abréVariations et notations utilisées tout au long de ce mémoire sont expliquées ci-dessous.

$(X_1, ..., X_n)$	échantillon de taille n de v.a's.
$\xrightarrow{P}$	convergence en probabilité.
$\xrightarrow{d}$	convergence en loi, convergence en distribution.
$\xrightarrow{p.s.}$	convergence presque sûre.
C_n^r	combinaison
$\mathbf{1}_A$	fonction indicatrice de l'ensemble A
$D(G_\gamma)$	domaine d'attraction
TVE	théorie des valeurs extrêmes
IVE	indice des valeurs extrêmes
GEV	distribution des valeurs extrêmes généralisée
GPD	distribution de paréto généralisée
$\mathbb{E}(X)$	l'esperance de X
$\hat{\gamma}_{k,n}^{(P)}$	estimateur de Pickands
$\hat{\gamma}_{k,n}^{(H)}$	estimateur de Hill

F	fonction de distribution de la variable X
$\bar{F}$	fonction de survie
$F^{\leftarrow}$	l'inverse généralisé de F
F_n	fonction de distribution empirique
f	densité de probabilité d'un va
$f_{X_{k,n}}$	fonction de densité de probabilité de $X_{k,n}$
$f_{(X_{i,n},\, X_{j,n})}$	fonction de densité jointe de $X_{i,n}$ et $X_{j,n}$
$i.i.d$	indépendantes et identiquement distribuées
G_γ	famille des lois des valeurs extrêmes généralisée
M_n	maximum de $X_1, ..., X_n$
$\mathcal{N}(0,1)$	loi normal standard
k	nombre de statistique d'ordre extrêmes
Q	fonction des quantiles
Q_n	fonction des quantiles empiriques
$\mathbb{R}$	ensemble des nombres réelles
x_F	point terminal de F
Φ_α	loi de Fréchet
Ψ_α	loi de Weibull
Λ	loi de Gumbel