République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
UNIVERSITÉ KASDI MERBAH OUARGLA

FACULTÉ DE MATHÉMATIQUES ET SCIENCES DE LA MATIÈRE

DÉPARTEMENT DE MATHÉMATIQUES

Mémoire présenté En Vue De L'obtention Du

DIPLÔME DE MASTER

EN MATHÉMATIQUES

Option : Probabilité et Statistique

Par

Abir HALASSA

Intitulé

L'estimation des indices extrêmes conditionelles dans le cas des données censurées

Membres du jury

Mohamed AGTI	M. A. A	UKMO	Président
Abdelmalek BOUSSAAD	M. A. A	UKMO	Examinateur
Fatima MEDDI	M. A. A	UKMO	Rapporteur

septembre 2020

Dédicace

Je dédie ce modeste travail

A ma très cher Mère et mon très cher Père
A mes chers sœurs et frères
A toutes la familles HLASSA

Sans oublier de dédier ce mémoire à mes très chères amies intimes
A tous mes collègues de ma promotion de Proba-Stat 2019

REMERCIEMENTS

Je remercie Dieu le tout-puissant de m'avoir donné la volonté, la force, et le courage patience et la santé pour bien mener et finir mon travail de mémoire.

Je voudrais d'abord et avant tout remercier ma encadreur vertueuse MEDDI fatima pour efforts en vue d'établir ce mémoire, elle a eu le rôle fondamental et essentiel et la grand mérite dans tout ce qui a été réalisé, me guider et corriger mes erreurs et me donner de précieux conseils et les conseils appropriés et les alertes considérés, je la répète mes remerciements pour tout ce qu'elle a fait pour moi à travers la mise en plan à toutes les étapes de la préparation de ce mémoire depuis le début jusqu'à la fin, étape par étape, était que le premier et dernier facteur pour le succès de ce travail, et je lui dis encore une fois merci beaucoup pour votre appréciation profonde.

Avec un grand honneur, j'aimerais présenter mes remerciements et ma gratitude aux membres du jury, Monsieur

Abdelmalek BOUSAAD,Mohamed AGTI , et Monsieur Said Zibar tout d'abord d'avoir accepté d'examiner mon mémoire, qui sans eux ce mémoire ne pourra jamais voir le jour pour l'intérêt et apport qu'ils ont apporté à mon travail.

J'exprime ma gratitude à ma famille qui m'a toujours soutenue et encouragée dans la voie que je m'étais fixée. Je remercie particulièrement mes parents qui m'ont stimulée et encouragé pendant mes études. qui étaient toujours prêts à fournir tous les moyens physique et morales pour la réussite de ce travail.

Table des matières

Table des figures

Liste des tableaux

0.1 Abréviations et Notations

EVD	Distribution des valeurs extrêmes
EVI, γ	Indice des valeurs extrêmes
F	Fonction de répartition
F_n	Fonction de répartition empirique
$F^{\leftarrow}$	Inverse généralisée de F
GEV	Distribution des valeurs extrêmes généralisée
GPD	Distribution de pareto généralisée
G_γ	Famille des la loi de valeurs extrêmes généralisée
$i.i.d$	Indépendantes et identiquement distribuées.
$\mathbb{I}_{\{A\}}$	Fonction indicatrice de l'ensemble A
Λ	Loi de Gumbel
$\ell(x)$	Fonction à variation lente
DA	Domaine d'attraction de maxcimum
$M_n = X_{n,n}$	Maxcimum de $X_1, ..., X_n$
VR	Vriation régulier
$v.a$	Variable aléatiore
$p.s$	Prèsque sûre
Φ	Loi de frêchet
Ψ	Loi de weibull
$resp$	Respectivement
$S = \overline{F}$	$1 - F$ fonction de survie
TEV	Théorème des valeurs extrêmes
$X_{1:n}, ..., X_{n,n}$	Statistique d'ordre associées à $X_1, ..., X_n$
$X \wedge Y$	$\min(X, Y)$
x_F	Point terminal
$\overset{\mathcal{L}}{=}$	Égalité en loi
$:=$	Égalité en définition
$\overset{D}{\rightarrow}$	Converge en distribution
$\overset{l}{\rightarrow}$	Converge en loi
$\overset{p}{\rightarrow}$	Converge en probabilité
$\overset{p.s}{\rightarrow}$	Converge prèsque sûre
$\overset{\rightarrow}{o_P}(.)$	Converge vers 0 en probabilité

$O_P(.)$	Être borné en probabilité
VR_α	Variation régulière d'indice α
$\hat{\gamma}^{(c,H)}$	Estimateur de Hill avec les données censurées
MSE	L'erreur quadratique moyenne
$Z^* = (Z_1^*, ..., Z_n^*)$	Échantillon Bootstrap
se	Erreur standard
$al.$	Autres
$(\Omega, \mathcal{A}, \mathbb{P})$	Espace probabilisé
$\sup A$	Supremum de l'ensemble A
$s.o$	Statistique d'ordre

0.2 Introduction

Dans la nuit du 31 janvier au 1 ér février 1953, de très fortes tempêtes traversant la mer du Nord balayèrent les côtes flamandes et néerlandaises d'ouest en est. Après que plusieurs digues eurent cédé, les provinces néerlandaises de la Hollande et de la Zélande furent particulièrement touchées. Les conséquences de ce raz-de-marée furent désastreuses. On dénombra plus de 2500 morts, 47000 habitations inondées et 10000 détruites, 200000 hectares de terres inondées, 30 000 têtes de bétail noyées, environ 9% des fermes des Pays-Bas inondées et plus de 400 brèches dans les digues.

A la suite de cette tempête, le gouvernement néerlandais décida la mise en place du "plan Delta" pour se prémunir contre une nouvelle inondation. Il fallut construire un nouveau réseau de digues renforcées de plus de 500 km le long de la côte de la mer du Nord.

Un comité composé de nombreux scientifiques, parmi lesquels des statisticiens, se réunit afin d'étudier le phénomène et proposer ainsi des recommandations sur les hauteurs des digues. Il fallut prendre en compte des facteurs économiques (coût de construction, coût des inondations,...), des facteurs physiques (rôle du vent sur la marée,...), mais aussi les hauteurs de marées enregistrées lors des précédentes inondations. En 1953, les inondations avaient entraîné une montée des eaux à 3.85 mètres au-dessus du niveau de la mer, soit largement en-deçà des 4 mètres atteints le 1 er novembre 1570, soit 382 ans auparavant.

Le but était de construire des digues assez grandes, de telle sorte qu'aucune vague ne les dépasse dans un horizon de 10000 ans. Autrement dit, il convenait de déterminer quelle serait la hauteur de la plus grande vague dans un horizon de temps de 10000 ans.

La difficulté résidait dans le fait que, pour calculer la hauteur des digues et donc la hauteur maximale d'une vague qui n'a jamais eu lieu, le comité d'experts devait se baser sur les informations des années précédentes ; or il n'y avait que très peu de données disponibles en particulier pour des événements de cette ampleur.

Le 4 octobre 1986 marqua l'achèvement du plan Delta aux Pays-Bas. Il s'agit du plus grand chantier de génie civil de tous les temps. Il permit de relier toutes les îles côtières de la province de Zélande par des digues.

Après calcul, le comité d'expert estima que ces digues devraient mesurer au moins 5 mètres, estimation fondée sur l'utilisation de techniques statistiques empruntées à la théorie des valeurs extrêmes qui constitue le sujet de la présente thèse.

Il est d'un grand intérêt de se prémunir contre les risques extrêmes, qu'ils résultent d'une crise financière, d'un accident nucléaire ou d'une catastrophe naturelle, compte tenu des répercussions humaines, économiques et financières que ces derniers peuvent avoir.

Les inondations de 1570 et de 1953 peuvent nous conduire à nous interroger sur une possible récurrence de ces événements. Le but serait d'être en mesure de prédire l'apparition de tels phénomènes, leurs impacts et le retour de ces derniers. Pour cela, il convient de définir précisément ce qu'est un événement extrême.

Un événement extrême est un événement qui a une faible probabilité de se produire mais qui, lorsqu'il se produit, prend de très petites ou de très grandes valeurs et a un grand impact. On notera la différence avec un événement rare qui, par définition, est un événement dont la probabilité d'occurence est faible. Le fait qu'un événement soit rare n'implique pas qu'il soit extrême ; il est dépourvu de la notion de quantifiabilité (petites ou grandes valeurs). A l'inverse, tout événement extrême est rare au sens où il a une faible probabilité de se produire.

La théorie des valeurs extrêmes trouve à s'appliquer dans de nombreux domaines tels qu'en fiabilité , en métallurgie [9] et en astrophysique [17]. Elle intéresse également les sciences de l'environnement, avec la modélisation de grands feux de forêts[2] ainsi que la climatologie[46] et la météorologie [[14], [15], [49]].

Dans un premier article, Einmahl et Magnus [21] proposèrent une application aux temps limites de records en athlétisme ; dans un second, ils s'intéressèrent plus particulièrement à l'estimation du temps minimal possible sur 100m [38].

Aarssen et de Haan proposèrent des résultats afin de calculer l'âge limite possible de l'être humain. Pour d'autres exemples d'applications, se référer au livre de Reiss et Thomas[43] . Au sujet de la mise en garde d'une mauvaise utilisation de la théorie des valeurs extrêmes, on citera Bouleau[11] .

Le domaine d'application historique reste l'hydrologie , notamment suite aux travaux de Jules Emile Gumbel en 1954[35] et son ouvrage [36]en 1958. Plus récemment l'utilisation de la théorie des valeurs extrêmes a été vivement recommandée par le rapport Flood Study Report NERC [41] afin de modéliser les lois de probabilités des maxima annuels de précipitations et de crues. Ce rapport, établi en 1975, utilise et teste de nombreux outils statistiques issus de la théorie des valeurs extrêmes sur un très grand nombre de jeux de données de crues de rivières et de pluviométrie en Grande-Bretagne.

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

Ce mémoire est réparti en trois chapitre.

Nous présentons dans 1^{eme} *chapitre* quelques rappels essentielles sur la théorie des valeurs extrêmes et la modélisation des valeurs extrêmes conditionnelles et de la notion de censure qui permettra de faciliter la lecture du mémoire. Ainsi, il s'agira de présenter brièvement les résultats essentiels rencontrés dans la littérature. Nous définissons rapidement les notions de domaine d'attraction, fonctions à variations régulières puis nous présentons ensuite quelques estimateurs de l'indice de queue, ainsi que ces propriétés asymptotiques. Puis nous définissions définir la fonction de répartition conditionnelle. Quand à la censure nous présenterons quelques définitions liées à la statistique des durées de survie.

Dans le 2^{eme} *chapitre,* nous considérons le problème de l'estimation de l'indice des valeurs extrêmes conditionnels pour covariable fixé et aléatoire pour des données complètes et censurées, nous présentons ensuite un aperçu historique des quelques estimateurs de l'indice de queue conditionnels, ainsi que leur propriétés asymptotiques dans le cas des données complètes et censurées. Puis nous présenterons la définition de l'éstimateur harmonique sur laquelle nous travaillerons dans le cas d'un échantillon conditionnel pour des donnés complètes et censuré, avant cela nous avons expliqué la méthode de fenêtre mobile pour sélectionner notre échantillon conditionnel par l'adaptation de l'estimateur de BEARAN .J et al 2013.
Pour la 3^{eme} *chapitre,* Nous présentons une définition simplifiée du logiciel R avec certains de ses avantages mentionnés et nous présentons le principe de bootstrap. Puis nous simulons les échantillons conditionnels et notre nouvel estimateur harmonique et nous appliquons la méthode du bootstrap sur ce dernier dans le cas des données complets et censurées pour une loi a variation régulière d'indice positife.Nnous avons présenteré à chaque algorithme sont programe corresspondant aboutissant des graphiques siginificatifs.

Notamment une distrubitiont limite empirique a eté donnée par la méthode du bootstrap dans le but de calculer des paramétres de diperission empirique comme le bias est l'erreure quadratique moiyenne.

Chapitre 1

Rappels sur la théorie des valeurs extrêmes et sur la censure

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

1.1 Comportement du maximum d'un échantillon

1.1.1 Densites et loi des statistiques d'ordre

Définition 1.1.1 *(Statistiques d'ordre).* soit $(X_1, ..., X_n)$ une suite de variable aléatoire indépendantes et identiquement distribuées, classée par ordre croissant. On écrit cette suite d'observation sous la notation $X_{i,n}$ tel que :

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

oú

-$X_{i,n}$:la $i^{éme}$ statistique d'orde (statistique d'ordre i) dans un échantillion de taille n.

-$X_{1,n}$:la plus petite valeur observée (oú statistique de minimum) avec

$$X_{1,n} = \min(X_{1,}..., X_n)$$

-$X_{n,n}$:la plus grand statistique d'ordre (oú statistique de maximum) avec

$$X_{n,n} = \max(X_1, ..., X_n)$$

Dans un échantillon de taille n, deux statistiques d'ordre sont particulièrement intéressantes pour l'étude des événements extrêmes, le minimum et le maximum : et

David [1970] et Balakrishnan et Clifford Cohen [1991] montrent que l'expression de la distribution de $X_{i,n}$ est

$$F_{i,n} = \mathbb{P}\{X_{i,n} \leq x\} = \sum_{r=i}^{n} \left(\begin{array}{c} n \\ r \end{array} \right) (F(x))^r (1 - F(x))^{n-r}.$$

Nous en déduisons que la fonction de densité est :

$$f_{i,n}(x) = \frac{n!}{(i-1)!\,(n-i)!}\left[F(n)\right]^{i-1}\left[1-F(x)\right]^{n-i} f(x),$$

où $f(x)$ est la densité de probabilité de X_i et F sa fonction de répartition associée. En utilisant la propriété d'indépendance des variables aléatoires $X_1, ..., X_n$, on obtient :

Loi de $X_{1,n}$.

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

d'où

$$f_{1,n}(x) = nf(x)(1 - F(x))^{n-1}.$$

Loi de $X_{n,n}$.

$$F_{n,n}(x) = \mathbb{P}\{X_{n,n} \leq x\} = (F(x))^n,$$

d'où

$$f_{n,n}(x) = nf(x)(F(x))^{n-1}.$$

Remarque 1.1.1

$$\mathbb{P}\{X_{n,n} \leq x\} = (F(x))^n \to 0 \text{ ou } 1 \text{ quand } n \to \infty.$$

Les expressions de $F_{1:n}$ et $F_{n:n}$ peuvent s'obtenir trés facilement en considérant les relations[37]

$$\begin{aligned}\{X_{1:n} \geq x\} &\Leftrightarrow \{\min(X_1, ..., X_n) \geq x\} \\ &\Leftrightarrow \bigcap_{i=1}^{n} \{X_i \geq x\}\end{aligned}$$

et

$$\begin{aligned}\{X_{n:n} \leq x\} &\Leftrightarrow \{\max(X_1, ..., X_n) \leq x\} \\ &\Leftrightarrow \bigcap_{i=1}^{n} \{X_i \leq x\}\end{aligned}$$

En utilisant la propriété d'indépendance des variables aléatoires $X_1, ..., X_n$ nous en déduisons que

$$\begin{aligned}F_{1:n}(x) &= \mathbb{P}\{X_{1:n} \leq x\} \\ &= 1 - \mathbb{P}\{X_{1:n} \geq x\} \\ &= 1 - \mathbb{P}\left\{\bigcap_{i=1}^{n} \{X_i \geq x\}\right\} \\ &= 1 - \prod_{i=1}^{n} \mathbb{P}\{X_i \geq x\} \\ &= 1 - \prod_{i=1}^{n} [1 - \mathbb{P}\{X_i \geq x\}] \\ &= 1 - [1 - F(x)]^n\end{aligned}$$

et

$$\begin{aligned} F_{n:n}(x) &= \mathbb{P}\{X_{n:n} \leq x\} \\ &= \mathbb{P}\left\{\bigcap_{i=1}^{n}\{X_i \leq x\}\right\} \\ &= \prod_{i=1}^{n}\mathbb{P}\{X_i \leq x\} \\ &= [F(x)]^n. \end{aligned}$$

Définition 1.1.2 (La fonction de répartition empirique). La fonction de répartition empirique de l'échantillon $(X_1, ..., X_n)$ notée F_n est donnée par :

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{]-\infty,x[}(X_i), \quad x \in \mathbb{R}$$

Il existe une autre version de la définition de F_n en utilisant les $(s.o)$ comme suit :

$$F_n(x) = \begin{cases} 0 & si, \quad x \leq X_{1,n} \\ \frac{i-1}{n} & si, \quad X_{i-1,n} < x \leq X_{i,n}, \quad 2 \leq i < n \\ 1 & si, \quad x > X_{n,n} \end{cases}$$

Définition 1.1.3 (Fonction de quantile)Pour tout $0 < s < 1$, la fonction de quantile associe à F est définie par

$$Q(s) = \inf\{t : F(t) \geq s\}, = F^{-1}(s)$$

oú F^{-1}représente la fonction inverse généralisée de F. On l'exprime en termes de la fonction de survie par :

$$Q(s) = \inf\{t : \overline{F}(t) < 1 - s\}, \quad 0 < s < 1.$$

Définition 1.1.4 (Quantile empirique) la fonction de quantile empirique *de l'échantillon $X_1, ..., X_n$ est défini par :*

$$Q_n(s) = F^{-1}(s) = \inf\{t : F_n(t) \geq s\}, \quad 0 < s < 1$$

oú

$$Q(s) = \overline{F}^{-1}(1 - s) = \inf\{t : \overline{F_n}(t) < 1 - s\}, \quad 0 < s < 1$$

1.1.2 Comportement asymptotique des extrêmes

On définit la variable aléatoire M_n, qui traduit le maximum d'un n-échantillon d'une variable aléatoire X, (les variables aléatoires X_i sont indépendantes et suivent la même loi que X) par :

$$M_n = \max(X_i)_{1 \leq i \leq n}$$

On pourrait aussi s'intéresser au minimum en utilisant la relation :

$$\min(X_i)_{1 \leq i \leq n} = -\max(-X_i)_{1 \leq i \leq n}$$

Le résultat central de la théorie des valeurs extrêmes concerne la distribution asymptotique du maximum (ou le minimum) en fonction de celle de la variable aléatoire X. Notons la fonction de répartition F_X de la variable aléatoire de loi de probabilité $\mathbb{P}$, à savoir $F_X(x) = \mathbb{P}(X < x)$. La fonction de répartition de M_n est alors définit par :

$$\begin{aligned} F_{M_n}(x) &= \mathbb{P}(M_n \leq x) \\ &= \mathbb{P}(X_1 \leq x, ..., X_n \leq x) \\ &= \mathbb{P}(X_1 \leq x) \times ... \times \mathbb{P}(X_n \leq x) \\ &= [F_X(x)]^n \end{aligned}$$

De ces résultats, nous tirons la conclusion que le maximum M_n est une variable aléatoire dont la fonction de répartition est égale à $(F_X)^n$. La fonction de répartition de X étant souvent inconnue et généralement pas possible d'être déterminée. Notons $x_F = \sup\{x \in \mathbb{R} : F_X(x) < 1\}$ le point terminal à droite (right-end point) de la fonction de répartition F_X. Ce point terminal peut être infini ou fini (Embrechts et al. 1997). On s'intéresse ici au distribution asymptotique du maximum, en faisant tendre n vers l'infini,

$$\lim_{n\to\infty} F_{M_n}(x) = \lim_{n\to\infty} [F_X(x)]^n = \begin{cases} 0 & \text{si } F(x) < 1 \\ 1 & \text{si } F(x) = 1 \end{cases}$$

On constate que la distribution asymptotique du maximum, donne une loi dégénérée, une masse de Dirac en x_F, puisque pour certaines valeurs de x, la probabilité peut être égale à 1 dans le cas où x_F est fini. et donc M_n tend vers x_F presque sûrement, Ce fait ne fournit pas assez d'informations, d'où l'idée d'utiliser une transformation afin d'obtenir des résultats plus exploitables pour les loi limites des maxima M_n. On s'intéresse par conséquent à une loi non dégénérée pour le maximum, la théorie des valeurs extrêmes permet de donner une réponse à cette problématique. Les premiers résultats sur la caractérisation du comportement asymptotique des maxima M_n convenablement normalisés et donnés par la suite.

1.2 Présentation de la théorie des valeurs extrêmes

1.2.1 Loi des valeurs extrêmes

Comme la fonction de répartition obtenue précédemment conduit à une loi dégénérée lorsque n tends vers l'infini, on recherche une loi non dégénérée pour le maximum de X. Cette loi limite non dégénérée est fournie par le "théorème des distributions extrêmes" qui donne une condition nécessaire et suffisante pour l'existence d'une loi limite non dégénérée pour le maximum. Ce théorème est proposé par Gnedenko (1943) qui donne la forme des lois limites et Jenkinson (1955) qui en donne l'expression générale.

Théorème 1.2.1 (Fisher et Tippett, 1928, Gnedenko, 1943). Soit $X_1, ..., X_n$ une suite de n variables aléatoires réelles indépendantes et identiquement distribuées de loi continue P et $M_n = \max(X_i)_{1\leq i\leq n}$. S'il existe deux suites réelles $(a_n)_{n\geq 1}$ et $(b_n)_{n\geq 1}$ avec $b_n > 0$, et une fonction de répartition non-dégénérée G_γ telle que,

$$\lim_{n\to\infty} \mathbb{P}\left[\frac{M_n - b_n}{a_n} \leq x\right] = \lim_{n\to\infty} F^n(a_n x + b_n) = G_\gamma(x) \quad \forall x \in \mathbb{R} \tag{1.1}$$

Alors G est du même type qu'une des trois lois suivantes *:*

$$\textit{loi de Gumbel : } \Lambda_\gamma(x) = \exp(-\exp(-x)) \qquad -\infty < x < +\infty \tag{1.2}$$

$$\textit{loi de Fréchet : } \Phi_\gamma(x) = \begin{cases} 0 & x < 0 \\ \exp\left(-x^{-1/\gamma}\right) & x \geq 0, \gamma > 0 \end{cases} \tag{1.3}$$

$$\textit{loi de Weibull : } \Psi_\gamma(x) = \begin{cases} \exp\left(-(-x)^{-1/\gamma}\right) & x < 0, \gamma < 0 \\ 1 & x \geq 0, \end{cases} \tag{1.4}$$

avec G_γ est la loi des valeurs extrêmes et γ est l'indice des valeurs extrêmes. a_n et b_n sont des paramètres de normalisation. Ce théorème est proposé par Gnedenko (1943) qui donne la forme des lois limites.

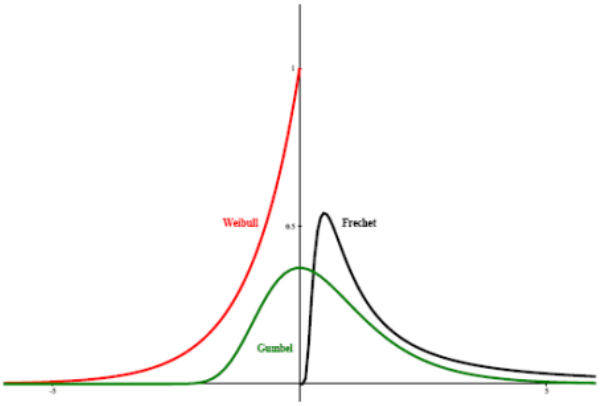


FIG. 1.1 – Densité de $\Phi_\gamma(x), \Psi_\gamma(x), \Lambda(x)$

Jenkinson (1955) donne l'expression générale notée GEV (Generalized Extreme Value Distribution) des trois distribution par,

$$G_\gamma(x) = \begin{cases} \exp(-(1+\gamma x)^{-1/\gamma}), \forall x \in \mathbb{R}, 1+\gamma x > 0 \text{ si } \gamma \neq 0 \\ \exp(-\exp(-x)), \quad \forall x \in \mathbb{R}, \quad \text{si } \gamma = 0 \end{cases} \tag{1.5}$$

1.2.2 Caractérisation des Domaines d'attraction

On définit les notions de fonctions à variations régulières et de fonctions à variations lentes qui nous seront utiles par la suite. Pour plus de détails, se référer à Bingham et al. [10] où de nombreux résultats sur les fonctions à variations régulières sont donnés.

Fonctions à variation régulière

Définition 1.2.1 . Une fonction mesurable $g : \mathbb{R}_+ \to \mathbb{R}_+$ est à variation régulière à l'infini si et seulement si, il existe un réel α tel que, pour tout $x > 0$

$$\lim_{t\to\infty} \frac{g(tx)}{g(t)} = x^\alpha.$$

Et on note $g \in VR_\alpha$, α est appelé indice (ou exposant) de la fonction à variation régulière.

Dans le cas particulier où $\alpha = 0$, on dit que g est à variation lente à l'infini, c'est à dire,

$$\lim_{t\to\infty} \frac{g\,(tx)}{g\,(t)} = 1 \quad \forall x > 0$$

Les fonctions à variation lente sont génériquement notées $\ell(x)$. Pour toute fonction à variation lente ℓ à l'infini, on a :

$$\lim_{x\to\infty} \frac{\log\,(\ell(x))}{\log\,(x)} = 0$$

Proposition 1.2.1 *Soient $\alpha \in \mathbb{R}$ et $g \in VR_\alpha$. Alors il existe une fonction à variation lente ℓ à l'infini telle que :*

$$\forall x > 0, \quad g(x) = x^\alpha \ell(x)$$

Théorème 1.2.2 *Une fonction $\ell : (0,\infty) \to (0,\infty)$ est à variation lente si et seulement si elle peut être écrite comme :*

$$\ell\,(x) = c\,(x) \exp\left\{\int_{x_0}^{x} \frac{\varepsilon\,(u)}{u} du\right\}, \qquad x \geq x_0,$$

pour certain $x_0 > 0$, où $\lim_{x\longrightarrow\infty} c\,(x) = c \in (0,\infty)$ et $\lim_{x\longrightarrow\infty} \varepsilon\,(x) = 0$.

Des exemples typiques de fonctions à variation lente sont des constantes positives ou des fonctions convergeant vers une constante positive, logarithmes, puissances des logarithmes et logarithmes itérés.

Définition 1.2.2 . Une fonction de répartition F sur $\mathbb{R}$ appartient à une classe à variation régulière VR s'il existe $\alpha \geq 0$ tel que $1 - F \in VR_{-\alpha}$ sur $\mathbb{R}$, ou d'une manière équivalente :

$$1 - F\,(x) \sim x^{-\alpha} \ell\,(x)\,, \text{ quand } x \to \infty$$

pour certaines $\ell \in RV_0$.

Domaine d'attraction de Fréchet

Théorème 1.2.3 *F appartient au domaine d'attraction de Fréchet avec un indice de valeur extrêmes $\gamma > 0$ si et seulement si $x_F = +\infty$ et $1 - F$ est une fonction à variation régulière d'indice $-1/\gamma$ c'est-à-dire,*

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

où ℓ est une fonction à variation lente. Dans ce cas, un choix possible pour les suites a_n et b_n est :

$$a_n = F^{-1}\left(1 - \frac{1}{n}\right) \quad \text{et} \quad b_n = 0.$$

Ce théorème permet de caractériser très simplement les distributions appartient au domaine d'attraction de Fréchet. En effet, elles doivent vérifier

$$\lim_{t\to\infty} \frac{1 - F(tx)}{1 - F(t)} = x^{-\alpha}$$

Prenons par exemple le cas de la distribution Pareto. Nous avons

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

Nous en déduisons que

$$\lim_{t\to\infty}\frac{1-F(tx)}{1-F(t)}=\lim_{t\to\infty}\frac{(tx)^{-1/\gamma}}{t^{-1/\gamma}}$$
$$=x^{-1/\gamma}$$

Donc $1-F\in RV_{-1/\gamma}$

Domaine d'attraction de Weibull

Théorème 1.2.4 *F appartient au domaine d'attraction de Weibul avec un indice de valeur extrême $\gamma<0$ si et seulement si $x_F<+\infty$ et $1-F^*$ est une fonction à variation régulière d'indice $1/\gamma$ c'est-à-dire,*

$$1-F=(x_F-x)^{-1/\gamma}\,\ell\left[(x_F-x)^{-1}\right].$$

avec,

$$F^*(x)=\begin{cases}0 & si \quad x\le 0,\\ F\left(x_F-x^{-1}\right) & si \quad x>0.\end{cases}$$

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

$$a_n=x_F-F^{-1}\left(1-\frac{1}{n}\right)\quad et\quad b_n=x_F.$$

Domaine d'attraction de Gumbel

Définition 1.2.3 Soit F une fonction de répartition de point terminal x_F fini ou infini. S'il existe $z<x$ tel que

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

où $c>0$ et a une fonction positive absolument continue de densité a' vérifiant $\lim\limits_{x\uparrow x_F}a'(x)=0$. Alors F est une fonction de Von-Mises et a est sa fonction auxiliaire.

Dans ce cas, un choix possible pour les suites (a_n) et (b_n) pour tout n >0 est :

$$a_n=F^{-1}\left(1-\frac{1}{n}\right)\text{ et }b_n=\frac{1}{\overline{F}(a_n)}\int_{a_n}^{yF}\overline{F}(z)dz.$$

Théorème 1.2.5 *F appartient au domaine d'attraction de Gumbel si et seulement si il existe une fonction de Von-Mises F^* telle que pour $z<x<x_F$ on ait :*

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

où $c(x)\to c>0$ lorsque $x\to x_F$.

Donc selon le signe de γ on distingue les trois cas de domaines d'attraction **(D.A.)** :

1. Si $\gamma > 0$, F appartient au **D.A. de Fréchet**, et l'on note $F \in D.A.(Fréchet)$**.** Il contient toutes les lois dont la fonction de survie décroit comme une fonction puissance. Ce sont les lois à **«queue lourde»**. Les distributions du domaine de Fréchet sont beaucoup utilisées en fiabilité mécanique, dans les phénomènes climatiques tels que la météorologie, l'hydrologie, la vitesse du vent enregistrée en continu dans les aéroports et en finance dans les études de risque.
2. Si $\gamma = 0$, F appartient au **D.A. de Gumbel**, et l'on note $F \in D.A.(Gumbel)$. Ce sont les lois dont la fonction de survie décroit vers zéro à une vitesse exponentielle.Ce sont les lois à **«queue légères**≫ Ces distributions sont souvent utilisées pour faire des prévisions dans les événements environnementaux tels que le séisme (le tremblement de terre), l'hydrologie (les inondations, la destruction des barrages), etc.
3. Si $\gamma < 0$, F appartient au **D.A. de Weibull**, et l'on note $F \in D.A.(Weibull)$. Ce domaine regroupe toutes les lois dont le point terminal, $x_F = \inf\{x, F(x) \geq 1\}$ est fini.Ce sont les lois à **«queue finie**≫ Les distributions de type de Weibull sont souvent utilisées pour décrire la résistance mécanique d'un matériau ou encore le temps de fonctionnement d'un appareil électronique ou mécanique.

Les Figures 1.1 et 1.2[13] ci-dessous illustre le comportement de différentes distributions GEV correspondant à différentes valeurs de γ. Les Tableaux (Tableau 1.1, Tableau 1.2, Tableau 1.3, ([50]) donnent quelques lois et leur domaine d'attraction.

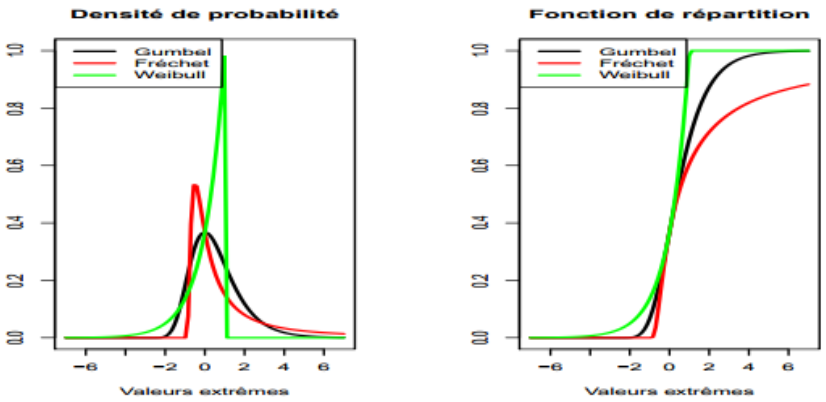


FIG. 1.2 – Représentation de la densité et fonction de répartition : Gumbel ($\gamma = 0$), Fréchet ($\gamma = 1$) et Weibull ($\gamma = -1$)

Distribution	$\mathbf{1-F(x)}$	$\boldsymbol{\gamma}$
Burr$(\beta,\tau,\lambda)\,,\beta>0,\tau>0,\lambda>0$	$\left(\frac{\beta}{\beta+x^{\tau}}\right)^{\lambda}$	$\frac{1}{\lambda\tau}$
Fréchet$\left(\frac{1}{\alpha}\right),\alpha>0$	$1-\exp\left(-x^{-\alpha}\right)$	$\frac{1}{\alpha}$
Loggamma$(m;\lambda)\,,m>0,\lambda>0$	$\frac{\lambda^m}{\Gamma(m)}\int_x^{\infty}\left(\log\left(u\right)\right)^{m-1}u^{-(\beta+1)}du$	$\frac{1}{\lambda}$
Loglogistic$(\beta,\alpha)\,,\beta>0,\alpha>0$	$\frac{1}{1+\beta x^{\alpha}}$	$\frac{1}{\alpha}$
Pareto$(\alpha)\,,\alpha>0$	$x^{-\alpha}$	$\frac{1}{\alpha}$

TAB. 1.1 – Quelques distributions appartenant au domaine d'attraction de Fréchet

Distribution	$\mathbf{1-F(x)}$	$\boldsymbol{\gamma}$
Gamma$(m,\lambda), m\in\mathbb{N},\lambda>0$	$\frac{\lambda^m}{\Gamma(m)}\int_x^{\infty}u^{m-1}\exp(-\lambda u)du$	0
Gumble$(\mu,\beta),\mu\in\mathbb{R},\beta>0$	$\exp(-\exp(-\frac{x-\mu}{\beta}))$	0
Logistic	$\frac{2}{1+\exp(x)}$	0
Log nomole$(\mu,\sigma),\mu\in\mathbb{R},\sigma>0$	$\frac{1}{\sqrt{2\pi}}\int_x^{\infty}\frac{1}{\mu}\exp(-\frac{1}{2\sigma^2}(\log u-\mu)^2du$	0
Weibull $(\lambda,\tau)\,,\lambda>0,\tau>0$	$\exp(-\lambda x^{\tau})$	0

TAB. 1.2 – Quelques distributions appartenant au domaine d'attraction de Gumble

Distribution	$\mathbf{1-F(x)}$	$\boldsymbol{\gamma}$
Uniforme $(0,1)$	$1-x$	-1
Reverse Burr$(\beta,\tau,\lambda,\tau_F),\beta>0,\tau>0,\lambda>0$	$\left(\frac{\beta}{\beta+(\tau_F-x)^{-\tau}}\right)^{\lambda}$	$-\frac{1}{\lambda\tau}$

TAB. 1.3 – Quelques distributions appartenant au domaine d'attraction de Weibull

Proposition 1.2.2 [10] *(Conditions du premier ordre, de Haan et Ferreira (2006)). Les assertions suivantes sont équivalentes :*

1. F est à queue lourde
$$F \in D.A(\text{fréchet}),\ \gamma > 0$$
2. $1 - F$ est une fonction à variation régulière à l'infini d'indice $-1/\gamma$
$$\lim_{t\to\infty} \frac{1 - F(tx)}{1 - F(t)} = x^{-1/\gamma},\ x > 0 \tag{1.6}$$
3. $Q(1 - s)$ est une fonction à variations régulières à zéro d'indice $-\gamma$
$$\lim_{s\to 0} \frac{Q(1 - sx)}{Q(1 - s)} = x^{-\gamma},\ x > 0$$
4. U est une fonction à variation régulière à l'infini d'indice γ
$$\lim_{t\to\infty} \frac{U(tx)}{U(t)} = x^{\gamma}\ ,\ x > 0$$

Proposition 1.2.3 [10] *(Conditions du second ordre de Haan et Ferreira (2006)). Une fonction de répartition $F(.) \in D.A(\text{fréchet}),\gamma > 0$, admet une condition du second ordre à l'infini si elle satisfait à l'une des assertions suivantes :*

1. Il existe un paramètre $\rho \leq 0$, et une fonction $A_1(.)$ qui tend vers 0 (ne change pas de signe à l'infini) définie par, $\forall x > 0$
$$\lim_{t\to\infty} \frac{(1 - F(tx))/(1 - F(t)) - x^{-1/\gamma}}{A_1(t)} = x^{-1/\gamma} \frac{x^{\rho} - 1}{\rho}$$
2. S'il existe un paramètre $\rho \leq 0$ et une fonction $A_2(.)$ qui tend vers 0 (ne change pas de signe à zéro) définie par, $\forall x > 0$
$$\lim_{s\to 0} \frac{Q(1 - sx)/Q(1 - s) - x^{-\gamma}}{A_2(s)} = x^{-\gamma} \frac{x^{\rho} - 1}{\rho},$$
3. S'il existe un paramètre $\rho \leq 0$, et une fonction $A(.)$ qui tend vers 0 (ne change pas de signe à l'infini) définie par, $\forall x > 0$
$$\lim_{t\to\infty} \frac{U(tx)/U(t) - x^{\gamma}}{A(t)} = x^{\gamma} \frac{x^{\rho} - 1}{\rho}$$

si $\rho = 0$, on remplace $(x^{\rho} - 1)/\rho$ par $\log x$
Les fonctions $A(.)$, $A_1(.)$, $A_2(.)$ sont à variations régulières à l'infini d'indices respectifs ρ, ρ/γ, et $-\rho$, avec $A_1(t) = A(1/(1 - F(t)))$ et $A_2(s) = A(1/s)$.

1.2.3 Estimation de l'indice des valeurs extrêmes γ

Dans la littérature de la TVE, il existe plusieurs méthodes et téchniques pour l'estimation de l'IVE, dans cette partie on reste limiter à trois méthodes.

Soit $X_1, .., X_n$ une suite de variable aléatoire indépendantes et identiquement distribuées et $X, ..., X_{n,n}$ les statistique d'ordre associées.$k = k_n$ une suite d'entier satisfaisant :

$$1 < k < n,\ k \longrightarrow \infty \text{ et } \frac{k}{n} \longrightarrow 0 \text{ quand } n \longrightarrow \infty \tag{1.7}$$

Estimateur de Pickands

L'estimateur de Pickands (1975) est le premier estimateur suggéré pour le paramètre $\gamma \in \mathbb{R}$ et il est donné par la formule suivante :

$$\hat{\gamma}^{p}_{k,n} = \frac{1}{\log(2)} \frac{X_{n-k,n} - X_{n-2k,n}}{X_{n-2k,n} - X_{n-4k,n}}$$

Théorème 1.2.6 *(Propriétés asymptotiques de $\hat{\gamma}^{p}_{n}$,[3])*

Pour $\gamma > 0$,$F \in DA(G_\gamma(x))$ et k que vérifie (1.7) on a :

(a) *consistance faible*

$$\hat{\gamma}^{p}_{k,n} \xrightarrow{p} \gamma \quad quand\ n \to \infty$$

(b) *consistance forte* $k/\log\log(n) \longrightarrow \infty, pour\ n \longrightarrow \infty$, alors

$$\hat{\gamma}^{H}_{k,n} \xrightarrow{p.s} \gamma \quad quand\ n \to \infty$$

(c) *Normalité asymptotique* sous certaines conditions sur k et F on a

$$\sqrt{k}(\hat{\gamma}^{p}_{k,n} - \gamma) \xrightarrow{D} N(0, \sigma^2) \quad quand\ n \to \infty$$

oú

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

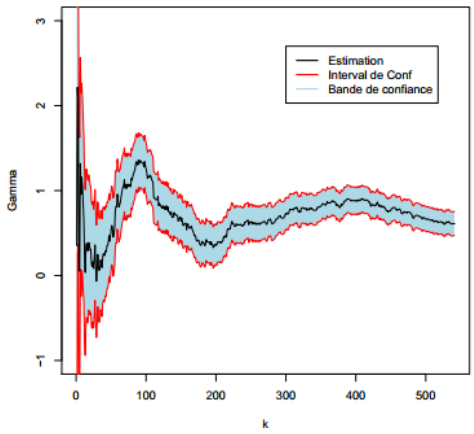


FIG. 1.3 – Comportement asymptotique de l'estimateur de Pickands avec des intervalles de confiance au niveau de confiance 0.95 sur les données d'assurance incendie Danoise pour la période allant de 01/01/1980 au 31/12/1990

Estimateur de Hill $\hat{\gamma}$

Cet estimateur qui a été présenté par *Hill* [39] en 1975 et l'estimateur le plus célébre parmi tous les estimateurs de l'indice de queue. Il s'applique seulement dans le cas oú l'indice de queue est positif ($\gamma > 0$), qui correspond aux distributions appartenant au domaine d'attraction de Fréchet. Cet estimateur donné par la d éfinition suivante :

Définition 1.2.4 *(Estimateur de Hill) :Pour* $\gamma < 0$ l'estimateur de Hill est défni par :

$$\hat{\gamma}_{k_n}^H = \frac{1}{k}\sum_{i=1}^{k} \log X_{n-i+1,n} - \log X_{n-k,n} \tag{1.8}$$

$$= \frac{1}{k}\sum_{i=1}^{k} i\,(\log X_{n-i+1,n} - \log X_{n-i,n})$$

1. Pour obtenir l'estimateur de $\hat{\gamma}_{k_n}^H$, en effet la condition(1.6) a une forme équivalente

$$\lim_{t\longrightarrow\infty} \frac{1}{\overline{F}(t)} \int_t^{\infty} \frac{F(x)}{x} dx = \gamma$$

Par l'integration par partie, on obtient

$$\lim_{t\longrightarrow\infty} \frac{1}{\overline{F}(t)} \int_t^{\infty} \log \frac{x}{t} dF(x) = \gamma$$

On remplace F par F_n et $t = X_{n-k,n}$, l'estimateur de Hill $\hat{\gamma}_{k_n}^H$défni par

$$\hat{\gamma}_{k_n}^H = \frac{1}{\overline{F}(X_{n-k,n})} \int_{X_{n-k,n}}^{\infty} \log \frac{x}{X_{n-k,n}} dF(x)$$

oú

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

Théoreme (Propriétés asymptotiques de $\hat{\gamma}_{k_n}^H$ **[40],**)

Pour $\gamma > 0$,$F \in DA(\Phi_\gamma)$ et k que vérifie (1.7) on a :

(a) *consistance faible*

$$\hat{\gamma}_{k_n}^H \xrightarrow{p} \gamma \quad quand\ n \to \infty$$

(b) *consistance forte*

$$\hat{\gamma}_{k_n}^H \xrightarrow{p.s} \gamma \quad quand\ n \to \infty$$

(c) *Normalité asymptotique*

Supposons que F $(\cdot)$ vérifie la condition (c) de la proposition **1.2.3** et $\sqrt{k}A(n/k) \longrightarrow \lambda$ si $n \to \infty$ alors

$$\sqrt{k}(\hat{\gamma}_{k_n}^H - \gamma) \xrightarrow{d} N(\frac{\lambda}{1-\rho}) \quad quand\ n \to \infty$$

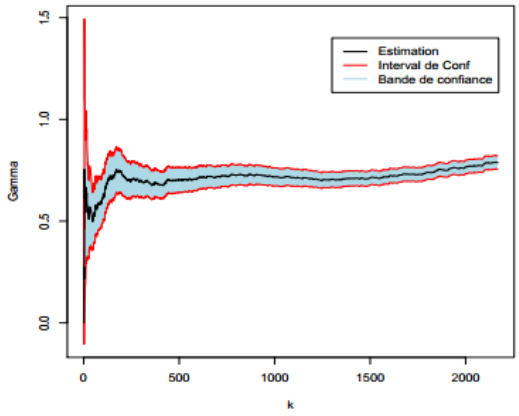


FIG. 1.4 – Comportement asymptotique de l'estimateur de Hill avec des intervalles de confiance au niveau de confiance 0.95 sur les données d'assurance incendie Danoise pour la période allant de 01/01/1980 au 31/12/1990.

Estimateur des Moments

Un autre estimateur qui peut être considéré comme une adaptation de l'estimateur de Hill, pour obtenir la consistance pour quelque soit le signe de l'indice, a étéproposé par *Dekkers et all* c'est l'estimateur de moment.

Définition 1.2.5 *(Estimateur des Moments de Dekkers et al)(1989,[13])*

pour $\gamma \in \mathbb{R}$,*l'estimateur de moment est défini par :*

$$\hat{\gamma}_n^M = M_{k,n}^{(1)} + 1 + \frac{1}{2}\left(1 - \frac{(M_{k,n}^{(1)})^2}{M_{k,n}^{(2)}}\right)^{-1}$$

$$= \hat{\gamma}_{k,n}^H + 1 + \frac{1}{2}\left(1 - \frac{(\hat{\gamma}_{k,n}^H)^2}{M_{k,n}^{(2)}}\right)^{-1}$$

$$M_{k,n}^{(r)} = \frac{1}{k}\sum_{ii=1}^{k}(\log X_{n-i+1,n} - \log X_{n-k,n}\ )^r \qquad r = 1, 2$$

Théorème 1.2.7 *(Propriétés asymptotiques de*$\hat{\gamma}_n^M$*,[40])*

Pour $\gamma > 0$,$F \in DA(\Phi_{1/\gamma})$ et k que vérifie (1.7) on a :

(a) *consistance faible*

$$\hat{\gamma}_n^M \xrightarrow{p} \gamma \quad quand\ n \to \infty$$

(b) *consistance forte :* si$(k/\log(n))^\delta \longrightarrow \infty$ qaund $n \longrightarrow \infty$ avec $\delta > 0$

$$\hat{\gamma}_n^M \xrightarrow{p.s} \gamma \quad quand\ n \to \infty$$

(c) *Normalité asymptotique*

Supposons que F $(\cdot)$ vérifie la condition (3) de la proposition **1.2.4** et $\sqrt{k}A(n/k) \longrightarrow \lambda$ si $n \to \infty$ alors

$$\sqrt{k}(\hat{\gamma}_n^M - \gamma) \xrightarrow{D} N(0, \eta^2\ ) \quad quand\ n \to \infty$$

oú

$$\eta \begin{cases} 1+\gamma^2 & \text{si } \gamma > 0 \\ (1-\gamma^2)(1-2\gamma)\left[4 - 8\frac{(1-2\gamma)}{(1-3\gamma)} + \frac{(5-11\gamma)(1-2\gamma)}{(1-3\gamma)(1-4\gamma)}\right] & \text{si } \gamma > 0 \end{cases}$$

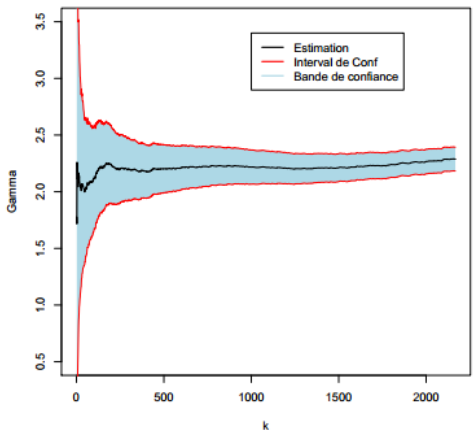


FIG. 1.5 – Comportement asymptotique de l'estimateur des moments avec des intervalles de confiance au niveau de confiance 0.95 sur les données d'assurance incendie Danoise pour la période allant de 01/01/1980 au 31/12/1990

Estimateur UH de Beirlant et al

Définition 1.2.6 ***(Estimateur UH de Beirlant et al)(1996,**[40]) :*

pour $\gamma \in \mathbb{R}$,l'estimateur de moment est défini par :

$$\widehat{\gamma}_{k,n}^H = \frac{1}{k}\sum_{i=1}^{k} \log UH_{(i)} - \log UH_{(k+1)}$$

$$\frac{1}{k}\sum_{i=1}^{k} \log(X_{(n-i,n)}\widehat{\gamma}_{i,n}^H) - \log(X_{(n-k-1,n)}\widehat{\gamma}_{k+1,n}^H)$$

oú

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

Théorème 1.2.8 ***(Propriétés asymptotiques de**$\hat{\gamma}_n^{UH}$**,[40])***

Supposons que F $(\cdot)$ vérifie la condition (**3**) de la proposition **1.2.4** $\sqrt{k}(A(n/k) \longrightarrow \lambda$ si $n \longrightarrow \infty$ alors

$$\sqrt{k}\left(\hat{\gamma}_n^{UH} - \gamma\right) \xrightarrow{D} N(\frac{\lambda}{1-\rho}, \sigma^2) \quad n \longrightarrow \infty$$

oú

$$\sigma^2 = \begin{array}{ll} 1+\gamma^2 & si\ \ \gamma \geq 0 \\ \frac{(1-\gamma)(1+\gamma+2\gamma^2)}{1-2\gamma} & si\ \ \gamma < 0 \end{array}$$

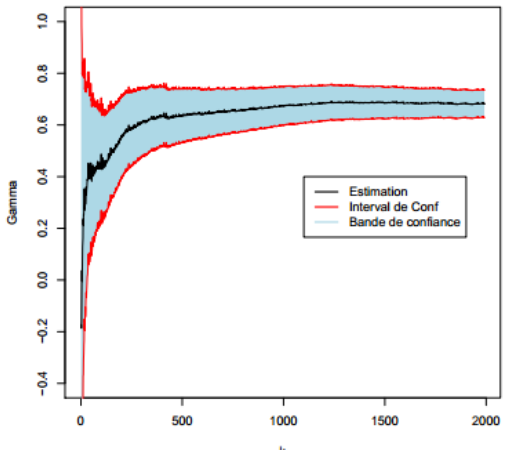


FIG. 1.6 – Comportement asymptotique de l'estimateur UH avec des intervalles de confiance au niveau de confiance 0.95 sur les données d'assurance incendie Danoise pour la période allant de 01/01/1980 au 31/12/1990.

1.3 Quelques rappeles sur la modélisation des valeurs extrêmes conditionnelles

1.3.1 Fonction de répartition conditionnelle

Soit Y une variable aléatoire réelle mesurée conjointement avec une covariable non-aléatoire $X \in X$, où X est un espace métrique muni d'une distance d. Soit { (X_i,Y_i), i = 1, . . . , n} un échantillon d'observations indépendantes et identiquement distribuées du couple $(X, Y) \in X \times R$. la fonction de répartition conditionnelle $F(x, y)$ de Y_i sachant X_i, donneée par

,

$$F(y \mid x) = p(Y \leq y \mid X = x)$$

et la fonction de survie conditionelle donneée par

$$\overline{F}(y \mid x) = p(Y > y \mid X = x)$$

dans le domaine de Fréchet où les queues de distribution sont lourdes : la fonction de distribution conditionnelle de Y sachant X se modélise alors de la façon suivante

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

où x est fixée dans un espace métrique muni d'une distance d, et $\ell(y; x)$ est une fonction à variations lentes.

1.3.2 Estimation de la fonction répartition conditionnelle

Méthode d'estimation indirecte

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

– Les estimateurs à noyaux qui peuvent être construits suivant :

(a) la méthode d'estimation du simple noyau ;

(b) la méthode d'estimation par noyau produit ou de Roussas (1969) ;

(c) la méthode du médianogramme ou de la fenêtre mobile.

– Les estimateurs au sens des plus proches voisins.
Nous nous intéressons à la premiére technique

Les estimateurs à noyaux

la méthode d'estimation du simple noyau :[1] l'estimateures par simple noyau de la fonction de répartiton condétionnelle défini par :

$$\widehat{F}_n(y \mid x) = \sum_{i=1}^{n} K(\frac{d(X_i, x)}{h_n}) \mathbb{I}\{Y_i \leq y\} \diagup \sum_{i=1}^{n} K(\frac{d(X_i, x)}{h_n}) \tag{1.9}$$

où K est une densité de probabilité appelée « noyau » et h_n un paramètre qui converge vers zéro lorsque n tend vers l'infini

la méthode d'estimation par noyau produit ou de Roussas (1969) :[1] L'estimateur de la densité conditionnelle étant défini par le rapport entre les estimateurs de la densité du couple f (x , y) et de la densité marginale g (x), il en découle que l'estimateur à noyau de la fonction de répartition conditionnelle[2] est :

$$\widehat{F}_n(y \mid x) = \int_{-\infty}^{y} \widehat{f}_n(u \mid x) du = \frac{\int_{-\infty}^{y} \widehat{f}_n(x, u) du}{\widehat{g}_n(x)} \tag{1.10}$$

ou

$$\widehat{f}_n(x, y) = \frac{1}{n h_n^{d*1}} \sum_{i=1}^{n} K_0(\frac{d(X_i, x)}{h_n}) K_1(\frac{d(Y_i, y)}{h_n}) \tag{1.11}$$

$$\widehat{g}_n(x) = \int_{\mathbb{R}^P} \widehat{f}_n(x, y) dy = \frac{1}{n h_n^d} \sum_{i=1}^{n} K_0(\frac{d(X_i, x)}{h_n}) \tag{1.12}$$

avec K_0 et K_1 des noyaux de probabilité

la méthode du médianogramme ou de la fenêtre mobile :[1] Cet estimateur est valable quel que soit le design. Pour un réel fixé $h_n > 0$, on se place en un point fixé x et on sélectionne les seuls Y_i pour lesquels les points d'observations X i (ou x_i) appartiennent à la boule centrée en x et de rayon h_n . Ceci revient alors à estimer la fonction de répartition conditionnelle par (1.9)en posant :

$$\begin{array}{l} w_{in}(x) = \mathbb{I}\{X_i \in B(x, h_n)\} \diagup \sum_{i=1}^{n} \mathbb{I}\{X_i \in B(x, h_n)\}, \text{(design aléatoire)} \\ w_{in}(x) = \mathbb{I}\{x_i \in B(x, h_n)\} \diagup \sum_{i=1}^{n} \mathbb{I}\{x_i \in B(x, h_n)\}, \qquad \text{(design fixe)} \end{array} \tag{1.13}$$

oú

$$w_{in}(x) = K(\frac{d(X_i, x)}{h_n}) \diagup \sum_{i=1}^{n} K(\frac{d(X_i, x)}{h_n}) \tag{1.14}$$

et B (x , h_n) est une boule centrée en x et de rayon $h_n \longrightarrow 0$ quand $n \longrightarrow \infty$.

1.3.3 Estimation des quantiles conditionnels

Estimation paramétrique des quantiles conditionnels

Elle est généralement utilisée quand l'on dispose d'un échantillon de petite taille. Afin d'estimer le quantile conditionnel, une façon de procéder consiste à supposer la fonction de répartition conditionnelle gaussienne. L'estimateur correspondant du quantile conditionnel $\widehat{q}(\alpha \mid x)$est défini par[1] :

$$\widehat{q}(\alpha \mid x) = \widehat{m}_n(x) + z_\alpha \widehat{\sigma}_n(x)$$

où $\widehat{m}_n(x)$(resp $\widehat{\sigma}_n(x)$) désigne l'estimateur de l'espérance conditionnelle (resp. l'écart type conditionnel) de Y sachant X =x et z_α le quantile d'ordre α de la loi normal centrée réduite.

Estimation non paramétrique des quantiles conditionnels

Dans le cadre de l'approche non-paramétrique, on distingue deux méthodes d'estimation. La première consiste à estimer au préalable la fonction de répartition conditionnelle puis à l'inverser pour en obtenir un estimateur du quantile conditionnel.. La seconde consiste quant à elle en une estimation directe basée sur le principe des moindres carrés.

Noud ne discuterons que de la premiére méthode

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

$$\widehat{q}(\alpha \mid x) = \inf\{y \mid \widehat{F}_n(y \mid x) \geq \alpha)$$

oú

$\widehat{F}_n(y \mid x)$ est Les estimateurs par simple noyau de la fonction de répartition conditionnelle qui nons donner don l'experéssion (1.9) ou L'estimateur de Roussas (1969) l'experéssion (1.10) ou L'estimateur de la fenêtre mobile l'experéssion (1.13)

1.4 Quelques généralités sur la censure

1.4.1 données de survie

L'analyse de survie, autrement dit la modélisation du temps de survenue d'un événement, apporte un outil principal d'évaluation théorique et pratique. L'analyse de ce type de données possède deux particularités intrinsèques, d'une part, celle ci ne concerne que des variables aléatoires positives et d'autre part, la présence de données non complètement observées comme nous l'expliquons ci dessous.

Désignons par X une variable d'intérêt, c'est à dire une variable aléatoire positive décrivant le temps qui s'écoule entre deux evénements par exemple [4]

- *En fiabilité* : durée de vie d'une lampe, durée de vie d'un matériel...
- *En biologie* : en culture de cellules les durées d'apparition de parasites...
- *En médecine* : durée de guérison d'un patient, durée de rémission d'un malade...
- *En economie* : durée de chômage...
- *En education* : durée d'apprentissage d'une langue...
- *En assurance* : durée de vie d'un contrat qui peut être défnie comme la diérence entre la date de résiliation et la date de création du contrat.

Définition 1.6. ***(Fonction de survie).*** *La fonction de survie qu'on note par S (t) ou* $\overline{F}(t)$ *est définie sur* R_+*par*

$$S(T) = \overline{F}(t)$$

$$= p(x > t) \ , \ t \geq 0$$

Pour t fixé c'est la probabilités de survivre jusqu'a l'instant t :

Définition 1.7. *La fonction de risque instantané, pour t tixé représente la probabilité de mourir dans un petit intervalle de temps après t, conditionnellement au fait d'avoir survécu jusqu'au temps t (c'est-à-dire le risque de mort instantané pour ceux qui ont survécu) :*

$$\lambda(t) = \lim_{h \to 0} \frac{\mathbb{P}(t \leq X < t + h \mid X \geq t)}{h}$$
$$= \frac{f(t)}{S(t)}$$

où f est la densité de probabilité de X

Définition 1.8. *(Fonctions empiriques de répartition et de survie)Soit X1;...,X*n un échantillon de taille n≥ 1 d'une variable positive X de fonction répartition F et de fonction de survie S . Les fonctions empiriques de répartition et de survie, Fn et Sn sont respectivement définies par

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left\{X_i \leq t\right\}, \quad \forall t \geq 0,$$

$$S_n(t) = 1 - F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left\{X_i \geq t\right\}, \quad \forall t \geq 0,$$

oú {A} est la fonction indicatrice de l'ensemble A

1.4.2 données censurées

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

Définition 1.10. *La variable de censure C est définie par la non-observation de l'événement étudié. Si au lieu d'observer X , on observe C, et que l'on sait que* $X > C$ *(respectivement* $X < C$*,* $C_1 < X < C_2$*, on dit qu'il y a censure à droite (respectivement censure à gauche, censure par intervalle).*

Caractéristiques

La censure est le phénomène le plus couramment rencontré lors du recueil de données en statistique. Pour un individu donné i, nous allons considérer [40]

. son temps de survie X_i de fonction de répartition F

. sa variable de censure C_i de fonction de répartition G

. sa variable réellement observée Z_i de fonction de répartition H.

Dans la littérature on distingue trois types de censure [40] :

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

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

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

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

La censure de type I

Définition 1.11. *La censure est dite non-aléatoire du type I si, étant donné un nombre positif fixé C et un n-échantillons $X_1, ..., X_n$ les observations consistent en (Z_i, δ_i) où*

$$\begin{cases} Z_i = X_i \wedge C \\ \delta_i = \mathbb{I}_{\{X_i \leq C\}} \end{cases}$$

La censure de type II

Elle est présente quand on décide d'observer les durées de survie des n patients jusqu'à ce que r d'entre eux soient décédés et d'arrêter l'étude à ce moment là. Soient $X_{(i)}$ et $Z_{(i)}$ les statistiques d'ordre des variables X_i et Z_i : La date de censure est donc $X_{(r)}$ et on observe les variables suivantes

$$\begin{aligned} Z_{(1)} &= X_{(1)} \\ Z_{(2)} &= X_{(2)} \\ &\vdots \\ Z_{(r)} &= X_{(r)} \\ Z_{(r+1)} &= X_{(r)} \\ &\vdots \\ Z_{(n)} &= X_{(r)} \end{aligned}$$

Définition 1.12. *La censure est dite non-aléatoire du type II si, étant donné un nombre positif fixé r et un n-échantillons $X_1, ..., X_n$ les observations consistent en (Z_i, δ_i) où*

$$\begin{cases} Z_i = X_i \wedge X_{(r)} \\ \delta_i = \mathbb{I}_{\{X_i \leq X_{(r)}\}} \end{cases}$$

La censure de type III : C'est la version aléatoire du type I
Définition 1.13. *La censure est dite aléatoire du type I si, étant donné un n-échantillons $X_1, ..., X_n$, il existe un v.a. n-dimensionnelle $(C_1, ..., C_n)$ de $(\mathbb{R}^+)^n$ telle que les observations consistent en (Z_i, δ_i) où*

$$\begin{cases} Z_i = X_i \wedge C_i \\ \delta_i = \mathbb{I}_{\{X_i \leq C_i\}} \end{cases}$$

1.4.3 Estimation de la fonction de survie

Dans la littérature plusieurs auteurs se sont intéressés pour l'estimation de la fonction de survie dans le cas où les données sont censurées. Parmi ces derniers nous pouvons citer *Kaplan* et *Meier* (1958) ont proposé un estimateur de la fonction de survie.

Estimateur de Kaplan-Meier

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

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

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

. Cet estimateur de Kaplan-Meier est une fonction étagée avec des sauts seulement aux observations non-censurées.

. La hauteur des sauts de cet estimateur est aléatoire.

. Quand toutes les observations sont non-censurées alors on obtient la fonction de répartition empirique.

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

Théorème 1.7 (Droesbeke et Saporta (2011)). Si les fonctions de répartition de la survie et de la censure n'ont aucune discontinuité commune, alors :

$$\sup_{t\geq 1}\left|\hat{S}_n(t)-S(t)\right| \overset{p.s}{\to} 0$$

et pour tout $t \geq 0$,

$$\sqrt{n}\left(\hat{S}_n(t)-S(t)\right) \overset{d}{\to} W_t$$

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

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

Chapitre 2

Estimation de l'indice des valeures extrêmes conditionnelles en présence de données Complétes et de données censurées

Dans ce chapitre, nous considérons le problème de l'estimation de l'indice des valeurs extrêmes conditionnels en présence de données compléts et de données censurées aléatoirement à droite. Les domaines privilégiés dans ce chapitre concernent ceux dont les fonctions de répartition sont à queues lourdes telles que l'hydrologie avec les inondations, la biologie avec la présence de données manquantes dans les bases de données, la finance avec les cracks boursiers, la télécommunication avec les trafics routiers, etc.

2.1 Estimation de l'indice des valeures extrêmes conditionnelles en design fixe en présence de données Complétes

La connaissance de γ est nécessaire pour résoudre un certain nombre de problèmes l'analyse des valeurs extrêmes. L'estimation de l'indice de queue a été largement étudiée dans la littérature et plusieurs estimateurs proposé. L'estimateur le plus populaire a été proposé par Hill (1975) dans le contexte de distributions à queue lourde. Diebolt et al. (2008) ont examiné le cas des distributions de queue de Weibull et le cas général a été étudié par Dekkers et al (1989).

Dans les applications pratiques, il arrive souvent que la variable d'intérêt Y puisse être lié à une covariable X Dans cette situation, l'indice de valeur extrême de la distribution conditionnelle de Y dépend de de la valeur x observée de la covariable$X$ et doit être mentioné dans le suivant ,comme indice de queue conditinnel. son estimation a été abordée dans la littérature récent sur les valeurs extrême, principalement dans le cas $\ll$ *design fixe* $\gg$, c'est-á-dire lorsque les covariables ne sont pas aléatoires.smith [48]et davison et smith ,[18] ont consideré un modèle de régression tandis que hall et tajvidi [8] ont utilisé une approche semi -paramétréque pour estimer l'indice de queue conditionnel . Des méthode entierement non paramétrique ont été envisagée en utilisant des splines.polynomes locaux une approche de fenêtre mobile .ou une aprroche de voisine.

2.1.1 Définitions des estimateurs

Soient $(x_i)_{1\leq i\leq n}$, un échantillon déterministe de X et $(y_i)_{1\leq i\leq n}$ des réalisations des variables aléatoires Y dans l'espace (Ω, A, p) et la fonction de répartition conditionnelle de Y sachant x et notée F (y, x) est à queue lourde. ,on veut définirquelque Estimateur de « indice de queue conditionnel » ou « indice des valeurs extrêmes conditionnel » et pour tout x non aléatoire

Estimateur de Hill adapté(1975,[38])

$$\gamma^H_{k_x m_{n,x}} = M^1_{k_x m_{n,x}} = \frac{1}{k_x}\sum_{i=1}^{k_x}\left(\frac{y^x_{(m_{n,x}-i+1)}}{y^x_{(m_{n,x}-i)}}\right) \ ,i = 1...m_{n,x} \tag{2.1}$$

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

$$\hat{\gamma}^M_{k_x m_{n,x}}(x) = M^{(1)}_{k_x m_{n,x}} + 1 - \frac{1}{2}\left(1 - \frac{(M^{(1)}_{k_x m_{n,x}})^2}{M^{(2)}_{k_x m_{n,x}}}\right)^{-1}, i = 1...m_{n.x} \tag{2.2}$$

où

$$M^{(2)}_{k_x m_{n,x}} = \frac{1}{k_x}\sum_{ii=1}^{k_x}\left(i\ \log\left(\frac{y^x_{(m_{n,x}-i+1)}}{y^x_{(m_{n,x}-i)}}\right)\right)^{\ 2}$$

Estimateur UH adapté (1996, [8])

$$\hat{\gamma}^{UH}_{k_x m_{n,x}}(x) = \frac{1}{k}\sum_{i=1}^{k}\log(y^x_{(m_{n,x}-i)}\widehat{\gamma}^H_{i,m_{n,x}}) - \log(y^x_{(m_{n,x}-k_x)}\widehat{\gamma}^H_{k_x,m_{n,x}}) \tag{2.3}$$

2.1.2 Propriétés asymptotiques de l'estimateur de l'indice

Pour tout $x \in X$,nous supposons les assertions suivantes :

a $F(.\mid x)$ et $G(.\mid x)$ sont absolument continues,

b Il existe une fonction $\rho(x) < 0$ et une fonction á variations réguliéres $b(.\mid x)$ d'indice $\rho(x)$ telles que pour tout $u > 0$,

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

c Il existe des constantes positives c_u , z_u et $\alpha_u \leq 1$ telles que pour tout $x \in B(x', 1)$

$$\sup_{z\geq z_u}\left|\frac{\log u(z\shortmid x)}{\log u(z\shortmid x')} - 1\right| \leq c_u d^{\alpha_u}(x, x')$$

oú $u(z\shortmid x) = \inf\left\{s; H(s\shortmid x) \geq 1-\frac{1}{z}\right\}$

Pour tout $x \in X$ nous supposons les assertions suivantes si $n \to \infty$,$k_x \longrightarrow \infty$, $k_x/m_{n,x} \longrightarrow 0$ et:

d $\sqrt{k_x}b\left(\frac{m_{n,x}}{k_x}, x\right) \longrightarrow \lambda(x) < \infty$

c $\frac{1}{\sqrt{k_x}}\sum_{i=1}^{k_x}\left[p_x\left(H^{\leftarrow}\left(1-\frac{i}{m_{n,x}}\mid x\right)\right)-p_x\right]\longrightarrow\varepsilon(x)<\infty$. Soient $c>0$ et $A(s,t):=\left\{1-k_x/m_{n,x}\leq t<1,|t-s|\leq c\sqrt{k_x}/m_{n,x},s<1\right\}$Nous assumons que si $n\to\infty$

e $\sqrt{k_x}\sup_{A(s,t)}\left|p_x\left(H^{\leftarrow}\left(t\shortmid x\right)\right)-p_x\left(H^{\leftarrow}\left(s\shortmid x\right)\right)\right|\longrightarrow 0.$

Sous ces conditions, nous avons les résultats asymptotiques de les estimateure

Théorème 2.1.1 *soit $x\in X$. Sous les conditions* ***c-e*** *et s'il existe des fonctions $m(.)$ et $\sigma(.)$ telles que $\sqrt{k_x}\left(\hat{\gamma}^{(.)}_{k_xm_{n,x}}(x)-\gamma(x)\right)\xrightarrow{d}N\left(m(x)\lambda(x),\sigma^2(x)\right)$, alors, nous avons :*

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

Corollaire 2.1.1 *Sous les hypothèses* ***a-e*** *et $k_x^{1/2}h_{n,x}^{\alpha_u}\longrightarrow 0$,nous avons*

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

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

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

2.2 Estimation de l'indice des valeures extrêmes conditionnelles en design aléatoire en présence de données Complétes

Soit Y une variable aléatoire ayant comme fonction de répartition F et Y des copies indépendantes de$Y$ Lorsqu'une covariable X est disponible et la distribution de Y dépend de X, le problème est d'estimer l'indice de queue conditionnel de la fonction de répartition conditionnelle $F(.\shortmid x)$ de Y sachant $X=x$. Motivés par les pluies extrêmes.Motivés par les pluies extrêmes mesurées conjointement avec leurs positions géographiques (longitude, altitude, latitude) considérées comme covariables, Gardes et Girard (2010, [27]) ont proposé un estimateur de l'indice de queue conditionnel et son quantile extrême conditionnel. En suivant la même logique, Ferrez et al. (2011, [25]) ont fait l'analyse des températures extrêmes avec des paramètres topologiques et Pisarenko et Sornette (2003, [42]) ont étudié les séismes extrêmes avec comme covariable leurs emplacements. L'estimation de l'indice des valeurs extrêmes conditionnel et quantiles extrêmes conditionnels avec covariables fixes (ou non aléatoires) a été étudiée assez largement dans la littérature récente de la statistique des valeurs extrêmes.

Nous nous référons à Beirlant et al. (2004,[6]), à Gardes et Girard (2008, [20]), à Gardes et al. (2010,[29]), à Stupfler (2013,[49]) pour un aperçu de la méthodologie disponible, y compris le cas où la variable explicative est fonctionnelle (Gardes et Girard, 2012, [28]). A ce jour, beaucoup d'auteurs s'attardent sur le cas à covariable aléatoire, en dépit de son intérêt pratique. Gardes et Stupfler (2014, [30]) et Goegebeur et al. (2014, [32]) adaptent l'estimateur

de Hill de l'indice de queue d'une distribution à queue lourde en présence d'une covariable aléatoire. L'estimateur des moments introduit par Dekkers et al. (1989,[19]) a été adapté aux covariables aléatoires par Goegebeur et al. (2014,[31]). Daouia et al. (2011, [16]) ont proposé un estimateur basé sur le noyau de quantiles extrêmes conditionnels avec des covariables aléatoires. Delafosse et Guillou (2002,[20]), Gomes et Oliveira (2003, [34]), Beirlant et al. (2007, [7]), Einmahl et al. (2008, [17]), Beirlant et al. (2010, [8]), Gomes et Neves (2011, [60]), Brahimi et al. (2013, [19]) et Worms et Worms (2014,[40]) ont proposé des estimateurs de l'indice des valeurs extrêmes conditionnel et parfois leurs quantiles extrêmes conditionnels. Par contre, Ndao et al. (2014,[40]), voir le chapitre précédent) ont proposé un estimateur de l'indice de queue conditionnel et un estimateur de quantile extrême conditionnel basé sur une adaptation de l'estimateur de Weissman (1978, [52]) et de la méthode de la fenêtre mobile

2.2.1 Définitions des estimateurs

Soient $(x_i)_{1\leq i\leq n}$, un échantillon déterministe de X et $(y_i)_{1\leq i\leq n}$ des réalisations des variables aléatoires Y dans l'espace (Ω, A, p) et la fonction de répartition conditionnelle de$Y$ sachant x et notée $F(y, x)$ est à queue lourde. ,on veut définirquelque Estimateur de « indice de queue conditionnel » ou « indice des valeurs extrêmes conditionnel » et pour tout$x$ aléatoire

(2014b, [59]) ont proposé un estimateur de l'indice des valeurs extrêmes conditionnel par la version du noyau de l'estimateur de Hill :

$$\gamma_{t_n}^H(x) = \sum_{i=1}^{n} K_h(x - X_i)(\log Z_i - \log t_n)\mathbb{I}_{\{z_i>t_n\}} \diagup \sum_{i=1}^{n} K_h(x - X_i)\mathbb{I}_{\{z_i>t_n\}} \tag{2.4}$$

oú $K_h(x) = h^{-p}K(x \diagup h), k$ est une densité de probabilité dans R^p ,$h = h_n$ est une suite positive non aléatoire appelée fenêtre telle que $h \to 0$ quand $n \longrightarrow \infty$ et t_n est une suite positive non aléatoire appelée seuil telle que $t_n \longrightarrow \infty$ si $n \to \infty$

2.2.2 Propriétés asymptotiques de l'estimateur de l'indice

Hypothèses lipschitziennes[40]

Pour tout$(x, x_0) \in X \times X$, nous supposons :

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

b Il existe $c_L > 0$ et u_0 tel que

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

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

Hypothèse classique (condition du second ordre)[10]

b Pour tout$x \in R$,il existe une fonction $\rho(x) < 0$ et une fonction à variations régulières à linfini $b(. \shortmid x)$ d'indice $\rho(x)$ telles que pour tout $u > 0$,

$$\lim_{t\longrightarrow\infty} \frac{H^{\leftarrow}\left(1 - \frac{1}{tu} \shortmid x\right) \diagup H^{\leftarrow}\left(1 - \frac{1}{t} \shortmid x\right) - u^{\gamma(x)}}{b(t \shortmid x)} = u^{\gamma(x)} \frac{u^{\rho(x)} - 1}{\rho(x)}$$

Hypothèse sur le noyau[40]

Proposition 2.2.1 *[10]Supposons les hypothèses* ***a-d*** *vérifiées. Soit* t_n *une suite non aléatoire telle que* $t_n \longrightarrow \infty$,$nh^p\overline{H}(t \mid x) \longrightarrow \infty$, $nh^{p+2}\overline{H}(t \mid x)(\log t_n)^2 \longrightarrow 0$ *et* $\sqrt{nh^p\overline{H}(t \mid x)b}\left(\frac{1}{\overline{H}(t_n \mid x)}\right) \longrightarrow 0$ *quand* $n \to \infty$.*Alors pour tout* $x \in X$

$$\sqrt{nh^p\overline{H}(t \mid x)}(\widehat{p_{t_n}}(x) - p_x) \xrightarrow{d} N(0, \frac{p_x(1-p_x)\left\|K\right\|_2^2}{g(x)} \quad \text{quand } n \longrightarrow \infty$$

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

Théorème 2.2.1 *Supposons les hypothèses* ***a-d*** *vérifiées. Soit* t_n *une suite non aléatoire telle que* $t_n \longrightarrow \infty$,$nh^p\overline{H}(t \mid x) \longrightarrow \infty$, $nh^{p+2}\overline{H}(t \mid x)(\log t_n)^2 \longrightarrow 0$ *et* $\sqrt{nh^p\overline{H}(t \mid x)b}\left(\frac{1}{\overline{H}(t_n \mid x)} \mid x\right) \longrightarrow 0$ *et* $\sqrt{nh^p\overline{H}(t \mid x)}b(t_n \mid x) \longrightarrow 0$ *quand* $n \to \infty$.*Alors, pour tout* $x \in X$ *tel que* $g(x) > 0$,

$$\sqrt{nh^p\overline{H}(t_n \mid x)}(\overline{\gamma}_{t_n}^{c.H}(x) - \gamma_1(x) \xrightarrow{d} N(0, \frac{\gamma_1^3(x)\left\|K\right\|_2^2}{\gamma(x)g(x)} \quad \text{quand } n \longrightarrow \infty$$

2.3 Estimation de l'indice des valeures extrêmes conditionnelles en présence de données censurées

Dans cette partie on va intéresser au problème de l'estimation de l'IVE mais cette fois c'est en présence de données censurées aléatoirement á droite. Ce problème,.est très récent dans la littérature, les premiers qui ont mentionné ce sujet sont Reiss et Thomas[43] en (2007) mais sans résultat asymptotique. En (2001) ; Beirlant et Guillou [5]ils ont proposé un estimateur mais pour les données tronquées. En 2007, Beirlant et al[4] ils ont introduit une méthode pour les estimateurs de Hill et de moment,... de plus, ils ont proposé les estimateurs des quantiles extrêmes et ont discuté leurs propriétés asymptotiques lorsque les données sont censurées pour un seuil déterministe et l'année suivante Einmahl et al.[23] ils ont adapté diffirents estimateurs de l'IVE au cas oú les données sont censurées par un seuil aléatoire et ils ont proposé une méthode unifiée pour établir leur normalité asymptotique.

Soit F et G sont supposées être dans le dommaine d'attraction maximales $D(G_{\gamma_1})$ et $D(G_{\gamma_2})$ respectivement où $\gamma_1, \gamma_2 \in \mathbb{R}$,l'échantillon $X_1 ... X_n$ et $Y_1 ... Y_n$. soit $\{(Z_i, \delta_i),\ 1 \leq i \leq n\}$ l'échantillon réellement observé oú $Z_i = X_i \wedge Y_i$ et $\delta_i = \mathbb{I}_{\{X_i \leq Y_i\}}$, avec points terminales τ_F et τ_G, où $\tau_F = \sup\{x, F(x) < 1\}$, alors cela signifie que $H \in D(G_\gamma)$. *Einmahl et al.*(2008, [7]) ont examiné les trois cas les plus intéressants suivants:

$$\begin{cases} cas\ 1 & : & \gamma_1 > 0, \gamma_2 > 0 & & \gamma = \frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2} \\ cas\ 2 & : & \gamma_1 < 0, \gamma_2 < 0 & \tau_F = \tau_G & \gamma = \frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2} \\ cas\ 3 & : & \gamma_1 = \gamma_2 = 0 & \tau_F = \tau_G = \infty & \gamma = 0 \end{cases} \tag{2.5}$$

les estimateurs de l' IVE de données censuré sont basés sur un estimateur standard de l'indice de queue divisé par l'estimateur de la proportion de données non censurées dans le plus grand k de Z

$$\hat{\gamma}_{X,k,n}^{(c,.)} = \frac{\hat{\gamma}_{Z,k,n}^{(.)}}{\hat{p}} \tag{2.6}$$

oú

$$\hat{p}_x = \frac{1}{k_x}\sum_{i=1}^{k_x}\delta_{[m_{n,x}-i+1]}^x \quad \text{pour la covariable } x \text{ fixe}$$

et

$$\hat{p}_{t_n}(x) = \frac{\overline{H}_n^1(t_n \shortmid x)}{\overline{H}(t_n \shortmid x)}.\text{pour la covariable } x \text{ aléatoire}$$

avec $\overline{H}(t_n \mid x) = \sum_{i=1}^{n} B_i(x)\mathbb{I}_{\{z_i > t_n\}}$, $\overline{H}^1(t_n \shortmid x) = \sum_{i=1}^{n} B_i(x)\mathbb{I}_{\{z_i > t_n, \delta_i = 1\}}$ et les poids de Nadarya-Watson définis par

$$B_i(x) = K(\frac{x - X_i}{h}) \diagup \sum_{i=1}^{n} K(\frac{x - X_i}{h})$$

donc on va défini l'éstimateure de l'indice de valeure extrême conditionelle en présence de données censuré

Soit $(x_i)_{1 \leq i \leq n}$, un échantillon déterministe de X dan l'espace métrique (x, d), $(y_i)_{1 \leq i \leq n}$ et $(c_i)_{1 \leq i \leq n}$ des réalisations respectives des variables aléatoires Y et C dans l'espace probabilisé $(\not\leqslant A\ P)$ avec F et G respectives la fonctions de répartition conditionnelles. Soient $(Z_i)_{1 \leq i \leq n} = (Y_i \wedge C_i)_{1 \leq i \leq n}$ et $\delta_i = \mathbb{I}_{\{Y_i \leq C_i\}}, \mathbb{I}_{\{.\}}$ est une fonction indicatrice avec Y et C sont indépendantes conditionnellement à X

– l'éstimateure de l'indice de valeure extrême conditionelle en design fixe en présence de données censuré défini par :

$$\hat{\gamma}_{X,k,n}^{(c,.)} = \frac{\hat{\gamma}_{k_x m_{n,x}}^{(.)}}{\hat{p}_x} \tag{2.7}$$

oú

$$\hat{p}_x = \frac{1}{k_x} \sum_{i=1}^{k_x} \delta_{[m_{n,x} - i + 1]}^{x}$$

et $\hat{\gamma}_{k_x m_{n,x}}^{(.)}$ est l'estimateure de l'indice de valeure extrême conditionelle qui définir pour Estimateur de Hill dan léxpritiont(2.1) ou Estimateur de Dekkers-Einmahl-de Haan dan léxpritiont(2.2) ou Estimateur UH dan l'éxprétiont (2.3)

– l'éstimateure de l'indice de valeure extrême conditionelle en desing aléatoire en présence de données censuré défini par :

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

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

2.4 Censtruction d'un nouvel estimateure de l'indice des valeures extrêmes conditionelles

Avant de parler de nouvelle estimateure nous allons voir L'estimateur d'indice de queue de moment harmonique qui nous allons defini sous

Définition 2.4.1 *(Estimateurs de l'indice de queue de moment harmonique) soit $X_1, ..., X_n$ suit de veriable aleatoire i.i.d avec F la distributions de paréto*

$$H^{(\beta)} = \frac{1}{\beta - 1} \left\{ \left[k^{-1} \sum_{i=1}^{n} \left(\frac{X_{n-k,n}}{X_{n-i+1,n}} \right)^{\beta-1} \right] - 1 \right\}$$

oú $1 \leq k \leq n-1$ et $\beta > 0$ est paramétre de réglage .pour $\beta = 1, H_{n,k}^{(\beta)}$ est interprété comme limlte pour $\beta \to 1$,i.e

$$H_{n,k}^{(1)} = \lim_{\beta \to 1} H_{n,k}^{(\beta)} = k^{-1} \sum_{i=1}^{k} \log(X_{n-i+1,n/X_{n-k,n}})$$

dans cette parté nous nous proposons d'étudier un nouvel estimateur de l'indice de valeure extrême conditinelle en présence de données complete et de données censuré. Pour cela nous'allons pas utiliser toutes les données mais simplement les observations dont leurs covariables sont proches de x Ainsi, l'approche que nous allons utiliser est la méthode de la **fenêtre mobile** comme dans Gardes et Girard (2008,), définie par une boule, $B(x, h_{n,x})$ où x est le centre de la boule et $h_{n,x}$ est le rayon de la boule :

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

Cette méthode renferme trois étapes :

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

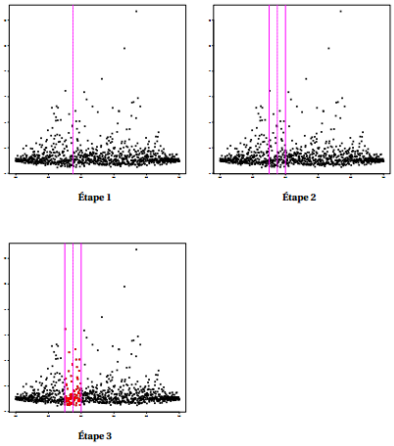


FIG. 2.1 – Les différentes étapes de sélection des données. En ordonnée on a la variable Y et en abscisse la covariable X.

2.4.1 Sélection de l'échantillon conditionnele

Soit $Y \in R$ une variable aléatoire associée à une covariable non-aléatoire $x \in E$, où E désigne un espace métrique, non nécessairement de dimension finie. Soient $\{(X_i, Y_i),\ i = 1, ..., n\}$ des observations indépendantes du couple $(x, Y) \in E \times R$

Nous nous plaçons dans le domaine de Fréchet où les queues de distribution sont lourdes, la fonction de distribution conditionnelle de$Y$ sachant Xse modélise alors de la façon suivante

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

oú $\gamma(x)$ est de paramètre fonctionnel positif de la covariable x fixé ,$\ell(.,x)$ est une fonction à variations lentes qui verifier ,pour $\lambda > 0$

$$\lim_{n\longrightarrow\infty} \frac{L(\lambda y, x)}{L(\lambda, x)}$$

nous utilisons une méthode d'estimation dite de la fenêtre mobile pour construire nos estimateurs. Pour cela, on introduit une boule centrée en x , de rayon $h_{n,x} > 0$, notée $B(x, h_{n,x})$ et définie par :

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

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

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

Le résultat de lobservation sample dan $(0,\infty) \times B(x, h_{n,x})$ donné par

$$\{Z_i(x), i = 1, ...m_{n,x}\}$$

oú $m_{n,x} = n\varphi(h_{n,x})$ et $Z_{1.m_{n,x}} \leq ... \leq Z_{m_{n,x},m_{n,x}}$ les statistiques ordonnées correspondantes

2.4.2 Estimateure pour des données complétes

pour le même échantillon consider dans la sous section ci -dessus oú $l(y,x) = 1$ et pour un seuil arbitaire $z_0 > 1$ soit

$$T = \frac{Z}{z_0}\mathbb{I}_{\{Z>z_0\}}$$

puis conditionnellement sur $Z > z_0$,

$$p(T \leq t / Z > z_0) = 1 - t^{-1/\gamma(x)}$$

De plus conditionellemnt $Z > z_0$, on a

$$T^{-1/\gamma(x)} = 1 - F(T) \overset{d}{=} U$$

oú U uniformément distrubué sur [0,1]. Étant donné une séqunce iid $T_1...T_k$ des dépssements relatifs sur un seuil z_0,la loit fort des grands nembres implique que

$$\frac{1}{k_{n,x}}\sum_{i=1}^{k_{n,x}} T_i^{1-\beta} = \frac{1}{k_{n,x}}\sum_{i=1}^{k_{n,x}} \left(T_i^{-1/\gamma(x)}\right)^{\gamma(x)(\beta-1)} \overset{d}{=} \frac{1}{k_{n,x}}\sum_{i=1}^{k_{n,x}} U_i^{\gamma(x)(\beta-1)}$$

$$\overset{\alpha.s}{\longrightarrow} E[U^{\gamma(x)(\beta-1)}] = \int_0^1 u^{\gamma(x)(\beta-1)}du = \frac{1}{\gamma(x)(\beta-1)+1}$$

á condition que $\gamma(x)(\beta-1)+1 \neq 0$

$$\frac{1}{\beta-1}\left(\frac{1}{k\sum_{i=1}^{k}T_i^{1-\beta}}-1\right)\xrightarrow{\alpha.s}\gamma(x)$$

$$\widehat{\gamma}=\frac{1}{\beta-1}\left\{\frac{1}{k_{n,x}}\sum_{i=1}^{k_{n,x}}\left[\frac{Z_{m_{n,x}-k_{n,x},m_{n,x}}(x)}{Z_{m_{n,x}-i+1,m_{n,x}}(x)}\right]^{\beta-1}-1\right\} \tag{2.8}$$

oú $1\leq k\leq m_{n,x}-1$ et $\beta>0$ est paramétre de réglage.

2.4.3 Estimateure pour des données censuré

Modèle et notation

On considère l'espace probabilisé (Ω,A,P) et l'espace métrique (x,d) , soient 'échantillon $(x_i)_{1\leq i\leq n}$ un échantillon déterministe de X , $(Y_i\ )_{1\leq i\leq n}$ et $(C_i)_{1\leq i\leq n}$ réalisations respectives des variables aléatoires Y et C dans l'espace(Ω,A,P).Nous supposons que Y et C ont respectivement comme fonctions de répartition conditionnelles F et G. soit $(Z_i)_{1\leq i\leq n}=(Y_i\wedge C_i)_{1\leq i\leq n}$ et $\delta_i=\mathbb{I}_{\{Y_i\leq C_i\}}$, oú $\mathbb{I}_{\{.\}}$ est une fonction indicatrice. Nous supposons aussi que Y et C sont indépendantes conditionnellement à X . Nous observons,$(z_1,\delta_1,x_1)...(z_n,\delta_n,x_n)$ des copies indépendantes de (Z,δ,X) . Dans la suite nous supposerons que la variable Z a comme fonction de répartition conditionnelle H.nous considérons que les fonctions de répartition conditionnelles de$Y$ et C pour tout $x\in X$ sont ainsi définies :

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

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

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

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

Si Y et$C$ sont indépendantes, alors la fonction de répartition conditionnelle$H(\cdot|x)$ de Z pour $X=x$ fixé, $\gamma(x)=\gamma_1(x)\gamma_2(x)\diagup(\gamma_1(x)+\gamma_2(x))$.Ce qui donne pour tout u et x

$$1-H(u\mid x)=(1-F(u\mid x))(1-G(u\mid x)).$$

$$=u^{-1/\gamma_1(x)}L_1(u,x)u^{-1/\gamma_2(x)}L_2(u,x)$$

$$u^{-(\frac{1}{\gamma(x)_1}+\frac{1}{\gamma(x)_2})}L_1(u,x)L_2(u,x)$$

$$u^{-(\frac{\gamma_1(x)+\gamma_2(x)}{\gamma_1(x)\gamma_2(x)})}L(u,x)$$

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

où $\gamma(x) = (\gamma_1(x)\gamma_2(x))/(\gamma_1(x)+\gamma_2(x))$ et $L(x) = L_1(x)\, L_2(x)$). Donc H est une fonction de répartition appartenant au domaine d'attraction de Fréchet

$$1 - H(u \mid x) \in RV(-1/\gamma(x)), \quad \text{avec } \gamma(x) = (\gamma_1(x)\gamma_2(x))\diagup((\gamma_1(x) + \gamma_2(x)) \ .$$

pour obtenu le neveau l'éstimateure pour de donne consuré on devisé 2.8 sur p_x

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

ou

$$\hat{p}_x = \frac{1}{k_x}\sum_{i=1}^{k_x} \delta^x_{[m_{n,x}-i+1]} \quad \text{pour la covariable } x \text{ fixe}$$

$$\hat{p}_{t_n}(x) = \frac{\overline{H}^1_n(t_n \shortmid x)}{\overline{H}(t_n \shortmid x)}. \text{pour la covariable } x \text{ aléatoire}$$

Chapitre 3

Etude de simulation

Dans ce chapiter nous consedérons la simulation de notre nouveau estimateure $\widehat{\gamma}$ dans le cas des données complet et censureés et nous appliquons le méthode du bootstrap pour une étude asymptotique

3.1 introduction du logiciel R

3.1.1 Qu'est-ce que le logiciel R ?

R est un logiciel permettant de faire des analyses statistiques et de produire des graphiques. Mais R est également un langage de programmation complet, c'est cet aspect qui fait que R est différent des autres logiciels statistiques. Les informations sur R sont disponibles sur la homepage du projet

http ://www.r-project.org/

R est un clône gratuit du logiciel S-Plus commercialisé par MathSoft et développé par Statistical Sciences autour du langage

3.1.2 les aventage de logiciel R

- permet de faire des graphiques très flexibles et d'une qualité exceptionnelle
- permet de construire facilement vos propres fonctions
- est facile à interfacer avec d'autres langages
- est aussi un logiciel mathématique (calcul matriciel, intégration numérique, optimisation, ..)
- les graphes sont bien plus jolis
- R est très facile à utiliser
- R est gratuit

3.2 Princip du Bootstrap

Le principe fondamental de cette technique de ré-échantillonnage est de substituer à la distribution de probabilité inconnue F, dont est issu l'échantillon d'apprentissage, la distribution empirique F_n qui donne un poids $\frac{1}{n}$ à chaque réalisation. Ainsi on obtient un échantillon de taille n dit échantillon bootstrap selon la distribution empirique F_n par n tirages aléatoires avec remise parmi les n observations initiales. Si n est grand, la distribution empirique $\hat{F}_n$ est proche de F, on aura donc une bonne approximation de la loi de X en utilisant $\hat{F}_n$ à la place de F. La méthode de bootstrap consiste à construire un nombre B (B entier)

d'échantillons bootstrap notée X^* (images de l'échantillon initial), afin de les utiliser pour faire des inférences, et plus le nombre d'images simulées est grand, plus la statistique est précise, pour chaque nouvel échantillon, on calcule de la même façon un nouvel estimateur (image simulée de l'estimateur initial). L'ensemble des images simulées de l'estimateur initial est considéré comme un modèle de sa distribution sur la population de l'échantillon initial.

3.3 Simulation pour des données complétes

3.3.1 Échantillon initial et paramètres de simulation

La loi de simulation utilisée dans ce cas est une loi de Pareto de paramètre γ de fonction de répartition conditionel,

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

Nous avons généré un échantillon $(Y_i)_{1\leq i\leq n} \frown \text{Pareto}(\gamma_1)$ de taille $n = 1000$, à partir d'une variable u de $U([0,1])$, le modèle ajusté sera :

$$F^{-1}(u) = (1-u)^{-\gamma_1}$$

les covariables $(X_i)_{1\leq i\leq n} \frown \text{Pareto}(\gamma_2)$ de taille $n = 1000$,à partir d'une variable v de $v([0,1])$, le modèle ajusté sera :

$$H^{-1}(v) = (1-v)^{-\gamma_2}$$

Par la méthode de fênétre mobile en obtient l'échantillon conditionnel Z

3.3.2 Seléctionement de l'échantillon condétionel

pour selectioner l'échentllion condétionelle on utulise la méthode de fenêtre mobille
Algorithme1

1. Nous avons génééré un échantillon Yet une covariable X suit la loit de préto de taill n=1000 de parametr $\gamma_1 = 1.5$ (resp $\gamma_2 = 0.6$),á partir d'une variable u de U ([0 ; 1]) resp(v de U ([0 ; 1])) , le modele ajusté sera :

$$F^{-1}(u) = 1 - u^{-1/\gamma}$$

$$F^{-1}(v) = 1 - v^{-1/\gamma}$$

1. On fixé une valeur $x1$ déterministe, à partire de la covariable x
2. Puis on choisit les variable x_i qui vérifier

$$d(x, x1) \leq h_{n,x} \tag{3.1}$$

oú $h_{n,x} = 0.3$
et on détermine la taill de x_i, $m_{n,x} = n\varphi(h_{n,x})$ oú

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

Enfin on sélectionne uniquement les observations (y_i) pour lesquelles x_i vérifiont la condition 3.1

Programme sous R

```
n= ?
gamma1= ?
gamma2=0.6
u<-runif(n)
v<-runif(n)
y<-(1-u)^(-gamma1)
x<-(1-v)^(-gamma2)
x1<-sample(x,1)
d<-numeric(n)
b<-numeric(n)
h<-(0.3)
for (i in 1 :n)
{
d[i]<-abs(x[i]-x1)
if (d[i]<h)
b[i]<-1 else b[i]<-0
}
Q<-(sum(b)/n)
m<-(n*Q)
z<-numeric(m)
for (i in 1 :n)
{
d[i]<-abs(x[i]-x1)
if (d[i]<h)
z[i]<-y[i] else z[i]<-0
}
z
z<-z[z!=0]
```

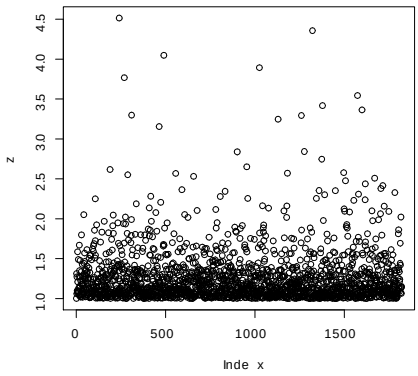


FIG. 3.1 – Echantillon conditionel Z pour ($\gamma_1 = 0.2$) et ($\gamma_2 = 0.8$)

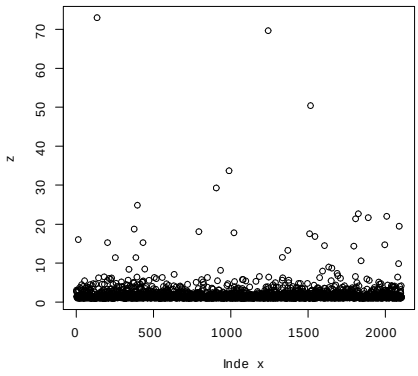


FIG. 3.2 – Échantillon conditionel Z pour ($\gamma_1 = 0.5$) et ($\gamma_2 = 0.8$)

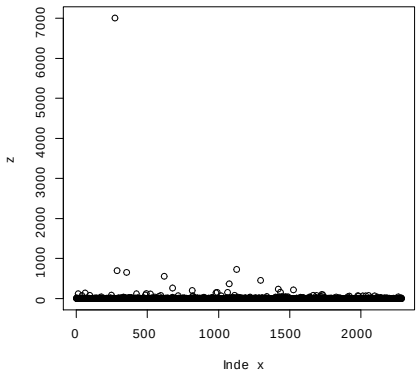


FIG. 3.3 – Échantillon conditionel Z pour ($\gamma_1 = 1$) et ($\gamma_2 = 0.8$)

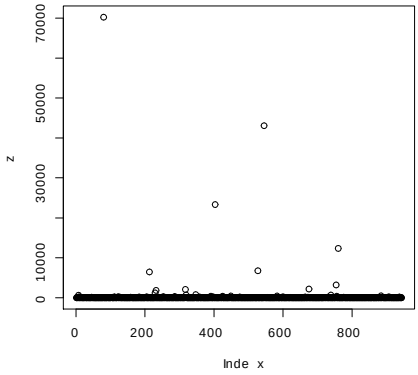


FIG. 3.4 – Échantillon conditionel Z pour ($\gamma_1 = 1.5$) et ($\gamma_2 = 0.8$)

3.3.3 Simulation de nouvel estimateure γ^h en fonction k

Algorithme2

1. on séléctione l'échantillon conditionnel z pour la méme méthode du l'algorithme1
2. on fixé $beta = 2$ et k varé de 1 á $m-1$, oú m la taill de l' échantillon Z ,et en simuler l'estimateur de queue $gama[k]$ oú

$$gama[k] = \frac{1}{beta-1}\left\{\frac{1}{k}\sum_{i=1}^{k}\left[\frac{Z_{m_{n,x}-k_{n,x},m_{n,x}}(x)}{Z_{m_{n,x}-i+1,m_{n,x}}(x)}\right]^{beta-1} - 1\right\}$$

3. Nous dessinons la courbe de gama en fonction "k"
4. Nous choisissons la valeur appropriée de "k" pour le coréspendant a la mieur valeur du gama que l'on note $kopt$

Programme sous R

```
#paramétra de simulation
n= ?
gamma=0.5
gamma2=0.8
#échentillon initial
u<-runif(n)
v<-runif(n)
y<-(1-u)^(-gamma1)
x<-(1-v)^(-gamma2)
y<-sort(y)
x1<-sample(x,1)
d<-numeric(n)
b<-numeric(n)
h<-(0.3)
for (i in 1 :n)
{
d[i]<-abs(x[i]-x1)
if (d[i]<h)
b[i]<-1 else b[i]<-0
}
Q<-(sum(b)/n)
m<-(n*Q)
z<-numeric(m)
for (i in 1 :n)
{
d[i]<-abs(x[i]-x1)
if (d[i]<h)
z[i]<-y[i] else z[i]<-0
}
z
z<-z[z!=0]
z
z<-sort(z)
```

```
z
#"graphe gama vs k"
beta=?
gama<-numeric(m-1)
for (k in 1 :m-1)
{
i<-seq(1,k)
gama[k]<-abs((1/(beta-1))*((1/k)*sum((z[m-k]/z[m-i+1])^(beta-1))-1))
}
k<-seq(1,m-1)
plot(k,gama[k],col="red",type="l",lwd="2",main="graphe de estimateure"
abline(h=gamma,col="blue",lwd="2")
# k optimal graphique
kopt<-k[which.min(abs(gamma-gama))]
kopt
gamaopt<-gama[kopt]
gamaopt
```

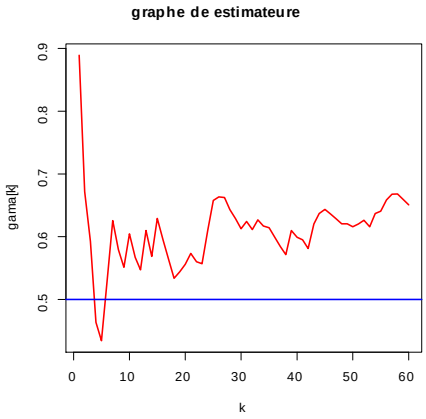


FIG. 3.5 – Comportement graphique de $\widehat{\gamma}$ vs k ,pour n=200 et beta =2

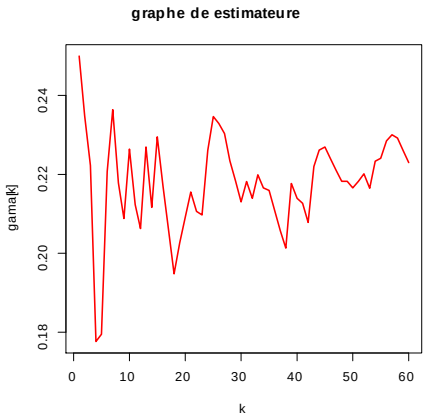


FIG. 3.6 – Comportement graphique de $\widehat{\gamma}$ vs k ,pour n=200 et beta =5

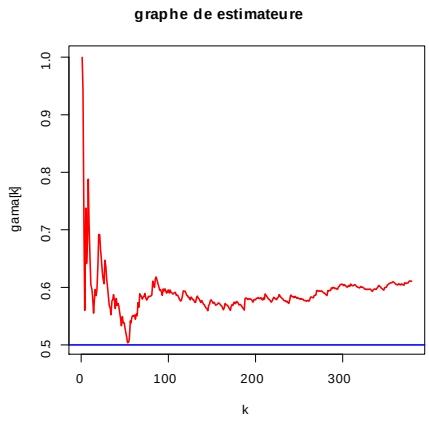


FIG. 3.7 – Comportement graphique de $\widehat{\gamma}$ vs k ,pour n=200 et beta 10

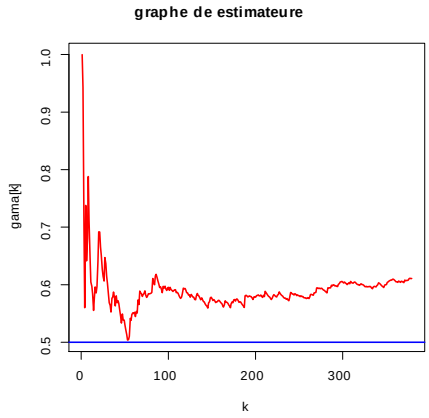


FIG. 3.8 – Comportement graphique de $\widehat{\gamma}$ vs k ,pour n=1000 et beta =2

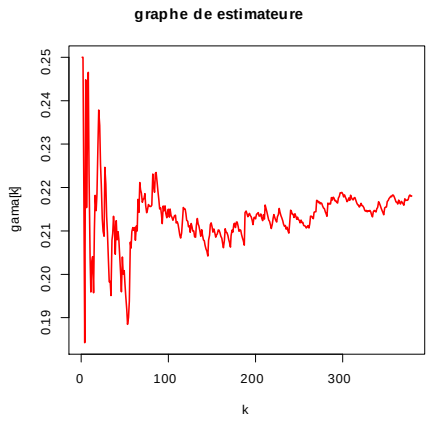


FIG. 3.9 – Comportement graphique de $\widehat{\gamma}$ vs k ,pour n=1000 et beta =5

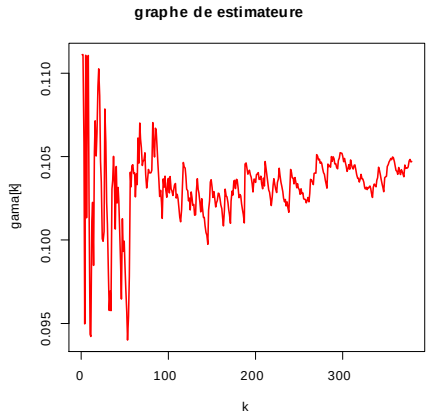


FIG. 3.10 – Comportement graphique de $\widehat{\gamma}$ vs k ,pour n=1000 et beta =10

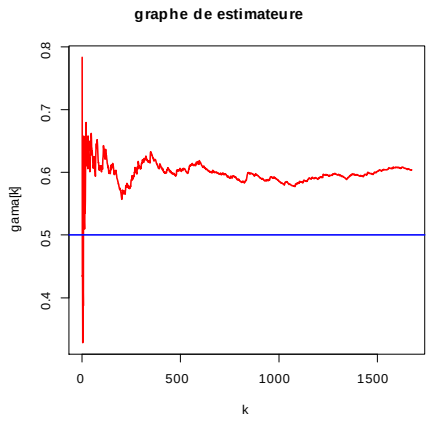


FIG. 3.11 – Comportement graphique de $\widehat{\gamma}$ vs k ,pour n=5000 et beta =2

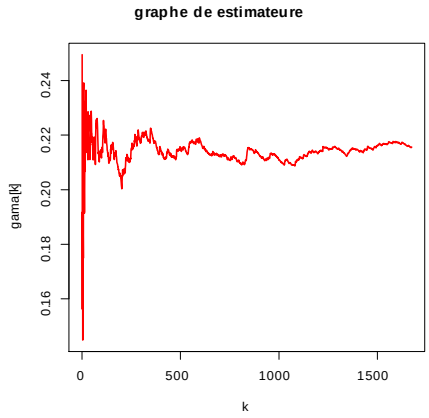


FIG. 3.12 – Comportement graphique de $\widehat{\gamma}$ vs k ,pour n=5000 et beta =5

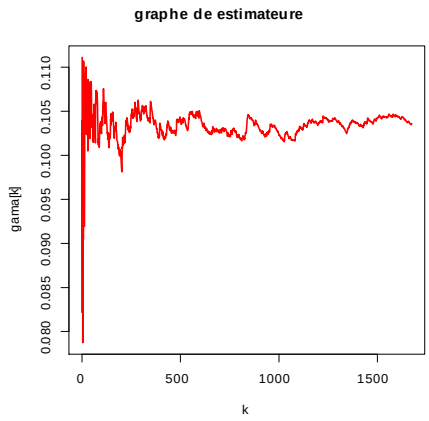


FIG. 3.13 – Comportement graphique de $\widehat{\gamma}$ vs k ,pour n=5000 et beta =10

3.3.4 En ploit de bootsrap pour de $\widehat{\gamma}$

Algorithme3

1. On séléctione l'échantillon conditionnel de la même méthode que l'lgorithme1
2. On choisit le valeure de $kopt$ de le même méthode que de l'lgorithme2
3. On simule l'estimateur de queue $gama1$ inétial pour $kopt$

$$gama1 = \frac{1}{beta-1}\left\{\frac{1}{kopt}\sum_{i=1}^{kopt}\left[\frac{Z_{m_{n,x}-k_{n,x},m_{n,x}}(x)}{Z_{m_{n,x}-i+1,m_{n,x}}(x)}\right]^{beta-1}-1\right\}$$

4. On crée B échantillons indépendants $Z^{*1}, Z^{*2}, ..., Z^{*B}$ où chaque échantillon Z^{*b} est obtenu en tirant n observations avec remise dans l'échantillon $Zboot = \{Zboot_1, ..., Zboo_n\}$
5. Pour chaque échantillon b tel que, $1 \leq b \leq B$, on calcule l'estimateur de queue :.$gmma[b]$,

$$gama[b] = \frac{1}{beta-1}\left\{\frac{1}{kopt}\sum_{i=1}^{kopt}\left[\frac{Zboot_{m_{n,x}-k_{n,x},m_{n,x}}(x)}{Zboot_{m_{n,x}-i+1,m_{n,x}}(x)}\right]^{beta-1}-1\right\}$$

   On obtient alors un échantillon de B valeurs de l'estimateur de queue $gama$
6. On estime alors la moiyenne $E(gama)$ et l'erreur standard $se_F(gama)$ et $QQnorme$ et du $biais$

Programme sous R

```
#paramétra de simulation
n=1000
gamma1=1.5
gamma2=0.6
beta=2
#échentillon initial
u<-runif(n)
v<-runif(n)
y<-(1-u)^(-gamma1)
x<-(1-v)^(-gamma2)
x1<-sample(x,1)
d<-numeric(n)
b<-numeric(n)
h<-(0.3)
for (i in 1 :n)
{
d[i]<-abs(x[i]-x1)
if (d[i]<h)
b[i]<-1 else b[i]<-0
}
Q<-(sum(b)/n)
m<-(n*Q)
z<-numeric(m)
for (i in 1 :n)
{
```

```
d[i]<-abs(x[i]-x1)
if (d[i]<h)
z[i]<-y[i] else z[i]<-0
}
z
z<-z[z!=0]
z
z<-(sort(z))
z
#choit kopt et gama inétial
gama<-numeric(m-1)
for (k in 1 :m-1)
{
i<-seq(1,k)
gama[k]<-abs((1/(beta-1))*((1/k)*sum((z[m-k]/z[m-i+1])^(beta-1))-1))
}
kopt<-which.min(abs(gamma-gama))
kopt
gama1<-gama[kopt]
#gama bootstrap
r<-2000
gama<-numeric(r)
for (b in 1 :r)
{
i<-seq(1,kopt)
zboot<-sample(z,length(z),TRUE)
zboot<-(sort(zboot))
gama[b]<-abs((1/(beta-1))*((1/kopt)*sum((zboot[m-kopt]/zboot[m-i+1])^(beta-1))-1))
}
#moiyenne gama
meangama<-mean(gama)
meangama
#graph de probabilié
qqnorm(gama)
#estimation bootstrap de l'ereure standar
sd<-sd(gama)
sd
#Estémation bootstrap de bias
bias<-mean(gama)-gama1
```

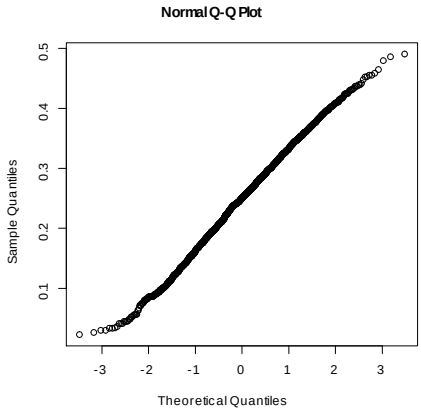


FIG. 3.14 – QQ-norm de la distribution limite bootstrap de 2000 répétition de Paréto ($\gamma_1 = 0.2), n = 1000$

	$R = 2000$	$R = 1000$	$R = 500$
k_{opt}	6	6	6
$\hat{\gamma}_1^{(h)}$	0.3248784	0.3248784	0.3248784
$\hat{\gamma}_{1_{boot}}^{(h)}$	0.1403471	0.2132133	0.1830524
$E(\hat{\gamma}^{(h)})$	0.2482064	0.246846	0.2481463
sd	0.8330169	0.08091615	0.08562753
$biais$	-0.07667207	-0.07803245	-0.07673214

TAB. 3.1 – Résultats de simulation pour n=1000

3.4 Simulation pour des données censureés

3.4.1 Échantillon initial et paramètres de simulation

La loi de simulation utilisée dans ce cas est une loi de Pareto de paramètre γ de fonction de répartition conditionel,

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

les covariables $(X_i)_{1\leq i\leq n} \backsim \text{Pareto}(\gamma_3)$ de taille $n = 1000$,à partir d'une variable w de $w([0,1])$, le modèle ajusté sera :

$$H^{-1}(w) = (1-w)^{-\gamma_3}$$

Nous avons généré un échantillon $(Z_{1_i})_{1\leq i\leq n} \backsim \text{Pareto}(\gamma_1)$ de taille $n = 1000$, à partir d'une variable u de $U([0,1])$, le modèle ajusté sera :

$$F^{-1}(u) = (1-u)^{-\gamma_1}$$

L'échantillon $(Z_{1_i})_{1\leq i\leq n}$est censuré par un deuxième échantillon $(Z_{2_i})_{1\leq i\leq n} \backsim \text{Pareto}(\gamma_2)$ à partir d'une variable v de $U([0,1])$:

$$G^{-1}(v) = (1-v)^{\gamma_2}$$

Les variables que nous observons sont d'une part les $Z_i \backsim \text{Pareto}(\gamma)$ définies par :

$$Y_i = Z_{1_i} \wedge Z_{2_i}$$

les indicateurs de censure sont,

$$\delta_i = \mathbb{I}_{\left\{Z_{1_i}\leq Z_{2_i}\right\}}$$

3.4.2 Seléctionement de l'échentillon condétionel censure

Algorithme4

1. Nous avons génééré deux échantillon $y1$ et $y2$ suivant la loit de préto de taille n=100 de paramétre $\gamma_1 = 1.5$ (resp $\gamma_2 = 0.6$),á partir d'une variable u de U ([0 ; 1]) resp(v de U ([0 ; 1])) , et une covariable x suivant la loi de préto de taill n=100 de parametr $\gamma_3 = 0.8$),á partir d'une variabler w de w ([0 ; 1] , le mod ele ajusté sera :

$$F^{-1}(u) = 1 - u^{-1/\gamma_1}$$

$$F^{-1}(v) = 1 - v^{-1/\gamma_2}$$

$$F^{-1}(w) = 1 - w^{-1/\gamma_3}$$

1. on séléctione deux échentillons condetionels z1 et z2 de même covariable que x, pour cela on étuliser la meme méthode de l' algoritme1

3 On séléctione l' échentillon initial oú $Y_i = Z_{1_i} \wedge Z_{2_i}$

4 On défini δ oú $\delta_i = \mathbb{I}_{\left\{Z_{1_i}\leq Z_{2_i}\right\}}$

5 On ordonnee les données pour obtenir l'échentillon censurés $Y = \{(Y_1,\delta_1),...,(Y_n,\delta_n)\}$,

Programme sous R

```
#paramétre de simulation

n=1000
gamma1=1.5
gamma2=0.6
gamma3=0.8
u<-runif(n)
v<-runif(n)
w<-runif(n)
y1<-(1-u)^(-gamma1)
x<-(1-v)^(-gamma2)
y2<-(1-w)^(-gamma2)
#séléctioner l' échantion condetionel z
x1<-sample(x,1)
d<-numeric(n)
b<-numeric(n)
h<-(0.3)
for (i in 1 :n)
{
d[i]<-abs(x[i]-x1)
if (d[i]<h)
b[i]<-1 else b[i]<-0
}
Q<-(sum(b)/n)
m<-(n*Q)
#selectioner l'échentillon z1
z1<-numeric(m)
for (i in 1 :n)
{
d[i]<-abs(x[i]-x1)
if (d[i]<h)
z1[i]<-y1[i] else z1[i]<-0
}
z1
z1<-z1[z1 !=0]
z1
seléctoner l'échentillon z2
z2<-numeric(m)
for (i in 1 :n)
{
d[i]<-abs(x[i]-x1)
if (d[i]<h)
z2[i]<-y2[i] else z2[i]<-0
}
z2
z2<-z2[z2 !=0]
z2
#séléctioner l'échentillon initial
delta<-numeric(m)
```

```
y<-pmin(z1,z2)
for (i in 1 :m)
{
if (z1[i]<=z2[i])
delta[i]<-1 else delta[i]<-0
}
delta
r<-rank(y)
#ordonner des donner
y<-sort(y)
r1<-sort(r)
L<-numeric(length(y))
B<-numeric(length(y))
Delta<-numeric(length(y))
for (i in 1 :length(y))
{
L[i]<-(delta[i]+length(y))*r[i]
}
L
A<-sort(L)
A
for (i in 1 :length(y))
{
B[i]<-A[i]/r1[i]
}
B
Delta<-B-length(y)
Delta
plot(y
```

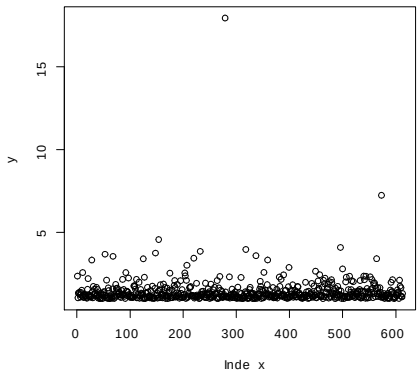


FIG. 3.15 – Échantillon conditionel Y, ($\gamma_1 = 0.35$) et ($\gamma_2 = 2.5$) (10%censurées)

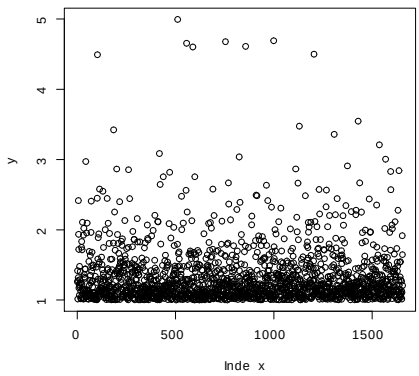


FIG. 3.16 – Échantillon conditionel Y, ($\gamma_1 = 0.35$) et ($\gamma_2 = 1$) (25%censurées)

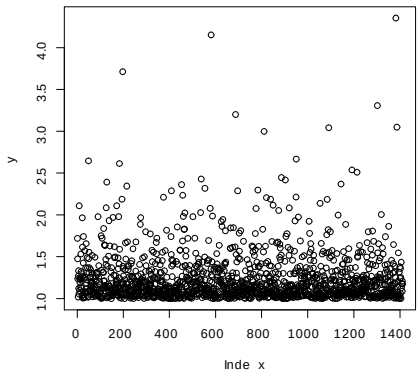


FIG. 3.17 – Échantillon conditionel Y, ($\gamma_1 = 0.35$) et ($\gamma_2 = 0.5$) (40%censurées)

3.4.3 Simulation de nouvel estimateur de l'indice en fonction de k

Algorithme5

1. On séléctione l'échantillon conditionnel censuré Y de la même méthode que l'lgorithme4
2. On fixe $beta = 2$ et k vari de 1 á $m-1$, où m la taill de léchantillon Y ,et on simule l'estimateur de queue $gama[k]$ où

$$gama[k] = \left(\frac{1}{beta-1} \left\{ \frac{1}{k} \sum_{i=1}^{k} \left[\frac{Y_{m_{n,x}-k_{n,x},m_{n,x}}(x)}{Y_{m_{n,x}-i+1,m_{n,x}}(x)} \right]^{beta-1} - 1 \right\} \right) \diagup P(x)$$

   oú

$$P(x) = \frac{1}{kopt} \sum_{i=1}^{kopt} \delta^{x}_{[m_{n,x}-i+1]}$$

3. Nous dessinons la courbe de gama en fonction "k"
4. Nous choisissons la valeur appropriée de "k" pour la meilleure valeure de gama corespondante gama et on noté $kopt$

Programme sous R

```
#paramétre de simulation
n=1000
gamma1=0.35
gamma2=
gamma3=0.8
u<-runif(n)
v<-runif(n)
w<-runif(n)
y1<-(1-u)^(-gamma1)
x<-(1-v)^(-gamma2)
y2<-(1-w)^(-gamma2)
#seléctioner l'écheantion condetionel z
x1<-sample(x,1)
d<-numeric(n)
b<-numeric(n)
h<-(0.3)
for (i in 1 :n)
{
d[i]<-abs(x[i]-x1)
if (d[i]<h)
b[i]<-1 else b[i]<-0
}
Q<-(sum(b)/n)
m<-(n*Q)
#sélectioner l'échentillon z1
z1<-numeric(m)
for (i in 1 :n)
{
```

```
d[i]<-abs(x[i]-x1)
if (d[i]<h)
z1[i]<-y1[i] else z1[i]<-0
}
z1
z1<-z1[z1 !=0]
z1
#seléctioner l'échantillon z2
z2<-numeric(m)
for (i in 1 :n)
{
d[i]<-abs(x[i]-x1)
if (d[i]<h)
z2[i]<-y2[i] else z2[i]<-0
}
z2
z2<-z2[z2 !=0]
z2
#séléctioner l'échantillon initial
delta<-numeric(m)
y<-pmin(z1,z2)
plot(y)
for (i in 1 :m)
{
if (z1[i]<=z2[i])
delta[i]<-1 else delta[i]<-0
}
delta
r<-rank(y)
#ordonner des donner
y<-sort(y)
r1<-sort(r)
L<-numeric(length(y))
B<-numeric(length(y))
Delta<-numeric(length(y))
for (i in 1 :length(y))
{
L[i]<-(delta[i]+length(y))*r[i]
}
L
A<-sort(L)
A
for (i in 1 :length(y))
{
B[i]<-A[i]/r1[i]
}
B
Delta<-B-length(y)
Delta
#"graphe gama vs k"
```

```
beta<-2
gama<-numeric(m-1)
for (k in 1 :m-1)
{
i<-seq(1,k)
gama[k]<-abs((1/(beta-1))*((1/k)*sum((y[m-k]/y[m-i+1])^(beta-1))-1))/((1/k)*sum(Delta[m-
i+1]))
}
k<-seq(1,m-1)
plot(k,gama[k],col="red",type="l",lwd="2",main="graphe de estimateure" )
abline(h=gamma1,col="blue",lwd="2")
# k optimal graphique
kopt<-k[which.min(abs(gamma-gama))]
kopt
gamaopt<-gama[kopt]
gamaopt
```

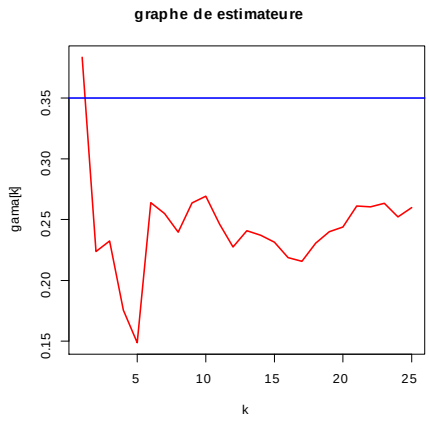


FIG. 3.18 – Comportement graphie de $\widehat{\gamma}$ vs k issu de la distribution de pareto ($\gamma_1 = 0.35$) censurée par pareto ($\gamma_2 = 2.5$), (10% de censure), pour n=200

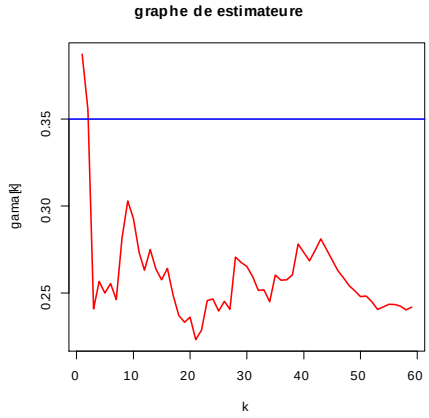


FIG. 3.19 – Comportement graphic de $\widehat{\gamma}$ vs k issu de la distribution de pareto ($\gamma_1 = 0.35$) censurée par pareto ($\gamma_2 = 2.5$), (10% de censure), pour n=1000

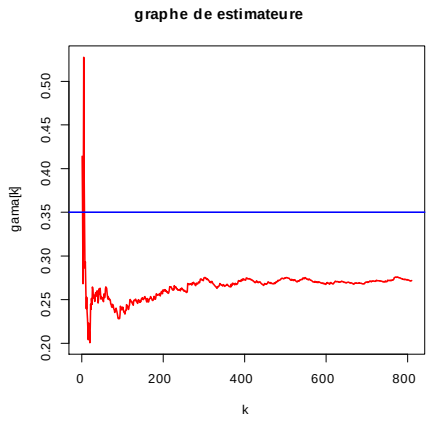


FIG. 3.20 – Comportement graphic de $\widehat{\gamma}$ vs k issu de la distribution de pareto ($\gamma_1 = 0.35$) censurée par pareto ($\gamma_2 = 2.5$), (10% de censure), pour n=5000

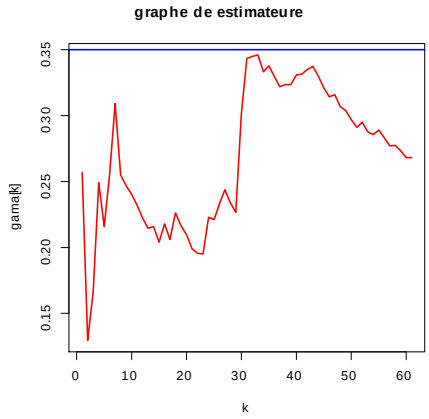


FIG. 3.21 – Comportement graphic de $\widehat{\gamma}$ vs k issu de la distribution de pareto ($\gamma_1 = 0.35$) censurée par pareto ($\gamma_2 = 1$), (25% de censure), pour n=200

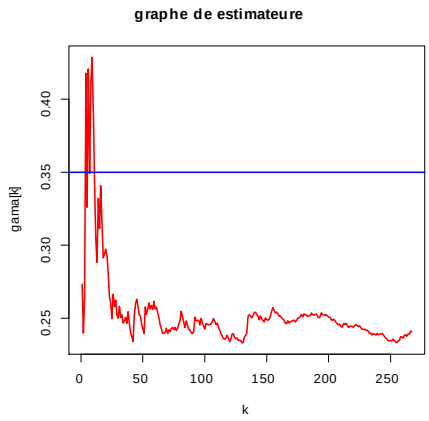


FIG. 3.22 – Comportement graphic de $\widehat{\gamma}$ vs k issu de la distribution de pareto ($\gamma_1 = 0.35$) censurée par pareto ($\gamma_2 = 1$), (25% de censure), pour n=1000

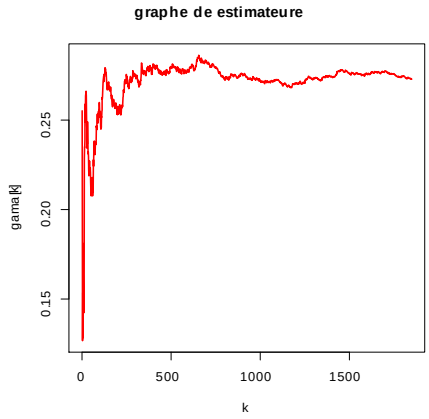


FIG. 3.23 – Comportement graphic de $\widehat{\gamma}$ vs k issu de la distribution de pareto ($\gamma_1 = 0.35$) censurée par pareto ($\gamma_2 = 1$), (25% de censure), pour n=5000

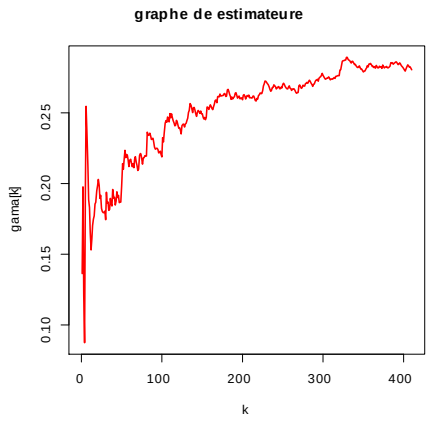


FIG. 3.24 – Comportement graphic de $\widehat{\gamma}$ vs k issu de la distribution de pareto ($\gamma_1 = 0.35$) censurée par pareto ($\gamma_2 = 0.5$), (40% de censure), pour n=200

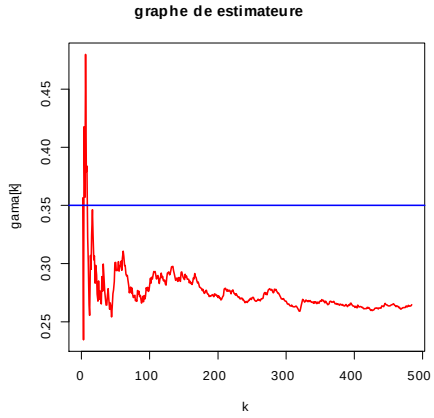


FIG. 3.25 – Comportement graphic de $\widehat{\gamma}$ vs k issu de la distribution de pareto ($\gamma_1 = 0.35$) censurée par pareto ($\gamma_2 = 0.5$), (40% de censure), pour n=1000

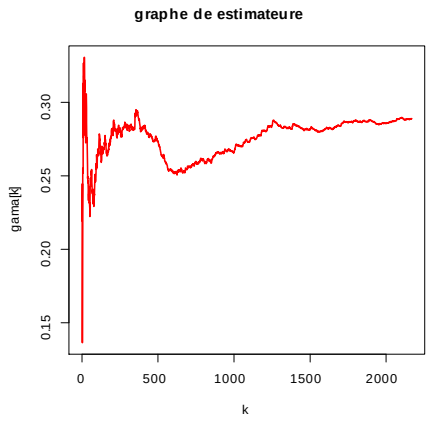


FIG. 3.26 – Comportement graphic de $\widehat{\gamma}$ vs k issu de la distribution de pareto ($\gamma_1 = 0.35$) censurée par pareto ($\gamma_2 = 0.5$), (40% de censure), pour n=5000

Comontaire

3.4.4 emploit du bootsrap

Algorithme

1. On séléctione l'échantillon conditionnel Y de la même méthode que l'lgorithme 4
2. On choit valeure de $kopt$ et $gama1$(estimateur inétiale) pour la méme méthode de l'lgorithme 5
3. On crée B échantillons indépendants $Z^{*1}, Z^{*2}, ..., Z^{*B}$ où chaque échantillon Z^{*b} est obtenu en tirant n observations avec remise dans l'échantillon $Yboot = \{(Yboot_1, \delta_1)..., (Yboo_n, \delta_n)\}$
4. Pour chaque échantillon b tel que, $1 \leq b \leq B$, on calcule l'estimateur de queue :.$gmma[b]$,

$$gama[b] = \frac{1}{beta-1}\left\{\frac{1}{kopt}\sum_{i=1}^{kopt}\left[\frac{Yboot_{m_{n,x}-k_{n,x},m_{n,x}}(x)}{Yboot_{m_{n,x}-i+1,m_{n,x}}(x)}\right]^{beta-1} - 1\right\} \diagup P(x)$$

$$P(x) = \frac{1}{kopt}\sum_{i=1}^{kopt}\delta^{x}_{[m_{n,x}-i+1]}$$

   On obtient alors un échantillon de B valeurs de l'estimateur de queue $gama$
5. On estime alors la moiyenne $E(gama)$ et l'erreur standard $se_F\,(gama)$ et $QQnorme$ et du $biais$

Programme sous R

```
 #paramétre de simulation
#paramétre de simulation
n=100
gamma1=1.5
gamma2=0.6
gamma3=0.8
beta=2
#séléctioner léchention condetionel z
u<-runif(n)
```

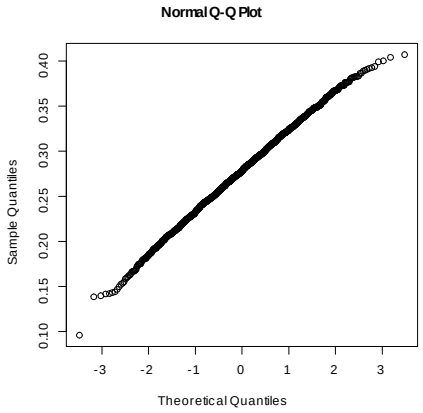
```
v<-runif(n)
w<-runif(n)
y1<-(1-u)^(-gamma1)
x<-(1-v)^(-gamma2)
y2<-(1-w)^(-gamma2)
x1<-sample(x,1)
d<-numeric(n)
b<-numeric(n)
h<-(0.3)
for (i in 1 :n)
{
d[i]<-abs(x[i]-x1)
if (d[i]<h)
b[i]<-1 else b[i]<-0
}
Q<-(sum(b)/n)
m<-(n*Q)
#selectioner l'échentillon z1
z1<-numeric(m)
for (i in 1 :n)
{
d[i]<-abs(x[i]-x1)
if (d[i]<h)
z1[i]<-y1[i] else z1[i]<-0
}
z1
z1<-z1[z1 !=0]
z1
séléctioner l'échentillon z2
z2<-numeric(m)
for (i in 1 :n)
{
d[i]<-abs(x[i]-x1)
if (d[i]<h)
z2[i]<-y2[i] else z2[i]<-0
}
z2
z2<-z2[z2 !=0]
z2
#séléctioner l'échentillon initial
delta<-numeric(m)
y<-pmin(z1,z2)
for (i in 1 :m)
{
if (z1[i]<=z2[i])
delta[i]<-1 else delta[i]<-0
}
delta
r<-rank(y)
#ordonner des données
```

```
y<-(sort(y))
r1<-(sort(r))
L<-numeric(length(y))
B<-numeric(length(y))
Delta<-numeric(length(y))
for (i in 1 :length(y))
{
L[i]<-(delta[i]+length(y))*r[i]
}
L
A<-(sort(L))
A
for (i in 1 :length(y))
{
B[i]<-A[i]/r1[i]
}
B
Delta<-B-length(y)
Delta
#chiosir de kopt et gama inétial
beta<-2
for (k in 1 :m-1)
{
i<-seq(1,k)
gama[k]<-abs((1/(beta-1))*((1/k)*sum((y[m-k]/y[m-i+1])^(beta-1))-1))/((1/k)*sum(Delta[m-
i+1]))
}
kopt<-which.min(abs(gamma1-gama))
kopt
gama1<-gama[kopt]
gama1
#gama bootstrap
R<-2000
gama<-numeric(R)
for (j in 1 :R)
{
indice <-sample(1 :length(y),length(y),replace=TRUE)
yboot<-y[indice]
deltaboot<-delta[indice]
yboot
deltaboot
r<-rank(y)
#ordonner des donner
yboot<-sort(yboot)
r1<-(sort(r))
L<-numeric(length(yboot))
B<-numeric(length(yboot))
Delta<-numeric(length(y))
for (i in 1 :length(yboot))
{
```

```
L[i]<-(deltaboot[i]+length(yboot))*r[i]
}
L
A<-(sort(L))
A
for (i in 1 :length(yboot))
{
B[i]<-A[i]/r1[i]
}
B
Deltaboot<-B-length(y)
Deltaboot
s<-seq(1,kopt)
gama[j]<-abs((1/(beta-1))*((1/kopt)*sum((yboot[m-kopt]/yboot[m-s+1])^(beta-1))-1))/((1/kopt)*su
s+1]))
}
#moiyenne gama
meangama<-mean(gama)
meangama
#graph de probabilié
qqnorm(gama)
#estimation bootstrap de l'ereure standar
sd<-sd(gama)
sd
#Estémation bootstrap de bias
bias<-mean(gama)-gama1
bias
```

	$c = 10\%$	$c = 25\%$	$c = 40\%$
k_{opt}	20	47	41
$\hat{\gamma}_1^{(h)}$	0.3035514	0.3376852	0.3495238
$\hat{\gamma}_{1_{boot}}^{(h)}$	0.3073708	0.3397037	0.3374236
$E(\hat{\gamma}^{(h)})$	0.2779916	0.3258021	0.3414811
sd	0.04537116	0.03366893	0.04272734
$biais$	-0.02555979	-0.01188314	-0.008042705

TAB. 3.2 – Résultats de simulation pour n=1000



QQ-norm de la distribution limite bootstrap de 2000 répétition de Paréto ($\gamma_1 = 0.35$),censurée par($\gamma_2 = 2.5$), (10% de censure),n=1000

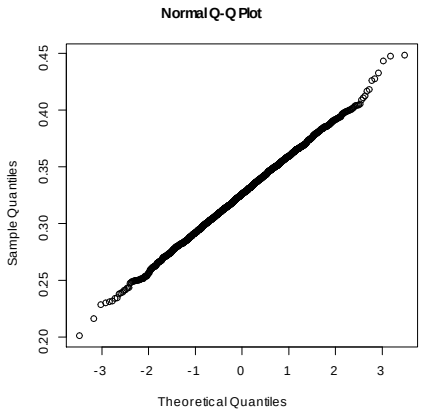


FIG. 3.27 – QQ-norm de la distribution limite bootstrap de 2000 répétition de Paréto ($\gamma_1 = 0.35$),censurée par($\gamma_2 = 1$), (25% de censure),n=1000

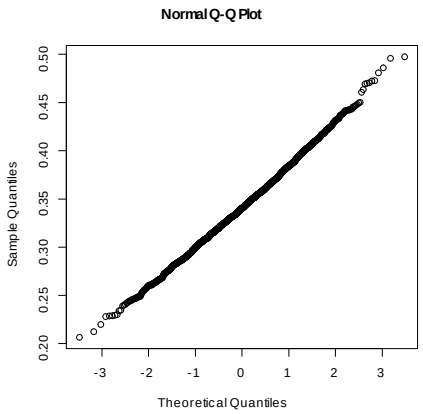


FIG. 3.28 – QQ-norm de la distribution limite bootstrap de 2000 répétition de Paréto ($\gamma_1 = 0.35$),censurée par($\gamma_2 = 0.5$), (40% de censure),n=1000

Conclusion

Nous avons calculé dans notre travail un nouvel estimateur de l'indice des valeurs extrêmes, en adaptant un estimateur harmonique au cas des données conditionnel, par la méthode du fenêtre mobile, pour une distribution à queue lourde dans le cas du domaine d'attraction de Fréchet.

Notre nouvel estimateur présente une forme bien explicite, q on a simulé en fixant certain paramètres initiaux.

Notre estimateur à été présenté en deux versions, dans le cas des données complètes ainsi que des données censurées

Nous avons consacré deux objectifs dans nos études de simulation dans la première partie en étudiant l'estimateur en fonction du nombre de valeurs extrêmes, ensuite en variant le paramètre de réglage pour avoir une idée sur une meilleur convergence empirique dans un premier temps, ainsi pour maitriser l'influence du censure selon l'échantillon étudié.

Dans la deuxième partie nous avons considéré le bootsrap pour étudier la normalité asymptotique empirique. Nous avons constaté que notre estimateur présente bel et bien une convergence en loi disant empiriquement, ce qui reste comme perspectifs à démontrer théoriquement ainsi que la fixation du nombre des valeurs extrêmes correspondant en minimisant l'erreur quadratique moyenne théorique correspondante.

Nous envigeons à présenter une application réelles pour des données de crédit majores des banques oû les echanciers du remboursements des benifisseurs presentent de la censure

Bibliographie

[1] Alexandre Lekina. Estimation non-paramétrique des quantiles extrêmes conditionnels. Mathématiques [math]. Université Joseph-Fourier - Grenoble I ; Université de Grenoble, 2010. Français. tel-00529476v2ffAlvarado, E.,Sandberg, D. etPickford, S. (1998). Modeling large forest firesas extreme events.Northwest Science, 72 :66-75.

[2] Alvarado, E.,Sandberg, D. etPickford, S. (1998). Modeling large forest firesas extreme events.Northwest Science, 72 :66-75

[3] Balkema, A., de Haan, L. Residual life at great age. Ann. Probab., 2 :792-804,1974. (Cité en pages xxi, 121, 124, 128 et 135.)

[4] Beirlant, J, Goegebeur, Y, Segers, J, and Teugels, J. (2006). Statistics of Extr ?mes :Theory and Applications. John Wiley.

[5] Beirlant, J., and Guillou, A. (2001). Pareto index estimation under moderate right censoring. Scand. Actuar. J., 111 125

[6] Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J. Statistics of extremes, theoryand applications. Wiley, New York, 2004. (Cité en pages xix, 22, 81 et 82.)-

[7] Beirlant, J., Guillou, A. , Dierckx, G., Fils-Villetard, A. . Estimation of the extreme value index and extreme quantiles under random censoring. SpringerScience+Business Media, LLC 2007, 10 :151-174, 2007. (Cité en pages xx, 73,74, 82, 86, 114 et 123.)-

[8] Beirlant, J., Vjnckier, P., Teugels, J.L. Tail index estimation, pareto quantile plots, and regression diagnostics. J. Amer Statist. Assoc., 91 :1659-1667, 1996.(Cité en pages xix, 14, 22, 40 et 123.)-

[9] Beirlant, J.,Goegebeur, Y.,Segers, J.,Teugels, J., deWaal, D. etFerro,C. (2004b).Statistics of Extremes : Theory and Applications. John Wiley Sons.

[10] Bingham, N., Goldie, C. et Teugels, J. (1987). Regular Variation. CambridgeUniversity Press. 22, 23, 24, 68, 111

[11] Bouleau, N. (1991). Splendeurs et misères des lois de valeurs extêmes.RevueRisques, 4 :85-92.

[12] Brahimi, B., Meraghni, D., Necir, A. On the asymptotic normality of hill's estimator of the tail index under random censoring. Preprint : arXiv-1302.1666,2013. (Cité en pages 44, 82 et 107.)

[13] Censure,th ?se de doctorat,universit ? Gaston berger de Saint-Louis.

[14] Ceresetti, D.,Ursu, E.,Carreau, J.,Anquetin, S.,Creutin, J.,Gardes, L.,Girard, S. etMolinié, G. (2012). Evaluation of classical spatial-analysis schemes ofextreme rainfall.Natural Hazards and Earth System Sciences, 12 :3229-3240.

[15] Coles, S. etTawn, J. (1996). A bayesian analysis of extreme rainfall data.AppliedStatistics, pages 463-478.

[16] Daouia, A., Gardes, L., Girard, S., Lekina, A. Kernel estimators of extreme level curves. Test, 20 :311-333, 2011. (Cité en pages xix, 82, 85, 88, 89 et 108.)Daouia, A.,Gardes,

L.,Girard, S. etLekina, A. (2011). Kernel estimators ofextreme level curves.Test, 20(2) :311-333.

[17] Daouia, A.,Gardes, L.,Girard, S. etLekina, A. (2011). Kernel estimators ofextreme level curves.Test, 20(2) :311-333

[18] Davison, A.C., Smith, R.L. (1990). Models for exeedanes over high thresholds,Journal of the Royal Statistial Soiety, series B , 52 , 393442.

[19] Dekkers, A.L.M., Einmahl, J.H.J., de Haan, L. A moment estimator for the index of an extreme value index. Annals of statistics, 17 :1833-1855, 1989. (Cité en pages xix, 13, 14, 18, 40, 45, 82 et 123.)

[20] Delafosse, E., Guillou, A. Almost sure convergence of a tail index estimatorestimator in the presence of censoring. comptes rendus mathématiques. Académie des Sciences. Paris, 335 :375-380, 2002. (Cité en page 82.)

[21] Einmahl, J. etMagnus, J. (2008). Records in athletics through extreme-valuetheory.Journal of the American Statistical Association, 103(484) :1382-1391.

[22] Einmahl, J. etSander, S. (2011). Ultimate 100-m world records through extreme-value theory.Statistica Neerlandica, 65(1) :32-42

[23] Einmahl, J. H. J., Fils-Villetard, A., Guillou, A. Statistics of extremes under Random Censoring. Bernoulli, 14(1) ; 207227.

[24] Einmahl, J. H., Fils-Villetard, A., and Guillou, A. (2008). Statistics of Extremes UnderRandom Censoring. Bernoulli, 14(1) ; 207227. et Gestion des Risques. Evenements Rares pour la Gestion des Risques, DESS 203 de l'Université

[25] Ferrez, J., Davison, A.C., Rebetez, M. Extreme temperature analysis under forest cover compared to an open field. Agricultural and Forest Meteorology,151 :992-1001, 2011. (Cité en page 81.)

[26] Gardes, L. etGirard, S. (2010). Conditional extremes from heavy-tailed distri-butions : an application to the estimation of extreme rainfall return levels.Extremes,13(2) :177-204.

[27] Gardes, L., Girard, S. Conditional extremes from heavy-tailed distributions : an application to the estimation of extreme rainfall return levels. Extremes, 13 :177-204, 2010. (Cité en page 81.)

[28] Gardes, L., Girard, S. Functional kernel estimators of large conditional quan-tiles. Electronic Journal of Statistics, 6 :1715-1744, 2012. (Cité en page 81.)

[29] Gardes, L., Girard, S., Lekina, A. Functional nonparametric estimation of conditional extreme quantiles. Journal of Multivariate Analysis, 101 :419-433, 2010.(Cité en pages xix, 23, 81 et 134.)

[30] Gardes, L., Stupfler, G. Estimation of the conditional tail index using a smoothedlocal hill estimator. Extremes, 17 :45-75, 2014. (Cité en page 82.)

[31] Goegebeur, Y., Guillou, A., Osmann, M. A local moment type estimator for the extreme value index in regression with random covariates. The Canadian Journal of Statistics, 42 :487-507, 2014a. (Cité en page 82.)

[32] Goegebeur, Y., Guillou, A., Schorgen, A. Nonparametric regression estimation of conditional tails : the random covariate case. Statistics, 48 :732-755, 2014b. (Cité en pages 82, 84, 85, 88, 91, 111, 117 et 134.)

[33] Gomes, M.I., Neves, M.M. Estimation of the extreme value index for randomly censored data. Biometrical Letters, 48 :1-22, 2011. (Cité en page 82.)

[34] Gomes, M.I., Oliveira, O. Censoring estimators of a positive tail index. Statistics. Statistics Probability Letter, 65 :147-159, 2003. (Cité en page 82.)

[35] Gumbel, E. (1954).Statistical theory of extreme values and some practical ap-plications : a series of lectures. Numéro Applied Mathematics Series, 33. NationalBureau of Standards, Washington.

[36] Gumbel, E. (1958).Statistics of extremes. Columbia University Press, Columbia.-109-

[37] Hall, P., Tajvidi, N. (2000). Nonparametric analysis of temporal trend when tting parametric models to extreme-value data, Statistial Siene , 15 , 153.167.

[38] Hill, B. A simple general approach to inference about the tail of a distribution. The Annals of Statistics, 3 :1163-1174, 1975. (Cité en pages xix, 13, 39 et 123.)

[39] Hill, B. M.(1975). A simple general approach to inference about the tail of a distribution.Ann. Statist., 3(5) ; 11631174 in the presence of censoring. comptes rendus mathématiques. Académie des Sciences. Paris, 335 :375-380, 2002. (Cité en page 82.)

[40] Ndao, P., Diop, A., Dupuy, J.-F. Nonparametric estimation of the conditional tail index and extreme quantiles under random censoring. Computational Statistics Data, 79 :63-79, 2014. (Cité en page 82.)

[41] NERC(1975). Flood studies report. Rapport technique, London, Natural Envi-ronement Research Council. Paris IX Dauphine Marchés Financiers, Marchés des Mati ?res Premi ?res

[42] Pisarenko, V.F., Sornette, D. Characterization of the frequency of extreme ear-thquake events by the generalized pareto distribution. Pure and Applied Geo-physics, 160 :2343-2364, 2003. (Cité en page 81.)

[43] Reiss, R.D, and Thomas, M.(2007). Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields.Birkh

[44] Reiss, R.-D. etThomas, M. (2001).Statistical analysis of extreme values : withapplications to insurance, finance, hydrology and other fields. Birkh ?user Verlag.

[45] Roncalli, T., 2002, Th ?orie des Valeurs Extr ?mes ou Mod ?lisation desdes Ev ?nements Rares pour la Gestion des Risques, DESS 203 de l ?Universit ? Paris IX Dauphine March ?s Financiers, March ?s des Mati ?res Premi ?res et Gestion des Risques

[46] Rootzén, H. etTajvidi, N. (2001). Can losses caused by wind storms be predictedfrom meteorological observations Scandinavian Actuarial Journal, 2001(2) :162-175.

[47] Sandra Plancade. Estimation de la fonction de répartition conditionnelle à partir de données censurées par intervalle, cas 1, par sélection de modèles. 42èmes Journées de Statistique, 2010, Marseille, France. pp.USB-key. inria-00494833ff.

[48] Stupfler, G. A moment estimator for the conditional extreme-value index. Elec-tronic Journal of Statistics, 7 :2298-2343, 2013. (Cité en pages 76, 81 et 129.)

[49] Smith, R. L. (1989). Extreme value analysis of environmental time series : an appliation to trend detetion in ground-level ozone (with disussion),Statistial Siene , 4 , 367393.-17-Stupfler, G. A moment estimator for the conditional extreme-value index. Elec-tronic Journal of Statistics, 7 :2298-2343, 2013. (Cité en pages 76, 81 et 129.)

[50] Toulemonde,G.,2008, Estimation et tests en th ?orie des valeurs extr ?mes,th ?se de doctorat de l'université Paris VI

[51] Worms, J., Worms, R. New estimators of the extreme value index under random right censoring, for heavy-tailed distributions. Extremes, 17 :337-358, 2014.(Cité en page 82.)

[52] Worms, J., Worms, R. New estimators of the extreme value index under random right censoring, for heavy-tailed distributions. Extremes, 17 :337-358, 2014.(Cité en page 82.)

Résuumé

En 2013 Bearan. J et all on présenté un neveux estimateur robuste des moment harmonique de l'indice des valeur extrême à variation régulière avec un paramètre de réglage notamment sous étude asymptique. Dans notre travail, nous avons adapté ce dernier dans la cas conditionnelle lorsque l'échantillon Y est lie an covariable X, notamment nous avons calculés notr neveux estimateur dans le cas des données complètes et censures. Nous avons présent quelque simulations simple ainsi bootstrap pour l'emploi du de sous compune éventuelle analyse ortement asymptotique

Mot clés : la théorie des valeurs extrêmes, les donnes censures, variation régulier , indice de valeurs extrêmes, variable
conditionnelle , indice de valeurs extrêmes, fenêtre mobile, bootstrap

Abstract

In 2013 Bearan. J et al. presented a new robust estimator of the harmonic moments of the index of the extreme values with regular variation with a tuning parameter, in particular its asymptotic study was established. In our work, we have adapted the latter in the conditional case when the sample Y is linked to the covariate X. Thus we have calculated our new estimator in the case of complete and censored data. We have presented some simple simulations as well as the use of the bootstrap for a possible asymptotic analysis

Keywords : extreme value theory, censored data, regular variation, extreme value index, conditional variable, extreme value index, moving window, bootstra

ملخص

في عام 2013 بيران. ج. وآخرون. قدمو مقدرًا جديدًا للعزوم الهندسية لمؤشر القيم القصوى المتغيرة بانتظام ذات وسيط تعديل ، ولا سيما تقديم دراسة التقارب اللانهائي المرافق. في عملنا ، قمنا بتكييف هذا الأخير في الحالة الشرطية عندما تكون العينة Y مرتبطة بالمتغير المشترك X. وهكذا قمنا بحساب مقدرنا الجديد في حالة البيانات الكاملة والخاضعة للرقابة. لقد قدمنا ببعض عمليات المحاكاة البسيطة بالإضافة إلى استخدام تقنية السحب لتحليل تقارب المقدر اللانهائي.

الكلمات المفتاحية: نظرية القيمة القصوى ، البيانات الخاضعة للرقابة ، التباين المنتظم ، مؤشر القيمة القصوى ، المتغير الشرطي ، مؤشر القيمة القصوى ، النافذة المتحركة ، السحب