PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
Ministry of Higher Education and Scientific Research
UNIVERSITY SAAD DAHLAB - BLIDA 1
Faculty of Sciences
Department of Mathematics

A Thesis Presented for the Degree of

DOCTOR in Mathematics - Statistics

Modeling of Conditional Extreme Values Distributions under Right Censored Data with Application

By
GUERMAH Toufik

Examination Committee Members :

Mohamed HACHAMA	Chairman	Prof.	Univ. Blida 1
Abdelaziz RASSOUL	Advisor	Prof.	ENSH, Blida
Hamid OULD ROUIS	Co-Advisor	Prof.	Univ. Blida 1
Diffalah LAISSAOUI	Examinator	MCA	Univ. Médéa
Halim ZAGHDOUDI	Examinator	Prof.	Univ. Annaba
Redouane BOUDJEMAA	Examinator	MCA	Univ. Blida 1

Blida, November 14, 2021

Dedication

If this thesis is of value, there are many who have shared it with me.

I am unable to thank you, mother, appropriately unless God helps me and grants me the ability to repay you,

My father, who spared no effort in my education and who has long awaited this research, may God forgive him and have mercy on him,

My wife and my four sons who have been patient with me throughout the years of research,

My brothers and sisters who cared for me and gave me material and moral support in this research,

All relatives and friends,

To all those that are living among them, I dedicate this work.

Acknowledgements

In the Name of Allah, the Most Gracious, the Most Merciful.
All praise is due to Allah, Lord of the universe and the Master of the Day of Judgment, his Blessings and peace be upon the Final Messenger, the Mercy to mankind, the Prophet Muhammad, as well as upon his family, his companions and those all follow the true path of the righteous predecessors until the Last Day of Judgment.

Narrated AbuHurayrah: The Prophet (b.p.u.h.) said: He who does not thank Allah does not thank people.
First of all, I would like to express my heartfelt thanks to my generous supervisor Pr. Abdelaziz Rassoul, without a compliment, who accepted me when they rejected me, opened the door for me to search after it was almost closed, encouraged me when I was cowardly, directed me when I went astray, and countered my wrongdoing with his kindness and my mistake with his pardon. I thank him for his great eagerness to complete this work. I ask God to thank him for his efforts.
All thanks and gratitude to my esteemed teacher Pr Hamid Ould Rouis who taught me and then accepted to sepervise this work; thank you.
I would like to thank the examination committee members for having devoted some of their precious time to very careful readings of this thesis and for their comments, criticisms and suggestions.
I would like also to express my gratitude to my dear brother and my friend Dr. Mohammed Laidi for all the comments, advice, and assistance he have given to me.

List of Abbreviations and Symbols

rv	: random variable;
$a.s$	: almost surely;
$i.i.d.$	: independent, identically distributed;
df	: distribution function;
$\bar{F}$	: Tail of the df F: $\bar{F}1 - F$;
GEV	: Generalized Extreme Value Distribution;
GPD	: Generalized Pareto Distribution;
BM	: Block Maxima;
POT	: Peak Over Threshold;
$\mathbb{P}(A)$	: Probability of event A;
$\mathbb{E}(X)$	: Expectation of the rv X;
$\xrightarrow{d}$	: Convergence in distribution;
$\xrightarrow{a.s.}$	: Convergence almost surely;
$Var(X)$	: Variance of the random variable X;
$Cov(X, Y)$	: Covariance between X and Y;
$\mathbb{1}_A$	: Indicator function of the set (event) A;
$\mathbb{R}$	: Real numbers set;
$\mathbb{R}^+$	: Positive real numbers set;
$\mathbb{N}$	: set of positive integers;
$F^{\leftarrow}(p)$	: $p - quantile$ of F;
$\mathcal{N}(\mu, \sigma^2)$, $\mathcal{N}(\mu, \Sigma)$	: Gaussian distribution with mean μ and variance σ^2 or covariance matrix Σ.

ملخص:

الهدف من هذا البحث هو إنشاء مقدر لمتوسط التوزيع ثقيل الذيل في حالة الاختفاء العشوائي الأيمن للبيانات ووجود المتغيرات المشتركة من خلال الجمع بين مقدر Kaplan-Meier المعمم قبل العتبة ونموذج وسيطي (توزيع باريتو المعمم GPD) والذي يقرب التجاوزات فوق العتبة، وهذا من أجل إيجاد مخرج للسلوك السيئ لمقدر KM في ذيل التوزيع الثقيل، ثم حددنا قانون التوزيع الطبيعي المقارب لمقدرنا في حالة المتغيرات المشتركة اللاعشوائية.

وكتطبيق لنظرية القيم المتطرفة على الهيدرولوجيا، وتحديدًا، على التساقط، فقد استطعنا معرفة التوزيعات الأنسب للتساقط في منطقة خميس مليانة (الجزائر) خلال الفترة 1975-2006، وذلك باستعمال طريقة الكتل القصوى (Block Maxima) وبالتالي استخدام التوزيع المعمم للقيم المتطرفة (GEV) لنمذجة البيانات، وكذا استعمال طريقة الاختيار فوق العتبة (Peak Over Threshold (POT)) عندما نستخدم توزيع Pareto المعمم (GPD) ، بعد اختبار استقراريه السلسلة الزمنية.

الكلمات المفتاحية: نظرية القيم المتطرفة، التوزيع ذو الذيل الثقيل، مقدر كابلن مبير Kaplan-Meier المعمم، توزيع باريتو المعمم (GPD)، الاختفاء العشوائي الأيمن، مقدر المتوسط الشرطي، الاختيار فوق العتبة، الكتل القصوى، الكميات المتطرفة، زمن الرجوع، مستوى الرجوع.

Abstract

In this thesis we are interested to the statistical analysis of extreme values (EVA) and Modeling of Conditional Extreme Values Distributions under Right Censored Data with Applications. Our goal, in the first place, is to propose an estimator of the mean of a heavy-tailed distribution under right random censored data in presence of covariates by combining the generalized Kaplan-Meier estimator before a threshold and a parametric model; Generalized Pareto Distribution (GPD) which approximates the excesses over threshold in order to overcome the bad behavior of Kaplan Meier estimator (K-M) in the heavy-tail of distribution, then we determined the asymptotic normality of our estimator in case of deterministic covariates. Secondly, as application of extreme value theory to hydrology [1], more specifically, to rainfalls. We have done a study to find out the most adequate fitting distributions of rainfalls taken in Khemis-Miliana region (Algeria) during the period 1975-2006. The method of Block Maxima (BM) is adopted when we use Generalized Extreme Value (GEV) distribution to fit the data, and the Peak Over Threshold (POT) method is applied when we use Generalized Pareto (GP) distribution, after testing stationarity of time serie in hand.

Keywords: Extreme values theory, Heavy-tailed distribution, Generalized Kaplan-Meier estimator, General Pareto Distribution (GPD), Random right censoring, conditional mean estimate, POT, Block Maxima, extreme quantiles, return period, return level.

[1] hydrology is the branch of science or geology that studies the Earth's water.

Résumé

Ce travail de recherche s'inscrit dans le cadre de deux grandes branches de statistique, à savoir la théorie des valeurs extrêmes et l'analyse de survie. Dans le cadre des données de survie, nous avons construit un estimateur de la moyenne d'une distribution à queue lourde avec des données lourdement censurées à droite en présence des covariables. Notre construction est basée sur l'estimateur de Kaplan-Meier généralisé avant un seuil choisi à priori, d'une part. D'autre part, nous avons utilisé le modèle paramétrique Pareto généralisé (GPD) qui approxime mieux les excès au-dessus du seuil afin d'éviter le mauvais comportement de l'estimateur KM pour les valeurs extrêmes, puis nous avons établi la normalité asymptotique de notre estimateur en cas de covariables déterministes. Notre étude pratique (étude de cas) consiste en l'analyse statistique des valeurs extrêmes (EVA) avec applications à l'hydrologie, plus précisément aux précipitations dans la région de Khemis-Miliana dont les valeurs manquantes ne sont pas nombreuses. Plus précisemment, nous avons estimé puis testé les paramètres des modèles probabilistes adéquats aux précipitations enregistrées à la station de Khemis-Miliana durant la période 1975-2006, puis nous avons fait un test d'hypothèse et déduire le modèle le plus approprié. Ceci montre la puissance et l'applicabilité de la théorie des valeur dans le domaine de l'hydrologie. Cette modélisation nous a permis d'estimer des valeurs d'une grande importance dans le monde réel, notamment pour la prévision et la décision, telles que le temps de retour d'une valeur extrême (dépassant un certain seuil) et le niveau de retour à court , moyen et long terme (i.e. 5 ans, 10 ans et 100 ans respectivement).

Mots clés : Théorie des valeurs extrêmes, distribution à queue lourde, estimateur de Kaplan-Meier généralisé, distribution généralisée de Pareto (GPD), censure aléatoire à droite, estimation de la moyenne conditionnelle, POT, Block Maxima, quantiles extrêmes, période de retour, niveau de retour.

Contents

List of Figures

List of Tables

General Introduction

"It is improbable for the impossible to never happen." ([77])

History: "The founders of the calculus of probabilities were too occupied with the general behavior of statistical masses to be interested in the extremes. (N. Bernoulli, in his actuarial problem (1709): n men of equal age die within t years. What is the mean duration of life of the last survivor? he reduces this question to the following: n points lie at random on a straight line of length t. Then he calculates the mean, largest distance from the origin.)"([77] p.2).

Both of extreme values and Poisson's law deal with small probabilities (rare events), where the first considers the size of rare events and the second gives their number.

L. von Bortkiewics studied Extreme Values for the first time in 1922. One year later, R. von Mise introduced the fundamental notion of the characteristic largest value (using an other name). In 1925, L.H.C. Tippett gave tables (Tippett's tables) which were the fundamental tools for all practical uses of largest values from normal distributions only.

E.L. Dodd (1923) was the first who studied largest values for non normal distributions in his work : the Greatest and the Least Variate Under General Laws of Error ([50])

M. Fréchet (1927) was the first who: published a paper based on the concept of a type of initial distributions (these were not very frequent)different from the normal one, obtained an asymptotic distribution of the largest value but more than that he proved that these largest values taken from different initial distributions sharing a common property may have a common asymptotic distribution.

In the next year (1928), R.A. Fisher and L.H.C. Tippet published paper which made the foundations of the asymptotic argument forming the backbone of extreme value theory. They found in addition to Fréchet's asymptotic distribution two others adequate for other initial types and showed the reason of the slow convergence of the distribution of the largest normal value toward its asymptote.

In 1936, R. von Mises classified the initial distributions possessing asymptotic distributions of the largest value, and gave sufficient conditions under which the three asymptotic distributions are valid. In 1943, B. Gnedenko gave necessary and sufficient conditions. In 1948, G. Elfving and E.J. Gumbel derived the relations of the asymptotic distribution of the normal range to certain Bessel functions. In his thesis (1954)R.A. da Silva Leme gave a systematic expository treatment of the asymptotic distributions of extreme values and their applications to engineering problems, especially the safety of structures, this problem was studied in great detail by Arne L. johnson in 1953. The first book treating extreme value theory and extreme value statistics is E.J. Gumbel's Statistics of Extremes [77]. In this book, E.J. Gumbel also provides a historical account of extreme value theory since its beginnings. There was little theoretical development until a major flood in the Netherlands killed more than 1800 people in 1953, this phenomena led Dutch mathematician attracted by the field of extreme value theory, among them Laurens de Haan and Guus Balkema. In 1974, Laurens de Haan and Guus Balkema, and independently J. Pickands, found the limit distribution of the excesses of an iid sequence above high thresholds, called Generalized Pareto Distribution (GPD). Their result gave a theoretical basis to the Peaks-Over-Threshold method which had been used by hydrologists for modeling extreme excesses since the 1950s.

In the 1970s, 1980s and 1990s, the foundations were laid for an extreme value theory of dependent sequences. Pioneering work was done by R. Leadbetter, H. Rootz´en, S. Resnick, J. Hüsler, T. de Olivera, R.A. Davis, T. Hsing, and many others. In 1983, Leadbetter, Lindgren and Rootz´en published the book that was the first one treating the extremes of stationary sequences (see [101]; it essentially solved the problem for Gaussian sequences in a rather complete way. They discussed extremal clusters and how to describe them in a quantitative way. Few years later, in 1987, S.I. Resnick published his important book entitled Extreme Values, Regular Variation, and Point Processes ([127]). This book focused on the relationship between the weak convergence of the point processes of the exceedances in a sample and the distributional convergence of the maxima and largest order statistics. It also provided a rigorous extreme value theory for a multivariate iid sequence. Castillo [30] has successfully updated [77] and gave many statistical applications of Extreme Value Theory. In 1997, P. Embrechts, C. Klüppelberg and T. Micosch wrote their book "Modelling Extremal Events for Insurance and Finance" not to solve problems re-

lated to calculus of very high quantiles of the return distributions of speculative assets, possibly outside the range of the data using extreme value theory but the book aimed to bring theoretical results about extremes to a very wide audience, including undergraduate students at many universities. The two best praises rightly said toward extreme value theory in a convincing way are cited by [61]: - Richard Smith: "There is always going to be an element of doubt, as one is extrapolating into areas one doesn't know about. But what EVT is doing is making the best use of whatever data you have about extreme phenomena." - Jonathan Tawn: " The key message is that EVT cannot do magic – but it can do a whole lot better than empirical curve–fitting and guesswork. My answer to the sceptics is that if people are not given well founded methods like EVT, they will just use dubious ones instead."
Since 1998 theory and practice of extreme value theory have been discussed through biannual international conferences, one can cite the one held in Delft at the end of June 2017 focused on topics related to stochastic processes related to extreme values. Those topics included the important class of max-stable processes and random fields introduced by L. de Haan in 1984 which have proven in a final way useful for the modeling and statistical analysis of extreme weather and climate phenomena. Fertility of extreme value theory pushed related community to become more open to fields like time series analysis, stochastic networks, telecommunications, branching,... Extreme value theory of high-dimensional structures is a hot topic, e.g. the extreme eigenvalues of random matrices or finding the important (extreme) components in complex stochastic systems.
Heavy tailed data frequently appear in insurance and finance, and a loss variable with a heavy tail rarely creates unusually huge losses. Unfortunately such an extreme loss often causes severe damages to our society. Extreme value theory has been developed to model, analyze and predict such an extreme event for decades. Several excellent books on extreme value theory have been available in the litterature such as Leadbetter et al. [102], Resnick [127], Embrechts et al. [61], Coles [34], Beirlant et al. [14], de Haan and Ferreira [45] and Novak [119].

In the case of complete data, there is a whole theory (extreme value theory; TVE). Analysis of extremes is done according to two approaches. The first one, called the GEV approach; allows to model the maximum blocks by a GEV distribution (generalized extreme value distribution) and the second, called GPD approach consists in fitting the observations exceeding a certain threshold (peaks over threshold POT)

by a GPD (generalized Pareto distribution).
The analysis of random censored extreme values is a new subject of research, especially when covariate is present. The modeling of censored extreme values is first seen in 1997 in the literature of extremes with the publication of the book Reiss and Thomas ([126]). In 2007, Beirlant et al. ([15]) really approached extreme values nonparametric statistics with censored data. Their estimator is based on a standard tail index estimator divided by the estimator of the proportion of uncensored data exceeding a certain given threshold. They have applied this theory to AIDS data (survival times of 100 HIV members in a follow up study). Then, Einmahl et al. (2008, [60]) used the same concept to propose an estimator of the tail index on the k-largest values, determined its asymptotic properties and finally illustrate its behavior on the same AIDS data. Later, research on the theory of censored extreme values has become a topical issue.
The main aim of this thesis is to extend the results of extreme value theory to the case where the sample consists of a censored data under fixed covariates, namely the mean of Pareto-type distributions.

Given samples $X_1,...,X_n$ from a distribution F, can we estimate the mean of F? This is the problem of mean estimation which is, alongside hypothesis testing, one of the most fundamental questions in statistics. As a result, answers to this problem are known in fairly general settings. For instance, the empirical mean is known to be an optimal estimate of a distribution's true mean under minimal assumptions. Unfortunately, it can do not always exist according to the stability of α, $\alpha = 1/\gamma$ where γ is the Extreme Value Index (EVI):

- $0 \leq \alpha \leq 1$: The mean and the variance are both infinite, as an example Cauchy distribution ($\alpha = 1$);
- $1 < \alpha < 2$: The mean is finite and the variance is infinite, as the heavy tailed distributions;
- $\alpha = 2$: The mean and the variance are finite, as Gauss distributions

When the condition of finiteness of second order mean is deprived, Peng [120], Johansson [96] and Peng [121] proposed Gaussian asymptotic estimator by exploiting the tools of extreme value theory. For both authors, the estimators are based on all observations. When the variable of interest is right censored by another variable,

estimation techniques based on complete data become inappropriate. In this situation, Stute [140] introduced an asymptotically normal estimator of the mean of a finite second order moment distribution using Kaplan-Meier estimator [97].

Our contribution in this thesis is to propose an estimator of the mean of heavy tailed distribution under censoring and presence of covariates taking advantage of several recent works restricted to conditional and right censoring case, we can cite for example [16], [114] concerning EVI estimators and [63] concerning survival function estimator.
This thesis is a combination between two fields of statistics: survival analysis and extreme value theory. We will provide a summary of the different notions and fundamental properties of these two domains of statistics that help reader to understand. So, this thesis is organized as follows:
Chapter 1 is devoted to generalities which defines the whole concepts and results that we will use in the rest of this work. We start by recalling the results on order statistics and extremes and we give exact distributions of order statistics, the latter allows us to introduce fundamental results on extreme value theory and asymptotic distributions in value theory extremes.

These distributions depend on an unknown real parameter γ called the tail index or the extreme values index (EVI) which its knowledge is of great importance in theory of extreme values. We will see that this parameter controls the behavior of the distribution tail. The larger is γ, the heavier the tail is. According to this shape parameter, we will also characterize the different maximum domains of attraction. Therefore its estimate has great interest. Then we recall some classical estimation methods of this parameter, namely Hill, Pickands, and Moments estimators in case of complete data where we toggle to estimating the extreme quantiles for each method and we conclude by giving available procedures for selecting the number of extreme values used for the estimate of this parameter. In Chapter 2, we place ourselves in the general case, that is we define estimators of extreme values index and extreme quantiles of heavy tailed functions when covariates are present and data are incomplete and then we give asymptotic properties. Within this section, we recall some survival analysis notions which are necessary to reach the objective of this section, as survival, hazard an risk functions and their estimators

The third chapter is about application of extreme value tools in the field of hydrology. In Section 2, we give the basic concepts of the classical block maxima

(for extremes) and threshold exceedances (for some high threshold) models and then making inferences for both models including its parameters and return levels (quantiles) by ML and PWM methods. Finaly, Section 3 is devoted to an application to the extreme rainfalls at Khemis -Miliana station (Algeria). This chapter is an article *"Toufik Guermah Abdelaziz Rassoul (2020) Study of extreme rainfalls using extreme value theory (case study: Khemis-Miliana region, Algeria), Communications in Statistics: Case Studies, Data Analysis and Applications, 6:3, 364-379, DOI: 10.1080/23737484.2020.1789901"* see [85]

In the last chapter, we give our main result consisting in adapting the mean estimator of heavy tailed distribution, established by Johansson [96], to right random censoring and to presence of fixed covariates inspired from Ndao PhD thesis [114]. In section 2 we construct our estimator by dividing data into parts; below threshold and above it, where data exceeding threshold are fitted by general Pareto Distribution (GPD). Section 3 is specified to our main result; first we give sufficient conditions and then we state asymptotic normality of the estimator as well as its proof. This chapter is an article submitted for publication.

Papers and Communications

Papers

1. Guermah, T., Rassoul, A. (2020). Study of extreme rainfalls using extreme value theory (case study: Khemis-Miliana region, Algeria). Communications in Statistics: Case Studies, Data Analysis and Applications, 6(3), 364-379.

2. Guermah, T., Rassoul, A. Estimating the Conditional Mean of Heavy-Tailed Distribution under random right censoring. (submitted for publication to Extremes; Statistical Theory and Applications in Science, Engineering and Economics).

Communication

Study of extreme rainfalls using extreme value theory (case study: Khemis-Miliana region, Algeria) Authors T. Guermah, A. Rassoul 11. International Statistics Days Conference, 3 - 7 October 2018. Muğla Sıtkı Koçman University, TURKEY.

Chapter 1

Extreme Value Theory

Abstract In this chapter, we will present principal definitions and classical results on Extreme Values Theory in uni-dimensional case.

1.1 Introduction

The goal of extreme value theory is to study large observations asymptotic behavior of a sample of independent random variables and identically distributed. The standard approach in probability theory places emphasis on average behavior and variability around the average, through probabilistic tools such as the law of large numbers or the central theorem limit. The fundamental theorem of extreme value theory (Known as Fisher-Tippett's theorem) gives the possible limit laws of the maximum of the sample and thus provides some knowledge of the tail distribution. The use of the laws of extreme values is based on statistical properties order and extrapolation methods. More precisely, it is based mainly on the limit distributions of the extremes and their domains of attraction. In this part, we recall some essential notions on the extreme values theory and on the notion of censorship which add reader to understand this thesis. Thus, we present briefly the essential results encountered in the literature. We quickly define the notions of attraction domain , regular variation functions then we characterize them in the one-dimensional case. As for censorship we will present some definitions linked to the statistics of survival times. Censorship is founded with a few functions such as distribution function, survival function and risk function. Many authors have been interested in the concept, in particular Kaplan and Meier (1958, [97]), who proposed an estimator of the survival function which Beran generalized ([22]) in the conditional case called the

Generalized or conditional Kaplan-Meier estimator.

1.2 Order Statistics and their Exceedances

"Don't trust random number" E. J. Gumbel.

1.2.1 Order statistics and Extremes

The study of order statistics is necessary because extreme values are special cases of order statistics. Conversely, certain problem involving order statistics can easily be solved from the theory of extreme values, especially the k-th largest observations problems, see [77] and [14], the latter explained why it would be unrealistic to assume that only the maximum of a sample contains valuable information about the tail of a distribution and that other large order statistics could do this as well. In practice, order statistics are very important and in particular the minimum and the maximum, because they are the critical values used in engineering, physics, medicine, etc.; see [31].

Definition 1.1 ***(Order statistics)***
Let $X_1, X_2, ...$ denote a sequence of iid non-degenerate rvs with common df F. Define the ordered sample of size n

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

Hence $X_{1,n} = min(X_1, X_2, ..., X_n)$ and $X_{n,n} = max(X_1, X_2, ..., X_n)$. The rv $X_{k,n}$ is called the kth upper order statistic. The notation of order statistics is not the same from an author to other; some denote by $X_{1,n}$ the maximum and by $X_{n,n}$ the minimum of a sample.

Definition 1.2 *(Extreme order statistics)*
The rv $X_{1,n}$ is the smallest order statistic (minimum statistic) and $X_{n,n}$ is the largest order statistic (maximum statistic). Note that it is very easy to switch from one to the other using the relationship

$$min(X_1, ..., X_n) = -max(-X_1, ..., -X_n)$$

1.2.2 Distribution function of the kth upper order statistic

Proposition 1.1 *For $k = 1, ..., n$ and $x \in \mathbb{R}$ let $F_{k,n}$ denote the df of $X_{k,n}$. Then*

(a)

$$F_{k,n}(x) = \sum_{r=0}^{k-1} \binom{n}{r} \bar{F}^r(x) F^{n-r}(x)$$

(b) If F is continuous, then

$$F_{k,n}(x) = \int_{-\infty}^{x} f_{k,n}(z)\, dF(z),$$

where

$$f_{k,n}(x) = \frac{n!}{(k-1)!(n-k)!} \bar{F}^{k-1}(x) F^{n-k}(x);$$

i.e. $f_{k,n}$ is a density of $F_{k,n}$ with respect to F.

Proof.[61]

(a) For $n \in N$ define

$$B_n = \sum_{i=1}^{n} I_{\{X>x\}}$$

Then B_n is a sum of n iid Bernoulli variables with success probability

$$E I_{\{X>x\}} = P(X > x) = \bar{F}(x).$$

Hence B_n is a binomial random variable with parameters n and $\bar{F}(x)$. Furthermore, we know that
$(X_{k,n} \le x)$ if and only if $\left(\sum_{i=1}^{n} I_{\{X>x\}} < k\right)$,
Consequently for $x \in \mathbb{R}$

$$\begin{aligned} F_{k,n} &= P(B_n < k) \\ &= \sum_{r=0}^{k-1} P(B_n = r) \\ &= \sum_{r=0}^{k-1} \binom{n}{r} \bar{F}^r(x) F^{n-r}(x). \end{aligned}$$

(b) Using the continuity of F, we calculate

$$\begin{aligned} &\frac{n!}{(k-1)!(n-k)!} \int_{-\infty}^{x} \bar{F}^{k-1}(z) F^{n-k}(z)\, dF(z) \\ &= \frac{n!}{(k-1)!(n-k)!} \int_{\bar{F}(x)}^{1} t^{k-1}(1-t)^{n-k}\, dt \\ &= \sum_{r=0}^{k-1} \binom{n}{r} \bar{F}^r(x) F^{n-r}(x) = F_{k,n}(x). \end{aligned}$$

■

By similar arguments, the joint distribution of a finite number of different order statistics. If F is absolutely continuous with density f, then the joint density of $(X_1,...,X_n)$ is

$$f_{(X_1,...,X_n)}(x_1,...,x_n) = \prod_{i=1}^{n} f(x_i),\ (x_1,...,x_n) \in R^n.$$

Since the n values of $(X_1,...,X_n)$ can be rearranged in $n!$ ways, every ordered sequence $(X_{k,n})_{k=1,...,n}$ could have come from $n!$ different samples, see for instance [61] and [124]. The joint density of the ordered samples is then:

$$f_{(X_{1,n},...,X_{n,n})}(x_1,...,x_n) = n! \prod_{i=1}^{n} f(x_i),\ (x_1 < ... < x_n). \tag{1.1}$$

The following theorem gives marginal densities as immediate consequence of (1.1).

Theorem 1.1 *(Joint density of k upper order statistics) If F is absolutely continuous with density f, then*

$$f_{(X_{1,n},...,X_{n,n})}(x_1,...,x_k) = \frac{n!}{(n-k)!} F^{n-k}(x_k) \prod_{i=1}^{n} f(x_i),\ (x_1 < ... < x_k).$$

Among quantities which are the basic building block of several estimators, Especially Hill's estimator, we find the *spacings*(or distances as called in [77]), that is the differences between successive order statistics; see [61]

Definition 1.3 *(Spacings of a sample) For a sample $X_1,...,X_n$ the spacings are defined by*

$$X_{k+1,n} - X_{k,n},\ k = 1,...,n-1.$$

For random variables with finite left (right) endpoint $\tilde{x}_F(x_F)$ we define the nth (0th) spacing as

$$X_{1,n} - X_{0,n} = X_{1,n} - \tilde{x}_F\ (X_{n+1,n} - X_{n,n} = x_F - X_{n,n}).$$

For examples; see [61](pp. 185-188).

The next notion is the so called *quantile transformation*. It is extremely useful since it often converts order statistics' problem to its version order statistics issue from a uniform sample.

Lemma 1.1 *(Quantile transformation) Let $X_1,...,X_n$ be iid with df F. Furthermore, let $U_1,...,U_n$ be iid random variables uniformly distributed on (0,1) and denote by $U_{1,n} < ... < U_{n,n}$ the corresponding order statistics. Then the following results hold:*

(a) $F^{\leftarrow}(U_1) \stackrel{d}{=} X_1$.

(b) For every $n \in \mathbb{N}$,

$$(X_{1,n},...,X_{n,n}) \stackrel{d}{=} (F^{\leftarrow}(U_{1,n},...,F^{\leftarrow}(U_{n,n})).$$

(c) The random variable $F(X_1)$ has a uniform distribution on (0,1) if and only if F is a continuous function.

Proof. Follows immediately from the definition of the uniform distribution. ■

Proposition 1.2 *(Almost sure convergence of order statistics)*
Let F be a df with right (left) endpoint $x_F \leq \infty (\tilde{x}_F \geq -\infty)$ and $(k(n))$ a non-decreasing integer sequence such that

$$\lim_{n\to\infty} n^{-1}k(n) = c \in [0,1].$$

(a) Then $X_{k(n),n} \stackrel{a.s.}{\to} x_F (\tilde{x}_F$ *according as* $c = 0 (c = 1)$.

(b) Assume that $c \in (0,1)$ is such that there is a unique solution $x(c)$ of the equation $\bar{F}(x) = c$. Then

$$X_{k(n),n} \stackrel{a.s.}{\to} x(c).$$

One can find proof in [61] (p.195).

1.3 Extreme value Distribution (EVD)

1.3.1 Max-stable distribution

The study of the uni-variate extreme values distributions is based on a parallel approach to that of the Central Limit Theorem, which is the basis of the inferential statistic. This theorem establishes that for a sequence of real random variables i.i.d. $X_1,...,X_n$, of finite variance σ^2 and mean μ, then

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \xrightarrow[n\to\infty]{\text{distribution}} \mathcal{N}(0,1);$$

or in terms of distribution function:

$$\lim_{n\to\infty} P\left\{\sqrt{n}\left(\frac{\bar{X}_n - E(X)}{\sigma_X}\right)\right\} = \Phi(x);$$

where Φ is the standard normal df and $\bar{X}_n$ is the sample mean random variable, that is

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

This theorem means that it exist normalization sequences

$$\{\mu_n = n\mu \in \mathbb{R}\} \quad and \quad \left\{\sigma_n = \sigma\sqrt{n} > 0\right\};$$

in such a manner that the sequence of standardized random variables

$$Y_n = \frac{S_n - \mu_n}{\sigma_n}$$

converges in distribution to standard normal distribution, i.e

$$P(Y_n) = \Phi(y),\ for\ all\ y \in \mathbb{R}.$$

In addition to the central limit theorem, one can ask the next natural question, for the maximum statistic: can we find normalization sequences $\mu_n \in \mathbb{R}$, $\sigma_n > 0$ and a non-degenerate distribution H such that

$$\frac{X_{n,n} - \mu_n}{\sigma_n} \xrightarrow[n\to\infty]{\mathrm{d}} G\ i.e\ \lim_{n\to\infty} P(M_n \le \sigma_n x + \mu_n) = H(x). \tag{1.2}$$

If the relation (1.2) is satisfied, we say that the respective variable is *max-stable*.

Definition 1.4 ***(Max-stability of a distribution)***
A non-degenerate variable X and its distribution function F are said to be max stables if it exists a normalization coefficients sequences $\mu_n \in \mathbb{R}$, $\sigma_n > 0$ such that $M_n \overset{d}{=} \sigma_n X + \mu_n$ for a sample of variables i.i.d. $(X_i; i = 1..n)$.

For more explication, see [14].

1.3.2 Fundamental theorem of extreme values

Let $(X_n)_n \ge 1$ a sequence of independent copies of a random variable X having distribution function $F(x) = P(X_1)$. The central result of extreme value theory concerns

the asymptotic distribution H of the maximum. Since the random variables are independent and identically distributed, then the distribution function $F_{X_{n,n}}$ of the maximum $X_{n,n}$ is given by

$$=[\mathrm{F(x)}]^n. \quad (1.3)$$

The standardization with σ_n and μ_n in (1.2) appears natural since otherwise and according to (1.3.2) $M_n \to x_F$ a.s. which make H a degenerate distribution. So we have a double problem: (1) find all possible non-degenerate distributions H that can occur as a limit for sample maxima of independent and identically distributed random variables as in (1.2); (2) specify the distributions F for which there exist sequences σ_n and μ_n such that (1.2) holds for each of those limit distributions.
The solution of problem (1) (called extremal limit problem) is given by Fisher and Tippett [69], Gnedenko [78] and de Haan [42]. These distributions are called *extreme value distributions.* The solution of problem (2) means that for any such specific limit distribution, one shall find necessary and sufficient conditions on the initial distribution F satisfying (1.2). The class of such distributions is called the maximum domain of attraction or simply domain of attraction of H and is often denoted by $D(H)$. In this context, Laurens de Haan reformulated relation (1.2) in two other ways and identified all extreme value distributions and their domains of attraction; see [45].

Definition 1.5 *A distribution function F is said to belong to the domain of attraction of a non-degenerate distribution function G (we write $F \in D(G)$) if there exist sequences of real numbers $a_n > 0$ and $b_n \in \mathbb{R}$ such that*

$$\lim_{n\to\infty} F^n(a_n x + b_n) = G(x) \text{ or } F^n(a_n x + b_n) \xrightarrow{d} G(x) \quad (1.4)$$

for all continuity points x of G.

Our first problem can thus be formulated as follows: find all distribution functions with non-empty domains of attraction. Due to Khinchine's convergence to type theorem (see [66], Ch. VIII.2, lemma 1), a df F cannot be in the domain of attraction of two essentially different df's. The results of the lemma lead to the following definition ([43]).

Definition 1.6 *The distribution functions F_1 and F_2 are the same type if there exist two constants $a > 0$ and $b \in \mathbb{R}$ such that $F_2(x) = F_1(ax + b)$ for all real x.*

It is clear from the definition that the relation *"are of the same type"* between *d f's* is an equivalence relation. Consequently it defines equivalence classes of distribution functions called *types*.
Now we are in a position to identify the class of non-degenerate distributions that can occur as a limit in the basic relation (1.4). This class of distributions was called the class of extreme value distributions as mentioned above.

Theorem 1.2 *Fisher and Tippet (1928), Gnedenko (1943)*
Let $(X_n)_{n\geq 1}$ a sequence of independent and identically distributed random variables with $F(x) = P(X_1 \leq x)$. If there exist two normalization sequences $a_n > 0$ and $b_n \in \mathbb{R}$ and a non-degenerative distribution H such that:

$$\begin{aligned} \lim_{n\to\infty} P(X_{n,n} \leq a_n x + b_n) &= \lim_{n\to\infty} F^n(a_n x + b_n) \\ &= H(x), \ for\ all\ x, \end{aligned} \tag{1.5}$$

Then H is of same type as one of the three following distributions:

$$\Phi_\alpha(x) = exp(-(x)^{-\alpha})\mathbb{1}_{\{x\geq 0\}},\ \alpha > 0 (Fréchet\ distribution)$$

$$\Gamma(x) = exp(-e^{-x}), (Gumbel\ distribution)$$

$$\Psi_\alpha(x) = exp(-(-x)^{-\alpha})\mathbb{1}_{\{x<0\}} + \mathbb{1}_{\{x\geq 0\}},\ \alpha > 0 (Weibull\ distribution)$$

where $\mathbb{1}_A$ is the indicator function of the set A.

For detailed proof of this theorem, see [127] (p.9) or with more expansions [61](p.152).
Sequences $a_n > 0$ and $b_n \in \mathbb{R}$ depends on distribution parameters of X. Figure 1.1 illustrates in case of standard normal distribution, the convergence of random variables sequence $(a_n^{-1}(X_{n,n}))_{n\geq 1}$ in distribution to a non-degenerate random variable Γ. As an example, de Haan [45](p.11 example 1.1.7) used the following theoretic re-normalization sequences associated to standard normal distribution:

$$a_n = (2/logn)^{-1/2}\ and\ b_n = (2/logn)^{1/2} - \frac{/log\log n + \log 4\pi}{2(2\log n)^{1/2}}.$$

Note that (1.5) is equivalent to the following statement (see for other equivalent statement and proof [45] (theorem 1.1.2 p.5)

$$\lim_{t\to\infty} t(1 - F(a(t)x + b(t))) = -logH(x),$$

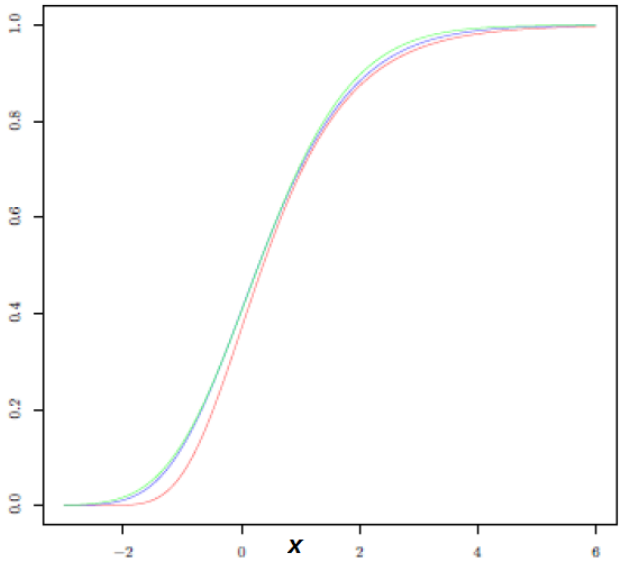

Figure 1.1: Illustration of extreme value theorem on standard normal distribution. comparison between $H_0(x)$(red), $\mathbb{P}\left[\frac{X_{n,n}-b_n}{a_n} \leq x\right]$ with $n = 100$(blue) and $\mathbb{P}\left[\frac{X_{n,n}-b_n}{a_n} \leq x\right]$ with $n = 10$ (green).

for each continuity point x of G for which $0 < G(x) < 1$, $a(t) := a_{[t]}$ and $b(t) := b_{[t]}$ (with $[t]$ the integer part of t .

Maximum Domains of Attraction

Definition 1.7 *(Maximum domain of attraction) The random variable X (resp. its distribution function F, is said to belong to the maximum domain of attraction of the extreme value distribution H if there exist two normalization sequences $a_n > 0$ and $b_n \in \mathbb{R}$ and a non-degenerative distribution H such that equation 1.5 holds. We write $X \in MDA(H)$ (resp. $F \in MDA(H)$.*

An important result which shows that the limit distribution functions form a simple explicit one -parameter family (called extreme value distributions family or general extreme value distributions (GEVD)) is the parametrization theorem due to von Mises [152] and Jenkinson [95], so we can write

$$\Lambda_\gamma(x) = \begin{cases} \exp\left[-(1+\gamma x)^{-1/\gamma}\right] & \text{if } \gamma \neq 0 \\ \exp\left[-\exp(-x)\right] & \text{if } \gamma = 0 \end{cases} \quad \text{and for all } x \text{ such that } 1+\gamma x > 0. \quad (1.6)$$

One can introduce the related location–scale family $H_{\gamma;\mu,\sigma}$ by replacing the argument x above by $(x-\mu)/\sigma$ for $\mu \in \mathbb{R}$, $\sigma > 0$.

Definition 1.8 *(Extreme Value Index (EVI)) The parameter γ in (1.6) is called the extreme value index.*

This result shows also that the class contains distributions with completely different features, let us consider the subclasses $\gamma > 0$, $\gamma = 0$ and $\gamma < 0$ separately:

(a) For $\gamma > 0$ clearly $\Lambda_\gamma(x) < 1$ for all x, i.e., the right endpoint of the distribution is infinity. Moreover, as $x \to \infty$, $1 - \Lambda_\gamma(x) \sim \gamma^{-1/\gamma} x^{-1/\gamma}$, which means that the distribution presents a heavy right tail, in this case, moments of order greater than or equal to $1/\gamma$ are infinite (see [45](exercise 1.16).

(b) For $\gamma = 0$ the right endpoint of the distribution is infinite where the distribution has a light right tail: $1 - \Lambda_0(x) \sim e^{-x}$ as $x \to \infty$, and therefore all moments are finite.

(c) for $\gamma < 0$ the right endpoint of the distribution is $-1/\gamma$ so it is short-tailed, where $1 - \Lambda_\gamma(-\gamma^{-1} - x) \sim \gamma^{-(\gamma} x)^{-1/\gamma}$, as $x \to 0$.

Figure 1.2 illustrates the behavior of different GEV densities.

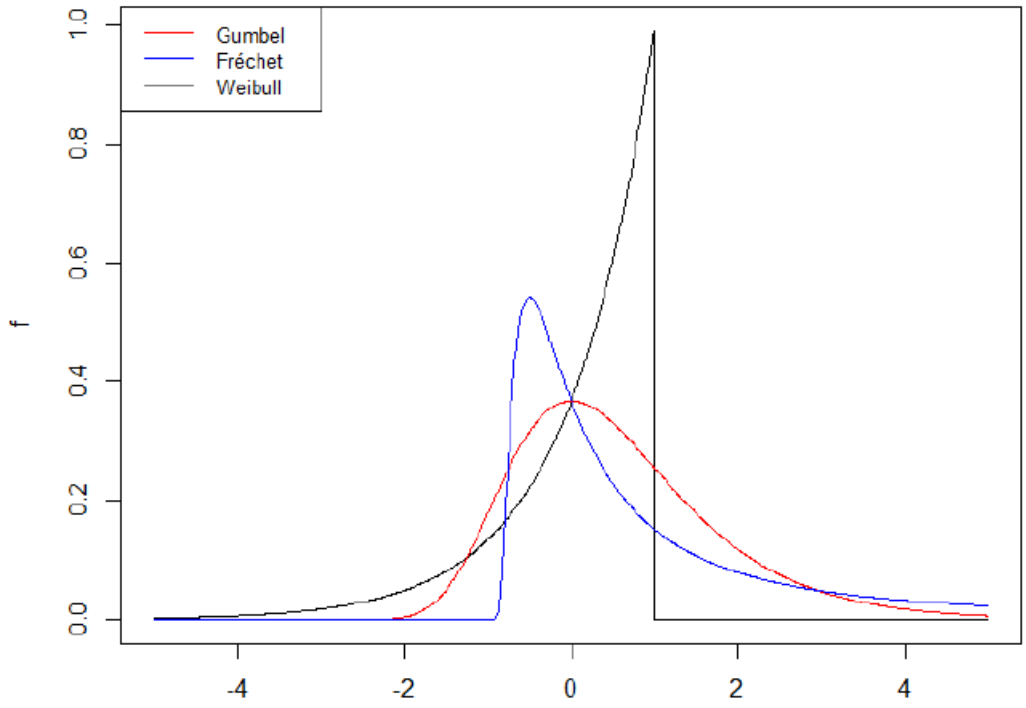

Figure 1.2: Example of extreme values densities $\gamma = -1$(black), $\gamma = 1$(blue) and $\gamma = 0$ (red).

Conditional distribution of excesses

The fact that $F \in MDA(H_\gamma)$ is equivalent to the following assertion; see [61] (Theorem 3.4.5 p.158):
There exists a positive, measurable function a(.) such that for $1 + \gamma x > 0$,

$$\lim_{u \uparrow x_F} \frac{\bar{F}(xa(u))}{\bar{F}(u)} = \begin{cases} (1+\gamma x)^{-1/\gamma} & \text{if } \gamma \neq 0, \\ \exp(-x) & \text{if } \gamma = 0. \end{cases} \tag{1.7}$$

This condition ((1.7)) has a probabilistic meaning; in fact let X be a random variable with distribution function $F \in MDA(H_\gamma)$, so (1.7) can be rewritten as

$$\lim_{u \uparrow x_F} P\left(\frac{X-u}{a(u)} > x / X > u\right) = \begin{cases} (1+\gamma x)^{-1/\gamma} & \text{if } \gamma \neq 0, \\ \exp(-x) & \text{if } \gamma = 0. \end{cases} \tag{1.8}$$

Let us define a distributional approximation for the scaled excesses over the threshold u, where the appropriate scaling factor is $a(u)$. Readers can see [61], section 6.5 for many applications of this interpretation. The next definition is motivated by the right-hand side limit in (1.8).

Definition 1.9 *(The generalized Pareto distribution (GPD)) Define the distribution function G_γ by*

$$G_\gamma(x) = \begin{cases} 1 - (1+\gamma x)^{-1/\gamma} & \text{if } \gamma \neq 0, \\ 1 - \exp(-x) & \text{if } \gamma = 0. \end{cases} \tag{1.9}$$

where

$$\begin{cases} x \geq 0 & \text{if } \gamma \geq 0, \\ 0 \leq x \leq -1/\gamma & \text{if } \gamma < 0. \end{cases}$$

G_γ is called a standard generalized Pareto distribution (GPD). One can introduce the related location-scale family $H_{\gamma;\mu,\sigma}$ by replacing the argument x above by $(x-\mu)/\sigma$ for $\mu \in \mathbb{R}$, $\sigma > 0$. The support has to be adjusted accordingly.

Without lost of generality, we will restrict ourselves to the distribution function $H_{\gamma;0,\sigma}$ which plays an important role in fitting excesses over a high threshold. The bridge that links GEV with GP distributions is based on the limit law for excess distributions over threshold due to Balkema and de Haan [9] and Pickands [123].

Theorem 1.3 *Denote by*

$$F_u(x) = P\left(X - u \leq x / X > u\right)$$

the conditional distribution of the excess of X over the threshold u, given that μ is exceeded.

If F belongs to one of three attraction domains of extreme value distributions (Fréchet, Gumbel or Weibull), then it exists a function σ(u) strictly increasing and a real number γ such that

$$\lim_{u \uparrow x_F} \sup_{0 \leq y \leq x_F - u} |F_u(y) - G_{\gamma,\sigma(u)}(y)| = 0 \tag{1.10}$$

where $G_{\gamma,\sigma}$ is the generalized Pareto distribution defined by

$$G_{\gamma,\sigma}(y) = \begin{cases} 1 - (1 + \gamma y/\sigma)^{-1/\gamma} & \text{if } \gamma \neq 0,\ \sigma > 0, \\ 1 - \exp(-y/\sigma) & \text{if } \gamma = 0,\ \sigma > 0. \end{cases} \tag{1.11}$$

$$\begin{aligned} y &\in [0,\ (x_F - u)] && \text{if } \gamma \geq 0, \\ y &\in [0,\ \min(-\sigma/\gamma,\ x_F - u)] && \text{if } \gamma < 0. \end{aligned}$$

Thus, for an enough high value of u, the excesses distribution is fitted by a generalized Pareto distribution:$F_u \approx G_{\gamma,\sigma(u)}$. The parameters γ and σ of the generalized Pareto distribution are the same as those of GEV distribution. It is interesting to note that the case $\gamma = 0$ corresponds to the exponential distribution of mean σ and the case $\gamma = 1$ corresponds to the uniform distribution on $[0, \sigma]$.

1.4 Attraction domains Characterization

One of the big problem in Extreme Value Theory is to ask for necessary and sufficient conditions for existence of centering constants a_n and norming constants b_n to get convergence in equation (1.2); one says that "F is attracted to H". According to Bingham [24], the first clear glimpse of the crucial role of regular variation in probability theory (especially in Extreme Value Theory) is provided by the striking answer to the question above: the truncated variance should be slowly varying; see [24](p.170) and [25](Theorem 8.3.1., p.346). This remarkable result has been achieved independently by Khinchin (1935), feller (1935) and Lévy (1935), but both of these works did not introduce the regular variation language, whereas this notion has been already provided by Karamata (1930). In this subsection, we present different characterizations of the three attraction domains given in [44], [127], [61] and

[45] using regular variation tool, where, they gave criterions to be taken such that convergence in (1.2) holds.

1.4.1 Regularly Varying (RV) Functions

Regular varying theory started with the pioneering work by Karamata [98] and passing by Feller [67] in the field of probability theory. theoretical developments in extreme value theory saw light with L. de Haan doctoral dissertation On Regular Variation and its Applications to the Weak Convergence of Sample Extremes in the 70's. We summarize here some of the main results on regular variation theory and for more details see the encyclopedic volume on the subject by Bingham, Goldie and Teugels [25].

Definition 1.10 ((Regularly Varying and slowly varying Functions)) *Let f be an ultimately positive and measurable function on $\mathbb{R}_+$. We will say that f is regularly varying at infinity if and only if there exists a real constant ρ for which*

$$\lim_{x\uparrow\infty} \frac{f(xt)}{f(x)} = t^{\rho} \quad for\ all\ t > 0.$$

We write $f \in RV_\rho$ and we call ρ the index of regular variation. In the case $\rho = 0$, the function will be called slowly varying or of slow variation at infinity. We will reserve the symbol ℓ for such functions.
The class of all regularly varying functions is denoted by RV

Remark 1.1 *1) In the definition above, we have defined regular variation at infinity, i.e. for $x \to \infty$. Analogously we can define regular variation at zero replacing $x \to \infty$ by $x \to 0$, or at any positive number.*
2) For $\alpha, \beta \in \mathbb{R}$ the functions x^α, $x^\alpha(\log x)^\beta$, $x^\alpha(\log\log x)^\beta$ are RV_α.
The functions $2+\sin(\log\log x)$, $\exp((\log x)^\alpha)$, with $0<\alpha<1$, are slowly varying.
The functions $2+\sin x$, $\exp[\log x]$, $2+\sin(\log x)$ are not regularly varying; where $[.]$ stands for integer part.
Typical examples of slowly varying functions are positive constants or functions converging to a positive constant, logarithms and iterated logarithms.
For more examples see [45] (Example B.1.2, p.362) [61](p.565) and [14] (p.78).

Now, we present some important properties which describe the class RV_0 of slowly varying functions, for others, we refer to the literature.

Proposition 1.3 *(**Properties of the class** RV_0)*

(a) RV_0 *is closed under addition, multiplication and division.*

(b) if ℓ *is slowly varying, then* ℓ^α *is also slowly varying for all* $\alpha \in \mathbb{R}$.

(c) If ℓ *is slowly varying,*

$$\lim_{x\to\infty} (\log \ell(x))/\log x = 0.$$

(d) If ℓ *is slowly varying and* $\rho > 0$,

$$\lim_{x\to\infty} x^\rho \ell(x) = \infty, \quad \lim_{x\to\infty} x^{-\rho} \ell(x) = 0.$$

One can transform a regular varying problem at infinity into slowly varying one through the next result.

Proposition 1.4 *Let* $\rho \in \mathbb{R}$ *and* $f \in RV_\rho$. *then it exists a slowly varying function* ℓ *at infinity such that*

$$\forall x > 0,\ f(x) = x^\rho \ell(x). \tag{1.13}$$

Mathematically, the two most important results about functions in RV_0 are given in the following theorem due to Karamata.

Theorem 1.4

(A) ***(Uniform convergence theorem)*** *if* $f \in RV_\rho$, *then relation* (1.12) *holds uniformly for* $t \in [a, b]$ *with* $0 < a < b < \infty$.

(B) ***(Representation theorem)*** *If* $f \in RV_\rho$, *there exist measurable functions* $a : \mathbb{R}^+ \to \mathbb{R}$ *and* $c : \mathbb{R}^+ \to \mathbb{R}$ *with*

$$\lim_{x\to\infty} c(x) = c_0\ (0 < c_0 < \infty) \ \text{ and } \ \lim_{x\to\infty} a(x) = \rho \tag{1.14}$$

and $x_0 \in \mathbb{R}^+$ *such that for* $x \geq x_0$,

$$f(x) = c(x) \exp\left(\int_{x_0}^{x} \frac{a(t)}{t} dt \right). \tag{1.15}$$

Conversely, if (1.15) *holds with a and c satisfying* (1.14), *then* $f \in RV_\rho$.

For the proof of theorem 1.4, one can refer to [45] (theorem B.1.4 and theorem B.1.6).

Remark 1.2

1. One can choose arbitrarily $x_0 \in [0,\infty]$ in expression (1.15) by choosing suitably the functions $c(t)$ and $a(t)$ on the interval $[0,x_0]$.

2. The functions $c(t)$ and $a(t)$ given in (1.15) are not uniquely determined. (for more details see [45]).

Theorem 1.5 ***(Karamata's theorem)*** *Suppose $f \in RV_\rho$. There exists $x_0 > 0$ such that $f(x)$ is positive and locally bounded for $x \geq x_0$. if $\rho \geq -1$ then*

$$\lim_{x\to\infty} \frac{xf(x)}{\int_{x_0}^{x} f(t)dt} = \rho + 1. \tag{1.16}$$

If $\rho < -1$, or $\rho = -1$ and $\int_0^\infty f(s)ds < \infty$, then

$$\lim_{x\to\infty} \frac{xf(x)}{\int_{x}^{\infty} f(t)dt} = -\rho - 1. \tag{1.17}$$

Conversely, if (1.16) holds with $-1 < \rho < \infty$, then $f \in RV_\rho$; if (1.17) holds with $\infty < \rho < -1$, then $f \in RV_\rho$.

For the proof see [45] (theorem B.1.5, p.364). Readers can consult [44], [45], [25] and [107] for more information on regular varying functions.

1.4.2 Fréchet Attraction Domain

Denote by $\bar{F}(.) = 1 - F(.)$ the survival function, the generalized inverse of F is defined by

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

$Q(.)$ is called also the quantile function of the distribution function F. The quantity $x_p = F^{\leftarrow}(p)$ define the $p - quantile$ of F; an other useful function extracted from the quantile function and which plays a role in extreme value theory comparable to the role of the characteristic function in the theory of the stable distributions and their domains of attraction is the so called the tail quantile function, given by

$$U(x) = Q(1 - 1/x)\, x \geq 1,$$

that is the function $U(.)$ is the inverse function of $1/(1-F)$

Proofs of the following results can be found in [45] (Theorem 1.2.1 (part 1), p.19) and in [128] (Proposition 1.13, p.59).

Theorem 1.6 *The distribution function F is in the domain of attraction of the extreme value distribution of Fréchet G_γ, $\gamma > 0$ ($AD(H_\gamma)$) if and only if its endpoint x_F is infinite and*

$$\lim_{t\to\infty} \frac{1-F(tx)}{1-F(t)} = x^{-1/\gamma} \tag{1.18}$$

for all $x > 0$. This means that the survival function $\bar{F}$ is regularly varying at infinity with index $-1/\gamma$. In this case,

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

holds for $x > 0$ with normalization sequences

$$a_n = U(n) \quad \text{and} \quad b_n = 0.$$

Remark 1.3 *(1) According to Proposition (1.4), F belongs to the domain $AD(H_\gamma), \gamma > 0$ if and only if $x_F = \infty$ and $\bar{F}(x) = x^{-1/\gamma}\ell_F(x)$ where ℓ_F is a slowly varying function at infinity.*

(2) (1.18) can be reformulated in terms of the quantile function Q and the tail quantile function U respectively as following
$Q(1-.)$ is a regular varying function of index $-\gamma$ at 0, that is $Q(1-s) = s^{-\gamma}\ell(1/s)$, with $\ell \in RV_0$.
$U(.)$ is a regular varying function of index γ at ∞.

(3) Theorem (1.4) (part (B)) and (1.4) give the next representation

$$\bar{F}(x) = c(x)x^{-1/\gamma}\exp\left(\int_{x_0}^{x} \frac{a(t)}{t}dt\right),\ x < x_F \tag{1.19}$$

with

$$\lim_{x\to\infty} c(x) = c_0\ (0 < c_0 < \infty) \quad \text{and} \quad \lim_{x\to\infty} a(x) = 0$$

1.4.3 Weibull Attraction Domain

Embrechts [61] (Chapter 2) states that there is a full equivalence between the cases $\gamma > 0$ and $\gamma < 0$ by a simple transformation. In this subsection we use the same references as the previous one, so readers can refer to them for more details.

Theorem 1.7 *The distribution function F is in the domain of attraction of the extreme value distribution of Weibull G_γ, $\gamma < 0$ ($AD(H_\gamma)$) if and only if its endpoint x_F is finite and*

$$\lim_{t\downarrow 0} \frac{1-F(x_F - tx)}{1-F(x_F - t)} = x^{-1/\gamma} \tag{1.20}$$

for all $x > 0$. This means that the function $1 - \bar{F}(x_F - \cdot)$ is regularly varying at zero with index $-1/\gamma$. in this case,

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

holds for $x < 0$ with normalization sequences

$$a_n = x_F - U(n) \quad \text{and} \quad b_n = x_F.$$

Remark 1.4 *(1) According to Proposition (1.4), F belongs to the domain $AD(H_\gamma), \gamma < 0$ if and only if $x_F < \infty$ and*

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

where ℓ is a slowly varying function at infinity.

(2) (1.18) can be reformulated in terms of the quantile function Q and the tail quantile function U respectively as following
$Q(1 - .)$ can be written as

$$Q(1 - s) = x_F - s^{-\gamma}\ell(1/s), \ell \in RV_0. \tag{1.21}$$

$U(.)$ is a regular varying function of index γ at ∞.

1.4.4 Gumbel Attraction Domain

This case, often called extremal type I, is more different than the two previous ones; its characterization problem is more complex than the other two cases. In this part, we present the solution to this problem given by the pioneering thesis of de Haan [42].

Theorem 1.8 *The distribution function F is in the domain of attraction of the extreme value distribution of Gumbel G_γ, $\gamma = 0$ $(AD(H_\gamma))$ if and only if its endpoint x_F can be finite or infinite and*

$$\lim_{t\uparrow x_F} \frac{1 - F(t + xf(t))}{1 - F(t)} = \exp(-x) \tag{1.22}$$

for all real x, where f is a suitable positive function (called auxiliary function). If (1.22) holds for some f, then $\int_x^{x_F}(1 - F(t))dt < \infty$ for $x < x_F$ and (1.22) holds with

$$f(x) := \frac{\int_x^{x_F}(1 - F(t))dt}{1 - F(x)}. \tag{1.23}$$

in this case,

$$\lim_{n\to\infty} F^n(a_n x + b_n) = \exp(-\exp(-x))$$

holds for all x with

$$a_n = f(U(n)) \ \ and \ \ b_n = U(n). \tag{1.24}$$

Further and general characterization of attraction domains are given in the next subsection.

1.4.5 General characterization of Attraction Domains

De Haan in [45] has reformulated (1.6), (1.7) and (1.8) in a seemingly more uniform way (see [45], Theorem 1.2.5, p.21)

Theorem 1.9 *The distribution function F is in the domain of attraction of the extreme value distribution G_γ if and only if for some positive function f,*

$$\lim_{t\uparrow x_F} \frac{1-F(t+xf(t))}{1-F(t)} = (1+\gamma x)^{-1/\gamma} \tag{1.25}$$

for all x with $1+\gamma x > 0$. If (1.25) holds for some $f > 0$, then it also holds with

$$f(x) = \begin{cases} \gamma x & if \ \ \gamma > 0, \\ -\gamma(x_F - x) & if \ \ \gamma < 0, \\ \int_x^{x_F} (1-F(t))dt/(1-F(x)) & if \ \ \gamma = 0 \end{cases}$$

Furthermore, any f for which (1.25) holds satisfies

$$\begin{cases} \lim_{x\to\infty} f(x)/x = \gamma & if \ \ \gamma > 0, \\ \lim_{x\uparrow x_F} f(x)/(x_F - x) = -\gamma & if \ \ \gamma < 0, \\ f(x) \sim f_1(x) \ where \ f_1(x). is \ some \ function & \\ for \ which \ f_1'(x) \to 0, \ x \uparrow x_F & if \ \ \gamma = 0 \end{cases}$$

A useful representation of survival function, according to index γ values, is given by the following Theorem.

Theorem 1.10 *The distribution function F is in $AD(G_\gamma)$ if and only if there exist positive functions c and f, f continuous, such that for all $x \in (x_0, x_F)$, $x_0 < x_F$,*

$$\bar{F}(x) = c(x)\exp\left\{-\int_{x_0}^{x} \frac{dt}{f(t)}\right\}$$

with

$$\lim_{x\uparrow x_F} c(x) = c \in (0,\infty)$$

and

$$\begin{cases} \lim\limits_{x\to\infty} f(x)/x = \gamma & \text{if } \gamma > 0, \\ \lim\limits_{x\uparrow x_F} f(x)/(x_F - x) = -\gamma & \text{if } \gamma < 0, \\ \lim\limits_{x\uparrow x_F} f^{'}(x) = 0 \text{ and } \lim\limits_{x\uparrow x_F} f(x) = 0 \text{ if } x_F < \infty & \text{if } \gamma = 0. \end{cases}$$

Remark 1.5 *The auxiliary functions f in Theorems (1.9) and (1.10) are asymptotically the same. One can take $f(x) = (1 - F(x))/F^{'}(x)$, if the following condition (called von Mises' condition) is satisfied for $\gamma = 0$*

$$\lim_{x\to\infty} \left(\frac{\bar{F}}{F^{'}}\right)^{'} = 0$$

The next table shows some distributions and its attraction domains.

Attraction domains	**gumbel** $\gamma = 0$	**Fréchet** $\gamma > 0$	**Weibull** $\gamma < 0$
Distributions	Gaussian	Cauchy	
	Exponential	Pareto	
	Log-normal	Student	Uniform
	Gamma	Burr	Beta
	Weibull	Chi-square	
		Fréchet	

Table 1.1: Examples of some distributions classified according to its attraction domains.

1.5 Estimation of Extreme quantiles and Tail Index (EVI) without censoring

As we saw, the extreme value index (or tail index) γ is a real quantity key in the domain of extreme value analysis. In this section , we consider the estimation of this parameter which gives information on tail form of extreme value distribution. This problem has been studied in great details in the literature, we can cite Hill ([90]) in case of positive index, which has been thoroughly studied in the literature and several generalizations have been proposed. Pickands ([123] in the same year

proposed an estimator of extreme values index in the general case. Dekkers et al. [49]) generalized the Hill estimator by the so called moment estimator. Later, Beirlant et al. [19] used Hill estimator and quantile function to construct a new tail index estimator. Works on the estimation of the tail index continues to develop in the semi-parametric frame. For a recent review of estimation procedures for the Extreme Value (or tail) index of a distribution see Gomes and Guillou [80]. Most of these estimators are based on the k-upper ordered statistics $X_{n-k,n} \leq ... \leq X_{n,n}$.

1.5.1 Hill estimator

The most famous estimator of $\gamma > 0$ is the Hill estimator. Hill used Maximum Likelihood (ML) method on the set of k_n upper observations of a sample. A large number of theoretical works have been devoted to the study of the properties of Hill's estimator. Mason in([110]) demonstrated the weak consistency and Deheuvels, Haeusler and Mason established the strong consistency in [46]. The asymptotic normality is due to Davis and Resnick ([39]), Csörgö and Mason ([37]), Haeusler and Teugels ([87]) and Smith ([136]).

Definition 1.11 ***(Hill's estimator)***
Let $X_1, ..., X_n$ a sequence of iid random variables with common distribution function $F \in (H_\gamma)$, where $\gamma > 0$. Let $k = k_n$ such that $0 < k < n$ and $k \to \infty$, Hill's estimator is defined by

$$\hat{\gamma}^H_{(k,n)} = \frac{1}{k-1} \sum_{i=1}^{k-1} \log X_{n-i+1,n} - \log X_{n-k+1,n}. \tag{1.26}$$

Theorem 1.11 *Asymptotic behaviors of $\hat{\gamma}^H$ Assume that $F \in (H_\gamma)$; γ 0, $k \to \infty$ and $k/n \to \infty$ as $n \to \infty$*

(a) ***Weak consistency:***
then $\hat{\gamma}^H_k$ converges ***in probability*** *to γ when $n \to \infty$,*

(b) ***Strong consistency:***
If in addition of above assumption, $k/\log\log n \to \infty$ as $n \to \infty$,
then $\hat{\gamma}^H_k$ converges almost surely to γ when $n \to \infty$,

(c) ***Asymptotic normality***

Assume that F satisfies second-order condition, i.e. for $x > 0$,

$$\lim_{t\to\infty} \frac{\frac{U(tx)}{U(t)} - x^{\gamma}}{A(t)} = x^{\gamma} \frac{x^{\rho} - 1}{\rho}, \tag{1.27}$$

or equivalently,

$$\lim_{t\to\infty} \frac{\frac{1-F(tx)}{1-F(t)} - x^{-1/\gamma}}{A\left(\frac{1}{1-F(t)}\right)} = x^{-1/\gamma} \frac{x^{\rho/\gamma} - 1}{\rho\gamma}, \tag{1.28}$$

where $\gamma > 0$, $\rho \leq 0$, and A is a positive or negative function with

$$\lim_{t\to\infty} A(t) = 0$$

then

$$\sqrt{k}(\hat{\gamma}_k^H - \gamma) \xrightarrow{d} \mathcal{N}\left(\frac{\lambda}{1-\rho}, \gamma^2\right), \text{ as } n \to \infty$$

with $\mathcal{N}$ standard normal and

$$\lim_{n\to\infty} \sqrt{k} A\left(\frac{n}{k}\right) = \lambda$$

with λ finite.

Remark 1.6 *One of the interesting facts concerning (1.26) is that various asymptotically equivalent versions of $\hat{\gamma}_k^H$ can be derived through essentially different methods (such as the ML method or the mean excess function approach), showing that the Hill estimator is very natural.*
This estimator is based on the assumption that the right tail function is heavy (Pareto type) for large x, that is, $1 - F(x) \sim cx^{-1/\gamma}$ at infinity, for some $\gamma > 0$ and $c > 0$. Hence, Hill's estimator is only applicable in case the EVI is known to be positive, that is, only in case the underlying distribution function presents a heavy tail.
The success of this estimator is due to the fact that it can be interpreted as an estimator of the slope of the Pareto quatile plot (see [19]).

1.5.2 Weissman estimator

Weissman [154] proposed the most famous estimator of the extreme quantile $q(\alpha_n)$ when only the k largest observations of a sample of size n are available, taking Hill's estimator of the shape parameter, Weissman estimator is given by

Definition 1.12 *(Weissman's estimator)*
The Weissman estimator, $\hat{q}^H$, is defined by

$$\hat{q}^H_{\alpha_n} = X_{n-k_n,n}\left(\frac{k_{n+1}}{(n+1)\alpha_n}\right)^{\hat{\gamma}^H_{k_n}}$$

The asymptotic properties of Weissman's estimator are discussed and a confidence interval has been constructed under certain conditions on the distribution function F, k_n and α_n in [61], [111], [68] and [109].
In practice, the choice of the threshold k_n poses problems. If we draw the Hill's diagram (see Figure (1.3)); the map of the function $k_n \mapsto \hat{\gamma}^H_k$, we observe an extreme volatility which makes it difficult to use this estimator in practice if there is no indication of the choice of k_n. Furthermore, this estimator is biased.

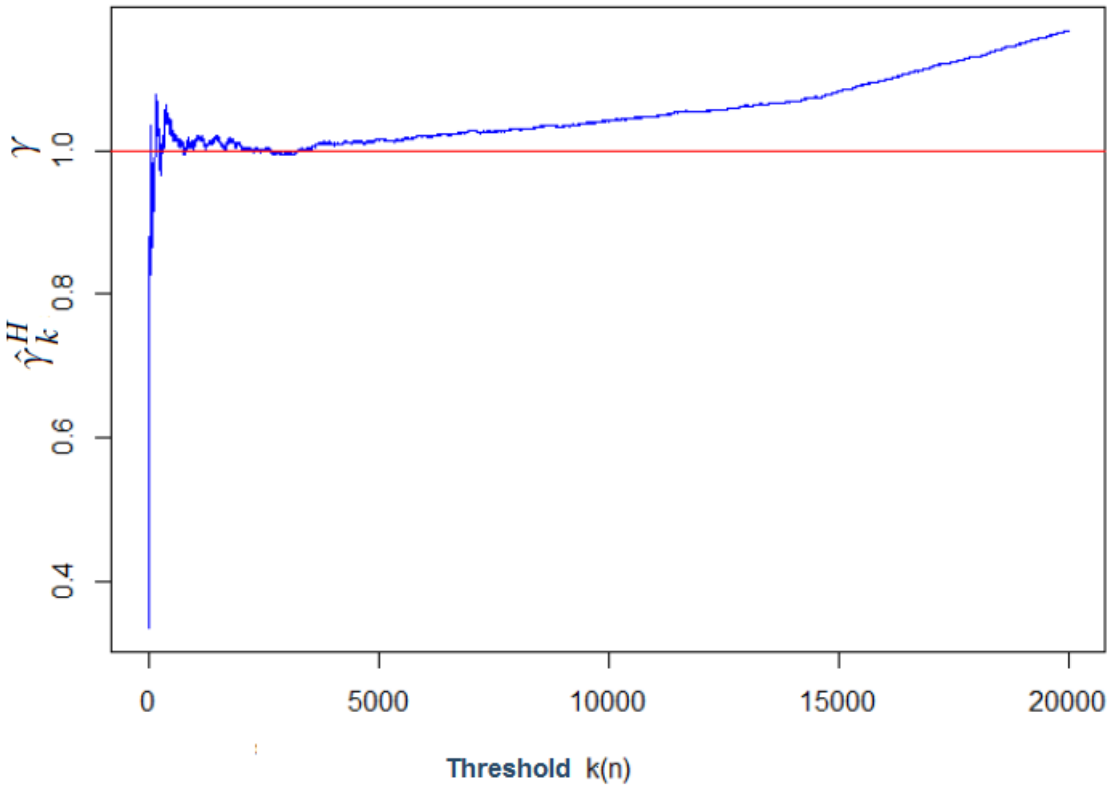

Figure 1.3: Graphical representation of Hill's estimator (observations derived from Cauchy distribution ($\in MDA(H_1)$) by simulation with $n = 40000$)

Comment

The graph above shows that if $k(n)$ is not negligible with respect to n, for large values of n, the Hill estimator does not converge.

1.5.3 Pickands estimator

The Pickands' estimator is the simplest and earliest estimator for extreme value index, it is constructed using three order statistics. This estimator is valid regardless of the attraction domain of the distribution, i.e. the extreme value index can takes any real value. Pickands ([123]) demonstrates the weak consistency of his estimator. Strong convergence as well as asymptotic normality have been demonstrated by Dekkers and de Haan ([48]).

Definition 1.13 *Let $(X_n, n \geq 1)$ to be a sequence of independent random variables having the same distribution function $F \in AD(H_\gamma)$, where $\gamma \in \mathbb{R}$. Let $k = k(n)$ a sequence of integers such that $1 < k < n$. Pickands' estimator is given by:*

$$\hat{\gamma}^P_{k(n)} = \log 2^{-1} \log\left(\frac{X_{n-k(n)+1,n} - X_{n-2k(n)+1,n}}{X_{n-2k(n)+1,n} - X_{n-4k(n)+1,n}}\right).$$

The next theorem summarize asymptotic properties of pickands' estimator.

Theorem 1.12 *Asymptotic properties of $\hat{\gamma}^P_{k(n)}$ Assume that $F \in AD(H_\gamma)$, where $\gamma \in \mathbb{R}$. Let $k = k(n)$ a sequence of integers such that $1 < k < [n/4]$, $k(n) \to \infty$ and $k(n)/n \to 0$ as $n \to \infty$.*

(a) Weak consistency
*then $\hat{\gamma}^P$ converges **in probability** to γ when $n \to \infty$,*

(b) Strong consistency
If in addition of above assumption, $k/\log\log n \to \infty$ as $n \to \infty$,
then $\hat{\gamma}^P_k$ converges almost surely to γ when $n \to \infty$,

(c) Asymptotic normality
Suppose that the tail function U associated to F satisfies the second-order condition, i.e.

$$\lim_{x \uparrow x_F} \frac{\frac{1 - F(x + tf(x))}{1 - F(x)} - Q_\gamma(t)}{\alpha(x)} = (Q_\gamma(t))^{1+\gamma} H_{\gamma,\rho}(Q_\gamma^{-1}(t)), \tag{1.29}$$

with $Q_\gamma(t) := (1+\gamma t)^{-1/\gamma}$, f some positive function, and α some positive or negative function with

$$\lim_{x \uparrow x_F} \alpha(x) = 0.$$

Recall the equivalent relation in terms of $U := (1/(1-F))^{\leftarrow}$*:*

$$\lim_{x\to\infty} \frac{\frac{U(tx)}{U(x)} - D_\gamma(t)}{A(x)} = H_{\gamma,\rho}(Q_\gamma(t) := \int_1^t s^{\gamma-1} \int_1^s u^{\rho-1} du ds, \tag{1.30}$$

for all $x > 0$ *with* $D_\gamma(t) = (t^{\gamma-1})/\gamma$, $a(x) = f(U(x))$, *and* $A(x) = \alpha(U(x))$.
Then, for

$$\lim_{n\to\infty} \sqrt{k} A\left(\frac{n}{k}\right) = \lambda$$

with λ *finite.*

$$\sqrt{k}\left(\hat{\gamma}_k^P - \gamma\right) \xrightarrow{d} \mathcal{N}\left(\lambda b_{\gamma,\rho}, \sigma_P^2(\gamma)\right), \text{ as } n \to \infty$$

with $\mathcal{N}$ *standard normal, where*

$$b_{\gamma,\rho} := \begin{cases} \dfrac{4^{-\rho}\gamma((4^{\gamma+\rho}-1)-(2^\gamma+1)(2^{\gamma+\rho}-1))}{\rho 2^\gamma(\rho+\gamma)(2^\gamma-1)\log 2}, & \rho < 0 \text{ and } \gamma \neq 0, \\ \dfrac{1-2^{-\rho+1}+4^{-\rho}}{\rho^2(\log 2)^2} & \rho < 0 \text{ and } \gamma = 0, \\ 1 \quad \rho = 0. \end{cases}$$

and

$$\sigma_P^2(\gamma) := \begin{cases} \dfrac{\gamma^2(2^{2\gamma+1}+1)}{4(\log 2)^2(2^\gamma-1)^2}, & \gamma \neq 0, \\ \dfrac{3}{4(\log 2)^4} & \gamma = 0. \end{cases}$$

For more details, see de Haan and Ferreira ([45]) p. 85-86. A more formal explication for the Pickands' estimator is provided in [61].

Remark 1.7 1. *A good property of Pickands estimator is that it is invariant by shifting the sample by a constant.*

2. *Using asymptotic normality of Pickands estimator in Theorem* (1.12) *when* $\sqrt{k(n)}A(n/k(n) \to 0$ *as* $n \to \infty$,

$$(\sqrt{k}/\sigma_P(\gamma))(\hat{\gamma}_k^P - \gamma) \xrightarrow{d} \mathcal{N}(0,1),$$

hence one can construct in this case a confidence interval of γ.

Some authors suggested improvements and generalization of Pickands estimator, we can cite:
Falk [64] who took a linear combination of two different numbers of observations

treated as the tail.

Drees [52] extended the Falk's refinement, proposed an other refinement in [53] and gave a generalized form in [54] and [55].

Yun [160] introduced a full generalization of the Pickands estimator encompassing estimators of Fraga Alves [2] and Yun [159] which were less general.

Segers [133] generalized it in a way that includes all of its previously known variants and proved by explicit formulas these this estimators have the same asymptotic variance as the maximum likelihood estimator. For more details on basic ideas and analytic expressions of many tail index estimators, readers can see [65].

Figure (1.4) illustrate Pickands' estimator with 95 % confidence interval

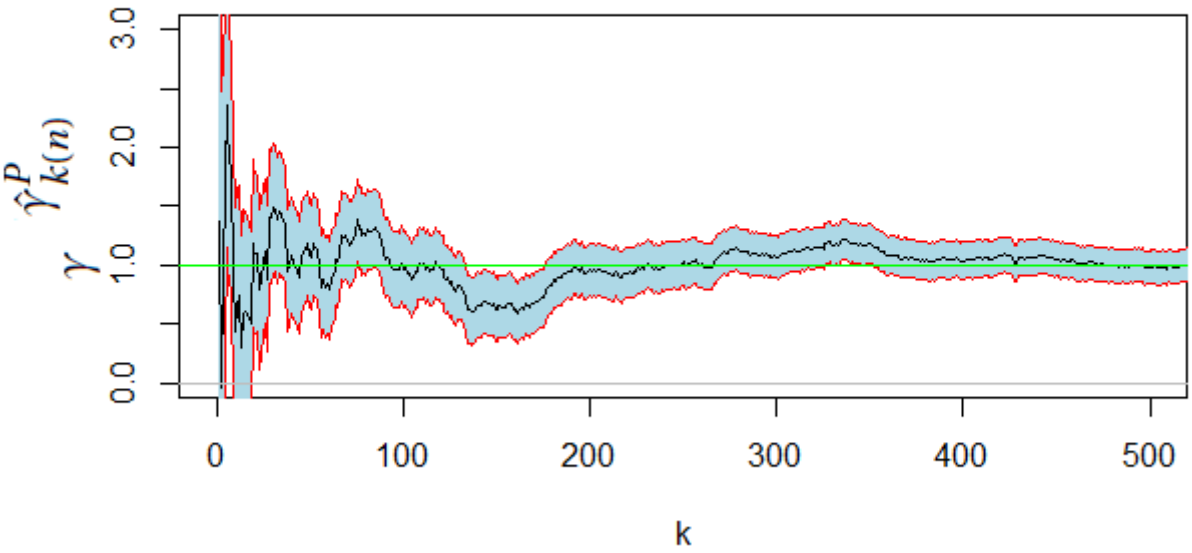

Figure 1.4: Graphical representation of Pickands's estimator (observations derived from Cauchy distribution$\in MDA(H_1)$ by simulation with $n = 40000$)

- For small k, there are large oscillations with a large confidence interval.

- For large k, we have a confidence interval that is narrower but not centered on the true value.

Definition 1.14 ***(Quantile estimator $\hat{Q}^P$)*** *The estimator $\hat{Q}^P$ of the quantile $Q(1-s)$ associated with the Pickands estimator is:*

$$\hat{Q}^P := X_{n-k+1,n} + \frac{(k/ns)^{\hat{\gamma}^P} - 1}{1 - 2^{-\hat{\gamma}^P}}(X_{n-k+1,n} - X_{n-2k+1,n}) \tag{1.31}$$

The asymptotic properties of the estimator 1.31, are discussed by Dekkers and de Haan [48] (Theorem 3.3, p.1809), see also Matthys and Beirlant [111].

1.5.4 Moment estimator

The moment estimator is similar to the Hill estimator but it can be used for general $\gamma \in \mathbb{R}$, not only for $\gamma > 0$, this extension of Hill estimator is due to Dekkers et al. [49] by applying the Hill estimator for the case $\gamma \leq 0$ but the problem is that $U(\infty) \leq 0$ is possible, in which case the logarithm of observations is not defined. This problem is overcame by assuming that $U(\infty) > 0$ by shifting the data and being aware that this shift influences the behavior of the estimator, see [45], Subsection 3.5

Definition 1.15 ***(Moment estimator)***
For $\gamma \in \mathbb{R}$, moment estimator is given by

$$\hat{\gamma}^M(k) := M_n^{(1)} + T_n := M_n^{(1)} + 1 - \frac{1}{2}\left(1 - \frac{\left(M_n^{(1)}\right)^2}{M_n^{(2)}}\right)^{-1} \tag{1.32}$$

where

$$M_n^{(r)}(k) = \frac{1}{k}\sum_{i=1}^{k}(\log X_{n-i+1,n} - \log X_{n-k,n})^r,\ r = 1, 2. \tag{1.33}$$

Note that $M_k^{(1)}$ corresponds to Hill estimator $\hat{\gamma}(k)^H$. Moment estimator is also known as Dekkers-Einmahl-de Haan estimator.

Asymptotic properties of moment estimator have been established by Dekkers at al.[49] (Theorem 3.1 and Corollary 3.2) .

Theorem 1.13 ***(Asymptotic properties of $\hat{\gamma}^M$)*** *Assume that $F \in AD(H_\gamma)$, where $\gamma \in \mathbb{R}$. Let $k = k(n)$ a sequence of integers such that $k(n) \to \infty$ and $k(n)/n \to 0$ as $n \to \infty$.*

(a) Weak consistency
*then $\hat{\gamma}^M$ converges **in probability** to γ when $n \to \infty$,*

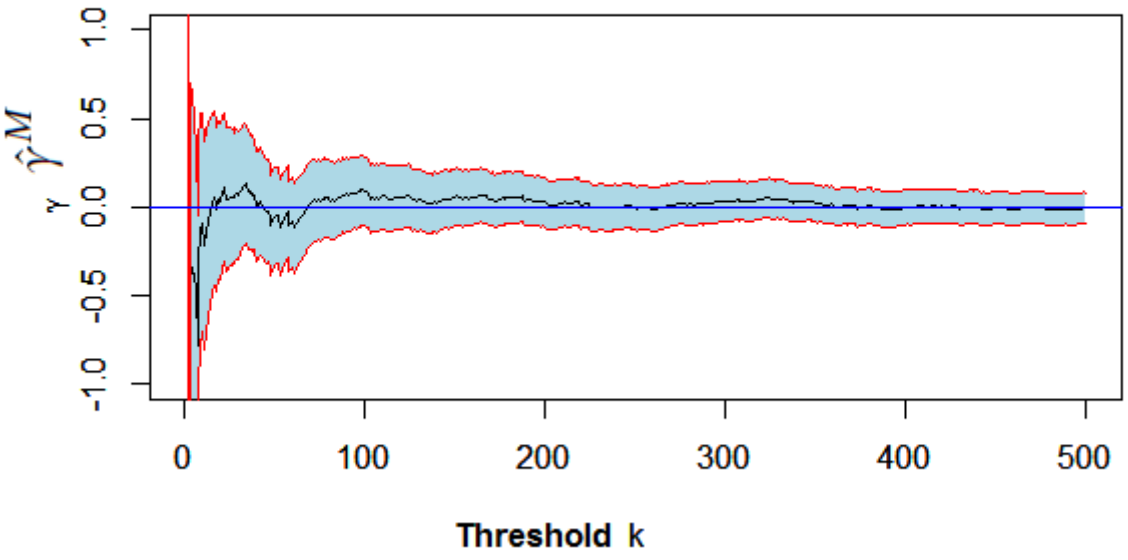

Figure 1.5: Graphical representation of Moment estimator (observations derived from Gumbel distribution$\in MDA(H_0)$ by simulation with $n = 100000$)

(b) *Strong consistency*

If in addition of above assumption, $k/(\log n)^\delta \to \infty$ *as* $n \to \infty$, *for* $\delta > 0$, *then* $\hat{\gamma}_k^M$ *converges almost surely to* γ *when* $n \to \infty$,

(c) *Asymptotic normality*

Under some conditions on distribution function F (see [49]),

$$\sqrt{k}(\hat{\gamma}_k^M - \gamma) \xrightarrow{d} \mathcal{N}(0, \sigma_M^2(\gamma)), \text{ as } n \to \infty$$

with $\mathcal{N}$ *standard normal, where*

$$\sigma_M^2(\gamma) := \begin{cases} 1 + \gamma^2, & \gamma \geq 0, \\ (1-\gamma)^2(1-2\gamma)\left[4 - 8\dfrac{1-2\gamma}{1-3\gamma)} + \dfrac{(5-11\gamma)(1-2\gamma)}{(1-3\gamma)(1-4\gamma)}\right] & \gamma < 0, \end{cases}$$

Remark 1.8 *The name of this estimator stems from the fact that* $M_k^{(1)}$ *and* $M_k^{(2)}$ *can be interpreted as empirical moments.*
The moment estimator has an excellent performance in general, however when it is applied to a full data set of exceedances, so one obtains irregular estimates.

Definition 1.16 ***(Quantile estimator $\hat{Q}^M$)***

Extreme quantile estimate basing on moment estimator is given by

$$\hat{Q}^M := X_{n-k+1,n} + \hat{a}_n^M \frac{(k/ns)^{\hat{\gamma}^M} - 1}{\hat{\gamma}^M}, \; for\, k < n, \tag{1.35}$$

with

$$\hat{a}_n^M = \frac{M_n^{(r)}}{\rho_1(\hat{\gamma}^M)} X_{n-k,n}, \;\; \rho_1(\hat{\gamma}) = \begin{cases} 1, & \gamma \geq 0, \\ \dfrac{1}{1-\gamma} & \gamma < 0, \end{cases}$$

1.5.5 Choice of Order Statistics number (k)

As we have seen, extreme value index estimators depend basically on the empirical tail size, so one must choose attentively the number k used in the implementation of these estimators which is an important problem. The choice of this number is clearly a question of trade-off between bias and variance: as k increases, the bias will grow because the tail satisfies less the convergence criterion, while if less data are used, the variance increases. The optimal value of k, then should minimize the mean-square error (the sum of the bias-squared and the variance). Theoretically, the optimal value of k depends on both the sample size and the unknown values of γ *and* ρ (see Hall Welsh [89]. In order to overcome this problem, some authors have chosen k graphically as the stable point in $(\hat{\gamma}, k)plot$ (see Drees et al. [57]), but this method is subjective to the practice as a guessing. Other graphical methods have been proposed by others; we can cite [88]; Draisma et al [51]; Danielsson et al. [38], Gomes Oliveira [82] and Beirlant et al. [18]; [11]. Some others authors' selection methods are based on the bias; see Drees Kaufmann ([56]) and Guillou Hall ([86]), Gomes Pestana ([94]), Gomes et al. ([79]) and Beirlant et al. ([12]).

Chapter 2

Modeling of conditional extreme values under censoring: A Review.

Abstract

In practice estimation of parameters depends on several factors or covariables for the reason that data under precise factor are homogeneous and statistics present less variability. In the presence of covariate information, it is interesting to include it in the estimation by modeling the parameters of extreme value distribution as a function of the covariate(s). Several works fall in this direction, for example, Davison and Smith (1990) fitted a Generalised Pareto (GP) distribution with parameters taken as an exponential function of the covariates; Gardes and Girard (2008) used moving-window methodology, then Gardes et al. [72] and Lekina [105] proposed conditional extreme quantiles estimators in the non-parametric framework; Beirlant and Goegebeur (2003) and Wang and Tsai (2009) used a conditional exponential regression model, and Beirlant Goegebeur (2004) employed repeated fitting of local polynomial maximum likelihood estimation.

2.1 Estimating conditional extreme values index and quantiles

we restrict ourselves to fix design, that is covariable is non random. Let Y a real random variable measured jointly with a non random covariable X. Given a sample $\{(x_i, Y_i),\ i = 1, ..., n\}$ of independent and identically distributed observations. Assume that the conditional distribution F of Y is heavy tailed, i.e.

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

where $\ell \in RV_0$ *and* $\gamma(.)$ is the conditional extreme value index (or conditional tail index). We will use the selection method proposed in [105] and [114] (see Chapter 4), Let Z_i^x, $i = 1,...,m_n^x$ be the response variables $Y_i^{'}s$ for which the associated covariate $X = x$ and let $Z_{(1)}^x \leq ...Z_{(m_n^x)}^x$ be the corresponding order statistics and m_n^x is the number of observations having $X = x$

A family of conditional tail index estimators is introduced in [71], they are based on weighted sum of the log spacings between the k_x largest order statistics $Z_{(m_n^x-k_x+1)},...,Z_{(m_n^x)}^x$. As in [72], this family is defined by

$$\hat{\gamma}_n(x,W) = \sum_{i=1}^{k_x} i \log\left(\frac{Z_{(m_n^x-i+1)}^x}{Z_{(m_n^x-i)}^x}\right) W(i/k_x,x) \Big/ \sum_{i=1}^{k_x} W(i/k_x,x), \tag{2.1}$$

where $W(.,x)$ is a weighted function defined on (0,1) and integrating to 1. Under some conditions on the weight function, L. Gardes and S. Girard ([71], Theorem 2) established asymptotic normality of $\hat{\gamma}_n(x,W)$ as following:

$$k_x^{1/2}(\hat{\gamma}_n(x,W) - \gamma(x)) \xrightarrow{D} \mathcal{N}(0, \gamma^2(x)V(x,W)),$$

where

$$V(x,W) = \int_0^1 W^2(s,x)ds.$$

Definition 2.1 ***(conditional extreme quantile estimator of*** $q(\alpha_{m_n^x},x)$***)***
Using (2.1) *and considering* $\beta_{m_n^x} = k_x/m_n^x$*, the conditional extreme quantile estimator of order* $\alpha_{m_n^x}$ *is defined by*

(C1)

$$\hat{q}_1(\alpha_{m_n^x},x) = Z_{(m_n^x-[m_n^x\alpha_{m_n^x}])}^x$$

if $\alpha_{m_n^x}$ *converges slowly to* 0*; i.e.* $\alpha_{m_n^x} \to 0$ *and* $m_n^x\alpha_{m_n^x} \to \infty$ *as* $m_n^x \to \infty$*. ([t] denotes the largest integer smaller than t)*

(C2)

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

if $\alpha_{m_n^x}$ *converges quickly to* 0*; i.e.* $\alpha_{m_n^x} \to 0$ *and* $m_n^x\alpha_{m_n^x} \to c \in [0,1]$ *as* $m_n^x \to \infty$*. where* $\beta_{m_n^x}$ *converges slowly to* 0 *and* $\hat{\gamma}_n(x)$ *is a conditional extreme value index estimator.*

L. Gardes in [72](Theorem 1, p.421) established the limit distribution of conditional extreme quantile, especially the estimator is asymptotically Gaussian with asymptotic variance proportional to $\gamma^2(x)/(m_n^x \alpha_{m_n^x})$ as shown below:

Theorem 2.1 ***(Theorem 1 in [72])***
Let $\left(\alpha_{m_n^x}\right)_n$ *a sequence satisfying (C1).*
If $(m_n^x \alpha_{m_n^x})^2 \omega_n(m_n^{x-(1+\delta)}) \to 0$ *for some* $\delta > 0$ *then*

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

$\omega_n(a)$ is the largest oscillation of the log-quantile function with respect to its second variable, defined for all $a \in (0, 1/2)$ as

$$\omega_n(a) = sup \left\{ \left| \log \frac{q(\alpha, x)}{q(\alpha, x')}, \alpha \in (a, 1-a), (x, x') \in B(x, h_x)^2 \right| \right\}$$

with $B(x,)$ is the ball centered at point x and with radius h_x.

2.2 Survival Analysis language

2.2.1 Introduction

Survival analysis is the study of survival times and of the factors that influence them, that is analyze time to event data. Survival times can be observed in several disciplines such that medicine, reliability, insurance, economy, biology, ...etc. Example of such times include time from birth until death, time from entry into a clinical trial until death or disease progression, life-time of a component, ...etc. The important difference between survival analysis and other statistical analyses is the presence of censoring or incomplete data. This really leads the survival function to be more important in setting up the models.
We begin with a reminder of some definitions.

2.2.2 Survival, Hazard and Risk functions

Y denotes the positive random variable representing time to event of interest. Cumulative distribution function is $F(y) = P(Y \leq y)$ with probability density function $f(y) = F'(y)$.

Definition 2.2 ***Survival function***

The survival function, also called tail distribution, denoted by $S(y)$ or $\bar{F}(y)$, is defined on $\mathbb{R}_+$ by

$$S(y) = 1 - F(y) := P(Y > y). \tag{2.2}$$

which is the probability of an individual surviving to time y;

In the context of equipment or manufactured item failures, $S(y)$ is referred to as the reliability function. If X is a continuous random variable, then $S(y)$ is a continuous and strictly decreasing function.
When Y is a continuous random variable, the survival function is the integral of the probability density function, $f(x)$.

Definition 2.3 ***(empirical survival function).***

Let $Y_1, ..., Y_n$ a sample of size $n \geq 1$ of a positive random variable Y. The empirical survival function $\bar{F}_n$ is given by

$$\bar{F}_n(y) = 1 - F_n(y) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{Y_i > y\}}, \quad \forall y \geq 0; \tag{2.3}$$

where $\mathbb{1}_{\{A\}}$ denotes the indicator function of the set A.

A basic quantity, fundamental in survival analysis, is *the hazard function*. This function is also known as the conditional failure rate in reliability, the force of mortality in demography, the intensity function in stochastic processes, the age-specific failure rate in epidemiology, the inverse of the Mill's ratio in economics, or simply as the hazard rate.

Definition 2.4 ***(Hazard rate function)***

The hazard rate is defined by

$$h(y) = \lim_{\delta y \to 0} \frac{P[y \leq Y < y + \delta y | Y \geq y]}{\delta y} \tag{2.4}$$

If Y is continuous random variable, then

$$h(y) = f(y)/S(y) = -d \ln[S(y)]/dy. \tag{2.5}$$

A related quantity is the **cumulative hazard function** $\Lambda(y)$, defined by

$$\Lambda(y) = \int_0^y h(x)dx = -\ln[S(y).] \tag{2.6}$$

Conversely

$$S(y) = \exp[-\Lambda(y)] = \exp\left[-\int_0^y h(x)dx\right] \tag{2.7}$$

From (2.4), $h(y)\delta y$ can be viewed as the approximate probability of an individual of age y undergoing the event in the next instant. This function is very useful in determining the appropriate survival distribution using qualitative information about the mechanism of failure (hazard rate function) and for describing how the chance of the occurring of the event changes with time, for illustrative examples see Lee and Wang [103], Wienke [155] or Klein Moeschberger [100].

2.2.3 Censoring and Truncation

Time-to-event data present a special problem in analyzing them, known as censoring, which occurs when there is not access to the whole information on lifetimes, that is some of these lifetimes occur only within certain intervals, whereas the remainder are known exactly. There are many types of censoring, such as right censoring, left censoring, and interval censoring.

Another problem which may be present with incomplete data is that of *truncation* under various categories such as left truncation, right truncation and interval truncation.

Censored Data

The interest of censoring:

The long experiment time is one of the main problems in classical life-time studies and to overcome this problem, different censoring schemes are proposed in the literature.

Censoring, arises when the starting or ending events are not precisely observed.

Definition 2.5 ***(Censoring variable)***

The censoring variable C is defined by the non-observation of studied event that is the time to the censoring event. If instead of observing Y, we observe C, and we know that $Y > C$ (respectively $Y < C$, $C1 < Y < C2$), we say that there is right censorship (respectively left censorship, interval censorship).

Intuitively, the right censoring, for example, results when the final endpoint is only known to exceed a particular value.

Formally, for a given individual, we consider:

- Y representing its survival time (time to event);
- C representing the time to a censoring event;
- Z is the time really observed.

Types of censoring

1. **Right censoring**
   Right censoring occurs when an individual leaves the study before an event occurs, or the study ends before the event has occurred. For example, we consider patients in a clinical trial to study the effect of treatments on certain disease occurrence. The study ends after 5 years. Those patients who have had no disease by the end of the years of study are censored, that is if the patient quits the study at time c, then the event occurs after c.

2. **Left censoring**
   The left censorship occurs when the individual has experienced the event before he was observed. We only know that the variable of interest is lower or equal to a known variable, but its exact event time is unknown. For example concerning reliability if one studies a certain electronic component which is connected in parallel with one or more other components: the system can continue to operate, although aberrantly, until this failure is detected (for example during a control or in case of system shutdown). Thus, the duration observed for this component is left censored. Often, if left censoring occurs in a study, right censoring may also occur, and the lifetimes are considered doubly censored (cf. [148])

3. **Double censoring**
   Double censoring occurs when both left censoring and right censoring are present. In addition some exact event times are observed. Many non-parametric models were suggested for the study of the double censoring. For example, the Turnbull model (see [148]) is the most used, and several researches are based on this model.

4. **Interval censoring**
   A more general type of censoring occurs when the lifetime is only known to

occur within an interval. In clinical trials, such interval censoring occurs when patients have periodic follow-up and the patient's event time is only known to belonging to an interval $(L, R]$ (L for left endpoint and R for right endpoint of the censoring interval). The same censoring type may also occur in industrial experiments when there is periodic inspection for proper functioning of equipment items. One advantage of this type of censoring is that it allows right or left censored data to be presented by intervals of the type $[c, 1[$and $[0, c]$ respectively, which means that interval censoring is a generalization of left and right censoring.

The above classes of censorship can be given according to the mode or mechanism of censorship. So Censoring may be classified into three types: Type I, Type II, and Type III or random.

- **Type I censoring (fixed)**
  In Type I censoring, the censoring times are pre-specified whereas the number of observed events is random. For example, in an animal experiment, a cohort of animals may start at a specific time, and all followed until a pre-specified ending time. Animals which have not experienced the event of interest before the end of the study are then censored at that time. Another example (discussed in detail in [113] Example 1.5, p.8) is a smoking cessation study, where by design each subject is followed until relapse (return to smoking) or 180 days, whichever comes first. Those individuals who did not relapse within the 180 day period were censored at that time.
  Formally, Let C be a fixed value. For example in right censorship, instead of observing the variables of interest $Y_1, ..., Y_n$, we observe Y_i only if it is less than or equal to C; else we observe C. One therefore observe a random variable Z_i such that $Z_i := \min(Y_i, C); i = 1, ..., n$.

- **Type II censoring (waiting)** Type II censoring occurs when the experimental individuals are followed until a fixed number of events among the individuals has occurred. Such a design is rare in biomedical studies, but may be used in industrial settings, where time to failure of a device is of primary interest. As an example (see [113]), where the study stops after, for instance, 25 out of 100 devices are observed to fail. The remaining 75 devices would then be censored. In this example, the smallest 25% of the ordered failure times are

observed, and the remainder are censored.

Formally, Let $Y_{(i)}$ and $Z_{(i)}$ be the order statistics of the variables Y_i and Z_i; the i^{th} individual survival time and the observed survival time respectively. Suppose that the censorship date $Y_{(k)}$ (until occurrence of k events) and we only observe the following variables:

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

- **Type III censoring (random)** This is the last general category of censoring; it is random censoring. One must be careful to the cause of the censoring in order to avoid biased survival estimates. For example, in biomedical field, one cause of random censoring is patient abandon. If the abandon occurs truly at random, and is unrelated to the disease process, such censoring may not cause any problems with bias in the analysis. But if patients who are near death are more likely to abandon than other patients, serious biases may arise in this situation. Another cause of random censoring is competing events, a good example is given in ([113], Example 1.4) in which patient dies of another cause different to the primary one, while that patient will be censored.

  Formally, Let $Y_1, ..., Y_n$ a sample of a positive random variable X, we say that there is a random censoring of this sample if there exists another positive random variable C of sample $C_1, ..., C_n$, in this case instead of observing the $Y_i^{'}s$; we observe a couple of random variables (Z_i, δ_i) with

$$Z_i := min(Y_i, C_i) \ \ and \ \ \delta_i := \mathbb{1}_{\{Y_i \leq C_i\}}, \ \ for\, i = 1, ..., n \tag{2.8}$$

  where δ is the censorship indicator; That is, δ is 0 or 1 according to whether Y is a censored time or an observed event time.

In the literature, several authors (see[62] ; [33] and [32]) introduced other types by combining the two ones censoring types; known as *hybrid censoring* types. Progressive censoring is a very flexible censoring scheme as it allows for the removal of live experimental units at various intermittent times during the experiment in addition to removal at the termination of the experiment, for more details one may refer to [8] and [7]. Note that hybrid censoring schemes have been introduced in the context

of progressive censoring as well. Ng et al. (2009, [118]) and Lin et al. (2009, [106]) have proposed adaptive progressive hybrid censoring schemes in order to allow the experimenter to modify the censoring scheme adaptively during the life-testing experiment. For more details, readers can see (2009, [6]).

In this thesis, we are only interested in the case of random right censorship.

Truncation

There is another type of incompleteness which is a second feature of many survival studies, sometimes confused with censoring, called "truncation". Truncation of survival data occurs when only those individuals whose event time lies within a certain observational window (C_L, C_R) are observed. An individual whose event time is not in this interval is not observed and no information in this case is available. Contrary to what we have seen concerning censoring where there is at least partial information on each individual. Since we are interested only to individuals with event times belongs to the observational window, the inference for truncated data is restricted to conditional estimation. There are three types of truncation: left, right and interval truncation.

- **Left truncation:**
  When C_R is infinite then we have left truncation. Here we only observe those individuals whose event time Y exceeds the truncation time C_L. That is we observe Y if and only if $C_L < Y$. A common example of left truncation is the problem of estimating the distribution of the diameters of microscopic particles. The only particles big enough to be seen based on the resolution of the microscope are observed and smaller particles do not come to the attention of the investigator.

- **Right truncation:**
  Right truncation is another form of length-biased sampling, but it is much more difficult to accommodate than left truncation. Right truncation occurs when C_L is equal to zero. That is, we observe the survival time Y only when $Y \leq C_R$. Right truncation arises, for example, in estimating the distribution of stars from the earth; in that stars too far away are not visible and are right truncated.

- **Interval truncation:**
  We say that we have interval truncation if the survival time is left and right truncated simultaneously.

Likelihood and Censoring

If the censoring mechanism is independent of the event process, then we have an easy way of dealing with it. Assume that Y is the time to event and that C is the time to the censoring event.
Suppose that all individuals may have an event or be censored, say for individual i one of a pair of observations (y_i, c_i) may be observed. Then since we observe the minimum time, we would have the following expression for the likelihood (using independence)

$$L = \prod_{y_i<c_i} f(y_i)S_C(y_i) \prod_{y_i>c_i} f(c_i)S_C(c_i)$$

Where S_C is survival function of the random variable C. Using notation above, for each individual we observe $z_i := min(y_i, c_i)$ and δ_i, as observations from a continuous random variable and a binary random variable. Hence L will be

$$L = \prod_i g(z_i)^{\delta_i} S(z_i) \prod_i g_C(z_i)^{1-\delta_i} S_C(z_i)$$

with g_C is the density function of C. where we have used

$$density = hazard \times survival\ function.$$

2.2.4 Estimation of survival function and cumulative hazard function

In the literature, several authors have been interested in estimating the survival function under censoring. Among these we can cite Kaplan and Meier ([97]) proposed an non-parametric estimator of the survival function. This estimator was generalized by Beran ([22]) in the conditional case. Some asymptotic properties of this generalized Kaplan-Meier estimator were presented by Gonzalez-Manteiga and Cadarso-Suarez ([84]).

Kaplan-Meier Estimator

Let $(Z_i, \delta_i)_{1\leq i\leq n}$ be the really observed sample and let $(Z_{(i)}, \delta_{(i)})_{1\leq i\leq n}$ its increasing order statistic. Kaplan-Meier estimator is defined by:

$$\begin{aligned}\bar{S}_n(t) = 1 - \bar{F}_n(t) &= \prod_{i=1}^{n}\left(\frac{n-i}{n-i+1}\right)^{\delta_{(i)}1_{\{Z_{(i)}\leq t\}}} \\ &= \prod_{i=1}^{n}\left(1 - \frac{\delta_{(i)}1_{\{Z_{(i)}\leq t\}}}{n-i+1}\right)\end{aligned} \tag{2.9}$$

It is also called "Product limit" due to the existence of product symbol in the formula of the estimator.

- Kaplan-Meier estimator is a step function having jumps only at uncensored observations.
- The jumps height of this estimator is random.
- If data is complete, we find the empirical distribution function.

Theorem below show that Kaplan-Meier estimator is asymptotically Gaussian and one can find the proof in [73] using the point processes theory.

Theorem 2.2 *(Droesbeke and Saporta (2011, [58]) If both of survival function* $(1-F)$ *and censorship distribution function* (G) *don't have any common discontinuity, then:*

$$\sup_{t\geq 1} \mid \hat{S}_n(t) - S(t) \mid \xrightarrow{a.s.} 0$$

and for every $t \geq 0$,

$$\sqrt{n}(\hat{S}_n(t) - S(t)) \mid \xrightarrow{D} W_t$$

where $(W_t)_{t\geq 0}$ *a mean zero Gaussian process satisfying for every* u *and* v *strictly positive*

$$Cov(W_u, W_v) = S(u)S(v)\int_0^{\min(u,v)} \frac{dF(s)}{(1-F(s))^2(1-G(s))}$$

In the survival analysis literature, many of the authors have been devoted to the study of the asymptotic properties Kaplan-Meier estimator. For example; Uniform consistency has been studied by Shorack and Wellner [134], Wang [153], Stute and Wang [143] and Gill [75]. The normality asymptotic has been studied by Breslow and Crowley [27], Gill [73] and [74].

In presence of covariates, Kaplan-Meier estimator has been generalized, taking the name Generalized Kaplan-Meier estimator or Beran estimator.

Generalized Kaplan-Meier Esimator (GKM)

Beran ([22]) proposed a local aspect to the Kaplan-Meier estimator using smoothing with Nadaraya-Watson weights. He was studying regression problems with incomplete data in a completely non-parametric framework. The estimator proposed is defined as follows:

$$1-\hat{F}_n(z/x)=\hat{S}_n^{GKM}(z/x)=\begin{cases}\prod_{i=1}^{n}\left[1-\frac{B_i(x)}{\sum_{j=1}^{n}\mathbb{1}_{\{Z_i\geq Z_j\}}B_j(x)}\right]^{\mathbb{1}_{\{Z_{(i)}\leq z,\,\delta_{(i)}=1\}}}, & if\ z<Z_{(n)},\\ 0 & else.\end{cases}$$

where

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

is the Nadaraya-Watson weights, $h_n\to 0$ the window and K the kernel.
In the following we will consider the metric space $(\mathcal{X},d)$. Let

$$H^u(z|x)=P(Z\leq z,\delta=1/X=x)$$

the conditional sub-distribution of the uncensored observations . In this case, conditional or generalized Kaplan-Meirer estimator has a complicated structure since it is the product of dependent factors. To avoid this problem in the case of the (unconditional) product-limit estimator, Lo and Singh (1986, [108]) furnished a representation as a sum of i.i.d. terms with some remainder term which is asymptotically negligible. For the conditional Kaplan-Meier estimator, decompositions similar to Lo and Singh (1986, [108]) were derived, all in the case where covariate is univariate, see e.g. Van Keilegom and Akritas (1999, [149]), Van Keilegom and Veraverberke (1997, [150]). In particular, Du and Akritas (2002, [59]) proposed an uniform i.i.d. representation that holds uniformly in terms of survival time and covariate .For asymptotic properties of this estimator, some assumptions will be assumed as in Van Keilegom and Veraverbeke (1998, [151]) and in Gonzalez-Manteiga and Cadarso- Suarez (1994, [84]).

- **(C1)** The functions $H(./x_i)$ and $H^u(./x_i)$, for $1\leq i\leq n$, belong to the families $(H(./x))_{x\in X}$ and $(H^u(./x))_{x\in X}$, continuous and differentiable with respect to the first variable and twice differentiable with respect to the second variable and that the derivatives are continuous.

- **(C2)** The kernel K is a symmetric density function and second order lipschitz function with bounded support satisfying:

$$\int K^2(u)du < \infty \; and \; \int \mid u \mid^2 K(u)du < \infty.$$

- **(C3)**

$$\max_i \mid s_i - s_{i-1} \mid \approx \max_i \mid x_i - x_{i-1} \mid = O(1/n).$$

Theorem 2.3 (Droesbeke and Saporta (2011,[58])) *Assume that conditions (C1), (C2) and (C3) satisfied as* $\frac{\log n}{nh_n} \to 0$ *and* $\frac{nh_n^5}{\log n} = O(1)$, *then for* $z < \tau(x)$ *with*

$$\inf_{x \in \mathcal{X}} (1 - H(\tau(x)/x)) > 0$$

,

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

where

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

and

$$\sup_{0 \le z \le (1 - H(\tau(x)/x))} \mid r_n(x,z) \mid = O\left((nh_n)^{-3/4}(log n)^{3/4}\right) \; a.s.$$

with f_X *is the density function of* X.

According to Theorem 2.1 and Theorem 2.2 in [84], the asymptotic covariance and bias of conditional Kaplan-Meier estimator are

$$Cov_{asymptotic}(\hat{F}_n(t/x), \hat{F}_n(s/x) = (nh_n)^{-1}\Gamma(s, t/x)$$

and

$$Bias_{asumptotic}(\hat{F}_n(t/x)) = h_n^2$$

where

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

and

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

where $\dot{H}(s/x)$ and $\ddot{H}(s/x)$ are first order and second order derivatives respectively of $H(s/x)$ with respect to x. Under these results we get the formulas on the parameter h_n which minimizes the asymptotic mean squared error defined by:

$$\begin{aligned}MSE_{asymptotic}(h_n)&=\mathbb{E}_\infty\left[(\hat{F}_n(t/x)-F_n(t/x))^2\right]\\&=Var_{asumptotic}(\hat{F}_n(t/x))+(bias_{asymptotic}(\hat{F}_n(t/x))^2\\&=(nh_n)^{-1}\Gamma(t,t/x)+h_n^4b^2(t/x).\end{aligned}$$

where $\mathbb{E}_\infty$ is the expectation according to the asymptotic distribution. The value of h_n which minimizes this function, $MSE_{asymptotic}(h_n)$, is given by:

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

Under these conditions, one can announce the following asymptotic property of Beran estimator:

Theorem 2.4 (Droesbeke and Saporta (2011, [58])) *Assume that conditions (C1)-(C3) are satisfied.*

- *If* $nh_n^5\to 0$ *and* $\dfrac{(\log n)^3}{nh_n}\to 0$, *then, for* $n\to\infty$

$$(nh_n)^{1/2}(\hat{F}_n(./x)-F_n(./x))\to W(./x)$$

- *If* $h_n=Cn^{-1/5}$ *for some* $C>0$, *then for* $n\to\infty$

$$(nh_n)^{1/2}(\hat{F}_n(./x)-F_n(./x))\to \tilde{W}(./x)$$

where $W(./x)$ *and* $\tilde{W}(./x)$ *are Gaussian processes with covariance function* $\Gamma(.,./x)$, *and for* $\tilde{W}(./x)$, *mean function is* $b(./x)C^{5/2}$.

Graphical representation under censored data

1. **Using Kaplan-Meier estimator**
   The "Pareto quantile plot" can be determined in case of random right-censored data without covariable. Therefore, it suffices to replace the empirical distribution function by its Kaplan-Meier estimator. We have, then the representation of the points:

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

2. **Using generalized Kaplan-Meier estimator**
   One can also use "Pareto quantile plot" to represent graphically random right-censored data in presence of covariable x:

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

   where $\hat{F}_n(\cdot/x)$ designates to generalized Kaplan-Meier estimator.

Nelson-Aalen Estimator

Cumulative hazard function estimator has be introduced by Nelson [117] in 1972 and generalized later by Aalen [1] in 1978, taken the name **Nelson-Aalen** estimator. First, one can observe that under general hypothesis of independence between survival and censoring times, we can decompose $H(t)$ as:

$$H(t) = 1-(1-F(t))(1-G(t)) = H^{(c)}(t) + H^{(u)}(t) \tag{2.10}$$

where

$$H^{(c)}(t) := P(Z \leq t,\, \delta = 0) = \int_0^t \bar{F}(y)dG(y) \tag{2.11}$$

and

$$H^{(u)}(t) := P(Z \leq t,\, \delta = 1) = \int_0^t \bar{G}(y)dF(y) \tag{2.12}$$

For $t \geq 0$, the cumulative hazard function (2.6) can be expressed as following:

$$\Lambda(t) = \int_0^t \frac{\bar{G}(y)dF(y)}{\bar{H}(y)} = \int_0^t \frac{dH^{(u)}(y)}{\bar{H}(y)}.$$

Definition 2.6 (Nelson-Aalen Estimator) *The non-parametric Nelson-Aalen estimator Λ_n of Λ based on the sample $\{(Z_i, \delta_i,\ 1 \leq i \leq n)\}$ is defined by*

$$\Lambda_n(t) = \int_0^t \frac{dH_n^{(u)}(y)}{\bar{H}_n(y)} = \begin{cases} \sum\limits_{Z_{(i)} \leq t}^{n} \dfrac{\delta_{(i)}}{n-i+1} & if\ t < Z_{(n)}, \\ 1 \quad else. \end{cases}$$

where

$$H_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{z_i \leq t\}} \text{ and } H_n^u(t) = \frac{1}{n}\sum_{i=1}^{n} \delta_i \mathbb{1}_{\{z_i \leq t\}}$$

are respectively the empirical distribution function of $H(t)$ and the empirical version of $H_n^u(t)$ from the sample $Z_1, ..., Z_n$.

Note that by substituting $\Lambda(t)$ by $\Lambda_n(t)$ in (2.7), we obtain a new estimator of survival function, relative to Nelson-Aalen cumulative hazard function estimator, called Breslow estimator (see [28]) given by

$$\hat{S}_n^{N-A}(t) = \begin{cases} \prod\limits_{Z_{(i)} \leq t} \exp\left\{-\dfrac{\delta_{(i)}}{n-i+1}\right\} & if\ t < Z_{(n)}, \\ 0 \quad else. \end{cases}$$

Fleming and Harrington [53] have shown the close relationship between Nelson-Aalen estimator and Kaplan-Meier one, they compared numerically for several sample sizes and pointed out that the two estimators, are asymptotically equivalent. For more details, see Huang and Strawderman [93]. Recently, Dragi Anevski [5] derived process limit distribution results for the Nelson-Aalen estimator of a hasard function and for the Kaplan-Meier estimator of a distribution function, under different dependence assumptions.

2.2.5 Estimation of Extreme Value Index (EVI) under censoring

In this subsection, we focus on the problem of estimating the extreme value index in the case of right censored data. Beirlant et al. ([15]) and Einmahl et al. ([60]) proposed an inverse probability-of-censoring weighted (IPCW) method to adapt classical extreme value index estimators to censoring. Similarly, Gomes and Neves ([81]) and Brahimi et al. ([26]) used this idea to adapt various estimators to censoring. In addition, Beirlant et al. ([16]) addressed the problem of censoring, obtaining maximum likelihood estimators by adapting the likelihood function of the generalized Pareto distribution to censoring. Also, Worms and Worms ([157]) considered

estimators based on Kaplan-Meier integration and censored regression. After that Ameraoui et al. ([3]) estimated the extreme value index from a Bayesian perspectives and Beirlant et al. ([17]) proposed a reduced-bias estimator based on an extended Pareto distribution. Recently, beirlant et al [20] revisited the estimation of the extreme value index for randomly censored data from a heavy tailed distribution by introducing a new class of estimators which encompasses the ones given in Worms and Worms ([157]) and Beirlant et al. ([17]) and proved good bias properties. They also derived an asymptotic representation and the asymptotic normality of the larger class of estimators and consider their finite sample behavior, however they obtained the asymptotic normality in case of heavy censoring, i.e. where the amount of censoring in the tail is at least 50 %.

In this case, suppose we have two samples $Y_1,...,Y_n$ and $C_1,...,C_n$ of independent and identically distributed for each sample, distributed respectively according to F and G, such that $F \in AD(H_{\gamma_1})$ and $G \in AD(H_{\gamma_2})$ for some γ_1 and $\gamma_2 \in \mathbb{R}$. Let $(Z_i, \delta_i)_{1\leq i\leq n}$ be the really observed sample. It is clear that the $Z_i^{'}s$ independent random variables of distribution function H given by (2.10). The extreme value index of H exists, denoted by, γ with $\gamma = \dfrac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}$. Let x_F, x_G and x_H be the endpoints of F, G and H respectively. A general adaptation of existing EVI estimators to censorship is provided by [60] in the following cases:

$$\begin{cases} case\ 1\ \gamma_1 > 0,\ \gamma_2 > 0, \\ case\ 2\ \gamma_1 < 0,\ \gamma_2 < 0,\ x_F = x_G, \\ case\ 3\ \gamma_1 = \gamma_2 = 0,\ x_F = x_G = \infty. \end{cases}$$

Their estimators are based on a standard estimator of the index tail divided by the estimator of the uncensored data proportion in the greatest k real observed order statistics, that is

$$\hat{\gamma}_1^{(\bullet,c)}(k) = \frac{\hat{\gamma}^{\bullet}}{\hat{p}}, and\ \hat{p} = \frac{1}{k}\sum_{i=1}^{k}\delta_{(n-i+1)} \tag{2.13}$$

The notation $\hat{\gamma}^{\bullet}$ can be any estimator not adapted for censorship and $\hat{p}$ is the estimator of the proportion of observed data at the right tail distribution. Beirlant et al. [15] are the first who introduced this methodology in the case of Hill's and Moment's estimators. Furthermore, they proposed the estimators of the extreme quantiles and have discussed their asymptotic properties when data are censored by a deterministic threshold. Einmahl et al. [60] proved that $\hat{\gamma}_1^{(\bullet,c)}$ is a consistent and asymptotically Gaussian estimator of γ_1 as soon as $\hat{\gamma}^{\bullet}$ and $\hat{p}$ are also consistent

and asymptotically Gaussian estimators of γ_1 and p respectively.
The main extreme quantile estimator $Q(1-s)$ under random censorship available in the literature has been proposed by Beirlant et al. [15] and by Einmahl et al. [60]. It is given by the following definition:

Definition 2.7 ***(Extreme quantile estimation under random censorship)***
Extreme quantile estimator under random censorship is defined by

$$\hat{Q}^{(\bullet,c)} = Z_{(n-k)} + \hat{a}^{(\bullet,c)} \frac{\left(\left(1 - F_n(\hat{Z_{(n-k)}})\right)/s\right)^{\hat{\gamma}_1^{(\bullet,c)}} - 1}{\hat{\gamma}_1^{(\bullet,c)}} \tag{2.14}$$

where $\hat{a}^{(\bullet,c)} = Z_{(n-k)} M_n^{(1)} (1 - T_n)/\hat{p}$, *with* $M_n^{(1)}$ *and* T_n *are defined in* (1.32).

2.2.6 Conditional Extreme quantiles and Tail Index (EVI) under censoring

Introduction

The study of estimation of conditional extreme quantile in incomplete data frameworks is of growing interest. Specially, the estimation of the extreme value index in a censorship framework has been the purpose of many investigations when finite dimension covariate information has been considered. In the case of the presence of both covariate information and censoring, Ndao et al. ([115]) proposed three estimators for the estimation of the conditional extreme value index and extreme quantiles for heavy-tailed distributions. In particular, the Hill, generalized Hill and moment type estimators were proposed using the moving window method (Gardes and Girard, [71]) and adapting the estimators to censoring by using the inverse probability of censoring weighted method (IPCW) (Beirlant et al. [15]; Einmahl et al. [60]). Whereas Stupfler ([138]) proposed a moment estimator valid for all domains of attraction. In addition, Ndao et al. ([116]) addressed the estimation of the extreme value index under censoring and the presence of random covariates.

Recently, Justin Ushize Rutikanga and Aliou Diop [129] and [130] discussed the estimation of the conditional extreme quantile and EVI of a heavy-tailed distribution when some functional random covariate (i.e. valued in some infinite-dimensional space) information is available and the scalar response variable is right-censored, they proposed a weighted kernel version of Hill's estimator of the extreme-value index and established its asymptotic normality under mild assumptions.

In this section, we present fix design conditional index and quantile estimators under radom right censoring, proposed by Ndao [114] and Lekina [105] in case of heavy tailed distributions by using mobile window method as in [71].

Some EVI estimators

Suppose that conditional distribution functions of Y and C given $x \in \mathcal{X}$ are heavy tailed functions with positive conditional tail indexes $\gamma_1(.)$ and $\gamma_2(.)$ respectively. Note by $Z_i^x, \delta_i^x, i = 1,...,m_n^x$ real observed data such that its covariates are in sufficiently small neighborhood of x, that is

$$(Z_i^x, \delta_i^x) = \{(Z_i, \delta_i)\, if \mid x_i - x \mid \leq h_{n,x}\},\ i = 1,...,n\ with\ h_{n,x} \to 0\ as\ n \to \infty.$$

The associate ordered sample is then denoted by

$$(Z_{(1)}^x, \delta_{(1)}^x),...,(Z_{(m_n^x)}^x, \delta_{(m_n^x)}^x).$$

In the general case, Ndao [116] defined the EVI estimator by

$$\hat{\gamma}_{k_x,m_n^x}^{(\cdot,c)}(x) = \frac{\hat{\gamma}_{k_x,m_n^x}^{(\cdot)}(x)}{\hat{p}_x} \tag{2.15}$$

where $\hat{p}_x = \frac{1}{k_x}\sum_{i=1}^{k_x} \delta_{(m_n^x-i+1)}^x$ estimates $p_x = \frac{\gamma_2(x)}{\gamma_1(x)+\gamma_2(x)}$ and $\hat{\gamma}_{k_x,m_n^x}^{(\cdot)}(x)$ can be:

1. **Hill adapted estimator**(1975, [90])

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

2. **Dekkers-Einmahl-de Haan adapted estimator** (1989, [49])

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

where

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

3. **UH (Beirlant et al.) adapted estimator**, 1996, [19]

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

For more details on these estimators, the reader can consult Chapter 1 in [114].

Conditional extreme quantile estimation

A conditional extreme quantile, $q(\alpha_{m_n^x}, x)$, of order $\alpha_{m_n^x}$ is obtained by solving the equation

$$\mathbb{P}(Y > q(\alpha_{m_n^x}, x)/X = x) = \alpha_{m_n^x}$$

where $\alpha_{m_n^x} \to 0$ as $n \to \infty$

Ndao [114] gave, by inverting distribution function at order $\alpha_{m_n^x}$ the solution of the above equation in case of heavy tailed functions:

$$q(\alpha_{m_n^x}, x) = Z^x_{(m_n^x - k_x)} \left(\frac{\bar{F}(Z_{(m_n^x - k_x)}/x)}{\alpha_{m_n^x}} \right)^{\gamma_1(x)}. \tag{2.16}$$

and the adapted family estimators are

$$\hat{q}_(\alpha_{m_n^x}, x)^{(\cdot,c)} = Z^x_{(m_n^x - k_x)} \left(\frac{1 - \hat{F}_{m_n^x}(Z^x_{(m_n^x - k_x)}/x)}{\alpha_{m_n^x}} \right)^{\hat{\gamma}^{(\cdot,c)}_{k_x, m_n^x}(x)}. \tag{2.17}$$

where $\hat{F}_{m_n^x}(./x)$ is generalized Kaplan-Meier estimator. For example, if one take Hill estimator for extreme value index, one obtained:

$$\hat{q}_(\alpha_{m_n^x}, x)^{(H,c)} = Z^x_{(m_n^x - k_x)} \left(\frac{1 - \hat{F}(Z^x_{(m_n^x - k_x)}/x)}{\alpha_{m_n^x}} \right)^{\hat{\gamma}^{(H,c)}_{k_x, m_n^x}(x)}.$$

Asymptotic properties

Under some conditions of regularity (see [114], **C1-C5**, p.44), the next theorem gives the asymptotic normality of $\hat{\gamma}^{(\cdot,c)}_{k_x, m_n^x}(x)$ in the general case and its corollary especially for hill, UH and moment estimators. (see its proofs in [115])

Theorem 2.5 ([115], Theorem 4.1) *Let $x \in \mathcal{X}$. Under conditions C1 - C6 and if it exists functions $m(.)$: and $\sigma(.)$ such that $\sqrt{k_x}(\hat{\gamma}^{(\cdot)}_{k_x, m_n^x}(x) - \gamma(x)) \xrightarrow{D} \mathcal{N}(m(x)\lambda(x), \sigma^2(x))$, then*

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

where

$$\lambda(x) = \lim_{k_x \to \infty} \sqrt{k_x} b\left(\frac{m_n^x}{k_x}, x\right)$$

and $b(., x)$ is a regularly varying function with index $\rho(x)$

Corollary 2.1 *Under hypothesis* ***C1-C6*** *and* $k_x^{1/2} h_{n,x}^{\alpha_U} \to 0$, *then*

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

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

$$\sqrt{k_x}\left(\hat{\gamma}^{(M,c)}_{k_x,m_n^x}(x) - \gamma_1(x)\right) \xrightarrow{D} \mathcal{N}\left(\frac{-\gamma_1(x)\epsilon(x)}{p_x} + \frac{\lambda(x)}{p_x(1-\rho(x))}, \frac{\gamma_1^2(x)}{\gamma^2(x)}(1+\gamma_1(x)\gamma(x))\right)$$

We now turn to the asymptotic properties of the estimator (2.17) of the conditional extreme quantiles. Under a further regularity assumption on $q(.,.)$ (condition C7 in [114], p. 46), Ndao ([115] established its asymptotic normality via the following theorem:

Theorem 2.6 ([114], Theorem 2.4.2) *Assume that conditions C1-C7 hold. Let* $(\alpha_{m_n^x})_{n\geq 1}$ *and* $(\beta_{m_n^x})_{n\geq 1} := (1 - \hat{F}(Z^x_{(m_n^x - k_x)}/x))_{n\geq 1}$ *two sequences such that* $\alpha_{m_n^x} < \beta_{m_n^x}$ *and let*

$$\xi_{m_n^x} = (m_n^x \beta_{m_n^x} \beta_{m_n^x})^{1/2} \log\left(\frac{\beta_{m_n^x}}{\alpha_{m_n^x}}\right).$$

If $n \to \infty$,, *it exists* $\delta > 0$ *such that*

$$(m_n^x \beta_{m_n^x} \beta_{m_n^x})^2 \omega_n((m_n^x)^{-(1+\delta)}) \to 0$$

and

$$k_x^{1/2} \max\left[\xi^{-1}_{(m_n^x}, \bar{\Delta}(\beta_{m_n^x}, x)\right] \to 0.$$

Then,

$$\frac{\sqrt{k_x}}{\log\left(\frac{\beta_{m_n^x}}{\alpha_{m_n^x}}\right)} \log\left(\frac{\hat{q}_(\alpha_{m_n^x}, x)^{(\cdot,c)}}{q_(\alpha_{m_n^x}, x)}\right) \xrightarrow{D} \mathcal{N}\left(\frac{\lambda(x)m(x) - \gamma_1(x)\epsilon(x)}{p_x}, \frac{\sigma^2(x) + \gamma_1^2(x)p_x(1-p_x)}{p_x^2}\right).$$

where $\omega_n(.)$ *is the greatest oscillation of log-quantile function (see ndaothesis, p.46)*

2.3 Mean lifetime Estimation under right censoring

2.3.1 Kaplan-Meier Integral

In a remarkable paper, Stute [140] extended the Central limit Theorem in full generality, where the mean is a special case, to the random censorship model. Let $Y_1, ..., Y_n$, the variables of interest, one observe

$$Z_i = \min(Y_i, C_i) \text{ and } \delta_i = 1_{\{Yi \leq C_i\}},\ i = 1, ..., n$$

as defined in Section 2.2. Let $\varphi : \mathbb{R} \to \mathbb{R}$ be any measurable function such that $\int \varphi^2 dF < \infty$ (i.e. φ is F square integrable function). One have also $\varphi(Y_1),...,\varphi(Y_n)$ are i.i.d. Put

$$S_n^{\varphi} = \int \varphi \, d\hat{S}_n,$$

where $\hat{S}_n$ is the kaplan-Meier product limit estimator given by (2.9). It is easily seen from (2.9) that

$$S_n^{\varphi} = \sum_{i=1}^{n} W_{i,n}\varphi(Z_{(i)}),$$

where for $1 \leq i \leq n$

$$W_{i,n} = \frac{\delta_{(i)}}{n-i+1}\prod_{j=1}^{i-1}\left(\frac{n-j}{n-j+1}\right)^{\delta_{(j)}}.$$

is the weight attached to the ith order statistic $Z_{(i)}$ under $\hat{S}_n$. If there is no censoring, all δ's equal 1 so that each W takes $1/n$ value. Whereas, under censoring, S_n^{φ} becomes function of $Z_{(i)}$'s correctly weighted by W's random quantities.

2.3.2 Sample mean under random right censoring

If we put $\varphi(z) = z$, we obtain $S_n^{\varphi} = \mu$

Definition 2.8 *Under random censoring, the sample mean estimator is defined by*

$$S_n^{\varphi} = \tilde{\mu}_n := \sum_{i=1}^{n} W_{i,n}Z_{(i)}.$$

2.3.3 Distributional Convergence of the Kaplan-Meier Integral

The asymptotic normality of the KM integral was investigated by different authors, we cite: Gill [74] who showed the distributional convergence for non negative, continuous and increasing φ's, Schick *et al.* [132] established a weak representation of S_n^{φ} in terms of a sum of i.i.d. random variables, Yang [158] extended the weak convergence of S_n^{φ}, under some regularity conditions on F, to φ's satisfying $\int \varphi^2/\bar{G}dF < \infty$. Stute [140] obtained a representation of S_n^{φ} as a sum of i.i.d. random variables plus a remainder without regularity conditions on F and G ,as shown in Theorem 2.7 below, under the following assumptions:

$$\int_0^{\infty} x^2\Gamma_0^2(x)dH^{(u)}(x) < \infty, \tag{2.18}$$

$$\int_0^\infty x\left(\int_0 \frac{dH^c(y)}{\left[\bar{H}(y)\right]^2}\right)^{1/2} dF(x) < \infty \tag{2.19}$$

where $H^{(c)}$ and $H^{(u)}$ are the subdistribution functions given by:

$$H^{(c)}(t) := \mathbb{P}(Z \le t, \delta = 0) = \int_0^t \bar{F}(x) dG(x),$$

and

$$H^{(u)}(t) := \mathbb{P}(Z \le t, \delta = 1) = \int_0^t \bar{G}(x) dF(x),$$

with

$$\gamma_0(x) := \exp\left\{\int_0^x \frac{dH^c(s)}{\bar{H}(s)}\right\}$$

$$\gamma_1(x) := \int_0^x \frac{s\Gamma_0(s)}{\bar{H}(s)} dH^u(s)$$

and

$$\gamma_2(x) := \int_0^x \int_s^\infty \frac{t\Gamma_0(t)}{\left[\bar{H}(s)\right]^2} dH^u(t) dH^c(s)$$

Theorem 2.7 (Stute ([140], Corollary 1.2.) *Under assumptions (2.18) and (2.19),*

$$\sqrt{n}(\tilde{\mu} - \mu) \xrightarrow{d} N(0, \sigma^2),$$

where

$$\sigma^2 := Variance[Z_1\Gamma_0(Z_1)\delta_1 + \Gamma_1(Z_1)(1-\delta_1) - \Gamma_2(Z_1)]$$

For the proof, one can see [140].

2.3.4 Further readings

- Concerning the bias of $\int \varphi dF_n$ in estimating $\int_0^\tau \varphi dF$ (where τ is the endpoint of d.f. H, Mauro [112] showed that for nonnegative φ's, the bias is negative.
- Zhou [161] obtained a lower bound of the bias for nonegative and continuous φ's.
- Stute [139] derived an expansion of the bias and showed that the bias decreases to zero exponentially fast as $n \to \infty$ if φ is bounded and vanishes right of some $T < \tau$.

- Informally speaking, Stute [142] indicated that the bias of a Kaplan-Meier integral may decrease to zero at any polynomial rate, if, e.g., $0 \leq \varphi(x) \uparrow \infty$ as $x \to \infty$ and censoring is heavy (as an example, in estimating mean lifetime; that is $\varphi(x) = x$, reader can see [140] (Section 6, p.247).

- In order to reduce the bias, Stute and Wang [144] proposed a jackknife modification of $\int \varphi dF_n$.

- Suzukawa [146] obtained a representation of the Kaplan-Meier integral in terms of the Kaplan-Meier estimator of a censoring distribution. Moreover, he considered a class of unbiased estimators of $\int_0^\tau \varphi dF$ under the condition that the censoring distribution is known.

- Soltane [137] proposed an estimating approach of the mean ensuring the asymptotic normality property for some class of heavy-tailed distributions for which The central limit theorem introduced by Stute [140] does not hold.

Chapter 3

Study of extreme rainfalls using Extreme Value Theory (case study: Khemis-Miliana region - Algeria)

Abstract

The main topic of this work is the statistical analysis of extreme values (EVA) with applications to hydrology, more specifically, to rainfalls. Statistical inference of rainfall is very important as we consider the risk of damage to agriculture, ecology, infrastructure systems and also risk of drought. The main aim of this study is to find out the most adequate fitting distributions of rainfalls taken in Khemis-Miliana region (Algeria) during the period 1975-2006. The method of Block Maxima (BM) is adopted when we use Generalized Extreme Value (GEV) distribution to fit the data, and the Peak Over Threshold (POT) method is applied when we use Generalized Pareto (GP) distribution, after testing of course stationarity of time serie in hand.

Concerning estimation of parameters, we use: Maximum Likelihood Estimation (MLE), Probability Weighted Moments (PWM) and Profile of Maximum Likelihood (PMLE) for both models (GEV and GP). With these models, we derive estimates of T-years return levels for different periods T and vice-versa.

Keywords: GEV, GPD, POT, Block Maxima, extreme values, extremes quantiles, return period, return level.

AMS 2000 Subject Classifications 62G32-62G05

3.1 Introduction

"However big floods get, there will always be a bigger one coming; so says one theory of extremes, and experience suggests it is true." (PRESIDENT'S WATER COMM., p. 141.)([77])

The oldest problems connected with extreme values arise from floods, so we should protect our life and property against the damages caused by inundations which are considerated as rare events that are more extreme than any that have already been observed. Meteorological data generally have no alarming aspects as long as they are situated in a narrow band around the average. The situation changes for instance when concentrations occur that overshoot a specific ecological threshold like Rainfall data with tremendous impact on society as they are among the most common themes for discussion, specifically global warming and climate change. Typically, one is interested in the analysis of maximal and minimal observations and records over time (often attributed to global warming) since these entail the negative consequences. Algeria, as a Mediterranean country, has undergone severe climatic changes during the last decades in terms of rainfall. The floods of Bab El Oued in 2001 and Ghardaia in 2008 were really catastrophic, caused by substantial extreme rainfalls. The role of extreme value theory is to develop procedures which are scientifically and statistically rational for estimating the extremal behaviour of such random variables or processes. In this context, the German mathematician Emil Gumbel (1891–1966), who was a pioneer in the application of EVT to engineering problems, in particular to hydrological phenomena such as annual flood flows, once wrote: *"It seems that the rivers know the theory. It only remains to convince engineers of the validity of this analysis."*[77]. The foundations of the asymptotic argument which forms the backbone of extreme value theory were set out by Fisher and Tippett in the 1920's though they were reluctant to propose the models for statistical application. This theory was unified and extended by Gnedenko in the 1940's. The statistical application of the probabilistic models for extremes was first studied and formalized by Gumbel in the 1950's. Also in the 1950's, Jenkinson[95] worked on the application of extreme value models to extreme wind speeds and developed a model parametrization which unified a number of previously disparate models thus adding clarity to the modelling procedure

In the 1970's the classical limit laws were generalized by Pickands leading to substantially improved modelling procedures which were developed in the 1980's

and 1990's. Also throughout the 1980's the extremal behaviour of a much more general class of processes admitting various types of non stationarity and dependence were investigated. Furthermore the characterization and statistical inference of multivariate extremes has been developed since the mid 1980's. The extreme levels of a river causing floods in hydrology were also introduced in the literature by Coles and Tawn (2005)[35]. Concerning rainfall, and the risk of floods caused by this phenomenon, a considerable number of studies aiming to model such events can be found in the literature: Coles (2001, [34]) provided a detailed discussion about the methods used to model such events, and extreme rainfall was modelled by Friederichs (2010)[70] in Germany, Benestad (2010)[21] in Norway, Kim et al. (2009) [99] in the south of Korea and Deka et al. (2011)[47] in India. Actually, extreme value theory is a blend of a variety of applications and sophisticated mathematical results on point processes and regular varying functions.

According to Fisher-Tippet's theorem , if the maximum value of a distribution function (d.f.) tends (in distribution) to a nondegenerate d.f. then this limiting d.f. can **only** be the Generalized Extreme Value (GEV) distribution.

The rest of the paper is organized as follows: in Section 2, we give the basic concepts of the classical block maxima (for extremes) and threshold exceedances (for some high threshold) models and then making inferences for both models including its parameters and return levels (quantiles) by ML and PWM methods. Finaly, Section 3 is devoted to an application to the extreme rainfalls at Khemis -Miliana station (Algeria).

3.2 Models and Methods

Two different methods are used to model extreme events, one is the Block Maxima Method (BMM) which involves the Generalized Extreme Value (GEV) distribution and another one is the Peak Over Threshold method which involves the Generalized Pareto Distribution (GPD).

3.2.1 The Generalized extreme value distribution

Let X_1, X_2,... be a sequence of independent random variables with common distribution function F, and let

$$M_n = max(X_1, X_2, ..., X_n)$$

Learning more about the (M_n) distribution would give substantial information about the extreme values of (X_n).

The entire range of possible limit distributions for M_n with adequate normalization is given by the *extremal types theorem* (also called limit theorem of Fisher and Tippet (1928)[69] and Gnedenko (1943)[78]).

A better analysis is offered by a reformulation of the models in the extremal types theorem. It is straightforward to check that the Gumbel, Frechet and Weibull families can be combined into a single family of models having distribution functions of the form

$$G(z) = \exp\{-[1 + \xi(\frac{z-\mu}{\sigma})]^{-1/\xi}\} \tag{3.1}$$

defined on the set $\{z : 1 + \xi(z-\mu)/\sigma > 0\}$, where the parameters satisfy $-\infty < \mu < \infty, \sigma > 0$ and $-\infty < \xi < \infty$. This is the generalized extreme value (GEV) family of distributions. The model has three parameters: a location parameter, μ; scale parameter, σ; and a shape parameter, ξ.

Suppose that data are blocked into sequences of observations of length n, for some large value of n, generating a series of block maxima, $M_{n,1}, \ldots, M_{n,m}$, say, to which the GEV distribution can be fitted. Often the blocks are chosen to correspond to a time period of length one year, in which case n is the number of observations in a year and the block maxima are annual maxima.

3.2.2 The Generalized Pareto Distribution

Suppose that F satisfies *extremal types theorem*, so that for large n, $\mathbb{P}\{M_n \leq z\} \approx G(z)$, where

$$G(z) = \exp\{-[1 + \xi(\frac{z-\mu}{\sigma})]^{-1/\xi}\}$$

for some $\mu, \sigma > 0$ and ξ. Then, for large enough u, the distribution function of $(X - u)$, conditional on $X > u$, is approximately

$$H(y) = 1 - (1 + \frac{\xi y}{\tilde{\sigma}})^{-1/\xi} \tag{3.2}$$

defined on $\{y : y > 0 \; and \; 1 + \xi y/\tilde{\sigma} > 0\}$, where

$$\tilde{\sigma} = \sigma + \xi(u - \mu)$$

The family of distributions defined by Eq. 3.2 is called the **generalized Pareto family**. The result above implies that if block maxima have approximating distribution G, then threshold excesses have a corresponding approximate distribution

within the generalized Pareto family. Moreover, the parameters of the generalized Pareto distribution of threshold excesses are uniquely determined by those of the associated GEV distribution of block maxima.

3.2.3 Threshold Selection

Threshold choice is similar to the choice of block size in the block maxima method, implying a balance between bias and variance. That is, too low a threshold is likely to violate the asymptotic properties of the model, leading to bias; too high a threshold will generate few excesses (insufficient data) with which the model can be estimated, carrying out a high variance.

Whenever we are interested in large values, estimation of model parameters is usually performed on the basis of the largest $k+1$ order statistics in the sample or on the excesses over a high level u. The question subtracted in practical applications of *extreme value theory* is the choice of either k or u. Such a choice can be either heuristic(explorative, see [83]) or based on sample paths stability or on the minimization of a mean squared error estimate as a function of k which highly sensitive to small changes in the threshold, we can state for the latest one (not exhaustively):

- Smooth Hill estimator ([128])
- Kernel EVI estimators ([36])
- Semi-parametric PWM-EVI estimator ([29])

Among EVI estimators that are less sensitive to the choice of k and easy to use, are Hill reduced-Bias (RB) estimators ([125] and [10]). For this purpose, we adopt the two (graphical) methods provided by Coles [34]; one is an exploratory technique carried out prior to model estimation; the other is an assessment of the stability of parameter estimates, based on the fitting of models across a range of different thresholds, for more details see [34]

3.2.4 Parameter Estimation

There are several numerical methods for the determination of the three parameters μ, σ and ξ, but the most commonly described in the literature are maximum likelihood (ML) and probability weighted moments (PWMs) solutions. We describe

here the PWM method. The ML method is described in detail elsewhere (Davison (1985) [41]; Smith (1986) [135]; Hosking and Wallis (1987) [92]; Davison and Smith (1990) [40]; Wilks (1995) [156]). The PWM method goes back to Hosking, Wallis and Wood [91] which is a competitor method to maximum likelihood for estimating the parameters. Define PWM of order r by:

$$w_r = E(XH_\theta^r(X)),\ r \in \mathbb{N} \tag{3.3}$$

where H_θ is the GEV distribution and X has df H_θ with parameter $\theta = (\xi, \mu, \sigma)$.

Recall that for $\xi \geq 1$, H_θ is regularly varying with index $1/\xi$ (see Embrecht [61]). Hence w_0 is infinite, therefore we restrict ourselves to the case $\xi < 1$. Define the empirical analogue to 3.3,

$$\hat{w}_r(\theta) = \int_{-\infty}^{+\infty} xH_\theta^r(x)dF_n(x),\ r \in \mathbb{N}$$

where F_n is the empirical df corresponding to the data $X_1, X_2, \ldots, X_n$. In order to estimate θ we solve the equations:

$$w_r(\theta) = \hat{w}_r(\theta), r = 0, 1, 2.$$

3.2.5 Return Level (Quantiles)

The return level is an interesting notion to determine the mean waiting time between extreme rainfalls. Estimates of *extreme quantiles* of the annual maximum distribution (BM approach) are then obtained by inverting Eq. 3.1:

$$x_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - \{-\log(1-p)\}^{-\xi}\right], for\ \xi \neq 0 \\ \mu - \sigma\log\{-\log(1-p)\}, for\ \xi = 0 \end{cases} \tag{3.4}$$

where $G(x_p) = 1 - p$. In common terminology, x_p is the *return level* associated with the return period $1/p$, since to a reasonable degree of accuracy, the level x_p is expected to be exceeded on average once every $1/p$ years. More precisely, x_p is exceeded by the annual maximum in any particular year with probability p.

Suppose now the case of POT approach, let we have a Generalized Pareto distribution with parameters σ and ξ as a suitable model for exceedances of a threshold u by a variable X.

It follows that

$$P(X > x) = \zeta_u[1 + \xi(\frac{x-u}{\sigma})]^{-1/\xi}$$

where $\zeta_u = \mathbb{P}(X > u)$. Hence the level x_m that is exceeded on average once every m observations is the solution of

$$\zeta_u[1 + \xi(\frac{x_m - u}{\sigma})]^{-1/\xi} = \frac{1}{m}$$

which yields to

$$x_m = u + \frac{\sigma}{\xi}[(m\zeta_u)^{\xi} - 1]^{-1/\xi}$$

provided m is sufficiently large to ensure that $x_m > u$, this all assumes that $\xi \neq 0$. If $\xi = 0$, by the same procedure we obtain

$$x_m = u + \sigma \log(m\zeta_u)$$

again provided m is sufficiently large.

3.3 Application

3.3.1 Data description

In this section we will fit our data by the two models cited above. Our dataset consists of the daily and annual maxima of rainfall from 1975 to 2006 registered in Khemis-Miliana weather station, the unit of measurement is milimeter.

We focus our attention on the estimation of the index of extreme values (EVI), return levels and return periods. For the computer tool we used R software which contains a large number of packages with several functions will be used and mentioned to model the data extremes, such as `evd`, `ismev`, `evir`, `POT`, `fExtremes`,..., etc. Gilleland et al.[76] give an excellent software review for extreme value analysis, they describe and compare packages available in R with other software.

3.3.2 A preliminary data analysis

The results of a preliminary graphical and descriptive analysis are shown in Figure 3.1 and Table 3.1.

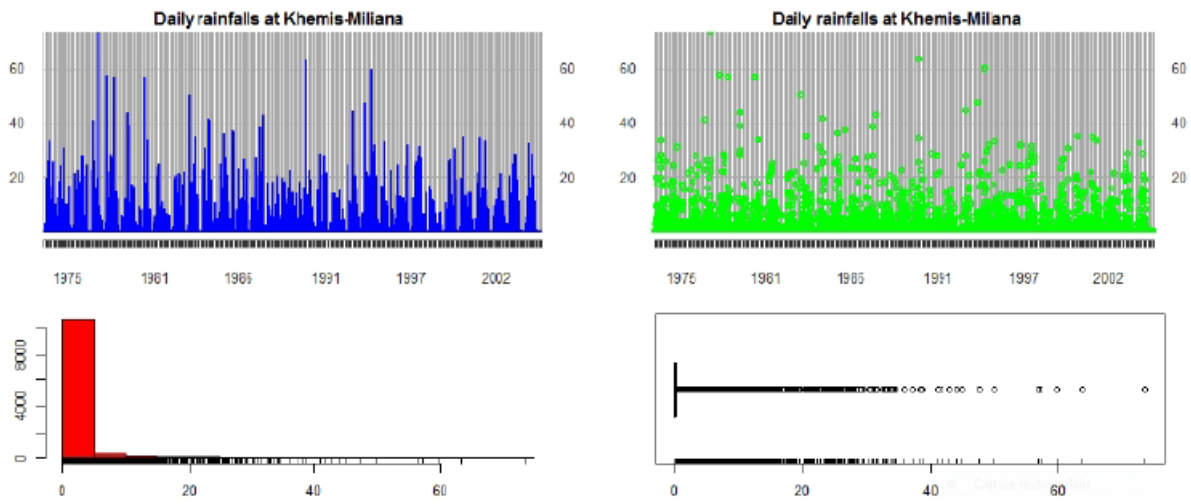

Figure 3.1: Chronogram (top left); scatter plot (top right); histogram (bottom left) and box-plot (bottom right))

n	Min	Max	1st Quart.	Median	Mean	3rd Quart.	St Dev.	Skew.	Kurt.
11324	0	73.7	0	0	1.11	0	3.95	6.37	58.12

Table 3.1: Descriptive statistics

The boxplot, the histogram and the descriptives statistics, in particular the skewness = 6.37 and the kurtosis = 58.12 indicate a heavier tail than the normal one.

3.3.3 Stationarity test

The stationarity was also studied by the Augmented Dickey–Fuller Test through the function **adf.test()**, available in the *package* **tseries**, in our case the test=-19.20 and p-value=0.01<0.05 so we accept the alternative hypothesis which allows us to admit the stationarity of our statistical serie.

3.3.4 Modeling using Generalized Extreme Value (GEV) distribution

In this framework, we have considered the years as blocks of observations and have picked the maximum values up in each block. So, we will use the maximum values of each of 31 years. For the yearly maximum rainfalls, the skewness = 0.998 and the kurtosis =2.265. Graphical analysis are shown in Figure 3.2.

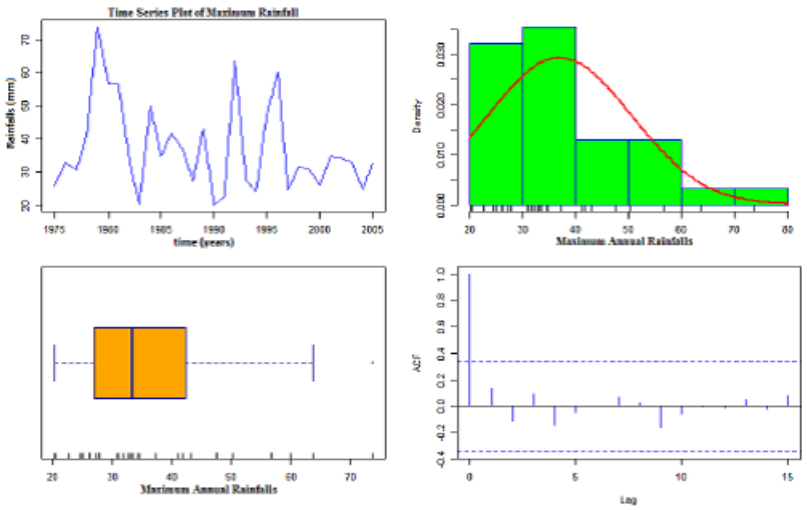

Figure 3.2: Time serie plot of Annual Maxima (top left); histogram of Annual Maxima (top right); box-plot of Annual Maxima (bottom left) and Partial auto-correlation function (bottom right)

The histogram, the boxplot and the skewness indicate a moderate positive asymmetry. From the partial autocorrelation function (ACF), it seems reasonable to assume that these data are not correlated.

Parameter estimation of GEV_ξ ($\xi \in \mathbb{R}$)

We use MLE and PWM to estimate the 3 parameters of GEV with giving its 95% confidence intervals (see results in tables 3.2, 3.3 and figure 3.3.)

	Shape parameter		Location parameter		Scale parameter	
Method	$\hat{\xi}$	SE	$\hat{\mu}$	SE	$\hat{\sigma}$	SE
MLE	0.21	0.18	30	1.81	8.55	1.47
PWM	0.14		30.10		9.19	

Table 3.2: Parameter estimates of GEV_ξ

	Shape parameter $\hat{\xi}$	Location parameter $\hat{\mu}$	Scale parameter $\hat{\sigma}$
CI. Normal Approximation	(-0.15;0.57)	(26.45;33.55)	(5.67;11.43)
CI. Log-Likelihood Profile	(-0.10;0.64)	(26.75;33.90)	(6.17;12.09)

Table 3.3: 95% Confidence Intervals of GEV_ξ parameters

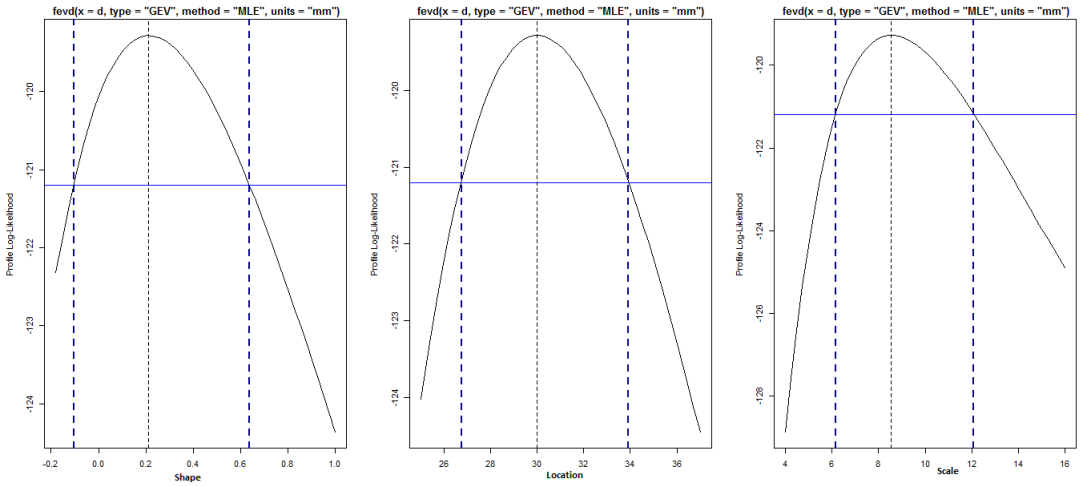

Figure 3.3: Profile log-Likelihood of GEV_ξ parameters by MLE.

For our data, a plot of the log-likelihood is shown in Figure 3.3. Using approximation to the sampling distribution for large sample sizes (see [34], Theorem 2.6, p. 35), a 95% confidence intervals for 3 parameters are obtained by drawing a line at a height of $0.5xc_{1,0.05}$ below the maximum of this graph, where $c_{1,0.05}$ is the 95% quantile of a χ^2_1 distribution, and reading off the points of intersection. This leads to a 95% confidence interval for $\hat{\xi}$ of $[-0.10, -0.64]$. Compared with the previous interval (Normal Approximation) of $[-0.15, 0.57]$ the profile likelihood interval is almost similar in width, but is shifted to the right, corresponding to the skewness observed in Figure. 3.3.

From Table 3.2, it is clear that the Shape parameter estimator is positive and it is close to zero, which implies that the GEV distribution in this case study is suspected to be of type *Fréchet* or *Gumbel* (see test below).

Graphic Diagnostics of GEV_ξ ($\xi \in \mathbb{R}$)

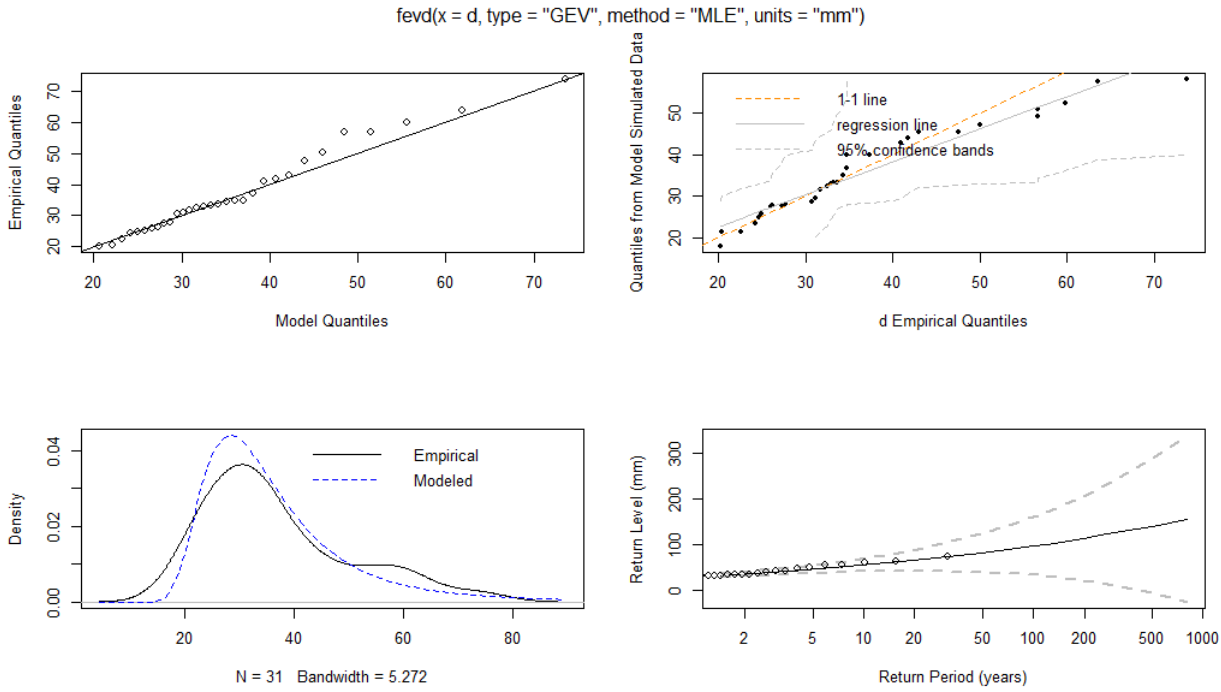

Figure 3.4: Graphic diagnostic of GEV_ξ Model with MLE.

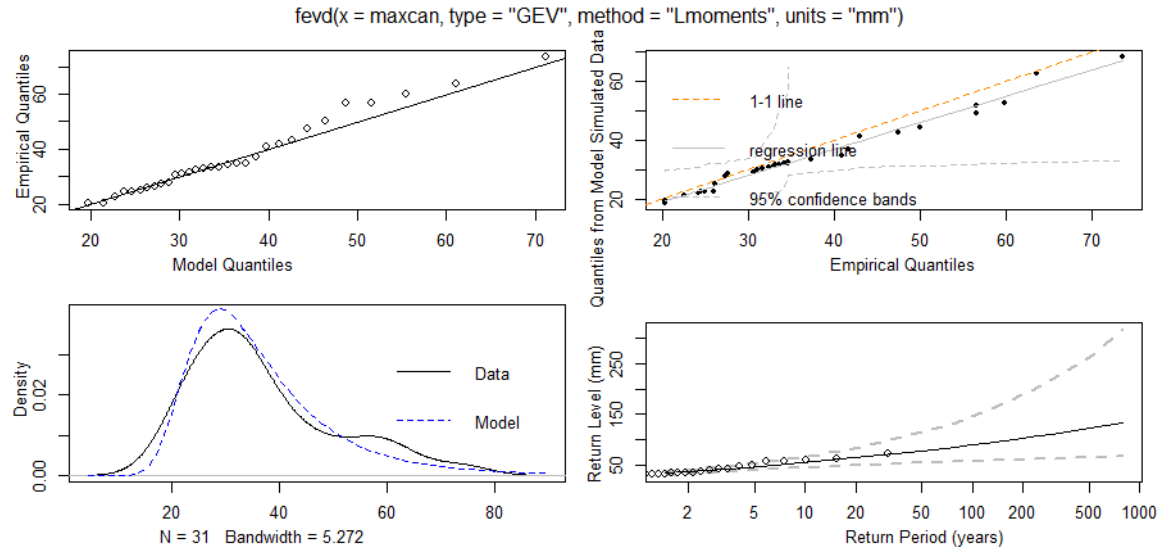

Figure 3.5: Graphic diagnostic of GEV_ξ Model with Lmoments (linear combination of PWM)

The various diagnostic plots for assessing the accuracy of the GEV model fitted to the rainfall data are shown in Figures. 3.4 and 3.5 using MLE and Lmoments methods respectively. Both probability plot and quantile plot give the validity of the fitted model : each set of plotted points is near-linear. The return level curve is convex and has no finite bound as a consequence of the positive estimate of ξ, though since the estimate is close to zero, the estimated curve is close to linear. Furthermore the corresponding density estimate seems consistent with the empirical density of the data. Consequently, all four diagnostic plots lend support to the fitted GEV model.

Testing GEV_ξ vsGEV_0 $(Gumbel)$

As the confidence interval of the shape parameter contains the value 0 (see table (3.4)) the GEV_0 is a candidate to be more adequate than GEV_ξ.To choose the best distribution (between GEV_ξ and GEV_0), the Likelihood-ratio test gives:

test	α	Chi square critical value	df	p-value
1.58	0.05	3.84	1	0.20

Table 3.4: Likelihood Ratio test (GEV_0 vs GEV_ξ)

According to the table above, the test value or the deviance (1.58) is less than the Chi square critical value which means that we suggest to accept that GEV_0 (Gumbel) Model is a plausible reduction of the GEV_ξ Model. So we accept the null hypothesis, that is the GEV_0 (Gumbel) model is more appropriate than the GEV_ξ model, hence we can continue modeling the rainfall at Khemis-Miliana using the GEV_0 model.

Parameter estimation of GEV_0 $(GumbelModel)$

Results are presented in Tables 3.5, 3.6 and graphics 3.6, 3.7

	Location parameter		Scale parameter	
Method	$\hat{\mu}$	SE	$\hat{\sigma}$	SE
MLE	31.03	1.79	9.50	1.41

Table 3.5: parameters estimation of GEV_0.

	Location parameter $\hat{\mu}$	Scale parameter $\hat{\sigma}$
CI. Normal Approximation	(27.53;34.53)	(6.73;12.26)
CI. Log-Likelihood Profile	(27.51;34.75)	(7.22;12.93)

Table 3.6: 95% Confidence Intervals of GEV_0 parameters

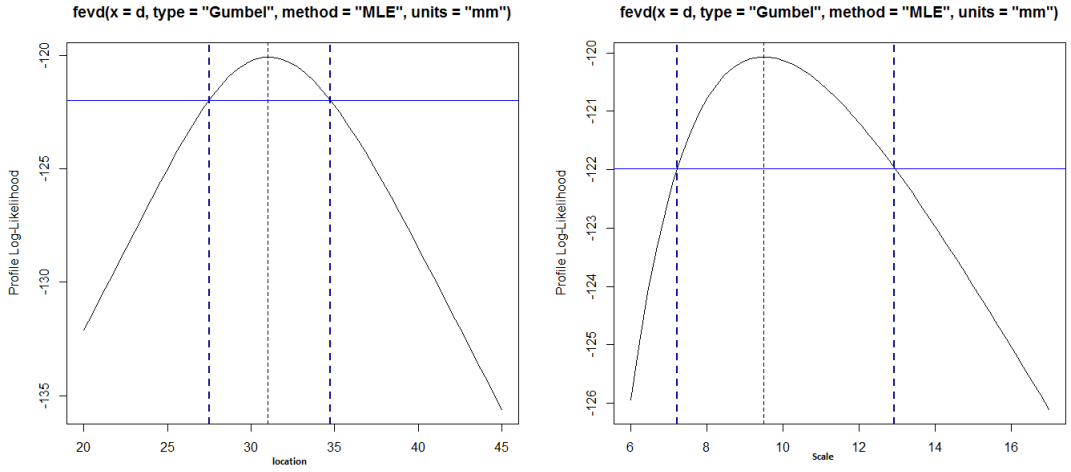

Figure 3.6: Profile log-Likelihood of GEV_0 parameters by MLE.

Graphic Diagnostics of GEV_0

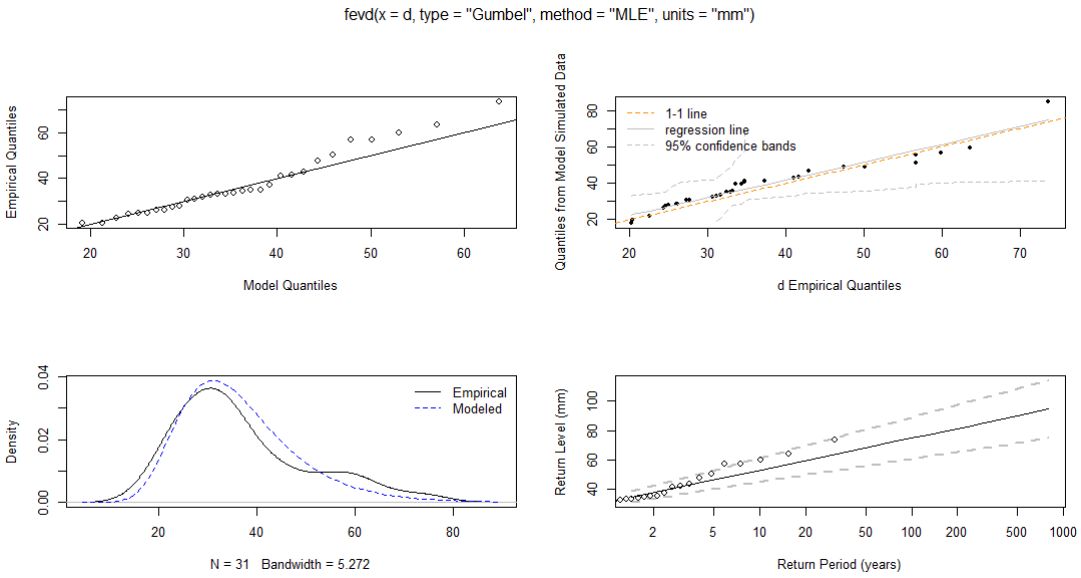

Figure 3.7: Graphic diagnostic of GEV_0 parameters by MLE.

Both GEV_ξ and GEV_0 models have similar estimated return level curves, but the confidence intervals are much wider for the GEV_ξ model, especially for long return periods. This signifies that Gumbel model is preferable.

Return levels and return periods estimates

After model validation, we can estimate return level at selected return periods for yearly maximum rainfalls. Table 3.7 shows the results, the 95% confidence intervals are included. From Table 3.7, the return level estimates increases as the return period increases because the matter is about rare events.

Selected period (year)	10	20	50	100
Return level (mm)	52.40	59.23	68.08	74.71
CI	(44.38;60.39)	(49.39;69.07)	(55.78;80.37)	(60.54;88.87)

Table 3.7: Return level estimation at selected return periods (GEV_0 model).

Table 3.8 shows the converse for the highest recorded data,

Highest levels (mm)	73.7	63.6	59.9	56.7
Return period (year)	89.98	31.39	21.42	15.44

Table 3.8: Return periods estimation

3.3.5 Modeling using Generalized Pareto (GP) distribution

It is a method based on the approximation of the distribution of excesses to a threshold by the Generalized Pareto Distribution.

Threshold selection

In a statistical framework, the choice of the threshold u is very important because it induces great variability in the estimation of extreme quantiles and parameters of the excesses distribution, it is determined graphically by the Mean Excess Plot. For our data we suspect the thresholds $u_1 = 32.5$ and $u_2 = 50.5$. (see figure 3.8), and in

order to confirm, we used the 2^{nd} method *"select threshold by estimate the model at a range of thresholds"* (or parameter stability plot). According to the figure 3.8, the selected threshold of $u = 32.5mm$ seems reasonable.

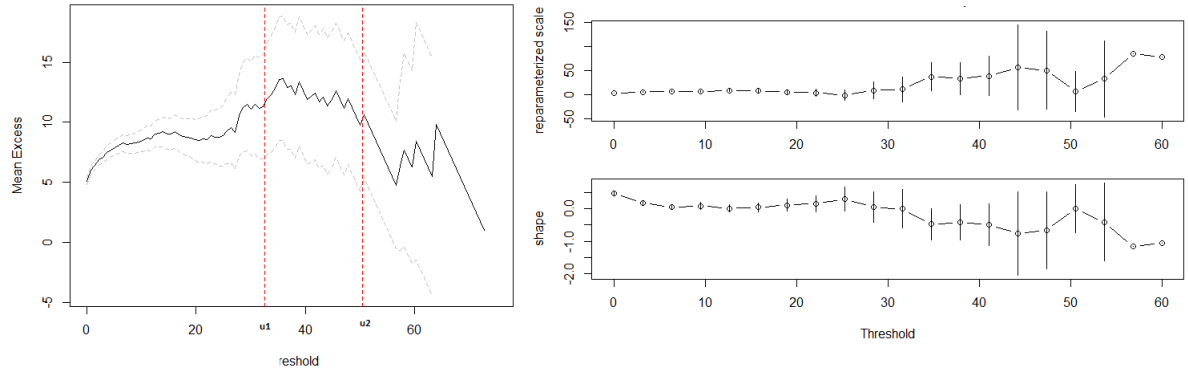

Figure 3.8: Mean Exess Plot (left); Estimate model at a range of thresholds (right).

Comments

Mean Excess Plot The graph appears to curve from $u = 0$ to $u = u_1 = 32.5mm$ beyond which it is approximately linear, but the approximate linearity is very clear after the value $u = u_2 = 50.5mm$. However, there are just 6 exceedances over the threshold u_2, too ew to make meaningful inferences.

meter Stability Plot Perturbations of parameter estimates are see to be small relative to sampling errors (confidence intervals) beyond the threshold $u_1 = 32.5mm$.

The next graph descibes the exceedances to the threshold selected.

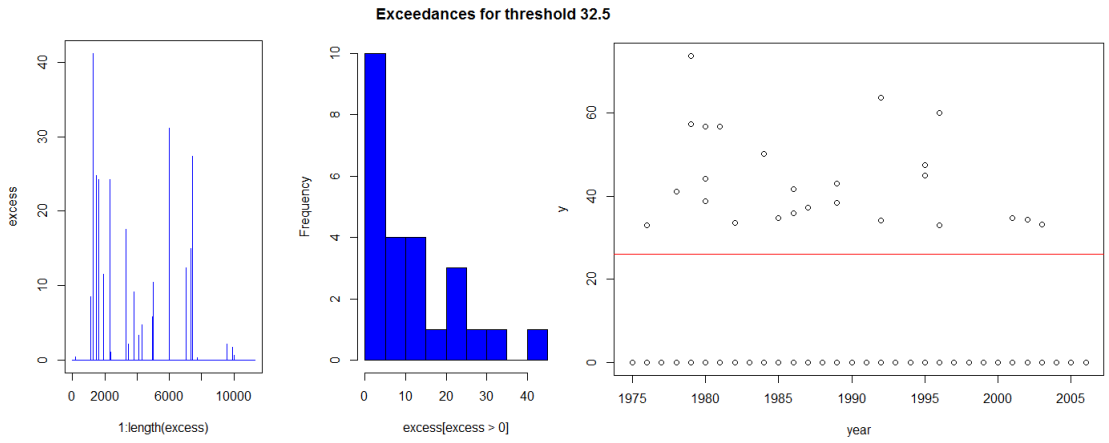

Figure 3.9: Threshold exceedances ($u = 32.5\ mm$)

Parameter estimation of GPD_ξ ($\xi \in \mathbb{R}$)

We estimate GPD parameters ξ and σ with MLE and PWD methods.

	Shape parameter		Scale parameter	
Method	$\hat{\xi}$	SE	$\hat{\sigma}$	SE
MLE	1.1×10^{-7}	0.32	11.3	4.16
PWM	0.06		9.82	

Table 3.9: Parameter estimates of GPD_ξ.

	Shape parameter $\hat{\xi}$	Scale parameter $\hat{\sigma}$
CI. Normal Approximation	(-0.64;0.64)	(5.67;11.43)
CI. Log-Likelihood Profile	(-0.27;0.13)	(5.70;23.85)

Table 3.10: 95% Confidence Intervals of GPD_ξ parameters.

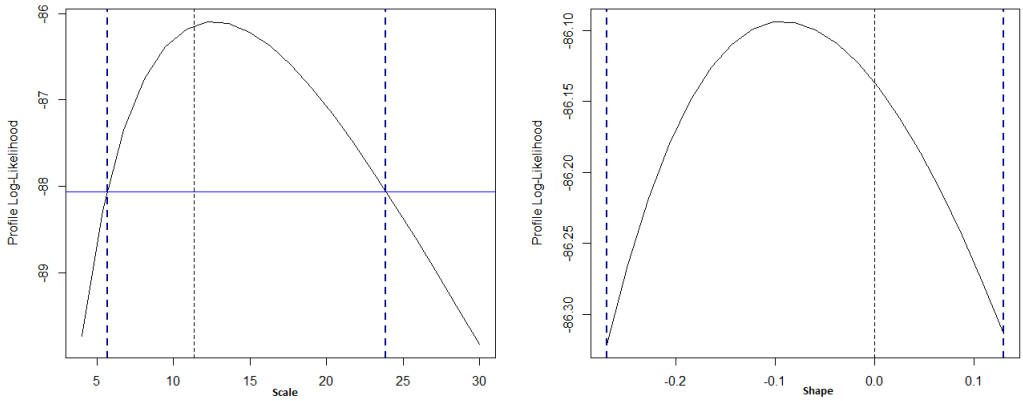

Figure 3.10: Profile log-Likelihood of GPD_ξ parameters by MLE.

The MLE therefore corresponds to an unlimited distribution (exponential distribution) (as $\xi \approx 0$) and a reasonably strong argument is that the confidence interval contains the value 0, for confirmation, *Likelihood ratio test* is given by the table 3.11.

test	α	Chi square ctrical value	df	p-value
-0.0069	0.05	3.84	1	1

Table 3.11: Likelihood Ratio test GPD_0 vs GPD_ξ.

Graphic Diagnostic of GPD_ξ and GPD_0 (Exponential distribution)

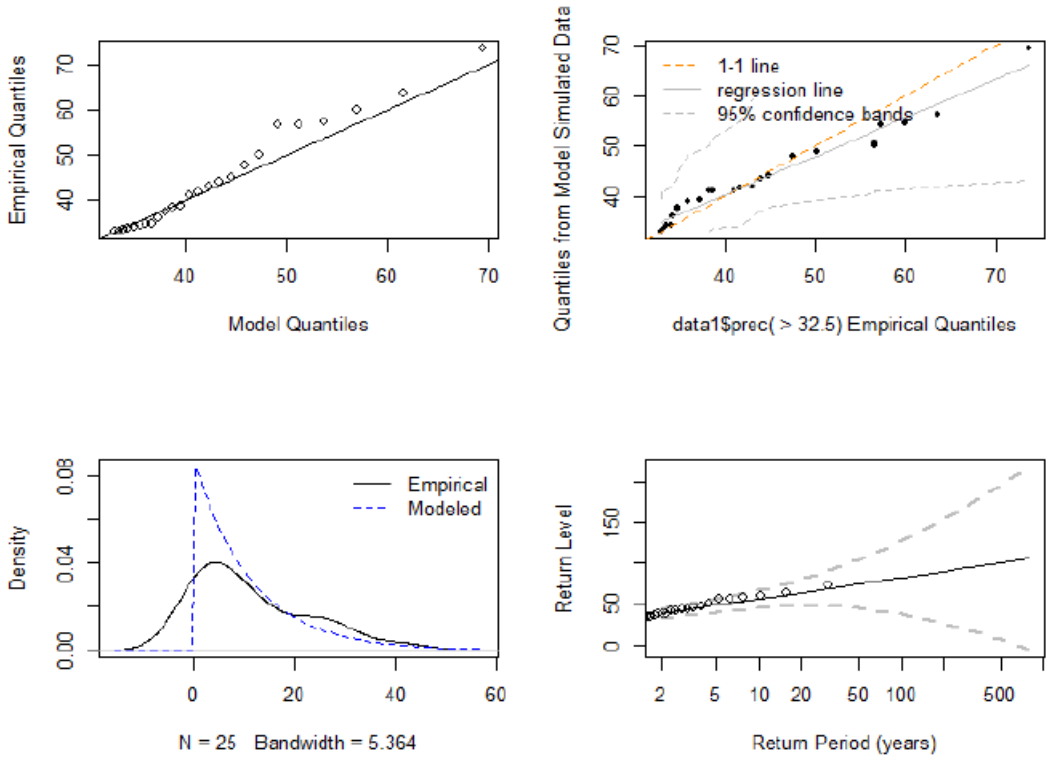

Figure 3.11: Graphic diagnostic of GPD_ξ by MLE

	Scale parameter	
Method	$\hat{\sigma}$	SE
MLE	11.54	2.31

Table 3.12: Scale parameter estimation of GPD_0

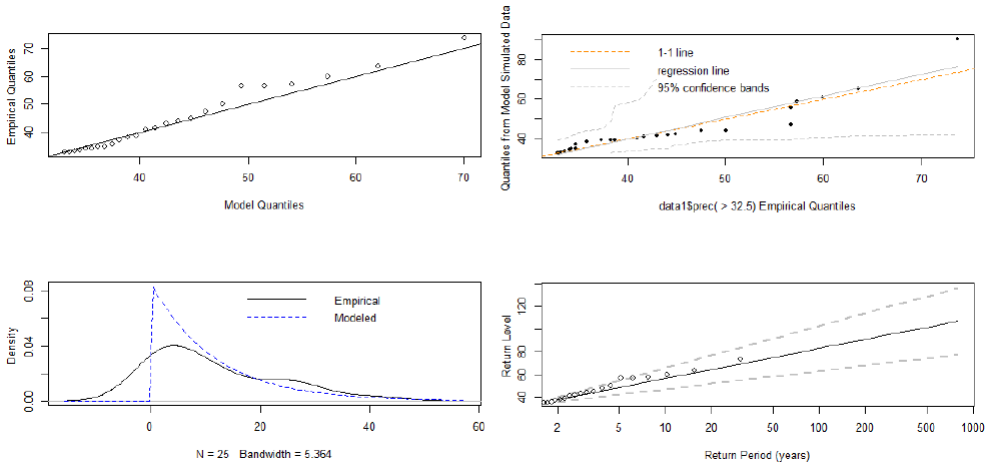

Figure 3.12: Graphic diagnostic of GPD_0 by MLE.

Parameter estimation of GPD_0 (*Exponential Model*)

Results are presented in Tables 3.12, 3.13

	Scale parameter $\hat{\sigma}$
CI. Normal Approximation	(7.01;16.06)
CI. Log-Likelihood Profile	(8.03;17.50)

Table 3.13: 95% Confidence Intervals of GPD_0 parameter.

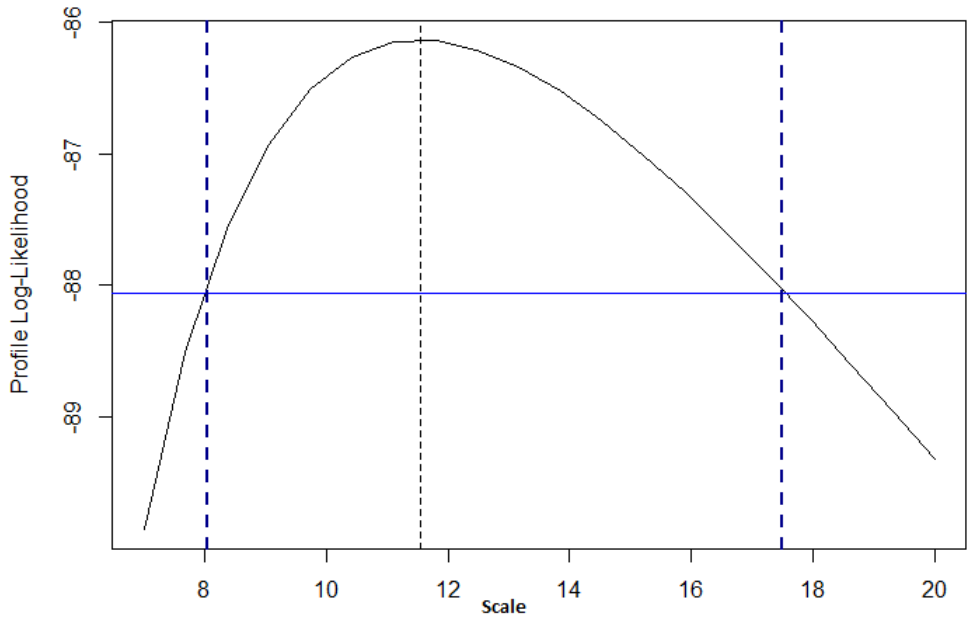

Figure 3.13: Profile log-Likelihood of GPD_0 parameters by MLE.

Return levels and return periods estimates

Return levels are estimated by MLE method for 10, 20, 50, and 100 years. The return periods and the 95% confidence intervals are shown in table (3.14) below.

Selected period (year)	10	20	50	100
Return level	52.39	59.23	68.07	74.70
CI	(44.39;60.39)	(49.38;69.07)	(55.78;80.37)	(60.54;88.87)

Table 3.14: Return level estimation at selected return periods GPD_0 model

Highest levels (mm)	73.7	63.6	59.9	56.7
Return period (year)	89.98	31.39	21.42	15.44

Table 3.15: Return periods estimation.

The corresponding return period estimate for the level $73.7mm$ is approximately $90\,years$ which means that the level $73.7mm$ will be exceeded once every $90\,years$ or the probability that the extreme value $73.7mm$ will be observed is 1/90.

3.3.6 Dependent Data Issue

The asymptotic results (GEV and GPD fitting) used in this application have assumed a sequence of independent random variables. However in real world, temporal independence is usually an unrealistic hypothesis, so extreme value data present the issue of dependence on covariate effects, short-term dependence and long-range dependence, this last is negligible so far as the standard asymptotic limits are concerned (the standard asymptotic limits still applied for yearly maxima) contrary to Threshold Excesses method which needs to be adapted because the extremes have some tendency to cluster, violating the assumption of independence among the individual excesses.

The most commonly used method for dealing with the problem of dependent exceedances in the threshold exceedance model is declustering, it means a filtering of the dependent observations to have a set of threshold excesses that are approximately independent.(for more details see [34]).

3.4 Conclusion

In our study, the maximum annual rainfall at Khemis-Miliana from 1975 to 2006 is modeled using the Extreme Value Distribution (GEV) to control and predict the behavior of rainfall and using also the GPD approach. The maximum likelihood and probability weighted moments methods are used to estimate the parameters of the models. Both approaches has its advantages, the GPD distribution has a big set of raw data because it uses all the maxima of the year (exceeding the threshold), while the GEV distribution uses a single maximum rainfall per year.It was found that the stationary Gumbel model (no trend) is more appropriate for the Khemis-Miliana station. According to the results, the two methods provide almost the same results; whether it is at the level of the estimation of the parameters of the models, or at the level of estimates of return levels and periods, we can interpret this by the absence of problems caused by autocorrelation and seasonality (non-stationarity), these latter are usually overcome by the POT method.

We can, not only fit extreme rainfalls for one station by a GEV or GP distribution, but also for several homogeneous regions (each region contains several stations) by a single GEV or GP using L-moment method, which is so called: *Regionalization of Extreme Rainfalls*.

Chapter 4

Estimating the Conditional Mean of Heavy-Tailed Distribution under Random Right censoring

Abstract

In this paper, we construct a new estimator of the mean of a heavy-tailed distribution under right random censored data in presence of covariates is proposed by combining the generalized Kaplan–Meier estimator before a threshold and a parametric model (GPD) which approximate the excesses over the threshold in order to overcome the bad behaviour of K-M estimator in the heavy-tail of distribution, and its asymptotic normality is established.

Keywords: Heavy-tailed distribution, Generalized Kaplan-Meier estimator, General Pareto Distribution (GPD), Random right censoring, conditional mean estimate.

AMS 2000 Subject Classifications

4.1 Introduction

Estimating the mean of heavy-tailed distributions F with tail index γ with complete data and absence of covariates encountered infinite second order mean problem where the simple sample mean estimator have a nonnormal limit in this case (see [66]. In order to overcome this problem, many authors gave contributions in this context; L. Peng in [120] proposed an alternative estimator whose limiting distribution, under a second order condition, is normal for any $\gamma < 1$ by dividing the mean into three parts. Under the same condition, L. Peng in [121] interested in the

same subject, his idea was to estimate the tail part parametrically and the middle part nonparametrically. Johansson [96] suggested, like L. Peng [121], a procedure that separated tail data which starting at some large level and approximated exesses over this level by GPD distribution in parametric framework. In case of censored data, Sander [131] obtained estimators for restricted mean untill a fixed finite time and indicated that is extremely difficult to obtain the distribution theory for mean life time estimator. Later, V. Susarla and J. Van Ryzin [145] gave a class of estimators of the mean survival time by replacing infinite time by time's sequence increasing to infinity. W. Stute [140] introduced the CLT which is a powerfull tool to establish asymptotic behaviours of mean life time estimator but this latter, as mentioned in Stute ([140], p.249), is slightly non-robust under right heavy censoring, in this case the KM estimator does not vanish to zero in the support tail. Our contribution takes place in case of censored data and presence of covariates where we try to combinate above works by proposing an alternative mean estimator ensuring the asymptotic normality property.

Let $(X_i; Y_i)$, $i = 1,...,n$ be independent copies of the random pair (X, Y) where Y is a non-negative random variable and $X \in \mathcal{X}$ (with $\mathcal{X}$ some bounded set of $\mathbb{R}^p$) is a p-dimensional covariate. We assume that Y can be right-censored by a non-negative random variable C. Thus we really observe independent triplets (X_i, δ_i, Z_i), $i = 1,...,n$, where $Z_i = \min(Y_i, C_i)$, $\delta_i = 1_{\{Yi \leq C_i\}}$ and 1_A is the indicator function of the event A. The random variable C is defined on the same probability space $(\Omega, \mathcal{F}, \mathbf{P})$ as Y. We assume that $C_1,...,C_n$ are independent of each other and that Y and C are independent given X. Let $F(.|x)$ and $G(.|x)$ denote the conditional cumulative distribution functions of Y and C given $X = x$, respectively. Let also $\bar{F}(.|x) = 1 - F(.|x)$ and $\bar{G}(.|x) = 1 - G(.|x)$ be the conditional survival functions of Y and C given $X = x$.

In this paper, we focus on heavy tails. Precisely, we assume that the conditional survival functions of Y and C given $X = x$ are

$$\bar{F}(u|x) = c_1(x)u^{-1/\gamma_1(x)}(1 + u^{-\delta_1}L_1(u|x))$$

and

$$\bar{G}(.|x) = c_2(x)u^{-1/\gamma_2(x)}(1 + u^{-\delta_2}L_2(u|x)),$$

where $\gamma_1(x)$ and $\gamma_2(x)$ are unknown positive continuous functions of the covariate x and for x fixed, $L_1(.|x)$ and $L_2(.|x)$ are slowly varying functions at infinity. This amounts to saying that $F(.|x)$ and $G(.|x)$ are regularly varying functions at infinity

with index $-1/\gamma_1(x)$ and $-1/\gamma_2(x)$ respectively. Condition (C1) also amounts to assuming that the conditional distributions of Y and C given $X = x$ are in the Fréchet maximum domain of attraction. In what follows, the functions $\gamma_1(.)$ and $\gamma_2(.)$ are referred to as conditional extreme-value index functions.

Remark 4.1 *By conditional independence of Y and C, then by straightforward proof, the conditional cumulative distribution function $H(.|x)$ of Z given $X = x$ is also* ***heavy-tailed****, with conditional extreme-value index*

$$\gamma(x) = \gamma_1(x)\gamma_2(x)/(\gamma_1(x) + \gamma_2(x)).$$

Here, $Z_{1:n} \leq \cdots \leq Z_{n:n}$ are the ordered $Z_i's$, and $\delta_{[i;n]}$ denote the indicator associated to $Z_{i:n}$, that is, the ith concomitant. Statistical properties of F_n^{KM} are well known (see, e.g., [4]). It is clearly seen from equation.(4.1) that the Kaplan–Meier estimator is a step function with jump points located at the uncensored observations, the jump size being a nondecreasing function of the $Z-$rank. In the uncensored case, F_n^{KM} reduces to the ordinary empirical cdf of the sample.

Let $x \in \varkappa$, so we take only observations (δ_i, Z_i) such that $X = x$, and denote m_n^x as the number of such observations. Let $Z_{(1)}^x, ..., Z_{(m_n^x)}^x$ be the ordered values of Z for these observations and let $\delta_{(1)}^x, ..., \delta_{(m_n^x)}^x$ be the corresponding $\delta's$ (that is, $\delta_{(i)}^x = \delta_j^x$ if $Z_{(i)}^x = Z_j$).

An efficient non-parametric estimator of the tail distribution of Y given $X = x$ based on the sample $(Z_{(1)}^x, \delta_{(1)}^x), ..., (Z_{(m_n^x)}^x, \delta_{(m_n^x)}^x)$, is given by the **time-honoured** Kaplan-Meier ([97]) product-limit estimator defined by:

$$\begin{aligned}\widehat{\bar{F}}_{m_n^x}^{KM}(y|x) = 1 - \widehat{F}_{m_n^x}^{KM}(y|x) &= \prod_{i=1}^{m_n^x}\left(1 - \frac{\delta_{(i)}^x}{m_n^x - i + 1}\right)^{\mathbf{1}_{\left\{Z_{(i)}^x \leq y\right\}}} \\ &= \prod_{i=1}^{m_n^x}\left(\frac{m_n^x - i}{m_n^x - i + 1}\right)^{\delta_{(i)}^x \mathbf{1}_{\left\{Z_{(i)}^x \leq y\right\}}} \end{aligned} \tag{4.1}$$

If all $\delta's$ equal 1, i.e., if there is no censorship, then $\widehat{\bar{F}}_{m_n^x}^{KM} = \widehat{\bar{F}}_{m_n^x}$.

4.2 Construction of estimate

The aim of this paper is to propose an asymptoticaly normal estimator for the mean of Y given $X = x$ ($\mathbf{E}(Y|X = x)$)

The standard estimate of $\mathbf{E}(Y|X=x)$ is

$$\mu(x) = \int_0^\infty \bar{F}_n(y|x)dy$$

building on this, we propose an estimate of the form (**for incovenients of KM estimate, see Reiss and Thomas [125] and** M. S. Pepe and T. R.Fleming [122])

$$\hat{\mathbf{E}}(Y|X=x) := \hat{\mu}_{KP}(x) + \hat{\mu}_T(x) := \int_0^{u_n(x)} \widehat{\bar{F}}^{KP}_{m_n^x}(y|x)dy + \int_{u_n(x)}^\infty y d\widehat{F}_n(y|x),$$

where $\hat{\mu}_T$ is the part of $\hat{\mathbf{E}}(Y|X=x)$ originating from the tail of distribution. The tail is assumed to start at some level $u_n(x)$, which in the analysis will be assumed to tend to infinity. Next, we will each term separately.

In order to obtain $\hat{\mu}_{KP}(x)$, we use Kaplan-Meier (K-M) integrals introduced by W. Stute ([143], [141], [142]) by putting the function φ equal to identity, which yield to:

$$\hat{\mu}_{KP}(x) = \int_0^{u_n(x)} y d\left[F^{KP}_{m_n^x}(y|x)\right] = \sum_{i=1}^{m_{n,u_n(x)}} W_{i,m_{n,u_n(x)}} Z^x_{(i)}, \tag{4.2}$$

(we can use $1_{\left\{Z^x_{(i)} \le u_n(x)\right\}}$) and summing till m_n^x), where the weight attached to $Z^x_{(i)}$ equals

$$W_{i,m_{n,u_n(x)}} = \frac{\delta^x_{(i)}}{m^x_{n,u_n(x)} - i + 1} \prod_{j=1}^{m^x_{n,u_n(x)}-1} \left(\frac{m_n^x - j}{m_n^x - j + 1}\right)^{\delta^x_{(j)}}.$$

(Beirlant and goegebeur ([13])) To estimate mean over threshold $u_n(x)$, we will first introduce conditional distribution of the excesses over the threshold $u_n(x)$ and then approximating it by GPD distribution. Let

$$F_{u_n(x)}(z|x) = P(Y - u_n(x) \le z | Y \ge u_n(x), X = x)$$

be conditional distribution of the excesses over the threshold $u_n(x)$. By definition

$$\bar{F}_{u_n(x)}(z|x) = \frac{\bar{F}_{u_n(x)}(z + u_n(x)|x)}{\bar{F}_n(u_n(x)|x)}$$

So, one can write

$$\begin{aligned}\hat{\mu}_T(x) &:= \int_{u_n(x)}^{\infty} \widehat{\bar{F}}_n(y|x)dy \\ &= \int_{u_n(x)}^{\infty} \widehat{\bar{F}}_n(u_n(x)|x)\widehat{\bar{F}}_{u_n(x)}(y-u_n(x)|x)dy \\ &= \widehat{\bar{F}}_n(u_n(x)|x) \underbrace{\int_{u_n(x)}^{\infty} \widehat{\bar{F}}_{u_n(x)}(y-u_n(x)|x)dy}_{\| \\ \hat{I}_n(x)}\end{aligned}$$

By generalized Kaplan-Meier estimate, we have

$$\widehat{\bar{F}}_n(u_n(x)|x) = \prod_{i=1}^{n}\left[1 - \frac{\delta_{(i)}^x 1_{\left\{Z_{(i)}^x \le u_n(x)\right\}}}{n-i+1}\right]$$

It remains the approximation of $F_{u_n(x)}(.|x)$ by GPD which has the form

$$G_{\gamma_1(x),\sigma_1(x)}(z|x) = 1 - \left(1 + \frac{\gamma_1(x)}{\sigma_1(x)}z\right)^{-\frac{1}{\gamma_1(x)}}.$$

Now results given by Balkema and de Haan ([9]) and Pickands ([123]) ensure, for large values of $u_n(x)$, $F_{u_n(x)}(z) \approx G_{\gamma_1(x),\sigma_1(x)}(z|x)$ in the sense that

$$\lim_{u_n(x) \nearrow z_F(x)} \sup_{0<z<z_F(x)-u_n(x)} \left|F_{u_n(x)}(z|x) - G_{\gamma_1(x),\sigma_1(x)}(z|x)\right| = 0$$

where $z_F(x)$ is the right end point of $F(.|x)$.

Therefore integral $\hat{I}_n(x)$ can be written as

$$\begin{aligned}\hat{I}_n(x) &= \int_{u_n(x)}^{\infty} \bar{G}_{\gamma_1(x),\sigma_1(x)}(y-u_n(x)|x)dy \\ &= \int_{u_n(x)}^{\infty} y\frac{1}{\hat{\sigma}_1(x)}\left[1 + \frac{\hat{\gamma}_1(x)}{\hat{\sigma}_1(x)}(y-u_n(x)\right]^{-\frac{1}{\hat{\gamma}_1(x)}-1} dy.\end{aligned}$$

By simple calculus, it yields (% it remains discussion of integral's convergence according to $\gamma_1(x)$ values%)

$$\hat{I}_n(x) = u_n(x) + \frac{\hat{\sigma}_1(x)}{1-\hat{\gamma}_1(x)}.$$

We are now able to express our estimte $\hat{\mu}_T(x)$ by

$$\hat{\mu}_T(x) = \prod_{i=1}^{n}\left[1 - \frac{\delta_{(i)}^x 1_{\{Z_{(i)}^x \le u_n(x)\}}}{n-i+1}\right]\left[u_n(x) + \frac{\hat{\sigma}_1(x)}{1-\hat{\gamma}_1(x)}\right]. \tag{4.3}$$

The next step is to approximate $(\hat{\gamma}_1(x), \hat{\sigma}_1(x))$ using P.O.T. method by adapting likelihood function to censorship (see Ndao [114]), Toulemonde ([147]) and Beirlant et al. ([16])). for this purpose, we will use approach proposed by beirlant ([16]) which consists of solving the ML-equations based on a one-step Newton-Raphson approximation (as discussed, for instance, in Lehmann [104]) and this to avoid the difficulty of the asymptotic normality in the case of censoring. According to Beirlant ([16]), this approach consists to adapt the likelihood to this purpose. This latter method relies on the results given by Balkema and de Haan ([9]) and Pickands ([123]), formally, let $E_i^x = Z_i^x - u(x)$, given $Z_i^x > u$, over a threshold u when $u \to \tau_{F,}$ in this case the adapted likelihood function is the following: (cf. Andersen et al. [4], p. 411)

$$L(\gamma_1(x), \sigma_1(x)) = \prod_{i=1}^{N_{u(x)}}\left[g_{\gamma_1(x),\sigma_1(x)}(E_i^x|x)\right]^{\delta_{(i)}^x}\left[1 - G_{\gamma_1(x),\sigma_1(x)}(E_i^x|x)\right]^{1-\delta_{(i)}^x},$$

where $1 - G_{\gamma_1(x),\sigma_1(x)}(z|x) = \left(1 + \frac{\gamma_1(x)}{\sigma_1(x)}z\right)^{-\frac{1}{\gamma_1(x)}}$ and $g_{\gamma_1(x),\sigma_1(x)}$ is the associated density.

Beirlant ([13], [16]) and Toulemonde ([147]) defined new estimators adapted to censoring for $(\gamma_1(x), \sigma_1(x))$ denoted in the sequel by $(\hat{\gamma}^{(c,os)}_{Z^x,u(x)}(x), \hat{\sigma}^{(c,os)}_{Z^x,u(x)}(x))$ called the one-step estimators, furthermore they have established its asymptotic normality after suitable normalization and have showed that at finite distance these estimators behave in a similar manner to those of maximum likelihood, however asymptotic normality of these latter is still an open problem. We are now ready to give an alternative estimate, $\hat{M}(x)$, of $\mathbf{E}(Y|X=x) := M$, as follow:

$$\begin{aligned}\hat{M}(x) :=& \sum_{i=1}^{m_{n,u}^x}\frac{\delta_{(i)}^x}{m_{n,u}^x - i + 1}\prod_{j=1}^{m_{n,u}^x - 1}\left(\frac{m_n^x - j}{m_n^x - j + 1}\right)^{\delta_{(j)}^x} Z_{(i)}^x 1_{\{Z_{(i)}^x \le u\}} + \\ &+ \prod_{i=1}^{n}\left[1 - \frac{\delta_{(i)}^x 1_{\{Z_{(i)}^x \le u\}}}{n-i+1}\right]\left[u(x) + \frac{\hat{\sigma}^{(c,os)}_{Z^x,u}(x)}{1-\hat{\gamma}^{(c,os)}_{Z^x,u}(x)}\right],\end{aligned} \tag{4.4}$$

where, by using de Haan and Ferreira ([45], Theorem 1.2.5, p.21), $\hat{\sigma}^{(c,os)}_{Z^x,u(x)}(x)$ can be estimated by

$$\hat{\sigma}^{(c,os)}_{Z^x,u(x)}(x) = u(x)\hat{\gamma}^{(c,os)}_{Z^x,u}(x).$$

4.3 Main results

4.3.1 Assumptions

(**C1**) $\bar{F}(y|x) = c_1(x)y^{-1/\gamma_1(x)}(1+y^{-\delta_1}L_1(y|x))$ and $\bar{G}(y|x) = c_2(x)y^{-1/\gamma_2(x)}(1+y^{-\delta_2}L_2(y|x))$, where $\gamma_1(x)$ and $\gamma_2(x)$ are unknown positive continuous functions of the covariate x and for x fixed, $L_1(.|x)$ and $L_2(.|x)$ are slowly varying functions at infinity, $c_i(x) and \delta_i \; for i = 1, 2$ are constants.

(**C2**) $F(.|x)$ and $G(.|x)$ are continuous distributions

(**C3**) $\bar{F}(.|x)$ and $\bar{G}(.|x)$ are Lipschitzian functions and their first derivatives exist.

(**C4**) *Conditional first order conditions: We denote by $U_F(.|x)$ (resp.by $U_G(.|x)$)(supposed continuous) the tail quantile function of F* (resp. of G) , that is $U_F(z|x) = F^{\leftarrow}(1 - \frac{1}{z}|x) = \inf\left\{y : F(y|x) \geq 1 - \frac{1}{z}\right\}$ (resp. $U_G(z|x) = G^{\leftarrow}(1-\frac{1}{z}|x) = \inf\left\{y : G(y|x) \geq 1 - \frac{1}{z}\right\}$). We assume that there exists two conditional positive auxiliary functions, a_F and a_G, such that

$$\lim_{z\to\infty} \frac{U_F(zs|x) - U_F(z|x)}{a_F(z|x)} = \int_1^s v^{\gamma_1(x)-1} dv =: h_{\gamma_1(x)}(s|x) \text{ for } s > 0$$

$$\lim_{z\to\infty} \frac{U_G(zs|x) - U_G(z|x)}{a_G(z|x)} = \int_1^s v^{\gamma_2(x)-1} dv =: h_{\gamma_2(x)}(s|x) \text{ for } s > 0$$

(**C5**) *Conditional second order conditions:*

$$\frac{U_F(zs|x) - U_F(z|x)}{a_F(z|x)} - h_{\gamma_1(x)}(s|x) \sim a_{2,F}(z|x)k_F(s|x), \; z \to \infty$$

$$\frac{U_G(zs|x) - U_G(z|x)}{a_G(z|x)} - h_{\gamma_2(x)}(s|x) \sim a_{2,G}(z|x)k_G(s|x), \; z \to \infty$$

where $a_{2,F}$ and $a_{2,G} \to 0$ are regulary varying functions with respective indices $\rho_1(x) \leq 0$ and $\rho_2(x) \leq 0$, and

$$k_F(s|x) = A_F h_{\gamma_1(x)+\rho_1(x)}(s|x) + c_F \int_1^s t^{\gamma_1(x)-1} h_{\rho_1(x)}(t|x)dt$$

$$k_G(s|x) = A_G h_{\gamma_2(x)+\rho_2(x)}(s|x) + c_G \int_1^s t^{\gamma_2(x)-1} h_{\rho_2(x)}(t|x)dt$$

for suitable constants A_F, $A_{G,}$ c_F and c_G..

Condition on bias

$$\sqrt{N_u^x}B_{1,u}(\gamma_1(x),\sigma_1(x)) \underset{u\to\infty}{\to} 0$$
$$\sqrt{N_u^x}B_{2,u}(\gamma_1(x),\sigma_1(x)) \underset{u\to\infty}{\to} 0$$

with

$$B_{1,u}(\gamma_1(x),\sigma_1(x)) = \frac{1}{\gamma_1(x)\sigma_1(x)}\frac{1}{\overline{H}(u|x)}\int_u^\infty \frac{\overline{H}(z|x)}{1+\frac{\gamma_1(x)}{\sigma_1(x)}(z-u)}dz$$
$$-\frac{1}{\gamma_1(x)\sigma_1(x)}\frac{1}{\overline{H}(u|x)}\int_u^\infty \frac{\overline{H}(z|x)}{(1+\frac{\gamma_1(x)}{\sigma_1(x)}(z-u))^2}dz$$
$$-\frac{1}{\sigma_1(x)}\frac{1}{\overline{H}(u|x)}\int_u^\infty \frac{\overline{H}^u(z|x)}{(1+\frac{\gamma_1(x)}{\sigma_1(x)}(z-u))^2}dz$$

$$B_{2,u}(\gamma_1(x),\sigma_1(x)) = -\frac{\overline{H}^u(u|x)}{\overline{H}(u|x)}\frac{1}{\sigma_1(x)}\frac{1}{\overline{H}(u|x)}\int_u^\infty \frac{\overline{H}(z|x)}{(1+\frac{\gamma_1(x)}{\sigma_1(x)}(z-u))^2}dz$$
$$+\frac{1}{\sigma_1(x)}\frac{1}{\overline{H}(u|x)}\int_u^\infty \frac{\overline{H}^u(z|x)}{(1+\frac{\gamma_1(x)}{\sigma_1(x)}(z-u))^2}dz$$

where $H^u(s|x) = P(Z^x > s, \delta^x = 1|X = x)$ called the conditional sub-distribution of noncensored observations.

Theorem 4.1 *Suppose the conditions (C1)-(C6) hold.*

$$\frac{\sqrt{m_n(x)}}{\alpha_n(x)\sqrt{k_n(x)}}(\hat{M}(x) - M(x)) \xrightarrow{d} N(0,1),$$

where

$$\alpha_n^2(x) = Var(\int_0^u z dF(z|x))$$

$F(.|x)$ is a continuous mean-zero Gaussian process (see (4.5) and (4.6))

$$k_n(x) = 1 + \frac{\left(\bar{H}(u|x)\right)^{2\frac{\gamma(x)}{\gamma_1(x)}-1}}{\alpha_n^2(x)}\times$$
$$\left[\frac{\gamma(x)}{\gamma_1(x)}\left(u+\frac{\sigma_1(x)}{1-\gamma_1(x)}\right)^2 + \frac{\sigma_1^2(x)(1+\gamma(x))^2}{(1-\gamma_1(x))^4}\left(\frac{\gamma_1(x)}{\gamma(x)}\right)^3 + \frac{2\sigma_1^2(x)(1+\gamma(x))}{(1-\gamma_1(x))^2}\frac{\gamma_1(x)}{\gamma(x)}\right]$$
$$= O_+(1).$$

4.3.2 Proofs of main results

First we present some results which will be used in the proof of the Theorem 4.1.

Breslow and Crowly ([27]) have shown that

$$F_{m_n(x)}(z|x) := \sqrt{m_n(x)}\Big(F_{m_n(x)}(z|x) - F(z|x)\Big) \overset{weakly}{\rightarrow} F(z|x), 0 \leq z \leq u, \tag{4.5}$$

where $F(.|x)$ is a continuous mean-zero Gaussian process with covariance

$$\Gamma_x(t,s) = \bar{F}(t|x)\bar{F}(s|x)\int_0^{s\wedge t} \frac{dH^u(y|x)}{(1-H(y|x))^2} \tag{4.6}$$

Proposition 4.1 *Under conditions (C1) and (C3) and let $x \in \varkappa$ and $u \rightarrow \infty$. Then*

$$\bar{F}(u|x) = P(X_1 > u|x) = O_+(1),$$

and

$$\alpha_n^2(x) = Var(\int_0^u z dF(z|x)) = O_+(u^{4+1/\gamma_2(x)+(1/\gamma_2(x)-1/\gamma_1(x))1_{\{\gamma_1(x)\geq\gamma_2(x)\}}}),$$

where $\varepsilon_n = O_+(a_n)$ denotes a sequence such that ε_n/a_n is bounded away from zero and infinity.

Proof of the proposition 4.1. We have

$$\bar{F}(y|x) = c_1(x)y^{-1/\gamma_1(x)}(1 + y^{-\delta_1}L_1(y|x)) = c_1(x)y^{-1/\gamma_1(x)}(1+o(1))$$

since $y^{-\delta_1}L_1(y|x)$ is non-increasing and $L_1(y|x)$ is locally bounded in $[y_0,\infty)$ for some $y_0 \geq 0$. So one can deduce that

$$\bar{F}(y|x) = y^{-1/\gamma_1(x)}O_+(1).$$

According to Sander ([131], Corollary 1, p.7) or Escobar-Bach ([63], Theorem 4, p.7);

$$\alpha_n^2(x) = \int_0^u\int_0^u std\left[\underbrace{\bar{F}(t|x)\bar{F}(s|x)\int_0^s \frac{dH^u(y|x)}{(1-H(y|x))^2}}_{\Gamma_x(t,s)}\right]; s \leq t.$$

Furthermore, under condition (C1) and by Ndao et al. ([116], proof of Lemma A.1., p13) we can write

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

which yields to

$$\Gamma_x(t,s) \sim \frac{\gamma_1(x)}{\gamma_2(x)} \frac{\bar{F}(t|x)}{\bar{G}(s|x)}$$

Assuming that condition (C3) is satisfied, result follows from straightforward calculations. ■

Theorem 4.4.1 in Ndao ([114], p.126, proof p.129) gives us the asymptotic distribution of the tail parameters $(\hat{\gamma}^{(c,os)}_{Z^x,u(x)}(x), \hat{\sigma}^{(c,os)}_{Z^x,u(x)}(x))$ as

$$\sqrt{N_u^x}\begin{pmatrix} \hat{\gamma}^{(c,os)}_{Z^x,u(x)}(x) - \gamma_1(x) \\ \hat{\sigma}^{(c,os)}_{Z^x,u(x)}(x) - \sigma_1(x) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma\right),$$

where N_u^x is the number of absolute excesses over u given the covariable value x, and

$$\Sigma = \begin{pmatrix} \left(\frac{\gamma_1(x)}{\gamma(x)}\right)^3 (1+\gamma(x))^2 & -\left(\frac{\gamma_1(x)}{\gamma(x)}\right)^2 (1+\gamma(x))\,\sigma_1(x) \\ -\left(\frac{\gamma_1(x)}{\gamma(x)}\right)^2 (1+\gamma(x))\,\sigma_1(x) & 2\frac{\gamma_1(x)}{\gamma(x)}(1+\gamma(x))\,\sigma_1^2(x) \end{pmatrix}, \tag{4.7}$$

under the assumptions (C4-C6).

The distribution of $\hat{\mu}_{KP}(x)$ is given by the following lemma:

Lemma 4.1 *Let $\mu(x) = \int_0^u z dF(z|x)$ and $\alpha_n^2(x) = Var(\int_0^u z dF(z|x))$, with $u_n = O_+(n^{\alpha\gamma_1(x)})$ for some $\alpha \in (0,1)$, where $\delta > 0$. Then*

$$\frac{\sqrt{m_n(x)}}{\alpha_n(x)}(\hat{\mu}_{KP}(x) - \mu) \xrightarrow{d} N(0,1), as\ n \to \infty,$$

or, equivalently,

$$\mathbf{E}\left[\exp\left\{it\frac{\sqrt{m_n(x)}}{\alpha_n(x)}(\hat{\mu}_{KP}(x) - \mu)\right\}\right] \to e^{-t^2/2},\ as\ n \to \infty,$$

where $\hat{\mu}_{KP}(x)$ is defined in equation (4.2).

Proof of Lemma 4.1. We will proceed as in Sander ([131]). We have

$$\sqrt{m_n(x)}(\hat{\mu}_{KP}(x) - \mu(x)) = \int_0^{u_n} z dF_{m_n(x)}(z|x),$$

by integration by parts,

$$\sqrt{m_n(x)}(\hat{\mu}_{KP}(x) - \mu(x)) = u_n F_{m_n(x)}(z|x) - \int_0^{u_n} F_{m_n(x)}(z|x)dz.$$

By equation (4.5) and continuous mapping theorem (Billingsly [23]),

$$u_n F_{m_n(x)}(z|x) \overset{weakly}{\to} u_n F(z|x)$$

and

$$\int_0^{u_n} F_{m_n(x)}(z|x)dz \overset{weakly}{\rightarrow} \int_0^{u_n} F(z|x)dz$$

so

$$\sqrt{m_n(x)}(\hat{\mu}_{KP}(x) - \mu(x)) \xrightarrow{d} u_n F(z|x) - \int_0^{u_n} F(z|x)dz$$

Corollary1 in Sander ([131], p.7) states that

$$\sqrt{m_n(x)}(\hat{\mu}_{KP}(x) - \mu(x)) \xrightarrow{d} N(0, \alpha_n(x)), as\ n \rightarrow \infty,$$

where, (see proof of proposition 1):

$$\alpha_n^2(x) = \int_0^u \int_0^u std\Gamma_x(t,s); s \leq t.$$

Now let us examine the joint distribution of the parameter estimates. ■

Lemma 4.2 (Joint distributon) *$u_n = O_+(n^{\alpha\gamma_1(x)})$ for some $\alpha \in (0,1)$, where $\delta > 0$ and $\gamma_1(x) \in (0,1)$. Then*

$$\phi_{m_n(x)}(t_1,t_2,t_3,t_4) = \mathbf{E}\left[\exp\left\{\begin{array}{c} it_1\frac{\sqrt{m_n(x)}}{\alpha_n(x)}(\hat{\mu}_{KP}(x) - \mu) \\ +i\sqrt{N_u^x}(t_2,t_3)\left(\Sigma^{-1}\right)^{1/2}\begin{pmatrix}\hat{\gamma}^{(c,os)}_{Z^x,u(x)}(x) - \gamma_1(x) \\ \hat{\sigma}^{(c,os)}_{Z^x,u(x)}(x) - \sigma_1(x)\end{pmatrix} \\ +it_4\sqrt{m_n(x)}\left(\bar{F}_{m_n(x)}(u_n|x) - \bar{F}(u_n|x)\right)\end{array}\right\}\right]$$
$$\rightarrow \exp\left\{-\frac{1}{2}(t_1^2+t_2^2+t_3^2+t_4^2)\right\}\ as\ n \rightarrow \infty,$$

where Σ is given by equation (4.7) , $\mu(x) = \int_0^u zdF(z|x)$, $\alpha_n^2(x) = Var(\int_0^u zdF(z|x))$ and

$$\bar{F}_{m_n(x)}(u_n|x) = \prod_{i=1}^{m_n(x)}\left[1 - \frac{\delta_{(i)}^x 1_{\{Z_{(i)}^x \leq u_n(x)\}}}{n-i+1}\right].$$

Proof. By Lemma 4.1, equation (4.5) and Theorem 4.4.1 in Ndao ([114], p.126), the claim follows using the same techiques as in Johansson ([96], proof of (Lemma A.2, p.107). ■

We are now able to prove Theorem 4.1.

Proof of Theorem 4.1. Equation (4.4) states that

$$\hat{M} - M = \hat{\mu}_{KP}(x) - \mu(x) + \hat{\mu}_T(x) - \mu_T(x)$$

where

$$\begin{aligned}\mu_T(x) &= \int_{u_n(x)}^{\infty} y dF(y|x) = -\int_{u_n(x)}^{\infty} y d\bar{F}(y|x) \\ &= u_n\bar{F}(u_n|x) + \bar{F}(u_n|x)\int_0^{\infty} \bar{F}_{u_n(x)}(z|x)dz \\ &= u_n\bar{F}(u_n|x) + \bar{F}(u_n|x)\int_{u_n(x)}^{\infty} \frac{\bar{F}(y|x)}{\bar{F}(u_n|x)}dy \end{aligned} \tag{4.8}$$

Furthermore

$$\bar{F}(y|x) = \exp\left\{-\int_0^y \frac{dH^u(s|x)}{1-H(s|x)}\right\}$$

By Theorem 1.2.2 in de Haan and Ferreira (2006) and Proof of Lemma 6.1 in Ndao ([116], p.13),

$$\bar{F}(y|x) \sim \left(\bar{H}(y|x)\right)^{\gamma/\gamma_1} \ when\ y\ is\ large. \tag{4.9}$$

Proposition 1.5.7 in Bingham ([25], p.26) allows us to conclude that

$$\left(\bar{H}(y|x)\right)^{\gamma/\gamma_1} \sim y^{-1/\gamma_1(x)} + y^{-1/\gamma_1(x)-\delta_3}L_3(y|x); \tag{4.10}$$

where $L_3(.|x)$ are slowly varying functions at infinity and $\delta_3 > 0$. So by substuting (4.9) and (4.10) in (4.8), we get

$$\mu_T(x) \sim u_n\bar{F}(u_n|x) + \bar{F}(u_n|x)\frac{u_n^{1/\gamma_1(x)}}{1+u_n^{-\delta_3}L_3(u_n|x)}\int_{u_n(x)}^{\infty}\left(y^{-1/\gamma_1(x)} + y^{-1/\gamma_1(x)-\delta_3}L_3(y|x)\right)dy.$$

Now assuming that $L_3(.|x)$ is locally bounded in $[x_0,\infty)$ for some $x_0 \geq 0$, then by Karamata's Theorem (see Resnick ([127], p.26))

$$\mu_T(x) \sim \bar{F}(u_n|x)\left(u_n + \frac{\sigma_1(x)}{1-\gamma_1(x)}\right) + R_1(x)$$

where

$$R_1(x) = \frac{\bar{F}(u_n|x)\sigma_1(x)}{1-\gamma_1(x)}\frac{\delta_3(x)\gamma_1(x)}{1-\gamma_1(x)+\delta_3(x)\gamma_1(x)}\frac{u_n^{-\delta_3}L_3(u_n|x)}{1+u_n^{-\delta_3}L_3(u_n|x)}$$

Using Taylor expansion, we find that

$$\begin{aligned}\hat{M}-M &\sim \hat{\mu}_{KP}(x)-\mu(x)+\left(\bar{F}_{m_n(x)}(u_n|x)-\bar{F}(u_n|x)\right)\left(u_n(x)+\frac{\hat{\sigma}_1(x)}{1-\hat{\gamma}_1(x)}\right)\\ &+\bar{F}(u_n|x)\left(\frac{\hat{\sigma}_1(x)}{1-\hat{\gamma}_1(x)}-\frac{\sigma_1(x)}{1-\gamma_1(x)}\right)-R_1(x)\\ &=\hat{\mu}_{KP}(x)-\mu(x)+\left(\bar{F}_{m_n(x)}(u_n|x)-\bar{F}(u_n|x)\right)\left(u_n(x)+\frac{\sigma_1(x)}{1-\gamma_1(x)}+R_2(x,1)\right)\\ &+\bar{F}(u_n|x)\left(\frac{\sigma_1(x)}{(1-\gamma_1(x))^2}(\hat{\gamma}_1(x)-\gamma_1(x))+\frac{\sigma_1(x)-\hat{\sigma}_1(x)}{1-\gamma_1(x)}+R_2(x,2)\right)-R_1(x)\end{aligned}$$

where

$$\begin{aligned}R_2(x,t)=&\,\sigma_1(x)\sum_{k=t}^{\infty}\frac{(\hat{\gamma}_1(x)-\gamma_1(x))^k}{(1-\gamma_1(x))^{k+1}}\\ &+\sum_{k=t}^{\infty}\frac{1}{(1-\gamma_1(x))^k}(\sigma_1(x)-\hat{\sigma}_1(x))(\hat{\gamma}_1(x)-\gamma_1(x))^{k-1}.\end{aligned}$$

Multiplying by $\frac{\sqrt{m_n(x)}}{\alpha_n(x)}$ and using Proposition 4.1 and Lemma 4.2 in addition of continuous mapping theorem, we can write

$$\begin{aligned}\frac{\sqrt{m_n(x)}}{\alpha_n(x)}\left(\hat{M}-M\right)=&\frac{\sqrt{m_n(x)}}{\alpha_n(x)}(\hat{\mu}_{KP}(x)-\mu(x))\\ &+\frac{\sqrt{m_n(x)}}{\alpha_n(x)}\left(u_n(x)+\frac{\sigma_1(x)}{1-\gamma_1(x)}\right)\left(\bar{F}_{m_n(x)}(u_n|x)-\bar{F}(u_n|x)\right)\\ &+\frac{\sqrt{m_n(x)}\bar{F}(u_n|x)\sigma_1(x)}{\alpha_n(x)(1-\gamma_1(x))^2}(\hat{\gamma}_1(x)-\gamma_1(x))\\ &+\frac{\sqrt{m_n(x)}\bar{F}(u_n|x)}{\alpha_n(x)(1-\gamma_1(x))}(\sigma_1(x)-\hat{\sigma}_1(x))+o_P(1).\end{aligned}$$

Therefore

$$\begin{aligned}&Var\left(\frac{\sqrt{m_n(x)}}{\alpha_n(x)}\left(\hat{M}-M\right)\right)\\ &\approx 1+\frac{\left(\bar{H}(u|x)\right)^{2\frac{\gamma(x)}{\gamma_1(x)}-1}}{\alpha_n^2(x)}\times\\ &\left[\frac{\gamma(x)}{\gamma_1(x)}\left(u+\frac{\sigma_1(x)}{1-\gamma_1(x)}\right)^2+\frac{\sigma_1^2(x)(1+\gamma(x))^2}{(1-\gamma_1(x))^4}\left(\frac{\gamma_1(x)}{\gamma(x)}\right)^3+\frac{2\sigma_1^2(x)(1+\gamma(x))}{(1-\gamma_1(x))^2}\frac{\gamma_1(x)}{\gamma(x)}\right]\\ &=k_n(x),\end{aligned}$$

where $k_n(x)=O_+(1)$ (see Proposition 4.1). ■

We wish to warmly thank the Algerian Directorate General of Scientific Research and Technological Development for their support and contribution to this work.

Conclusion and Perspectives

Two main problems relating to the univariate theory of extreme values and extreme values analysis are addressed in this thesis. Before getting to the heart of the matter, the essential concepts for understanding our work is presented in Chapter 1, concepts that cover different key notions on extreme value theory in the simple case; i.e. complete data and absence of covariates.
In Chapter 2, we recalled some survival analysis notions namely censorship notion, Kaplan-Meier estimator and its generalized version. Also we gave a review of conditional extreme quantiles and conditional extreme value index estimators in case of right random censoring and its asymptotic behaviors that exist in literature up to now.
Chapter 3 is the main practical topic in this thesis, it is about the field of statistical analysis of extreme values (EVA) applied to hydrology domain , more especially, to rainfalls. In this study, we found the most adequate distributions that fit rainfalls in Khemis-Miliana region recorded between 1975 and 2006, we used the two existent methods, namely Generalized Extreme Value (GEV) distribution and Generalized Pareto (GP) distribution. We concluded this chapter by deriving the most important estimates, used to overcome risk of drought, called T-years return levels for different periods using both mentioned models.
Finally, in Chapter 4, we proposed a new estimator of the mean of a heavy-tailed distribution under right random censored data in presence of fixed covariates by combining the generalized KaplanMeier estimator before a threshold and a parametric model (GPD) which approximate the excesses over the threshold in order to overcome the bad behavior of K-M estimator in the heavy-tail of distribution, and its asymptotic normality is established.
This thesis offers interesting perspectives from a point of view both theoretical and practical.

In our future research we will try first to apply our estimate in real-world after

testing its performance by simulation. Secondly, an open problem is to construct an estimator of the mean of a heavy-tailed distribution under right random censored data where the covariate is random and to establish its asymptotic behaviors.
In practical part, we will use recent estimators to modeling our rainfall data in case of censorship geographical coordinates as covariates.

Bibliography

[1] Odd Aalen. Nonparametric estimation of partial transition probabilities in multiple decrement models. *The Annals of Statistics*, pages 534–545, 1978.

[2] MI Fraga Alves. Estimation of the tail parameter in the domain of attraction of an extremal distribution. *Journal of statistical planning and inference*, 45(1-2):143–173, 1995.

[3] Abdelkader Ameraoui, Kamal Boukhetala, and Jean-François Dupuy. Bayesian estimation of the tail index of a heavy tailed distribution under random censoring. *Computational Statistics & Data Analysis*, 104:148–168, 2016.

[4] Per K Andersen, Ornulf Borgan, Richard D Gill, and Niels Keiding. *Statistical models based on counting processes*. New York. Springer, 1993.

[5] Dragi Anevski. Functional central limit theorems for the nelson-aalen and kaplan-meier estimators for dependent stationary data. *arXiv e-prints*, pages arXiv–1609, 2016.

[6] N Balakrishnan and Debasis Kundu. Hybrid censoring: Models, inferential results and applications. *Computational Statistics & Data Analysis*, 57(1):166–209, 2013.

[7] Narayanaswamy Balakrishnan. Progressive censoring methodology: an appraisal. *Test*, 16(2):211, 2007.

[8] Narayanaswamy Balakrishnan and Rita Aggarwala. *Progressive censoring: theory, methods, and applications*. Springer Science & Business Media, 2000.

[9] August A Balkema and Laurens De Haan. Residual life time at great age. *The Annals of probability*, 2:792–804, 1974.

[10] Jan Beirlant, Frederico Caeiro, and M Ivette Gomes. An overview and open research topics in statistics of univariate extremes. *Revstat*, 10(1):1–31, 2012.

[11] Jan Beirlant, Goedele Dierckx, A Guillou, and C Staăricaă. On exponential representations of log-spacings of extreme order statistics. *Extremes*, 5(2):157–180, 2002.

[12] Jan Beirlant, Goedele Dierckx, and Armelle Guillou. Bias-reduced estimators for bivariate tail modelling. *Insurance: Mathematics and Economics*, 49(1):18–26, 2011.

[13] Jan Beirlant and Yuri Goegebeur. Local polynomial maximum likelihood estimation for pareto-type distributions. *Journal of Multivariate Analysis*, 89(1):97–118, 2004.

[14] Jan Beirlant, Yuri Goegebeur, Johan Segers, and Jozef L Teugels. *Statistics of extremes: theory and applications*. John Wiley & Sons, 2006.

[15] Jan Beirlant, Armelle Guillou, Goedele Dierckx, and Amélie Fils-Villetard. Estimation of the extreme value index and extreme quantiles under random censoring. *Extremes*, 10(3):151–174, 2007.

[16] Jan Beirlant, Armelle Guillou, and Gwladys Toulemonde. Peaks-over-threshold modeling under random censoring. *Communications in Statistics—Theory and Methods*, 39(7):1158–1179, 2010.

[17] Jan Beirlant, Gaonyalelwe Maribe, and Andrehette Verster. Penalized bias reduction in extreme value estimation for censored pareto-type data, and long-tailed insurance applications. *Insurance: Mathematics and Economics*, 78:114–122, 2018.

[18] Jan Beirlant, Petra Vynckier, Josef L Teugels, et al. Excess functions and estimation of the extreme-value index. *Bernoulli*, 2(4):293–318, 1996.

[19] Jan Beirlant, Petra Vynckier, and Jozef L Teugels. Tail index estimation, pareto quantile plots regression diagnostics. *Journal of the American statistical Association*, 91(436):1659–1667, 1996.

[20] Jan Beirlant, Julien Worms, and Rym Worms. Estimation of the extreme value index in a censorship framework: Asymptotic and finite sample behavior. *Journal of Statistical Planning and Inference*, 202:31–56, 2019.

[21] Rasmus E Benestad. Downscaling precipitation extremes. *Theoretical and Applied Climatology*, 100(1):1–21, 2010.

[22] Rudolf Beran. Nonparametric regression with randomly censored survival data. *Technical Report, Univ. California, Berkeley*, 1981.

[23] Patrick Billingsley. Convergence of probability measures john wiley & sons. *New York*, 157, 1968.

[24] NH Bingham. Regular variation in probability theory. *Publications de l'Institut Mathématique*, 48(68):169–180, 1990.

[25] Nicholas H Bingham, Charles M Goldie, and Jef L Teugels. *Regular variation*. Number 27. Cambridge university press, 1987.

[26] B Brahimi, D Meraghni, and A Necir. On the asymptotic normality of hill's estimator of the tail index under random censoring. *Preprint: arXiv-1302.1666*, page 44, 2013.

[27] Norman Breslow and John Crowley. A large sample study of the life table and product limit estimates under random censorship. *The Annals of statistics*, pages 437–453, 1974.

[28] Norman E Breslow. Contribution to discussion of paper by dr cox. *J. Roy. Statist. Soc., Ser. B*, 34:216–217, 1972.

[29] Frederico Caeiro and M Ivette Gomes. Semi-parametric tail inference through probability-weighted moments. *Journal of statistical planning and inference*, 141(2):937–950, 2011.

[30] E Castillo. Extreme value theory in engineering. aca. *Press, New York*, 1988.

[31] Enrique Castillo and Ali S Hadi. Fitting the generalized pareto distribution to data. *Journal of the American Statistical Association*, 92(440):1609–1620, 1997.

[32] B Chandrasekar, A Childs, and N Balakrishnan. Exact likelihood inference for the exponential distribution under generalized type-i and type-ii hybrid censoring. *Naval Research Logistics (NRL)*, 51(7):994–1004, 2004.

[33] A Childs, B Chandrasekar, N Balakrishnan, and D Kundu. Exact likelihood inference based on type-i and type-ii hybrid censored samples from the exponential distribution. *Annals of the Institute of Statistical Mathematics*, 55(2):319–330, 2003.

[34] Stuart Coles, Joanna Bawa, Lesley Trenner, and Pat Dorazio. *An introduction to statistical modeling of extreme values*, volume 208. Springer, 2001.

[35] Stuart Coles and Jonathan Tawn. Bayesian modelling of extreme surges on the uk east coast. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 363(1831):1387–1406, 2005.

[36] Sandor Csorgo, Paul Deheuvels, and David Mason. Kernel estimates of the tail index of a distribution. *The Annals of Statistics*, pages 1050–1077, 1985.

[37] Sándor Csörgoő and David M Mason. Central limit theorems for sums of extreme values. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 98, pages 547–558. Cambridge University Press, 1985.

[38] Jon Danielsson, Laurens de Haan, Liang Peng, and Casper G de Vries. Using a bootstrap method to choose the sample fraction in tail index estimation. *Journal of Multivariate analysis*, 76(2):226–248, 2001.

[39] Richard Davis, Sidney Resnick, et al. Tail estimates motivated by extreme value theory. *Annals of Statistics*, 12(4):1467–1487, 1984.

[40] Anthony C Davison and Richard L Smith. Models for exceedances over high thresholds. *Journal of the Royal Statistical Society: Series B (Methodological)*, 52(3):393–425, 1990.

[41] Anthony Christopher Davison. *A statistical model for contamination due to long-range atmospheric transport of radionuclides*. PhD thesis, University of London, 1985.

[42] Laurens de Haan. On regular variation and its application to the weak convergence of sample extremes. *Amsterdam, Mathematisch Centrum*, 32, 1970.

[43] Laurens De Haan. Sample extremes: an elementary introduction. Technical report, 1976.

[44] Laurens de Haan. Slow variation and characterization of domains of attraction. In *Statistical extremes and applications*, pages 31–48. Springer, 1984.

[45] Laurens De Haan and Ana Ferreira. *Extreme value theory: An introduction*. New York, Springer, 2006.

[46] Paul Deheuvels, Erich Haeusler, and David M Mason. Almost sure convergence of the hill estimator. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 104, pages 371–381. Cambridge University Press, 1988.

[47] Surobhi Deka, Munindra Borah, and Sarat Chandra Kakaty. Statistical analysis of annual maximum rainfall in north-east india: an application of lh-moments. *Theoretical and applied climatology*, 104(1):111–122, 2011.

[48] Arnold LM Dekkers, Laurens De Haan, et al. On the estimation of the extreme-value index and large quantile estimation. *The annals of statistics*, 17(4):1795–1832, 1989.

[49] Arnold LM Dekkers, John HJ Einmahl, and Laurens De Haan. A moment estimator for the index of an extreme-value distribution. *The Annals of Statistics*, pages 1833–1855, 1989.

[50] Edward Lewis Dodd. The greatest and the least variate under general laws of error. *Transactions of the American Mathematical Society*, 25(4):525–539, 1923.

[51] Gerrit Draisma, Laurens de Haan, Liang Peng, and T Themido Pereira. A bootstrap-based method to achieve optimality in estimating the extreme-value index. *Extremes*, 2(4):367–404, 1999.

[52] Holger Drees. Refined pickands estimators of the extreme value index. *The Annals of Statistics*, pages 2059–2080, 1995.

[53] Holger Drees. Refined pickands estimators wtth bias correction. *Communications in Statistics-Theory and Methods*, 25(4):837–851, 1996.

[54] Holger Drees. A general class of estimators of the extreme value index. *Journal of Statistical Planning and Inference*, 66(1):95–112, 1998.

[55] Holger Drees. On smooth statistical tail functionals. *Scandinavian Journal of Statistics*, 25(1):187–210, 1998.

[56] Holger Drees and Edgar Kaufmann. Selecting the optimal sample fraction in univariate extreme value estimation. *Stochastic Processes and their applications*, 75(2):149–172, 1998.

[57] Holger Drees, Sidney Resnick, and Laurens de Haan. How to make a hill plot. *The Annals of Statistics*, 28(1):254–274, 2000.

[58] Jean-Jacques Droesbeke and Gilbert Saporta. *Approches non paramétriques en régression*. Editions Technip, 2011.

[59] Y Du and MG Akritas. I.i.d representations of the conditional kaplan-meier process for arbitrary distributions. *Math. Methods Statist*, 11:152–182, 2002.

[60] John HJ Einmahl, Amélie Fils-Villetard, Armelle Guillou, et al. Statistics of extremes under random censoring. *Bernoulli*, 14(1):207–227, 2008.

[61] Paul Embrechts, Claudia Klüppelberg, and Thomas Mikosch. Modelling extremal events, volume 33 of applications of mathematics, 1997.

[62] Benjamin Epstein. Truncated life tests in the exponential case. *The Annals of Mathematical Statistics*, pages 555–564, 1954.

[63] Mikael Escobar-Bach and Olivier Goudet. On the study of the beran estimator for generalized censoring indicators. *arXiv preprint arXiv:2009.01726*, 2020.

[64] Michael Falk. Efficiency of convex combinations of pickands estimator of the extreme value index. *Journal of Nonparametric Statistics*, 4(2):133–147, 1994.

[65] Igor Fedotenkov. A review of more than one hundred pareto-tail index estimators. *Statistica*, 80(3):245–299, 2020.

[66] William Feller. An introduction to probability theory and its applications. *Second edition John Wiley Sons Inc., New York*, 1971.

[67] Willy Feller. Über den zentralen grenzwertsatz der wahrscheinlichkeitsrechnung. ii. *Mathematische Zeitschrift*, 42(1):301–312, 1937.

[68] Ana Ferreira, Laurens de Haan, and Liang Peng. On optimising the estimation of high quantiles of a probability distribution. *Statistics*, 37(5):401–434, 2003.

[69] Ronald Aylmer Fisher and Leonard Henry Caleb Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical proceedings of the Cambridge philosophical society*, volume 24, pages 180–190. Cambridge University Press, 1928.

[70] Petra Friederichs. Statistical downscaling of extreme precipitation events using extreme value theory. *Extremes*, 13(2):109–132, 2010.

[71] Laurent Gardes and Stephane Girard. A moving window approach for nonparametric estimation of the conditional tail index. *Journal of Multivariate Analysis*, 99(10):2368–2388, 2008.

[72] Laurent Gardes, Stéphane Girard, and Alexandre Lekina. Functional nonparametric estimation of conditional extreme quantiles. *Journal of Multivariate Analysis*, 101(2):419–433, 2010.

[73] RD Gill. Censoring and stochastic integrals, mathematical centre tracts 124. amsterdam: Mathematisch centrum. *Mathematical Reviews (MathSciNet): MR596815 Zentralblatt MATH*, 456, 1980.

[74] Richard Gill et al. Large sample behaviour of the product-limit estimator on the whole line. *The annals of statistics*, 11(1):49–58, 1983.

[75] Richard D Gill. Glivenko-cantelli for kaplan-meier. *Mathematical Methods of Statistics*, 3(1):76, 1994.

[76] Eric Gilleland, Mathieu Ribatet, and Alec G Stephenson. A software review for extreme value analysis. *Extremes*, 16(1):103–119, 2013.

[77] Emil Julius Gmbel. *Statistics of extremes*. New York, Columbia University Press, 1958.

[78] Boris Gnedenko. Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of mathematics*, pages 423–453, 1943.

[79] M Ivette Gomes, Luísa Canto e Castro, M Isabel Fraga Alves, and Dinis Pestana. Statistics of extremes for iid data and breakthroughs in the estimation of the extreme value index: Laurens de haan leading contributions. *Extremes*, 11(1):3–34, 2008.

[80] M Ivette Gomes and Armelle Guillou. Extreme value theory and statistics of univariate extremes: a review. *International statistical review*, 83(2):263–292, 2015.

[81] M Ivette Gomes and M Manuela Neves. Estimation of the extreme value index for randomly censored data. *Biometrical Letters*, 48(1):1–22, 2011.

[82] M Ivette Gomes and Orlando Oliveira. The bootstrap methodology in statistics of extremes—choice of the optimal sample fraction. *Extremes*, 4(4):331–358, 2001.

[83] M Ivette Gomes and Dinis Pestana. A sturdy reduced-bias extreme quantile (var) estimator. *Journal of the American Statistical Association*, 102(477):280–292, 2007.

[84] Wenceslao Gonzalez-Manteiga and Carmen Cadarso-Suarez. Asymptotic properties of a generalized kaplan-meier estimator with some applications. *Communications in Statistics-Theory and Methods*, 4(1):65–78, 1994.

[85] Toufik Guermah and Abdelaziz Rassoul. Study of extreme rainfalls using extreme value theory (case study: Khemis-miliana region, algeria). *Communications in Statistics: Case Studies, Data Analysis and Applications*, 6(3):364–379, 2020.

[86] Armelle Guillou and Peter Hall. A diagnostic for selecting the threshold in extreme value analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):293–305, 2001.

[87] Erich Haeusler and Jozef L Teugels. On asymptotic normality of hill's estimator for the exponent of regular variation. *The Annals of Statistics*, pages 743–756, 1985.

[88] Peter Hall. Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *Journal of multivariate analysis*, 32(2):177–203, 1990.

[89] Peter Hall, Alan H Welsh, et al. Adaptive estimates of parameters of regular variation. *The Annals of Statistics*, 13(1):331–341, 1985.

[90] Bruce M Hill. A simple general approach to inference about the tail of a distribution. *The annals of statistics*, pages 1163–1174, 1975.

[91] Jonathan Richard Morley Hosking, James R Wallis, and Eric F Wood. Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics*, 27(3):251–261, 1985.

[92] Jonathan RM Hosking and James R Wallis. Parameter and quantile estimation for the generalized pareto distribution. *Technometrics*, 29(3):339–349, 1987.

[93] Xuelin Huang and Robert L Strawderman. A note on the breslow survival estimator. *Journal of Nonparametric Statistics*, 18(1):45–56, 2006.

[94] M Ivette Gomes and Dinis Pestana. A simple second-order reduced bias' tail index estimator. *Journal of Statistical Computation and Simulation*, 77(6):487–502, 2007.

[95] Arthur F Jenkinson. The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society*, 81(348):158–171, 1955.

[96] Joachim Johansson. Estimating the mean of heavy-tailed distributions. *Extremes*, 6(2):91–109, 2003.

[97] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

[98] J Karamata. Sur un mode de croissance reguliere des functions, mathematica (cluj), 4 (1930), 38-53. *Karamata384Mathematica (Cluj)*, 1930.

[99] Chansoo Kim, Myoung-Seok Suh, and Ki-Ok Hong. Bayesian changepoint analysis of the annual maximum of daily and subdaily precipitation over south korea. *Journal of Climate*, 22(24):6741–6757, 2009.

[100] John P Klein and Melvin L Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer Science & Business Media, 2005.

[101] M Ross Leadbetter, Georg Lindgren, and Holger Rootzén. Extremes and related properties of random sequences and processes. 1983.

[102] MR Leadbetter. Extremes and local dependence in stationary sequences, z. wahrscheinlichkeit, 65, 291–306, 1983.

[103] Elisa T Lee and John Wang. *Statistical methods for survival data analysis*, volume 476. John Wiley & Sons, 2003.

[104] Erich Leo Lehmann, George Casella, and George Casella. Theory of point estimation. wadsworth & brooks. *Cole Advanced Books and Software, Pacific Grove, California*, 16:24–25, 1991.

[105] Alexandre Lekina. Estimation non-paramétrique des quantiles extrêmes conditionnels. *PhD thesis, Grenoble University*, 2010.

[106] Chien-Tai Lin, Hon Keung Tony Ng, and Ping Shing Chan. Statistical inference of type-ii progressively hybrid censored data with weibull lifetimes. *Communications in Statistics—Theory and Methods*, 38(10):1710–1729, 2009.

[107] Gane Samb Lo. Asymptotic behavior of hill's estimate and applications. *Journal of applied probability*, pages 922–936, 1986.

[108] Shaw-Hwa Lo and Kesar Singh. The product-limit estimator and the bootstrap: some asymptotic representations. *Probability Theory and Related Fields*, 71(3):455–465, 1986.

[109] Natalia M Markovich. High quantile estimation for heavy-tailed distributions. *Performance Evaluation*, 62(1-4):178–192, 2005.

[110] David M Mason. Laws of large numbers for sums of extreme values. *The Annals of Probability*, pages 754–764, 1982.

[111] Gunther Matthys and Jan Beirlant. Estimating the extreme value index and high quantiles with exponential regression models. *Statistica Sinica*, pages 853–880, 2003.

[112] David Mauro. A combinatoric approach to the kaplan-meier estimator. *The Annals of Statistics*, pages 142–149, 1985.

[113] Dirk F Moore. *Applied survival analysis using R*. Springer, 2016.

[114] Pathé Ndao. Modélisation de valeurs extrêmes modélisation de valeurs extrêmes conditionnelles en présence de censure. *PhD thesis, Gaston Berger University*, 2015.

[115] Pathé Ndao, Aliou Diop, and Jean-François Dupuy. Nonparametric estimation of the conditional tail index and extreme quantiles under random censoring. *Computational Statistics & Data Analysis*, 79:63–79, 2014.

[116] Pathé Ndao, Aliou Diop, and Jean-François Dupuy. Nonparametric estimation of the conditional extreme-value index with random covariates and censoring. *Journal of Statistical Planning and Inference*, 168:20–37, 2016.

[117] Wayne Nelson. A short life test for comparing a sample with previous accelerated test results. *Technometrics*, 14(1):175–185, 1972.

[118] Hon Keung Tony Ng, Debasis Kundu, and Ping Shing Chan. Statistical analysis of exponential lifetimes under an adaptive type-ii progressive censoring scheme. *Naval Research Logistics (NRL)*, 56(8):687–698, 2009.

[119] Serguei Y Novak. *Extreme value methods with applications to finance*. CRC Press, 2011.

[120] Liang Peng. Estimating the mean of a heavy tailed distribution. *Statistics & Probability Letters*, 52(3):255–264, 2001.

[121] Liang Peng and Yongcheng Qi. *Inference for Heavy-Tailed Data: Applications in Insurance and Finance*. Academic Press, 2017.

[122] Margaret Sullivan Pepe and Thomas R Fleming. Weighted kaplan-meier statistics: a class of distance tests for censored survival data. *Biometrics*, pages 497–507, 1989.

[123] James Pickands III et al. Statistical inference using extreme order statistics. *Annals of statistics*, 3(1):119–131, 1975.

[124] R-D Reiss. Approximate distributions of order statistics: with applications to nonparametric statistics. 1989.

[125] Rolf-Dieter Reiss and Michael Thomas. Flood frequency analysis. *Statistical Analysis of Extreme Values: with Applications to Insurance, Finance, Hydrology and Other Fields*, pages 337–351, 2007.

[126] Rolf-Dieter Reiss, Michael Thomas, and RD Reiss. *Statistical analysis of extreme values*, volume 2. Springer, 1997.

[127] SI Resnick. Extreme values, point processes and regular variation, 1987.

[128] Sidney Resnick and Ctlin Stric. Smoothing the hill estimator. *Advances in Applied Probability*, pages 271–293, 1997.

[129] Justin Rutikanga and Aliou Diop. Functional kernel estimation of the conditional extreme value index under random right censoring. 2020.

[130] Justin Ushize Rutikanga, Aliou Diop, et al. Functional kernel estimation of the conditional extreme quantile under random right censoring. *Open Journal of Statistics*, 11(01):162, 2021.

[131] Joan M Sander. *Asymptotic normality of linear combinations of functions of order statistics with censored data*. 1975.

[132] A Schick, V Susarla, and H Koul. Efficient estimation of functionals with censored data. *Statistics & Risk Modeling*, 6(4):349–360, 1988.

[133] Johan Segers. Generalized pickands estimators for the extreme value index. *Journal of Statistical Planning and Inference*, 128(2):381–396, 2005.

[134] Galen R Shorack and Jon A Wellner. *Empirical processes with applications to statistics*. John Wiley & Sons, 1986.

[135] Richard L Smith. Extreme value theory based on the r largest annual events. *Journal of Hydrology*, 86(1-2):27–43, 1986.

[136] Richard L Smith et al. Estimating tails of probability distributions. *The annals of Statistics*, 15(3):1174–1207, 1987.

[137] Louiza Soltane, Djamel Meraghni, and Abdelhakim Necir. Estimating the mean of a heavy-tailed distribution under random censoring. *arXiv preprint arXiv:1507.03178*, 2015.

[138] Gilles Stupfler. Estimating the conditional extreme-value index under random right-censoring. *Journal of Multivariate Analysis*, 144:1–24, 2016.

[139] Winfried Stute. The bias of kaplan-meier integrals. *Scandinavian Journal of Statistics*, pages 475–484, 1994.

[140] Winfried Stute. The central limit theorem under random censorship. *The Annals of Statistics*, pages 422–439, 1995.

[141] Winfried Stute. Distributional convergence under random censorship when covariables are present. *Scandinavian journal of statistics*, pages 461–471, 1996.

[142] Winfried Stute. Kaplan–meier integrals. *Advances in Survival Analysis*, page 87, 2004.

[143] Winfried Stute and J-L Wang. The strong law under random censorship. *The Annals of statistics*, pages 1591–1607, 1993.

[144] Winfried Stute and Jane-Ling Wang. The jackknife estimate of a kaplan—meier integral. *Biometrika*, 81(3):602–606, 1994.

[145] V Susarla, J Van Ryzin, et al. Large sample theory for an estimator of the mean survival time from censored samples. *Annals of Statistics*, 8(5):1002–1016, 1980.

[146] Akio Suzukawa. Unbiased estimation of functionals under random censorship. *Journal of the Japan Statistical Society*, 34(2):153–172, 2004.

[147] Gwladys Toulemonde. *Estimation et tests en théorie des valeurs extrêmes*. PhD thesis, Université Pierre et Marie Curie-Paris VI, 2008.

[148] Bruce W Turnbull. Nonparametric estimation of a survivorship function with doubly censored data. *Journal of the American statistical association*, 69(345):169–173, 1974.

[149] Ingrid Van Keilegom and Michael G Akritas. Transfer of tail information in censored regression models. *Annals of Statistics*, pages 1745–1784, 1999.

[150] Ingrid Van Keilegom and Noël Veraverbeke. Estimation and bootstrap with censored data in fixed design nonparametric regression. *Annals of the Institute of Statistical Mathematics*, 49(3):467–491, 1997.

[151] Ingrid Van Keilegom and Noël Veraverbeke. Bootstrapping quantiles in a fixed design regression model with censored data. *Journal of Statistical Planning and Inference*, 69(1):115–131, 1998.

[152] Richard Von Mises. La distribution de la plus grande de n valuers. *Rev. math. Union interbalcanique*, 1:141–160, 1936.

[153] Jia-Gang Wang et al. A note on the uniform consistency of the kaplan-meier estimator. *The Annals of Statistics*, 15(3):1313–1316, 1987.

[154] Ishay Weissman. Estimation of parameters and large quantiles based on the k largest observations. *Journal of the American Statistical Association*, 73(364):812–815, 1978.

[155] Andreas Wienke, Samuli Ripatti, Juni Palmgren, and Anatoli Yashin. A bivariate survival model with compound poisson frailty. *Statistics in Medicine*, 29(2):275–283, 2010.

[156] Daniel S Wilks. Statistical methods in the atmospheric sciences: An introduction. *Academic Press, San Diego*, page 467, 1995.

[157] Julien Worms and Rym Worms. New estimators of the extreme value index under random right censoring, for heavy-tailed distributions. *Extremes*, 17(2):337–358, 2014.

[158] Song Yang. A central limit theorem for functionals of the kaplan—meier estimator. *Statistics & probability letters*, 21(5):337–345, 1994.

[159] Seokhoon Yun. A class of pickands-type estimators for the extreme value index. *Journal of Statistical Planning and Inference*, 83(1):113–124, 2000.

[160] Seokhoon Yun. On a generalized pickands estimator of the extreme value index. *Journal of statistical planning and inference*, 102(2):389–409, 2002.

[161] M Zhou. Two-sided bias bound of the kaplan-meier estimator. *Probability theory and related fields*, 79(2):165–173, 1988.