People's Democratic Republic of Algeria

Ministry of Higher Education and Scientific Research

Mohamed Khider University, Biskra - Algeria

Faculty of Exact Sciences, Natural Sciences and Life Sciences

Laboratory of Applied Mathematics
Department of Mathematics

A Third Cycle Doctoral Thesis

Presented for the Degree of

DOCTOR in Applied Mathematics

In the field of STATISTICS

By

ZAHNIT Abida

Title :

On robust tail index estimation under incomplete data

Examination Committee Members :

Mr.	**Zouhir MOKHTARI**	**Professor**	**U. Biskra**	**President**
Mr.	**Brahim BRAHIMI**	**Professor**	**U. Biskra**	**Supervisor**
Mr.	**Fatah BENATIA**	**Professor**	**U. Biskra**	**Examiner**
Mr.	**Adel AISSAOUI**	**Professor**	**U. El-Oued**	**Examiner**

Mars 22, 2022

People's Democratic Republic of Algeria

Ministry of Higher Education and Scientific Research

Mohamed Khider University, Biskra – Algeria

Faculty of Exact Sciences, Natural Sciences and Life Sciences

Laboratory of Applied Mathematics

Department of Mathematics

Third Cycle Doctoral LMD Thesis

Submitted in partial fulfillment for the requirements of the

Doctorate Degree in "Applied Mathematics"

Presented by Mm. *Z*AHNIT Abida

Title :

On robust tail index estimation under incomplete data

Examination Committee :

Mr.	Zouhir MOKHTARI	Professor	U. Biskra	President
Mr.	Brahim BRAHIMI,	Professor	U. Biskra	Supervisor
Mr.	Fatah BENATIA	Professor	U. Biskra	Examiner
Mr.	Adel AISSAOUI	Professor	U. El-Oued	Examiner

defended publicly the 22/03/2022

Dédicace

To my parents,

To my husband, *my son and daughters*,

To all my family,

To all those who are dear to me.

*A*cknowledgments

I extend my sincere thanks, appreciation and respect to my supervisor, Prof. Brahim Brahimi for the instructions and directions he gave me and who accompanied me throughout the realization of the thesis.

I would like to thank the Members of Examination Committee : Prof. Mokhtari Z., Prof. Benatia F. and Prof. Aissaoui A. who accepted the reading and evaluation of my thesis.

To all those who have contributed to the realization of this thesis : Thank's

Achieved Works

I) Paper:

1) Zahnit A., Brahimi B. ,Yahia D. (2021). Robust estimation of the extreme value index of Pareto-type distributions under random truncation with applications. Pak. J. Stat. Oper.Res. Vol.17 No.1, pp. 235-245.

II) Presentations at Conferences :

1) Zahnit A. (2017). Robust estimation of the tail index of Pareto-type distributions under random truncation. $61st$World Statistics Congress of the ISI (16 – 21 July 2017), Marrakech, Morocco.

2) Zahnit A. (2017). On robust tail index estimation procedure for Pareto-type distributions in the framework of randomly censored samples. Journées de Mathématiques Appliquées, 18-19 Décembre 2017, Université de Biskra.

Abstract

In this thesis, we propose a new robust estimation procedure for the tail index for Pareto-type distributions under incomplete data (censorship or truncation). Under truncation, the extreme quantile estimation is also derived and applied to an actual data set on automobile brake pad life.

Simulation study using R statistical software is carried out to evaluate the performance and the robustness of the proposed estimators for small and large sample size and for both uncontaminated and contaminated cases. Our newly estimators have been shown to be more robust and perform better than existing Hill-type estimators based on upper order statistics, in both cases of incomplete data (censorship or truncation).

Notations and symbols

$\xrightarrow{a.s}$	almost sure convergence
$\xrightarrow{P}$	convergence in probability
$\xrightarrow{d}$	convergence in distribution
cdf	cumulative distribution function
$F(.)$	distribution function
$\bar{F}(.)$	tail function
$D(A)$	domain of attraction of A
$F_n(.)$	empirical distribution function
evi	extreme value index
EVT	extreme value theory
$E(X)$	expectation of X
iid	independent and identically distributed
1_A	indicator function of a set A
i.e.	in other words
$f(.)$	probability density function
$Q(.)$	quantile function
rv or rv's	random variable(s)

$\mathcal{RV}_{-a}$	regularly varying functions with index a
$\mathcal{N}(0,1)$	standard Gaussian distribution
$L(.)$	slowly varying function at infinity
x_F	upper endpoint
$\mathcal{F}$	a set of distributions with support in $\mathbb{R}_+$
$(\Omega, \mathcal{A}, P)$	probability space
$abias$	absolute bias
$rmse$	root mean squared error
RRT	randomly right-truncated
$X_{1,n} \leq ... \leq X_{n,n}$	order statistics pertaining to the sample $(X_1, ..., X_n)$
k	numbers of top statistics (upper observations)
Q_ε	high quantile corresponding to upper tail probability $(1-\varepsilon)$
$F_n^{(KM)}$	Kaplan-Meier nonparametric estimator of F
$F_n^{(LB)}$	Lynden-Bell nonparametric estimator of F
$F_n^{(W)}$	Woodroofe's nonparametric estimator of F
α	tail index
γ	extreme value index
$\widehat{\alpha}^H$	Hill (1975) tail index estimator
$\widehat{\alpha}_1^{(H,c)}$	Einmahl et al. (2008) adapted tail index estimator
T_n	Beran and Shell (2012) tail index estimator
$\widehat{\alpha}_1^{(S)}$	Sayah et al. (2014) tail index estimator
$\widehat{\alpha}_1^{(Z)}$	Zahnit et al. (2021) tail index estimator
$\hat{\gamma}_1^{(GS)}$	Gardes and Stupfler (2015) extreme value index estimator
$\hat{\gamma}_1^{(W)}$	Worms and Worms (2016) extreme value index estimator
$\hat{\gamma}_1^{(B)}$	Benchaira et al. (2016) extreme value index estimator
$\hat{\gamma}_1^{(Z)}$	Zahnit et al. (2021) extreme value index estimator

Contents

List of Figures

List of Tables

Résumé de la thèse

Dans cette thèse, nous proposons une nouvelle procédure d'estimation robuste de l'indice des queues pour les distributions de type Pareto sous données incomplètes (censure ou troncature). Après la construction des estimateurs de l'indice des queues de distribution et des quantiles extrêmes, nous étudions leurs propriétés asymptotiques.

Nos considérations sont basées sur l'intégrale de Lynden-Bell (pour les données de troncature aléatoire) et les idées d'intégration de Kaplan-Meier (sous le modèle de censure aléatoire) en utilisant l'estimateur M-Huberisé de l'indice de queue.

Sous troncature, l'estimation des quantiles extrêmes est également dérivée et appliquée à un ensemble de données réelles sur la durée de vie des plaquettes de frein automobile.

Des travaux de simulation à l'aide du logiciel de traitement statistique R sont réalisés pour confirmer le bon comportement et pour évaluer la performance et la robustesse des estimateurs proposés pour des échantillons de petite et grande taille ainsi que dans le cas de contamination.

Il a été démontré que nos nouveaux estimateurs d'indice de valeur extrême des distributions de type Pareto sont plus robustes et fonctionnent mieux que les esti-

mateurs de type Hill existants basés sur les statistiques d'ordre et les observations supérieures, pour les petits échantillons et dans les deux cas de données incomplètes (censure ou troncature).

La thèse est organisée en quatre chapitres comme suit :

Le premier chapitre est consacré à la présentation du concept de distributions à queue lourde et des différentes classes de ce type de distributions. Les distributions à queue lourde sont liées à la théorie des valeurs extrêmes et permettent de modéliser plusieurs phénomènes rencontrés dans différentes disciplines : finance, hydrologie, télécommunications, géologie ... etc. Plusieurs définitions ont été associées à ces distributions en fonction de critères de classification.

Nous donnons dans un deuxième chapitre, quelques résultats importants et utiles, des concepts de base et des exemples dans la littérature sur le contexte des données incomplètes (troncature et censure).

Dans le troisième chapitre, nous proposons une nouvelle procédure robuste d'estimation de l'indice de queue pour les distributions de type Pareto dans le cadre d'échantillons censurés aléatoirement, basée sur les idées de l'intégration de Kaplan-Meier en utilisant l'M-estimateur hubérisé de l'indice de queue. Nous en dérivons les résultats asymptotiques et nous illustrons les performances et la robustesse de cet estimateur pour des échantillons de petite et grande taille dans une étude de simulation.

Dans le quatrième et dernier chapitre de cette thèse, nous introduisons un nouvel estimateur robuste pour l'indice des valeurs extrêmes des distributions de type Pareto sous des données tronquées à droite et établissons sa cohérence et sa normalité asymptotique. Nos considérations sont basées sur l'intégrale de Lynden-Bell

et un M-estimateur hubérisé de l'indice de queue. Une étude de simulation est réalisée pour évaluer la robustesse et le comportement en échantillon de taille finie de l'estimateur proposé. L'estimation des quantiles extrêmes est également dérivée et appliqué à un ensemble de données réelles sur la durée de vie des plaquettes de frein automobile.

Enfin, nous décrivons quelques remarques de conclusion et quelques perspectives de recherche dans la partie Conclusion et Perspectives.

Introduction

Heavy tailed (or Pareto-type) distributions are related to extreme value theory (EVT) and allow to model several phenomena encountered in different disciplines: Finance and business, internet traffic, hydrology, economics and have been accepted as realistic models for various phenomena, flood levels of rivers, major insurance claims, low and high temperatures...

Let (X_j), $1 \leq j \leq n$, *denote a sample of positive and independent random variables (rv's) defined over some probability space* $(\Omega, \mathcal{A}, P)$, *with continuous cumulative distribution functions (cdf)* F. *We assume that the survival function* $\overline{F} := 1 - F$ *is regularly varying at infinity, with index* $(1/\gamma =: \alpha)$, *i.e,* $\overline{F} \in RV_{-1/\gamma}$. *That is, for any* $t > 0$,

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

where $\gamma > 0$ *is the so-called extreme value index (e.v.i) is a well-known parameter to measure the tail heaviness of a distribution. Distributions satisfying (1) play a very crucial role in extreme value analysis. They include many commonly used models such as Pareto, Burr, Fréchet and Lévy-stable distributions, known to be suitable models for adjusting large insurance claims, log-returns, large fluctuations, etc... see, for instance, [Resnick (2006)].*

Estimation of e.v.i. is very important in the determination of high quantiles,

upper tail probabilities, mean excess functions, and excess-of-loss and stop-loss reinsurance premiums. Consequently, small errors in estimation of this quantity can produce substantial impact in applications. Thus, for robust estimation of quantities based on γ robust estimation of γ itself is crucial. In other words, for a heavy tailed distributions, robust estimation of the high quantile Q_ε corresponding to upper tail probability ε, becomes of interest, and this may be carried out by robust estimation of γ. [Brazauskas and Serfling (2000)] gives a detailed account of this issue.

The more popular estimator for e.v.i γ, is the well known Hill estimators [Hill (1975)] denoted by $\hat{\gamma}$ as functions of the numbers of upper observations:

$$\hat{\gamma}(k) := \frac{1}{k}\sum_{j=1}^{k} \log\left(X_{(n-j+1)}/X_{(n-k)}\right),$$

$X_{(1)} \leq ... \leq X_{(n)}$ denote the usual order statistics of the observed samples, k is the number of top statistics (upper observations). This estimator suffers from some kind of robustness, in the sense that it very sensitive to extreme observations, data contamination or model deviation and tend to be highly volatile for small samples. Also, the rate of convergence of this estimator is based on the optimal value of the numbers of top statistics k, but this rate is slower than the parametric rate $\sqrt{n}$. Moreover, estimating the optimal value of k is virtually impossible when the sample size n is small and this leads to unstable estimates for small samples and large confidence intervals, see for a discussion [Resnick (1997)].

In many real applications, such as survival analysis, reliability theory or insurance..., the variable of interest X is not necessarily completely available. This is the case of incomplete data (presence of random right censoring or random right

truncation). The usual way to model the situation of random right censoring is to introduce a random variable Y called censoring rv, independent of X, and then to consider the rv $Z := \min(X, Y)$ and the indicator variable $\delta := 1(X \leq Y)$, which determines whether or not X has been observed. Statistics of extremes of randomly censored data is a new research field. The topic was first mentioned in [Reiss and Thomas (1997)], where an estimator of a positive extreme value index was introduced, but with no asymptotic results. Recently, [Beirlant et al. (2007)] proposed an estimators for the general extreme value index and for the extreme quantile with their asymptotic properties. [Einmahl et al. (2008)] adapted various extreme value index estimators to the case where the data are censored, by a random threshold and establish their asymptotic normality by imposing some assumptions that are rather unusual to the context of extreme value theory. More recently [Sayah et al. (2014)], using the empirical process theory to approximate the adapted Hill estimator, for censored data, and derived its asymptotic normality.

In case of presence of random right truncation (RRT), the rv of interest X may not be fully available. This truncation can occur in many areas, for example, it is usual that the insurer's claim data do not correspond to the underlying losses, because they are truncated from above. For a recent paper on insurance claims under RRT, one refers to [Escudero and Ortega (2008)].

As a consequence of truncation, the size of actually observed sample, n, is a binomial rv with parameters N and $p := P(X \leq Y)$. We shall assume that $p > 0$, otherwise, nothing will be observed. Recently, [Gardes and Stupfler (2015)] defined an estimator for the parameter of interest γ by considering the classical Hill estimators [Hill (1975)]. Recently, [Worms and Worms (2016)] proposed an

asymptotically normal estimator for γ_1 by considering a Lynden-Bell integrals with a deterministic threshold. The case of a random threshold, is addressed by [Benchaira et al. (2016)] who propose a Hill-type estimator under RRT based on a Woodroofe integration.

The alternative approach is inspired by the theory of robust inference (see, for instance, [Huber (1981)] and [Hampel et al. (1986)]) instead of exact consistency, this theory aim at stability for small samples, possibly at the cost of a small asymptotic bias. However, as observed by [Beran and Shell (2012)], in some practical cases, such as natural disasters, operational risk assessment or reinsurance data are sparse (with n often somewhere between 20 and 50) and distributions are expected to be heavy tailed with an unknown e.v.i. Robust estimation of e.v.i. focuses primarily on complete data case, see [Brazauskas and Serfling (2000)], [Beran and Shell (2012)] and references therein. The incomplete data case has first been considered by [Sayah et al. (2014)], who dealt with heavy-tailed and right censored data and they gives a robust estimation of e.v.i. for the exactly Pareto distribution. The general case is discuted in the remainder of this thesis.

The thesis is organized into four chapters as follow:

The first chapter is devoted to the presentation of the concept of heavy-tailed distributions and different classes of this type of distributions. Heavy tailed distributions are related to extreme value theory and allow to modeling several phenomena in different disciplines: finance, hydrology, telecommunications, geology... etc.

Since our work carries on the incomplete data, and in order to give back easy the reading of this thesis, we give in a second chapter, some important and useful results, basic concepts and examples in the literature on the incomplete data context

(truncation and censoring).

In the third chapter, we propose a new robust tail index estimation procedure for Pareto-type distributions in the framework of randomly censored samples, based on the ideas of Kaplan-Meier integration using the huberized M-estimator of the tail index. We derive their asymptotic results. We illustrate the performance and the robustness of this estimator for small and large sample size in a simulation study.

In the fourth chapter of this thesis, we introduce a new robust estimator for the extreme value index of Pareto-type distributions under randomly right-truncated data and establish its consistency and asymptotic normality. Our considerations are based on the Lynden-Bell integral and a useful huberized M-functional and M-estimators of the tail index. A simulation study is carried out to evaluate the robustness and the finite sample behavior of the proposed estimator. Extreme quantiles estimation is also derived and applied to real dataset of lifetimes of automobile brake pads.

Finally we outline some concluding remarks and topic for future investigations in the Conclusion and Perspectives part.

Chapter 1

Heavy-tailed distributions and extreme value theory

1.1 Heavy-tailed distributions

In statistical theory, heavy-tailed distributions are functions whose tail is heavier than the exponential distribution (see, Fig1.1).

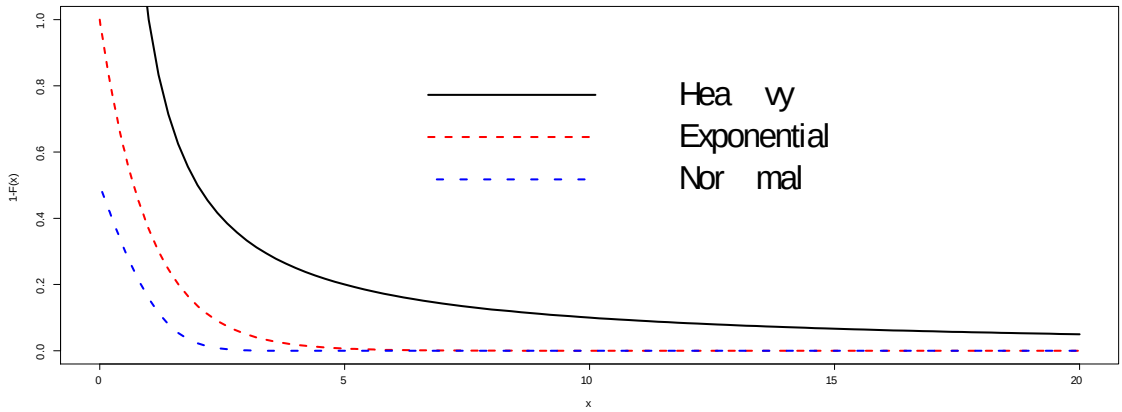

Figure 1.1: Decay of the tail functions: Normal, Exponential and Heavy.

In many applications it is the right tail that is of interest, but a cast can have a heavy left tail, or both tails can be heavy. There are a few different definitions of heavy tailedness, all are related to the decay of the tail function.

For example, in real data application, financial returns are known to be non-normal and tend to have heavy-tailed distributions. Fig1.2 contains the daily returns R_t (closing prices P_t) of Germany DAX (Ibis) stock indice between 1991 and 1998 :

$$R_t = \log\left(\frac{P_t}{P_{t-1}}\right), \qquad t = (1991, 130) : (1998, 169) = 7440 \; obs.$$

The data are sampled in business time (260 days/year), i.e., weekends and holidays are omitted. The data were kindly provided by Erste Bank AG, Vienna, Austria.

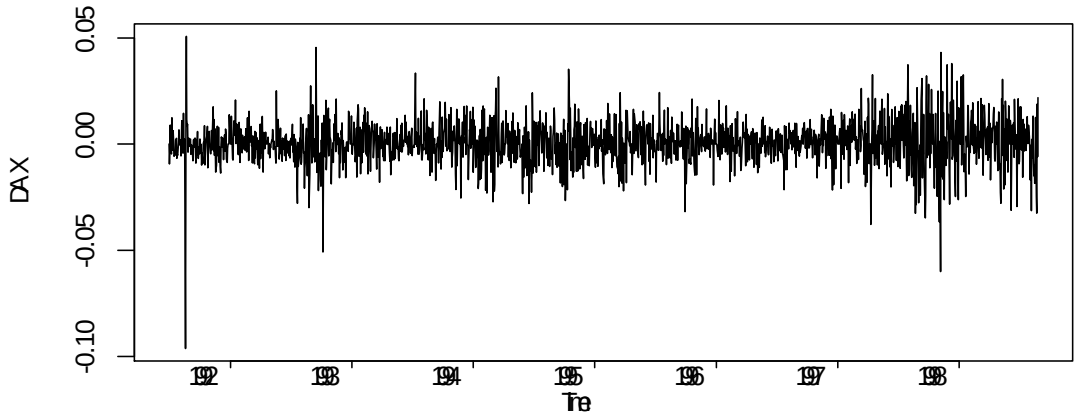

Figure 1.2: Daily returns of Germany DAX stock indice (1991-1998).

Normal quantiles-quantiles plot in Fig1.3 show the non-normality of the Germany DAX stock indice daily return's, due to possible heavy-tailed distributions.

Let (X_j), $1 \leq j \leq n$, denote a sample of independent and identically distributed (iid) random variables (rv's) defined over some probability space $(\Omega, \mathcal{A}, P)$, with

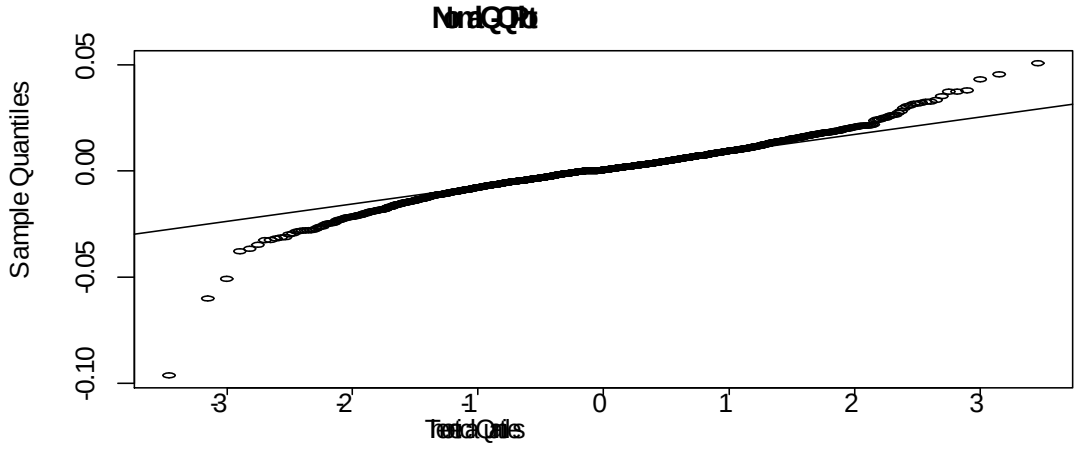

Figure 1.3: QQ-plot of daily returns of Germany DAX stock indice.

continuous cumulative distribution functions (cdf) F :

$$F(x) = P(X \leq x).$$

We consider nonnegative rv's X, such as losses in investments or claims in insurance. For arbitrary rv's, we should consider both right and left tails. Concerning about large losses leads us to consider $P(X > x)$ for x large. If F is the distribution function of X, we define the tail (survival) function $\bar{F}$ by

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

The tail of a distribution represents the probability for large (extremes) values of the variable. When these large values appear in a dataset, their probabilities of occurrence are not zero.

Definition 1.1.1 *It is said that the distribution has heavy tail if the kurtosis*

coefficient κ *defined by:*

$$\kappa = \frac{\mu_4}{\mu_2^2} > 3, \tag{1.1}$$

where $\mu_j := E\left(X - E\left(X\right)\right)^j$ *is the centred* $j-th$ *moment, and* $\kappa = 3$ *in the case of normal rv* X.

Remark 1.1.1 *The characterization given by equation (1.1) is very general and can be applied only if the moment of order* 4 *exists, therefore no discrimination, for distributions with infinite* $4-th$ *moment can be made. The following table 1.1 shows the historical kurtosis of some real data sets (see, [Gouriéroux (2012)], pages 68-69).*

Data	Kodak	G.Electric	G.Motors	Gold	Zinc
Period	$1966-1976$	$1966-1976$	$1966-1976$	$1975-1982$	$1970-1981$
κ	6.3	5.1	7.2	11.4	15.0

Table 1.1: Kurtosis coefficients of some real data sets

Definition 1.1.2 *Let* X *a rv with a distribution function* F. *This distribution is said to have a heavy tail if*

$$\bar{F}\left(x\right) = P\left(X > x\right) \sim x^{-\alpha}, \ \ as\ x \to \infty,$$

where the parameter $\alpha > 0$ *is called the tail index, which measure the tail heaviness of a distribution (see, Figure 1.4).*

Remark 1.1.2 *If a distribution is heavy-tailed then its tail function is heavy-tailed also. The distribution* F *is heavy tailed if its tail function goes slowly to zero at infinity. Distributions satisfying this definition are called Pareto type distribution given by:*

$$F(x) = 1 - x^{-\alpha}, \qquad \alpha > 0.$$

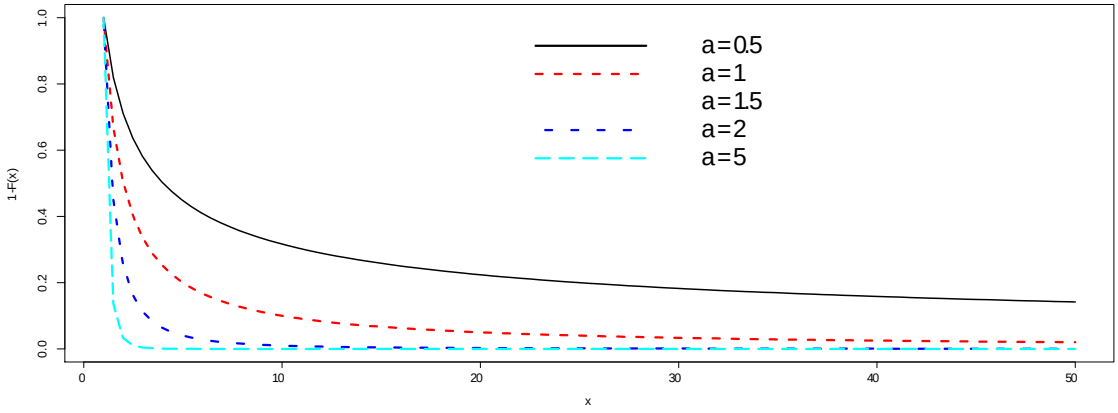

Figure 1.4: Tail function of Pareto distribution for $\alpha = 0.5, 1, 1.5, 2$ and 5.

1.1.1 Examples of heavy-tailed distributions

i) Pareto distribution : This distribution has tail function $\bar{F}$ given by

$$\bar{F}(x) = \left(\frac{c}{x+c}\right)^{\alpha},$$

for parameters $c > 0$ and $\alpha > 0$. Clearly we have $\bar{F}(x) \sim (x/c)^{-\alpha}$ as $x \to \infty$. The Pareto distribution has all moments of order $\lambda < \alpha$ finite, while all moments of order $\lambda \geq \alpha$ are infinite.

ii) Burr distribution (a model for losses in insurance): Has tail function $\bar{F}$ given by

$$\bar{F}(x) = \left(\frac{c}{x^{\tau}+c}\right)^{\alpha},$$

for parameters $\alpha, c, \tau > 0$. We have $\bar{F}(x) \sim c^{\alpha} x^{-\tau\alpha}$ as $x \to \infty$, thus the Burr distribution is similar in its tail to the Pareto distribution, of which it is otherwise a generalization. All moments of order $\lambda < \alpha\tau$ are finite, while those of order $\lambda \geq \alpha\tau$ are infinite.

iii) Cauchy distribution : Recall that the density of the standard Cauchy distribution is

$$f(x) = \frac{1}{\pi(1+x^2)}, \qquad x \in \mathbb{R},$$

and its distribution function is

$$F(x) = \frac{1}{2} + \frac{\arctan x}{\pi},$$

and hence

$$\bar{F}(x) = \frac{1}{2} - \frac{\arctan x}{\pi},$$

we see that $\bar{F}(x) \approx (\pi x)^{-1}$, as $x \to \infty$, its tail goes to zero like the power function x^{-1}. All moments are infinite.

iv) Lognormal distribution : The tail of the distribution F is then

$$\bar{F}(x) = \bar{\Phi}\left(\frac{\log x - \mu}{\sigma}\right) \qquad \text{for } x > 0,$$

for parameters μ and $\sigma > 0$, where $\bar{\Phi}$ is the tail of the standard normal random variable $N(0,1)$. All moments of the lognormal distribution are finite.

v) Weibull distribution: This has tail function $\bar{F}$ given by

$$\bar{F}(x) = e^{-(x/c)^{\alpha}}, \qquad x \geq 0.$$

for some scale parameter $c > 0$ and shape parameter $\alpha > 0$. This is a heavy-tailed distribution if and only if $\alpha < 1$.

1.1.2 Regularly varying distribution functions

An important class of heavy tailed distributions is the class of regularly varying distribution functions. A more detail is found in [Bingham et al. (1987)].

Definition 1.1.3 *A distribution F (or r.v. X) is called a regularly varying at infinity with index $(1/\gamma =: \alpha)$, i.e, $\overline{F} \in RV_{-1/\gamma}$. That is*

$$\bar{F}(x) = P(X > x) = x^{-1/\gamma}\ell(x), \qquad \forall x > 0,$$

where $\gamma > 0$ is the so-called extreme value index (e.v.i), the parameter α is called the tail index. $\ell(x)$ is a slowly varying function $(\ell \in \mathcal{RV}_0)$.

Remark 1.1.3 *1) A positive, measurable function $\ell(x)$ on $(0, \infty)$ is called a slowly varying function at infinity if*

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

2) Examples of $\ell(x)$ are given by $clnx$, $cln(lnx)$ and all functions converging to positive constants.

Remark 1.1.4 *For the regularly varying distribution functions, the $k-th$ moment does not exist whenever $k \geq 1/\gamma$. This has a few important implications. When we consider the sum of iid rv's that has a regularly varying distributions with a tail index $\alpha < 2$, the variance of these rv's is infinite, and hence the central limit theorem does not hold.*

Example 1.1.1 *1) If X has a Pareto distribution with tail $\bar{F}(x) = \left(\frac{c}{x+c}\right)^{\alpha}$, then $\overline{F} \in RV_{-\alpha}$.*
2) If X has a Burr distribution with tail $\bar{F}(x) = \left(\frac{c}{x^{\tau}+c}\right)^{\alpha}$, then $\overline{F} \in RV_{-\tau\alpha}$.

1.2 Extreme value theory

Extreme value theory (EVT) has emerged as one of the most important statistical disciplines for the applied sciences. Their techniques are also becoming widely used in many other disciplines. In particular, extreme value analyses usually require estimation of the probability of events that are more extreme than any that have already been observed. For example, EVT might be used in the field of hydrology to estimate the probability of an unusually large flooding event.

1.2.1 Extreme value distributions

Let $X_1, .., X_n$ be iid rv's representing risks or losses with unknown cumulative distribution function (cdf), $F(x) = P(X \leq x)$. Examples of random risks are returns on financial markets or portfolios, operational losses, catastrophic insurance claims, credit losses, natural disasters, traffic prediction in telecommunications etc. See [Coles (2001)], [McNeil and Frey (2000)], [Rachev (2003)], and [Embrechts et al. (1997)].

A traditional statistical discussion on the mean is based on the central limit theorem (CLT) and hence often returns to the normal distribution as a basis for statistical inference. The classical CLT states that the distribution of

$$\sqrt{n}\frac{\bar{X} - E(X)}{\sqrt{Var(X)}} = \frac{X_1 + ... + X_n - nE(X)}{\sqrt{nVar(X)}},$$

converges for $n \to \infty$ to a standard normal distribution $N(0,1)$. In general, the central limit problem deals with the sum $S_n := X_1 + ... + X_n$ and tries to find constants $a_n > 0$ and b_n such that $\frac{S_n - b_n}{a_n}$ tends in distribution to a non-

degenerate distribution. Typically the normal distribution is attained as a limit for this sum of iid rv's, except when the underlying distribution F possesses a heavy tail with infinite variance which yield non-normal limits for the average.

In what follows, we will replace the sum S_n by the maximum. The model focuses on the statistical property of :

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

Clearly, results for the minimum or the maximum can be immediately transferred to the other through the relation

$$X_{(1)} = \min(X_1, ..., X_n) = -\max(-X_1, ..., -X_n).$$

In theory the distribution of $X_{(n)}$ can be derived exactly for all values of n :

$$F_{X_{(n)}}(x) = P\left(X_{(n)} \leq x\right) = P(X_1 \leq x, ..., X_n \leq x) = F^n(x). \tag{1.2}$$

However, this is not immediately helpful in practice, since the distribution function F is unknown. An alternative approach is to accept that F is unknown, and to look for approximate families of models defined in (1.2), which can be estimated on the basis of the extreme data only. This is similar to the usual practice of approximating the distribution of sample means by the normal distribution, as justified by the central limit theorem.

It is natural to consider the probabilistic problem of finding the possible limit distributions of the maximum $X_{(n)}$. Hence, the main mathematical problem posed in extreme value theory concerns the search for distributions of X for which there

exist a sequence of numbers $\{b_n; n \geq 1\}$ and a sequence of positive numbers $\{a_n; n \geq 1\}$ such that for all real values x (at which the limit is continuous)

$$P\left(\frac{X_{(n)} - b_n}{a_n} \leq x\right) \to G(x) \qquad \text{as } n \to \infty.$$

This problem has been solved in [Fisher and Tippett (1928)], and [Gnedenko (1943)] and was later revived and streamlined by [de Haan (1970)], by the following theorem which is an extreme value analog of the CLT.

Theorem 1.2.1 (Fischer-Tippett, 1928 and Gnedenko, 1943) *Let (X_i) be independent identically distributed random variables with distribution function F. If there exist two real valued sequences $a_n > 0$ and $b_n \in \mathbb{R}$ and a distribution function G such that:*

$$\frac{X_{(n)} - b_n}{a_n} \xrightarrow{\mathcal{D}} G_\alpha.$$

Then, if $\alpha > 0$

$$G_\alpha(x) = e^{(-x)^{-\alpha}} 1_{(x>0)}(x),$$

if $\alpha < 0$

$$G_\alpha(x) = \begin{cases} e^{-(-x)^{\alpha}}, & x \leq 0 \\ 1, & x > 0 \end{cases}$$

and if $\alpha = 0$

$$G_0(x) = e^{-e^{-x}}, \quad x \in \mathbb{R}.$$

Remark 1.2.1 *1) The previous theorem is true for most of the usual laws.*

2) The distribution function G_α is called ***Generalized Extreme Value Distribution.*** *The parameter α is called the extreme value index. If F verifies the precedent Theorem, we say that belongs to the domain of attraction of G_α.*

3) Within the sign of α there are three areas of attraction.

• *If $\alpha > 0$ we say that F belongs to the domain of attraction of Frechet. This domain of attraction contains the heavy tailed distribution functions (with polynomial decay) such as the Cauchy distribution, Pareto, Burr, inverse gamma and log gamma distributions.*

• *If $\alpha < 0$ we say that F belongs to the domain of attraction of Weibull. This domain of attraction contains the majority of distribution functions whose end point is finite (uniform law, $Beta(p, q)$, Weibull distributions etc.)*

• *If $\alpha = 0$ we say that F belongs to the domain of attraction of Gumbel. This domain of attraction contains the distribution functions (with exponential decay) such as Gaussian, exponential, gamma, lognormal, Logistic, etc).*

The Fischer-Tippett Theorem is stating that the distribution function describing the dynamic of extreme events belongs to Maximum Domain of Attraction of a Generalized Extreme Value Distribution, that is:

Definition 1.2.1 *The Generalized Extreme Value Distribution $G_{\tau,\mu,\sigma}(z)$, is defined by*

$$G_{\alpha,\mu,\sigma}(z) = \begin{cases} \exp\left(-\left(1+\alpha\dfrac{z-\mu}{\sigma}\right)^{-1/\alpha}\right) & \alpha \neq 0 \\ \exp\left(-\exp\left(-\dfrac{z-\mu}{\sigma}\right)\right) & \alpha = 0, \end{cases}$$

$G_{\alpha,\mu,\sigma}(z)$ is defined on $\{z : 1+\alpha(z-\mu)/\sigma > 0\}$, where $\mu \in \mathbb{R}$, $\sigma > 0$, and the real parameter α is a shape parameter that determines the tail behavior of $G_\alpha(z)$.

[Gnedenko (1943)] accomplished an important result on this issue. He proved that The Fischer-Tippett theorem is applicable for heavy tailed distributions functions. More precisely, he shown that heavy tailed distribution functions belong to the maximum domain of attraction of the Frechet distribution.

1.2.2 Tail index estimation

The estimation of α has a great interest and common applications in a big variety of domains, as for example in economics, applied finance, insurance, business, industry, traffic, telecommunications, sociology and geology, as one might see the textbook, [Dekkers *et al.* (1989)], [Bacro and Brito (1995)], [Beirlant *et al.* (2007)] and references therein.

The tail index is used for the estimation of high quantiles of observed rv's. For many applications, it is important to know αas well as to determine the number of finite moments. For example, if $\alpha < 2$, than $EX^2 = \infty$ holds.

There are numerous tail-index estimators. They are based on various assumptions, have diverse asymptotic and finite-sample properties. The paper of [Fedotenkov (2018)] reviews more than one hundred univariate Pareto-type (and equivalent) tail index estimators : Hill, Kernel, Pickands, Peng,...

In the case of a positive-valued tail index $(\alpha > 0)$, the most celebrated estimator of α is that proposed by [Hill (1975)], for $\gamma = 1/\alpha$, is determined by

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

$X_{1,n} \leq ... \leq X_{n,n}$ denote the usual order statistics of the observed samples, $k = k_n$ is the number of top statistics (upper observations), k is an integer sequence satisfying

$$1 < k < n, \quad k \to \infty \quad \text{and } k/n \to 0 \quad \text{as } n \to \infty. \tag{1.3}$$

The asymptotic properties of $\hat{\gamma}^H$ have been much studied. In the independent context, it is well known that, under some regularity conditions, $\hat{\gamma}^H$ is strongly con-

sistent with asymptotic normal distribution [Haeusler and Teugels (1985)] when properly normalized. The consistency of $\hat{\gamma}^H$ was proved by [Mason (1982)] by only assuming the regular variation condition while its asymptotic normality was established under a suitable extra assumption, known as the second-order regular variation condition (see [de Haan and Stadtmüller (1996)] and [de Haan and Ferreira (2006)]). In particular, under (1.3) Hill's estimator is weakly consistent

$$\hat{\gamma}^H \xrightarrow{P} \gamma, \qquad \text{as } n \to \infty.$$

and asymptotically normal with mean γ and variance γ^2/k :

$$\sqrt{k}\left(\frac{\hat{\gamma}^H - \gamma}{\gamma}\right) \xrightarrow{D} N(0,1), \qquad \text{as } n \to \infty.$$

Plots of Hill $\hat{\gamma}^H$ estimator against k is shown in Figure 1.5 for Pareto distribution, with parameter $\gamma = 0,7$. The sample size is $n = 1000$.

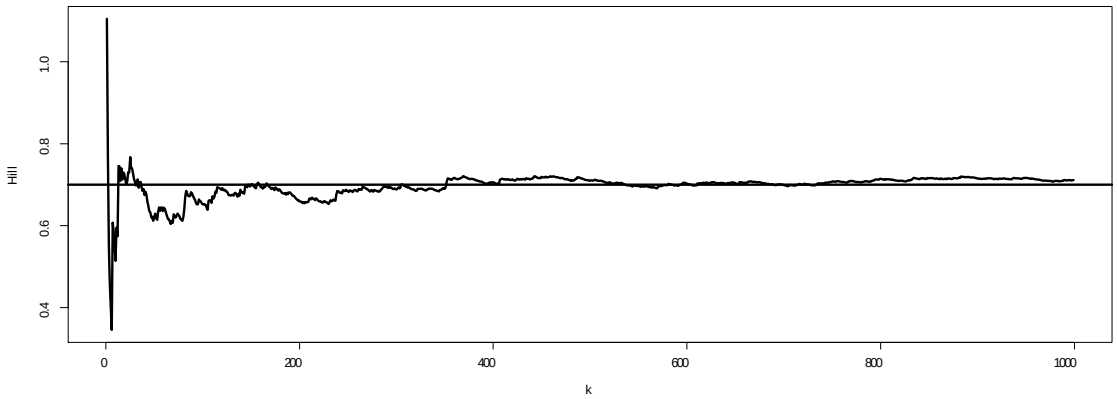

Figure 1.5: Hill $\hat{\gamma}^H$ estimator against k for Pareto distribution, with parameter $\gamma = 0,7$.

1.2.3 Robust estimation vs upper quantile

A given estimator is said to be robust if small changes in its distribution have a relatively small effect on its value. The theory of robustness aim at :

i) Stability for small samples : In general, small sample size n leads to unstable estimates.

ii) Resistant to errors in the results: Produced by deviations from assumptions on the ϵ-contaminated model known by mixture distributions:

$$F_{\epsilon}(x) = (1-\epsilon) F(x) + \epsilon G(x).$$

iii) Sensitivity to the presence of outliers: An outlier may be due to measurement error, experimental error or it may indicate that the population has a heavy-tailed distribution with extremes values.

Example 1.2.1 *The sample mean $\bar{X}$, is not robust. Taking the dataset $\{1,3,5,8,13\}$. If we add another datapoint with value 100 to the data, the resulting mean $\left(\bar{X} = \frac{130}{6} = 21.67\right)$ will be very different to the mean of the original data $\left(\bar{X} = \frac{30}{5} = 6\right)$.*

Example 1.2.2 *The median is a robust statistic : Taking the same dataset $\{1,3,5,8,13\}$, if we add another datapoint with value 100 then the median will change slightly $(Med = 6.5)$, but it will still be similar to the median of the original data $(Med = 5)$.*

Readers interested in mathematical issues on robustness can refer to the excellent books by [Huber (1981)], [Hampel *et al.* (1986)] as well as [Barnett and Lewis (1995)].

Estimation of e.v.i. is very important in the determination of high quantiles, upper tail probabilities, mean excess functions, and excess-of-loss and stop-loss reinsurance premiums. Consequently, small errors in estimation of this quantity can produce substantial impact in applications. Thus, for robust estimation of quantities based on α robust estimation of α itself is crucial.

A useful parametric model with relatively high probability in the upper tail is the Pareto distribution having cdf

$$F(x) = 1 - x^{-\alpha}, \quad x \geq 1. \tag{1.4}$$

defined for $\alpha > 0$.

For estimation of upper (high) quantile, let $\varepsilon \downarrow 0$ above a specified threshold $Q = F^{-1}$, it follows from equation (1.4) that

$$x = (1 - F(x))^{-1/\alpha} \Rightarrow Q_\varepsilon = \varepsilon^{-1/\alpha}.$$

Thus, for the estimator $\hat{Q}_\varepsilon$, defined by putting $\hat{\alpha}$ for α, we have

$$\frac{\hat{Q}_\varepsilon}{Q_\varepsilon} = \varepsilon^{1/\alpha - 1/\hat{\alpha}}.$$

1) For $\varepsilon = 0.001$, $\alpha = 1$ and $\hat{\alpha} = 1.1$ we get

$$\frac{\hat{Q}_\varepsilon}{Q_\varepsilon} = \varepsilon^{1/\alpha - 1/\hat{\alpha}} = (0.001)^{1 - 1/1.1} = 0.53.$$

Consequently, overestimation of $\alpha = 1$ by 10% produces underestimation of $Q_{0,001}$ by 47%.

2) For $\varepsilon = 0.001$, $\alpha = 1.5$ and $\hat{\alpha} = 1.65$ we get

$$\frac{\hat{Q}_\varepsilon}{Q_\varepsilon} = (0.001)^{1/1.5-1/1.65} = 0.66.$$

Thus, overestimation of $\alpha = 1.5$ by 10% produces underestimation of $Q_{0,001}$ by 34%.

3) Likewise, underestimation of $\alpha = 1$by only 5%produces overestimation of $Q_{0,001}$by 44% :

$$\hat{\alpha} = 0.95 \quad \text{and} \quad \frac{\hat{Q}_\varepsilon}{Q_\varepsilon} = (0.001)^{1-1/0.95} = 1.44.$$

4) In the case of underestimation of $\alpha = 1.5$ by 5%, we get

$$\hat{\alpha} = 1.425 \quad \text{and} \quad \frac{\hat{Q}_\varepsilon}{Q_\varepsilon} = (0.001)^{1/1.5-1/1.425} = 1.27.$$

which produces overestimation of $Q_{0,001}$ by 27%.

In conclusion, for a heavy tailed distributions, robust estimation of the high quantile Q_ε corresponding to upper tail probability ε, becomes of interest, and this may be carried out by robust estimation of α. [Brazauskas and Serfling (2000)] gives a detailed account of this issue.

Chapter 2

Incomplete data context

A survival model is based on lifetime. This term (lifetime) is a positive and continuous random variable Xdesignating the necessarily time to the appearance of a precise event or to pass from a state A to a state B in general. The survival model and lifetime study can occur in many areas of application: such as in medicine (length of survival after a heart attack), in engineering (operating time of a component), finance (time between 2 successive failures of a device), social sciences (duration of unemployment),...

Generally, the data to treat are not completely observed. In this case classical techniques don't adjust correctly to the incomplete data (truncation or censoring). Since our work carries on the incomplete data, and in order to give back easy the reading of this thesis, we give some definitions, examples, important and useful results on this issue of the incomplete data context which occur quite naturally in lifetime data analysis.

2.1 Censoring

Censoring is when a data is incompletely observed due to same random case (problem of missing data, end of the study,...). In such case, the observed value of same variables is only partially known.

2.1.1 Censoring Types

1) **Right censoring:** Commonest form of censoring is right censoring. Subjects followed until some time, at which the event has yet to occur, but then talks no further part in the study. This may be because:

- the subject dies from another cause, independently of the cause of interest,
- the study ends while the subject survives,
- the subject is lost to the study, by dropping out, moving to a different area.

So the lifetime X is only known to be greater than a censoring time denoted C. In this case, we have:

The exact lifetime X of an individual will be known if $X \leq C$.

An individual is survivor and his event time is censored from the right at $X > C$.

2) **Left censoring:** A lifetime X is left censoring if it is less than a censoring time denoted C. The event of interest has already occurred for individual before that person is observed in the study at the censoring time:

X while be known if $X \geq C$.

X is censored from the left at C if $X < C$.

Example 2.1.1 *In a study of the learning time of a task for some children. This lifetime is a random variable X and C is the age of the child. For children who already know perform the task, C censors X from the left because for them the learning time X is unknown but inferior at $C : X < C$.*

3) **Interval censoring:** A data point is somewhere on an interval between two dates C_1, C_2 in which the event of interest occurred $(C_1 < X < C_2)$. Left and right censoring are special cases of interval censoring, with the beginning of the interval at zero or the end at infinity, respectively.

   This type occurs in medical study (when a person have a periodic follow-up), in industrial experiments (where there is a periodic inspection for proper functioning or ecupement items),...

2.1.2 Framework of randomly right censoring

Let $X_1, ..., X_n$ be n copies of iid rv X, with cdf F assumed to be heavy-tailed. In many real applications, such as survival analysis, reliability theory or insurance..., the variable of interest X is not necessarily completely available. This is the case in the presence of random right censoring.

The usual way to model this situation is to introduce another random variable C called censoring rv, independent of X. In this case the data can be represented by pairs of rv's (Z_i, δ_i) where

$$Z_i := \min(X_i, C_i), \qquad \delta_i := \mathbf{1}_{(X_i \leq C_i)}, \qquad i = 1, ..., n.$$

Remark 2.1.1 *If our data contain only uncensored and right-censored data, we can represent all individuals by the triple (i, Z_i, δ_i) :*

i : *indexes subjects,*

Z_i : *is the time at which the death or censoring event occurs to individual* j*,*

δ_i : *is an indicator variable which determines whether or not* X_i *has been observed:* $\delta_i = 1$ *if* i *is uncensored and* $\delta_i = 0$ *if censored.*

Example 2.1.2 *An artificial example of right-censored data involving* $n = 10$ *data points is given in Table 2.1. We see that the observed sample is given by*

$$(Z_1, Z_2, Z_3, Z_4, Z_5, Z_6, Z_7, Z_8, Z_9, Z_{10}) = (C_1, X_2, C_3, X_4, X_5, C_6, X_7, C_8, X_9, X_{10})$$

for which $Z_j = \min(X_j, C_j)$ *. The indicator variable*

$$\delta_j = 1_{(X_j < C_j)} = \begin{cases} 1, & \text{if } X_j \text{ has been observed} \\ 0, & \text{if } X_j \text{ is censored} \end{cases}$$

$$= (0, 1, 0, 1, 1, 0, 1, 0, 1, 1).$$

j	X_j	C_j	$Z_j = \min(X_j, C_j)$	δ_j
1	11.12	9.15	9.15	0
2	10.85	11.45	10.85	1
3	10.25	9.68	9.68	0
4	10.02	10.25	10.02	1
5	9.08	9.88	9.08	1
6	7.65	7.54	7.54	0
7	6.63	7.65	6.63	1
8	5.41	4.03	4.03	0
9	5.21	8.03	5.21	1
10	5.03	5.77	5.03	1

Table 2.1: An artificial example of right-censored data.

In the context of this randomly right censoring, the nonparametric maximum likelihood estimator of F in the case of censored data equals the famous the

Kaplan-Meier estimator [Kaplan and Meire (1958)], given by

$$1 - F_n^{(KM)}(z) := \prod_{i:Z_{i,n} \leq z} \left(1 - \frac{\delta_{i,n}}{n-i+1}\right), \quad \text{for } z \in \mathbb{R}$$

where $Z_{i,n}$ denote the order statistics associated to $Z_1, ..., Z_n$ and $\delta_{i,n}$ is the concomitant of the $ith-$order statistics, for which $\delta_{i,n} = \delta_i$ if $Z_{i,n} = Z_i$.

Remark 2.1.2 *When there is no censorship, the Kaplan-Meier estimator* $F_n^{(KM)}(z)$ *is reduced to the empirical distribution function* $F_n(z) = \frac{1}{n}\sum_{i=1}^{n} 1_{(X_i \leq z)}$.

2.2 Truncation

The censored data are not the unique type of incomplete data. The other case is the one of the so-called truncated data. Truncation appears when time to event is only observed in study. A rv X is said to be truncated from below if, for some threshold value T, the exact value of X is known for all cases $X > T$, but unknown for all cases $X \leq T$. Similarly, truncation from above means the exact value of X is known in cases where $X < T$, but unknown for all cases $X \geq T$.

2.2.1 Truncation Types

1) **Right truncation:** In many real applications, the rv of interest X may be not fully available (truncated from the right by a rv denoted Y). In this case, only individuals with event time less than some threshold are included in the study.

Example 2.2.1 *If we ask a group of smoking school pupils at what age they started smoking, we have necessarily a truncated data. Individuals who start*

smoking after leaving school are not included in the study, and therefore right truncated.

Example 2.2.2 *An artificial example of right-truncated data involving* $N = 10$ *data points is given in Table 2.2. We see that* $(X_1, X_3, X_5, X_6, X_7, X_8, X_9)$ *are observed, but not* (X_2, X_4, X_{10})*. In this case, the observed sample is given by*

$$(X_1^*, X_2^*, X_3^*, X_4^*, X_5^*, X_6^*, X_7^*) = (0.56, 0.10, 2.61, 3.55, 6.09, 4.01, 8.18)$$

j	X_j	Y_j	X_j^*
1	0.56	8.05	0.56
2	0.14	0.11	
3	0.10	1.12	0.10
4	1.5	0.98	
5	2.61	3.35	2.61
6	3.55	4.58	3.55
7	6.09	8.07	6.09
8	4.01	5.03	4.01
9	8.19	9.81	8.19
10	7.15	6.14	

Table 2.2: An artificial example of right-truncated data.

The truncation probability is about 30%, thus: $p = P(X \leq Y) \simeq \frac{n}{N} = \frac{7}{10} = 0.7$.

2) **Left truncation:** due to structure of the study design, we can only observe those individuals whose event time is greater than some truncation threshold.

Example 2.2.3 *Imagine you wish to study how long people who have been hospitalized for a heart attack survive taking some treatment at home. The start time*

is taken to be the time of the heart attack. Only those individuals who survive their stay in hospital are able to be included in the study.

2.2.2 Estimation under right truncation

Let X, denote a rv of interest, with continuous cdf F and let Y, denote the truncation rv with cdf G. Consequently, the observed subsequence denote (X_i^*, Y_i^*); $1 \leq i \leq n$ subject to $X_i^* \leq Y_i^*$ from the N-sample (X_j, Y_j); $1 \leq j \leq N$, $(n \leq N)$. Then, the size of actually observed sample,

$$n := \sum_{j=1}^{N} 1_{(X_j \leq Y_j)}$$

is a binomial rv with parameters N and $p := P(X \leq Y)$. We shall assume that $p > 0$, otherwise, nothing will be observed.

Since, the joint cdf of (X^*, Y^*) is

$$\begin{aligned} H(x,y) &= P(X^* \leq x, Y^* \leq y) = P(X \leq \min(x,Y), Y \leq y | X \leq Y) \\ &= p^{-1} \int_0^y F(\min(x,t)) dG(t). \end{aligned}$$

The marginal cdf's of the observed rv's X^* and Y^* (subject to $X^* \leq Y^*$), respectively denoted by F^* and G^*, becomes

$$F^*(x) = P(X^* \leq x) = P(X \leq x | X \leq Y) := p^{-1} \int_0^x \overline{G}(t) dF(t)$$

and

$$G^*(y) = P(Y^* \leq y) = P(Y \leq y | X \leq Y) := p^{-1} \int_0^y F(t) dG(t),$$

Therefore, the Woodroofe's nonparametric estimator, see [Woodroofe (1985)] of

F, is defined by

$$F_n^{(W)}(x) := \prod_{j:X_j^*>x} \exp\left(-\tfrac{1}{nC_n(X_j^*)}\right), \quad \text{where } C_n(x) := \frac{1}{n}\sum_{j=1}^{n} 1_{\left(X_j^* \le x \le Y_j^*\right)},$$

in which C_n is the empirical estimator of

$$C(z) := P(X^* \le z \le Y^*) = P(X \le z \le Y | X \le Y) = p^{-1}\overline{G}(z)F(z).$$

Another more popular estimator for F, is the well known Lynden-Bell nonparametric maximum likelihood estimator, originally proposed by [Lynden-Bell (1971)], defined by

$$F_n^{(LB)}(x) := \prod_{j:X_j^*>x} \left(1 - \tfrac{1}{nC_n(X_j^*)}\right).$$

Finally, readers interested on the incomplete data, lifetime and survival analysis can refer to [Lawless (2011)], [Escudero and Ortega (2008)] and references therein.

Chapter 3

Robust tail index estimation under censoring

Statistics of extremes under random censoring is a new research field. The topic was first mentioned in [Reiss and Thomas (2007)], where an estimator of a positive extreme value index was introduced, but with no asymptotic results. Recently, [Beirlant *et al.* (2007)] proposed an estimators for the general extreme value index and for the extreme quantile with their asymptotic properties. [Einmahl *et al.* (2008)] adapted various extreme value index estimators to the case where the data are censored, by a random threshold and establish their asymptotic normality by imposing some assumptions that are rather unusual to the context of extreme value theory. Robust estimation of e.v.i. focuses primarily on complete data case, see [Brazauskas and Serfling (2000)], [Beran and Shell (2012)] and references therein. The incomplete data case has first been considered by [Sayah et al. (2014)], who dealt with single Pareto distributions under right censored data.

The aim of the current chapter is to provide a robust e.v.i. estimator for the

general case of heavy tailed Pareto-type distributions under random censorship. Based on the ideas of Kaplan-Meier integration using the huberized M-estimator of the tail index. We derive the asymptotic results and we illustrate (in a simulation study) the performance and the robustness of this newly estimator.

3.1 Hill-type estimator under censoring

Let $X_1, ..., X_n$ be n copies of iid rv X, with cdf F assumed to be heavy-tailed. In other words, the distribution tail $\overline{F} := 1 - F$ is regularly varying, with index $(-\alpha_1)$, notation: $\overline{F} \in RV_{(-\alpha_1)}$. That is

$$\lim_{t\to\infty} \frac{\overline{F}(tx)}{\overline{F}(t)} = x^{-\alpha_1}, \text{ for any } x > 0,$$

where $\alpha_1 > 0$ is the so-called shape parameter or tail index.

The estimation of α_1 has a great interest (see, chapter 1). The most celebrated estimator of α_1 is that proposed by [Hill (1975)]

$$\hat{\alpha}_1^H := \hat{\alpha}_1^H(k) = \left(\frac{1}{k} \sum_{i=1}^{k} \log(X_{n-i+1,n}) - \log(X_{n-k,n}) \right)^{-1},$$

for $k = k_n$ is an integer sequence satisfying

$$1 < k < n, \quad k \to \infty \quad \text{and } k/n \to 0 \quad \text{as } n \to \infty. \tag{3.1}$$

The asymptotic properties of $\hat{\alpha}_1^H$ have been much studied. In the independent context, it is well known that, under some regularity conditions, $\hat{\alpha}_1^H$ is strongly consistent with asymptotic normal distribution when properly normalized [Haeusler and Teugels (1985)]. The consistency of $\hat{\alpha}_1^H(k)$ was proved by

[Mason (1982)] by only assuming the regular variation condition while its asymptotic normality was established under a suitable extra assumption, known as the second-order regular variation condition (see [de Haan and Stadtmüller (1996)] and [de Haan and Ferreira (2006)]).

As shown in chapter 2, in many real applications, such as survival analysis, reliability theory or insurance...(in the presence of random right censoring), the variable of interest X is not necessarily completely available. The usual way to model this situation is to introduce a random variable C called censoring rv, independent of X, and then to consider the rv $Z := \min(X, C)$ and the indicator variable $\delta := 1(X \leq C)$, which determines whether or not X has been observed. The cdf's of C and Z will be denoted by G and H respectively.

The tail of the censoring distribution G is assumed to be regularly varying too, that is $1 - G \in RV_{(-\alpha_2)}$, for some $\alpha_2 > 0$. By virtue of the independence of X and C, we have

$$1 - H(x) = (1 - F(x))(1 - G(x)).$$

Therefore $1 - H \in RV_{(-\alpha)}$, with $\alpha := \alpha_1 + \alpha_2$. Let $\{(Z_i, \delta_i),\ 1 \leq i \leq n\}$ be a sample from the couple of rv's (Z, δ) and $Z_{1,n} \leq ... \leq Z_{n,n}$ represent the order statistics pertaining to $(Z_1, ..., Z_n)$.

Definition 3.1.1 *The adapted Hill estimator proposed by [Einmahl et al. (2008)] of the tail index α_1 is defined by*

$$\widehat{\alpha}_1^{(H,c)} := \frac{\widehat{\alpha}^H(k)}{\widehat{p}}, \tag{3.2}$$

where

$$\widehat{\alpha}^{H}(k) := \frac{1}{k}\sum_{i=1}^{k} \log Z_{n-i+1,n} - \log Z_{n-k,n} \tag{3.3}$$

and

$$\widehat{p} := \frac{1}{k}\sum_{i=1}^{k} \delta_{n-i+1,n}, \tag{3.4}$$

the proportion of non-censored observations in the largest k order statistics of Z, with $\delta_{i,n}$ (i.e. $\delta_{i,n} = \delta_j$ if $Z_{i,n} = Z_j$) denote the concomitant of the ith order statistic $Z_{i,n}$ and $k := k_n$ satisfying (3.1).

Remark 3.1.1 *The adapted Hill estimator (3.2) is equal to the quotient of the classical Hill estimator to the proportion of non censored data $p = \frac{1}{n}\sum 1_{(X \leq C)}$ based on k largest order statistics.*

3.2 Robust tail index estimation and asymptotic results

The adapted Hill estimator (3.2), as well as the classical Hill-type (in complete data case) are non robust, in the sense that they are very sensitive to extreme observations, data contamination or model deviation and tend to be highly volatile for small samples (this is illustrated in our simulation study). Also, the rate of convergence of these estimators are based on the optimal value of the numbers of top statistics k, but this rate are slower than the parametric rate $\sqrt{n}$. Moreover, estimating the optimal value of k is virtually impossible when the sample size n is small and this leads to unstable estimates for small samples. The alternative approach is inspired from the theory of robust inference, which aim at stability for small samples. However, as observed by [Beran and Shell (2012)], in some prac-

tical cases, such as natural disasters, operational risk assessment or reinsurance data are sparse (with n often somewhere between 20 and 50) and distributions are expected to be heavy tailed with an unknown e.v.i.

Historically, several approaches to robust estimation were proposed, including R-estimators, L-estimators and M-estimators as a result of a generalization of the (non-robust) maximum likelihood estimators. Robust estimation of e.v.i. focuses primarily on complete data case, see [Brazauskas and Serfling (2000)], [Beran and Shell (2012)] and references therein. This can be obtained by the definition of the following class of M-functional and M-estimators respectively that are defined as follows (see, [Beran and Shell (2012)]).

Definition 3.2.1 *Let* $F_{Par}(x,\alpha) = 1 - x^{-\alpha}$ $(x \geq 1)$ *and*

$$\begin{aligned}\psi_v(x,\alpha) &= [\alpha \log(x) - 1]_v - \int [\alpha \log(z) - 1]_v dF_{Par}(z,\alpha) \\ &= [\alpha \log(x) - 1]_v - (v + \exp(-(v+1))),\end{aligned}$$

where $[y]_v = \max(y, v)$, *and denote by* F *a set of distributions with support in* $\mathbb{R}_+$. *Then the functional* $T =: T^v$ *defined on* F *as the solution* t_0 *of the equation*

$$\beta_F(t) = \int \psi_v(x,t)\, dF(x) = 0, \qquad (F \in \mathcal{F})$$

is called huberized tail index M*-functional. The corresponding* M*-estimator* $T_n =: T_n^v$, *defined by*

$$\sum_{j=1}^{n} \psi_v(X_j, T_n) = 0, \tag{3.5}$$

is called huberized M*-estimator of the tail index. Moreover,* $\psi_v(.,\alpha)$ *is defined for any choice of* $\alpha > 0$ *and* $v \in \mathbb{R}$. *Thus, as shown by [Beran and Shell (2012)],*

robustness on the left is obtained only if $v \geq -1$.

Remark 3.2.1 *This approach of huberized M-estimator is inspired from the theory of robust inference (see, for instance, [Huber (1981)]).*

The nonparametric maximum likelihood estimator of F in the case of censored data equals the famous estimator of [Kaplan and Meire (1958)] also called the product limit estimator, given by

$$1 - F_n^{(KM)}(z) := \prod_{i:Z_{i,n} \leq z} \left(1 - \frac{\delta_{i,n}}{n-i+1}\right), \text{ for } z \in \mathbb{R}. \tag{3.6}$$

[Stute and Wang (1993)] and [Stute (1995)] studied the almost sure and distributional behavior of the so-called Kaplan-Meier integrals

$$I_n := \int \varphi(z)\, dF_n^{(KM)}(z),$$

where φ is an arbitrary integrable function. It is easily seen from (3.6) that

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

where for $1 \leq i \leq n$

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

[Stute (1995)] obtained under random censoring and under some assumptions the central limit theorem for a general transformation φ, that is

$$\sqrt{n} \int \varphi d\left(F_n^{(KM)} - F\right) \xrightarrow{D} \mathcal{N}\left(0, \sigma^2\right), \qquad \left(\sigma^2 < \infty\right).$$

To formulate our main results, the following assumptions are required:

$$\int \psi_v^2(x,\alpha_1)\,\lambda_0^2(x)\,\tilde{H}_1(dx) < \infty, \tag{3.7}$$

and

$$\int |\psi_v(x,\alpha_1)|\,A^{1/2}(x)\,F(dx) < \infty, \tag{3.8}$$

where

$$A(x) := \int_{-\infty}^{x} \frac{G(dy)}{(1-H(y))(1-G(y))}.$$

The functions $\tilde{H}_0$,$\tilde{H}_1$, λ_0, λ_1 and λ_2 are defined as below

$$\tilde{H}_0(z) := P(Z \le z, \delta = 0) = \int_{-\infty}^{z} \bar{F}(t)\,G(dt),$$

$$\tilde{H}_1(z) := P(Z \le z, \delta = 1) = \int_{-\infty}^{z} \bar{G}(t)\,F(dt),$$

$$\lambda_0(z) := \exp\left(\int_{-\infty}^{z} \frac{\tilde{H}_0(dx)}{\bar{H}(x)}\right),$$

$$\lambda_1(z) := \frac{1}{\bar{H}(z)} \int \psi_v(x,\alpha_1)\,\lambda_0(x)\,\mathbf{1}_{\{z<x\}}\tilde{H}_1(dx),$$

and

$$\lambda_2(z) := \int\int \frac{\psi_v(x,\alpha_1)\,\lambda_0(x)\,\mathbf{1}_{\{y<z,\ y<x\}}}{\bar{H}(y)}\tilde{H}_0(dy)\,\tilde{H}_1(dx).$$

Theorem 3.2.1 ([Sayah et al. (2014)]) *Let $X_i \sim F_{Par}(x,\alpha_1)$ and $C_i \sim F_{Par}(y,\alpha_2)$, $x \ge 1,\ y \ge 1$ where $\alpha_1 > 0$ and $\alpha_2 > 0$, with $\alpha_2 < \alpha_1$. Moreover, let $F_n = F_n^{(KM)}$ be the Kaplan-Meier estimator of the df F and $\widehat{\alpha}_1^{(S)}$ a sequence of solutions of*

$$\beta_{F_n}(t) = \sum_{j=1}^{n} W_{in}\psi_v(Z_{i,n},t) = 0, \qquad (n \in \mathbb{N}).$$

Then, under assumptions (3.7) and (3.8) we have

$$\sqrt{n}\left(\widehat{\alpha}_1^{(S)} - \alpha_1\right) \xrightarrow{D} \mathcal{N}\left(0, \sigma^2\right).$$

where $\sigma^2 < \infty$ *and expletively given by the authors in the proof of their theorem.*

Proof of Theorem 3.2.1. See the proof of Theorem 1, page 678-682 in Sayah et al. (2014).

■

Remark 3.2.2 *Condition (3.7) is the properly modified variance assumption on* ψ_v *and (3.8) only incorporates the first* ψ_v-*moment. It is mainly to control the bias of* $\int \psi_v d\hat{F}_n$*, which is a function of* ψ_v *rather than* ψ_v^2*. Stute (1994) and Stute (1995) gives a detailed account of this issue. In our case, this two assumptions are satisfied when* $\alpha_2 < \alpha_1$*.*

Next, we investigate the asymptotic properties of the estimator of the tail index α_1 under the large class of Pareto-type distributions assumptions.

Theorem 3.2.2 *Let* $\overline{F} \in RV_{(-\alpha_1)}$ *and* $\overline{G} \in RV_{(-\alpha_2)}$ *where* $\alpha_1 > 0$ *and* $\alpha_2 > 0$*, with* $\alpha_2 < \alpha_1$*. Assume that assumptions (3.7) and (3.8) hold. Then, provided the existence of* α_1 *as a unique solution of* $\beta_F(t) = 0$*, any solution sequence* $\widehat{\alpha}_1^{(Z)}$ *of*

$$\hat{\beta}_{F_n}(t) = \sum_{j=1}^{n} W_{in}\psi_v\left(Z_{i,n}, t\right) = 0 \qquad (n \in \mathbb{N})$$

is a consistent estimator of α_1*. Assume further that* $\int \frac{\partial}{\partial t}\psi_v(x,t)\, dF(x) \neq 0$ *hold in a neighborhood of* α_1*. Then*

$$\sqrt{n}\left(\widehat{\alpha}_1^{(Z)} - \alpha_1\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma_v^2\right), \quad \text{as } n \to \infty$$

where $\xrightarrow{d}$ *stands for convergence in distribution and*

$$\sigma_v^2 := \sigma^2 \left(\int \frac{\partial}{\partial t} \psi_v (x,t)\, dF(x) \right)^{-2}$$

in which

$$\sigma^2 = Var \left\{ -\psi(z,\alpha_1)\, \lambda_0(z)\, \delta + \lambda_1(z)(1-\delta) - \lambda_2(z) \right\}.$$

Remark 3.2.3 *It is worth mentioning that for complete data case (no censoring), we have,* $W_{in} = 1/n$ *and* $Z_i = X_i$ *so that* I_n *becomes the sample mean. It follows that* $\hat{\beta}_{F_n}\left(\widehat{\alpha}_1^{(Z)}\right) = \sum_{i=1}^{n} \psi_v(X_i, T_n)$ *and consequently* $\widehat{\alpha}_1^{(Z)}$ *reduce to the Beran and Shell estimator* T_n *see, e.g. [Beran and Shell (2012)].*

Proof of Theorem 3.2.2. The proof is essentially based on Theorem 1.1 in [Stute and Wang (1993), page 1594], Corollary 1.2 in [Stute (1995), page 426] and [Beran and Shell (2012), page 3432]. Recall that,

$$\int \psi_v\left(z, \widehat{\alpha}_1^{(Z)}\right) dF_n(z) - \int \psi_v(z,\alpha_1)\, dF(z) - \int \psi_v(z,\alpha_1)\, dF_n(z) + \int \psi_v(z,\alpha_1)\, dF_n(z)$$
$$= \int \left(\psi_v\left(z, \widehat{\alpha}_1^{(Z)}\right) - \psi_v(z,\alpha_1) \right) dF_n(z) + \int \psi_v(z,\alpha_1)\, d\left(F_n(z) - F(z)\right) = 0,$$

The assumed differentiability of $\psi_{v,u}(x,t)$ in t allows a Taylor expansion around α_1 which yields

$$\sqrt{n}\left(\widehat{\alpha}_1^{(Z)} - \alpha_1\right) \int \frac{\partial}{\partial t} \psi_v(z,t)\, dF_n(z) = \sqrt{n} \int \left(-\psi_v(z,\alpha_1)\right) d\left(F_n(z) - F(z)\right).$$

Then,

$$\sqrt{n}\left(\widehat{\alpha}_1^{(Z)} - \alpha_1\right) = \sqrt{n} \left(\int \left(\frac{\partial}{\partial t} \psi_v(z,t) \right) dF_n(z) \right)^{-1} \int \left(-\psi_v(z,\alpha_1)\right) d\left(F_n(z) - F(z)\right).$$

It was shown in Theorem 1.1 in [Stute and Wang (1993)] that for any measurable real function φ, and under the condition $\int |\varphi| \, dF < \infty$, we get

$$\int \varphi dF_n = \int \varphi dF + op\,(1)\,.$$

From [Stute (1995)] under assumptions (3.7) and (3.8) we have

$$\sqrt{n} \int \left(-\psi_v\left(z, \alpha_1\right)\right) d\left(F_n\left(z\right) - F\left(z\right)\right) \xrightarrow{D} \mathcal{N}\left(0, \sigma^2\right).$$

The stated assumptions are sufficient for consistency and asymptotic normality.

■

3.3 Performance and comparative study

In this section we examines the performance of our estimator and compare with the adapted Hill estimator given in 3.2 proposed by [Einmahl *et al.* (2008)]. For this reason, we follows the steps below.

Step 1: We generate 1000 pseudorandom samples X and C of size $n = 100$, 200, 500 from Pareto cdfs with $\alpha_1 = 0.6$ and $\alpha_2 = 0.25$ respectively:

$$\overline{F}\left(x\right) = x^{-\alpha_1} \quad and \quad \bar{G}\left(x\right) = x^{-\alpha_2}, \qquad x \geq 0.$$

Here $v = 1$ and $p = 0.70$ that means the percentage of censorship is 30%.

Step 2: We obtained 1000 pseudorandom samples $Z = \min\left(X, C\right)$ and the indicator variable $\delta := 1_{(X \leq C)}$ of size $n = 100$, 200 and 500.

Step 3: We estimate the tail index parameter by the two estimators based on

the observed data Z.

Step 4: For choosing the optimal number $k = k_n$ of upper order statistics used in the computation of the adapted Hill estimator given in 3.2 we adopt the Reiss and Thomas algorithm [Reiss and Thomas (2007), page 137].

Step 5: Finally, we compute the absolute bias (abias) and root mean squared error (rmse) of these estimators :

$$abias\left(\widehat{\alpha}_1^{(Z)}\right) = \left|E\left(\widehat{\alpha}_1^{(Z)}\right) - \alpha_1\right| \simeq \left|\frac{1}{N}\sum_{j=1}^{N}\widehat{\alpha}_{1j}^{(Z)} - \alpha_1\right|$$

$$\text{and} \quad rmse = \sqrt{bias^2\left(\widehat{\alpha}_1^{(Z)}\right) + var\left(\widehat{\alpha}_1^{(Z)}\right)}.$$

The results are summarized in Table 3.1.

	$\widehat{\alpha}_1^{(Z)}$		k	$\widehat{\alpha}_1^{(H,c)}(k)$	
n	$abias$	$rmse$		$abias$	$rmse$
100	.0611	.2511	17	.1143	.2586
200	.0431	.1821	34	.0845	.1013
500	.0153	.1142	86	.0245	.0684

Table 3.1: Bias and RMSE of the two estimators based on 1000 samples of Pareto-distributed with tail index 0.6.

We see that our new estimator shows good performance for small sample sizes.

3.4 Comparative robustness study

We study the sensitivity to outliers of our estimator and compare with the adapted Hill estimator. We consider an ϵ-contaminated model known by mixture of Pareto distributions

$$F_\epsilon(x) = 1 - (1-\epsilon)\,x^{-\alpha_1} + \epsilon x^{-\gamma}, \tag{3.9}$$

where α_1, $\alpha_2 > 0$ and $0 < \epsilon < 0.5$ is the fraction of contamination. Note that for $\epsilon = 0$, $\widehat{\alpha}_1^{(Z)}$ and $\widehat{\alpha}_1^{(H,c)}(k)$ are asymptotically unbiased. Therefore, for $\epsilon > 0$, the effect of contamination becomes immediately apparent. If $\alpha_1 < \gamma$ and $\epsilon > 0$, (3.9) corresponds to a Pareto distribution contaminated by a longer tailed distribution.

For the implementation of mixtures models to the study outliers one refers, for instance, to [Barnett and Lewis (1995), page 43]. In this context, we proceed our study as follows.

Step 1: We consider $\alpha_1 = 0.6$, $\gamma = 0.5$ and $\alpha_2 = 0.25$, to have the contaminated model.

Step 2: Then we consider three contamination scenarios according to $\epsilon = 5\%$, 10%, 15%.

Step 3: For each value ϵ, we generate 1000 samples of size $n = 100$, 200 and 500 from the model (3.9).

Step 4: Finally, we compare the two estimators with the true value ($\alpha_1 = 0.6$), by computing for each estimator, the appropriate abias and rmse and summarize the results in Table 3.2.

The values of the first line are those of the case where $\epsilon = 0$ (i.e., uncontaminated case). It has been shown that our estimator is more robust and perform better than the adapted Hill estimator proposed by [Einmahl *et al.* (2008)]. In fact, the adapted Hill estimator depends on the choice of the optimal number $k = k_n$ of upper order statistics and turn out to be more sensitive to this type of contaminations, for example, in 0% contamination for $n = 100$ the (abias, rmse) of the adapted Hill estimator equals $(0.1143, 0.2568)$, while for 15% contamination is $(1.0452, 1.1256)$. We may conclude that the adapted Hill estimator is note robust. However for 0% contamination the (abias, rmse) of our estimator equals

		$\widehat{\alpha}_1^{(Z)}$		$\widehat{\alpha}_1^{(H,c)}(k)$	
n	% contamination	$abias$	$rmse$	$abias$	$rmse$
100	0	.0611	.2511	.1143	.2586
	5	.0685	.2865	.1325	.3251
	10	.0754	.3561	.6485	.7546
	15	.0791	.3940	1.0452	1.1256
200	0	.0431	.1821	.0845	.1013
	5	.0487	.1965	.0911	.2511
	10	.0510	.2213	.2496	.3217
	15	.0614	.3889	.4518	.9941
500	0	.0153	.1142	.0245	.0684
	5	.0099	.1021	.1231	.1254
	10	.0239	.2289	.2211	.4024
	15	.0556	.3496	.4154	.4372

Table 3.2: Abais and rmse of the two estimators based on 1000 samples of mixture of Pareto distributions with tail index 0.6.

$(0.0611, 0.2511)$, while for 15% contamination is $(0.0791, 0.3940)$. We can conclude the robustness of our estimator, giving us, an excellent level of protection against contaminated data.

Chapter 4

Robust tail index estimation under truncation

In this chapter[1], we introduce a new robust estimator for the extreme value index of Pareto-type distributions under randomly right-truncated data and establish its consistency and asymptotic normality. Our considerations are based on the Lynden-Bell integral and a useful huberized M-functional and M-estimators of the tail index. A simulation study is carried out to evaluate the robustness and the finite sample behavior of the proposed estimator. Extreme quantiles estimation is also derived and applied to real dataset of lifetimes of automobile brake pads.

4.1 Introduction

Let (X_j, Y_j), $1 \leq j \leq N$, denote a sample of bivariate positive and independent rv's defined over some probability space $(\Omega, \mathcal{A}, P)$, with cdf's F and G respectively.

[1] Zahnit A., Brahimi B. ,Yahia D. (2021). Robust estimation of the extreme value index of Pareto-type distributions under random truncation with applications. *Pak. J. Stat. Oper.Res.* Vol.17 No.1, pp. 235-245.

Suppose that X is right-truncated by Y, in the sense that the rv of interest X_j is only observed when $X_j \leq Y_j$. We assume that both survival functions $\overline{F} := 1 - F$ and $\overline{G} := 1 - G$ are regularly varying at infinity, with respective indices $(-1/\gamma_1)$ and $(-1/\gamma_2)$, i.e, $\overline{F} \in RV_{-1/\gamma_1}$ and $\overline{G} \in RV_{-1/\gamma_2}$. That is, for any $t > 0$,

$$\lim_{x\to\infty} \overline{F}(tx)/\overline{F}(x) = t^{-1/\gamma_1} \quad \text{and} \quad \lim_{x\to\infty} \overline{G}(tx)/\overline{G}(x) = t^{-1/\gamma_2} \tag{4.1}$$

where $\gamma_j > 0$ $(j = 1, 2)$ is the so-called extreme value index (e.v.i) is a well-known parameter to measure the tail heaviness of a distribution. Distributions satisfying (4.1) play a very crucial role in extreme value analysis. They include many commonly used models such as Pareto, Burr, Fréchet and Lévy-stable distributions, known to be suitable models for adjusting large insurance claims, log-returns, large fluctuations, etc... see, for instance, [Resnick (2006)]. In many real applications, in case of presence of random right truncation (RRT), the rv of interest X may not be fully available. This truncation can occur in many areas, for example, it is usual that the insurer's claim data do not correspond to the underlying losses, because they are truncated from above. For a recent paper on insurance claims under RRT, one refers to [Escudero and Ortega (2008)].

In what follows, the star notation $(*)$ relates to any characteristic of the observed subsequence denoted by (X_i^*, Y_i^*) ; $1 \leq i \leq n$, $(n \leq N)$ subject to $X_i^* \leq Y_i^*$ from the N-sample. As a consequence of truncation, the size of actually observed sample, n, is a binomial rv with parameters N and $p := P(X \leq Y)$. We shall assume that $p > 0$, otherwise, nothing will be observed. Consequently, the marginal cdf's of X^* and Y^*, respectively denoted by F^* and G^*, becomes

$$F^*(x) := p^{-1} \int_0^x \overline{G}(t) dF(t) \quad \text{and} \quad G^*(y) := p^{-1} \int_0^y F(t) dG(t),$$

the corresponding tails are

$$\overline{F^*}(x) = -p^{-1}\int_x^\infty \overline{G}(t)d\overline{F}(t) \quad \text{and} \quad \overline{G^*}(y) = -p^{-1}\int_y^\infty F(t)d\overline{G}(t).$$

We can easily show that see, Proposition B.1.10 in [de Haan and Ferreira (2006)] $\overline{F^*} \in RV_{-1/\gamma_1^*}$ and $\overline{G^*} \in RV_{-1/\gamma_2^*}$ with respective indices

$$\gamma_1^* = \frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2} \quad \text{and} \quad \gamma_2^* = \gamma_2. \tag{4.2}$$

Since F and G are heavy-tailed. Therefore, the Woodroofe's nonparametric estimator, see [Woodroofe (1985)] of F, is defined by

$$F_n^{(W)}(x) := \prod_{j:X_j^*>x} \exp\left(-\tfrac{1}{nC_n(X_j^*)}\right), \qquad \text{where } C_n(x) := \frac{1}{n}\sum_{j=1}^n 1_{(X_j^*\le x\le Y_j^*)},$$

in which C_n is the empirical estimator of

$$C(z) := P(X \le z \le Y | X \le Y) = p^{-1}\overline{G}(z)F(z).$$

Another more popular estimator for F, is the well known Lynden-Bell nonparametric maximum likelihood estimator, originally proposed by [Lynden-Bell (1971)], defined by

$$F_n^{(LB)}(x) := \prod_{j:X_j^*>x} \left(1 - \tfrac{1}{nC_n(X_j^*)}\right).$$

Recently, [Gardes and Stupfler (2015)] exploited the above relations (4.2) to define an estimator for the parameter of interest γ_1 by considering the classical Hill estimators, see [Hill (1975)] of γ_1^* and γ_2^* as functions of two distinct numbers of

upper observations. Thus, from (4.2) we have

$$\gamma_2 = \gamma_2^* \quad \text{and} \quad \gamma_1 = \frac{\gamma_1^* \gamma_2^*}{\gamma_2^* - \gamma_1^*}.$$

Therefor,

$$\hat{\gamma}_1^{(GS)}(k_1,k_2) = \frac{\hat{\gamma}_1^*(k_1)\hat{\gamma}_2^*(k_2)}{(\hat{\gamma}_2^*(k_2) - \hat{\gamma}_1^*(k_1))} \tag{4.3}$$

where

$$\hat{\gamma}_1^*(k_1) := \frac{1}{k_1}\sum_{j=1}^{k_1} \log\left(\frac{X^*_{n-j+1,n}}{X^*_{n-k_1,n}}\right) \quad \text{and} \quad \hat{\gamma}_2^*(k_2) := \frac{1}{k_2}\sum_{j=1}^{k_2} \log\left(\frac{Y^*_{n-j+1,n}}{Y^*_{n-k_2,n}}\right),$$

$X^*_{1,n} \leq ... \leq X^*_{n,n}$ and $Y^*_{1,n} \leq ... \leq Y^*_{n,n}$ denote the usual order statistics of both observed samples, k_1 and k_2 are the numbers of top statistics (upper observations) which are kept for estimating γ_1^* and γ_2^*.

The estimator given by (4.3) suffer from some kind of calibration problem, because of the difficulty in assessing the correlation between $\hat{\gamma}_1^*$ and $\hat{\gamma}_2^*$, [Gardes and Stupfler (2015)] they don't consider the situation where the upper statistics are equal. [Benchaira et al. (2015)] considered the case where $k := k_1 = k_2$ in the expression (4.3) of $\hat{\gamma}_1^{(GS)}$. They proved the asymptotic normality of this estimator under the tail dependence and the second-order regular variation conditions. Recently, [Worms and Worms (2016)] proposed an asymptotically normal estimator for γ_1 by considering a Lynden-Bell integrals with a deterministic threshold $t_n > 0$ given by

$$\hat{\gamma}_1^{(W)}(t_n) := \frac{1}{nF_n^{(LB)}(t_n)}\sum_{j=1}^{n} \frac{F_n^{(LB)}(X_j^*)}{C_n(X_j^*)} \log\left(\frac{X_j^*}{t_n}\right) 1_{(X_j^* > t_n)}. \tag{4.4}$$

The case of a random threshold, is addressed by [Benchaira et al. (2016)] who propose a Hill-type estimator under RRT based on a Woodroofe integration as

follows:

$$\hat{\gamma}_1^{(B)}(k) := \frac{1}{nF_n^{(W)}\left(X^*_{n-k,n}\right)} \sum_{i=1}^{k} \frac{F_n^{(W)}\left(X^*_{n-i+1,n}\right)}{C_n\left(X^*_{n-i+1,n}\right)} \log\left(\frac{X^*_{n-i+1,n}}{X^*_{n-k,n}}\right). \tag{4.5}$$

All of these e.v.i estimators, as well as the classical Hill-type (in complete data case) are non robust, in the sense that they are very sensitive to extreme observations, data contamination or model deviation and tend to be highly volatile for small samples (this is illustrated in our simulation study). Also, the rate of convergence of these estimators are based on the optimal value of the numbers of top statistics k or the threshold t_n, but this rate are slower than the parametric rate $\sqrt{n}$. Moreover, estimating the optimal value of k is virtually impossible when the sample size n is small and this leads to unstable estimates for small samples and large confidence intervals, see [Resnick (1997)] for a detailed discussion. The alternative approach is inspired by the theory of robust inference (see, for instance, [Huber (1981)] and [Hampel *et al.* (1986)]) instead of exact consistency, this theory aim at stability for small samples, possibly at the cost of a small asymptotic bias. However, as observed by [Beran and Shell (2012)], in some practical cases, such as natural disasters, operational risk assessment or reinsurance data are sparse (with n often somewhere between 20 and 50) and distributions are expected to be heavy tailed with an unknown e.v.i. Robust estimation of e.v.i. focuses primarily on complete data case, see [Brazauskas and Serfling (2000)], [Beran and Shell (2012)] and references therein. The incomplete data case has first been considered by [Sayah et al. (2014)], who dealt with heavy-tailed and right censored data. The aim of the current paper is to provide a robust e.v.i. estimator for heavy tailed data under RRT.

4.2 Framework and statement of the results

Recall that the condition (4.1) can be rephrased as $\bar{F}(x) = x^{-1/\gamma_1} L_F(x)$ and $\overline{G}(x) = x^{-1/\gamma_2} L_G(x)$, where L_F and L_G are slowly varying functions at infinity. Assuming that $\lim_{x\to\infty} L_F(x) = c > 0$ leads to the class of so-called Pareto-like (or heavy-tailed) distributions, i.e. distribution satisfying $1 - F(x) \sim cx^{-1/\gamma_1}$ as x tends to infinity. Then, the tail of such distribution behaves asymptotically like the tail of Pareto distribution. Thus, we suggests to robustify the Pareto maximum likelihood estimator of γ_1 in order to obtain sensible estimates for the class of Pareto-type distributions despite possible deviations from the single-parameter Pareto model see, [Beran and Shell (2012)] for a detailed discussion. A natural estimate of γ_1 can therefore be based on a Huberized Pareto score function :

$$\begin{aligned}\psi_{v,u}(x,\gamma) &= [\gamma^{-1}\log(x) - 1]_v^u - \int [\gamma^{-1}\log(z) - 1]_v^u dF_{Par,\gamma}(z) \\ &= [\gamma^{-1}\log(x) - 1]_v^u - (v + \exp(-(v+1)) - \exp(-(u+1))), \quad (4.6)\end{aligned}$$

where $F_{Par,\gamma}(x) := 1 - x^{-1/\gamma}$, for $x \geq 1$ and $[y]_v^u := \min(\max(y, v), u)$. The reason for huberization is that the Pareto distribution is only an approximation of the true cdf. By huberizing, the estimate becomes robust against a large class of deviations from this approximation. Since deviations are mainly relevant in the center of the distribution, the lower truncation parameter v is more important. As an alternative to Hill-type estimation, [Beran(1997)] proposed to use all data but huberize the Pareto score function at lower quantiles. This method has been investigated in the complete data case in [Beran and Shell (2012)]. Moreover, $\psi_{v,u}(x,\gamma)$ is defined for any choice of $\gamma > 0$ and $-1 \leq v < u \leq \infty$. Thus, as shown by [Beran and Shell (2012)], robustness needs to be achieved for lower

quantiles whereas extreme observations on the right are those we are interested in. In particular, $\psi_{-1,\infty}(x,\gamma) = \gamma^{-1}\log(x) - 1$ for $x \geq 1$. Consequently, a natural choice is $u = \infty$ and robustification on the left is obtained only if $v > -1$.

Under the assumptions above, and denote by $\mathcal{F}$ a set of distributions with support in R_+. Then the functional $T(F)$ defined on F as the solution $t = t_0$ of the equation

$$\beta_F(t) = \int \psi_{v,u}(x,t)\, dF(x) = 0, \quad (F \in \mathcal{F})$$

is called huberized tail index M-functional. Consequently, by using relations **(1.9)** and **(1.10)** in [Stute and Wang (2008)] in the left truncation case, a natural adaptation of this integral $\beta_F(t)$ in the framework of RRT, leads to the corresponding Huberized Lynden-Bell integral estimator of the e.v.i. γ_1 as any solution sequence T_n of the empirical equation

$$\hat{\beta}_{F_n}(T_n) := \sum_{j=1}^{n} \psi_{v,u}\left(X_j^*, T_n\right) \frac{F_n^{(LB)}\left(X_j^*\right)}{C_n\left(X_j^*\right)} = 0. \tag{4.7}$$

Remark 4.2.1 *It is worth mentioning that for complete data case (no truncation), we have $n = N$, $X^* = X$ and $C_n = F_n = F_n^*$, it follows that $\hat{\beta}_{F_n}(T_n) = \sum_{i=1}^{n} \psi_{v,u}(X_i, T_n)$ and consequently T_n reduce to the Beran and Shell estimator see, e.g. [Beran and Shell (2012)].*

Next, we investigate the asymptotic properties of the estimator of the tail index γ_1 under the large class of Pareto-type distributions assumptions. To formulate our main result, the following conditions are required:

(A1) Let $\overline{F} \in RV_{-1/\gamma_1}$ and $\overline{G} \in RV_{-1/\gamma_2}$ with $0 < \gamma_1 < \gamma_2$.

(A2) $\int \frac{1}{\overline{G}(x)} \psi_{v,u}^2(x,t)\, dF(x) < \infty$ and $\int \frac{1}{\overline{G}(x)} dF(x) < \infty$.

Theorem 4.2.1 *Assume that assumptions (A1) and (A2) hold. Moreover, let* $F_n := F_n^{(LB)}$ *be the Lynden-Bell estimator of the cdf* F*. Then, provided the existence of* γ_1 *as a unique solution of* $\beta_F(t) = 0$*, any solution sequence* $\hat{\gamma}_1^{(Z)} := \hat{\gamma}_{1n}^{(Z)}(v,u)$ *of*

$$\hat{\beta}_{F_n}(t) = \int \psi_{v,u}(x,t)\, dF_n(x) = 0 \qquad (n \in \mathbb{N})$$

is a consistent estimator of γ_1*. Assume further that* $\int \frac{\partial}{\partial t}\psi_{v,u}(x,t)\, dF(x) \neq 0$ *hold in a neighborhood of* γ_1*. Then*

$$\sqrt{n}\left(\hat{\gamma}_1^{(Z)} - \gamma_1\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma_{v,u}^2\right), \quad \text{as } n \to \infty$$

where $\xrightarrow{d}$ *stands for convergence in distribution and*

$$\sigma_{v,u}^2 := \sigma^2 \left(\int \frac{\partial}{\partial t}\psi_{v,u}(x,t)\, dF(x) \right)^{-2} \tag{4.8}$$

in which

$$\sigma^2 = Var\left\{ \frac{\Lambda(X^*)}{C(X^*)} + \int_{X^*}^{Y^*} \frac{\Lambda(z)}{C^2(z)} dF^*(z) \right\},$$

where

$$\Lambda(z) := \int_{z>x} \left[\psi_{v,u}(z,\gamma_1) - \psi_{v,u}(x,\gamma_1)\right] dF(x).$$

Remark 4.2.2 *Condition (A1) is standard in heavy-tailed and RRT context. The condition* $\gamma_1 < \gamma_2$ *ensures that the tail of the truncated rv of interest* X *is note too contaminated by the truncation rv* Y*. In addition, (A1) implies that, the right endpoints of* X *and* Y *are infinite and thus they are equal. Assumption (A2) already appeared in [Stute and Wang (2008)], they showed that,* $\sigma^2 < \infty$ *under (A2), therefore,* $\sigma_{v,u}^2 < \infty$ *too. Since* $\overline{G} \leq 1$*, it implies* $\int \psi_{v,u}^2(x,t)\, dF(x) < \infty$*, which is the assumption when no truncation occurs, see Theorem 2 in [Beran and Shell (2012)].*

In our case, (A2) is satisfied when $0 < \gamma_1 < \gamma_2$.

Remark 4.2.3 *In comparison with the optimal value of the numbers of top statistics* k *in the Hill-type estimators, the parameter* v *play a less crucial role, since the rate of convergence does not depend on* v. *In contrast to Hill-type estimators under truncation (see, equations 4.3 and 4.5), all data are used. The role of* v *is only to determine a threshold below which data have a bounded influence on the estimator. Note also that, the equation (4.7) defining our estimator has a solution for* $n \geq 2$ *almost surely.*

4.3 Simulation study

In this section we examines the performance of our estimator $\hat{\gamma}_1^{(Z)}$ given by solving the empirical equation (4.7), in which, the huberizing constants are $v = 0$ and $u = \infty$, and compare it with estimators proposed by [Gardes and Stupfler (2015)], [Worms and Worms (2016)] and [Benchaira et al. (2016)] given by (4.3), (4.4) and (4.5) respectively. Firstly, we generate 1000 pseudo-random samples X and Y of size $N = 100$, 150 and 200 from Burr's models, $\overline{F}(x) = \left(1 + x^{1/\theta}\right)^{-\theta/\gamma_1}$ and $\bar{G}(x) = \left(1 + x^{1/\theta}\right)^{-\theta/\gamma_2}$, $x \geq 0$. We fix $\theta = 1/4$ and choose the values 0.6 and 0.8 for γ_1 and $p = 0.7$ (resp. 0.9), that means the percentage of truncation is 30% (resp. 10%). The pertaining γ_2-value is obtained by solving the equation $p = \gamma_2/(\gamma_1 + \gamma_2)$, for each couple (γ_1, p). We obtained 1000 pseudo-random samples X^* and Y^* of size $n \simeq pN$. Next, we calculate the estimators values from the observed data X^* and Y^*. For choosing the optimal number k_n of upper order statistics used in the computation of $\hat{\gamma}_1^{(GS)}$, $\hat{\gamma}_1^{(W)}$ and $\hat{\gamma}_1^{(B)}$ we adopt the Reiss and Thomas algorithm [Reiss and Thomas (2007), page 137]. In those simulations, we used the random threshold $X^*_{n-k_n,n}$ instead of t_n in the definition of $\hat{\gamma}_1^{(W)}$. Also

note that we only consider $k_n := k_1 = k_2$ in the expression (4.3), in this case $\hat{\gamma}_1^{(GS)}$ is the one considered in [Benchaira et al. (2015)]. Finally, we compute the absolute bias (abias) and root mean squared error (rmse) of these estimators, the results are summarized in Table 4.1 and Table 4.2. We see that our new estimator shows good performance compared to existing methods for small sample sizes.

Table 4.1: Bias and rmse of the estimators based on 1000 samples of Burr's models with $\gamma_1 = 0.6$, for p=0.7 (top) and p=0.9 (bottom).

			$\hat{\gamma}_1^{(Z)}$		$\hat{\gamma}_1^{(GS)}$		$\hat{\gamma}_1^{(W)}$		$\hat{\gamma}_1^{(B)}$	
p	N	n	$abias$	$rmse$	$abias$	$rmse$	$abias$	$rmse$	$abias$	$rmse$
	100	70	.008	.028	.422	7.310	.014	.243	.197	.447
0.7	150	104	.006	.013	.225	1.892	.011	.212	.154	.399
	200	139	.003	.010	.227	.993	.009	.187	.148	.363
	100	90	.004	.171	.122	4.751	.007	.178	.050	.556
0.9	150	135	.005	.073	.072	.537	.007	.143	.061	.392
	200	179	.006	.019	.084	.651	.006	.121	.068	.309

Table 4.2: Bias and rmse of the estimators based on 1000 samples of Burr's models with $\gamma_1 = 0.8$, for p=0.7 (top) and p=0.9 (bottom).

			$\hat{\gamma}_1^{(Z)}$		$\hat{\gamma}_1^{(GS)}$		$\hat{\gamma}_1^{(W)}$		$\hat{\gamma}_1^{(B)}$	
p	N	n	$abias$	$rmse$	$abias$	$rmse$	$abias$	$rmse$	$abias$	$rmse$
	100	70	.006	.019	.315	9.594	.017	.379	.247	.617
0.7	150	104	.009	.011	.308	2.803	.018	.365	.190	.515
	200	139	.008	.012	.256	1.192	.019	.291	.200	.513
	100	90	.023	.027	.093	5.440	.037	.183	.090	.713
0.9	150	135	.018	.020	.138	.786	.036	.161	.137	.467
	200	179	.010	.014	.110	.487	.034	.138	.102	.407

Now, in order to study the sensitivity to outliers of our newly estimator, we consider an ϵ-contaminated model known by mixture of Pareto distributions

$$F_{\gamma_1,\lambda,\epsilon}(z) = 1 - (1-\epsilon)\, z^{-1/\gamma_1} + \epsilon z^{-1/\lambda}, \qquad \gamma_1, \lambda > 0 \quad and \quad 0 < \epsilon < 0.5 \quad (4.9)$$

Note that, for $\gamma_1 < \lambda$ and $\epsilon > 0$, (4.9) corresponds to a Pareto distribution contaminated by a longer tailed distribution. In this context, we proceed our

study as follows. We fix $\lambda = 2$ and consider four different contamination levels $\epsilon = 0.05, 0.15, 0.25, 0.35$, and we vary γ_1 among 0.6 and 0.8. For each value of ϵ, 1000 samples of size $N = 200$ were generated from the model (4.9) truncated by a simple Pareto model $\bar{G}(x) = x^{-1/\gamma_2}$, with $p = 0.7$ and 0.9. Our illustration, made with respect to the biases and rmse's, are summarized in Table 4.3. The values of the first line are those of the case where $\epsilon = 0$ (i.e., uncontaminated case). Both the bias and the rmse of our estimator are note sensitive to outliers. Then we can conclude its robustness, giving us, in fact, an excellent level of protection against contamination data.

Table 4.3: Bais and rmse of the estimators based on 1000 samples of a contaminated Pareto distribution, with tail index $\gamma_1 = 0.6$ (left) and $\gamma_1 = 0.8$ (right), $N = 200$.

	γ_1= 0.6				γ_1= 0.8			
p	0.7		0.9		0.7		0.9	
$\epsilon\%$	*bias*	*rmse*	*bias*	*rmse*	*bias*	*rmse*	*bias*	*rmse*
0	.0088	.0137	.0052	.0998	.0265	.0180	.0189	.0558
5	.0104	.0558	.0644	.1112	.0562	.0591	.0698	.0954
15	.0153	.0921	.0905	.1568	.0872	.0938	.0954	.1589
25	.0256	.3336	.1256	.4451	.1010	.7470	.1615	.4785
35	.1414	.5330	.2115	.6121	.1726	.9221	.2121	.7787

4.4 Applications

4.4.1 Estimation of an upper quantile

Estimation of e.v.i. is very important in the determination of high quantiles, upper tail probabilities, mean excess functions, and excess-of-loss and stop-loss reinsurance premiums. Consequently, small errors in estimation of this quantity can produce substantial impact in applications. Thus, for robust estimation of quantities based on γ_1 robust estimation of γ_1 itself is crucial. In other words, for a

heavy tailed distributions, robust estimation of the high quantile Q_ε corresponding to upper tail probability ε, becomes of interest, and this may be carried out by robust estimation of γ_1. [Brazauskas and Serfling (2000)] gives a detailed account of this issue.

Let (α_n) be some sequence of quantiles orders tending to 0, such that $\alpha_n = o\left(\overline{F}\left(s_n\right)\right)$, where (s_n) is a sequence of positive deterministic thresholds growing to infinity with n. Consequently, the quantile of F of order $(1-\alpha_n)$ is such that $\overline{F}\left(Q_{\alpha_n}\right) = \alpha_n$. We can then define an estimator $\hat{Q}_{\alpha_n,s_n}$ of Q_{α_n} :

$$\hat{Q}_{\alpha_n,s_n} = s_n \left(\alpha_n^{-1}\left(1 - F_n^{(LB)}\left(s_n\right)\right)\right)^{\hat{\gamma}_1^{(Z)}}.$$

A similar estimator is proposed by [Worms and Worms (2016)], but instead of $\hat{\gamma}_1^{(Z)}$ they consider the estimator $\hat{\gamma}_1^{(W)}\left(t_n\right)$ given by (4.4). Before we state the asymptotic normality of $\hat{Q}_{\alpha_n,s_n}$, we set $d_n := \overline{F}\left(s_n\right)/\alpha_n$ and assume that

$$d_n \to \infty \quad \text{and} \quad \sqrt{n}/\log\left(d_n\right) \to \infty, \qquad \text{as } n \to \infty. \tag{4.10}$$

Theorem 4.4.1 *Under (4.10) and the assumptions of Theorem 4.2.1, we have*

$$\frac{\sqrt{n}}{\log\left(d_n\right)}\left(\frac{\hat{Q}_{\alpha_n,s_n}}{Q_{\alpha_n}} - 1\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma_{v,u}^2\right), \qquad \text{as } n \to \infty.$$

4.4.2 Real data example : automobile brake pad lifetime

In reliability, a real dataset of lifetimes of automobile brake pads already considered by [Lawless (2011)], was recently analyzed in [Gardes and Stupfler (2015)] and [Benchaira et al. (2016)] as an application of heavy-tailed and RRT data. We follow the same steps as those of [Gardes and Stupfler (2015)] who trans-

formed this sample into a right-truncated data, which originally is left-truncated. We are dealing with a dataset of small size ($n = 98$), consequently, robust estimation of γ_1 can produce substantial robust estimation of the high quantile. Then, our procedure should be preferred to that based on no robust estimation of γ_1. In these situation, we used the random threshold $X^*_{n-k_n,n}$ instead of s_n in the definition of $\hat{Q}_{\alpha_n,s_n}$. We select the optimal number of top statistics, via the numerical procedure of [Reiss and Thomas (2007), page 137] and we get $k = 10$ and we estimate the tail index γ_1 given in (4.5) and (4.7) we get $\hat{\gamma}_1^{(B)} = 0.4701$ and $\hat{\gamma}_1^{(Z)} = 0.4925$ respectively. The estimation results of our based (right-panel) and that of [Benchaira et al. (2016)] based (left-panel) extreme quantiles estimators with three different quantile levels $(0.990, 0.995, 0.99)$ corresponding to $\alpha_n = 0.001, 0.005, 0.010$ are summarized in Table 4.4. For instance, we conclude that the brake pad lifetime is estimated to be less than 17063 km for 1% of the cars while only one out of a thousand brake pads lasts less than 10200 km.

Table 4.4: Extreme quantiles for car brake pad lifetimes.

Quantile level	$\hat{Q}_{\alpha_n}$ via $\hat{\gamma}_1^{(B)}$	$\hat{Q}_{\alpha_n}$ via $\hat{\gamma}_1^{(Z)}$
0.990	17604	17063
0.995	14641	14138
0.999	10559	10203

4.5 Proofs

Proof of Theorem 4.2.1. The proof is essentially based on Theorem 4.3 in [He and Yang (1998)] and Corollary 1.1. in [Stute and Wang (2008)]. Note that $\psi_{v,u}(x,t)$ is monotone and continuous in t and $\beta_F(t)$ possesses an isolated root

at $t = \gamma_1$. Let $\varepsilon > 0$, then under (A1) and (A2) by strong low of large numbers under truncation, see Theorem 4.3 in [He and Yang (1998)], we have

$$\hat{\beta}_{F_n}(\gamma_1 - \varepsilon) = \int \psi_{v,u}(x, \gamma_1 - \varepsilon)\, dF_n(x) \to \beta_F(t_0 - \varepsilon) > 0 \quad \text{almost surely}$$

and

$$\hat{\beta}_{F_n}(\gamma_1 + \varepsilon) = \int \psi_{v,u}(x, \gamma_1 + \varepsilon)\, dF_n(x) \to \beta_F(t_0 + \varepsilon) < 0 \quad \text{almost surely.}$$

Hence, there exists some $n \in \mathbb{N}$ such that

$$P\left(\hat{\beta}_{F_m}(\gamma_1 + \varepsilon) < 0 < \hat{\beta}_{F_m}(\gamma_1 - \varepsilon), \quad \forall m \geq n\right) \to 1 \quad as\ n \to \infty. \tag{4.11}$$

According to the monotonicity of $\psi_{v,u}(x,t)$ in t, together with the assumption of the existence of a solution sequence $\hat{\gamma}_1^{(Z)}$ of the empirical equation

$$\hat{\beta}_{F_n}(t) = \int \psi_{v,u}(x,t)\, dF_n(x) = 0 \qquad (n \in \mathbb{N})$$

we then get

$$P\left(\gamma_1 + \varepsilon < \hat{\gamma}_1^{(Z)} < \gamma_1 - \varepsilon, \quad \forall n \geq m\right) \to 1 \quad as\ n \to \infty.$$

The existence of such a solution sequence for a continuous function in a neighborhood of γ_1 follows from (4.11) for n large enough. Thus, $\hat{\gamma}_1^{(Z)}$ is a consistent estimator of γ_1.

Let us now focus on the asymptotic normality of $\hat{\gamma}_1^{(Z)}$. Recall that,

$$\int \psi_{v,u}\left(x,\hat{\gamma}_1^{(Z)}\right) dF_n(x) - \int \psi_{v,u}(x,\gamma_1)\, dF(x) = \int \left(\psi_{v,u}\left(x,\hat{\gamma}_1^{(Z)}\right) - \psi_{v,u}(x,\gamma_1)\right) \times dF_n(x) + \int \psi_{v,u}(x,\gamma_1)\, d\left(F_n(x) - F(x)\right), \tag{4.12}$$

The assumed differentiability of $\psi_{v,u}(x,t)$ in t allows a Taylor expansion around γ_1 which yields

$$\sqrt{n}\left(\hat{\gamma}_1^{(Z)} - \gamma_1\right) \int \frac{\partial}{\partial t}\psi_{v,u}(x,t)\, dF_n(x) = \sqrt{n} \int \left(-\psi_{v,u}(x,\gamma_1)\right) d\left(F_n(x) - F(x)\right).$$

Then,

$$\sqrt{n}\left(\hat{\gamma}_1^{(Z)} - \gamma_1\right) = \sqrt{n}\left(\int \frac{\partial}{\partial t}\psi_{v,u}(x,t)\, dF_n(x)\right)^{-1} \int \left(-\psi_{v,u}(x,\gamma_1)\right) d\left(F_n(x) - F(x)\right).$$

It was shown in Theorem 4.3 in [He and Yang (1998)] that for any nonnegative measurable real function $\varphi := \frac{\partial}{\partial t}\psi$, and under the condition $\int \varphi_{v,u}(x,t)\, dF(x) \neq 0$ hold in a neighborhood of γ_1, we get

$$\int \varphi_{v,u}(x,t)\, dF_n(x) = \int \varphi_{v,u}(x,t)\, dF(x) + op(1). \tag{4.13}$$

Under assumptions (A1) and (A2), we can apply the central limit theorem under right truncation, see Corollary 1.1 in [Stute and Wang (2008)] for the Lynden-Bell integral, obtaining

$$\sqrt{n} \int \left(-\psi_{v,u}(x,\gamma_1)\right) d\left(F_n(x) - F(x)\right) \xrightarrow{d} \mathcal{N}\left(0,\sigma^2\right), \quad \text{as } n \to \infty \tag{4.14}$$

where σ^2 is given by (4.8). Consequently, the limit variance follows from (4.13) and (4.14). This concludes the proof of Theorem 4.2.1. ■

Proof of Theorem 4.4.1. The result follows by analogous arguments as in the proof of Theorem 2 in [Worms and Worms (2016)]. Recall that the high quantile Q_{α_n} corresponding to order $(1-\alpha_n)$ is such that $\overline{F}\left(Q_{\alpha_n}\right)=\alpha_n$, and its estimator is defined by

$$\hat{Q}_{\alpha_n,s_n}=s_n\left(\frac{\bar{F}_n\left(s_n\right)}{\alpha_n}\right)^{\hat{\gamma}_1^{(Z)}}.$$

For convenience, we set $\Lambda_n:=\bar{F}_n\left(s_n\right)/\bar{F}\left(s_n\right)$. Indeed, we have

$$\begin{aligned}\frac{\hat{Q}_{\alpha_n,s_n}}{Q_{\alpha_n}}-1&=\frac{s_n}{Q_{\alpha_n}}\left(\frac{\bar{F}_n\left(s_n\right)}{\alpha_n}\Lambda_n\right)^{\hat{\gamma}_1^{(Z)}}-1\\&=\Lambda_n^{\hat{\gamma}_1^{(Z)}}\left\{\left(\frac{s_n}{Q_{\alpha_n}}d_n^{\gamma_1}d_n^{\left(\hat{\gamma}_1^{(Z)}-\gamma_1\right)}-1\right)+\left(1-\Lambda_n^{-\hat{\gamma}_1^{(Z)}}\right)+\left(\frac{s_n}{Q_{\alpha_n}}d_n^{\gamma_1}-1\right)\right\}\\&=:\Lambda_n^{\hat{\gamma}_1^{(Z)}}\left\{I_{n1}+I_{n2}+I_{n3}\right\}.\end{aligned}$$

We will show that $\frac{\sqrt{n}}{\log(d_n)}I_{n1}$ is asymptotically centred Gaussian rv with variance $\sigma_{v,u}^2$ and $\frac{\sqrt{n}}{\log(d_n)}I_{nj}\xrightarrow{P}0$, $j=2,3$. Concerning the term I_{n1}, by using the mean value theorem, it follows that

$$\frac{\sqrt{n}}{\log\left(d_n\right)}I_{n1}=\sqrt{n}\left(\hat{\gamma}_1^{(Z)}-\gamma_1\right)\exp\left(\delta_n\right),$$

where $\delta_n\leq\left|\hat{\gamma}_1^{(Z)}-\gamma_1\right|\log\left(d_n\right)$. Assumption (4.10) and Theorem 4.2.1, allows us to conclude that δ_n tends to 0. We use then Theorem 4.2.1 to get.

$$\frac{\sqrt{n}}{\log\left(d_n\right)}I_{n1}\to N\left(0,\sigma_{v,u}^2\right),\qquad\text{as } n\to\infty.$$

Let us now focus on the negligible terms I_{n2} and I_{n3}. By using the mean value

theorem, we get

$$I_{n2} = \hat{\gamma}_1^{(Z)} M_n^{-\hat{\gamma}_1^{(Z)}-1} \left(\Lambda_n - 1\right),$$

with M_n tending to 1. In view of assumptions (A1) and (4.10), the sequence (Λ_n) converge to 1 in probability (see, Lemma 2 in [Worms and Worms (2016)]), we have then

$$\frac{\sqrt{n}}{\log\left(d_n\right)} \left(\Lambda_n - 1\right) = o_p\left(1\right).$$

Hence

$$\frac{\sqrt{n}}{\log\left(d_n\right)} I_{n2} = o_p\left(1\right).$$

For I_{n3}, in view of the regular variation of $\bar{F}$, (4.1) can be rephrased as $\bar{F}\left(x\right) = x^{-1/\gamma_1} L_F\left(x\right)$, where L_F is slowly varying function at infinity and by definition of Q_{α_n}, we get

$$I_{n3} = \frac{s_n}{Q_{\alpha_n}} \left(\frac{\bar{F}\left(s_n\right)}{\bar{F}\left(Q_{\alpha_n}\right)}\right)^{-\gamma_1} - 1 = \left(\frac{L_F\left(Q_{\alpha_n}\right)}{L_F\left(s_n\right)}\right)^{-\gamma_1} - 1.$$

Therefore, we use the following representation of L_F (see, [Smith (1987)], page 1195)

$$L_F\left(x\right) = c\left(1 + \rho^{-1} h\left(x\right) + o\left(h\left(x\right)\right)\right), \qquad \text{for } x \to \infty$$

where h is a positive measurable function, slowly varying with index $\rho < 0$. We have, Q_{α_n}/s_n tends to infinity, then $h\left(Q_{\alpha_n}\right)/h\left(s_n\right)$ tends to 0 and

$$\left|\frac{h\left(Q_{\alpha_n}\right)}{h\left(s_n\right)} - \left(\frac{Q_{\alpha_n}}{s_n}\right)^{\rho}\right| \leq \sup_{w \geq 1} \left|\frac{h\left(w s_n\right)}{h\left(s_n\right)} - w^{\rho^{-1}}\right| \to 0.$$

It follows that

$$\begin{aligned}\frac{L_F\left(Q_{\alpha_n}\right)}{L_F\left(s_n\right)} &= 1-\rho^{-1}h\left(s_n\right)\left(1-\frac{h\left(Q_{\alpha_n}\right)}{h\left(s_n\right)}+o\left(\frac{h\left(Q_{\alpha_n}\right)}{h\left(s_n\right)}\right)+o_p\left(1\right)\right)\\ &= 1-\rho^{-1}h\left(s_n\right)\left(1+o_p\left(1\right)\right).\end{aligned}$$

Therefore $\left|I_{n3}\right| \leq C\left|L_F\left(Q_{\alpha_n}\right)/L_F\left(s_n\right)-1\right|$, then

$$\frac{\sqrt{n}}{\log\left(d_n\right)}\left|I_{n3}\right| \leq C\frac{\sqrt{n}}{\log\left(d_n\right)}\rho^{-1}h\left(s_n\right)\left(1+o_p\left(1\right)\right)$$

and then the desired negligibility of I_{n3} follows from assumption (4.10), which ends the proof of the Theorem. ■

Conclusion & Perspectives

In this Thesis, we propose a new robust tail index estimation procedure for Pareto-type distributions under incomplete data (random censorship or random truncation). Our considerations are based on the Lynden-Bell integral (for randomly truncation data) and the ideas of Kaplan-Meier integration (under random censorship model) using the huberized M-estimator of the tail index. We derive their asymptotic results. Extreme quantiles estimation is also derived and applied to real dataset of lifetimes of automobile brake pads. A simulation study is carried out to evaluate the performance and the robustness of the proposed estimators.

It has been shown that our newly extreme value index estimators of Pareto-type distributions are more robust and perform better than the existing Hill-type estimators based on the order statistics and numbers of upper observations, for small sample sizes and for both uncontaminated and contaminated cases. Therefore, it can be used in practice as an alternative when one has to deal with small samples, in contaminated cases or in the presence of outliers.

In our further research we will study this robust estimator in more detail. Note that, the degree of robustness is determined by the tuning parameters v and u. It remains a likely topic for future investigations to treat the choice of these parameters.

Bibliography

[1] Applebaum, D. (2005). Levy-type stochastic integrals with regurlary varying tails. *Stochastic Analysis and Applications.* 23(3), 595-611.

[2] Bacro, J.N. and Brito, M. 1995. Weak limiting behaviour of a simple tail Pareto-index estimator. *J. Statist. Plann. Inference* 45, 7-19.

[3] Balkema, A. A. and de Haan, L. (1974). Residual life time at great age. *Annals of Probability.* 2: 792–804.

[4] Barnett, V. and Lewis, T. (1995). *Outliers in statistical data.* Third ed. John Willey & Sons. New York.

[5] Beirlant, J., Guillou, A., Dierckx, G. and Fils-Villetard, A. (2007). Estimation of the extreme value index and extreme quantiles under random censoring. *Extremes.* 10, no. 3, 151-174.

[6] Benchaira, S., Meraghni, D. and Necir, A. (2015). On the asymptotic normality of the extreme value index for right-truncated data. *Statist. Probab. Lett.* 107, 378-384.

[7] Benchaira, S., Meraghni, D., Necir, A. (2016). Tail product-limit process for truncated data with application to extreme value index estimation. *Extremes*, 19, 219-251.

[8] Beran, J. (1997). On heavy tail modeling and teletraffic Data by S.I. Resnick. *Anna. Statist.* 25 (5), 1852–1855.

[9] Beran, J., Schell, D. (2012). On robust tail index estimation. *Comput. Statist. Data Anal.* 56(11), 3430-3443.

[10] Bingham, N, Goldie, C. and Teugels, J. (1987). *Regular Variation.* Cambridge University Press.

[11] Brazauskas, V., Serfling, R. (2000). Robust and efficient estimation of the tail index of a single-parameter Pareto distribution. *North American Actuarial J.* 4(4), 12–27.

[12] Castillo, E. (1988). *Extreme value theory in engineering. Statistical Modeling and Decision Science.* Academic Press, Inc.

[13] Coles S., Heffernan J. and Tawn J. (1999). Dependence Measures for Extreme Value Analyses. *Extremes* 2:4, 339-365.

[14] Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values.* Springer, New York.

[15] David, H.A. (1981). *Order Statistics*, 2nd Edition, New York: John Wiley.

[16] Dekkers, A. L. M., Einmahl, J.H.J. and de Haan, L. (1989). A moment estimator for the index of an extreme-value distribution. *Ann. Statist.* 17, 1833-1855.

[17] Denuit, M., Dhaene, J., Goovaerts, M.J. and Kaas, R. (2005). A*ctuarial Theory for Dependent Risk: Measures, Orders and Models.* Wiley, New York.

[18] Einmahl, J.H.J., Fils-Villetard, A. and Guillou, A. (2008). Statistics of extremes under random censoring. *Bernoulli.* 14, no.1, 207-227.

[19] Embrechts, P. Klueppelberg ,C. and Mikosch T. (1997). *Modelling Extremal Events for Insurance and Finance.* Springer.

[20] Embrechts, P. Mc Neil and Straumann. (1999). Correlation: Pitfalls and alternatives. A short, non-technical article, *RISK Magazine*, May, 69-71.

[21] Escudero, F., Ortega, E. (2008). Actuarial comparisons for aggregate claims with randomly right-truncated claims. Insurance Math. Econom. 43, 255-262.

[22] Fedotenkov, I. (2018). A review of more than one hundred Pareto-tail index estimators. *Munich Personal RePEc Archive*. Paper No. 90072.

[23] Fisher, R. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of largest or smallest member of a sample. *Proceedings of the Cambridge philosophical society*. 24,180–190.

[24] Galambos, J. (1987). *The Asymptotic Theory of Extreme Order Statistics.* Krieger, Malabar, Florida.

[25] Gardes, L., Stupfler, G. (2015). Estimating extreme quantiles under random truncation. *Test*, 24, 207-227.

[26] Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of Mathematics*, 423-453.

[27] Gouriéroux, C. (2012). *ARCH models and financial applications*. Springer Science & Business Media.

[28] Gumbel, E. J. (1958). *Statistics of extremes*. Columbia University Press.

[29] de Haan L. (1970). *On Regular Variation and its Applications to the Weak Convergence of Sample Extremes.* Mathematical Centre Tract, 32, Amsterdam.

[30] de Haan, L. and Stadtmüller, U. (1996). Generalized regular variation of second order. *J. Australian Math. Soc. (Series A)* 61, 381-395.

[31] de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory: An Introduction*. Springer.

[32] Haeusler, E. and Teugels, J.L. (1985). On asymptotic normality of Hill's estimator for the exponent of regular variation. *Ann. Statist.* 13, 743-756.

[33] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986). *Robust statistics. The approach based on influence functions*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York.

[34] He, S., Yang, G. L. (1998). The strong law under random truncation. *Anna. statist.* 992-1010.

[35] Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution. *Anna. Statist.* 3(5), 1163-1174.

[36] Huber, P.J. (1981). *Robust Statistics.* John Wiley and Sons, New York.

[37] Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* 53, 457-481.

[38] Lawless, J.F. (2011). *Statistical models and methods for lifetime data*, (vol 362). John Wiley and Sons.

[39] Lynden-Bell, D. (1971). A method of allowing for known observational selection in small samples applied to 3CR quasars. *Monthly Notices Roy. Astronom. Soc.* 155, 95-118.

[40] Mason, D.M. (1982). Laws of large numbers for sums of extreme values. *Ann.Probab.* 10, 756-764.

[41] McNeil, A.J. (1999). *Extreme value theory for risk managers. Internal Modelling and CAD II*, published by Risk Books. 93-113.

[42] McNeil, A. J. and Frey, R. (2000). Estimation of Tail-Related Risk Measures for Heterocedastic Financial Times Series: an Extreme Value Approach. *Journal of Empirical Finance.* 7, 271-300.

[43] Pickands, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics*. 3, 119-131.

[44] Rachev, S. (2003). *Handbook of heavy-tailed distributions in finance*. Elsevier.

[45] Reiss, R.D. and Thomas, M. (1997). *Statistical analysis of extreme values. From insurance, finance, hydrology and other fields*. Birkhäuser Verlag, Basel.

[46] Reiss, R.D., Thomas, M. (2007). *Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and other Fields*, 3rd ed. Birkhäuser Verlag, Basel, Boston, Berlin.

[47] Resnick, S. (1997). Heavy tail modelling and teletraffic data: special invited papr. *Anna. Statist.* 25 (5), 1805–1869.

[48] Resnick, S. (2006). *Heavy-tail phenomena: Probabilistic and Statistical Modeling*. Springer.

[49] Sayah, A., Yahia, D., Brahimi, B. (2014). On robust tail index estimation under random censorship. *Afrika Statistika*, 9(1), 671-683.

[50] Smith, R. (1987). Estimating tails of probability distributions. *The Annals of Statistics*, pages 1174-1207.

[51] Stute, W. (1994). The bias of Kaplan-Meir integrals. *Scand. J. Statist.* 21, 475-484.

[52] Stute, W. (1995). The central limit theorem under random censorship. *Ann. Statist.* 23, 422-439.

[53] Stute, W. and Wang, J.L. (1993). The strong law under random censorchip. *Ann. Statist.* 21, 1591-1607.

[54] Stute, W., Wang, J.L. (2008). The central limit theorem under random truncation. *J. Bernoulli.* 14, 604 622.

[55] Worms, J., Worms, R. (2016). A Lynden-Bell integral estimator for extremes of randomly truncated data. *Statist. Probab. Lett.* 109, 106-117.

[56] Woodroofe, M. (1985). Estimating a distribution function with truncated data. *Ann. Statist.* 13, 163-177.

ملخـــــص

في هذه الأطروحة، نقترح إجراء تقديرات قوية وصلبة جديدة لمؤشر الذيل لتوزيعات من نوع باريتو في ظل بيانات غير كاملة (الرقابة أو الاقتطاع). في حالة الاقتطاع، يتم أيضًا اشتقاق تقدير الربيعيات الحادة وتطبيقها على مجموعة بيانات حقيقية حول عمر صفائح فرامل السيارات.

تم إجراء دراسة محاكاة باستخدام البرنامج الإحصائي R لتقييم أداء وقوة المقدرات المقترحة لحجم عينة صغير وكبير. لقد ثبت أن المقدرات الجديدة أكثر قوة وأفضل أداء من المقدرات من نوع هيل استنادًا إلى إحصائيات الترتيب الحادة، في كلتا حالتي البيانات غير المكتملة (الرقابة أو الاقتطاع).

Résumé

Dans cette thèse, nous proposons une nouvelle procédure d'estimation robuste de l'indice de queue pour les distributions de type Pareto sous données incomplètes (censure ou troncature). Sous troncature, l'estimation des quantiles extrêmes est également dérivée et appliquée à un ensemble de données réel sur la durée de vie des plaquettes de frein automobile.

Une étude de simulation à l'aide du logiciel statistique R est réalisée pour évaluer la performance et la robustesse des estimateurs proposés pour des échantillons de petite et grande taille. Nos nouveaux estimateurs se sont révélés plus robustes et plus performants que les estimateurs de type Hill existants basés sur des statistiques d'ordre supérieur, dans les deux cas de données incomplètes (censure ou troncature).

Abstract

In this thesis, we propose a new robust estimation procedure for the tail index for Pareto-type distributions under incomplete data (censorship or truncation). Under truncation, the extreme quantile estimation is also derived and applied to an actual data set on automotive brake pad life.

Simulation study using R statistical software is carried out to evaluate the performance and the robustness of the proposed estimators for small and large sample size. Our newly estimators have been shown to be more robust and perform better than existing Hill-type estimators based on upper order statistics, in both cases of incomplete data (censorship or truncation).