Estimation of Tail Indices of Heavy-Tailed Distributions with Application

Ph.D. thesis

Amenah AL-Najafi

Supervisors:

Dr. Péter Kevei

Dr. László Viharos

Doctoral School of Mathematics
and Computer Science
University of Szeged, Bolyai Institute

Szeged

2021

Acknowledgements

I would like to thank Prof. Czédli Gábor, Professor László Stachó for their professional contribution, and advise, and my supervisors Dr. Péter Kevei and Dr. László Viharos for providing thoughtful supervision; to include valued comments, and expert guidance throughout this project. I wish to express my sincere thanks to Dr. Attila Dénes, and Mr Mahmoud Ibrahim for joint publications. I would like to express my sincere gratitude to the Stipendium Hungaricum Foundation for awarding me a scholarship. I am also thankful to the Bolyai Institute and all its member's staff who contributed to my career development. I am also grateful for the support Ministry of Education, and Ministry of Higher Education and Scientific Research in Iraq for giving me their permission to study in Hungary. I would like to express my gratitude and appreciation to my guru N. Dakhil for her dedicated support and encouragement that has been invaluable throughout this study. I would also like to say a special thanks to my family for all their love and support. I cannot forget to thank my friends for their moral support.

Contents

List of Figures

1

Introduction

The normal distribution is important because of its application in many sciences, including the natural and social sciences, especially because of its shape, which fits many experimental data. Moreover, the central limit theorem states that if we have a sufficiently large number of independent and identically (i.i.d.) random variables with finite variance, the centred and normed sum approximates the normal distribution. However, there are many phenomena that do not follow a normal distribution, and the central limit theorem cannot be applied because of the behaviour of the data, which is dominated by large values. For example, damages caused by hurricanes, financial assets, the intensities of earthquakes, file sizes stored on a server, losses caused by floods and fire insurance losses etc.

A common phenomena followed by all the above events is the heavy-tailed distribution, since the experimental data cannot be described by its mean. Heavy-tailed distributions are probability distributions do not have an exponential moment. Heavy-tailed distributions have a number of applications in computer science, finance, insurance, and economics, for more details see [Res07]. In addition, they are common in physics, astronomy, biology, economics, and the social sciences, see [New05, Sor06]. Frequently it is difficult to choose an appropriate theoretical distribution for a given application. Accordingly, a nonparametric (semiparametric) method is used, see e.g. [Nov12]. The problem of estimating the tail index of probability distributions has received enormous attention in the statistical literature. In 1975, Hill [Hil75] provided a robust estimator based on the asymptotic behavior of extreme values that has been widely used. Alternatively, another estimator was proposed by Pickands [Pic75].
Various modifications have been recommended for both estimators: Gomes and Martin [GM01], Drees [Dre95], Csörgő, Deheuvels and Mason [CDM85] established estimates that are considered special cases of the estimator proposed by Hill [Hil75] and de Haan

[dH81]. While A.L.M. Dekkers, J.H.J. Einmahl and L. de Haan [DEdH89, DdH93] extended Hill's estimator to an estimator for the index of an extreme value distribution. In addition, de Haan and Resnick provided [dHR80] estimators based on order statistics. In 1979, Parzen [Par79] proposed a complementary approach using the density-quantile function as a measure of tail ordering. Later, Schuster [Sch84] developed Parzen's classification system associated with the extreme distance probability limit, and Rojo [Roj96] developed an approach that relaxed the smoothness constraints required in Schuster [Sch84]. Holan and McElroy [HM10] developed an approach based on a Fourier series estimator that provides separate estimates of the left and right tail exponents, and evaluated practical performance through simulation studies. Viharos [Vih99] proposes a whole class of weighted least-squares estimators for the tail index of a regularly varying upper tail of a distribution, proving universal asymptotic normality of the estimators over the entire model.

This dissertation summarises results of research papers [ANV20], [ANSV], [IAND20] and [IAN20]. that have been carried out during my PhD studies, which are presented in chapters 5, 6 and 7. We propose a class of weighted least squares estimators for the tail index of a distribution function with a regularly varying. Our approach is based on the method for the Parzen tail index developed by Holan and McElroy [HM10] with some applications.

Chapter 2, contains some background about estimation techniques, such as weighted least square estimation with some basic definitions.

Chapter 3, this chapter concerns with classical extreme value theory, and its importance and application to many rare phenomena.

Chapter 4, it contains briefly explanation of some estimators that have been proposed for tail index estimation.

Chapter 5, we focused on comparing our suggested a class of weighted least squares (WLS) estimators for the Parzen tail index to the Hill, Pickands, DEdH (Dekkers, Einmahl and de Haan) and ordinary least squares (OLS) estimators using the mean square error.

Chapter 6, by means of simulation, our class of (WLS) estimators for the tail index of a distribution function with a regularly varying upper tail and prior estimators are compared in the Pareto and Hall models using the mean squared error as a criterion.

Chapter 7, we explore the spread of Corona Virus Disease-2019 (COVID19) in Iraq and Egypt. The logistic and Gaussian models were applied to forecast and predict the number of confirmed cases from both countries. An expand generalized SEIR model

for the spread of COVID19 was used, taking into account mildly and symptomatically infected individuals. The extreme value theory approach was also employed to discover and modelling Covid-19 peaks, and the return level.

2

Preliminaries

Let $F(x) = P\{X \leq x\}$ denote the distribution function of the random variable X. The probability that the random variable X exceeds a given value of x is the survival function used in reliability, $P\{X > x\} = 1 - F(x)$. The empirical distribution function estimates the cumulative distribution function (cdf) $F(x)$.

Definition 1 (Empirical distribution function)**.** *Let $X_1, X_2, \ldots$ be independent identically distributed random variables (i.i.d. rv) with distribution function $F(x)$. Denote $X_{1,n} \leq X_{2,n} \leq \cdots \leq X_{n,n}$ the order statistics of the random sample $X_1, \ldots, X_n$, for each $n \geq 1$. The empirical distribution function $F_n(x), x \in R$, is defined by.*

$$F_n(x) = \begin{cases} 0, & if \quad X_{1,n} > x \\ K/n & if \quad X_{K,n} \leq x < X_{K+1,n} \quad (K = 1, 2, \ldots, n-1) \\ 1 & if \quad X_{n,n} \leq x \end{cases}$$

Definition 2 (Quantile function)**.** *Let X be a random variable with distribution function F. Define the quantile function $Q(u)$, $0 \leq u \leq 1$ by*

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

is the left continuous inverse of the right continuous $F(.)$. If $F(.)$ is continuous, $Q(.)$ satisfies

$$Q(u) = F^{-1}(u) = \inf\{x : F(x) = u\}, \qquad F(Q(u)) = u \in [0, 1].$$

As with the distribution function, we have to estimate the quantile by the empirical quantile function.

Definition 3 (Empirical quantile function)**.** *Let $X_1, X_2, \ldots$ a sequence of i.i.d. rv with continuous distribution function F and order statistics $X_{1,n} \leq X_{2,n} \leq \ldots \leq X_{n,n}$ of the*

random sample $X_1, \ldots, X_n$. Let u uniform $(0,1)$ random variables, define the empirical quantile function $Q_n(u)$ for a sample by

$$\begin{aligned} Q_n(u) = F_n^{-1}(u) &= \inf\{x : F_n(x) \geq u\} \\ &= X_{K,n} \quad if \quad \frac{K-1}{n} < u \leq \frac{K}{n}, \;\; K = 1, \ldots, n, \end{aligned}$$

$Q_n(.) = F^{-1}(.)$ is the left continuous inverse of the right continuous $F_n(.)$.

2.1 Estimation Technique

2.1.1 Weighted least square estimation

Ordinary least squares (OLS) regression-based estimation is a simpler method that has attracted much attention. Assuming X is a random variable denoting observable values from a population, the probability model for X is estimated using a sample from the population. The statistical analysis depends on the properties of this estimated probability model that correspond to the population characteristics of interest. Regression analysis is a statistical method that examines the relationships between two or more quantitative variables. The model can be stated as follows:
Let $X_1, \ldots, X_{p-1}$ represent predictor variables and let Y_i be the value of the response variable, the multiple regression model is represented by the following equation,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i(p-1)} + \varepsilon_i, \tag{2.1}$$

where $X_{i1}, X_{i2}, \ldots, X_{i(p-1)}$ are the values of $p-1$ predictor variables in the ith trial (known constants), ε_i is the normal error term with mean zero and constant variance σ^2, $i = 1, 2, \ldots$ and $\beta_0, \ldots \beta_{p-1}$ are regression coefficients (unknown parameters) that we have to estimate based on information from random samples. Assuming that $E(\varepsilon_i) = 0$ implies that the Y_i are independent and $N(0, \sigma^2)$, then the regression model (2.1) is

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1}.$$

One of the classical estimation methods to find the best estimators of the regression parameters is the least squares method. The least squares method is a statistical technique introduced by Gauss (1777-1855) to find the best fit for a set of data points by minimizing the sum of the residuals of the points from the plotted curve and offering the least sum of the squares of the errors. Considering the equation (2.1) and the method of regression coefficients, an estimate of the coefficients is obtained as

$$\hat{\beta} = (X'X)^{-1} X'Y,$$

where X that shows in the above regression model, is defined as a matrix, which consisted of a column of ones as well as columns of the $X_{i,p-1}$ variables.

The usual way to measure the accuracy of an estimator is via its mean square error (MSE):

$$mse(\hat{\theta}) = E(\hat{\theta} - \theta)^2,$$

where $\hat{\theta}$ is the estimator value for the true value θ.
One of the assumptions of the ordinary least squares estimation method is the assumption of constant variance, but in situations where the underlying distribution is continuous but skewed, constant variance cannot be assumed. This situation can best be solved by modifying ordinary least squares using a weighted least square, which allows the variance of the error term to be almost constant.
Let W be a diagonal matrix with weights $W = diag(w_1 \ldots, w_n)$, then the minimization of the weighted sum of squares is denoted by

$$\sum_{i=1}^{n} w_i(Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \cdots - \beta_{p-1} X_{ip-1})^2.$$

The weighted least squares estimator is

$$\hat{\beta} = (X'WX)^{-1}X'WY.$$

2.1.2 Types of convergences

We start as usual with a sequence of random variables $X_1, X_2, \ldots, X_n$ with cdf F. The limiting behavior of the sequence of random variables associate with the concept of convergence. Among the most common concepts of convergence are convergence in distribution, convergence in probability and almost sure convergence. Let X_n is the sample mean, these concepts are associated with the classical central limit theorem, the weak law of large numbers, and the strong law of large numbers, respectively.

Definition 4. *Let $X_1, X_2, \ldots$ be a sequence of random variables, denoted by $X_n, n \geq 1$ with distribution functions F_n. The sequence $X_1, X_2, \ldots$ converges in distribution to the random variable X if*

$$F_n(x) \to F(x), \qquad as \ n \to \infty,$$

for all x where F is continuous, we denote this as $X_n \xrightarrow{D} X$.

In some literature, weak convergence is called convergence in the distribution. Convergence in probability implies convergence in the distribution.

Definition 5. *Let $X, X_1, X_2, \ldots$ random variables defined on the same sample space. The sequence $X_1, X_2, \ldots$ converges in probability to X as $n \to \infty$ if for all $\varepsilon > 0$,*

$$\lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0.$$

Notation: $X_n \xrightarrow{P} X$ as $n \to \infty$

Definition 6. *Let ω be the set of sample points, $X, X_1, X_2, \ldots$ random variables. X_n converges almost surely or with probability 1 to X as $n \to \infty$ if*

$$P(\{\omega : X_n(\omega) \to X(\omega)\}) = 1.$$

We denote this as $X_n \xrightarrow{a.s} X$ as $n \to \infty$.

Convergence with probability 1 implies convergence in probability and convergence in distribution.

2.1.3 Criterion for estimator

Consistency and asymptotic normality are important criteria for estimation.

Definition 7 (Consistency)**.** *Let θ be the true unknown parameter of the distribution of the sample. A sequence of estimators $\hat{\theta}_n$ weakly consistent if it converges to θ in probability, that is, if for all $\varepsilon > 0$*

$$\lim_{n\to\infty} P\left\{|\hat{\theta_n} - \theta| > \varepsilon\right\} = 0.$$

It can be expressed as $\hat{\theta_n} \xrightarrow{P} \theta$.

Definition 8 (Asymptotic Normality)**.** *Suppose that $\hat{\theta}$ is an estimator for true unknown parameter θ, we say that $\hat{\theta}$ is asymptotically normal if it converges to Standard Normal when $n \to \infty$,*

$$\sqrt{n}(\hat{\theta} - \theta) \to N(0, \sigma_\theta^2),$$

where σ_θ^2 is the asymptotic variance of the estimate $\hat{\theta}$. At a rate $1/\sqrt{n}$ the estimator converges fast enough to the unknown parameter θ.

3

Fundamentals of Extreme Value Theory

This chapter deals with classical extreme value theory, its importance and application to many rare phenomena. It concerned with the study and properties of the limit distribution of the maximum.

3.1 Extreme Value Theory

We often focus on the behavior of the average, this average would then be described through the expected value of the distribution, this is called the classical theory approach. But extremes can be more important in some situations. Extreme events are rare by definition, but often their consequences have significant impacts on finance, hydrology, meteorology, geology and public health. The mean or variance of extreme events is not finite. The classical theory and technique based on the empirical distribution function does not provide useful information. Therefore, these events must be treated by other statistical methods, and one of these methods is the extreme value theory (EVT).

EVT is an important branch of statistics, founded by Leonard Tippett (1902-1985). With the help of Fisher, Tippet 1928 [FT28] obtained three asymptotic limits describing the distributions of extremes under the assumption of independent variables. Gumbel [Gum04] collected this theory in his 1958 book Statistics of Extremes, including the Gumbel distributions that bear his name.

The basic goal of EVT is to determine from sequences of observations the probability of events that are more extreme than those previously recorded. Extreme value analysis is widely used in many disciplines and has shown hopeful results in predicting rare events, such as income distribution, flood protection, extreme winds, mining area identification, and mortality.

Extreme value theory (EVT) is the theory of modeling and measuring events that occur with very small probability. Let $X_1, X_2, \ldots,$ be a sequence of independent identically distributed (i.i.d.) random variables with common distribution function F, the theory of sample extremes is concerned with the limit behavior of sample extremes $X_{n:n} = M_n = \max(X_1, \ldots, X_n), n \geq 2$, the maximum value theorem is concerned with the distribution of M_n and its properties when $n \to \infty$. The results for minima $X_{1:n} = m_n = \min(X_1, \ldots, X_n)$ can be obtained from converting the results about maxima, since the sample minimum has the same distribution as the negative of the sample maximum, by applying the rule

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

Let $X_{1,n} \leq X_{2,n} \leq \cdots \leq X_{n,n}$ denote the order statistics of the random sample $X_1, \ldots, X_n$, the random variable $X_{k,n}$ is called the kth upper order statistic.

There is clearly some resemblance between central limit theory and extreme value theory, but central limit theory is concerned with the limit normal distribution for the sums i.i.d. random variables, $X_1 + \ldots + X_n$ as $n \to \infty$, independent of the original distribution function. The analogous situation in extreme value theory with the asymptotic distribution of sample extremes M_n or m_n is one of three possible families known as extreme value distributions as $n \to \infty$. Finding possible limit distributions for sample maxima of independent and identically distributed random variables is most important

Thus, from the concept of an extreme, it hence intuitively that M_n refers to the right upper endpoint. Let x^* its right endpoint, i.e., $x^* := \sup\{x \in \mathbb{R} : F(x) < 1\}$, which may be infinite. The asymmetric behaviour of M_n must be associated with the underlying distribution function F in its right tail near the right endpoint, then

$$\max(X_1, X_2, \ldots, X_n) \xrightarrow{p} x^*, \quad n \to \infty, \quad x^* \leq \infty.$$

The distribution function of $max(X_1, X_2, \ldots, X_n)$ is F^n, under the assumed independence of X_i, the distribution function of M_n is obtained as

$$\begin{aligned} P(M_n \leq x) &= P(X_1 \leq x, X_2 \leq x, \ldots, X_n \leq x) \\ &= P(X_1 \leq x) \times P(X_2 \leq x) \times \cdots \times P(X_n \leq x) \\ &= F^n(x), \end{aligned} \tag{3.1}$$

where F is an unknown distribution function, so we investigate the behaviour of $F^n(x)$ such that: for all $x < x^*$,

$$P(M_n \leq x) = F^n(x) \to 0$$

and for $x \geq x^*$ that

$$P(M_n \leq x) = F^n(x) \to 1.$$

Since M_n is a nondecreasing sequence in n, implies that

$$M_n \xrightarrow{a.s.} x^*, \quad n \to \infty.$$

Illustrate the previous relation that the nondegenerate limit distribution does not exist if we do not normalize M_n.
Let $a_n > 0$ represent a location shift, where the choice of norming constants a_n is not unique such that

$$P\left[\frac{M_n - b_n}{a_n} \leq x\right] = F^n(a_n x + b_n) \xrightarrow{d} G(x), \tag{3.2}$$

converges weakly as $n \to \infty$ for any continuity point x of G. We can say that F belongs to the maximum domain of attraction of a non-degenerate distribution function G where G are all distribution functions that can occur as the limit of $F^n(a_n x + b_n)$ for $n \to \infty$, these distributions are called extreme value distributions that are continuous on $\mathbb{R}$, we will write $F \in MDA(G)$. The following result identifies the class of extreme value distributions. Fisher and Tippett (1928) identified the all limit distribution of (3.2), the result is summarized in the following theorem, a detailed proof of this theorem can be found in Leadbetter et al. (1983) [LLR83].

Theorem 3.1.1. *Let (X_n) be a sequence of (iid) random veriables. If there are norming constants $a_n > 0, d_n \in \mathbb{R}$ such that 3.2 holds as $n \to \infty$ and some nondegenerate distribution function G. Then G is of type of one of the following three classes:*

$$\Phi_\alpha(x) = \begin{cases} 0, & x \leq 0 \\ \exp\left\{-x^{-\alpha}\right\}, & x > 0 \end{cases}$$

for some $\alpha > 0$, this class is called the Frèchet class of distributions (Frèchet (1927)).

$$\Psi_\alpha(x) = \begin{cases} \exp\left\{-(-x)^{\alpha}\right\}, & x \leq 0 \\ 1, & x > 0 \end{cases}$$

for some $\alpha < 0$, this class is sometimes called the reverse-Weibull class of distributions.

$$\Lambda(x) = \exp\left\{-e^{-x}\right\}, \quad x \in \mathbb{R}$$

this class is called the double-exponential or Gumbel distribution. All Φ_α, Ψ_α and Λ represent the extreme value distributions.

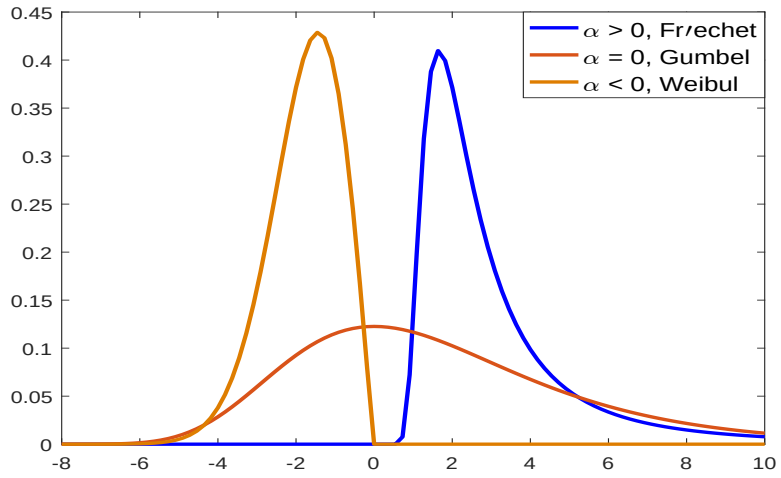


Figure 3.1: Density function for Frèchet, Gumbel and weibull

The first two class are related to regular variation tail behavior. The distributions function F in $MDA(\Phi_\alpha)$ have infinite endpoint, while $x^* < 0$ in the second and the third class it could be the endpoint is finite or infinite. Figure 3.1 shows the shape of the probability density functions for three types of G. Fisher and Tippett (1928) noted that only these three families of distributions are the possible limiting distributions for linear normalization (3.2), regardless of the population and its unknown distribution F. In general, any time extreme value theory is used to analyse a data set, in most cases a prior decision is made as to which of the three families to apply. This is a necessary procedure to follow when deciding which distribution to choose when estimating the distribution of a parameter. This method has drawbacks because the choice of an ideal family may not be the right one. Also, the results may be biased and the model may misrepresent the data.

Jenkinson-von Mises [Jen69], [dHF06] represented that all previous types of distributions are significantly shortened by combining the three models into a single family with the following distribution function:

$$G_\gamma(x) = \exp(-(1+\gamma x)^{-1/\gamma}), \;\; 1+\gamma x > 0, \;\; -\infty < \gamma < \infty,$$

$G_\gamma(x)$ is known as the family of generalised extreme value (GEV) distributions. This distribution is obtained by setting $\gamma = \alpha^{-1}$ and $\gamma = -\alpha^{-1}$ for Fréchet and Weibull, respectively. The parameter γ is called the extreme value index and different values of γ lead to the different extreme value distributions and whether the upper endpoint x^* is finite ($\gamma < 0$) or infinite ($\gamma \geq 0$).

3.2 Regularly varying functions

Regular variation is important for a proper understanding of the extreme value theory founded by Jovan Karamata in a famous paper of 1930 [Kar33]. In his studies, he developed a whole new theory of slowly and regularly varying functions and important properties of these functions, while Eugene Seneta 1976 [Sen76] in his monograph gave a treatment of the basic theory of the subject. Regularly varying functions were later applied in various branches of analysis: Abelian theorems, analytic number theory, etc. The detailed discussion in this section is found in Bingham et al.[BGT89] and Resnick [Res07].

Definition 9 (Slowly varying functions)**.** *Let ℓ be a positive measurable function $\ell : \mathbb{R}^+ \to \mathbb{R}^+$, and satisfying*

$$\lim_{x\to\infty} \ell(ux)/\ell(x) = 1, \qquad \forall u > 0,$$

then ℓ is called slowly varying at infinity for all $u > 0$.

Note that a function that has a finite non zero limit is slowly varying. An example of a slowly varying function f such that $f(x) \to \infty$ as $x \to \infty$ is $f(x) = \log^\alpha(x)$ for some $\alpha \in \mathbb{R}$.

Theorem 3.2.1. *Let f be a nonnegative measurable function. For all $u > 0$,*

$$\frac{f(ux)}{f(x)} \to g(u) \in (0, \infty),$$

where $x \to \infty$, $u \in S$ for some set S, then

1. *$g(u) = u^\alpha$ for some $\alpha \in \mathbb{R}$,*
2. *$f(x) = x^\alpha \ell(x)$ with ℓ slowly varying.*

Definition 10 (Regularly varying)**.** *A positive measurable function $f : \mathbb{R}^+ \to \mathbb{R}^+$, is called regularly varying (at infinity) with index $\alpha \in \mathbb{R}$, written $f \in \mathcal{R}_\alpha$, provided that*

$$\lim_{x\to\infty} \frac{f(ux)}{f(x)} = u^\alpha, \qquad \forall u > 0.$$

Let f be a measurable function, then the above definition can be expressed in various ways, f is regularly varying if and only if $\lim_{u\to\infty} f(xu)/f(u)$ is finite and positive. Moreover, f belongs to regular variation at the origin with index α if f is positive,

$$\lim_{x\to 0} \frac{f(ux)}{f(x)} = u^\alpha, \quad x \downarrow 0, \quad \forall u > 0.$$

In the analysis of phenomena with heavy tails, α, the tail index is always positive. It is also a parameter often used to characterize the type of the tail of distribution: The smaller the value of α, the slower $P(X > x)$ decays with $x \to \infty$ to 0 and the more likely extreme values are to emerge.

We give a brief description of the class Fréchet. The aim is to provide some basics and necessary and sufficient conditions for F in domain of attraction of G where $G = \Phi_\alpha$ ($F \in MDA(\Phi_\alpha)$) and also to characterize a_n and b_n

3.3 Frèchet Distribution Φ_α

The domain of attraction of the Fréchet distribution includes the distribution F whose right tail is regularly varying function with index $-\alpha$. Gnedenko, 1943 [BGT89, Theorem 8.13.2] showed that for some $\alpha > 0$, $F \in MDA(\Phi_\alpha) \iff \bar{F} \in \mathcal{R}_{-\alpha}$, moreover d_n can be chosen to be zero which mean that we have

$$P[M_n/a_n \leq x] = F^n(a_n x) \to \Phi_\alpha(x), \tag{3.3}$$

and we can choose $a_n = \inf\left\{x : \bar{F}(x) \leq \frac{1}{n}\right\}$. In what follows, we show that (3.3) implies that $\bar{F} \in \mathcal{R}_{-\alpha}$. Applying value for $x > 0, \Phi_\alpha(x) = \exp\{-x^{-\alpha}\}$ in (3.3) and take logarithms to get $\lim_{n\to\infty} n(-\log F(a_n x)) = x^{-\alpha}$, use the relation $-\log(1-z) \sim z$ as $z \to 0$, we obtain $\lim_{n\to\infty} n(1 - F(a_n x)) = x^{-\alpha}, x > 0$, and according to de Haan, 1970 [Res08, Proposition 0.4] we obtain $1 - F(x) \sim x^{-\alpha}\ell(x), \quad x \to \infty$ for some $\alpha > 0$.

This class of distribution functions contains very heavy-tailed distributions. Examples of distributions that belong to this class are Pareto, Cauchy, Burr, and Stable with index $\alpha < 2$.

The following theorem gives a sufficient condition for belonging to a maximum domain of attraction of (Φ_α). The condition is called the von Mises condition.

Theorem 3.3.1. *Let $\{X_n\}$ sequence of independent and identically random variables have absolutely continuous with density function $f > 0$ if*

$$\lim_{x\to\infty} \frac{xf(x)}{\bar{F}(x)} = \alpha > 0$$

then F is belong to the domain of attraction of (Φ_α).

We will deal only with the distributions that belong to the maximum domain of attraction of $\Phi_\alpha(x) = \exp\{-x^{-\alpha}\}, x > 0$. One of the most important classes of distributions belonging to MDA(Φ_α) is the Pareto type.

Definition 11. *If X is a random variable that has a Pareto distribution (type I) (PD) with $\alpha \in R$, then the probability that X is greater than a number x is given by*

$$\bar{F}(x) := 1 - F(x) = cx^{-\alpha}, \;\; x \geq 1, \;\; \alpha > 0.$$

Note that the PD (type I) can also be called a power function, so it is considered a special case of a regularly varying distribution.
By inverting the distribution function, the quantile function is.

$$Q(u) = (\frac{1-u}{c})^{-1/\alpha},$$

and it follows that the density function is

$$f(x) = \alpha c x^{-\alpha-1}.$$

Definition 12. *The Generalized Pareto Distribution (GPD) with parameters $\gamma \in \mathbb{R}$ is defined by the cdf*

$$GP_\gamma(x) = \begin{cases} 1 - (1+\gamma x)^{-1/\gamma}, & \gamma \neq 0 \\ 1 - \exp^{-x}, & \gamma = 0 \end{cases}$$

where $x \geq 0$ for $\gamma \geq 0$, and $0 \leq x \leq -1/\gamma$ for $\gamma < 0$.

4

Tail Index Estimators

In general, there are two methods for estimating the extreme value index: parametric estimators, meaning that the data follow an exact GEV distribution, and semiparametric estimators, where the parameter has both a finite-dimensional and an infinite-dimensional and are therefore based on partial properties of the underlying distribution, such as the Pickands, Hill, and DEdH (moment) estimators. In the following, we briefly review the estimators that have been proposed for tail index estimation.

4.1 Hill estimator

Hill's estimator, is one of the most common estimators for the tail index of heavy tailed distributions, which is a type of the maximum likelihood estimator. Numerous applications can be found, for example, in insurance reliability theory, econometrics, geology, and climatology. His approach to extraction inference about tail behaviour was simple and general, for details see De Haan and Ferreira (2007) [dHF06]. Pickands (1975) [Pic75], proposed alternative methods to Hill's, and also one of the modifications of Hill's estimator is the so-called moment estimator proposed by Dekkers et al. (1989) [DEdH89]. A more detailed description is given in the next subsections.

Let $X_{1,n} \leq \ldots \leq X_{n,n}$ be the order statistics based on the sample $X_1 \ldots, X_n$, with distribution F and $X_{k,n}$ is the k^{th} upper order statistic. The intermediate order statistics $X_{n-k,n} \to \infty \ a.s$ and $k = k_n$ be a sequence of positive numbers satisfying the conditions

$$1 \leq k_n < n, k_n \to \infty \ \ and \ \ \frac{k_n}{n} \to 0 \ \ as \ \ n \to \infty, \tag{4.1}$$

Suppose for simplicity that X has a distribution of Pareto type

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

Recall that

$$F(x) \in MDA(\Phi_\alpha) \Longleftrightarrow 1 - F(x) = x^{-\alpha}\ell(x),$$

where ℓ is a slowly varying function, then the Pareto distribution is a special case of the semiparametric approach (regularly varying distribution). To estimate α, Hill proposed an estimator based on upper order statistic k and the sample size as follows

$$\hat{\alpha} = \left(\frac{1}{k_n} \sum_{j=1}^{k_n} \log X_{n-j+1,n} - \log X_{n-k_n,n} \right)^{-1}.$$

The Hill estimator is a function of the number of upper order statistics that depend on the threshold k, choosing the appropriate value of tail observations k is not straightforward and it is not obvious how it should be chosen, the estimate is often very sensitive to the choice k.

In the literature, various methods have been used to select k. Danielsson, Jon, et al. 2016 [DEdHdV16] provided a methodology based on fitting the tail of heavy tailed distribution by minimizing the maximum deviation in the quantile dimension. Clauset and et al. 2009 [CSN09] use the Kolmogorov-Smirnov metric to find the optimal k. Hill plots consider the simplest procedure for choosing the optimal k is, i.e., plots of $\tilde{\alpha}$ against k, and choose the 'optimal' k from a region of the graph at which the relationship seems to have settled down Bol, Georg, et al. 2012 [BNR$^+$12]. Beirlant et al. (1996) [HKKP01] proposed a methodology that selects an optimal k. They proposed a weighted average to derive information about the tail from more than a single Hill estimate, where their procedure uses a number of conventional Hill estimates for different k as input. This is a simple method to obtain unbiased estimates of the tail- index in small samples. Several new estimates of the tail index of the Hill's estimator have been proposed that avoid the k selection problem, for more details see [BNR$^+$12].

The main properties of Hill's estimator are consistency and asymptotic normality. Mason [Mas82] proved that if k_n satisfies (4.1), then the estimator $\hat{\alpha}$ is weakly consistent as an estimator of α

$$\lim_{n\to\infty} P(|\hat{\alpha} - \alpha| \leq \epsilon) = 1, \forall \epsilon > 0.$$

Deheuvels, Haeusler and Mason 1988 [DHM88] provided the almost sure consistency behaviour of $\hat{\alpha}$ such that if $\bar{F} \in \mathcal{R}_\alpha$ and k_n satisfies (4.1), $\hat{\alpha} \to \alpha$ almost surely if and only if $k_n / \log\log n \to \infty$ for all sequences k_n.

The asymptotic normality of $\hat{\alpha}$ was first proved by Hall in 1982 [Hal82], assuming that F satisfies the Hall condition:

$$1 - F(x) = Cx^\alpha[1 + Dx^{-\beta} + o(x^{-\beta})],$$

for some constants $C, \alpha, \beta > 0, D \neq 0$ as $x \to \infty$. This amounts to allowing only a special type of slowly varying function ℓ. The quantile function according to the Hall model is defined as

$$Q_+(1-s) = Cs^{1/\alpha}[1 + Ds^{-\beta/\alpha} + o(s^{-\beta/\alpha})]$$

Hall [Hal82] showed that for $k_n/n \to 0$ as $k_n, n \to \infty$ the $\sqrt{k_n}[\hat{\alpha} - \alpha]$ is asymptotically normal distributed $N(0, \sigma^2)$. The asymptotic normality of Hill,s estimator has led researchers to develop an alternative method of Hill estimator values that is more informative than the standard method, and these modifications work well in the Pareto case see Drees, de Haan, and Resnick 2000 [DdHR00] and Resnick and Stărică 1997 [RS97]. Although it works very well for Pareto-distributed data, the Hill estimator becomes less effective for other regularly varying distribution functions. To illustrate this we have drawn two different samples from two different distributions, Pareto and Burr, with parameters such that the tail index is equal to one. Figure 4.1 (a), (b) shows the Hill estimator for the two data sets. Note that the Hill estimate is plotted against the various values of k. As can be seen in Figure 4.1 (a), the Hill estimator provides a good estimate of the tail index, but it is clear from Figure 4.1 (b) that the tail index is greater than one. The Hill estimator is widely used, however, in practice it is not easy due to the estimation parameter k for which the optimal value is unknown. Therefore, researchers have been attracted to alternative methods to the Hill estimator, one of these methods is based on OLS, see. Gabaix, 1999 [Gab99] and Gabaix and Ibragimov2012 [GI11].

4.2 Dekkers-Einmahl-de Haan's estimator (DEdH)

Hill's estimator is essentially designed for F with regularly varying function, $\alpha > 0$. In Dekkers, Einmahl and de Haan 1989 [DEdH89], the Hill's estimator was developed into an estimator for the index of extreme value distribution, i.e., $\alpha \in \mathbb{R}$. Let $X_1, X_2, \ldots, X_n$ be finite sample with distribution function F such that $\bar{F} \in \mathcal{R}_{-\alpha}$ where $X_{1,n} \leq X_{2,n}, \ldots, \leq X_{n,n}$ are order statistics of $X_1, X_2, \ldots, X_n$ and for $k = k(n) \to \infty, k(n)/n \to 0$ as $n \to \infty$ the moment estimator for the Dekkers, Einmahl and de Haan is defined as

$$\widehat{\alpha}_n = \Big(M_n^{(1)} + 1 - \frac{1}{2}\Big(1 - \frac{(M_n^{(1)})^2}{M_n^{(2)}}\Big)^{-1}\Big)^{-1},$$

where for $j = 1, 2$

$$M_n^{(j)} = \frac{1}{k_n} \sum_{i=0}^{k_n - 1} \left(\log X_{n+1-i,n} - \log X_{n-k_n,n}\right)^j,$$

when $j = 1$ this lead to Hill's estimator.
Dekkers et. al. [DEdH89] proved the following asymptotic properties. of this estimator,

$\hat{\alpha}$ is weak consistent for α,

$$\hat{\alpha}_n \xrightarrow{P} \alpha,$$

when $k(n)/(\log n)^\delta \to \infty$ for some $\delta > 0$, $\hat{\alpha}_n$ is strongly consistent

$$\hat{\alpha}_n \xrightarrow{a.s.} \alpha.$$

The asymptotic normality of $\hat{\alpha}_n$ is

$$\sqrt{k(n)}(\hat{\alpha}_n - \alpha) \to N(0, V),$$

where

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

4.3 Pickands estimator

The simplest estimator for α is the estimator of Pickands (1975) [Pic75]. Denote by $X_{1,n} \leq X_{2,n}, \ldots, \leq X_{n,n}$ the order statistics of $X_1, X_2, \ldots, X_n$ from F such that $\bar{F} \in \mathcal{R}_{-\alpha}$, for $\alpha \in \mathbb{R}$ and $1 \leq k \leq [n/4]$, Pickands [Pic75] proposed his estimator as

$$\widehat{\alpha}_{k,n}^{-1} = \frac{1}{\log 2} \log\left(\frac{X_{n-k_n+1,n} - X_{n-2k_n+1,n}}{X_{n-2k_n+1,n} - X_{n-4k_n+1,n}}\right), \text{ for } k_n = 1, \ldots, \lfloor n/4 \rfloor,$$

where $\lfloor x \rfloor$ denotes the integer part of x. As with the Hill estimator, the choice of k is unclear.

Dekkers and de Haan 1989 [DdH89] gave a fairly natural and general condition to study the properties as follows:

If $k \to \infty, k/n \to 0$ for $n \to \infty$, then $\widehat{\alpha}_{k,n}^{-1}$ is weakly consistent, i.e.

$$\frac{1}{\widehat{\alpha}_{k,n}} \xrightarrow{P} \frac{1}{\alpha}$$

If $k/n \to 0, k/\log\log n \to \infty$ for $n \to \infty$, and X_n is sequence of i.i.d., then

$$\frac{1}{\widehat{\alpha}_{k,n}} \xrightarrow{a.s} \frac{1}{\alpha}$$

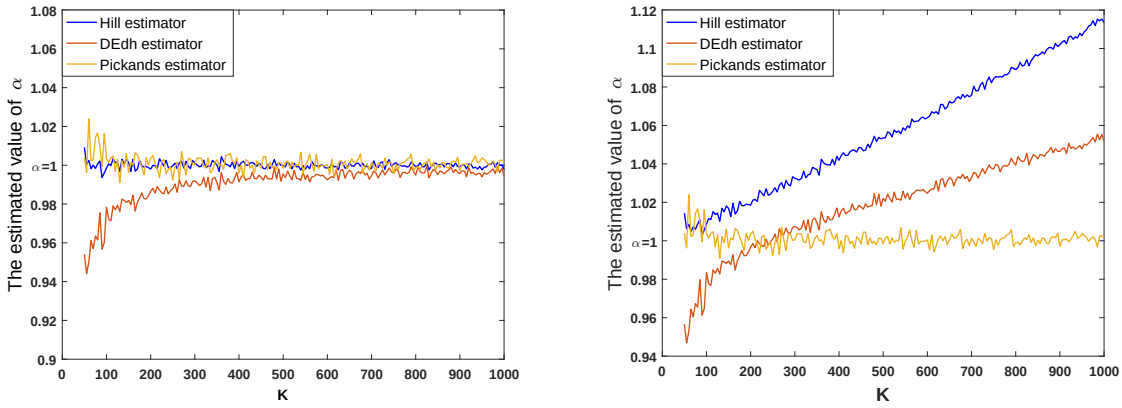
The asymptotic normality of $\widehat{\alpha}_{k,n}^{-1}$, if $k \to \infty, k/n \to 0$ for $n \to \infty$ and X_n is sequence of i.i.d., then

$$\sqrt{k}(\frac{1}{\widehat{\alpha}_{k,n}} - \frac{1}{\alpha}) \to N(0, V),$$

where

$$V = \frac{\alpha^2(2^{2\alpha+1}+1)}{(2(2^{\alpha}-1)\ln 2)^2}.$$

Figure 4.1(a) and (b) show that the Pickands estimator performs well with both the Pareto and Burr distributions. As we have noted, these estimates depend on a relatively



(a) Pareto distributed data (b) Burr distributed data

Figure 4.1: Hill estimator for samples of a Pareto and Burr distributions with tail index 1

small proportion of upper order statistics, so that an estimate is limited to certain values even if the sample size is large. Consequently, an alternative approach to estimate the tail index has been used, see, Politis (2002) [Pol02] where using diverging statistics, Meerschaert and Scheffler (1998) [MS98] using the sample variance and sample size, and McElroy and Politis (2007) using over subsets of the whole data set [MP07].

5

Weighted least squares estimators for the Parzen tail index

Estimation of the tail index of heavy-tailed distributions and its applications are essential in many research areas. We propose a class of weighted least squares (WLS) estimators for the Parzen tail index. Our approach is based on the method developed by Holan and McElroy [HM10]. We investigate consistency and asymptotic normality of the WLS estimators. Through a simulation study, we make a comparison with the Hill, Pickands, DEdH (Dekkers, Einmahl and de Haan) and ordinary least squares (OLS) estimators using the mean square error as criterion. The results show that in a restricted model some members of the WLS estimators are competitive with the Pickands, DEdH and OLS estimators. The results presented in this chapter are based on [ANV20].

5.1 The tail index estimation

The problem of estimating the index of heavy-tailed distributions has received enormous attention in the last decades. Several estimators have been proposed for the tail index among which Hill's [Hil75] estimator is the most classical. A considerable part of the large literature is centered around the asymptotic properties of Hill's estimator. Several generalizations of the Hill estimator have been suggested. A recent generalization was proposed by Ciuperca and Mercadier [CM10] based on weighted power sums of extreme values. More recently, McElroy and Nagaraja [MN16] studied tail index estimators when the sample fraction parameter is fixed. Other estimators were proposed by Pickands [Pic75] and Dekkers et al. [DEdH89], to name a few.

In classical tail index estimation it is assumed that the tail of the distribution function is regularly varying at infinity with some positive index. Parzen [Par79,Par04] studied an alternative model for the tail of the distribution. Let F be an absolutely continuous prob-

ability distribution function with density function f and let Q denote the corresponding quantile function defined as

$$Q(s) := \inf\{x : F(x) \geq s\}, \quad 0 < s \leq 1, \quad Q(0) := Q(0+).$$

Parzen [Par79] used the density-quantile function $fQ(\cdot) = f(Q(\cdot))$ to classify probability distributions. Parzen [Par79] assumed that the limit

$$\nu_1 := \lim_{u \to 1} \frac{(1-u)J(u)}{fQ(u)} \tag{5.1}$$

exists, where J is the score function defined as $J(u) = -(fQ)'(u)$. Assumption (5.1) yields the following approximation for u values near 1:

$$fQ(u) \approx C(1-u)^{\nu_1},$$

for some positive constant C. Based on the parameter ν_1, Parzen [Par79] classified the probability distributions. Heavy tailed distributions correspond to $\nu_1 > 1$.

Parzen [Par04] assumed that $fQ(\cdot)$ is regularly varying at 0 and 1:

$$fQ(u) = u^{\nu_0} L_0(u), \quad u \in [0, 1/2), \tag{5.2}$$

$$fQ(u) = (1-u)^{\nu_1} L_1(1-u), \quad u \in (1/2, 1], \tag{5.3}$$

where $\nu_0, \nu_1 > 0$ are finite constants and L_0 and L_1 are slowly varying at zero. The parameters ν_0 and ν_1 are called the left and right tail exponents of the density-quantile function.

Using Karamata's representation theorem for slowly varying functions ([BGT89, Theorem 1.3.1]), Holan and McElroy [HM10] proved the following result ([HM10, Lemma 1]): If K is a slowly varying function at infinity and $L(x) = K(1/x)$ for $x \in (0, 1)$, then $\log L$ is square integrable. It follows that L_i can be expressed as

$$L_i(u) = \exp\left\{\theta_{i,0} + 2\sum_{k=1}^{\infty} \theta_{i,k} \cos(2\pi k u)\right\}, \quad i = 0, 1. \tag{5.4}$$

In order to estimate the tail exponents, Holan and McElroy [HM10] assumed that L_i satisfies the representation

$$L_i(u) = L_i^{(p_i)}(u) = \exp\left\{\theta_{i,0} + 2\sum_{k=1}^{p_i} \theta_{i,k} \cos(2\pi k u)\right\}, \quad i = 0, 1, \tag{5.5}$$

where p_i is fixed and unknown. In the representation (5.2) and (5.3) they considered $fQ(u)$ for $u \in (0, u_l]$ and $u \in [u_r, 1)$, where $u_l \leq 1/2$ and $u_r \geq 1/2$ are chosen by the

statistician, and they assumed that $p_i < \widetilde{p}_i$, where $\widetilde{p}_i$ is a prespecified integer. Using representation (5.5), we obtain the equations

$$\log fQ(u) = \nu_0 \log u + \theta_{0,0} + 2\sum_{k=1}^{p_0} \theta_{0,k} \cos(2\pi k u), \quad u \in (0, u_l],$$
$$\log fQ(u) = \nu_1 \log(1-u) + \theta_{1,0} + 2\sum_{k=1}^{p_1} \theta_{1,k} \cos(2\pi k (1-u)), \quad u \in [u_r, 1).$$

Based on some estimator $\widehat{fQ}(u)$ of the density-quantile $fQ(u)$, this leads to the regression equations

$$\log \widehat{fQ}(u_j) = \nu_0 \log u_j + \theta_{0,0} + 2\sum_{k=1}^{p_0} \theta_{0,k} \cos(2\pi k u_j) + \varepsilon(u_j),$$
$$\log \widehat{fQ}(1-u_j) = \nu_1 \log u_j + \theta_{1,0} + 2\sum_{k=1}^{p_1} \theta_{1,k} \cos(2\pi k u_j) + \varepsilon(1-u_j),$$

where $\varepsilon(u) = \log\Big(\widehat{fQ}(u)/fQ(u)\Big)$ is the residual process, $u_j = j/n, j = u_{\lceil na \rceil}, \ldots, u_{\lfloor nb \rfloor}$ and $0 < a < b < 1$, so the percentiles u_j are chosen from a subset $[a, b]$ of the interval $(0, 1)$. Holan and McElroy [HM10] obtained some estimators $\widehat{\nu}_0$ and $\widehat{\nu}_1$ for the tail exponents ν_0 and ν_1 using ordinary least squares regression.

We propose a more general class of estimators using weighted least squares regression. We choose some nonnegative weights of the form $w_{j,n} = R(j/n)$ with some weight function R. Set $y_j := \log \widehat{fQ}(u_j)$,

$$y := (y_{\lceil na \rceil}, \ldots, y_{\lfloor nb \rfloor})',$$
$$W := \operatorname{diag}(w_{\lceil na \rceil, n}, \ldots, w_{\lfloor nb \rfloor, n}),$$

and let $X := [G^*, G_0, 2G_1, \ldots, 2G_{\widetilde{p}_0}]$, where

$$G^* = \Big(\log(u_{\lceil na \rceil}), \ldots, \log(u_{\lfloor nb \rfloor})\Big)'$$
$$G_k = \Big(\cos(2\pi k u_{\lceil na \rceil}), \ldots, \cos(2\pi k u_{\lfloor nb \rfloor})\Big)', \quad k = 0, \ldots, \widetilde{p}_0.$$

Set $\beta_{\widetilde{p}_0} := (\nu_0, \theta_{0,0}, \theta_{0,1}, \ldots, \theta_{0,\widetilde{p}_0})'$, where $\theta_{0,j} = 0$ if $j > p_0$. By minimizing the weighted sum of squares

$$\sum_{j=\lceil na \rceil}^{\lfloor nb \rfloor} w_{j,n}\Big(y_j - \nu_0 \log u_j - \theta_{0,0} - 2\sum_{k=1}^{\widetilde{p}_0} \theta_{0,k} \cos(2\pi k u_j)\Big)^2,$$

we obtain the following estimator of $\beta_{\widetilde{p}_0}$:

$$\widehat{\beta}_{\widetilde{p}_0} = (X'WX)^{-1} X'Wy.$$

Then the weighted least squares estimator of ν_0 can be written in the form

$$\widehat{\nu}_0 = e_1' \widehat{\beta}_{\widetilde{p}_0} = e_1'(X'WX)^{-1}X'Wy,$$

where e_1 is the $\widetilde{p}_0 + 2$ dimensional vector defined as $e_1 = (1, 0, 0, \ldots, 0)'$. The right tail exponent ν_1 can be estimated similarly.

A crucial point of this method is to choose a good estimator for the density-quantile $fQ(u)$. Letting $q(u) := Q'(u)$ denote the quantile density function, and using the identity

$$fQ(u)Q'(u) = 1, \tag{5.6}$$

one wish to estimate $q(u)$ instead of $fQ(u)$. Given a sample $X_1, \ldots, X_n$ with distribution function F, let F_n denote its empirical distribution function and define $Q_n := F_n^{-1}$ to be the empirical quantile function. Holan and McElroy [HM10] used the kernel quantile estimator of $q(u)$:

$$\widehat{q}_n(u) = \frac{d}{du} \int_0^1 Q_n(t) K_n(u,t) d\mu_n(t), \quad u \in (0,1), \tag{5.7}$$

where the kernel function $K_n(u,t)$ and the measure μ_n satisfy the following conditions of Cheng [Che95]: (K_1) For every n, $0 < \mu_n([0,1]) < \infty$, and $\mu_n(\{0,1\}) = 0$.

(K_2) For every n and each (u,t), $K_n(u,t) \geq 0$, and for every $u \in [a,b]$, $\int_0^1 K_n(u,t) d\mu_n(t) = 1$.
(K_3) For every n, $\int_0^1 t K_n(u,t) d\mu_n(t) = u, u \in [a,b]$.
(K_4) There is a sequence $\delta_n \downarrow 0$ such that $\sup_{u \in [a,b]} \left| \int_{u-\delta_n}^{u+\delta_n} K_n(u,t) d\mu_n(t) - 1 \right| \downarrow 0$ as $n \uparrow \infty$.

Let S_n be the unique closed subset of (0,1) such that $\mu_n\Big((0,1) \backslash S_n\Big) = 0$ and $\mu_n\Big((0,1) \backslash S_n'\Big) > 0$ for any $S_n' \subset S_n$. For the sequence δ_n in (K_4), let $I_n(u) = [u - \delta_n, u + \delta_n]$, $I_n^c(u) = (0,1) \backslash I_n(u)$, for $u \in [a,b]$. Define $\Lambda(u; K_n) = \int_{I_n(u)} |K_n'(u,t)| d\mu_n(t)$, $u \in [a,b]$, and for a well-defined function g on (0,1), let $\Psi(g; K_n) = \sup_{u \in [a,b]} \int_{I_n^c(u)} |g(t) K_n'(u,t)| d\mu_n(t)$. It is also assumed that the derivative $K_n'(u,t) = \partial K_n(u,t)/\partial u$ satisfies the conditions $(K_5) - (K_7)$ below:

(K_5) For every n, $\sup_{u \in [a,b]} \int_0^1 |K_n'(u,t)| d\mu_n(t) < \infty$.
(K_6) For every n and each $u \in [a,b]$, $K_n(u,t) \equiv 0$, $t \in I_n^c(u)$; or $S_n \subseteq [\varepsilon, 1-\varepsilon] \subset (0,1)$, with $[a,b] \subset [\varepsilon, 1-\varepsilon]$ for some $0 < \varepsilon < 1/2$.
(K_7) For the sequence δ_n in (K_4), $\delta_n^2 \sup_{u \in [a,b]} \Lambda(u; K_n) \to 0$ and $\Psi(1; K_n) \to 0$ as $n \uparrow \infty$.

Similarly as in [HM10], in some cases we assume that the kernel function has the form $K_n(u,t) = K\Big(h_n^{-1}(t-u)\Big)h_n^{-1}$ and satisfies the condition

$$(K_8) \quad \sup_{u\in[a,b]} \left| h_n^{-1} K\Big(\frac{s-u}{h_n}\Big) - h_n^{-1} K\Big(\frac{t-u}{h_n}\Big)\right| \le C_n|t-s|^{\beta} \quad \text{and} \quad |K''(x)| \le C/|x|$$

for some constants $C, \beta > 0$ and $|x|$ sufficiently large, and C_n are positive constants such that $\sup_{n\ge 1} C_n < \infty$.

Moreover, Holan and McElroy [HM10] used the following assumptions of Cheng [Che95] on $q(u)$:

(Q_1) The quantile density function is twice differentiable on (0,1).
(Q_2) There exists a positive constant γ such that $\sup_{u\in(0,1)} u(1-u)|J(u)|/fQ(u)$ $\le \gamma$, where J is the score function in (5.1).
(Q_3) Either $q(0) < \infty$ or $q(u)$ is nonincreasing in some interval $(0, u_*)$, and either $q(1) < \infty$ or $q(u)$ is nondecreasing in some interval $(u^*, 1)$.

We will show that the limit matrix $M(a,b,R) := \lim_{n\to\infty} n^{-1}X'WX$ exists (see the proof of Theorem 5.1.1). Let $(v^*, v_0, \ldots, v_{\widetilde{p_i}})$ be the first row of $M(a,b,R)^{-1}$, and set $G_R(u) := R(u)\Big(v^* \log u + v_0 + 2\sum_{k=1}^{\widetilde{p_i}} v_k \cos(2\pi k u)\Big)$, $i = 0, 1$.

Finally, we assume that the weight function R satisfies the following condition:
(R) R is nonnegative and Riemann integrable on $[a,b]$.

Let $\xrightarrow{P}$ denote convergence in probability, $\xrightarrow{\mathcal{D}}$ denote convergence in distribution, and let $N(\mu,\sigma^2)$ stand for the normal distribution with mean μ and variance σ^2. Limiting and order relations are always meant as $n \to \infty$ if not specified otherwise. Our main results are contained in the following two theorems:

Theorem 5.1.1. *Suppose that the conditions $(Q_1)-(Q_3)$ are satisfied for the quantile density $q(u)$, and $\widehat{q}(u)$ is a kernel smoothed estimator with kernel function satisfying $(K_1)-(K_7)$, the weight function R satisfies the condition (R), and the matrix $M(a,b,R)$ is invertible. Moreover, assume that the percentiles u_j are chosen from a set $[a,b] \subset (0,1)$ such that $u_j = j/n$, $j = \lceil na\rceil, \ldots, \lfloor nb\rfloor$, and $\tilde{p}_i > p_i$, $i = 0, 1$. Then $\widehat{\nu}_i \xrightarrow{P} \nu_i$, $i = 0, 1$.*

Theorem 5.1.2. *Assume that the conditions of Theorem 5.1.1 are satisfied, and suppose that the kernel function is symmetric and differentiable on $[-1,1]$, and satisfies the condition (K_8). Suppose that the derivative $g_R(u) := G_R'(u)$ exists, and g_R and G_R are*

uniformly bounded on $[a,b]$. *Let* h_n *be a sequence such that* $nh_n^2 \to \infty$, $nh_n^4 \to 0$ *and* $h_n \to 0$, *and assume that* $\tilde{p}_i > p_i$, $i = 0, 1$. *Then*

$$\sqrt{n}(\widehat{\nu}_i - \nu_i) \xrightarrow{\mathcal{D}} N(0, V), \quad i = 0, 1,$$

where

$$V = \int_a^b G_R^2(u)du + \int_a^b \int_a^b G_R(u)G_R(v)\left(1 + [(u \wedge v) - uv]\frac{q'(u)q'(v)}{q(u)q(v)}\right)dudv. \tag{5.8}$$

In the special case when the weight function R is identically 1, the two theorems above reduces to Theorems 1 and 2 of [HM10].

5.2 Proofs

The proof of Theorems 5.1.1 and 5.1.2 follows the general outline of the proof of Theorems 1 and 2 in [HM10]. We give a more detailed proof for Theorem 5.1.2.

Proof of Theorem 5.1.1. We deal only with the left tail exponent ν_0, the proof for ν_1 is similar. Set $\gamma = (\gamma_{\lceil na \rceil}, \ldots, \gamma_{\lfloor nb \rfloor})' := \sqrt{W}X(X'WX)^{-1}e_1$ and $\underline{\varepsilon} := \left(\varepsilon(u_{\lceil na \rceil}), \ldots, \varepsilon(u_{\lfloor nb \rfloor})\right)'$. Then $\widehat{\nu}_0 - \nu_0 = \gamma'\sqrt{W}\underline{\varepsilon}$, and hence, using the Cauchy-Schwarz inequality,

$$|\widehat{\nu}_0 - \nu_0| = \left|\sum_{j=\lceil na \rceil}^{\lfloor nb \rfloor} \gamma_j \sqrt{w_{j,n}}\, \varepsilon(u_j)\right| \leq \left(\sum_{j=\lceil na \rceil}^{\lfloor nb \rfloor} \gamma_j^2\right)^{1/2} \left(\sum_{j=\lceil na \rceil}^{\lfloor nb \rfloor} w_{j,n}\, \varepsilon^2(u_j)\right)^{1/2}.$$

We have $\sum_{j=\lceil na \rceil}^{\lfloor nb \rfloor} \gamma_j^2 = \gamma'\gamma = e_1'(n^{-1}X'WX)^{-1}e_1 n^{-1}$ with the matrix

$$X'WX = \begin{bmatrix} \sum_{j=\lceil na \rceil}^{\lfloor nb \rfloor} \log^2 u_j R(u_j) & \sum_{j=\lceil na \rceil}^{\lfloor nb \rfloor} \log u_j R(u_j) & 2\sum_{j=\lceil na \rceil}^{\lfloor nb \rfloor} \log u_j \cos(2\pi u_j) R(u_j) \ldots \\ \sum_{j=\lceil na \rceil}^{\lfloor nb \rfloor} \log u_j R(u_j) & \sum_{j=\lceil na \rceil}^{\lfloor nb \rfloor} R(u_j) & 2\sum_{j=\lceil na \rceil}^{\lfloor nb \rfloor} \cos(2\pi u_j) R(u_j) \ldots \\ \vdots & \vdots & \vdots \end{bmatrix}.$$

Then by Riemann sum approximation

$$\begin{aligned} \lim_{n\to\infty} n^{-1}X'WX &= M(a,b,R) \\ &:= \begin{bmatrix} \int_a^b \log^2 u\, R(u)du & \int_a^b \log u\, R(u)du & 2\int_a^b \log u \cos(2\pi u) R(u)du \ldots \\ \int_a^b \log u\, R(u)du & \int_a^b R(u)du & 2\int_a^b \cos(2\pi u) R(u)du \ldots \\ \vdots & \vdots & \vdots \end{bmatrix}. \end{aligned} \tag{5.9}$$

It follows that for all n large enough $e_1'(n^{-1}X'WX)^{-1}e_1 \leq C$ for some constant C, and hence

$$|\widehat{\nu}_0 - \nu_0| \leq \sqrt{C}\left(n^{-1}\sum_{j=\lceil na\rceil}^{\lfloor nb\rfloor} R(u_j)\,\varepsilon^2(u_j)\right)^{1/2}.$$

Let $C' > 0$ be a constant such that $R(u) \leq C'$, $0 \leq u \leq 1$. Then

$$|\widehat{\nu}_0 - \nu_0| \leq \sqrt{CC'}\left(n^{-1}\sum_{j=\lceil na\rceil}^{\lfloor nb\rfloor} \varepsilon^2(u_j)\right)^{1/2}.$$

Now, by Theorem 2.1 in [Che95], $n^{-1}\sum_{j=\lceil na\rceil}^{\lfloor nb\rfloor}\varepsilon^2(u_j) = o_P(1)$ (cf. the proof of Theorem 1 in [HM10]). $\square$

Proof of Theorem 5.1.2. Write

$$\begin{aligned}\sqrt{n}(\widehat{v_0} - v_0) &= \frac{1}{\sqrt{n}}e_1'(n^{-1}X'WX)^{-1}X'W\underline{\varepsilon}\\ &= \frac{1}{\sqrt{n}}e_1'M(a,b,R)^{-1}X'W\underline{\varepsilon} + \frac{1}{\sqrt{n}}e_1'\Big((n^{-1}X'WX)^{-1} - M(a,b,R)^{-1}\Big)X'W\underline{\varepsilon}.\end{aligned}$$

By straightforward calculation,

$$A_n := \frac{1}{\sqrt{n}}e_1'M(a,b,R)^{-1}X'W\underline{\varepsilon} = \frac{1}{\sqrt{n}}\sum_{j=\lceil na\rceil}^{\lfloor nb\rfloor}\varepsilon(u_j)G_R(u_j). \tag{5.10}$$

It follows from Theorem 5 in [HM10] that

$$A_n \xrightarrow{\mathcal{D}} G_R(b)W(b) - G_R(a)W(a) - \int_a^b W(u)\left(g_R(u) - G_R(u)\frac{q'(u)}{q(u)}\right)du,$$

where $W(u)$ is a Brownian bridge process. The limiting variance is given by (5.8). Next we show that

$$B_n := \frac{1}{\sqrt{n}}e_1'\Big((n^{-1}X'WX)^{-1} - M(a,b,R)^{-1}\Big)X'W\underline{\varepsilon} = o_P(1).$$

Let $(v_n^*, v_{0,n}, \ldots, v_{\widetilde{p}_0,n})$ be the first row of $(n^{-1}X'WX)^{-1} - M(a,b,R)^{-1}$. By (5.9), $(v_n^*, v_{0,n}, \ldots, v_{\widetilde{p}_0,n}) \to \mathbf{0}$. Set

$$G^{(n)}(u) = R(u)\Big(v_n^*\log u + v_{0,n} + 2\sum_{k=1}^{\widetilde{p}_0} v_{k,n}\cos(2\pi k u)\Big).$$

Similarly as in (5.10),

$$\begin{aligned} B_n &= \frac{1}{\sqrt{n}} \sum_{j=\lceil na \rceil}^{\lfloor nb \rfloor} \varepsilon(u_j) G^{(n)}(u_j) \\ &= v_n^* \frac{1}{\sqrt{n}} \sum_{j=\lceil na \rceil}^{\lfloor nb \rfloor} \varepsilon(u_j) R(u_j) \log u_j + v_{0,n} \frac{1}{\sqrt{n}} \sum_{j=\lceil na \rceil}^{\lfloor nb \rfloor} \varepsilon(u_j) R(u_j) \\ &\quad + 2 \sum_{k=1}^{\widetilde{p}_0} v_{k,n} \frac{1}{\sqrt{n}} \sum_{j=\lceil na \rceil}^{\lfloor nb \rfloor} \varepsilon(u_j) R(u_j) \cos(2\pi k u_j). \end{aligned}$$

Each term in the last sum tends to zero, e.g., in the first term $v_n^* \to 0$ and using again Theorem 5 in [HM10], the sequence $\frac{1}{\sqrt{n}} \sum_{j=\lceil na \rceil}^{\lfloor nb \rfloor} \varepsilon(u_j) R(u_j) \log u_j$ has a weak limit. □

5.3 Classical tail index estimation

A distribution function F has right heavy tail with tail index $\alpha_1 > 0$ if $1 - F(x)$ is regularly varying at infinity with index $-1/\alpha_1$, i.e.,

$$1 - F(x) = x^{-1/\alpha_1} \ell_1(x), \quad 0 < x < \infty, \tag{5.11}$$

where ℓ_1 is a function slowly varying at infinity. Similarly, F has left heavy tail with tail index $\alpha_0 > 0$ if $F(-x)$ is regularly varying at infinity with index $-1/\alpha_0$:

$$F(-x) = x^{-1/\alpha_0} \ell_0(x), \quad -\infty < x < 0. \tag{5.12}$$

Let $X_1, X_2, \ldots$ be independent random variables with a common distribution function F having right heavy tail with tail index α_1, and for each $n \in \mathbb{N}$, let $X_{1,n} \le \cdots \le X_{n,n}$ denote the order statistics pertaining to the sample $X_1, \ldots, X_n$. Hill [Hil75] proposed the following estimator for the tail index α_1:

$$\widehat{\alpha}_1 = \frac{1}{k_n} \sum_{j=1}^{k_n} \log X_{n-j+1,n} - \log X_{n-k_n,n} = \frac{1}{k_n} \sum_{j=1}^{k_n} \log \frac{X_{n-j+1,n}}{X_{n-k_n,n}},$$

where the k_n are some integers satisfying

$$1 \le k_n < n, \quad k_n \to \infty \quad \text{and} \quad k_n/n \to 0 \quad \text{as } n \to \infty.$$

The left tail analogue of the Hill estimator is the following:

$$\widehat{\alpha}_0 = \frac{1}{k_n} \sum_{j=1}^{k_n} \log \frac{X_{j,n}}{X_{k_n+1,n}}.$$

We discuss the relation between the Parzen tail model described in equations (5.2), (5.3) and the classical tail model in (5.11) and (5.12). Holan and McElroy [HM10] pointed out that the heavy tailed condition (5.11) imply

$$fQ(1-1/n) = n^{-(1+\alpha_1)}L_1(1/n) \tag{5.13}$$

for some function L_1 slowly varying at zero, if F has a density function with ultimately monotone right tail. Equation (5.13) is the Parzen model (5.3) with $\nu_1 = 1+\alpha_1$ at discrete points $u = 1/n$. A similar relation exists between the left Parzen tail and the classical left tail. Conversely, assume that the Parzen condition (5.3) is satisfied for some $\nu_1 > 1$. Then, using the identity (5.6), we obtain $Q'(1-u) = u^{-\nu_1}/L_1(u)$, $0 < u < 1/2$. Moreover, for $0 < u < u_0 < 1/2$, we have

$$Q(1-u) - Q(1-u_0) = \int_u^{u_0} x^{-\nu_1}/L_1(x)dx \sim \frac{1}{\nu_1 - 1}u^{1-\nu_1}/L_1(u) \quad \text{as } u \downarrow 0$$

using Karamata's theorem in the last step (see [BGT89, Theorem 1.5.11]). It follows that

$$Q(1-u) \sim \frac{1}{\nu_1 - 1}u^{1-\nu_1}/L_1(u) \quad \text{as } u \downarrow 0.$$

This implies that (5.11) holds with $\alpha_1 = \nu_1 - 1$ and some function ℓ_1 slowly varying at infinity (see [BGT89, Theorem 1.5.12]).

On the other hand, the Parzen model (5.3) with $\nu_1 \leq 1$ does not imply that the distribution has right heavy tail: the exponential distribution satisfies condition (5.3) with $\nu_1 = 1$.

Based on the equality $\nu_i = 1 + \alpha_i$, the classical tail index estimators also can be used to estimate the Parzen tail index.

5.4 Comparison of tail index estimators

5.4.1 Asymptotic variances

We evaluate the limiting variance (5.8) for $\widetilde{p}_0 = 1$, different weight functions and tail indices to compare the WLS and the unweighted (ordinary least squares) estimators in the following submodel of (5.4):

$$L_0(u) = \exp\Big\{2\cos(2\pi u)\Big\}, \quad u \in [a, b].$$

The limiting variances are contained in Table 5.1. For the calculations we used numerical integration performed by the Wolfram Mathematica software. We see that in some cases the use of the weights makes the asymptotic variance smaller.

5.4.2 Simulation results

In order to make a comparison with existing proposals, simulations were done performed by the Matlab software. The samples were generated from the model (5.2) with $L_0 \equiv 1$ using different tail indices ν_0. The Hill, Pickands, DEdH (Dekkers, Einmahl and de Haan) and the least squares estimators were included in the simulation study. Similarly as in [HM10], for the simulations we used the Bernstein polynomial estimator of $q(u)$. Let $0 < \varepsilon < 1/2$ be a constant, and assume that $[a, b] \subset [\varepsilon, 1-\varepsilon]$. Set $L_\varepsilon := 1 - 2\varepsilon$ and $t_j := \varepsilon + (j/k)L_\varepsilon$, $j = 0, 1, \ldots, k$. The Bernstein polynomial estimator is defined as

$$\widehat{q}_n^B(u) = \frac{1}{L_\varepsilon^k} \sum_{j=0}^{k-1} \frac{Q_n(t_{j+1}) - Q_n(t_j)}{1/k} \binom{k-1}{j} (u-\varepsilon)^j (1-\varepsilon-u)^{k-1-j}.$$

This estimator belongs to the class (5.7) and satisfies the conditions $(K_1)-(K_7)$. We used the values $k = n = 700$, $\varepsilon = 0.001$, $a = 0.001$ and $b = 0.4$ for the regression estimators, and the weight function $R(u) = u/300$ for the WLS estimator. Tables 5.2 and 5.3 contain the average simulated estimates (mean) and the calculated empirical mean square errors (MSE). We used the sample fraction size $k_n = 100$ for the Hill, Pickands and DEdH estimators. All the simulations were repeated 200 times. We conclude that in the submodel $L_0 \equiv 1$ for α values between 0.8 and 1.5 the WLS estimator has better performance than the OLS estimator. Thus for thinner tails we propose the WLS estimator instead of the OLS estimator. The Hill estimator is the best among the examined estimators. This good performance is not surprising since the Hill estimator was obtained in the special case of (5.11) when the slowly varying function $\ell_1(x)$ is constant for all $x \geq x_{\alpha_1}$, for some threshold x_{α_1}. The Pickands estimator has also good performance. On the other hand, we emphasize that the WLS method can be applied not only for the estimation of the tail index but for the estimation of the slowly varying functions L_i in (2) and (3).

Table 5.1: Limiting variances for different weight functions and tail indices.

$\nu_0 = 1.2$	R(u)				unweighted
	$1+\cos u$	e^{-u}	$-\log u$	$1/u$	
$a=0.1, b=0.4$	821.232	816.812	823.778	851.364	822.13
$a=0.1, b=0.3$	1512.62	1513.46	1538.35	1600.46	1512.83
$a=0.2, b=0.3$	269523	269655	270796	272081	269524

$\nu_0 = 1.8$	R(u)				unweighted
	$1+\cos u$	e^{-u}	$-\log u$	$1/u$	
$a=0.1, b=0.4$	821.962	819.166	829.786	860.498	822.66
$a=0.1, b=0.3$	1521.58	1523.69	1551.68	1617.04	1521.66
$a=0.2, b=0.3$	267666	267807	268969	270267	267666

$\nu_0 = 1.667$	R(u)				unweighted
	$1+\cos u$	e^{-u}	$-\log u$	$1/u$	
$a=0.1, b=0.4$	819.423	816.278	826.109	856.14	820.164
$a=0.1, b=0.3$	1516.49	1518.31	1545.6	1610.22	1516.6
$a=0.2, b=0.3$	268011	268151	269308	270604	268012

$\nu_0 = 2.25$	R(u)				unweighted
	$1+\cos u$	e^{-u}	$-\log u$	$1/u$	
$a=0.1, b=0.4$	840.595	838.929	825.157	885.102	841.151
$a=0.1, b=0.3$	1551.91	1555.02	1585.51	1653.45	1551.89
$a=0.2, b=0.3$	266776	266924	268099	269406	266775

Table 5.2: Average simulated tail index estimates (Mean) for sample size $n = 700$ and for $L_0 \equiv 1$.

	Mean								
	WLS			OLS			Hill	Pickands	DEdH
$\nu(\alpha)$	$\widetilde{p}_0 = 1$	$\widetilde{p}_0 = 2$	$\widetilde{p}_0 = 3$	$\widetilde{p}_0 = 1$	$\widetilde{p}_0 = 2$	$\widetilde{p}_0 = 3$			
2.25(1.25)	2.3777	2.4751	2.5088	2.4271	2.4803	2.4825	2.2396	2.2703	2.7346
2(1)	2.0741	2.1231	2.2423	2.0902	2.1162	2.1177	2.0038	1.9998	2.4988
1.833(0.833)	1.9119	1.9249	1.9405	1.9248	1.904	1.8959	1.8404	1.8471	2.3354
1.667(0.667)	1.7163	1.6915	1.7274	1.7217	1.7019	1.7058	1.6743	1.6902	2.1692
1.556(0.556)	1.5949	1.6294	1.5951	1.6017	1.5822	1.5637	1.5534	1.5567	2.0483
1.5(0.5)	1.5239	1.5448	1.5518	1.5222	1.5613	1.5668	1.5005	1.4942	1.9955
1.333(0.333)	1.3639	1.389	1.3874	1.3598	1.3335	1.3136	1.3347	1.3294	1.8296
1.25(0.25)	1.2956	1.2471	1.242	1.2741	1.2585	1.2629	1.2476	1.2474	1.7426
1.2(0.2)	1.2281	1.2483	1.2189	1.1967	1.2204	1.2089	1.1993	1.2144	1.6942
1.182(0.182)	1.1742	1.1891	1.199	1.1776	1.1725	1.1677	1.1833	1.174	1.6783
1.167(0.167)	1.1628	1.1953	1.1826	1.162	1.158	1.1452	1.167	1.1624	1.662
1.1(0.1)	1.1116	1.0926	1.1538	1.0899	1.0755	1.0725	1.1006	1.0952	1.5955
1.067(0.067)	1.0761	1.106	1.0895	1.0456	1.0597	1.0431	1.0673	1.0562	1.5622
1.05(0.05)	1.0674	1.0607	1.0866	1.0527	1.0476	1.0438	1.0496	1.048	1.5445

Table 5.3: Empirical mean square errors (MSE) of tail index estimates for sample size $n = 700$ and for $L_0 \equiv 1$.

	MSE								
	WLS			OLS			Hill	Pickands	DEdH
$\nu(\alpha)$	$\widetilde{p}_0 = 1$	$\widetilde{p}_0 = 2$	$\widetilde{p}_0 = 3$	$\widetilde{p}_0 = 1$	$\widetilde{p}_0 = 2$	$\widetilde{p}_0 = 3$			
2.25(1.25)	0.0953	0.1565	0.2224	0.1540	0.2701	0.3855	0.0177874	0.0592	0.2525
2(1)	0.0794	0.1121	0.1865	0.1029	0.1244	0.1942	0.0112351	0.0491	0.2600
1.833(0.833)	0.0599	0.1134	0.1550	0.0714	0.1257	0.1673	0.0075016	0.0427	0.2598
1.667(0.667)	0.0594	0.0817	0.1164	0.0565	0.0832	0.1218	0.0062222	0.0412	0.2471
1.556(0.556)	0.0515	0.0935	0.0938	0.0404	0.0593	0.0845	0.0056131	0.0405	0.2482
1.5(0.5)	0.0465	0.1105	0.1352	0.0471	0.0640	0.0909	0.0036438	0.0395	0.2501
1.333(0.333)	0.0400	0.0679	0.1064	0.0292	0.0350	0.0627	0.0033354	0.0397	0.2432
1.25(0.25)	0.0413	0.0754	0.0878	0.0229	0.0445	0.0580	0.0009903	0.0436	0.2447
1.2(0.2)	0.0388	0.0716	0.1090	0.0196	0.0301	0.0456	0.0007893	0.0358	0.2468
1.182(0.182)	0.0335	0.0620	0.0894	0.0216	0.0284	0.0365	0.0007318	0.0335	0.2453
1.167(0.167)	0.0304	0.0708	0.1008	0.0160	0.0341	0.0476	0.0005918	0.0372	0.2462
1.1(0.1)	0.0356	0.0788	0.1001	0.0191	0.0384	0.0489	0.00048686	0.0332	0.2454
1.067(0.067)	0.0358	0.0652	0.1013	0.0169	0.0318	0.0455	0.00024720	0.0313	0.2445
1.05(0.05)	0.0308	0.0625	0.0845	0.0149	0.0238	0.0315	0.00022473	0.0351	0.2443

6

Regression estimators for the tail index

We propose a class of weighted least squares estimators for the tail index of a distribution function with a regularly varying upper tail. Our approach is based on the method developed by Holan and McElroy (2010) for the Parzen tail index. We prove asymptotic normality and consistency for the estimators under suitable assumptions. Through a simulation study, these and earlier estimators are compared in the Pareto and Hall models using the mean squared error as criterion. The results show that the weighted least squares estimator is better than the other estimators investigated. The results presented in this chapter are based on [ANSV].

6.1 Introduction and main result

Let $X_1, X_2, \ldots$ be independent random variables with a common right-continuous distribution function F, and for each $n \in \mathbb{N}$, let $X_{1,n} \leq \cdots \leq X_{n,n}$ denote the order statistics pertaining to the sample $X_1, \ldots, X_n$. Let $\mathcal{R}_\alpha$ be the class of all distribution functions F such that $1 - F$ is regularly varying at infinity with index $-1/\alpha$, that is,

$$1 - F(x) = x^{-1/\alpha}\ell(x), \quad 1 < x < \infty,$$

where ℓ is some positive function on the half line $[1, \infty)$, slowly varying at infinity and $\alpha > 0$ is a fixed unknown parameter to be estimated. Introducing the quantile function Q of F defined as

$$Q(s) := \inf\{x : F(x) \geq s\}, \quad 0 < s \leq 1, \quad Q(0) := Q(0+),$$

it is well known that $F \in \mathcal{R}_\alpha$ if and only for some function L slowly varying at zero,

$$Q(1 - s) = s^{-\alpha}L(s), \quad 0 < s < 1. \tag{6.1}$$

Several estimators exist for the tail index α among which Hill's estimator is the most classical. Hill (1975) [Hil75] proposed the following estimator for the tail index α:

$$\widehat{\alpha}_n^{(H)} = \frac{1}{k_n}\sum_{j=1}^{k_n} \log X_{n-j+1,n} - \log X_{n-k_n,n},$$

where the k_n are positive integers, which in theoretical asymptotic considerations will satisfy the conditions

$$1 \le k_n < n, \quad k_n \to \infty \quad \text{and} \quad k_n/n \to 0 \quad \text{as } n \to \infty.$$

The asymptotic normality of $\widehat{\alpha}_n^{(H)}$ was first considered by Hall (1982) [Hal82] in the following submodel of $\mathcal{R}_\alpha$:

$$1 - F(x) = x^{-1/\alpha} C_1[1 + C_2 x^{-\beta/\alpha}\{1 + o(1)\}], \quad \text{as } x \to \infty,$$

for some constants $C_1 > 0$ and $C_2 \neq 0$. This is equivalent to

$$Q(1-s) = s^{-\alpha} D_1[1 + D_2 s^{\beta}\{1 + o(1)\}], \quad s \to 0, \tag{6.2}$$

where $D_1 = C_1^{\alpha}$ and $D_2 = C_2/C_1^{\beta}$.

Another estimators were proposed by Pickands (1975) [Pic75], Dekkers et al. (1989) [DEdH89], to name a few.

Assuming that F is absolutely continuous with density function f, Parzen (2004) studied the following alternative model for the right tail of the distribution:

$$fQ(s) := f(Q(s)) = (1-s)^{\nu} L_1(1-s), \quad s \in (1/2, 1],$$

where $\nu > 0$ is a finite constant and L_1 is slowly varying at zero. The parameter ν is called the Parzen tail index of the density-quantile function $fQ(\cdot)$.

Based on an orthogonal series expansion for L_1, Holan and McElroy (2010) [HM10] introduced a regression estimator for the Parzen tail index using ordinary least squares. AL-Najafi and Viharos (2020) [ANV20] obtained a more general class of estimators for ν using weighted least squares. We adopt this method to estimate the classical tail index α. Following the idea of Holan and McElroy (2010) [HM10], we assume that the slowly varying function L in (6.1) admits the truncated orthogonal series expansion

$$L(s) = \exp\left\{\theta_0 + 2\sum_{k=1}^{p} \theta_k \cos(2\pi k s)\right\},$$

where $p > 0$ is a fixed integer, and $\theta_0, \ldots, \theta_p$ are unknown parameters. We suppose that $p \le \widetilde{p}$, where $\widetilde{p}$ is a prespecified integer. The knowledge of p is not assumed, condition $p \le \widetilde{p}$ gives only an upper bound for p. It follows that

$$\log Q(1-s) = -\alpha \log s + \theta_0 + 2\sum_{k=1}^{p} \theta_k \cos(2\pi k s). \tag{6.3}$$

Let Q_n be the empirical quantile function defined as

$$Q_n(s) = X_{k,n} \qquad \text{if} \quad \frac{k-1}{n} < s \le \frac{k}{n}, \quad k = 1, 2, \dots, n.$$

Based on the representation (6.3), we obtain the regression equations

$$\log Q_n(1 - s_j) = -\alpha \log s_j + \theta_0 + 2 \sum_{k=1}^{\widetilde{p}} \theta_k \cos(2\pi k s_j) + \varepsilon(s_j),$$

where

$$\varepsilon(s) = \log(Q_n(1-s)/Q(1-s)) \tag{6.4}$$

is the residual process, $s_j = j/n$, $j = \lceil na \rceil, \dots, \lfloor nb \rfloor$, $a < b$ are fixed constants taken from the interval (0,1), and $\theta_k = 0$ for $k > p$. The value $\widetilde{p}$ is chosen by the statistician. We propose a class of estimators for α using weighted least squares. We choose some nonnegative weights of the form $w_{j,n} = R(s_j)$ with some weight function R. Set $y_j := \log Q_n(1 - s_j)$,

$$y := (y_{\lceil na \rceil}, \dots, y_{\lfloor nb \rfloor})',$$

$$W := \operatorname{diag}(w_{\lceil na \rceil, n}, \dots, w_{\lfloor nb \rfloor, n}),$$

and let $X := [G^*, G_0, 2G_1, \dots, 2G_{\widetilde{p}}]$, where

$$G^* = \Big(-\log(s_{\lceil na \rceil}), \dots, -\log(s_{\lfloor nb \rfloor}) \Big)',$$

$$G_k = \Big(\cos(2\pi k s_{\lceil na \rceil}), \dots, \cos(2\pi k s_{\lfloor nb \rfloor}) \Big)', \quad k = 0, \dots, \widetilde{p}.$$

Set $\beta_{\widetilde{p}} := (\alpha, \theta_0, \theta_1, \dots, \theta_{\widetilde{p}})'$. By minimizing the weighted sum of squares

$$\sum_{\lceil na \rceil}^{\lfloor nb \rfloor} w_{j,n} \Big(y_j + \alpha \log s_j - \theta_0 - 2 \sum_{k=1}^{\widetilde{p}} \theta_k \cos(2\pi k s_j) \Big)^2,$$

we obtain the following estimator of $\beta_{\widetilde{p}}$:

$$\widehat{\beta}_{\widetilde{p}} = (X'WX)^{-1} X'Wy.$$

Then the weighted least squares estimator of α can be written in the form

$$\widehat{\alpha}_n^{(W)} := e_1' \widehat{\beta}_{\widetilde{p}} = e_1' (X'WX)^{-1} X'Wy, \tag{6.5}$$

where e_1 is the $\widetilde{p} + 2$ dimensional vector defined as $e_1 = (1, 0, 0, \dots, 0)'$.

We assume the following conditions on the underlying distribution:

(Q_1) The distribution function F is continuous and twice differentiable on (a^*, b^*), where $a^* = \sup\{x : F(x) = 0\}$, $b^* = \inf\{x : F(x) = 1\}$, $-\infty \le a^* < b^* \le \infty$ and $f(x) := F'(x) \neq 0$ on (a^*, b^*).
(Q_2) $\sup_{a^*<x<b^*} F(x)(1 - F(x))|f'(x)/f^2(x)| < \infty$.
(Q_3) $\sup_{1-b\le s\le 1-a} 1/|Q(s)| < \infty$, $\sup_{1-b\le s\le 1-a} 1/fQ(s) < \infty$ and
$\sup_{1-b\le s\le 1-a} 1/|fQ(s)Q(s)| < \infty$.

We will show that the limit matrix $M(a, b, R) := \lim_{n\to\infty} n^{-1}X'WX$ exists (see the proof of Theorem 6.1.1 in Chapter 3). Let $(v^*, v_0, \ldots, v_{\widetilde{p}})$ be the first row of $M(a, b, R)^{-1}$, and set $G_R(u) := R(u)\Big(-v^* \log u + v_0 + 2\sum_{k=1}^{\widetilde{p}} v_k \cos(2\pi k u)\Big)$ for $u \in (0, 1)$.

Moreover, we suppose the following conditions:

(R) The weight function R is nonnegative and Riemann integrable on $[a, b]$.
(M) The matrix $M(a, b, R)$ is invertible.

Now we state our main result for the estimator $\widehat{\alpha}^{(W)}$. Throughout, $\xrightarrow{D}$ denotes convergence in distribution, $\xrightarrow{P}$ denotes convergence in probability, and limiting and order relations are always meant as $n \to \infty$ if not specified otherwise.

Theorem 6.1.1. *Assume that the conditions $Q_1 - Q_3$ are satisfied for the underlying distribution and suppose that the quantile function Q admits the representation (6.3). Moreover, assume the conditions (R) and (M), and assume also that the percentiles s_j are chosen from a closed set $U = [a, b]$, $0 < a < b < 1$, such that $s_j = j/n$, $j = \lceil na \rceil, \ldots, \lfloor nb \rfloor$, and $p \le \widetilde{p}$. Then*

$$\sqrt{n}(\widehat{\alpha}_n^{(W)} - \alpha) \xrightarrow{D} N(0, V), \tag{6.6}$$

where

$$V = \int_a^b \int_a^b \frac{G_R(s)G_R(t)\Big((1-s) \wedge (1-t) - (1-s)(1-t)\Big)}{Q(1-s)Q(1-t)fQ(1-s)fQ(1-t)} dsdt. \tag{6.7}$$

The proof is in Chapter 6.4.

6.2 Asymptotics for $\widetilde{p} \to \infty$

The estimation method proposed in Section 6.1 is heavily based on the assumption $p \le \widetilde{p}$. However, choosing $\widetilde{p} < p$ inflicts a bias. To overcome this difficulty, we adjust our method

to study asymptotics when $\widetilde{p} \to \infty$. In this section our investigation is based on the following series expansion:

$$\log L(s) \sim \sum_{k=0}^{\infty} \theta_k \varphi_k(s),$$

where

$$\varphi_0(s) = \frac{1}{\sqrt{(b-a)R(s)}},$$
$$\varphi_k(s) = \cos\left(\pi k \frac{s-a}{b-a}\right) \frac{1}{\sqrt{(b-a)R(s)/2}}, \quad k = 1, 2, \ldots,$$

and $\theta_k = \int_a^b \log L(x)\varphi_k(x)R(x)dx$. The sequence $\varphi_k\sqrt{R}$, $k = 0, 1, \ldots$, is a complete orthonormal system in $L^2[a, b]$. For convenience, in this section we use the percentiles $s_j = a + j\frac{b-a}{n}$, $j = 0, \ldots, n-1$. Similarly as in Section 6.1, with $y_j := \log Q_n(1 - s_j)$ and $w_{j,n} = R(s_j)$ define

$$y := (y_0, \ldots, y_{n-1})',$$
$$W := \mathrm{diag}(w_{0,n}, \ldots, w_{n-1,n}),$$

and let $X := [G^*, G_0, G_1, \ldots, G_{\widetilde{p}}]$, where

$$\begin{aligned} G^* &= \big(-\log s_0, \ldots, -\log s_{n-1}\big)', \\ G_k &= \big(\varphi_k(s_0), \ldots, \varphi_k(s_{n-1})\big)', \quad k = 0, \ldots, \widetilde{p}. \end{aligned} \tag{6.8}$$

Set

$$b_{\widetilde{p}}(s) := \log L(s) - \sum_{k=0}^{\widetilde{p}} \theta_k \varphi_k(s). \tag{6.9}$$

Recall (6.4). Then we have

$$\log Q_n(1 - s_j) = -\alpha \log s_j + \sum_{k=1}^{\widetilde{p}} \theta_k \varphi_k(s_j) + b(s_j) + \varepsilon(s_j).$$

By minimizing the weighted sum of squares

$$\sum_{\lceil na \rceil}^{\lfloor nb \rfloor} w_{j,n}\Big(y_j + \alpha \log s_j - \sum_{k=0}^{\widetilde{p}} \theta_k \varphi_k(s_j)\Big)^2,$$

we obtain the following estimator of α:

$$\widehat{\alpha}_n^{(W)} = e_1'(X'WX)^{-1}X'Wy.$$

In order to formulate the result for $\widehat{\alpha}_n^{(W)}$, we need the series expansion of the $-\log(\cdot)$ function:

$$-\log s \sim \sum_{j=0}^{\infty} c_j \varphi_j(s), \tag{6.10}$$

where $c_j = \int_a^b (-\log x)\varphi_j(x)R(x)dx$. We assume the following conditions on the sequences $\widetilde{p}$, θ_n and c_n:

(P_1) $\widetilde{p} \to \infty$ and $\widetilde{p}/n \to 0$.
(P_2) For each n, $3(\widetilde{p}+1)/n < 1$.
(P_3) $n\sum_{i=\widetilde{p}+1}^{\infty} c_i^2 \to \infty$.
(P_4) $\theta_n/c_n \to 0$.

Theorem 6.2.1. *Suppose the conditions* $(P_1)-(P_4)$ *are satisfied. Then* $\widehat{\alpha}_n^{(W)} \xrightarrow{P} \alpha$.

6.3 Simulation results

In order to make a comparison with existing proposals, simulations were done performed by the Matlab software. The samples were generated from the strict Pareto model $L \equiv 1$ in (6.1) and from the Hall model (6.2). The Hill, Pickands, DEdH (Dekkers, Einmahl and de Haan) and the weighted least squares (WLS) estimators were included in the simulation study. We used the values $n = 5000$, $a = 0.001$, $b = 0.4$ and $\widetilde{p} = 1, 2, 3$, and the weight function $R(s) = s/500$ for the WLS estimator. In case of $R \equiv 1$, we refer to as ordinary least squares (OLS) estimator. The tail indexes were chosen between 0.5 and 20. For the Hill, Pickands and DEdH estimators the simulations were done for sample size $n = 5000$ and sample fraction size $k_n = 200$. All the simulations were repeated 1000 times.

Tables 6.1 and 6.2 contains the empirical mean square errors (MSE) and the average simulated estimates (mean) for the strict Pareto model. We conclude that in the submodel $L \equiv 1$ for all α values, the WLS estimator performs better than the other estimators investigated.

Tables 6.3 and 6.4 presents the simulation results for the Hall model. Specifically, we used the parameters $D_1 = 0.4$, $D_2 = 1$ and $\beta = 0.01$. We see from Table 6.3 that the WLS estimator performs better than the other estimators, and the OLS estimator is competitive with the Hill estimator especially for $\widetilde{p} = 3$.

Given the values of $[a, b]$, which determines the number of values taken from the simulation data, we experimented with some expanding intervals to find an appropriate

range, and we stop when we obtain reasonable stability of the estimator of α. Figure 6.1 shows the tail index estimates for WLS approach for different values of (a) for the Preto distribution with $\alpha = 1.8$ (left panel) and the $\alpha = 5$ (right panel), the values of the remaining α with both Pareto distribution and Hall model give fairly similar results. The results are almost stable when b=0.45 and (a) is very close to zero, otherwise, the values start to scatter and move away from the true alpha value.

Table 6.1: Empirical mean square errors (MSE) of tail index estimates for the Pareto model and for sample size $n = 5000$.

	MSE								
	WLS			OLS			Hill	Pickands	DEdh
α	$\widetilde{p}=1$	$\widetilde{p}=2$	$\widetilde{p}=3$	$\widetilde{p}=1$	$\widetilde{p}=2$	$\widetilde{p}=3$			
0.5	0.00049	0.000668	0.000945	0.00065	0.00098	0.001357	0.001172	0.017866	0.006558
0.8	0.001183	0.001572	0.002261	0.00161	0.002368	0.00325	0.003325	0.02146	0.008336
1	0.001756	0.002394	0.003668	0.002425	0.003697	0.005203	0.005457	0.024083	0.010687
1.2	0.002821	0.003826	0.005298	0.003641	0.005365	0.007366	0.007532	0.025102	0.01219
1.5	0.00451	0.006126	0.008397	0.005867	0.008671	0.01188	0.01052	0.03013	0.016092
1.8	0.006049	0.007993	0.011399	0.007694	0.011178	0.015334	0.016801	0.035497	0.021695
2	0.007639	0.010499	0.014921	0.010842	0.016055	0.022093	0.020194	0.034981	0.025421
3	0.017668	0.024202	0.034858	0.023523	0.034985	0.047931	0.044665	0.063986	0.049712
4	0.029136	0.040729	0.05895	0.03926	0.058641	0.080589	0.0807	0.094346	0.089062
5	0.047688	0.063472	0.096547	0.064079	0.094958	0.13097	0.114725	0.13557	0.121162
5.5	0.055014	0.076889	0.106532	0.074036	0.110494	0.151476	0.142506	0.16283	0.144236
6	0.071694	0.103854	0.141469	0.089924	0.129628	0.171023	0.173129	0.188113	0.175776
10	0.191172	0.262768	0.375258	0.233466	0.339353	0.45505	0.525182	0.558138	0.527627
15	0.402501	0.535825	0.802723	0.582015	0.884501	1.226799	1.169978	1.167519	1.176961
20	0.792631	1.095608	1.579634	0.996911	1.434474	1.916717	2.100758	1.981171	2.101663

Table 6.2: Average simulated tail index estimates (Mean) for sample size $n = 5000$ and for the Pareto model.

	Mean								
	WLS			OLS			Hill	Pickands	DEdh
α	$\widetilde{p}=1$	$\widetilde{p}=2$	$\widetilde{p}=3$	$\widetilde{p}=1$	$\widetilde{p}=2$	$\widetilde{p}=3$			
0.5	0.500964	0.501233	0.502571	0.503044	0.504023	0.505077	0.501476	0.495427	0.489674
0.8	0.801937	0.802524	0.803656	0.805577	0.807293	0.809021	0.800238	0.801774	0.783686
1	1.001483	1.001634	1.00246	1.005316	1.00711	1.009101	1.001825	1.004785	0.98694
1.2	1.201603	1.201804	1.202563	1.206612	1.208947	1.211492	1.197918	1.195252	1.185589
1.5	1.502324	1.502346	1.502635	1.509168	1.512328	1.515847	1.501775	1.492907	1.485452
1.8	1.805614	1.807831	1.808328	1.812501	1.815819	1.818663	1.801355	1.80158	1.787262
2	2.006075	2.008649	2.012745	2.016946	2.022076	2.026978	2.004505	2.004395	1.988554
3	3.004755	3.002857	3.007692	3.013462	3.017458	3.022898	3.007171	3.002503	2.996076
4	4.00635	4.009942	4.017468	4.028563	4.039037	4.049668	3.985504	3.98685	3.966318
5	5.007934	5.007172	5.011766	5.020999	5.027234	5.034629	5.004943	5.012502	4.98503
5.5	5.521636	5.523414	5.535038	5.54912	5.562017	5.576119	5.498843	5.49632	5.48765
6	6.010705	6.020936	6.035309	6.042542	6.057651	6.071267	6.00263	6.012857	5.987134
10	10.03551	10.0453	10.04212	10.06879	10.0851	10.099	9.997173	10.04161	9.981231
15	15.00041	15.02029	15.05347	15.07633	15.11221	15.14596	15.05984	15.02914	15.0449
20	20.0481	20.05749	20.09294	20.11033	20.14008	20.17114	20.01204	20.04928	19.99807

Table 6.3: Empirical mean square errors (MSE) of tail index estimates for the Hall model and for sample size $n = 5000$.

	MSE								
	WLS			OLS			Hill	Pickands	DEdh
α	$\tilde{p}=1$	$\tilde{p}=2$	$\tilde{p}=3$	$\tilde{p}=1$	$\tilde{p}=2$	$\tilde{p}=3$			
0.5	0.000495	0.000667	0.00092558	0.000632	0.000946	0.001306	0.001159	0.017902	0.00665892
0.8	0.001174	0.001552	0.00222172	0.00156	0.002292	0.003147	0.003306	0.02142	0.00847904
1	0.001749	0.002379	0.00363231	0.002374	0.003616	0.005088	0.00541	0.024003	0.01078627
1.2	0.002806	0.003801	0.00525345	0.003571	0.005259	0.007218	0.007516	0.025114	0.01229618
1.5	0.004482	0.006087	0.00834029	0.005763	0.008519	0.011673	0.010459	0.030153	0.01618835
1.8	0.005985	0.007897	0.01127938	0.007554	0.010987	0.015093	0.016721	0.035417	0.02175322
2	0.007566	0.010387	0.01474723	0.010648	0.015785	0.021747	0.020076	0.034877	0.02545883
3	0.017587	0.024119	0.03469301	0.023338	0.034725	0.047576	0.044474	0.063841	0.04963012
4	0.029026	0.040556	0.0586581	0.038909	0.058141	0.079932	0.08067	0.094312	0.08921482
5	0.04754	0.063301	0.09626703	0.063773	0.094531	0.130401	0.114477	0.135233	0.12110866
5.5	0.054727	0.076546	0.10602299	0.073448	0.109716	0.150488	0.142289	0.162625	0.14413155
6	0.071496	0.103502	0.14091586	0.089385	0.128878	0.170073	0.172846	0.187722	0.17564752
10	0.190659	0.262089	0.37450066	0.232588	0.338214	0.453664	0.524723	0.557207	0.52732507
15	0.402258	0.5353	0.80169824	0.580913	0.882852	1.2246	1.168656	1.166491	1.17578666
20	0.791792	1.094529	1.57797168	0.995368	1.432428	1.914136	2.099641	1.979735	2.10068457

Table 6.4: Average simulated tail index estimates (Mean) for sample size $n = 5000$ and for the Hall model.

	Mean								
	WLS			OLS			Hill	Pickands	DEdh
α	$\tilde{p}=1$	$\tilde{p}=2$	$\tilde{p}=3$	$\tilde{p}=1$	$\tilde{p}=2$	$\tilde{p}=3$			
0.5	0.49603	0.496302	0.497636	0.498107	0.499084	0.500135	0.496567	0.490542	0.484814
0.8	0.797	0.79759	0.798724	0.800636	0.802349	0.804074	0.795342	0.796859	0.77882
1	0.996551	0.996707	0.997539	1.000382	1.002176	1.004164	0.996921	0.999856	0.982061
1.2	1.196672	1.196878	1.197643	1.201678	1.204011	1.206553	1.193032	1.190336	1.180723
1.5	1.497391	1.49742	1.497717	1.50423	1.507388	1.510903	1.496874	1.487989	1.480568
1.8	1.800674	1.802891	1.803397	1.807559	1.810876	1.81372	1.796457	1.796655	1.782377
2	2.001136	2.003709	2.007804	2.011997	2.017123	2.02202	1.999599	1.999456	1.98366
3	2.999823	2.997934	3.00277	3.008533	3.01253	3.017969	3.002265	2.99757	2.991178
4	4.001418	4.005012	4.012537	4.023621	4.03409	4.044716	3.980627	3.981932	3.961447
5	5.003001	5.002247	5.006845	5.016071	5.022308	5.029703	5.000043	5.007562	4.980135
5.5	5.516692	5.518475	5.530098	5.544169	5.557062	5.57116	5.493949	5.491392	5.482761
6	6.005772	6.016001	6.03037	6.037599	6.052704	6.066316	5.997733	6.007918	5.982241
10	10.03057	10.04036	10.03719	10.06385	10.08015	10.09406	9.99228	10.03666	9.97634
15	14.99548	15.01536	15.04854	15.07139	15.10728	15.14102	15.05493	15.0242	15.03999
20	20.04316	20.05255	20.08801	20.1054	20.13515	20.16621	20.00714	20.04434	19.99317

6.4 Proofs

Let $q_n(s)$ be the quantile process defined as

$$q_n(s) = \sqrt{n}(Q_n(s) - Q(s)), \quad 0 < s < 1.$$

The proof is based on the strong approximation of the quantile process.

Theorem 6.4.1. *(Csörgő and Révész (1978) [CR78], Theorem 6.) Suppose that the conditions Q_1 and Q_2 are satisfied. Then on some probability space one can define a sequence $\{B_n(t) : 0 \le t \le 1\}_{n=1}^{\infty}$ of Brownian bridges such that*

$$\sup_{\delta_n \le s \le 1-\delta_n} \left| fQ(s)q_n(s) - B_n(s) \right| \stackrel{a.s.}{=} O(n^{-1/2} \log n),$$

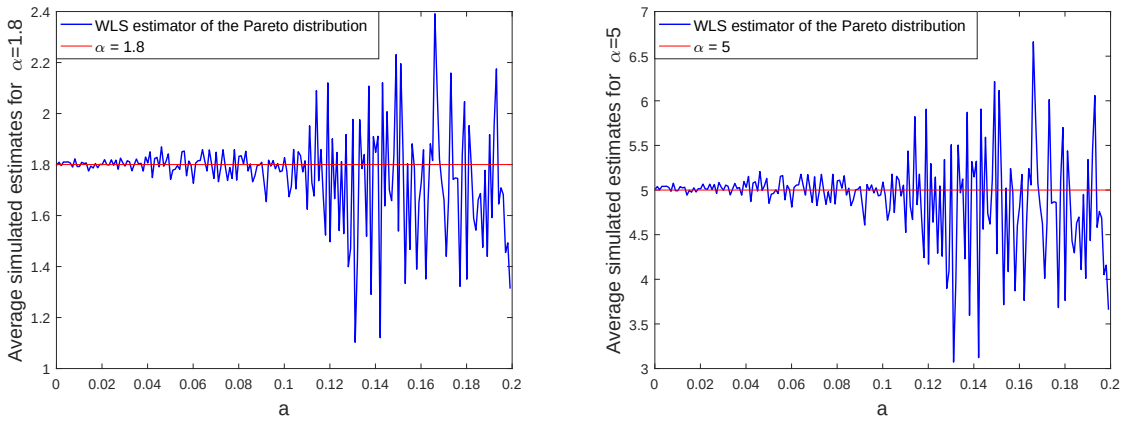


Figure 6.1: Tail index estimates for WLS approach with Pareto distribution in (left panel) from $\alpha = 1.8$ and in (right panel) from $\alpha = 5$.

where $\delta_n = 25n^{-1}\log\log n$.

Proof of Theorem 6.1.1. We assume that the random variables $X_1, X_2, \ldots$ are defined on the probability space given in Theorem 6.4.1. By a simple calculation,

$$X'WX = \begin{bmatrix} \sum_{j=\lceil na\rceil}^{\lfloor nb\rfloor} \log^2 s_j R(s_j) & -\sum_{j=\lceil na\rceil}^{\lfloor nb\rfloor} \log s_j R(s_j) & -2\sum_{j=\lceil na\rceil}^{\lfloor nb\rfloor} \log s_j \cos(2\pi s_j) R(s_j)\ldots \\ -\sum_{j=\lceil na\rceil}^{\lfloor nb\rfloor} \log s_j R(s_j) & \sum_{j=\lceil na\rceil}^{\lfloor nb\rfloor} R(s_j) & 2\sum_{j=\lceil na\rceil}^{\lfloor nb\rfloor} \cos(2\pi s_j) R(s_j)\ldots \\ \vdots & \vdots & \vdots \end{bmatrix}.$$

By Riemann sum approximation, we get

$$\lim_{n\to\infty} n^{-1}X'WX = M(a,b,R) \tag{6.11}$$
$$:= \begin{bmatrix} \int_a^b \log^2 u\, R(u)du & -\int_a^b \log u\, R(u)du & -2\int_a^b \log u \cos(2\pi u) R(u)du\ldots \\ -\int_a^b \log u\, R(u)du & \int_a^b R(u)du & 2\int_a^b \cos(2\pi u) R(u)du\ldots \\ \vdots & \vdots & \vdots \end{bmatrix}.$$

Set $\underline{\varepsilon} := \big(\varepsilon(s_{\lceil na\rceil}), \ldots, \varepsilon(s_{\lfloor nb\rfloor})\big)'$ and

$$y^* := \big(\log Q(1-s_{\lceil na\rceil}), \ldots, \log Q(1-s_{\lfloor nb\rfloor}))\big).$$

Then we have $y^* = X\beta_{\widetilde{p}}$, $\beta_{\widetilde{p}} = (X'WX)^{-1}X'Wy^*$ and hence $\alpha = e_1'\beta_{\widetilde{p}} = e_1'(X'WX)^{-1}X'Wy^*$. It follows that $\underline{\varepsilon} = y - y^*$ and

$$\sqrt{n}(\widehat{\alpha}_n^{(W)} - \alpha) = \frac{1}{\sqrt{n}} e_1'(n^{-1}X'WX)^{-1}X'W\underline{\varepsilon} = Y_n + A_n,$$

where $Y_n = n^{-1/2} e_1' M(a,b,R)^{-1} X' W \underline{\varepsilon}$ and

$$A_n = n^{-1/2} e_1' \Big((n^{-1} X' W X)^{-1} - M(a,b,R)^{-1} \Big) X' W \underline{\varepsilon}.$$

A straightforward calculation yields

$$Y_n = \frac{1}{\sqrt{n}} \sum_{j=\lceil na \rceil}^{\lfloor nb \rfloor} \varepsilon(s_j) G_R(s_j). \tag{6.12}$$

The main point of the proof is to show that

$$\frac{1}{\sqrt{n}} \sum_{j=\lceil na \rceil}^{\lfloor nb \rfloor} \varepsilon(s_j) G_R(s_j) \xrightarrow{D} N(0,V). \tag{6.13}$$

With $\gamma_n(s) := (Q_n(1-s) - Q(1-s))/Q(1-s)$, the residual process can be written as

$$\varepsilon(s) = \log(1 + \gamma_n(s)). \tag{6.14}$$

Set $\eta(x) := \log(1+x) - x$, and let C and δ be some constants such that $\eta(x) \le Cx^2$, if $|x| \le \delta$. Then we obtain $Y_n = Y_{n,1} + A_{n,1}$, where

$$Y_{n,1} = \frac{1}{\sqrt{n}} \sum_{j=\lceil na \rceil}^{\lfloor nb \rfloor} \gamma_n(s_j) G_R(s_j), \quad A_{n,1} = \frac{1}{\sqrt{n}} \sum_{j=\lceil na \rceil}^{\lfloor nb \rfloor} \eta(\gamma_n(s_j)) G_R(s_j).$$

First we show that $A_{n,1} = o_P(1)$. On the event

$$E_n := \left\{ \max_{\lceil na \rceil \le j \le \lfloor nb \rfloor} |\gamma_n(s_j)| \le \delta \right\},$$

we have

$$|A_{n,1}| \le C\sqrt{n} \max_{\lceil na \rceil \le j \le \lfloor nb \rfloor} \gamma_n^2(s_j) \, \frac{1}{n} \sum_{j=\lceil na \rceil}^{\lfloor nb \rfloor} |G_R(s_j)|.$$

With $\kappa_1 := \sup_{1-b \le s \le 1-a} 1/|Q(s)|$, we obtain

$$\max_{\lceil na \rceil \le j \le \lfloor nb \rfloor} \gamma_n^2(s_j) \le \kappa_1^2 \sup_{1-b \le s \le 1-a} (Q_n(s) - Q(s))^2.$$

Set $e_n(s) := fQ(s) q_n(s) - B_n(s)$. With the Brownian bridges in Theorem 6.4.1 and $\kappa_2 := \sup_{1-b \le s \le 1-a} 1/fQ(s)$ we get

$$\begin{aligned} \sup_{1-b \le s \le 1-a} |Q_n(s) - Q(s)| &= \frac{1}{\sqrt{n}} \sup_{1-b \le s \le 1-a} \frac{|e_n(s) + B_n(s)|}{fQ(s)} \\ &\le \frac{\kappa_2}{\sqrt{n}} \sup_{1-b \le s \le 1-a} (|e_n(s)| + |B_n(s)|). \end{aligned}$$

It follows that

$$\sqrt{n}\max_{\lceil na\rceil\le j\le\lfloor nb\rfloor}\gamma_n^2(s_j)\le\frac{\kappa_1^2\kappa_2^2}{\sqrt{n}}\left(\sup_{1-b\le s\le 1-a}|e_n(s)|+\sup_{1-b\le s\le 1-a}|B_n(s)|\right)^2.$$

Applying Theorem 6.4.1, we obtain $\sqrt{n}\max_{\lceil na\rceil\le j\le\lfloor nb\rfloor}\gamma_n^2(s_j)=o_P(1)$. This, in combination with $P(E_n)\to 0$ and $\frac{1}{n}\sum_{j=\lceil na\rceil}^{\lfloor nb\rfloor}|G_R(s_j)|\to\int_a^b|G_R(s)ds$ implies $A_{n,1}=o_P(1)$.

Now we decompose $Y_{n,1}$ as $Y_{n,1}=Y_{n,2}+A_{n,2}$, where

$$Y_{n,2}=\frac{1}{n}\sum_{j=\lceil na\rceil}^{\lfloor nb\rfloor}\frac{B_n(1-s_j)G_R(s_j)}{fQ(1-s_j)Q(1-s_j)},$$
$$A_{n,2}=\frac{1}{n}\sum_{j=\lceil na\rceil}^{\lfloor nb\rfloor}\frac{e_n(1-s_j)}{fQ(1-s_j)Q(1-s_j)}G_R(s_j).$$

To prove that $A_{n,2}=o_P(1)$, we use the inequality

$$A_{n,2}\le\kappa_3\sup_{1-b\le s\le 1-a}|e_n(s)|\,\frac{1}{n}\sum_{j=\lceil na\rceil}^{\lfloor nb\rfloor}|G_R(s_j)|,$$

where

$$\kappa_3=\sup_{1-b\le s\le 1-a}1/|fQ(s)Q(s)|.$$

By Theorem 6.4.1 we have $A_{n,2}=o_P(1)$. We prove that the limit of $Y_{n,2}$ is $N(0,V)$ given in (6.6). By the distributional equality

$$Y_{n,2}\stackrel{D}{=}\frac{1}{n}\sum_{j=\lceil na\rceil}^{\lfloor nb\rfloor}\frac{B(1-s_j)G_R(s_j)}{fQ(1-s_j)Q(1-s_j)},\quad n=1,2,\ldots,$$

where $B(\cdot)$ is a Brownian bridge process, we obtain

$$Y_{n,2}\xrightarrow{D}\int_a^b\frac{B(1-s)G_R(s)}{fQ(1-s)Q(1-s)}ds.$$

The variance of the limit random variable is described in (6.7).

The last step is to prove that $A_n=o_P(1)$. Let $(v_n^*,v_{0,n},\ldots,v_{\widetilde{p},n})$ be the first row of $(n^{-1}X'WX)^{-1}-M(a,b,R)^{-1}$. Using statement (6.11), we have $(v_n^*,v_{0,n},\ldots,v_{\widetilde{p},n})\to\mathbf{0}$. Set

$$G^{(n)}(u):=R(u)\Big(-v_n^*\log u+v_{0,n}+2\sum_{k=1}^{\widetilde{p}}v_{k,n}\cos(2\pi ku)\Big),\quad u\in(0,1).$$

Similarly as in (6.12),

$$\begin{aligned}A_n &= \frac{1}{\sqrt{n}}\sum_{j=\lceil na\rceil}^{\lfloor nb\rfloor}\varepsilon(s_j)G^{(n)}(s_j)\\ &= -v_n^*\frac{1}{\sqrt{n}}\sum_{j=\lceil na\rceil}^{\lfloor nb\rfloor}\varepsilon(s_j)R(s_j)\log s_j + v_{0,n}\frac{1}{\sqrt{n}}\sum_{j=\lceil na\rceil}^{\lfloor nb\rfloor}\varepsilon(s_j)R(s_j)\\ &\quad + 2\sum_{k=1}^{\widetilde{p}} v_{k,n}\frac{1}{\sqrt{n}}\sum_{j=\lceil na\rceil}^{\lfloor nb\rfloor}\varepsilon(s_j)R(s_j)\cos(2\pi k s_j).\end{aligned}$$

Each term in the last sum tends to zero, e.g., in the first term $v_n^* \to 0$ and applying (6.13), in which $G_R(s_j)$ is replaced by $R(s_j)\log s_j$, the sequence $\frac{1}{\sqrt{n}}\sum_{j=\lceil na\rceil}^{\lfloor nb\rfloor}\varepsilon(s_j)R(s_j)\log s_j$ has a weak limit.

□

To prove Theorem 6.2.1, we need three lemmas.

Lemma 6.4.2. *With G_k given in* (6.8) *for $i, j \geq 0$ we have*

$$h_n(i,j) := \frac{b-a}{n}G_i' W G_j = \begin{cases}\frac{1}{n} + \frac{2}{n}(-1)^{i+j}, & \text{if } i \neq j,\\ 1 + \frac{1}{2n} - \frac{(-1)^j}{n}, & \text{if } i = j.\end{cases} \tag{6.15}$$

Proof. If $1 \leq i < j$ then

$$\begin{aligned}h_n(i,j) &= \sum_{k=0}^{n-1}\cos\left(\pi i\frac{s_k-a}{b-a}\right)\cos\left(\pi j\frac{s_k-a}{b-a}\right)\frac{2}{b-a}\\ &= \sum_{k=0}^{n-1}\cos\left(\pi i\frac{k}{n}\right)\cos\left(\pi j\frac{k}{n}\right)\frac{2}{n}\\ &= \frac{1}{n}\sum_{k=0}^{n-1}\left(\cos\left(\pi(i+j)\frac{k}{n}\right) + \cos\left(\pi(i-j)\frac{k}{n}\right)\right)\\ &= \frac{1}{2n}\left(\frac{\sin((n-\frac{1}{2})(\pi\frac{i+j}{n}))}{\sin(\pi\frac{i+j}{2n})} + 1\right) + \frac{1}{2n}\left(\frac{\sin((n-\frac{1}{2})(\pi\frac{i-j}{n}))}{\sin(\pi\frac{i-j}{2n})} + 1\right)\\ &= \frac{1}{n} + \frac{\sin((1-\frac{1}{2n})\pi(i+j))}{n\sin(\frac{i+j}{2}\frac{1}{n})} + \frac{\sin((1-\frac{1}{2n})\pi(i-j))}{n\sin(\frac{i-j}{2}\frac{1}{n})}.\end{aligned}$$

Using the identity $\sin(m\pi - z) = (-1)^{m+1}\sin z$, $m \in \mathbb{Z}$, we have

$$h_n(i,j) = \frac{1}{n}\Big((-1)^{i+j} + (-1)^{i-j} + 1\Big) = \frac{1}{n} + \frac{2}{n}(-1)^{i+j}.$$

The proof for the remaining cases is similar, therefore, is omitted. □

Lemma 6.4.3. *Recall* (6.10). *If* $R(\cdot) \in C^2[a,b]$ *then*

$$c_j = O(j^{-2}) \quad \text{as } j \to \infty.$$

Proof. With $g(z) = \log(a + z(b-a))\sqrt{R(a+b(z-a))}\sqrt{2(b-a)}$,

$$\begin{aligned} -c_j &= \int_a^b (\log x)\sqrt{R(x)}\cos(j\pi\frac{x-a}{b-a})\frac{\sqrt{2}}{\sqrt{b-a}}dx \\ &= \int_0^1 \log(a+z(b-a))\sqrt{R(a+b(z-a))}\sqrt{2(b-a)}\cos(j\pi z)dz \\ &= \int_0^1 g(z)\cos(j\pi z)dz. \end{aligned}$$

Then, integrating by parts, we have

$$\begin{aligned} -c_j &= \frac{1}{j\pi}\sin(j\pi z)g(z)\Big|_{z=0}^1 - \int_0^1 \frac{1}{j\pi}\sin(j\pi z)g'(z)dz \\ &= -\frac{1}{j\pi}\int_0^1 \sin(j\pi z)g'(z)dz \\ &= -\frac{1}{j\pi}[\frac{1}{j\pi}(-\cos(j\pi z))g'(z)\Big|_{z=0}^1 + \int_0^1 \frac{1}{j^2\pi^2}\cos(j\pi z)g''(z)dz] \\ &= \frac{1}{j^2\pi^2}[(-1)^j g'(1) - g'(0) + \int_0^1 \cos(j\pi z)g''(z)dz]. \end{aligned}$$

It follows that

$$|c_j| \le \frac{1}{j^2\pi^2}[|g'(1)| + |g'(0)| + \max_{0\le x\le 1}|g''(x)|] = O(j^{-2}).$$

□

From Lemma 6.4.3 we obtain that the series $\sum_{j=0}^\infty c_j\varphi_j(s)$ converges uniformly on $[a,b]$, and hence,

$$-\log s = \sum_{j=0}^\infty c_j\varphi_j(s), \quad x \in [a,b]. \tag{6.16}$$

Lemma 6.4.4. *If* $\theta_n/c_n \to 0$ *then*

$$\frac{\sum_{j=n}^\infty |c_j\theta_j|}{\sum_{j=n}^\infty c_j^2} \to 0.$$

Proof. Fix $\varepsilon > 0$ and choose N such that $|\theta_n/c_n| < \varepsilon$ is satisfied for all $n > N$. Then for all $n > N$,

$$\frac{\sum_{j=n}^\infty |c_j\theta_j|}{\sum_{j=n}^\infty c_j^2} \le \frac{\sum_{j=n}^\infty \varepsilon c_j^2|}{\sum_{j=n}^\infty c_j^2} = \varepsilon.$$

□

Proof of Theorem 6.2.1. The proof is inspired by the proof of Theorem 3 of [HM10]. Recall (6.4) and (6.9) and set

$$\underline{\varepsilon} := \Big(\varepsilon(s_0), \dots, \varepsilon(s_{n-1})\Big)', \quad \underline{b_{\widetilde{p}}} := \Big(b_{\widetilde{p}}(s_0), \dots, b_{\widetilde{p}}(s_{n-1})\Big)'. \tag{6.17}$$

Similarly as in the proof of Theorem 6.1.1, we have

$$\widehat{\alpha}_n^{(W)} - \alpha = \frac{b-a}{n} e_1' M_n^{-1} X' W (\underline{\varepsilon} + \underline{b_{\widetilde{p}}}), \tag{6.18}$$

where

$$M_n = \frac{b-a}{n} X' W X. \tag{6.19}$$

Now the matrix M_n can be written as

$$M_n = \begin{bmatrix} m_n & r_n' \\ r_n & H_n \end{bmatrix},$$

where $m_n = \frac{b-a}{n} G_*' W G_*$,

$$r_n = \frac{b-a}{n} (G_*' W G_0, \dots, G_*' W G_{\widetilde{p}})', \tag{6.20}$$

and H_n is a $\widetilde{p}+1 \times \widetilde{p}+1$ matrix with elements $h_n(i,j) = \frac{b-a}{n} G_i' W G_j$, $0 \le i, j \le \widetilde{p}$. The inverse of M_n is given by

$$M_n^{-1} = \begin{bmatrix} S^{-1} & -S^{-1} r_n' H_n^{-1} \\ -H_n^{-1} r_n S^{-1} & H_n^{-1} + H_n^{-1} r_n S^{-1} r_n' H_n^{-1} \end{bmatrix},$$

where

$$S = m_n - r_n' H_n^{-1} r_n \tag{6.21}$$

(see e.g. Seber [Seb08, p. 293]). It follows that

$$e_1' M_n^{-1} X' W = (S^{-1}, -S^{-1} r_n' H_n^{-1}) X' W = S^{-1} (G_*' W - r_n' H_n^{-1} R_n), \tag{6.22}$$

where R_n is the last $\widetilde{p}+1$ rows of $X'W$. Let f and g be real functions defined on the interval $[a,b]$ and set $\langle f | g \rangle_n := \frac{b-a}{n} \sum_{j=0}^{n-1} f(s_j) g(s_j)$. Then $h_n(i,j) = \langle \varphi_i | R \varphi_j \rangle_n$, and using (6.16)

$$\frac{b-a}{n} G_*' W G_j = \langle -\log | R \varphi_j \rangle_n = \sum_{i=0}^{\infty} c_i \langle \varphi_i | R \varphi_j \rangle_n = \sum_{i=0}^{\infty} c_i h_n(i,j). \tag{6.23}$$

Thus r_n can be written as

$$r_n = \Big(\sum_{i=0}^{\infty} c_i h_n(i,0), \dots, \sum_{i=0}^{\infty} c_i h_n(i,\widetilde{p}) \Big). \tag{6.24}$$

Define the vectors

$$\underline{c}(\widetilde{p}) = (c_0, \dots, c_{\widetilde{p}})', \quad \underline{d}(\widetilde{p}) = (c_{\widetilde{p}}, c_{\widetilde{p}+1}, \dots)',$$

and let $\widetilde{H}_n$ be the $\widetilde{p}+1 \times \infty$ matrix with elements $h_n(i,j)$, $0 \le i \le \widetilde{p}$, $\widetilde{p} < j$. Equations (6.20) and (6.23) yield

$$r_n' = \underline{c}(\widetilde{p})' H_n + \underline{d}(\widetilde{p})' \widetilde{H}_n'. \tag{6.25}$$

It follows that

$$G_*'W - r_n' H_n^{-1} R_n = G_*'W - \underline{c}(\widetilde{p})' R_n - T_n.$$

where

$$T_n = \underline{d}(\widetilde{p})' \widetilde{H}_n' H_n^{-1} R_n. \tag{6.26}$$

Again by (6.16), $G_*'W = \sum_{j=0}^{\infty} c_j G_j'W$, and by a routine calculation $c(\widetilde{p})' R_n = \sum_{j=0}^{\widetilde{p}} c_j G_j'W$. Therefore, we obtain

$$G_*'W - r_n' H_n^{-1} R_n = \sum_{j=\widetilde{p}+1}^{\infty} c_j G_j'W - T_n. \tag{6.27}$$

Next we examine H_n^{-1}. Let $I_{\widetilde{p}}$ be the $\widetilde{p}+1 \times \widetilde{p}+1$ identity matrix and set $O_n = I_{\widetilde{p}} - H_n$. For an $m \times n$ matrix A with elements a_{ij} define

$$\|A\|_\infty = \max_{\substack{1 \le i \le m \\ 1 \le j \le n}} |a_{ij}|.$$

By Lemma 6.4.2, we have $\|O_n\|_\infty \le 3/n$. This implies that if $3(\widetilde{p}+1)/n < 1$, then $H_n^{-1} = I_{\widetilde{p}} + \sum_{k=1}^{\infty} O_n^k$. By induction, $\|O_n^k\|_\infty \le (\widetilde{p}+1)^{k-1}(3/n)^k$. It follows that for $\widetilde{O} := \sum_{k=1}^{\infty} O_n^k$,

$$\|\widetilde{O}\|_\infty \le \frac{3}{n} \frac{1}{1 - (\widetilde{p}+1)\frac{3}{n}}.$$

Recall (6.26). We show that $\|T_n\|_\infty = O(1/n)$. Using the decomposition $H_n^{-1} = I_{\widetilde{p}} + \widetilde{O}$, we have

$$T_n = \underline{d}(\widetilde{p})' \widetilde{H}_n' R_n + \underline{d}(\widetilde{p})' \widetilde{H}_n' \widetilde{O} R_n. \tag{6.28}$$

Let $(\underline{d}(\widetilde{p})' \widetilde{H}_n')_k$ denote the k-th component of the vector $\underline{d}(\widetilde{p})' \widetilde{H}_n'$. $k = 0, \dots, \widetilde{p}$. Applying Lemmas 6.4.2 and 6.4.3, for some $K_1 > 0$,

$$|(\underline{d}(\widetilde{p})' \widetilde{H}_n')_k| = \Big| \sum_{j=\widetilde{p}+1}^{\infty} h_n(k,j) c_j \Big| \le \frac{3}{n} \sum_{j=\widetilde{p}+1}^{\infty} |c_j| \le \frac{K_1}{n} \sum_{j=\widetilde{p}+1}^{\infty} \frac{1}{j^2} \le \frac{K_1}{n\widetilde{p}}.$$

Furtheremore, letting K_2 be an upper bound of $\varphi_k(\cdot)R(\cdot)$, for some $K_3 > 0$ and $k = 0, \ldots, n-1$ we have

$$\begin{aligned} |(\underline{d}(\widetilde{p})'\widetilde{H}_n' R_n)_k| &= \Big| \sum_{i=0}^{\widetilde{p}} \Big(\sum_{j=\widetilde{p}+1}^{\infty} h_n(k,j) c_j \Big) \varphi_i(s_k) R(s_k) \Big| \\ &\leq K_2 \sum_{i=0}^{\widetilde{p}} \sum_{j=\widetilde{p}+1}^{\infty} |h_n(k,j) c_j| \leq K_2 \sum_{i=0}^{\widetilde{p}} \frac{K_1}{n\widetilde{p}} \leq \frac{K_3}{n}. \end{aligned} \tag{6.29}$$

Thus we have $\|\underline{d}(\widetilde{p})'\widetilde{H}_n' R_n\|_\infty = O(1/n)$. A similar argument yields $\|\underline{d}(\widetilde{p})'\widetilde{H}_n'\widetilde{O} R_n\|_\infty = O(1/n^2)$. We then have

$$\|T_n\|_\infty = O(1/n). \tag{6.30}$$

Next, we turn to examine S in (6.21). The order of the Riemann sum approximation yields

$$m_n = \frac{b-a}{n} \sum_{j=0}^{n-1} R(s_j) \log^2 s_j = \int_a^b R(x) \log^2 x dx + O\Big(\frac{1}{n}\Big). \tag{6.31}$$

Applying (6.25),

$$r_n' H_n^{-1} r_n = \underline{c}(\widetilde{p})' r_n + \underline{d}(\widetilde{p})'\widetilde{H}_n' H_n^{-1} r_n =: r_n^{(1)} + r_n^{(2)}. \tag{6.32}$$

For the first term, similarly an in (6.23), we have

$$\begin{aligned} r_n^{(1)} &= \sum_{j=0}^{\widetilde{p}} c_j \langle R\varphi_j | - \log \rangle_n = \sum_{j=0}^{\widetilde{p}} c_j \sum_{k=0}^{\infty} c_k h_n(k,j) \\ &= \sum_{j=0}^{\widetilde{p}} c_j \sum_{k=0}^{\widetilde{p}} c_k h_n(k,j) + \sum_{j=0}^{\widetilde{p}} c_j \sum_{k=\widetilde{p}+1}^{\infty} c_k h_n(k,j) =: t_n^{(1)} + t_n^{(2)}. \end{aligned} \tag{6.33}$$

A similar argument as in (6.29) yields

$$t_n^{(2)} = O\Big(\frac{1}{n\widetilde{p}}\Big). \tag{6.34}$$

Letting δ_{ij} denote the Kronecker delta, we have

$$t_n^{(1)} = \sum_{j=0}^{\widetilde{p}} c_j^2 + \sum_{j=0}^{\widetilde{p}} \sum_{k=0}^{\widetilde{p}} c_j c_k (h_n(j,k) - \delta_{jk}) =: \sum_{j=0}^{\widetilde{p}} c_j^2 + t_n^{(3)}. \tag{6.35}$$

By Lemmas 6.4.2 and 6.4.3, we have

$$|t_n^{(3)}| \leq \frac{3}{n} \Big(\sum_{j=0}^{\widetilde{p}} c_j \Big)^2 = O\Big(\frac{1}{n}\Big). \tag{6.36}$$

To treat $r_n^{(2)}$ in (6.32) write

$$r_n^{(2)} = \underline{d}(\widetilde{p})'\widetilde{H}_n' r_n + \underline{d}(\widetilde{p})'\widetilde{H}_n'\widetilde{O} r_n.$$

Using again Lemmas 6.4.2 and 6.4.3, we obtain

$$\underline{d}(\widetilde{p})'\widetilde{H}_n' r_n = O\Big(\frac{1}{n\widetilde{p}}\Big) \quad \text{and} \quad \|\underline{d}(\widetilde{p})'\widetilde{H}_n'\widetilde{O}\|_\infty = O\Big(\frac{1}{n^2}\Big). \tag{6.37}$$

The sequence r_n is bounded, since

$$|(r_n)_k| = \Big|\sum_{i=0}^{\infty} c_i h_n(i,k)\Big| \le 3\sum_{i=0}^{\infty}|c_i|.$$

By (6.37) it follows that $\underline{d}(\widetilde{p})'\widetilde{H}_n'\widetilde{O} r_n = O(\widetilde{p}/n^2)$ and hence

$$r_n^{(2)} = O\Big(\frac{1}{n\widetilde{p}}\Big). \tag{6.38}$$

Equations (6.21)-(6.38) above imply

$$S = \int_a^b R(x)\log^2 x dx - \sum_{j=0}^{\widetilde{p}} c_j^2 + O\Big(\frac{1}{n}\Big).$$

By Parseval's equality, S can be written as

$$S = \sum_{\widetilde{p}+1}^{\infty} c_j^2 + O\Big(\frac{1}{n}\Big). \tag{6.39}$$

Recall (6.17) and (6.26). By (6.18), (6.22) and (6.27),

$$\begin{aligned}
\widehat{\alpha}_n^{(W)} - \alpha =& \frac{1}{nS}\Big(\sum_{j=\widetilde{p}+1}^{\infty} c_j G_j' W - T_n\Big)(\underline{\varepsilon} + \underline{b}_{\widetilde{p}}) \\
=& \frac{1}{nS}\Big(\sum_{j=\widetilde{p}+1}^{\infty} c_j G_j'\sqrt{W}\Big)\sqrt{W}\underline{\varepsilon} - \frac{1}{nS}T_n\,\underline{\varepsilon} \\
&+ \frac{1}{nS}\Big(\sum_{j=\widetilde{p}+1}^{\infty} c_j G_j' W\Big)\underline{b}_{\widetilde{p}} - \frac{1}{nS}T_n\,\underline{b}_{\widetilde{p}} \\
=:& \,\xi_n^{(1)} + \xi_n^{(2)} + \xi_n^{(3)} + \xi_n^{(4)}.
\end{aligned} \tag{6.40}$$

where T_n satisfies $\|T_n\|_\infty = O(1/n)$.

Let $\langle x|y\rangle = x'x$ be the inner product of the vectors $x, y \in \mathbb{R}^n$ and let $\|x\|_2 = \sqrt{\langle x|x\rangle}$ denote the induced norm of x. Using the representation (6.14) and Theorem 6.4.1, one can obtain $\sup_{a\le s\le b}|\varepsilon(s)| = O_P(1/\sqrt{n})$, from which it follows that $\big\|\sqrt{W}\underline{\varepsilon}\big\|_2 = O_P(1)$.

Next we determine the order of $\sum_{j=\widetilde{p}+1}^{\infty} c_j G_j' \sqrt{W}$. Then we have

$$\begin{aligned}
\Big\| \sum_{j=\widetilde{p}+1}^{\infty} c_j G_j' \sqrt{W} \Big\|_2^2 &= \Big\langle \sum_{j=\widetilde{p}+1}^{\infty} c_j G_j' \sqrt{W} \Big| \sum_{j=\widetilde{p}+1}^{\infty} c_j G_j' \sqrt{W} \Big\rangle \\
&= \sum_{i=\widetilde{p}+1}^{\infty} \sum_{j=\widetilde{p}+1}^{\infty} c_i c_j \langle G_i' \sqrt{W} | G_j' \sqrt{W} \rangle \\
&= \sum_{i=\widetilde{p}+1}^{\infty} \sum_{j=\widetilde{p}+1}^{\infty} c_i c_j \sum_{m=0}^{n-1} \varphi_i(s_m) \varphi_j(s_m) R(s_m) \\
&= \frac{n}{b-a} \sum_{i=\widetilde{p}+1}^{\infty} \sum_{j=\widetilde{p}+1}^{\infty} c_i c_j h_n(i,j) \\
&= \frac{n}{b-a} \sum_{i=\widetilde{p}+1}^{\infty} \sum_{j=\widetilde{p}+1}^{\infty} c_i c_j (h_n(i,j) - \delta_{ij}) + \frac{n}{b-a} \sum_{i=\widetilde{p}+1}^{\infty} c_i^2.
\end{aligned}$$

By Lemmas 6.4.2 and 6.4.3, for some $K > 0$,

$$\Big\| \sum_{j=\widetilde{p}+1}^{\infty} c_j G_j' \sqrt{W} \Big\|_2^2 \le K \sum_{i=\widetilde{p}+1}^{\infty} \sum_{j=\widetilde{p}+1}^{\infty} \frac{1}{i^2} \frac{1}{j^2} + \frac{n}{b-a} \sum_{i=\widetilde{p}+1}^{\infty} c_i^2 \le \frac{K}{\widetilde{p}^2} + \frac{n}{b-a} \sum_{i=\widetilde{p}+1}^{\infty} c_i^2.$$

Therefore, using the Cauchy-Schwarz inequality, for $\xi_n^{(1)}$ in (6.40) we have

$$\begin{aligned}
|\xi_n^{(1)}| &\le \frac{1}{n|S|} \Big\| \sum_{j=\widetilde{p}+1}^{\infty} c_j G_j' \sqrt{W} \Big\|_2 \, \|\sqrt{W} \underline{\varepsilon}\|_2 \\
&\le \frac{O_P(1)}{\big| n \sum_{\widetilde{p}+1}^{\infty} c_j^2 + O(1) \big|} \sqrt{\frac{K}{\widetilde{p}^2} + \frac{n}{b-a} \sum_{i=\widetilde{p}+1}^{\infty} c_i^2}.
\end{aligned} \tag{6.41}$$

Hence, by condition (P_3) it follows that

$$|\xi_n^{(1)}| = o_P(1). \tag{6.42}$$

For $\xi_n^{(2)}$, by (6.30) and condition (P_3), we have

$$|\xi_n^{(2)}| = \frac{1}{n|S|} |\sum_{m=0}^{n-1} (T_n)_m \varepsilon(s_m)| \le \frac{1}{n|S|} O\Big(\frac{1}{n}\Big) n O_P\Big(\frac{1}{\sqrt{n}}\Big) = \frac{1}{n|S|} O_P\Big(\frac{1}{\sqrt{n}}\Big). \tag{6.43}$$

Next we turn to examine $\xi_n^{(3)}$:

$$\begin{aligned}\left(\sum_{j=\widetilde{p}+1}^{\infty} c_j G_j' W\right)\underline{b}_{\widetilde{p}} &= \sum_{j=\widetilde{p}+1}^{\infty} c_j \sum_{m=0}^{n-1} \varphi_j(s_m) R(s_m) b_{\widetilde{p}}(s_m) \\ &= \sum_{j=\widetilde{p}+1}^{\infty} c_j \sum_{m=0}^{n-1} \varphi_j(s_m) R(s_m) \sum_{k=\widetilde{p}+1}^{\infty} \theta_k \varphi_k(s_m) \\ &= \frac{n}{b-a} \sum_{j=\widetilde{p}+1}^{\infty} \sum_{k=\widetilde{p}+1}^{\infty} c_j \theta_k h_n(k,j) \\ &= \frac{n}{b-a} \sum_{j=\widetilde{p}+1}^{\infty} \sum_{k=\widetilde{p}+1}^{\infty} c_j \theta_k (h_n(k,j) - \delta_{kj}) + \frac{n}{b-a} \sum_{j=\widetilde{p}+1}^{\infty} c_j \theta_j \\ &=: \eta_n^{(1)} + \frac{n}{b-a} \sum_{j=\widetilde{p}+1}^{\infty} c_j \theta_j.\end{aligned}$$

By Lemmas 6.4.2 and 6.4.3, for some $K > 0$,

$$|\eta_n^{(1)}| \le \frac{3}{b-a} \sum_{j=\widetilde{p}+1}^{\infty} \sum_{k=\widetilde{p}+1}^{\infty} |c_j \theta_k| \le K \sum_{j=\widetilde{p}+1}^{\infty} \frac{1}{j^2} \sum_{k=\widetilde{p}+1}^{\infty} |\theta_k| \le \frac{K}{\widetilde{p}} \sum_{k=\widetilde{p}+1}^{\infty} |\theta_k|.$$

Thus

$$|\xi_n^{(3)}| \le \frac{1}{n|S|} \frac{K}{\widetilde{p}} \sum_{k=\widetilde{p}+1}^{\infty} |\theta_k| + \frac{1}{(b-a)|S|} \sum_{j=\widetilde{p}+1}^{\infty} |c_j \theta_j|.$$

Condition (P_3) and Lemma 6.4.4 implies

$$\xi_n^{(3)} \to 0. \tag{6.44}$$

Finally, we examine $\xi_n^{(4)}$. We have

$$T_n \underline{b}_{\widetilde{p}} = \sum_{m=0}^{n-1} (T_n)_m b_{\widetilde{p}}(s_m) = \sum_{m=0}^{n-1} (T_n)_m \sum_{k=\widetilde{p}+1}^{\infty} \theta_k \varphi_k(s_m)$$

Applying (6.30) and condition (P_3), for some $K > 0$ we obtain

$$|\xi_n^{(4)}| \le \frac{K}{n^2|S|} \sum_{m=0}^{n-1} \sum_{k=\widetilde{p}+1}^{\infty} |\theta_k| = \frac{K}{n|S|} \sum_{k=\widetilde{p}+1}^{\infty} |\theta_k| \to 0. \tag{6.45}$$

Equations (6.42)-(6.45) imply the statement of the theorem. □

7

Application

In this chapter, we investigate various different models to study the spread of Corona Virus Disease-2019 (COVID19) in Iraq and Egypt. The logistic and Gaussian models were utilized to forecast and predict the number of confirmed cases from both countries. We estimate the parameters which give the best fit to the incidence data, the results provide severe forecasts for Iraq from February 15 to October 8, 2020 and for Egypt from February 22 to October 8, 2020. Using Gaussian and logistic regression models, a reasonable concord with officially reported cases was shown by the forecasted cases. We developed a generalized SEIR model for the spread of COVID–19 taking into account mildly and symptomatically infected individuals. The extreme value theory approach for finding and modeling Covid-19 peaks was studied, and one of the prime successes EVT is the return level idea. Our sensitivity analyses of the basic reproduction number conclude that the most effective way to prevent COVID–19 cases is decreasing the transmission rate. The findings of this study could therefore assist Iraqi and Egyptian officials to intervene with the appropriate safety measures to handle the increase of the COVID–19 cases. The results presented in this chapter are based on [IAN20, IAND20]

7.1 Methods

7.1.1 Logistics Growth Model

The logistic growth model in mathematical epidemiology is a regression model frequently used to estimate the growth of a population as exponential, then followed by a reduction in growth, with a bound provided by a carrying capacity. The logistic population growth happens if the population growth rate declines with an increase in the number of

individuals. The logistic model introduced in [Bac11, Bat20] takes the form:

$$C' = rC\Big(1 - \frac{C}{K}\Big), \tag{7.1}$$

where C denotes the accumulated number of cases, $r > 0$ is the rate of infection and $K > 0$ is the final epidemic size. The number of infected cases is define as the solution of (7.1) and given by

$$C(t) = \frac{K}{1 + be^{-rt}}, \tag{7.2}$$

where $b = \frac{K - C_0}{C_0}$ and C_0 is the initial population. The parameters r and K can be estimated from the data of the epidemic. The maximum growth rate peaks can be estimated at the time $t_p = \frac{\ln(b)}{r}$, and the number of cases at this time is $C_p = \frac{K}{2}$. Thus, the growth rate at the maximum peak is given by

$$C'(t_p) = \frac{rK}{4}.$$

7.1.2 Gaussian model

To model the time-dependent daily change of infections, we employ a simple Gaussian model. Let $I(t)$ denotes the time-dependent Gaussian function [SSSK20, SS20] and takes the following form:

$$I(t) = I_0 e^{-\left(\frac{t-\mu}{\sigma}\right)^2}, \tag{7.3}$$

where I_0 denotes the maximum value at time μ and σ controls the width. Gaussian model is important despite its simplicity, still have predictive power and use it to predict the peak number of infected per day and peak date, in addition, a numerical study explained that the Gaussian model is a special case of SIR model when imposing gradual of social distancing through a linear drop in the infection rate [BT20]. Gaussian model characterized by three independent parameters: a variance, a maximum height and a time of the maximum height (peak date). Some may argue that with increasing time, the rate of growth should decrease, the behavior of the logistic model and Gaussian model are therefore very similar results: epidemics are initially exponential, and later approaches zero when the population size approaches the carrying capacity thus give rise to bell–shaped daily quantities. The main difference between them is the way that the functions approaching zero. The Gaussian function converges to zero quickly, like the function e^{-x^2} while the logistic function is simply the exponential in that aspect e^x, and the logistic growth model is often used to provide the future total epidemic size of the variable.

7.1.3 Compartmental model for COVID–19 transmission

We spilt the human population into seven compartments: susceptible $S(t)$, exposed $E(t)$, symptomatically infected $I_s(t)$, mildly infected $I_m(t)$, treated $H(t)$, recovered individuals $R(t)$, and $D(t)$ is the individuals who lose their lives due to the COVID–19. The total size of the population at any time t is given by

$$N(t) = S(t) + E(t) + I_m(t) + I_s(t) + H(t) + R(t) + D(t).$$

We do not add separate compartments for the quarantined individuals to keep our model simpler. Susceptible humans (S) are those who can be infected by COVID–19. Once having contracted the disease, an individual progress to the exposed class (E), these individuals do not have any symptoms yet. Following the incubation period, exposed individuals move to one of the symptomatically infected class (I_s) and the mildly infected class (I_m), based on whether that person shows symptoms or not. Mildly infected progress to the recovered class (R) or the symptomatically compartment (I_s). Symptomatically infected move to the treated compartment (H) which includes those who reported hospitalized. After the infectious period treated individuals move the recovered class (R).

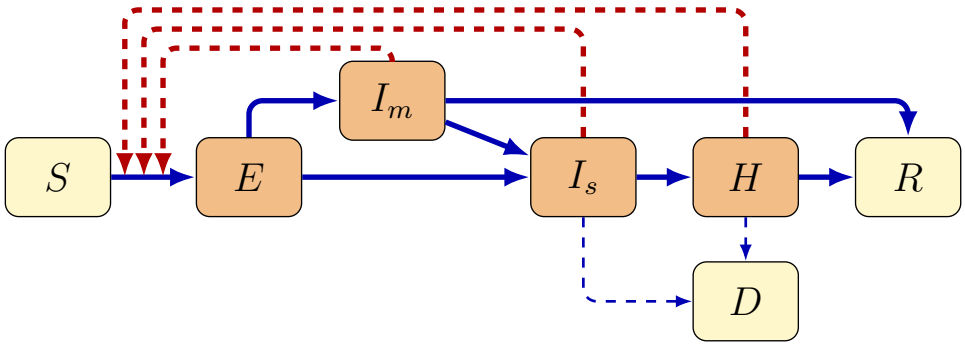


Figure 7.1: Follow diagram of the COVID–19 transmission. Blue arrows indicate transition from one compartment to another. Light and dark coloured nodes depict noninfected and infected compartments, respectively.

The transmission dynamics is shown in the flow diagram (see Figure 7.1) and our

model takes the form

$$\begin{aligned}
S'(t) = & -\beta\frac{\beta_e E(t)+\beta_m I_m(t)+I_s(t)+\beta_h H(t)}{N(t)-D(t)}S(t),\\
E'(t) = & \beta\frac{\beta_e E(t)+\beta_m I_m(t)+I_s(t)+\beta_h H(t)}{N(t)-D(t)}S(t)-\nu E(t),\\
I'_m(t) = & \theta\nu E(t)-\sigma_m I_m(t)-\sigma I_m(t),\\
I'_s(t) = & (1-\theta)\nu E(t)+\sigma I_m(t)-\sigma_s I_s(t)-\delta_s I_s(t),\\
H'(t) = & \sigma_s I_s(t)-\sigma_h H(t)-\delta_h H(t),\\
R'(t) = & \sigma_m I_m(t)+\sigma_h H(t),\\
D'(t) = & \delta_s I_s(t)+\delta_h H(t).
\end{aligned} \tag{7.4}$$

The description of the model parameters are summarized in Table 7.1. Particularly, β stands for transmission rate from symptomatically infected to susceptible, while for transmission rates from exposed, mildly infected, and treated to susceptible are obtained by multiplying β by β_e, β_m and β_h, respectively. The parameter θ is the fraction of asymptomatically infected among all infected people. The length of latent period for humans is $1/\nu$ and $1/\sigma_m, 1/\sigma_h$ denote the length of infected period for mildly and symptomatically infected people, respectively, while $1/\sigma$ is the length of the mildly infected period until one gets severely infected.

Table 7.1: Description of the model (7.4) parameters.

Parameters	**Description**
β	Transmission rate from infectious classes to susceptible
$\beta_e, \beta_m, \beta_h$	The relative transmissibility of E, I_m and H, respectively
θ	Proportion of asymptomatic infections
σ	Progression rate from I_m to I_s
σ_s	Progression rate from I_s to H
σ_m, σ_h	Recovery rates
δ_s, δ_h	Disease-induced death rates
ν	Incubation rate

7.1.4 Parameters estimation and Sensitivity

To estimate the parameters of the logistic growth model (7.1) and the Gaussian model (7.3), which give the best fit to the epidemic data, we employ the nonlinear least-square curve fitting to the incidence curve given by equation $C'(t)$ and $I(t)$, respectively. As per the first observation, the initial number of C_0 cases was set. The parameters of the model (7.4) giving the best fit to data can be estimated by applied Latin Hypercube Sampling. This

is a sampling method which is used to measure the variation of several parameter values at the same time (see [MBC79] for details). The key idea of the method is to produce a representative sample from the ranges for all parameters shown in Table 7.6. To every element of this sample set, the solutions of the model (7.4) are (numerically) calculated. Finally, we apply the least-squares method to find the parameters providing the best fit. To identify the parameters w.r.t their effect on the basic reproduction number, we will apply Partial Rank Correlation Coefficients analysis (PRCC, see, e.g. [BD94]), to perform sensitivity analysis. The PRCC-based sensitivity analysis measures the effect of the parameters on the basic reproduction number, while we change the parameters in the given ranges.

7.1.5 Return level estimation

The application of EVT offers different techniques to study the behavior of a sample with very high or very low levels. One of the important techniques of extreme value theory is the idea of the return level. The return level is strongly related to the generalized Pareto distribution (GPD). We will use it to investigate the upper tail distribution properties of the infection of the COVID-19 epidemic. In this section, we assume that the daily recorded new confirmed cases are independent and identically distributed. This assumption is needed for our statistical analysis, and it is also assumed in [AKI+20]. In this subsection, we follow the methods and definitions given in [CBTD01, TR19].
Let X be a random variable with unknown cumulative distribution function F, the distribution function of the excesses over the threshold u of this random variable is called excess distribution function over the threshold u denoted by F_u, defined as

$$F_u(x) = P(X - u \leq x \mid X > u) = \frac{F(u+x) - F(u)}{1 - F(u)}, \qquad 0 \leq x \leq x^* - u, \tag{7.5}$$

where $1 - F(u)$ is the exceedance probability, and x^* right endpoint it could be finite or infinite. The mean excess function of X denote the mean residual life function is

$$e(u) = E(X - u \mid X > u), \qquad 0 \leq u < x^*, \tag{7.6}$$

the plot threshold u against $e(u)$ is linear in case F_u approximate to GPD [EKM13, section 6.2.2].
The method is based on exceedances of a certain threshold, this method is preferred by practitioners because it uses the data more efficiently. Provided that the appropriate distribution is chosen and then the parameters are estimated, it is reasonable to calculate the return level. Pickands (1975) [Pic75] provided a very helpful result which is stated in the following theorem:

Theorem 7.1.1. *Let $X_l, X_2, ...$ be a sequence of independent random variables with distribution function F, and suppose that $F \in MDA(G_\gamma(x))$. For sufficiently large u, the conditional excess distribution function $F_u(x)$ is approximately by GPD,*

$$\lim_{u \uparrow x^*} \sup_{0<x<x^*-u} |F_u(x) - GP(x)| = 0.$$

Let N_u the number of observations over the threshold u and m is the total number of observations, the empirical estimator for $\bar{F}(u)$ define by $\hat{\zeta}_u$ where $\hat{\zeta}_u = N_u/m$. Replacing x by $y - u$, we can write (7.5) as

$$\bar{F}(y) = \bar{F}(u)\bar{F}_u(y-u), \qquad y > u.$$

By Theorem 7.1.1, we approximate F_u by GPD and replacing $F(u)$ by empirical estimator, we obtain

$$\hat{F}(y) = 1 - \hat{\zeta}_u(1 + \hat{\gamma}\frac{y-u}{\hat{\sigma}})^{-1/\hat{\gamma}}, \qquad y > u. \tag{7.7}$$

let $\{X_n\}$ denote a time series of the maximum of n observations of our quantity of interest. The return level estimate is the level expected to be exceeded by the maximum of n observations with probability $1 - \alpha$ is estimated by $\hat{y}_\alpha$ of $\hat{F}(y)^n$. If $\gamma \neq 0$ and from (7.7), we obtain $\hat{y}_\alpha$ as

$$\hat{y}_\alpha = \frac{\hat{\sigma}}{\hat{\gamma}}[(\frac{1}{\hat{\zeta}_u}(1 - \alpha^{1/n}))^{-\gamma} - 1] + u \tag{7.8}$$

7.1.6 Reproduction numbers

The basic reproduction number $\mathcal{R}_0$ is an important threshold parameter for estimating the effort required to eliminate the contagious diseases, and it is perceived as the expected number of secondary infections produced by one infected individual in a population where all individuals are susceptible to infection.

By using the next generation method introduced in [DHR10], we derive a formula for the basic reproduction number of (7.4).

Then by considering the infectious states E, I_m, I_s and H in (7.4) and substituting the values in the disease-free equilibrium $(N, 0, 0, 0, 0, 0, 0)$ and as per [DHR10], the basic reproduction number is given by

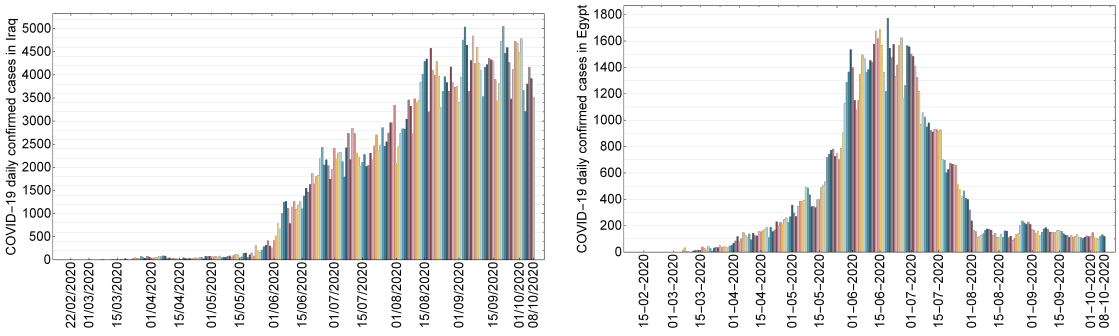
$$\mathcal{R}_0 = \frac{\beta\beta_e}{\nu} + \frac{\theta\beta\beta_m}{(\sigma + \sigma_m)} + \frac{\beta(\sigma + (1-\theta)\sigma_m)}{(\sigma + \sigma_m)(\sigma_s + \delta_s)} + \frac{\beta\beta_h\sigma_s(\sigma + (1-\theta)\sigma_m)}{(\sigma + \sigma_m)(\sigma_s + \delta_s)(\sigma_h + \delta_h)}. \tag{7.9}$$

Besides calculating the basic reproduction number $\mathcal{R}_0$ of the model (7.4), the time dependent reproduction number can be calculated from incidence data (see e.g. [OHB12] for details).

7.2 Results

7.2.1 COVID–19 data from Iraq and Egypt

The data are collected from Worldometer website which is available online [data, datb], we focus on the Iraq data from 22 February until the 08 October 2020 and from 15 February until the 08 October 2020 in Egypt.



(a) Iraq (b) Egypt

Figure 7.2: The daily number of confirmed cases in (a) Iraq from 22 February 2020 to 08 October 2020 and in (b) Egypt from 15 February 2020 to 08 October 2020.

Figure 7.2 shows the daily confirmed cases in Iraq and Egypt from the start of the pandemic in Iraq and Egypt until 08 October 2020, respectively. The statistics of the data are given in Table 7.2.

Table 7.2: Statistics for the COVID–19 data from Iraq and Egypt.

Descriptive statistics	Iraq Total sample size = 231 22 February 2020 Cumulative cases	Egypt Total sample size = 238 15 February 2020 Cumulative cases
Min	1	1
Max	394, 566	104, 156
Median	22, 008	41, 303
Mean	92, 023	48, 004
Standard error	119, 256.6	42, 988.4

7.2.2 Forecast of the COVID–19 spread in Iraq and Egypt

To fit the confirmed cumulative cases from both country Iraq and Egypt starting from the beginning of the outbreak on 22 February and 15 February 2020 until 08 October

2020, respectively, we applied the logistic model (7.1), which used to predict the short-term forecast. Figure 7.3 shows the logistics growth model (7.1) fitted to in (left panel)

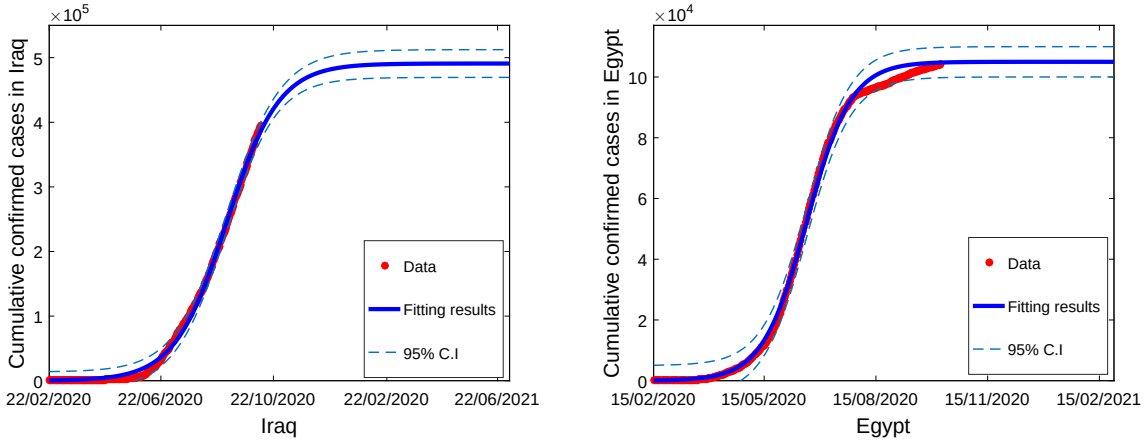


Figure 7.3: The logistic model (7.1) fitted to the cumulative number of infected cases in Iraq (Left panel) and in Egypt (right panel).

the cumulative number of infected cases from Iraq and in (right panel) the cumulative number of infected cases from Egypt with parameters given in Table 7.3.

We note that the logistic model fitted the incidence data with a root mean square error (RMSE) of $5,229.7$, R^2 of 0.9981 for Iraq data and with (RMSE) of $1,924.4$, R^2 of 0.9980 for Egypt data, as shown in Tables 7.3. The logistic model gives a reasonable good fit for both countries.

Table 7.3: Estimated parameter results of the logistics model (7.1) to Iraq and Egypt.

Parameters	**Iraq**		**Egypt**	
	$\mathcal{R} = 1.0659$	$C.I_{0.95}$	$\mathcal{R} = 1.0318$	$C.I_{0.95}$
Estimated epidemic size K (cumulative cases)	$490,900$	$(478300, 503500)$	$105,000$	$(104500, 105900)$
Growth Rate r	0.03787	$(0.03685, 0.03889)$	0.05634	$(0.05546, 0.05721)$
Estimated start of ending phase date	05/05/2021		04/11/2020	
Goodness of fit (R^2)	0.9981		0.9980	
Root Mean Square Error (RMSE)	$5,229.7$		$1,924.4$	

The Gaussian model was fitted to data from Iraq and Egypt with reproduction numbers 1.0659 and 1.0318, respectively. Figure 7.4 shows the Gaussian model fitted to in (left panel) the daily number of confirmed cases from Iraq, and in (right panel) the daily number of confirmed cases from Egypt with parameters given in Table 7.4. The model fits the actual data well with a root mean square error (RMSE) of 335.607, R^2 of 0.9614 for Iraq data and with (RMSE) of 110.33, R^2 of 0.9528 for Egypt data, as listed in Tables 7.4.

The peak of the COVID–19 in Egypt occurs on 16 June 2020 with $1,534$ day cases. Afterword, the daily confirmed cases gradually decreased and the estimated epidemic

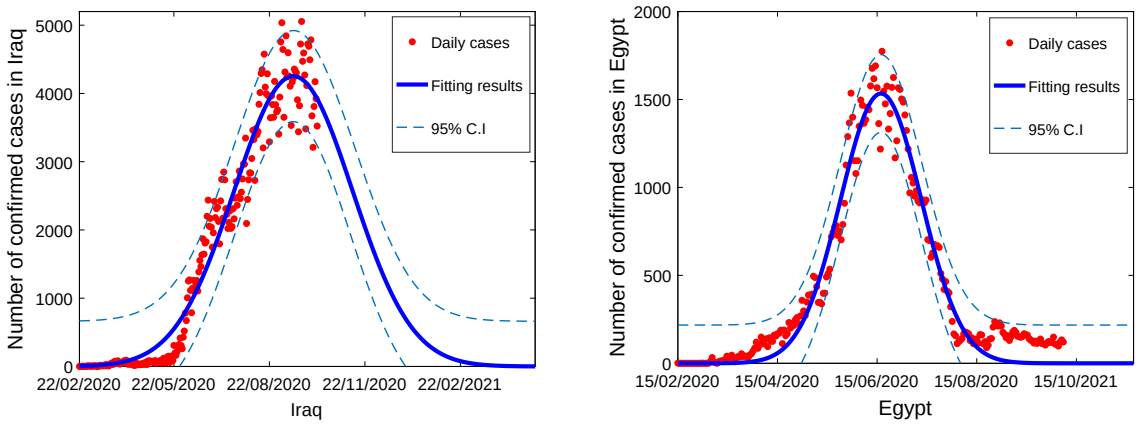


Figure 7.4: The Gaussian model fitted to the daily confirmed cases in Iraq (Left panel) and in Egypt (right panel).

Table 7.4: Estimated parameter results of the Gaussian model to Iraq and Egypt.

Parameters	**Iraq**		**Egypt**	
	$\mathcal{R} = 1.0659$	$C.I_{0.95}$	$\mathcal{R} = 1.0318$	$C.I_{0.95}$
Estimated peak day cases I_0	$4,254$	$(4161, 4347)$	$1,534$	$(1493, 1574)$
σ	80.16	$(74.62, 85.69)$	34.99	$(33.94, 36.04)$
Estimated peak date	14/09/2020		16/06/2020	
Goodness of fit (R^2)	0.9614		0.9528	
Root Mean Square Error (RMSE)	335.607		110.33	

size was $105,000$ on 04 November 2020. In the coming days in Iraq, the forecast curve shows a significant increase in the estimated epidemic size can be noted $490,900$. It also indicates that the epidemic size is rapidly increasing, which shows that the number of infections continues to rise steadily.

The daily confirmed cases are expected to have a significant increase due to ziarat the holy shrines in Iraq from various countries.

To predict the spread of COVID–19 in Iraq, we apply the Gaussian model (7.3) to estimate the value and time of the expected peak. The Logistic model was employed to estimate the growth rate at each time of the expected peak.

A short-term forecasting of the confirmed cases and cumulative predicted from Iraq is presented in Table 7.5.

7.2.3 Parameters estimation for Iraq and Egypt

Using the method described in Subsection 7.1.4, we fitted our model to symptomatically infected cases in Iraq and Egypt. Figure 7.5 shows the model (7.4) fitted to the daily number of confirmed cases in (left panel) from Iraq, 22 February 2020 until 08 October

Table 7.5: Prediction and confirmed cases in Iraq.

Date	Daily cases			Cumulative cases		
	Predicted	Confirmed	Error (%)	Predicted	Confirmed	Error (%)
5-Oct-20	4,011.94	3,808	5.36	376,351.09	382,949	1.72
6-Oct-20	3,987.28	4,172	4.43	379,430.15	387,121	1.99
7-Oct-20	3,961.56	3,923	0.98	382,450.32	391,044	2.20
8-Oct-20	3,934.80	3,522	11.72	385,411.44	394,566	2.32
9-Oct-20	3,907.02	-	-	388,313.42	-	-
10-Oct-20	3,878.25	-	-	391,156.25	-	-
11-Oct-20	3,848.52	-	-	393,940.00	-	-
12-Oct-20	3,817.84	-	-	396,664.80	-	-
13-Oct-20	3,786.24	-	-	399,330.85	-	-
14-Oct-20	3,753.76	-	-	401,938.42	-	-
15-Oct-20	3,720.42	-	-	404,487.83	-	-
16-Oct-20	3,686.24	-	-	406,979.47	-	-
17-Oct-20	3,651.26	-	-	409,413.76	-	-
18-Oct-20	3,615.50	-	-	411,791.19	-	-
19-Oct-20	3,579.00	-	-	414,112.30	-	-
20-Oct-20	3,541.77	-	-	416,377.66	-	-
21-Oct-20	3,503.87	-	-	418,587.89	-	-
22-Oct-20	3,465.30	-	-	420,743.64	-	-
23-Oct-20	3,426.12	-	-	422,845.60	-	-

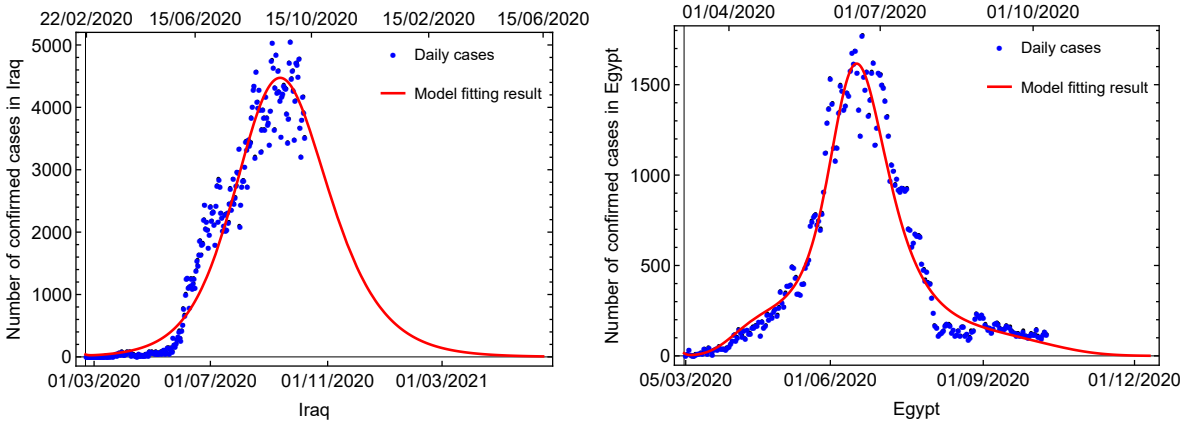


Figure 7.5: The model (7.4) fitted to the daily confirmed cases in (left panel) from Iraq and in (right panel) from Egypt with parameters given in Table 7.6.

2020, and in (right panel) from Egypt, 05 March 2020 until 08 October 2020. Our model gives a reasonable good fit for both countries, predicting the peak in Iraq and showing the peak in Egypt. The fitting parameter results are listed in Table 7.6.

Table 7.6: Parameters and fitted values of model (7.4) in the case of Iraq and Egypt.

Parameters	**Value for Iraq**	**Value for Egypt**	**Source**
	$\mathcal{R}_0 = 1.323$	$\mathcal{R}_0 = 1.11$	
β	0.753	0.56	Fitted
β_e	0.082	0.053	Fitted
β_m	0.475	0.587	Fitted
β_h	0.2057	0.443	Fitted
θ	0.778	0.875	Fitted
σ	0.307	0.104	Fitted
σ_s	0.3247	0.213	Fitted
σ_m	0.239	0.661	Fitted
σ_h	0.446	0.508	Fitted
δ_s	0.127	0.131	Fitted
δ_h	0.298	0.268	Fitted
ν	0.54	0.266	Fitted

7.2.4 Prediction of the second wave of the COVID-19 epidemic

In this section, the analyses were performed for COVD-19 daily cases from 22r February 2020 to 5to February 2021 for Iraq and from 15 February 2020 to 5 February 2021for Egypt.

The application of the return level required choosing an optimal threshold assuming that data exceeding a specified threshold follows a GP distribution to determine an accurate return level estimate. It is very important to choose a plausible threshold value, because choosing a threshold value that is too small leads to an imprecise estimate and choosing a threshold value that is too high leads to a biased estimate. The mean excess plot graphical tool is very helpful for the selection of the threshold u defined by the points $(u, \hat{e}_X(u))$, where $\hat{e}_X(u)$ is empirical mean excess function of (7.6),

$$\hat{e}_X(u) = \frac{\sum_{i=1}^{m}(X_i - u)^+}{\sum_{i=1}^{m} I_{\{X_i > u\}}}.$$

The generalized Pareto distribution (GPD) of two-parameter was used to model exceedances over a threshold, the Maximum likelihood estimators was preferred, the estimated parameters are gamma, sigma of the GPD, where $\gamma = -0.616$ and $\sigma = 686.19$ for Iraq and $\gamma = -0.648$ and $\sigma = 316.796$ for Egypt. Figure 7.6 shows pick the suitable threshold u for infections, which are 4000 and 1300 for COVID-19 data in Iraq and Egypt, respectively, which gave two corresponding observations: 35 and 37 over the threshold. Hence the estimate of the exceedance probability $\hat{\zeta}_u = 0.1003$ for Iraq and $\hat{\zeta}_u = 0.1039$ for Egypt. Moreover, the mean excess plot with a downwards sloping line indicated thin tailed behaviour with $\gamma < 0$.

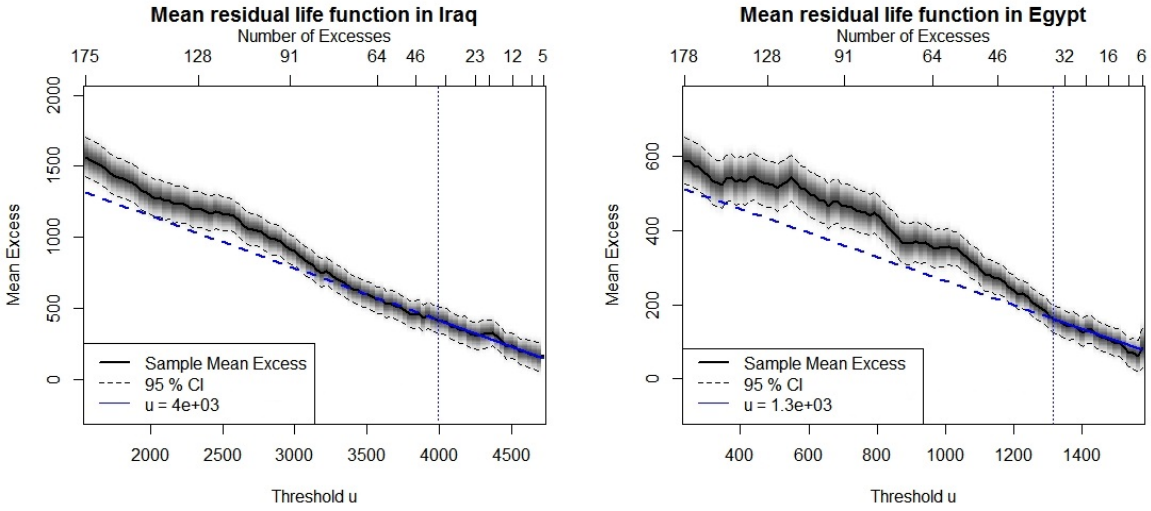


Figure 7.6: Mean excess plot with threshold in Iraq and Egypt,2020.

We focus on estimate the return level during the following year and the following 2 years with two value of probability 0.1 and 0.01. These estimates were computed using Equation (7.8). The results indicate that there is a possibility 0.1 that the infection cases will exceed 5083 once during the next year and 5107 within two years for Iraq, while in Egypt the epidemic will exceed 1788 during the two years with probability 0.01, all results are presented in table 7.7.

Table 7.7: Estimated levels that the maximum of COVED-19 epidemic will exceed with probability 0.1 and 0.01 for the one year and two years for Iraq and Egypt.

Probability	One year		Two year	
$(1-\alpha)$	0.1	0.01	0.1	0.01
Iraq	5083	5107	5094	5109
Egypt	1778	1787	1782	1788

We have listed our results acquired in Subsections 7.2.2–7.2.3 in Table 7.9 to summarize our findings for Iraq and Egypt obtained by using a compartmental mathematical model (7.4), logistic growth model (7.1) and Gaussian model (7.3). Table 7.9 shows a comparison between the estimated parameters obtained by three different models.

7.2.5 Sensitivity analysis and possible control measure

To estimate how easily the virus is spreading, the reproduction number $\mathcal{R}_0$ is estimated from COVID–19 cumulative number of cases using Exponential Growth (EG) method

and Maximum Likelihood method (ML) (see e.g., [OHB12] for details), the results are presented in Table 7.8. The reproduction number in both countries is greater than one and the disease persist

Table 7.8: The reproduction number $\mathcal{R}$ is calculated using cumulative cases

Methods	Iraq		Egypt	
	$\mathcal{R}$	$C.I_{0.95}$	$\mathcal{R}$	$C.I_{0.95}$
EG	1.1017	$(1.10166, 1.101768)$	1.0604	$(1.060386, 1.060481)$
ML	1.0659	$(1.065207, 1.066639)$	1.0318	$(1.030841, 1.032637)$

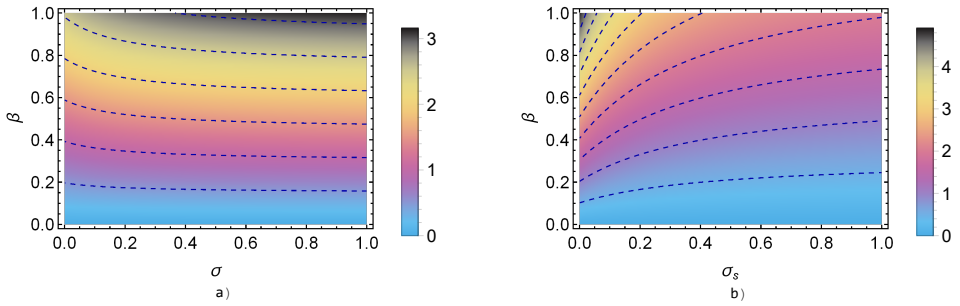
Besides calculating the reproduction number from the incidence data, we derive a formula for the reproduction number from our compartmental model (7.4). Formula (7.9) provides us the basic reproduction number in any time point by substituting the parameter values into it.

Table 7.9: Summery results obtained for Iraq and Egypt in Subsections 7.2.3–7.2.4.

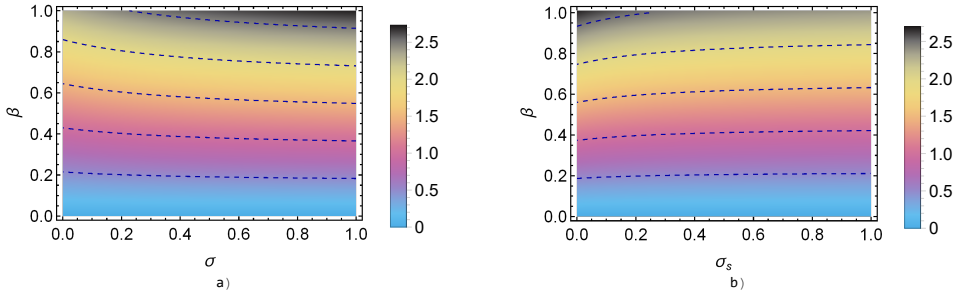
Models/Parameters	**Iraq**	**Egypt**
	Value(C.I)	Value(C.I)
Compartmental model		
$\mathcal{R}_0$	1.323	1.11
β	0.753	0.56
β_e	0.082	0.053
β_m	0.475	0.587
β_h	0.2057	0.443
θ	0.778	0.875
σ	0.307	0.104
σ_s	0.3247	0.213
σ_m	0.239	0.661
σ_h	0.446	0.508
δ_s	0.127	0.131
δ_h	0.298	0.268
ν	0.54	0.266
Logistics Growth Model		
$\mathcal{R}_0$	1.0659	1.0318
Estimated epidemic size	$490,900(478300, 503500)$	$105,000(104500, 105900)$
Estimated start of ending phase date	05/05/2021	04/11/2020
R^2	0.9981	0.9980
RMSE	5,229.7	1,924.4
Gaussian model		
$\mathcal{R}_0$	1.0659	1.0318
Estimated peak day cases	4,254(4161, 4347)	1,534(1493, 1574)
Estimated peak date	14/09/2020	16/06/2020
R^2	0.9614	0.9528
RMSE	335.607	110.33

To assess the dependence of the basic reproduction number on the parameters which can be subject to control the spread of the virus, the contour plot of the basic reproduction number in term of the transmission rate (β), progression rate from mildly infected to symptomatically infected (σ) and progression rate from symptomatically infected to

hospitalized individuals for Iraq and Egypt are shown in Figure 7.7(a) and Figure 7.7(b), respectively. Reducing the transmission rate (β) can decrease the number of severely infected, and consequently the number of infected individuals who need treatment and intensive care at hospitals.



(a) Iraq



(b) Egypt

Figure 7.7: The contour plot of the basic reproduction number for Iraq and Egypt as a function of (β) and in a) progression rate from I_m to I_s (σ) and in b) progression rate from I_s to H (σ_s), respectively.

Figure 7.8 shows a comparison of the PRCC values obtained for the parameters in the basic reproduction number $\mathcal{R}_0$ for Iraq and Egypt. The results of the sensitivity analysis show that any positive change in the parameters $(\beta, \beta_e, \beta_m, \beta_h, \theta, \nu, \sigma, \sigma_h)$ gives a corresponding ratio in the riskiness of the disease, while $\mathcal{R}_0$ can be decreased by increasing the values of the parameters (σ_m, σ_s) in both countries. The parameters (δ_s, δ_h) can not be used as a control measure because they are death rates.

We noticed from the PRCC that the most influential parameter is β, which can be used to control the spread of the COVID–19. Decreasing the transmission rate can decrease the number of infected and even turn the disease to a complete extinction.

In addition to decreasing the transmission rate from severely infected to susceptible (β), decreasing the parameters $\beta_e, \beta_m, \beta_h$ would also decrease the basic reproduction number but this control is not enough to drive $\mathcal{R}_0$ below one. These parameters can be decreased by putting any person has tested positive for COVID–19, and not have

any symptoms yet or have just mild symptoms in quarantine for a sufficient time period (10–14 days). Our model also has its limitation, where we can not estimate the effect of quarantine on the spread of the virus.

It was also observed that by decreasing the progression rate from mildly to severely infected σ also decreases the number of serve infected, but just this measure is unable to drive the disease to extinction.

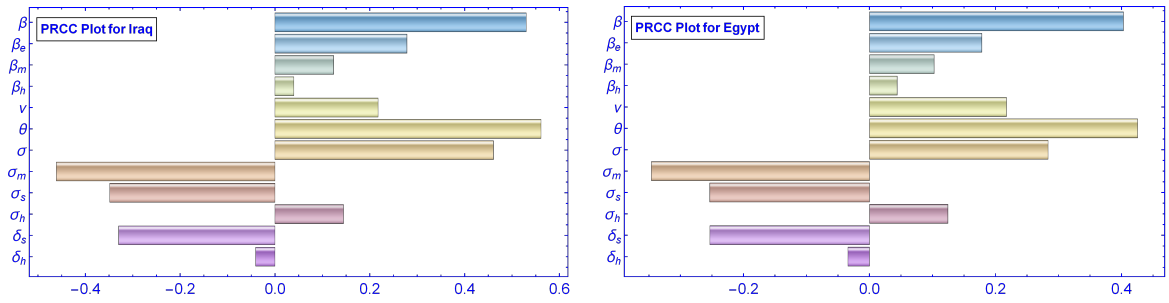


Figure 7.8: The PRCC plot of the parameters of $\mathcal{R}_0$ for Iraq (left panel) and for Egypt (right panel).

8

Summary

In this thesis, we aim to use a class of weighted least squares estimators for the tail index of a distribution function with a regularly varying. Our approach is based on the method for the Parzen tail index developed by Holan and McElroy [HM10]. Analysis of a simulation used to study the performance of weighted least squares estimators. Finally, an approach extreme value theory was applied along with the weighted least squares estimators on real data. This thesis is based on papers [ANV20], [ANSV], [IAND20] and [IAN20].

In chapter 5, we suggest a class of weighted least squares (WLS) estimators for the Parzen tail index. Our approach is built on the method developed by Holan and McElroy. We investigate consistency and asymptotic normality of the WLS estimators and assess the limiting variance 5.8 for $p_0 = 1$, different weight functions and tail indices to compare the WLS and the unweighted (ordinary least squares) estimators in the sub-model of (5.4). Our results show that in some cases the use of the weights makes the asymptotic variance smaller. Simulations are performed and the samples were generated from the model (5.2) with $L_0 \equiv 1$ using different tail indices. We conclude that in the submodel $L_0 \equiv 1$ for α values between 0.8 and 1.5 the WLS estimator has better performance than the OLS estimator. Thus, for thinner tails we propose the WLS estimator instead of the OLS estimator. The Hill estimator is the best among the examined estimators, and the Pickands estimator has also good performance.

In chapter 6, we propose a class of weighted least squares estimators for the classical tail index of a distribution function with a regularly varying upper tail. Asymptotic normality of the estimators is proved. A simulation is performed to compare selected well-known tail index techniques with existing proposals using MSE The samples were generated from the strict Pareto model $L \equiv 1$ and from the Hall model. The Hill, Pickands, DEdH and the weighted least squares (WLS) estimators were included in the

simulation study. We conclude that in the sub-model $L \equiv 1$ for all α values, the WLS estimator performs better than the other estimators investigated. Specifically, we used the parameters $D_1 = 0.4, D_2 = 1$ and $\beta = 0.01$. We notice that the WLS estimator performs better than the other estimators, and the OLS estimator are competitive with the Hill estimator especially for value $p = 3$.

In chapter 7, we have studied the spread of COVID–19 pandemic in Iraq and Egypt using Gaussian model, logistics growth model and compartmental (generalized SEIR) model. The parameters were estimated by using our compartmental model, which used to understand the spread of COVID–19 in both countries. Our model provides a reasonable good fit to the incidences data. We fitted the logistic model to the COVID–19 cumulative number of confirmed cases in Iraq and Egypt. The simulation results can be utilized to decline the at-risk susceptible population by control interventions such as social distancing and lock-down, and/or changing the population behaviour. The Gaussian model was used to obtain statistical predictions for COVID–19 pandemic in Iraq and Egypt, we fitted the Gaussian model to the COVID–19 daily confirmed cases in both countries. A large rise in the predicted epidemic size in Iraq is seen by the forecast curve, and for Egypt data, the model gives a reasonable good fit.

The Gaussian model indicates that the peak value is $4,254$ in Iraq and $1,534$ in Egypt, While the logistic model shows the peak value is $490,900$ and $105,000$ in Iraq and Egypt, respectively. It is vital to emphasize that the lock-down was imposed on 30 July 2020 by the Iraqi government. The basic reproduction number over a period is greater than one, suggesting an exponential growth in the number of cumulative confirmed cases in Iraq, which may indicate that the lock-down regulations are not properly implemented, that might contribute to a rise in the size and spread of the epidemic. Therefore, in order to lift the restriction, the goal of Iraq's health authority is to keep the reproduction number below one.

The return level for the peaks indicates that infection cases are expected to be exceeded 5083 and 5109 once in the following year and following two years with probability 0.1 and 0.01 respectively in Iraq, while in Egypt, will exceed 1778 at least once during the next year with probability 0.1 and 1788 for following two years with probability 0.01.

The reproduction number was estimated based on the confirmed cumulative cases by using Exponential Growth (EG) method and Maximum Likelihood method (ML) and is found 1.0659–1.1017 and 1.0318–1.0604 for Iraq and Egypt, respectively. Using our compartmental model a formula for the basic reproduction number was obtained, which allow us to calculate the value of $\mathcal{R}_0$. With the estimated parameters set obtained from fitting our model to the incidence data in both countries, we found that $\mathcal{R}_0 = 1.323$ and

$\mathcal{R}_0 = 1.11$ for Iraq and Egypt, respectively. The basic reproduction number is greater than one which indicates the virus still persist in both countries.

The sensitivity analysis and the contour plots of the basic reproduction number (see Figure 7.7 and Figure 7.8) suggest that to control the spread of COVID–19 outbreak, both countries should work to decrease the transmission rate enough by educate the population on how to keep away from contracting the disease, raising the population awareness to fight the virus, wearing face mask is necessary in the public places and making more restriction on the traveling between cities which have large number of infected people.

Bibliography

[AKI+20] M. Aadhityaa, K. S. Kasiviswanathan, Idhayachandhiran Ilampooranan, B. Soundharajan, M. Balamurugan, and Jianxun He. A global scale estimate of novel coronavirus (covid-19) cases using extreme value distributions. *medRxiv*, 2020.

[ANSV] A. AL-Najafi, L. Stachó, and L. Viharos. Regression estimators for the tail index. *Available on arXiv: https://arxiv.org/abs/2002.12634*.

[ANV20] A. AL-Najafi and L. Viharos. Weighted least squares estimators for the parzen tail index. *Periodica Mathematica Hungarica.*, 2020.

[Bac11] N. Bacaër. Verhulst and the logistic equation (1838). *In A Short History of Mathematical Population Dynamics*, pages 35–39, 2011.

[Bat20] M. Batista. Estimation of the final size of the second phase of coronavirus epidemic by the logistic model. *MedRxiv*, 2020.

[BD94] S.M. Blower and H. Dowlatabadi. Sensitivity and uncertainty analysis of complex models of disease transmission: an hiv model, as an example. *International Statistical Review/Revue Internationale de Statistique*, pages 229–243, 1994.

[BGT89] N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular variation*, volume 27 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 1989.

[BNR+12] G. Bol, G. Nakhaeizadeh, S.T. Rachev, T. Ridder, and K.H. eds. Vollmer. *Credit risk: measurement, evaluation and management*. Springer Science & Business Media, 2012.

[BT20] G.D. Barmparis and G.P. Tsironis. Estimating the infection horizon of covid-19 in eight countries with a data-driven approach. *Chaos Solitons Fractals*, page 109842, 2020.

[CBTD01] S. Coles, J. Bawa, L. Trenner, and P. Dorazio. *An introduction to statistical modeling of extreme values*, volume 208. London: Springer, 2001.

[CDM85] Sándor Csörgő, Paul Deheuvels, and David Mason. Kernel estimates of the tail index of a distribution. *Ann. Statist.*, 13(3):1050–1077, 1985.

[Che95] C. Cheng. Uniform consistency of generalized kernel estimators of quantile density. *Ann. Statist.*, 23(6):2285–2291, 1995.

[CM10] Gabriela Ciuperca and Cécile Mercadier. Semi-parametric estimation for heavy tailed distributions. *Extremes*, 13(1):55–87, 2010.

[CR78] Miklós Csörgő and Pál Révész. Strong approximations of the quantile process. *Ann. Statist.*, 6(4):882–894, 1978.

[CSN09] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Rev.*, 51(4):661–703, 2009.

[data] Worldometer. available online:. *https://www.worldometers.info/coronavirus/country/iraq/*.

[datb] Worldometer. available online:. *https://www.worldometers.info/coronavirus/country/egypt/*.

[DdH89] Arnold L. M. Dekkers and Laurens de Haan. On the estimation of the extreme-value index and large quantile estimation. *Ann. Statist.*, 17(4):1795–1832, 1989.

[DdH93] Arnold L. M. Dekkers and Laurens de Haan. Optimal choice of sample fraction in extreme-value estimation. *J. Multivariate Anal.*, 47(2):173–195, 1993.

[DdHR00] Holger Drees, Laurens de Haan, and Sidney Resnick. How to make a Hill plot. *Ann. Statist.*, 28(1):254–274, 2000.

[DEdH89] A. L. M. Dekkers, J. H. J. Einmahl, and L. de Haan. A moment estimator for the index of an extreme-value distribution. *Ann. Statist.*, 17(4):1833–1855, 1989.

[DEdHdV16] J. Danielsson, L.M. Ergun, L. de Haan, and C.G. de Vries. Tail index estimation: Quantile driven threshold selection. *Available at SSRN 2717478*, 2016.

[dH81] Laurens de Haan. Estimation of the minimum of a function using order statistics. *J. Amer. Statist. Assoc.*, 76(374):467–469, 1981.

[dHF06] Laurens de Haan and Ana Ferreira. *Extreme value theory*. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2006. An introduction.

[DHM88] Paul Deheuvels, Erich Haeusler, and David M. Mason. Almost sure convergence of the Hill estimator. *Math. Proc. Cambridge Philos. Soc.*, 104(2):371–381, 1988.

[dHR80] L. de Haan and S. I. Resnick. A simple asymptotic estimate for the index of a stable distribution. *J. Roy. Statist. Soc. Ser. B*, 42(1):83–87, 1980.

[DHR10] O. Diekmann, J.A.P. Heesterbeek, and M.G. Roberts. The construction of next-generation matrices for compartmental epidemic models. *Journal of the Royal Society Interface*, 7(47):873–885, 2010.

[Dre95] Holger Drees. Refined Pickands estimators of the extreme value index. *Ann. Statist.*, 23(6):2059–2080, 1995.

[EKM13] P. Embrechts, C. Klüppelberg, and T Mikosch. *Modelling extremal events: for insurance and finance*, volume 33. 2013.

[FT28] R.A. Fisher and L.H.C. Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. *In Mathematical Proceedings of the Cambridge Philosophical Society*, 24(2):180–190, 1928.

[Gab99] X. Gabaix. Zipfś law for cities: an explanation. *The Quarterly journal of economics*, 114(3):739–767, 1999.

[GI11] Xavier Gabaix and Rustam Ibragimov. Rank $-$ 1/2: a simple way to improve the OLS estimation of tail exponents. *J. Bus. Econom. Statist.*, 29(1):24–39, 2011.

[GM01] M. Ivette Gomes and M. João Martins. Generalizations of the Hill estimator-asymptotic versus finite sample behaviour. *J. Statist. Plann. Inference*, 93(1-2):161–180, 2001.

[Gum04] E. J. Gumbel. *Statistics of extremes*. Dover Publications, Inc., Mineola, NY, 2004. Reprint of the 1958 original [Columbia University Press, New York; MR0096342].

[Hal82] Peter Hall. On some simple estimates of an exponent of regular variation. *J. Roy. Statist. Soc. Ser. B*, 44(1):37–42, 1982.

[Hil75] Bruce M. Hill. A simple general approach to inference about the tail of a distribution. *Ann. Statist.*, 3(5):1163–1174, 1975.

[HKKP01] Ronald Huisman, Kees G. Koedijk, Clemens J. M. Kool, and Franz Palm. Tail-index estimates in small samples. *J. Bus. Econom. Statist.*, 19(2):208–216, 2001.

[HM10] Scott H. Holan and Tucker S. McElroy. Tail exponent estimation via broadband log density-quantile regression. *J. Statist. Plann. Inference*, 140(12):3693–3708, 2010.

[IAN20] M.A. Ibrahim and A. Al-Najafi. Modeling, control, and prediction of the spread of covid-19 using compartmental, logistic, and gauss models: A case study in iraq and egypt. *Processes*, 8(11):1400, 2020.

[IAND20] M.A. Ibrahim, A. Al-Najafi, and A. Dénes. Predicting the covid-19 spread using compartmental model and extreme value theory with application to egypt and iraq. *in press, Trends in Biomathematics: Chaos and Control in Epidemics, Ecosystems, and Cells.*, 2020.

[Jen69] A. Jenkinson. Statistics of extremes, estimation of maximum flood. *World Meterological Organisation*, Tech, 98:193–227, 1969.

[Kar33] J. Karamata. Sur un mode de croissance régulière. Théorèmes fondamentaux. *Bull. Soc. Math. France*, 61:55–62, 1933.

[LLR83] M. R. Leadbetter, Georg Lindgren, and Holger Rootzén. *Extremes and related properties of random sequences and processes*. Springer Series in Statistics. Springer-Verlag, New York-Berlin, 1983.

[Mas82] David M. Mason. Laws of large numbers for sums of extreme values. *Ann. Probab.*, 10(3):754–764, 1982.

[MBC79] M. D. McKay, R. J. Beckman, and W. J. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2):239–245, 1979.

[MN16] Tucker McElroy and Chaitra H. Nagaraja. Tail index estimation with a fixed tuning parameter fraction. *J. Statist. Plann. Inference*, 170:27–45, 2016.

[MP07] Tucker McElroy and Dimitris N. Politis. Moment-based tail index estimation. *J. Statist. Plann. Inference*, 137(4):1389–1406, 2007.

[MS98] Mark M. Meerschaert and Hans-Peter Scheffler. A simple robust estimation method for the thickness of heavy tails. *J. Statist. Plann. Inference*, 71(1-2):19–34, 1998.

[New05] M.E. Newman. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351, 2005.

[Nov12] Serguei Y. Novak. *Extreme value methods with applications to finance*, volume 122 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL, 2012.

[OHB12] T. Obadia, R. Haneef, and P.Y. Boëlle. The r0 package: a toolbox to estimate reproduction numbers for epidemic outbreaks. *BMC medical informatics and decision making*, 12(1):1–9, 2012.

[Par79] Emanuel Parzen. Nonparametric statistical data modeling. *J. Amer. Statist. Assoc.*, 74(365):105–131, 1979.

[Par04] Emanuel Parzen. Quantile probability and statistical data modeling. *Statist. Sci.*, 19(4):652–662, 2004.

[Pic75] James Pickands, III. Statistical inference using extreme order statistics. *Ann. Statist.*, 3:119–131, 1975.

[Pol02] Dimitris N. Politis. A new approach on estimation of the tail index. *C. R. Math. Acad. Sci. Paris*, 335(3):279–282, 2002.

[Res07] Sidney I. Resnick. *Heavy-tail phenomena*. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2007. Probabilistic and statistical modeling.

[Res08] Sidney I. Resnick. *Extreme values, regular variation and point processes*. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2008. Reprint of the 1987 original.

[Roj96] Javier Rojo. On tail categorization of probability laws. *J. Amer. Statist. Assoc.*, 91(433):378–384, 1996.

[RS97] Sidney Resnick and Cătălin Stărică. Asymptotic behavior of Hill's estimator for autoregressive data. volume 13, pages 703–721. 1997. Heavy tails and highly volatile phenomena.

[Sch84] Eugene F. Schuster. Classification of probability laws by tail behavior. *J. Amer. Statist. Assoc.*, 79(388):936–939, 1984.

[Seb08] George A. F. Seber. *A matrix handbook for statisticians*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2008.

[Sen76] E. Seneta. Functions of regular variation. *In Regularly Varying Functions*, pages 1–52, 1976.

[Sor06] Didier Sornette. *Critical phenomena in natural sciences*. Springer Series in Synergetics. Springer-Verlag, Berlin, second edition, 2006. Chaos, fractals, selforganization and disorder: concepts and tools.

[SS20] R. Schlickeiser and F. Schlickeiser. A gaussian model for the time development of the sars-cov-2 corona pandemic disease. predictions for germany made on 30 march 2020. *Physics*, 2(2):164–170, 2020.

[SSSK20] J. Schüttler, R. Schlickeiser, F. Schlickeiser, and M. Kröger. Covid-19 predictions using a gauss model, based on data from april 2. *Physics*, 2(2):197–212, 2020.

[TR19] M. Thomas and H. Rootzén. Real-time prediction of severe influenza epidemics using extreme value statistics. *arXiv preprint arXiv:1910.10788.*, pages 754–764, 2019.

[Vih99] László Viharos. Weighted least-squares estimators of tail indices. *Probab. Math. Statist.*, 19(2, Acta Univ. Wratislav. No. 2198):249–265, 1999.