Estimation of Tail Indices of Heavy-Tailed Distributions with Application

Outline of 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

1 Weighted least squares estimators for the Parzen tail index

The results presented in this chapter are based on [ANV20].
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.

1.1 The tail index estimation

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

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

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

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

exists, where J is the score function defined as $J(u) = -(fQ)'(u)$. Assumption (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:

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

$$fQ(u) = (1-u)^{\nu_1} L_1(1-u), \quad u \in (1/2, 1], \tag{3}$$

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

$$L_i(u) = \exp\left\{\theta_{i,0} + 2\sum_{k=1}^{\infty} \theta_{i,k}\cos(2\pi k u)\right\}, \quad i = 0, 1. \tag{4}$$

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

$$L_i(u) = L_i^{(p_i)}(u) = \exp\left\{\theta_{i,0} + 2\sum_{k=1}^{p_i} \theta_{i,k}\cos(2\pi k u)\right\}, \quad i = 0, 1, \tag{5}$$

where p_i is fixed and unknown. In the representation (2) and (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), 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 $\varepsilon(u) = \log\left(\widehat{fQ}(u)/fQ(u)\right)$ 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

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

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

$$fQ(u)Q'(u) = 1, \tag{6}$$

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 $q(u)$:

$$\widehat{q}_n(u) = \frac{d}{du}\int_0^1 Q_n(t)K_n(u,t)d\mu_n(t), \quad u \in (0,1), \tag{7}$$

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, $\int_0^1 tK_n(u,t)d\mu_n(t) = u, u \in [a,b]$.
(K_4) There is a sequence $\delta_n \downarrow 0$ such that $\sup_{u\in[a,b]} \big| \int_{u-\delta_n}^{u+\delta_n} K_n(u,t)d\mu_n(t) - 1\big| \downarrow 0$ 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 (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 show that the limit matrix $M(a,b,R) := \lim_{n\to\infty} n^{-1} X'WX$ exists. 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 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 2. *Assume that the conditions of Theorem 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

$$V = \int_a^b G_R^2(u)du + \int_a^b \int_a^b G_R(u)G_R(v)\left(1 + [(u \wedge v) - uv]\frac{q'(u)q'(v)}{q(u)q(v)}\right)dudv. \quad (8)$$

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

1.2 Comparison of tail index estimators

1.2.1 Asymptotic variances

We evaluate the limiting variance (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 (4):

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

The limiting variances are contained in Table 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.

Table 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

1.2.2 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 (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 (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 2 and 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 $1 - F(x) = x^{-1/\alpha_1}\ell_1(x)$, $0 < x < \infty$ 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 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 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

2 Regression estimators for the tail index

This chapter is based on [ANSV].
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.

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

$$Q(1-s) = s^{-\alpha} L(s), \quad 0 < s < 1. \tag{9}$$

The asymptotic normality of Hill estimator 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

$$Q(1-s) = s^{-\alpha} D_1 [1 + D_2 s^{\beta} \{1 + o(1)\}], \quad s \to 0, \tag{10}$$

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

Following the idea of Holan and McElroy (2010) [HM10], we assume that the slowly varying function L in (9) 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 $p \leq \widetilde{p}$, where $\widetilde{p}$ is a prespecified integer. The knowledge of p is not assumed, condition $p \leq \widetilde{p}$ gives only an upper bound for p. It follows that

$$\log Q(1-s) = -\alpha \log s + \theta_0 + 2 \sum_{k=1}^{p} \theta_k \cos(2\pi k s). \tag{11}$$

Let Q_n be the empirical quantile function defined as

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

Based on the representation (11), 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

$$\varepsilon(s) = \log(Q_n(1-s)/Q(1-s)) \tag{12}$$

is the residual process, $s_j = j/n$, $j = \lceil na \rceil, \ldots, \lfloor nb \rfloor$, $a < b$ 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 $y_j := \log Q_n(1 - s_j)$,

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

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

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

$$\widehat{\alpha}_n^{(W)} := e_1'\widehat{\beta}_{\widetilde{p}} = e_1'(X'WX)^{-1}X'Wy, \tag{13}$$

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 show that the limit matrix $M(a, b, R) := \lim_{n\to\infty} n^{-1}X'WX$ exists. Let $(v^*, v_0, \dots, v_{\widetilde{p}})$ 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.

Theorem 3. *Assume that the conditions $Q_1 - Q_3$ are satisfied for the underlying distribution and suppose that the quantile function Q admits the representation (11). 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 $s_j = j/n$, $j = \lceil na\rceil, \dots, \lfloor nb\rfloor$, and $p \le \widetilde{p}$. Then*

$$\sqrt{n}(\widehat{\alpha}_n^{(W)} - \alpha) \xrightarrow{D} N(0, V), \tag{14}$$

where

$$V = \int_a^b \int_a^b \frac{G_R(s)G_R(t)\big((1-s)\wedge(1-t) - (1-s)(1-t)\big)}{Q(1-s)Q(1-t)fQ(1-s)fQ(1-t)} dsdt. \tag{15}$$

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

The estimation method proposed in previous section 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

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

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

$$\begin{aligned} y &:= (y_0, \dots, y_{n-1})',\\ W &:= \operatorname{diag}(w_{0,n}, \dots, w_{n-1,n}),\end{aligned}$$

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

$$\begin{aligned} G^* &= \big(-\log s_0, \dots, -\log s_{n-1}\big)',\\ G_k &= \big(\varphi_k(s_0), \dots, \varphi_k(s_{n-1})\big)', \quad k = 0, \dots, \widetilde{p}. \end{aligned} \tag{16}$$

Set

$$b_{\widetilde{p}}(s) := \log L(s) - \sum_{k=0}^{\widetilde{p}} \theta_k \varphi_k(s). \tag{17}$$

Recall (12). 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:

$$-\log s \sim \sum_{j=0}^{\infty} c_j \varphi_j(s), \tag{18}$$

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

2.3 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 (9) and from the Hall model (10). 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 4 and 5 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 and 7 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 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 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 4: 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 5: 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: 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 7: 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

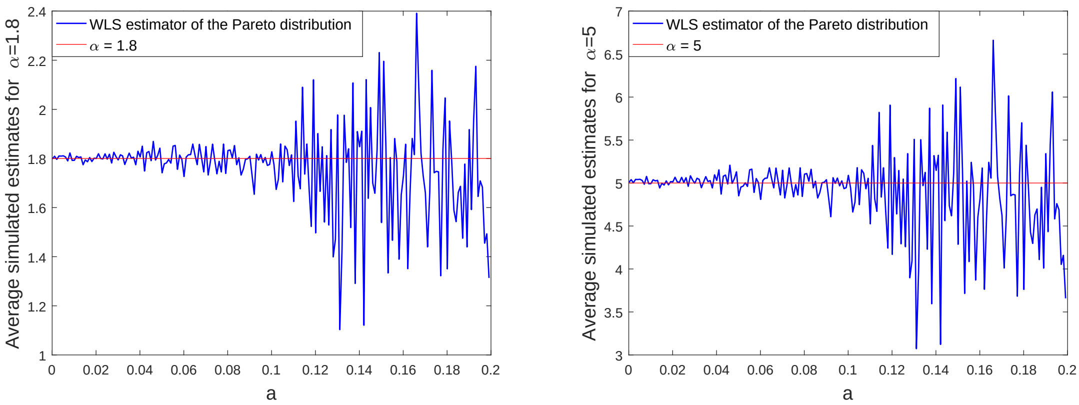

Figure 1: Tail index estimates for WLS approach with Pareto distribution in (left panel) from $\alpha = 1.8$ and in (right panel) from $\alpha = 5$.

3 Application

The results presented in this chapter are based on [IAN20, IAND20].
we study the prevalence of the COVID-19 pandemic in Iraq and Egypt using a generalised (SEIR) compartmental mathematical model, a logistic regression model, and a simple Gaussian model. 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.

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

The logistic growth takes the form:

$$C(t) = \frac{K}{1 + be^{-rt}}, \tag{19}$$

where $r > 0$ is the rate of infection, $K > 0$ is the final epidemic size and $b = \frac{K - C_0}{C_0}$ and C_0 is the initial population. Figure 2 shows the logistics growth model (19) fitted to in (left 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 8. We note that the logistic model fitted the incidence data with a root mean square

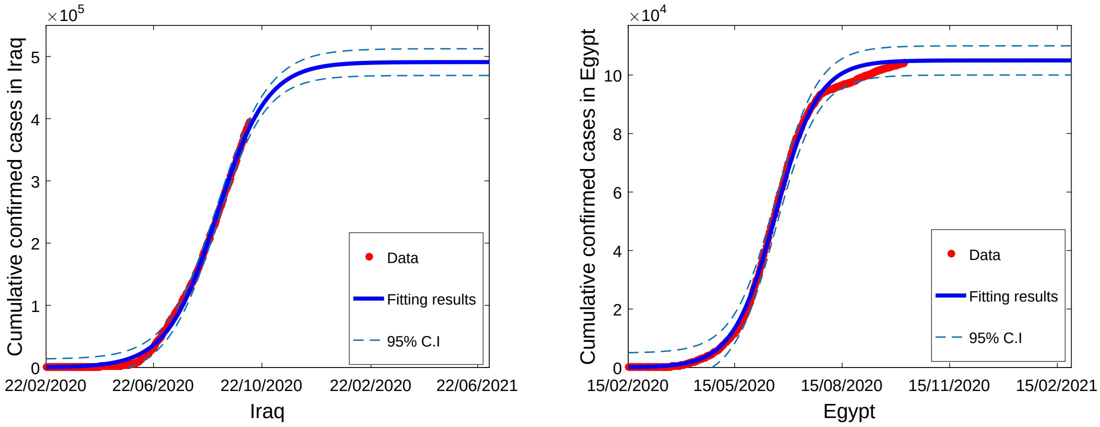

Figure 2: The logistic model (19) fitted to the cumulative number of infected cases in Iraq (Left panel) and in Egypt (right panel).

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 8. The logistic model gives a reasonable good fit for both countries.

Table 8: Estimated parameter results of the logistics 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 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)	$5,229.7$		$1,924.4$	

We employed a simple Gaussian model, to model the time-dependent daily change of infections. Let $I(t)$ denotes the time-dependent Gaussian function and takes the following form:

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

where I_0 denotes the maximum value at time μ and σ controls the width. The Gaussian model was fitted to data from Iraq and Egypt with reproduction numbers 1.0659 and 1.0318, respectively. Figure 3 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 9. 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 9.

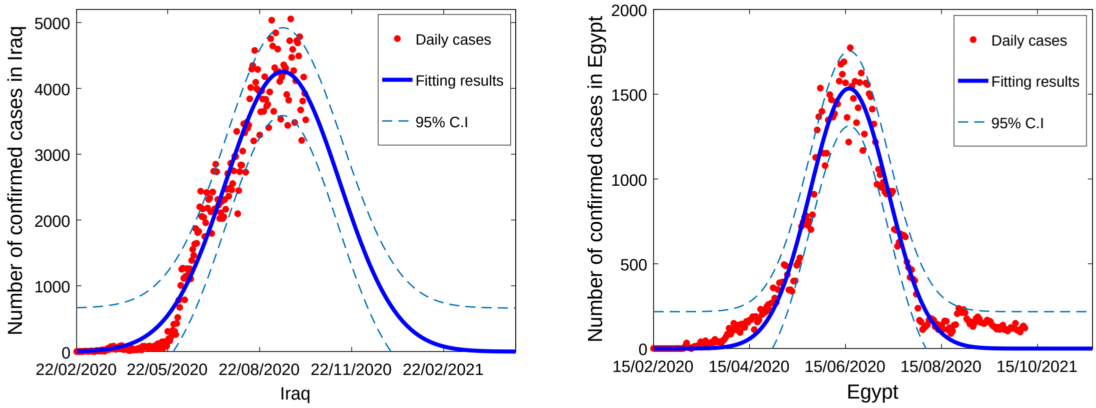

Figure 3: The Gaussian model fitted to the daily confirmed cases in Iraq (Left panel) and in Egypt (right panel).

Table 9: 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	

3.2 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. Hence, we consider the following SEIR model:

$$\begin{aligned}
S'(t) &= -\beta\frac{\beta_e E(t) + \beta_m I_m(t) + I_s(t) + \beta_h H(t)}{N(t) - D(t)} S(t),\\
E'(t) &= \beta\frac{\beta_e E(t) + \beta_m I_m(t) + I_s(t) + \beta_h H(t)}{N(t) - D(t)} S(t) - \nu E(t),\\
I_m'(t) &= \theta\nu E(t) - \sigma_m I_m(t) - \sigma I_m(t),\\
I_s'(t) &= (1-\theta)\nu E(t) + \sigma I_m(t) - \sigma_s I_s(t) - \delta_s I_s(t),\\
H'(t) &= \sigma_s I_s(t) - \sigma_h H(t) - \delta_h H(t),\\
R'(t) &= \sigma_m I_m(t) + \sigma_h H(t),\\
D'(t) &= \delta_s I_s(t) + \delta_h H(t).
\end{aligned} \tag{20}$$

Figure 4 shows the model (20) fitted to the daily number of confirmed cases in (left panel) from Iraq, 22 February 2020 until 08 October 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 10.

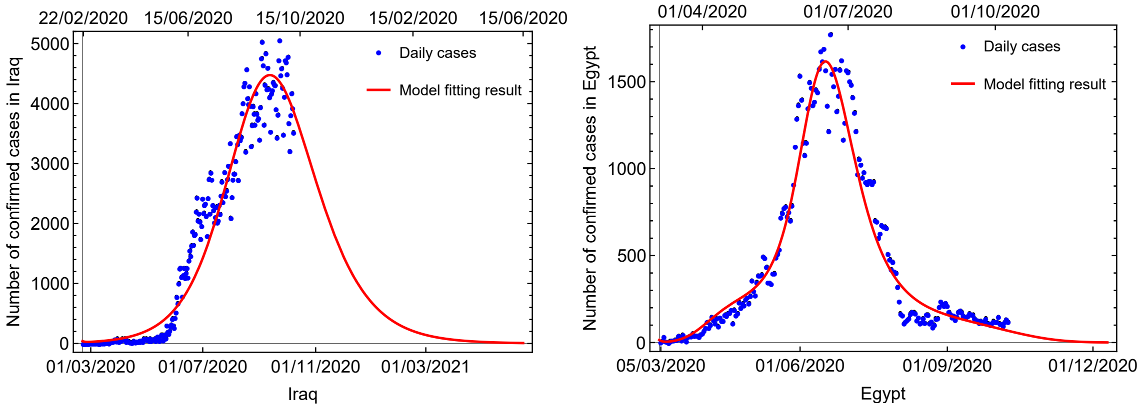

Figure 4: The model (20) fitted to the daily confirmed cases in (left panel) from Iraq and in (right panel) from Egypt with parameters given in Table 10.

Table 10: Parameters and fitted values of model (20) 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

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

We assume that the observations are independent and identically distributed with common cdf F. For $y > u$, $F(y)$ is estimated, by $\hat{F}(u) = 1 - \hat{\zeta}_u(1 - \hat{G}(y - u))$, where $\hat{G}$ is the GPD and $\hat{\zeta}_u$ the empirical estimator of observations that exceed the threshold u. 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 $\gamma \neq 0$, we obtain $\hat{y}_\alpha$ as

$$\hat{y}_\alpha = \frac{\hat{\sigma}}{\hat{\gamma}}[(\frac{1}{\hat{\zeta}_u}(1 - \alpha^{1/n}))^{-\gamma} - 1] + u \tag{21}$$

The mean excess function of X denote the mean residual life function is

$$e(u) = E(X - u \mid X > u), \qquad 0 \leq u < x^*. \tag{22}$$

The generalized Pareto distribution (GPD) of two-parameter was used to model exceedances over a threshold, the Maximum likelihood estimators was preferred, the

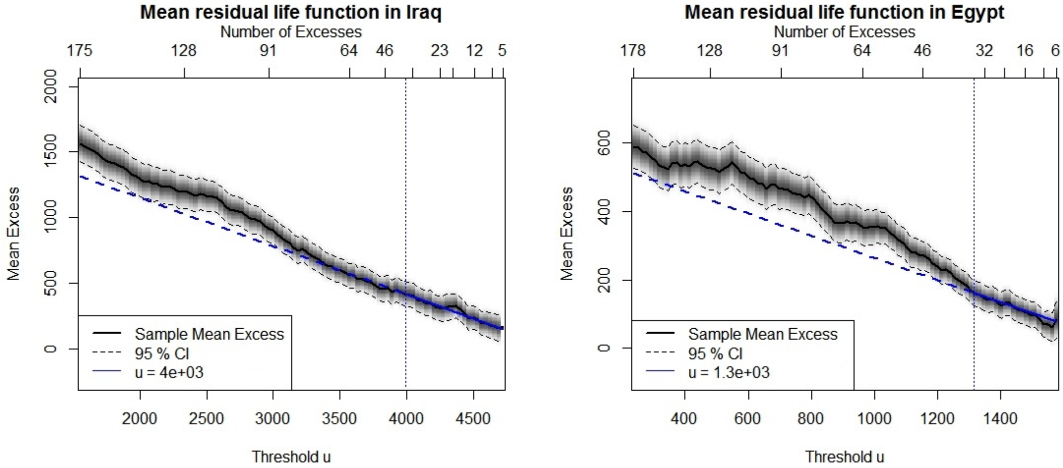

Figure 5: Mean excess plot with threshold in Iraq and Egypt,2020.

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 5 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 $\gamma < 0$. We focus on estimate the return level during the following year and the following two years with two value of probability 0.1 and 0.01. These estimates were computed using Equation (21). 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 11.

Table 11: 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

References

[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.

[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.

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

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

[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.

[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.