Limit laws of weighted power sums of extreme values and Statistical analysis of partition lattices

Abstract of Ph.D. Thesis

by

LILLIAN ACHOLA OLUOCH

Thesis advisors:
Professor Dr. László Viharos,
Professor Dr. László Zádori

Doctoral School of Mathematics and Computer Science, Bolyai Institute, University of Szeged, Faculty of Science and Informatics
Szeged, 2021

1 Asymptotic distributions for weighted power sums of extreme values

This chapter is based on a joint paper [9]. Here we considered proving the asymptotic normality for the weighted power sums over the whole heavy-tail model under some constraints on the weights $d_{i,n}$. The results obtained are crucial in the construction of a new class of estimators for the parameter γ.

1.1 Formulation of the weighted power sums

Let $X, X_1, X_2, \ldots$ be independent random variables with a common distribution function $F(x) = P\{X \leq x\},\ x \in \mathbb{R}$, and for each integer $n \geq 1$ let $X_{1,n} \leq \cdots \leq X_{n,n}$ denote the order statistics pertaining to the sample $X_1, \ldots, X_n$. For a constant $\gamma > 0$, let $\mathcal{R}_\gamma$ be the class of all probability distribution functions F such that

$$1 - F(x) = x^{-1/\gamma} L(x), \quad 0 < x < \infty,$$

where L is a function slowly varying at infinity. Without loss of generality we assume that $F(1-) = 0$ for all $F \in \mathcal{R}_\gamma$. If $Q(\cdot)$ denotes the quantile function of F defined as

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

then $F \in \mathcal{R}_\gamma$ if and only if

$$Q(1-s) = s^{-\gamma} \ell(s), \tag{1}$$

where ℓ is a slowly varying function at 0. Let k_n be a sequence of integers such that

$$1 \leq k_n < n, \quad k_n \to \infty \quad \text{and} \quad k_n/n \to 0 \quad \text{as } n \to \infty. \tag{2}$$

For some constants $d_{i,n}$, $1 \leq i \leq n$, consider the weighted power sums of the extreme values $X_{n-k_n+1,n}, \ldots, X_{n,n}$:

$$S_n(p) := \sum_{i=1}^{k_n} d_{n+1-i,n} \log^p X_{n+1-i,n},$$

where $p > 0$ is a fixed number. Our aim is to study the asymptotic behavior of $S_n(p)$ as $n \to \infty$ whenever $F \in \mathcal{R}_\gamma$.

We will assume as in [13] that the weights $d_{i,n}$ are of the form

$$d_{i,n} = n \int_{(i-1)/n}^{i/n} \bar{L}(t) dt, \quad 1 \leq i \leq n,$$

for some non-negative continuous function $\bar{L}$ defined on (0,1).

1.2 Main results

We state now the main limit theorem of this chapter.

Theorem 1.1. (i) *Assume that* $F \in \mathcal{R}_\gamma$, (2) *holds and suppose that condition* $\bar{\mathbf{L}}$ *is satisfied for the weighs* $d_{i,n}$. *Then*

$$\frac{1}{\sqrt{n}a_n}\left\{\sum_{i=1}^{k_n} d_{n+1-i,n}\log^p X_{n+1-i,n} - \bar{\mu}_n\right\} \xrightarrow{\mathcal{D}} N(0,1). \tag{3}$$

(ii) *If in addition to the conditions of* (i) *we have* $(\log n)/k_n^\varepsilon \to 0$ *for some* $0 < \varepsilon < \rho + 1/2$, *then* (3) *holds with* μ_n *replacing* $\bar{\mu}_n$.

The special case $p = 1$ of Theorem 1.1(i) was stated in Theorem 1.2 of [14]. Several estimators exist for the tail index γ among which Hill's estimator is the most classical (see Hill [6]). Dekkers et al. [4] proposed a moment estimator based on the statistics

$$\frac{1}{k_n}\sum_{i=1}^{k_n}\left(\log\frac{X_{n+1-i,n}}{X_{n-k_n,n}}\right)^j, \quad j = 1, 2. \tag{4}$$

The case $j = 1$ yields the Hill estimator.

The next corollary describes the asymptotic behavior of the weighted norms $R_n(p) := (S_n(p))^{1/p}$.

Corollary 1.2. *Assume the conditions of Theorem* 1.1(ii). *Then*

$$\frac{1}{\gamma\sqrt{2}}\left(\frac{1+2\rho}{1+\rho}\right)^{1/2}\sqrt{k_n}\log\frac{n}{k_n}\left\{\frac{1}{\alpha_n^{1/p}}R_n(p) - \left(\frac{\mu_n}{\alpha_n}\right)^{1/p}\right\} \xrightarrow{\mathcal{D}} N(0,1).$$

By Corollary 1.2,

$$\widehat{\gamma}_n := \frac{1}{\alpha_n^{1/p}}R_n(p)$$

is an asymptotically normal estimator for γ. This is a generalization of the estimator proposed in [15]. Asymptotic normality was proved for the Hill estimator and for the estimators in [1] and [12] under general conditions but not for every distribution in $\mathcal{R}_\gamma$. However, $\widehat{\gamma}_n$ is asymptotically normal over the whole model $\mathcal{R}_\gamma$.

To investigate the asymptotic bias of the estimator $\widehat{\gamma}_n$, we assume the following conditions:

(B_1) $\quad \sqrt{k_n}\log\dfrac{n}{k_n}\sup\limits_{0\le u\le k_n/n}\left|\dfrac{\log\ell(u)}{\log u}\right| \to 0.$

(B_2) $\sqrt{k_n}/\log n \to 0.$

(B_3) $(\log n)/k_n^{\rho+\frac{1}{2}}n \to 0.$

(B_4) $J(s) = s^\rho, \quad 0 < s < 1.$

Corollary 1.3. *Assume the conditions* (B_1)–(B_4), *and the conditions of Theorem* 1.1(i), *and set* $t_n := (\rho+1)\log(n/k_n)$. *Then we have*

(i)

$$\frac{1}{\gamma^p p\sqrt{2}}\left(\frac{1+2\rho}{1+\rho}\right)^{1/2}\sqrt{k_n}\log\frac{n}{k_n}\left\{\frac{S_n(p)}{\alpha_n}-\gamma^p\big(1+pt_n^{-1}\big)\right\}\xrightarrow{\mathcal{D}}N(0,1),$$

(ii)

$$\frac{1}{\gamma\sqrt{2}}\left(\frac{1+2\rho}{1+\rho}\right)^{1/2}\sqrt{k_n}\log\frac{n}{k_n}\big\{\widehat{\gamma}_n-\gamma\big(1+t_n^{-1}\big)\big\}\xrightarrow{\mathcal{D}}N(0,1). \tag{5}$$

1.3 Simulation results

This section evaluates the performance of the estimator $\widehat{\gamma}_n$ through simulations. In the first simulation study, we compare $\widehat{\gamma}_n$ to the Hill, Pickands ([10]) and moment estimators. For the simulation we use the following model proposed by Hall [5]:

$$Q(1-s)=s^{-\gamma}D_1[1+D_2s^{\beta}(1+o(1))]\quad\text{as } s\to 0, \tag{6}$$

where $D_1>0$, $D_2\neq 0$ and $\beta>0$ are constants. The Hall model satisfies condition (B_1) if $D_1=1$ and $k_n^{\beta+\frac{1}{2}}/n^{\beta}\to 0$.

We repeated the simulations 1000 times and we assumed $n=1000$ for the sample size and $k_n=136$ for the sample fraction size. We used $\bar{\ell}\equiv 1$ for the weights $d_{i,n}$. We examined the following two cases of the Hall model:

Case 1: $\beta=2$, $D_2=1$ and $D_1=1/\sqrt{e}$.

Case 2: $\beta=1$, $D_2=4/3$ and $D_1=e^{-2/3}$.

In both cases we assume $o(1)\equiv 0$ in (6). Tables 1 and 2 contain the average simulated estimates (mean) and the calculated empirical mean square errors (MSE) for *Case 1*. Using the mean square error as criterion, we see that for $\rho\leq 1$ the performance of $\widehat{\gamma}_n$ generally increases as γ decreases from 2 to 0.5. For $\gamma\geq 1$ the weights improve the performance of $\widehat{\gamma}_n$ significantly ($\rho=0.5,1,2$). For the thin tail pertaining to $\gamma=0.5$ we also see a trend that the performance of $\widehat{\gamma}_n$ improves as the value of p increases from 1 to 3. The same conclusion holds for $\gamma=1$ when $\rho=2$. It can be also seen that $\widehat{\gamma}_n$ with $p=1,2,3$ and appropriate ρ value performs better than the Pickands and the moment estimator. The Pickands estimator has poor performance for $\gamma=2$. Nonetheless, the Hill and the moment estimator tend to have good estimates.

Table 1: Mean in the Hall model for *Case 1*.

mean							
		$\widehat{\gamma}_n$			Hill	Pickands	Moment
ρ	γ	$p=1$	$p=2$	$p=3$			
0	0.5	0.502461	0.5598067	0.6278012	0.4874154	0.5388793	0.4832535
	1	1.252406	1.347012	1.461455	0.9872326	1.021725	0.9745838
	1.5	2.002351	2.136447	2.299039	1.48705	1.52004	1.471576
	2	2.752296	2.926308	3.137432	1.986867	2.022467	1.969981
0.5	0.5	0.4207121	0.4523482	0.4918764	0.4874154	0.5388793	0.4832535
	1	1.088022	1.138332	1.200928	0.9872326	1.021725	0.9745838
	1.5	1.755332	1.826024	1.913608	1.48705	1.52004	1.471576
	2	2.422641	2.514022	2.626971	1.986867	2.022467	1.969981
1	0.5	0.37965551	0.3994002	0.4240878	0.4874154	0.5388793	0.4832535
	1	1.005246	1.03595	1.073641	0.9872326	1.021725	0.9745838
	1.5	1.630837	1.673773	1.726098	1.48705	1.52004	1.471576
	2	2.256427	2.311814	2.379069	1.986867	2.022467	1.969981
2	0.5	0.33886111	0.3486395	0.3606289	0.4874154	0.5388793	0.4832535
	1	0.9227323	0.9375759	0.9552161	0.9872326	1.021725	0.9745838
	1.5	1.506604	1.527265	1.551595	1.48705	1.52004	1.471576
	2	2.090475	2.117078	2.148269	1.986867	2.022467	1.969981

Table 2: MSE in the Hall model for *Case 1*.

MSE							
		$\widehat{\gamma}_n$			Hill	Pickands	Moment
ρ	γ	$p=1$	$p=2$	$p=3$			
0	0.5	0.008489717	0.004758226	0.01848487	0.001920372	0.1238975	0.008732585
	1	0.06713682	0.124994	0.2205786	0.007254561	0.1510138	0.01456819
	1.5	0.2601122	0.415274	0.6550551	0.01616043	0.191689	0.02365229
	2	0.579775	0.8761246	1.322759	0.02863798	0.2457045	0.0362088
0.5	0.5	0.006915487	0.002963965	0.0009153005	0.001920372	0.1238975	0.008732585
	1	0.01033168	0.02195434	0.04367552	0.007254561	0.1510138	0.01456819
	1.5	0.07105951	0.1126648	0.1784538	0.01616043	0.191689	0.02365229
	2	0.189099	0.2755773	0.4061787	0.02863798	0.2457045	0.0362088
1	0.5	0.01503467	0.01069469	0.006382895	0.001920372	0.1238975	0.008732585
	1	0.002311005	0.003667682	0.007952766	0.007254561	0.15101382	0.01456819
	1.5	0.02231372	0.03559411	0.05684494	0.01616043	0.191689	0.02365229
	2	0.07504283	0.1068695	0.1538997	0.02863798	0.2457045	0.0362088
2	0.5	0.02645074	0.02340072	0.01992387	0.001920372	0.1238975	0.008732585
	1	0.007996087	0.005951954	0.004100054	0.007254561	0.1510138	0.01456819
	1.5	0.004666964	0.005432634	0.007437678	0.01616043	0.191689	0.02365229
	2	0.01646337	0.02210106	0.03052874	0.02863798	0.2457045	0.0362088

By Corollary 1.3(ii) we infer that

$$Z_n := \frac{1}{\widehat{\gamma}_n\sqrt{2}}\left(\frac{1+2\rho}{1+\rho}\right)^{1/2}\sqrt{k_n}\log\frac{n}{k_n}\{\widehat{\gamma}_n - \gamma(1+t_n^{-1})\} \xrightarrow{\mathcal{D}} N(0,1). \qquad (7)$$

Asymptotic confidence intervals for γ can be constructed using either (5) or (7). In the second simulation study we investigated how fast the distribution result (7) kicks in. We simulated the quantity Z_n 5000 times. According to condition (B_2), we used k_n values less than $\log^2 n$. First, we investigated the Fréchet distribution with shape parameter $1/\gamma$ that belongs to the Hall model with parameters $D_1 = 1$, $D_2 = -\gamma/2$ and $\beta = 1$. The simulation was done for $\gamma = 1$, $\rho = 1$, $p = 1$, $n = 900$ and $k_n = 10$. We found empirically that $n = 900$ is the threshold sample size to obtain a good normal approximation in (7). Figure 1 contains the histogram with the fitted normal curve and the Q–Q plot of the simulated Z_n quantities with estimated parameters. The mean of the simulated Z_n values is -0.06, the simulated standard deviation is 0.8974. The mean of the simulated $\widehat{\gamma}_n$ values is 1.1116. The bias of the mean is in accordance with the bias term γt_n^{-1} in (7). Due to the biased estimator in the leading factor $1/(\widehat{\gamma}_n\sqrt{2})$ of Z_n, the simulated standard deviation of Z_n is smaller than the asymptotic value 1. We performed the chi-square test for normality, and we obtained the p-value 0.2965.

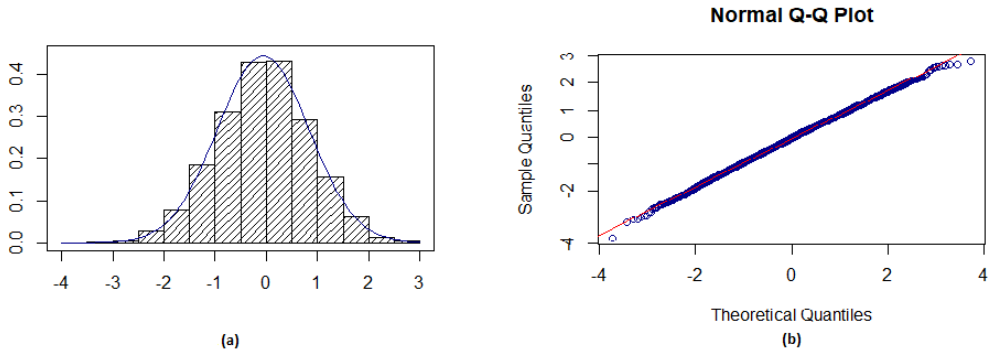

Figure 1: Histogram (a) and Q-Q plot (b) for Fréchet Distribution, $n = 900, kn = 10$.

We investigated two more distributions from the Hall model: Case 1: $\gamma = 1$, $D_1 = 1$ and $D_2 = 1/2$, $\beta = 3/4$; Case 2: $\gamma = 2$, $D_1 = 1$ and $D_2 = 1$, $\beta = 1$. We used $\rho = 3$, $p = 2$, $n = 500$ and $k_n = 7$ for Case 1, and $\rho = 1$, $p = 1$, $n = 900$ and $k_n = 10$ for Case 2. These n values are the threshold sample sizes to obtain a good normal approximation in (7). We obtained the following numerical results. Case 1: mean of the simulated Z_n values: 0.0013, standard deviation of the Z_n values: 0.9127, mean of the simulated $\widehat{\gamma}_n$ values: 1.0667; Case 2: mean of the simulated Z_n values: -0.0393, standard deviation of the Z_n values:

0.8878, mean of the simulated $\widehat{\gamma}_n$ values: 2.2267. The p-value of the chi-square test for normality is 0.323 for Case 1, and 0.6428 for Case 2. Figures 2 contain the histograms with the fitted normal curves and the Q–Q plots of the simulated quantities for Case 2.

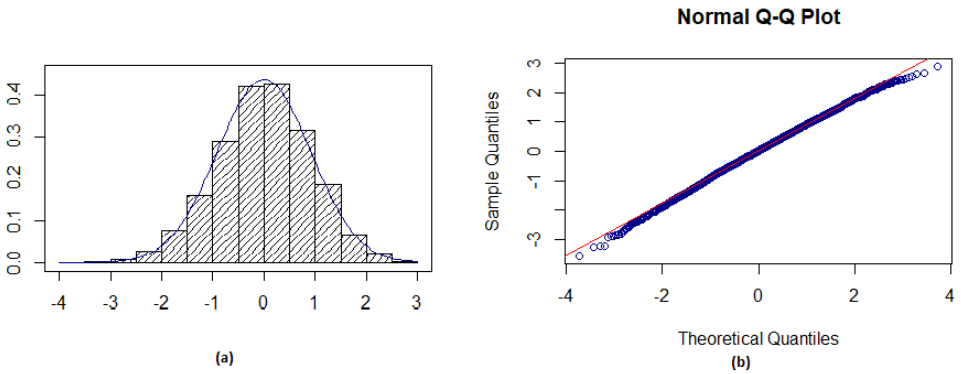

Figure 2: Histogram (a) and Q-Q plot (b) for Hall Model Case1, at $n = 500, kn = 7$.

2 Limit laws for the norms of extremal samples

This chapter is based on the paper [7]. We considered a class of estimator $\hat{\gamma}(n)$ which is an extension of the Hill estimator. We investigated the asymptotic properties of $\hat{\gamma}(n)$ under conditions of regular varying upper tail.

2.1 Introduction

For $p > 0$ introduce the notation

$$S_n(p) = \frac{1}{k_n} \sum_{i=1}^{k_n} \left(\log \frac{X_{n+1-i,n}}{X_{n-k_n,n}} \right)^p . \tag{8}$$

Our main objective is to estimate

$$\widehat{\gamma}(n) = \left(\frac{S_n(p)}{\Gamma(p+1)} \right)^{\frac{1}{p}} \tag{9}$$

of the tail index, where Γ is the usual gamma function. In what follows we always assume that $1 \leq k_n \leq n$ is a sequence of integers such that $k_n \to \infty$ and $k_n/n \to 0$.

As a special case for $p = 1$, we obtain the well-known Hill estimator of the tail index $\gamma > 0$ introduced by Hill in 1975 [6]. For $p = 2$, the estimator was suggested by Dekkers et al. [4]. To the best of our knowledge the possibility $p = p_n \to \infty$ in (9) was not considered before, which is the main focus of our

study. The estimate $\widehat{\gamma}(n)$ can be considered as $p_n \to \infty$ as the limit law for the norm of the extremal sample.

In this study, we investigate the asymptotic properties of $S_n(p_n)$ and $\widehat{\gamma}(n)$ both for $p > 0$ fixed and for $p = p_n \to \infty$. Although the focus of the paper is to obtain asymptotics for large p, in the course we obtain new results for p fixed.

2.2 Results for fixed p

In what follows, $U, U_1, U_2, \ldots$ are iid uniform$(0,1)$ random variables, and $U_{1,n} \le U_{2,n} \le \ldots \le U_{n,n}$ stand for the corresponding order statistics. To ease notation we frequently suppress the dependence on n and simply write $k = k_n$. Define $X = Q(1-U)$, $X_i = Q(1-U_i)$ for $i = 1, 2, \ldots$. According to the well-known quantile representation, $X, X_1, X_2, \ldots$ is an iid sequence with common distribution function F, which implies that S_n in (8) can be written as

$$S_n(p) = \frac{1}{k} \sum_{i=1}^{k} \left(\log \frac{Q(1-U_{i,n})}{Q(1-U_{k+1,n})} \right)^p \quad \text{for each } n \ge 1, \text{ a.s.} \tag{10}$$

First we show strong consistency for $S_n(p)$. Our assumption on the sequence k_n is the same as in Theorem 2.1 in [4]. This is not far from the optimal condition $k_n / \log\log n \to \infty$, which was obtained by Deheuvels et al. [3] for $p = 1$. In what follows any nonspecified limit is meant as $n \to \infty$.

Theorem 2.1. *Assume that* (1) *holds and* $k_n/n \to 0$, $(\log n)^\delta / k_n \to 0$ *for some* $\delta > 0$. *Then* $S_n(p) \to \gamma^p \Gamma(p+1)$ *a.s., that is for* $p > 0$ *fixed the estimator* $\widehat{\gamma}(n)$ *is strongly consistent.*

Weak consistency holds under weaker assumption on k_n. The following result is a special case of Theorem 2.1 in [12], and it follows from representation (10) and from the law of large numbers.

Theorem 2.2. *Assume that* (1) *holds, and the sequence* (k_n) *is such that* $k_n \to \infty$, $k_n/n \to 0$. *Then* $S_n(p) \xrightarrow{\mathbb{P}} \gamma^p \Gamma(p+1)$, *that is for* $p > 0$ *fixed the estimator* $\widehat{\gamma}(n)$ *is weakly consistent.*

Assume that there exist a regularly varying function a and a Borel set $B \subset [0,1]$ of positive measure such that

$$\lim_{v \downarrow 0} \frac{a(v)}{\ell(v)} = 0, \quad \limsup_{v \downarrow 0} \frac{|\ell(uv) - \ell(v)|}{a(v)} < \infty \quad \text{for } u \in B. \tag{11}$$

Theorem 2.3. *Assume that* (1) *holds, and* $k_n \to \infty$, $k_n/n \to 0$. *Then*

$$\sqrt{k_n} \left(S_n(p_n) - m_p(U_{k+1,n}) \right) \xrightarrow{\mathcal{D}} N(0, \sigma_p^2).$$

If ℓ belongs to the de Haan class Π (defined at 0) then condition (11) holds. Therefore, even in the special case $p = 1$, that is, for the Hill estimator, our next result is a generalization of Theorem 3.1 in [4]. The asymptotic normality of various generalizations of the Hill estimator are obtained under second-order regular variation for ℓ. Our conditions in the next result are weaker.

Theorem 2.4. *Assume that* (11) *holds for* ℓ*, and* k_n *is such that* $k_n \to \infty$*,* $k_n/n \to 0$*, and*

$$\sqrt{k_n}\frac{a(k_n/n)}{\ell(k_n/n)} \to 0. \tag{12}$$

Then, with $\sigma_p^2 = \gamma^{2p}(\Gamma(2p+1) - \Gamma^2(p+1))$*,*

$$\frac{\sqrt{k_n}}{\sigma_p}\left(S_n(p_n) - \gamma^p\Gamma(p+1)\right) \xrightarrow{\mathcal{D}} N(0,1),$$

and

$$\frac{p\sqrt{k_n}}{\gamma^{1/p-1}\sigma_p}\left(\widehat{\gamma}(n) - \gamma\right) \xrightarrow{\mathcal{D}} N(0,1).$$

2.3 Asymptotics for large p

Conditioned on $U_{k+1,n}$ the sum $k_nS_n(p_n)$ in (10) is the sum of k_n iid random variables distributed as $Y(U_{k+1,n})$. The limit theorems with *random centering and norming* for $S_n(p_n)$ are obtained.

2.3.1 Weak laws and Gaussian limit

Let us define the parameter ζ as

$$\zeta = \liminf_{n\to\infty} \frac{\log k_n}{p_n}. \tag{13}$$

For $\zeta \le 2$ we need precise assumption on the power sequence, and we assume that

$$k_n \sim e^{\zeta p_n}. \tag{14}$$

Also define the centering and norming functions for $v \in [0,1)$,

$$\widetilde{m}_p(v) = \begin{cases} 0, & \zeta \in (0,1), \\ m_p^1(v), & \zeta = 1, \\ m_p(v), & \zeta \in (1,2), \end{cases} \qquad \widetilde{\sigma}_p(v) = \begin{cases} \sigma_p(v), & \zeta > 2, \\ \sigma_p^1(v), & \zeta = 2. \end{cases} \tag{15}$$

To ease notation put $m_p^1 = m_p^1(0)$, $\sigma_p^1 = \sigma_p^1(0)$, $\widetilde{m}_p = \widetilde{m}_p(0)$, and $\widetilde{\sigma}_p = \widetilde{\sigma}_p(0)$.

Weak consistency holds for $\zeta \ge 1$, while asymptotic normality holds for $\zeta \ge 2$.

Theorem 2.5. *Assume that $k_n \to \infty$, $k_n/n \to 0$, and $p_n \to \infty$. If $\zeta > 1$ in (13) or $\zeta = 1$ in (14) then*

$$\left(\widetilde{m}_{p_n}(U_{k_n+1,n})\right)^{-1} S_n(p_n) \xrightarrow{\mathbb{P}} 1. \tag{16}$$

In both cases $\widehat{\gamma}(n)$ is weakly consistent. Furthermore, if $\zeta > 2$ in (13) or $\zeta = 2$ in (14) then

$$\frac{\sqrt{k_n}}{\widetilde{\sigma}_{p_n}(U_{k+1,n})} \left(S_n(p_n) - \widetilde{m}_{p_n}(U_{k+1,n})\right) \xrightarrow{\mathcal{D}} N(0,1), \tag{17}$$

and

$$\frac{\sqrt{k_n}\widetilde{m}_{p_n}(U_{k+1,n})}{\widetilde{\sigma}_{p_n}(U_{k+1,n})} p_n \left[\left(\frac{S_n(p_n)}{\widetilde{m}_{p_n}(U_{k+1,n})}\right)^{1/p_n} - 1\right] \xrightarrow{\mathcal{D}} N(0,1). \tag{18}$$

Theorem 2.6. *Assume that for the slowly varying function ℓ (11) holds and $\beta_1 > 0$. If $\zeta > 1$ in (13) or $\zeta = 1$ in (14) then*

$$\left(\widetilde{m}_{p_n}\right)^{-1} S_n(p_n) \xrightarrow{\mathbb{P}} 1. \tag{19}$$

If $\zeta > 2$ in (13) or $\zeta = 2$ in (14) then assume additionally that for some $\varepsilon > 0$,

$$\limsup_{n\to\infty} p_n^{-1} \log\left(\sqrt{k_n}\left(\frac{a(k_n/n)}{\ell(k_n/n)}\right)^{(\nu_\beta - \varepsilon)\wedge 1}\right) < \log 2.$$

Then

$$\frac{\sqrt{k_n}}{\widetilde{\sigma}_{p_n}} \left(S_n(p_n) - \widetilde{m}_{p_n}\right) \xrightarrow{\mathcal{D}} N(0,1), \tag{20}$$

and

$$\frac{\sqrt{k_n}\widetilde{m}_{p_n}}{\gamma\widetilde{\sigma}_{p_n}} p_n \left(\widehat{\gamma}(n) - \gamma\right) \xrightarrow{\mathcal{D}} N(0,1). \tag{21}$$

Note that $m_p/\sigma_p \sim 2^{-p}(p\pi)^{1/4}$ as $p \to \infty$.

2.4 Non-Gaussian stable limits

Next, we explore the regime $\zeta < 2$. Here we need the precise asymptotic assumption (14) on the power sequence p_n. We obtain non-Gaussian limits, where the characteristic exponent of the stable law equals ζ, coming from the growth rate of the power sequence p_n. Therefore, in what follows we use the notation $\zeta = \alpha$.

Let Z_α denote a one-sided α-stable random variable with characteristic function

$$\mathbb{E}e^{\mathrm{i}tZ_\alpha} = \begin{cases} \exp\left\{-\Gamma(1-\alpha)|t|^\alpha e^{-\mathrm{i}\frac{\pi\alpha}{2}\mathrm{sgn}u}\right\}, \\ \exp\left\{\mathrm{i}t(1-a) - \frac{\pi}{2}|t|\left(1 + \mathrm{i}\mathrm{sgn}u\,\frac{2}{\pi}\log|t|\right)\right\}, \end{cases}$$

where $a = 0.577\ldots$ stands for the Euler–Mascheroni constant.

Theorem 2.7. *Assume that $k_n \to \infty$, $k_n/n \to 0$, and $p_n \to \infty$ such that (14) holds for some $\zeta = \alpha \in (0,2)$. Then*

$$\frac{k_n}{\eta_{U_{k_n+1,n}}(p_n)^{p_n}} \left(S_n(p_n) - \widetilde{m}_{p_n}(U_{k+1,n})\right) \xrightarrow{\mathcal{D}} Z_\alpha.$$

Moreover, for $\zeta = \alpha \in (0,1)$,

$$p_n \left(\frac{[k_n S_n(p_n)]^{1/p_n}}{\eta_{U_{k_n+1,n}}(p_n)} - 1 \right) \xrightarrow{\mathcal{D}} \log Z_\alpha, \tag{22}$$

in particular,

$$\widehat{\gamma}(n) \xrightarrow{\mathbb{P}} \gamma\, \alpha e^{1-\alpha}. \tag{23}$$

While for $\alpha \in [1,2)$,

$$p_n \frac{k_n \widetilde{m}_{p_n}(U_{k_n+1,n})}{\eta_{U_{k_n+1,n}}(p_n)^{p_n}} \left[\left(\frac{S_n(p_n)}{\widetilde{m}_{p_n}(U_{k_n+1,n})} \right)^{1/p_n} - 1 \right] \xrightarrow{\mathcal{D}} Z_\alpha. \tag{24}$$

Theorem 2.8. *Assume (14) and that (11) holds. Furthermore, $k_n \to \infty$, $k_n/n \to 0$, and*

$$\widetilde{\ell}\,(n^\gamma \ell(k/n)) \sim \widetilde{\ell}\,((n/k)^\gamma \ell(k/n)) \tag{25}$$

and for $\alpha \in [1,2)$ assume that

$$\nu_\beta \beta_1 > \alpha - 1 - \log \alpha = H(\alpha). \tag{26}$$

Then for $\alpha \in (0,2)$,

$$\frac{k_n}{(\alpha\gamma p_n)^{p_n}} \left(S_n(p_n) - \widetilde{m}_{p_n}\right) \xrightarrow{\mathcal{D}} Z_\alpha. \tag{27}$$

For the estimator $\widehat{\gamma}(n)$ if $\alpha \in (0,1)$,

$$\frac{e^{\alpha-1}}{\alpha\gamma} p_n \left[\widehat{\gamma}(n) \left(1 + \frac{\log p_n}{2p_n} \right) - \gamma\alpha e^{1-\alpha} \right] \xrightarrow{\mathcal{D}} \log Z_\alpha - \frac{\log 2\pi}{2}. \tag{28}$$

while for $\alpha \in (1,2)$,

$$\frac{\sqrt{2\pi}}{\gamma} e^{p_n(\alpha-1-\log\alpha)} p_n^{3/2} \left[\widehat{\gamma}(n) - \gamma\right] \xrightarrow{\mathcal{D}} Z_\alpha, \tag{29}$$

and for $\alpha = 1$,

$$\frac{\sqrt{2\pi}}{2\gamma} p_n^{3/2} \left[\widehat{\gamma}(n) \left(1 + \frac{\log 2}{p_n} \right) - \gamma \right] \xrightarrow{\mathcal{D}} Z_1. \tag{30}$$

2.5 Simulation study

The purpose of this simulation study is to show that the use of larger p values sometimes is beneficial in practical situation. Note that for $p = 1$ we obtain the usual Hill estimator. We also see from (21) that the asymptotic variance increases with p. However, in practical situation higher p values turns out to be useful. It is more common that the large values fit to a Pareto-type distribution, while the smaller values behave as a light-tailed distribution. Consider the quantile function

$$Q(1-s) = \begin{cases} s^{-\gamma}, & \text{if } s \le 0.1, \\ \frac{10^{\gamma}}{\log 10} \log s^{-1}, & \text{if } s \ge 0.1, \end{cases} \tag{31}$$

which is a mixture of an exponential quantile and a strict Pareto quantile. The parameter of the exponential is chosen such that Q is continuous. Figures 3 and 4 contain the simulation results for $\gamma = 1$ and $\gamma = 2$. In this simple model we already see the advantage of larger p values. Note that the Hill estimator is very sensitive to the change of k_n for those values where the quantile function changes. Indeed, for $k_n \le 100$ we basically have a sample from a strict Pareto distribution, and for those values the Hill estimator is the best. For $k_n = 200$ we already see the exponential part of the sample, and the Hill estimator changes drastically (for $\gamma = 1$ from 0.98 to 0.76), while for $p = 5$ the change is not as large (from 0.92 to 0.88).

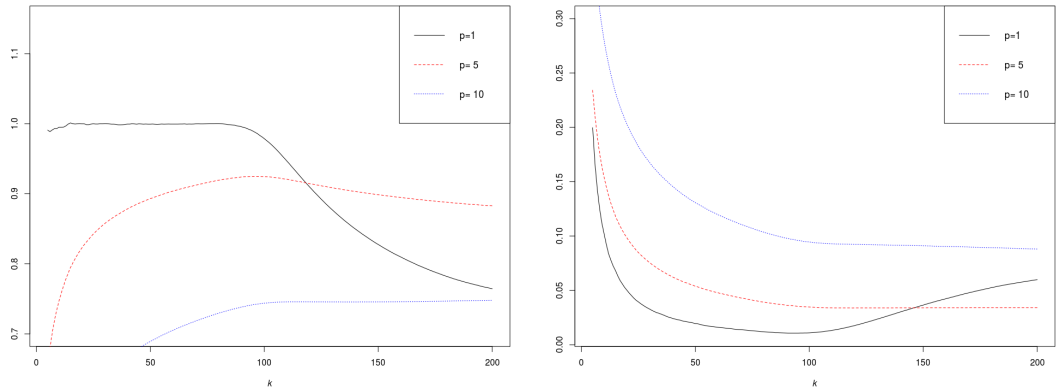

Figure 3: Mean and MSE for a sample with quantile function (31) with $\gamma = 1$.

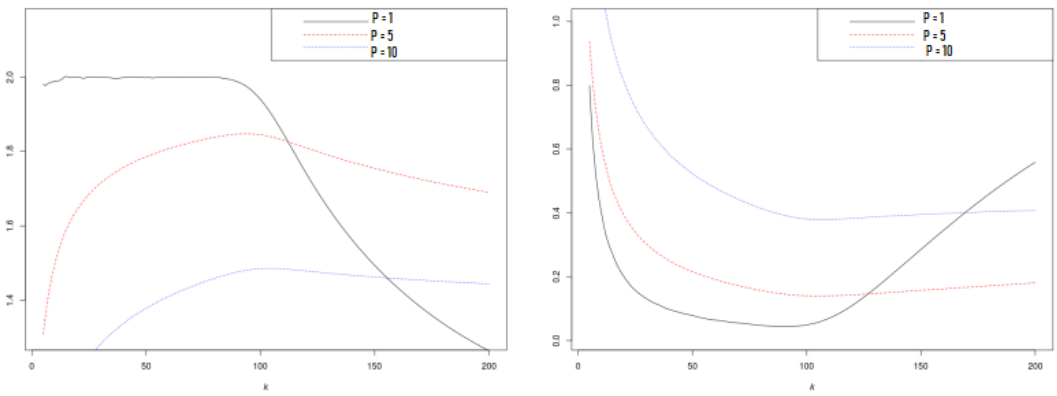

Figure 4: Mean and MSE for a sample with quantile function (31) with $\gamma = 2$.

We also apply the estimator with different p values to real data. We chose the data set of Danish fire insurance losses. In Figure 5 we plotted the estimate for $1/\gamma$, i.e. we plotted $1/\hat{\gamma}(n)$ against k_n, to obtain the Hill plot in [11] for $p = 1$. In our setting larger p values naturally produce smoother plots.

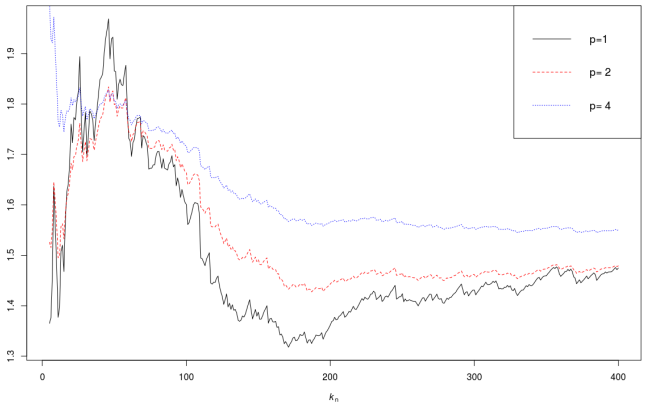

Figure 5: Hill type plots of $\hat{\gamma}(n)^{-1}$ for the Danish fire insurance claim with different p values.

3 A statistical approach to partition lattices with some theoretical "by-products"

3.1 Introduction

This chapter is based on papers [2] and [8] which entails investigating four-element generating sets of a partition lattice and establishing a lower bound for the number of four-element generating sets of direct products of two neighbouring partition lattices.

3.2 Zádori's problem on $(1+1+2)$-generation

We know from Zádori [16] that, for $n \geq 7$, the lattice Part(n) of all partitions of the n-element set 1,...,n has a so-called $(1+1+2)$-*generating set*, that is, a four-element generating set of which two elements (and only two elements) are comparable. The question whether Part(5) and Part(6) have (1+1+2)-generating sets was left open in Zádori [16]. The purpose of this section is to prove the following two statements, which solve Zádori's problem.

Proposition 3.1. *The partition lattice* Part(6) *has a* $(1+1+2)$*-generating set.*

Proposition 3.2. *Every four-element generating set of* Part(5) *is an antichain. Hence,* Part(5) *has no* $(1+1+2)$*-generating set.*

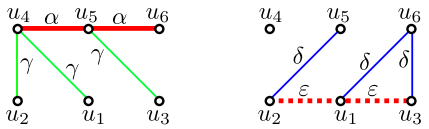

Figure 6: With $\beta := \alpha + \varepsilon$, the set $\{\alpha, \beta, \gamma, \delta\}$ is a $(1+1+2)$-generating set of Equ(6).

Each edge of the graph is colored by one of the colors α, β, γ, δ, and ε. On the vertex set of such a graph A, With $A = u_1, ..., u_6$, Figure 6 defines an *equivalence* (relation) $\alpha \in \mathrm{Equ}(A)$ in the following way: deleting all edges but the α-colored ones, the components of the remaining graph are the blocks of the partition associated with α. In other words, $\langle x, y \rangle \in \alpha$ if and only if there is an α-coloured path from vertex x to vertex y in the graph, that is, a path (of possibly zero length) all of whose edges are α-colored. The equivalences γ and δ are defined analogously while $\beta := \alpha + \varepsilon$.

We have a mathematical proof for Proposition 3.1. However, we used computer programs on the websites [1] that list all four-element generating sets of Part(5); there are exactly 5305 such sets. And we have another program that checks if these 5305 sets are antichains. Note that Proposition 3.1 can also be proved by these programs.

3.3 Computer assisted results and statistical analysis

3.3.1 Estimating confidence intervals

Assume that an experiment has only two possible outcomes: "success" with probability p and "failure" with probability $q := 1 - p$ but none of p and q is known. In order to obtain some information on p, a random *sample* is taken,

[1] http://www.math.u-szeged.hu/~czedli/ and http://www.math.u-szeged.hu/~oluoch/

that is, the experiment is repeated N times independently. Let s denote the number of those experiments that ended up with "success". Then, of course,

$$\text{we estimate } p \text{ by } \widehat{p} := s/N, \tag{32}$$

but we would also like to know how much we can rely on this estimation. Therefore, let $\widehat{q} := 1 - \widehat{p}$, pick a "confidence level" $1 - \alpha_{\text{conf}} \in (0,1) \subset \mathbb{R}$, we let

$$\widehat{\sigma} := \sqrt{\frac{\widehat{p} \cdot \widehat{q}}{N-1}}, \tag{33}$$

and determine the positive real number $z(\alpha_{\text{conf}})$ from the equation

$$1 - \alpha_{\text{conf}} = \int_{-z(\alpha_{\text{conf}})}^{z(\alpha_{\text{conf}})} \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2} dx. \tag{34}$$

Note that the function to be integrated in (34) is the density function of the *standard normal distribution* and the $z(\alpha_{\text{conf}})$ for many typical values of α_{conf} are given in practically all books on statistics. We define the

$$\textit{confidence interval } I(\alpha_{\text{conf}}) \text{ to be } [\widehat{p} - z(\alpha_{\text{conf}})\widehat{\sigma},\, \widehat{p} + z(\alpha_{\text{conf}})\widehat{\sigma}\,]. \tag{35}$$

Let us emphasize that while p is a concrete real number, the confidence interval is *random*, because it depends on a randomly chosen sample. Taking another N-element sample with the same N (that is, repeating the experiments N times again), then (with very high probability in general) a different confidence interval is obtained.

We cannot claim that the confidence interval $I(\alpha_{\text{conf}})$ surely contains the unknown probability p. Furthermore, as it has been pointed by an anonymous referee, it may even happen that $I(\alpha_{\text{conf}})$ contains p only with *very little* probability. For example, if $p = 10^{-100}$ and $N = 2$, then a random N-element sample yields that $\widehat{p} = 0$ and $p \notin I(\alpha_{\text{conf}}) = [0,0]$ with probability $q^2 = 1 - 2 \cdot 10^{-100} + 10^{-200} \approx 1$. However, the Moivre-Laplace theorem, which is a particular case of the central limit theorem, implies that whenever $p \notin \{0,1\}$, then

$$\text{the probability of } p \in I(\alpha_{\text{conf}}) \text{ tends to } 1 - \alpha_{\text{conf}} \text{ as } N \to \infty. \tag{36}$$

3.3.2 Computer programs

Two disjoint sets of computer programs were developed and all data to be reported in (this) Section 3.3 were achieved by these programs. Furthermore, a sufficient amount of these data, including $\nu(4) = 50$ and $\nu(5) = 5\,305$ from Table 3, were achieved independently by different persons (namely, by both authors of [2], with different programs, different computers, and different attitudes

to computer programming. This fact gives us a lot of confidence in our programs and the results obtained by them even if some results that needed too much performance from our computers and programs were achieved only by one of the above-mentioned two settings.

3.3.3 Data obtained by computer programs

The results obtained by computers are given in Table 3 which gives the number $\nu(n)$ of four-element generating sets for $n \in \{4, 5, 6\}$. Clearly, $\nu(7)$ cannot be determined by our programs and computers, although this task might be possible with thousands or millions of similar computers working jointly for a few years or so.

n	4	5	6
$\binom{\mathrm{Bell}(n)}{4}$	1 365	270 725	68 685 050
$\nu(n)$	50	5 305	1 107 900
%, i.e., $100p(n)$	3.663003663	1.959553052	1.613014768
computer time	0.11sec	68 sec	38 hours

Table 3: The (exact) number $\nu(n)$ of the four-element generating sets of $\mathrm{Part}(n)$ for $n \in \{4, 5, 6\}$

Using (35), a fifteen million element sample yielded that, for $n \geq 7$

$$p(7) \in [0.0157753,\ 0.0159877] \text{ with approximate probability } 0.999 \text{ and} \quad (37)$$

$$\nu(7) \in [3.86180 \cdot 10^8,\ 3.91381 \cdot 10^8] \text{ with approximate probability } 0.999. \quad (38)$$

Analogous results based on smaller samples have been obtained for $n = 8$ and $n = 9$.

3.4 Direct products of two neighbouring partition lattices

This section is taken from [8], but we follow closely the approach presented in [2], where Theorem 4.4 states that certain direct products of direct powers of partitions lattices are still 4-generated. In particular, for any integer $5 \leq n$, the direct product $\mathrm{Part}(n) \times \mathrm{Part}(n+1)$ is four-generated. Of course, a much larger lower bound is presented here for large values of n.

From the main result of [2], we could derive the following lemma, which plays an important role in the proof of the theorem that comes thereafter.

Lemma 3.3. *With* $t^* = \binom{n-6}{(n-5)/2}$, *if* $n \geq 7$ *and* n *is odd, then the lattice* $\mathrm{Part}(n)^{t^*} \times \mathrm{Part}(n+1)^{t^*}$ *is 4-generated.*

The exponent t^* given above is not the best (=largest) possible value; simply because the exponent supplied by Theorem 4.4 of [2] is not the best either. Now we state the main result of this section.

Theorem 3.4. *Let* $n \geq 7$ *be an integer number and define*

$$t_n := \begin{cases} \binom{n-6}{(n-5)/2}, & \text{if } n \text{ is odd, and} \\ \\ \min\left((n-2)(n-4)/8, \binom{n-6}{n/2-3}\right), & \text{if } n \text{ is even.} \end{cases} \tag{39}$$

Then $\mathrm{Part}(n) \times \mathrm{Part}(n+1)$ *has at least* $t_n^2 \cdot n! \cdot (n+1)!/2$ *many* 4*-element generating sets.*

References

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

[2] G. Czédli and **Lillian Oluoch**: Four-element generating sets of partition lattices and their direct products. *Acta Sci. Math. (Szeged),* 86:405–448, 2020.

[3] P. Deheuvels, E. Haeusler, and D. M. Mason: Almost sure convergence of the Hill estimator. *Math. Proc. Cambridge Philos. Soc.*, 104(2):371– 381, 1988.

[4] 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.*, 17:1833–1855, 1989.

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

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

[7] P. Kevei, **L. Oluoch**, and L. Viharos: Limit laws for the norms of extremal samples. Submitted to *Journal of Statistical Planning and Inference*, http://arxiv.org/abs/2004.12736 .

[8] **L. Oluoch** and A. Al-Najafi: Lower bound for the number of 4-element generating sets of direct products of two neighboring partition lattices. *Discussiones Mathematicae — General Algebra and Applications*, 2021. Accepted.

[9] **L. Oluoch** and L. Viharos: Asymptotic distributions for weighted power sums of extreme values. *Acta Sci. Math. (Szeged)*, 2021. Accepted.

[10] J. Pickands III.: Statistical Inference Using Extreme Order Statistics. *The Annals of Statistics.*, 3(1):119–131, 1975.

[11] S. Resnick: Discussion of the Danish data on large fire insurance losses. *ASTIN Bulletin: The Journal of the IAA*, 27(1):139–151, 1997.

[12] J. Segers: Residual estimators. *J. Statist. Plann. Inference*, 98(1–2):15–27, 2001.

[13] L. Viharos: Asymptotic distributions of linear combinations of extreme values. *Acta Sci. Math. (Szeged)*, 58(1-4):211–231, 1993.

[14] L. Viharos: Limit theorems for linear combinations of extreme values with applications to inference about the tail of a distribution. *Acta Sci.Math. (Szeged)*, 60(3–4):761– 777, 1995.

[15] L. Viharos: Estimators of the exponent of regular variation with universally normal asymptotic distributions. *Math. Methods Statist.*, 6(3):375–384, 1997.

[16] L. Zádori: *Generation of finite partition lattices.* In: Lectures in universal algebra (Proc. Colloq. Szeged, 1983), Colloq. Math. Soc. János Bolyai ,43, North-Holland, 573–586.