Estimation of the second order parameter for heavy-tailed distributions

Stéphane Girard

Inria Grenoble Rhône-Alpes, France

joint work with El Hadji Deme (Université Gaston-Berger, Sénégal) and Laurent Gardes (Université de Strasbourg, France)
1 Extreme value theory
2 Estimation of the second order parameter
3 Asymptotic properties
4 Link with existing estimators
5 Illustration on simulations
Let $X_1, \ldots, X_n$ be independent copies of a real random variable $X$ with survival function $\bar{F} = 1 - F$. The order statistics associated to this sample are denoted by: $X_{1,n} \leq \cdots \leq X_{n,n}$.

**Fréchet Maximum domain of attraction**

The cumulative distribution function $F$ belongs to the Fréchet maximum domain of attraction if and only if

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

where $\gamma > 0$ is the extreme-value index and $\ell$ is a slowly varying function i.e.

$$\frac{\ell(\lambda x)}{\ell(x)} \to 1 \text{ as } x \to \infty \text{ for all } \lambda \geq 1.$$ 

This condition is equivalent to $\bar{F}$ is regularly varying with index $-1/\gamma$ (heavy-tailed distribution).

The asymptotic distribution of estimators of $\gamma$ is obtained under a second order condition.
Extreme value theory

Second order condition

There exist a function $A(x) \to 0$ and a second order parameter $\rho \leq 0$ such that, for all $\lambda > 0$,

$$
\lim_{x \to \infty} \frac{1}{A(x)} \log \left( \frac{\ell(\lambda x)}{\ell(x)} \right) = K_\rho(\lambda) := \int_1^\lambda u^{\rho-1} du.
$$

- $|A|$ is regularly varying with index $\rho$.
- If $\rho$ is small, the rate of convergence of $\ell(\lambda x)/\ell(x)$ to one is high (and conversely).
- $\rho$ controls the bias of the estimators of $\gamma$.
- $\rho$ is of primordial importance in the adaptative choice of $k$ which is the number of upper order statistics $X_{n-k,+1,n} \leq \cdots \leq X_{n,n}$ used in the estimation of $\gamma$.
- A third order condition is needed to deal with the asymptotic distribution of $\rho$ estimators.
Extreme value theory

Third order condition

There exist functions $A(x) \to 0$ and $B(x) \to 0$, a second order parameter $\rho < 0$ and a third order parameter $\beta < 0$ such that, for every $\lambda > 0$,

$$\lim_{x \to \infty} \frac{(\log \ell(\lambda x) - \log \ell(x)) / A(x) - K_\rho(\lambda)}{B(x)} = L_{(\rho,\beta)}(\lambda)$$

with

$$L_{(\rho,\beta)}(\lambda) = \int_1^\lambda s^{\rho-1} \int_1^s u^{\beta-1} \, du \, ds,$$

and where the functions $|A|$ and $|B|$ are regularly varying with index $\rho$ and $\beta$ respectively.

Contributions

- A new class of estimators for the second order parameter $\rho$,
- Asymptotic properties,
- Links with existing estimators,
- New estimators.
Definition of the family of estimators for the second order parameter

Model

The two main ingredients of our approach are

- a random vector \( T_n = T_n(X_1, \ldots, X_n) \in \mathbb{R}^d \)
- a function \( \psi : \mathbb{R}^d \rightarrow \mathbb{R} \)

verifying the following assumptions:

- There exist a random variable \( \omega_n \) such that

\[
\omega_n^{-1}(T_n - I) \xrightarrow{\mathbb{P}} f(\rho),
\]

where \( I = \begin{pmatrix} 1 & \ldots & 1 \end{pmatrix} \in \mathbb{R}^d \).

- **Invariance properties**

\[
\psi(x + \lambda I) = \psi(x) \quad \text{and} \quad \psi(\lambda x) = \psi(x)
\]

for all \( x \in \mathbb{R}^d \) and \( \lambda \in \mathbb{R} \setminus \{0\} \).
Definition of the family of estimators for the second order parameter

Idea

- Invariance (and regularity) properties entail

\[ \psi(T_n) = \psi(\omega_n^{-1}(T_n - I)) \xrightarrow{P} \psi(f(\rho)) \]

- Letting \( Z_n := \psi(T_n) \) and \( \varphi := \psi \circ f \), one obtains \( Z_n \xrightarrow{P} \varphi(\rho) \).

- Suppose there exist \( J_0 \subseteq \mathbb{R}^- \) and \( J \subseteq \mathbb{R} \) such that \( \varphi \) is a bijection \( J_0 \rightarrow J \).

Definition

The family of estimators of the second order parameter is thus defined by:

\[ \hat{\rho}_n = \begin{cases} \varphi^{-1}(Z_n) & \text{if } Z_n \in J, \\ 0 & \text{otherwise.} \end{cases} \]
Asymptotic properties

**Theorem**

Under the invariance (and regularity) conditions,

- If \( \omega_n^{-1}(T_n - I) \xrightarrow{p} f(\rho) \) then \( \hat{\rho}_n \xrightarrow{p} \rho \).

- If, moreover, \( \nu_n(\omega_n^{-1}(T_n - I) - f(\rho)) \xrightarrow{d} \mathcal{N}(m(\rho), \gamma^2 \Sigma) \) where \( \nu_n \to \infty \), \( m \in \mathbb{R}^d \) and \( \Sigma \) is a regular \( d \times d \) matrix then

\[
\nu_n(\hat{\rho}_n - \rho) \xrightarrow{d} \mathcal{N}
\left(
\begin{pmatrix}
t m \nabla \psi(f(\rho)) \\
\varphi'(\rho)
\end{pmatrix},
\gamma^2 \begin{pmatrix}
t \nabla \psi(f(\rho)) \Sigma \nabla \psi(f(\rho)) \\
(var'(\rho))^2
\end{pmatrix}
\right).
\]
In the literature, at least two ways of estimating the second order parameter can be found:

- **Estimators based on rescaled log-spacings** $j(\log X_{n-j+1,n} - \log X_{n-j,n})$, $j = 1, \ldots, k$
  Hall & Welsh (Annals of Statistics, 1985),
  Goegebeur et al. (JSPI, 2010),
  De Wet et al. (SPL, 2012), ...

- **Estimators based on log-excesses**, $(\log X_{n-j+1,n} - \log X_{n-k,n})$, $j = 1, \ldots, k$
  Gomes et al. (Extremes, 2002),
  Fraga-Alves et al. (Portugaliae Mathematica, 2003),
  Ciuperca & Mercadier (Extremes, 2010), ...
1. Estimators based on rescaled log-spacings: \( j(\log X_{n-j+1} - \log X_{n-j}) \)

\[
R_k(\tau) = \frac{1}{k} \sum_{j=1}^{k} H_{\tau} \left( \frac{j}{k+1} \right) j(\log X_{n-j+1,n} - \log X_{n-j,n}),
\]

- \( H_{\tau} \) is a kernel function indexed by a parameter \( \tau > 0 \).
- This statistics is used for instance by Beirlant et al. (Extremes, 1999) to estimate the extreme-value index \( \gamma \) and by Hall & Welsh (Annals of Statistics, 1985), Goegebeur et al. (JSPI, 2010), De Wet et al. (SPL, 2012) to estimate the second order parameter \( \rho \).
- They proved asymptotic normality of these estimators under a technical condition on the kernel, denoted by \( (C1) \) hereafter.
## Extreme value theory

### Estimation of the second order parameter

#### Asymptotic properties

#### Link with existing estimators

#### Illustration on simulations

### Links with existing estimators based on rescaled log-spacings

**Statistics $T_n$**

Suppose the **third order condition** and **(C1)** hold. If the sequence $k$ satisfies

\[ k \to \infty, \; n/k \to \infty, \; k^{1/2}A(n/k) \to \infty, \]

\[ k^{1/2}A^2(n/k) \to \lambda_A \text{ and } k^{1/2}A(n/k)B(n/k) \to \lambda_B, \]

then the random vector

\[ T_n := \left( (R_k(\tau_i)/\gamma)^{\theta_i}, \; i = 1, \ldots, d \right), \]

properly normalised in asymptotically Gaussian. More precisely,

\[ \omega_n = A(n/k)/\gamma(1 + o_P(1)), \; \nu_n = k^{1/2}A(n/k) \text{ and} \]

\[ f(\rho) = \left( \theta_i \int_0^1 H_{\tau_i}(u)u^{-\rho} \, du, \; i = 1, \ldots, d \right). \]
Link with existing estimators based on rescaled log-spacings

Statistics $T_n$
- Let $d = 8$, $T_n$ depends on 16 parameters $\theta_1, \ldots, \theta_8, \tau_1, \ldots, \tau_8$.
- Suppose $\theta_1 = \theta_2$, $\theta_3 = \theta_4$, $\theta_5 = \theta_6$ and $\theta_7 = \theta_8$.

Function $\psi$

The chosen function $\psi$ is given by:

$$
\psi(x_1, \ldots, x_8) = \frac{x_1 - x_2}{x_3 - x_4} \left( \frac{x_7 - x_8}{x_5 - x_6} \right)^{(\theta_1 - \theta_3)/(\theta_5 - \theta_7)}
$$

and thus

$$
Z_n = \frac{R_{\theta_1}^k(\tau_1) - R_{\theta_1}^k(\tau_2)}{R_{\theta_3}^k(\tau_3) - R_{\theta_3}^k(\tau_4)} \left( \frac{R_{\theta_7}^k(\tau_7) - R_{\theta_7}^k(\tau_8)}{R_{\theta_5}^k(\tau_5) - R_{\theta_5}^k(\tau_6)} \right)^{(\theta_1 - \theta_3)/(\theta_5 - \theta_7)}
$$

Asymptotic normality

In this situation, the estimator of $\rho$ is asymptotically Gaussian.

- The estimator still depends on 12 parameters.
- The estimator is not necessarily explicit, the inverse of $\varphi$ has to be computed numerically.
Examples

- Let \( H_\tau(u) = (\tau + 1)u^\tau \).
- To simplify, we assume that \( \tau_2 = \tau_3, \tau_4 = \tau_8 \) and \( \tau_6 = \tau_7 \). There are 9 remaining parameters and \( \varphi \) is given in this case by:

\[
\varphi(\rho) = \text{cste} \left[ \frac{\tau_4 - \rho}{\tau_1 - \rho} \right] \left[ \frac{\tau_5 - \rho}{\tau_4 - \rho} \right]^{(\theta_1 - \theta_3)/(\theta_5 - \theta_7)}
\]

Three explicit estimators can be derived:

- \( \theta_1 - \theta_3 = \theta_5 - \theta_7 \), Googebeur et al. (JSPI, 2010), 8 free parameters,
- \( \theta_1 = \theta_3 \), new estimator, 8 free parameters,
- \( \hat{\rho} = \frac{\tau_1 Z_n - \text{cste} \tau_4}{Z_n - \text{cste}} \)
- \( \tau_1 = \tau_5 \), new estimator, 8 free parameters,
- \( \hat{\rho} = \frac{\tau_4 Z_{n}^{1/(\delta-1)} - \text{cste}^{1/(\delta-1)} \tau_1}{Z_{n}^{1/(\delta-1)} - \text{cste}^{1/(\delta-1)}} \)
Link with existing estimators based on log-excesses

2. Estimators based on log-excesses: \( (\log X_{n-j+1,n} - \log X_{n-k,n}) \)

\[
S_k(\tau, \alpha) = \frac{1}{k} \sum_{j=1}^{k} G_{\tau,\alpha} \left( \frac{j}{k+1} \right) (\log X_{n-j+1,n} - \log X_{n-k,n})^\alpha, \ \alpha > 0,
\]

- \( G_{\tau,\alpha} \) is a positive function.
- This statistics is used for instance by Dekkers et al. (Annals of statistics, 1989), Gomes & Martins (JSPI, 2001), Segers (JSPI, 2001) to estimate the extreme-value index \( \gamma \) and by Hall & Welsh (Annals of Statistics, 1985), Peng (SPL, 1998), Fraga et al. (MMS, 2003), Ciuperca & Mercadier (Extremes, 2010), to estimate the second order parameter \( \rho \).
- They proved the asymptotic normality under a technical condition on the function \( G_{\tau,\alpha} \), denoted by (C2) hereafter.
Statistics $T_n$

Suppose the third order condition and (C2) hold. If the sequence $k$ satisfies

$$k \to \infty, \ n/k \to \infty, \ k^{1/2} A(n/k) \to \infty,$$

$$k^{1/2} A^2(n/k) \to \lambda_A \quad \text{and} \quad k^{1/2} A(n/k) B(n/k) \to \lambda_B,$$

then the random vector

$$T_n = \left( \left( \frac{S_k(\tau_i, \alpha_i)}{\gamma^{\alpha_i}} \right)^{\theta_i}, \ i = 1, \ldots, d \right)$$

properly normalised in asymptotically Gaussian. More precisely, $\omega_n = A(n/k)/\gamma(1 + o_P(1))$, $v_n = k^{1/2} A(n/k)$ and

$$f(\rho) = \left( -\theta_i \alpha_i \int_0^1 G_{\tau_i, \alpha_i}(u)(\log(1/u))^{\alpha_i-1} K_{-\rho}(u)du, \ i = 1, \ldots, d \right),$$
## Link with existing estimators based on log-excesses

### Statistics $T_n$
- Let $d = 8$, $T_n$ depends on 24 parameters $\theta_1, \ldots, \theta_8, \tau_1, \ldots, \tau_8, \alpha_1, \ldots, \alpha_8$.
- Suppose $\theta_1 \alpha_1 = \theta_2 \alpha_2, \theta_3 \alpha_3 = \theta_4 \alpha_4, \theta_5 \alpha_5 = \theta_6 \alpha_6$ and $\theta_7 \alpha_7 = \theta_8 \alpha_8$.

### Function $\psi$

The chosen function $\psi$ is given by:

$$\psi(x_1, \ldots, x_8) = \frac{x_1 - x_2}{x_3 - x_4} \left( \frac{x_7 - x_8}{x_5 - x_6} \right)^{\left(\frac{\theta_1 \alpha_1 - \theta_3 \alpha_3}{\theta_5 \alpha_5 - \theta_7 \alpha_7}\right)}$$

and thus

$$Z_n = \frac{S_k^{\theta_1}(\tau_1, \alpha_1) - S_k^{\theta_2}(\tau_2, \alpha_2)}{S_k^{\theta_3}(\tau_3, \alpha_3) - S_k^{\theta_4}(\tau_4, \alpha_4)} \left( \frac{S_k^{\theta_7}(\tau_7, \alpha_7) - S_k^{\theta_8}(\tau_8, \alpha_8)}{S_k^{\theta_5}(\tau_5, \alpha_5) - S_k^{\theta_6}(\tau_6, \alpha_6)} \right)^{\left(\frac{\theta_1 \alpha_1 - \theta_3 \alpha_3}{\theta_5 \alpha_5 - \theta_7 \alpha_7}\right)}$$

### Asymptotic normality

In this situation, the estimator of $\rho$ is asymptotically Gaussian.

- The estimator still depends on 20 parameters.
- The estimator is not necessarily explicit, the inverse of $\varphi$ has to be computed numerically.
Link with existing estimators based on log-excesses

### Examples

- Let \( G_{\tau,\alpha}(u) = (1 - u^\tau) / \int_0^1 (1 - x^\tau)(-\log x)^\alpha \, dx \),

  - To simplify, we assume that \( \tau_2 = \tau_3 = \tau_5 = \tau_6 = \tau_7 = \tau_8 = \alpha_7 = 1 \), \( \alpha_6 = 3 \) and \( \alpha_8 = 2 \). There are 11 remaining parameters.

**Three explicit estimators can be recovered:**

- \( \theta_1 = \theta_3, \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 1, \tau_1 = 2 \) and \( \tau_4 = 3 \),
  
  *Ciuperca & Mercadier* (Extremes, 2010), no free parameter

- \( \theta_1 = \theta_3, \alpha_1 = \alpha_3 = \alpha_4 = 1 \) and \( \tau_1 = \tau_4 = \alpha_2 = 2 \),
  
  *Ciuperca & Mercadier* (Extremes, 2010), no free parameter

- \( \theta_1 = \theta_3 = \theta_6 = \theta_8 = \alpha_2 = \alpha_4 = \alpha_5 = \tau_1 = \tau_4 = 1, \alpha_3 = \theta_4 = \theta_7 = 2, \tau_5 = 3 \) and \( \alpha_1 = 4 \),
  
  *Gomes et al.* (Extremes, 2002), no free parameter
Examples

Three new estimators can be built:

- \( \theta_1 - \theta_2 = 2\theta_5 - \theta_7, \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 1, \tau_1 = \alpha_5 = 2 \) and \( \tau_4 = 3 \),
  3 free parameters
  \[
  \hat{\rho} = \frac{6Z_n + 4 \text{cste}}{3Z_n + 4 \text{cste}}
  \]

- \( \theta_1 - \theta_2 = 2\theta_5 - \theta_7, \alpha_1 = \alpha_3 = \alpha_4 = 1 \) and \( \tau_1 = \tau_4 = \alpha_2 = \alpha_5 = 2 \),
  3 free parameters
  \[
  \hat{\rho} = \frac{6Z_n - 4 \text{cste}}{2Z_n - 1 \text{cste}}
  \]

- \( \tau_1 = \tau_4 = \alpha_1 = 1, \alpha_2 = \alpha_3 = \alpha_5 = 2 \) and \( \alpha_4 = 3 \),
  4 free parameters
  \[
  \hat{\rho} = \frac{3Z_n^{1/((\delta+1))} - \text{cste}^{1/((\delta+1))}}{Z_n^{1/((\delta+1))} - \text{cste}^{1/((\delta+1))}}
  \]

If \( \delta = 0 \), we find back the estimator introduced in Fraga-Alves et al. (Portugaliae Mathematica, 2003)
Illustration on simulations

Estimators based on rescaled log-spacings

- \( H_{\tau_i}(u) = (\tau_i + 1)u^{\tau_i}, \ i = 1, \ldots, 8 \)
- \( \tau_1, \ldots, \tau_8 \) and \( \theta_1, \theta_2, \theta_3, \theta_5, \ldots, \theta_8 \) chosen as in Goegebeur et al. (JSPI, 2010) and De Wet et al. (SPL, 2012) : \( \tau_1 = 1.25, \tau_2 = \tau_3 = 1.75, \tau_4 = \tau_8 = 2, \tau_5 = 1.5, \tau_6 = \tau_7 = 1.75, \theta_1 = \theta_2 = 0.01, \theta_5 = \theta_6 = 0.02 \) and \( \theta_7 = \theta_8 = 0.04. \)
- \( \theta_3 = \theta_4 = 0.01 + 0.02\delta \) for \( \delta \geq 0. \)

A simple expression if obtained for \( \varphi : \)

\[
\varphi(\rho) = \text{cste} \left[ \frac{2 - \rho}{1.25 - \rho} \right] \left[ \frac{1.5 - \rho}{2 - \rho} \right]^{\delta}
\]

Its inverse is explicit when \( \delta = 0 \) (new explicit estimator) or \( \delta = 1 \) (Goegebeur et al. (JSPI, 2010)). We shall also consider the case \( \delta = 1.5 \) which can be shown to be in some sense ”optimal” when \( \rho = 0 \) (new implicit estimator).
Illustration on simulations

**Burr distribution**

Survival distribution function :

\[ 1 - F(x) = (1 + x^{-\rho})^{1/\rho} \]

with \( x \geq 0 \) and \( \rho < 0 \).

- Extreme-value index \( \gamma = 1 \), second order parameter \( \rho < 0 \).
- The third order condition holds with \( \beta = \rho \), \( A(x) = \gamma x^\rho / (1 - x^\rho) \) and \( B(x) = \rho x^\rho / (1 - x^\rho) \).

**Experimental design**

- Sample size \( n = 5000 \), 500 replications.
- Intermediate sequence \( k = 1500, ..., 4995 \).
- Second order parameter \( \rho = -0.25 \) and \( \rho = -1 \).
Asymptotic mean-squared error & empirical mean-squared error

Left: Asymptotic mean-squared error, Right: empirical mean-squared error.
Asymptotic mean-squared error & mean-squared error

Left : Asymptotic mean-squared error, Right : empirical mean-squared error.
Conclusion

+ **General framework** for building estimators of the second order parameter.
+ **Asymptotic normality** of the estimators is directly derived from the asymptotic behavior of rescaled log-spacings or log-excesses.
+ **Efficient tool** for studying existing estimators or defining new ones.
  - But ... How to compare in practice estimators depending on so many parameters?