Statistics and Probability Letters 119 (2016) 186–193

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Kernel estimation of the tail index of a right-truncated Pareto-type distribution

Souad Benchaira, Djamel Meraghni, Abdelhakim Necir *

Laboratory of Applied Mathematics, Mohamed Khider University, Biskra, Algeria

ARTICLE INFO

Article history:
Received 12 February 2016
Received in revised form 9 August 2016
Accepted 9 August 2016
Available online 17 August 2016

MSC:
60F17
62G30
62G32
62P05

Keywords:
Extreme value index
Heavy-tails
Kernel estimation
Product-limit estimator
Random truncation

ABSTRACT

An asymptotically normal kernel estimator for the positive tail index of right-truncated data is introduced. A simulation study shows that the proposed estimator performs much better than the existing ones in terms of bias.

1. Introduction

Let $(\mathbf{X}_i, \mathbf{Y}_i)$, $1 \leq i \leq N$ be a sample of size $N \geq 1$ from a couple $(\mathbf{X}, \mathbf{Y})$ of independent random variables (rv's) defined over some probability space $(\Omega, \mathcal{A}, \mathbf{P})$, with continuous marginal distribution functions (df's) $\mathbf{F}$ and $\mathbf{G}$ respectively. Suppose that $\mathbf{X}$ is truncated to the right by $\mathbf{Y}$, in the sense that $\mathbf{X}_i$ is only observed when $\mathbf{X}_i \leq \mathbf{Y}_i$. We assume that both survival functions $\overline{\mathbf{F}} := 1 - \mathbf{F}$ and $\overline{\mathbf{G}} := 1 - \mathbf{G}$ are regularly varying at infinity with negative indices $-1/\gamma_1$ and $-1/\gamma_2$ respectively. That is, for any $x > 0$,

$$\lim_{z\to\infty} \frac{\overline{\mathbf{F}}(xz)}{\overline{\mathbf{F}}(z)} = x^{-1/\gamma_1} \quad \text{and} \quad \lim_{z\to\infty} \frac{\overline{\mathbf{G}}(xz)}{\overline{\mathbf{G}}(z)} = x^{-1/\gamma_2}. \tag{1.1}$$

This class of distributions, which includes models such as Pareto, Burr, Fréchet, stable and log-gamma, plays a prominent role in extreme value theory. Also known as heavy-tailed, Pareto-type or Pareto-like distributions, these models have important practical applications and are used rather systematically in certain branches of non-life insurance as well as in finance, telecommunications, geology, and many other fields (see, e.g., Resnick, 2006). Let us denote (X_i, Y_i), $i = 1, \ldots, n$ to be the observed data, as copies of a couple of rv's (X, Y), corresponding to the truncated sample $(\mathbf{X}_i, \mathbf{Y}_i)$, $i = 1, \ldots, N$, where

* Corresponding author.
E-mail addresses: benchaira.s@hotmail.fr (S. Benchaira), djmeraghni@yahoo.com (D. Meraghni), necirabdelhakim@yahoo.fr (A. Necir).

http://dx.doi.org/10.1016/j.spl.2016.08.004
0167-7152/

$n = n_N$ is a sequence of discrete rv's which, in virtue of the weak law of large numbers, satisfies $n_N/N \xrightarrow{\mathbf{P}} p := \mathbf{P}(\mathbf{X} \le \mathbf{Y})$, as $N \to \infty$. We denote the joint distribution of X and Y by $H(x, y) := \mathbf{P}(X \le x, Y \le y)$. The df H is given by

$$H(x, y) = \mathbf{P}(\mathbf{X} \le \min(x, \mathbf{Y}), \mathbf{Y} \le y \mid \mathbf{X} \le \mathbf{Y}) = p^{-1} \int_0^y \mathbf{F}(x, z)\, d\mathbf{G}(z).$$

The marginal distributions of the rv's X and Y, respectively denoted by F and G, are equal to $F(x) = p^{-1} \int_0^x \overline{\mathbf{G}}(z)\, d\mathbf{F}(z)$ and $G(y) = p^{-1} \int_0^y \mathbf{F}(z)\, d\mathbf{G}(z)$. The tail of df F simultaneously depends on $\overline{\mathbf{G}}$ and $\overline{\mathbf{F}}$ while that of $\overline{G}$ only relies on $\overline{\mathbf{G}}$. By using Proposition B.1.10 in de Haan and Ferreira (2006), to the regularly varying functions $\overline{\mathbf{F}}$ and $\overline{\mathbf{G}}$, we show that both $\overline{F}$ and $\overline{G}$ are regularly varying at infinity as well, with respective indices $-1/\gamma := -(\gamma_1 + \gamma_2)/(\gamma_1\gamma_2)$ and $-1/\gamma_2$. By using the definition of γ, Gardes and Stupfler (2015) derived a consistent estimator, for the extreme value index γ_1, whose asymptotic normality is established in Benchaira et al. (2015), under the tail dependence and the second-order regular variation conditions. Also, Worms and Worms (2016) proposed an estimator for γ_1 and proved its asymptotic normality, by considering a Lynden-Bell integration with a deterministic threshold. More recently, Benchaira et al. (2016) treated the case of a random threshold and introduced a Hill-type estimator for the tail index γ_1 of randomly right-truncated data. The asymptotic normality of the latter is established by considering the second-order regular variation conditions (2.6) and (2.7) and the assumption $\gamma_1 < \gamma_2$. This condition is required in order to ensure that it remains enough extreme data for the inference to be accurate. In other words, we consider the situation where the tail of the rv of interest $\mathbf{X}$ is not too contaminated by that of the truncating rv $\mathbf{Y}$. In this paper, we derive a kernel-type estimator for γ_1, under random right truncation, in the spirit of the work of Csörgő et al. (1985) in the complete data case. Thereby, for a suitable choice of the kernel function, we obtain an improved estimator of γ_1 in terms of bias. To this end, let $\mathbb{K} : \mathbb{R} \to \mathbb{R}_+$ be a fixed function, that will be called kernel, satisfying the following conditions:

[$\mathbb{C}$1] $\mathbb{K}$ is non increasing and right-continuous on $\mathbb{R}$.
[$\mathbb{C}$2] $\mathbb{K}(s) = 0$ for $s \notin [0, 1)$ and $\mathbb{K}(s) \ge 0$ for $s \in [0, 1)$.
[$\mathbb{C}$3] $\int_{\mathbb{R}} \mathbb{K}(s) ds = 1$.
[$\mathbb{C}$4] $\mathbb{K}$ and its first and second Lebesgue derivatives $\mathbb{K}'$ and $\mathbb{K}''$ are bounded on $\mathbb{R}$.

As examples of such functions (see, e.g., Groeneboom et al., 2003), we have the indicator kernel $\mathbb{K} = \mathbf{1}_{[0,1)}$ and the biweight and triweight kernels respectively defined by

$$\mathbb{K}_2(s) := \frac{15}{8}\left(1 - s^2\right)^2 \mathbf{1}_{\{0 \le s < 1\}}, \qquad \mathbb{K}_3(s) := \frac{35}{16}\left(1 - s^2\right)^3 \mathbf{1}_{\{0 \le s < 1\}}. \tag{1.2}$$

For an overview of kernel estimation of the extreme value index with complete data, one refers to, for instance, Hüsler et al. (2006) and Ciuperca and Mercadier (2010). By using Potter's inequalities, see e.g. Proposition B.1.10 in de Haan and Ferreira (2006), to the regularly varying function $\overline{\mathbf{F}}$ together with assumptions [$\mathbb{C}$1] – [$\mathbb{C}$3], we may readily show that

$$\lim_{u \to \infty} \int_u^\infty x^{-1} \frac{\overline{\mathbf{F}}(x)}{\overline{\mathbf{F}}(u)} \mathbb{K}\left(\frac{\overline{\mathbf{F}}(x)}{\overline{\mathbf{F}}(u)}\right) dx = \gamma_1 \int_0^\infty \mathbb{K}(s) ds = \gamma_1. \tag{1.3}$$

An integration by parts yields

$$\lim_{u \to \infty} \frac{1}{\overline{\mathbf{F}}(u)} \int_u^\infty g_{\mathbb{K}}\left(\frac{\overline{\mathbf{F}}(x)}{\overline{\mathbf{F}}(u)}\right) \log \frac{x}{u} d\mathbf{F}(x) = \gamma_1, \tag{1.4}$$

where $g_{\mathbb{K}}$ denotes the Lebesgue derivative of the function $s \to \Psi_{\mathbb{K}}(s) := s\mathbb{K}(s)$. Note that, for $\mathbb{K} = \mathbf{1}_{[0,1)}$, we have $g_{\mathbb{K}} = \mathbf{1}_{[0,1)}$, then the previous two limits meet assertion (1.2.6) given in Theorem 1.2.2 by de Haan and Ferreira (2006). For kernels $\mathbb{K}_2$ and $\mathbb{K}_3$, we have

$$g_{\mathbb{K}_2}(s) := \frac{15}{8}\left(1 - s^2\right)\left(1 - 5s^2\right) \mathbf{1}_{\{0 \le s < 1\}}, \qquad g_{\mathbb{K}_3}(s) := \frac{35}{16}\left(1 - s^2\right)^2\left(1 - 7s^2\right) \mathbf{1}_{\{0 \le s < 1\}}.$$

Let $X_{1:n} \le \cdots \le X_{n:n}$ be the order statistics pertaining to the sample $(X_1, \ldots, X_n)$ and $k = k_n$ be a (random) sequence of discrete rv's such that given $n = m$, $k_m \to \infty$ and $k_m/m \to 0$ as $N \to \infty$. Since $\overline{F}$ is regularly varying at infinity, then $X_{n-k:n}$ tends to ∞ almost surely. By replacing, in (1.4), u by $X_{n-k:n}$ and $\mathbf{F}$ by the well-known Woodroofe's product-limit estimator (Woodroofe, 1985) $\mathbf{F}_n(x) := \prod_{i:X_i > x} \exp\{-1/(nC_n(X_i))\}$, with $C_n(x) := n^{-1} \sum_{i=1}^n \mathbf{1}(X_i \le x \le Y_i)$, we get

$$\widehat{\gamma}_{1,\mathbb{K}} = \frac{1}{\overline{\mathbf{F}}_n(X_{n-k:n})} \int_{X_{n-k:n}}^\infty g_{\mathbb{K}}\left(\frac{\overline{\mathbf{F}}_n(x)}{\overline{\mathbf{F}}_n(X_{n-k:n})}\right) \log \frac{x}{X_{n-k:n}} d\mathbf{F}_n(x),$$

as a kernel estimator for γ_1. Next, we give an explicit formula for $\widehat{\gamma}_{1,\mathbb{K}}$. Since $\overline{\mathbf{F}}$ and $\overline{\mathbf{G}}$ are regularly varying at infinity, then their right endpoints are infinite and so they are equal. Hence, from Woodroofe (1985), we may write $\int_x^\infty d\mathbf{F}(y)/\mathbf{F}(y) = \int_x^\infty dF(y)/C(y)$, where $C(z) := \mathbf{P}(X \le z \le Y)$ is the theoretical counterpart of $C_n(z)$ defined above. Differentiating