Journal of Multivariate Analysis 144 (2016) 1–24

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Estimating the conditional extreme-value index under random right-censoring

Gilles Stupfler *

Aix Marseille Université, CERGAM, EA 4225, 15-19 allée Claude Forbin, 13628 Aix-en-Provence Cedex 1, France

ARTICLE INFO

Article history:
Received 9 November 2013
Available online 4 November 2015

AMS 2010 subject classifications:
62G05
62G20
62G30
62G32
62N01
62N02

Keywords:
Extreme-value index
Random covariate
Random right-censoring
Consistency
Asymptotic normality

ABSTRACT

In extreme value theory, the extreme-value index is a parameter that controls the behavior of a cumulative distribution function in its right tail. Estimating this parameter is thus the first step when tackling a number of problems related to extreme events. In this paper, we introduce an estimator of the extreme-value index in the presence of a random covariate when the response variable is right-censored, whether its conditional distribution belongs to the Fréchet, Weibull or Gumbel domain of attraction. The pointwise weak consistency and asymptotic normality of the proposed estimator are established. Some illustrations on simulations are provided and we showcase the estimator on a real set of medical data.

1. Introduction

Studying extreme events is relevant in numerous fields of statistical applications. For instance, one can think about hydrology, where it is of interest to estimate the maximum level reached by seawater along a coast over a given period, or to study extreme rainfall at a given location; in actuarial science, a major problem for an insurance firm is to estimate the probability that a claim so large that it represents a threat to its solvency is filed. The focus in this type of problem is not in the estimation of "central" parameters of the random variable of interest, such as its mean or median, but rather in the understanding of its behavior in its right tail. The basic result in extreme value theory, known as the Fisher–Tippett–Gnedenko theorem (Fisher and Tippett [13], Gnedenko [17]) states that if (Y_n) is an independent sequence of random copies of a random variable Y such that there exist normalizing nonrandom sequences of real numbers (a_n) and (b_n), with $a_n > 0$ and such that the sequence

$$\frac{1}{a_n}\left(\max_{1\leq i\leq n} Y_i - b_n\right)$$

converges in distribution to some nondegenerate limit, then the cumulative distribution function (cdf) of this limit has the form $y \mapsto G_{\gamma_Y}(ay+b)$, with $a > 0$ and $b, \gamma_Y \in \mathbb{R}$ where

$$G_{\gamma_Y}(y) = \begin{cases} \exp\left(-(1+\gamma_Y y)^{-1/\gamma_Y}\right) & \text{if } \gamma_Y \neq 0 \text{ and } 1+\gamma_Y y > 0, \\ \exp\left(-\exp(-y)\right) & \text{if } \gamma_Y = 0. \end{cases}$$

* Correspondence to: Aix Marseille Université, CNRS, EHESS, Centrale Marseille, GREQAM UMR 7316, 13002 Marseille, France.
E-mail address: gilles.stupfler@univ-amu.fr.

http://dx.doi.org/10.1016/j.jmva.2015.10.015
0047-259X/

If the aforementioned convergence holds, we shall say that Y (or equivalently, its cdf F_Y) belongs to the domain of attraction (DA) of G_{γ_Y}, with γ_Y being the so-called extreme-value index of Y, and we write $F_Y \in \mathcal{D}(G_{\gamma_Y})$. The parameter γ_Y drives the behavior of G_{γ_Y} (and thus of F_Y) in its right tail:

- if $\gamma_Y > 0$, namely Y belongs to the Fréchet DA, then $1 - G_{\gamma_Y}$ is heavy-tailed *i.e.* it has a polynomial decay;
- if $\gamma_Y < 0$, namely Y belongs to the Weibull DA, then $1 - G_{\gamma_Y}$ is short-tailed *i.e.* it has a support bounded to the right;
- if $\gamma_Y = 0$, namely Y belongs to the Gumbel DA, then $1 - G_{\gamma_Y}$ has an exponential decay.

This makes it clear that the estimation of γ_Y is a first step when tackling various problems in extreme value analysis, such as the estimation of extreme quantiles of Y. Recent monographs on extreme value theory and especially univariate extreme-value index estimation include Beirlant et al. [2] and de Haan and Ferreira [10].

In practical applications, it may happen that only incomplete information is available. Consider for instance a medical follow-up study lasting up to time t which collects the survival times of patients for a given chronic disease. If a patient is diagnosed with the disease at time s, his/her survival time is known if and only if he/she dies before time t. If the patient survives until the end of the study, the only information available is that his/her survival time is not less than $t - s$. This situation is the archetypal example of right-censoring, which shall be the focus of this paper. An interesting problem in this particular case is the estimation of extreme survival times or, in other words, how long an exceptionally strong individual can survive the disease. A preliminary step necessary to give an answer to this question is to estimate the extreme-value index of the survival time Y; this problem, which is much more complex than the estimation of the extreme-value index when the data set is complete, has been investigated quite recently by Beirlant et al. [3] where asymptotic results for an extreme-value index estimator using the data above a nonrandom threshold are derived in the context of the Hall model (see Hall [20]), Einmahl et al. [12] in which the authors also suggest an estimator of extreme quantiles under random right-censoring so as to provide extreme survival times for male patients suffering from AIDS, Beirlant et al. [4] where maximum likelihood estimators are discussed, Sayah et al. [25] who focus on the heavy-tailed case and introduce a robust estimator with respect to contamination and Worms and Worms [29] where the consistency of several estimators, coming either from Kaplan–Meier integration or censored regression techniques, is studied. This situation should not be confused with right-truncation, in which case no information is available at all when Y is not actually observed: a recent reference in this case is Gardes and Stupfler [16].

Besides, it may well be the case that the survival time of a patient depends on additional random factors such as his/her age or the pre-existence of some other medical condition. Our goal in this study is to make it possible to integrate such information in the model by taking into account the dependency of Y on a covariate X. The problem thus becomes to estimate the conditional extreme-value index $\gamma_Y(x)$ of Y given $X = x$. Recent papers on this subject when Y is noncensored include Wang and Tsai [28] who introduced a maximum likelihood approach, Daouia et al. [8] who used a fixed number of nonparametric conditional quantile estimators to estimate the conditional extreme-value index, Gardes and Girard [14] who generalized the method of [8] to the case when the covariate space is infinite-dimensional, Goegebeur et al. [18] who studied a nonparametric regression estimator whose uniform asymptotic properties are examined in Goegebeur et al. [19] and Gardes and Stupfler [15] who introduced a smoothed local Hill estimator (see Hill [21]). All these papers consider the case when Y given $X = x$ belongs to the Fréchet DA; the case when the response distribution belongs to an arbitrary domain of attraction is considered in Daouia et al. [7], who generalized the method of [8] to this context and Stupfler [26] who introduced a generalization of the popular moment estimator of Dekkers et al. [11]. To the best of our knowledge, the only paper tackling this problem when Y is right-censored is Ndao et al. [23]; their work is, however, restricted to the case when Y is heavy-tailed. Our focus here is to devise an estimator which works regardless of whether or not the tail of Y is heavy.

The outline of this paper is as follows. In Section 2, we give a precise definition of our model. In Section 3, we define our estimator of the conditional extreme-value index. The pointwise weak consistency and asymptotic normality of the estimator are stated in Section 4. The finite sample performance of the estimator is studied in Section 5. In Section 6, we revisit the medical data set of [12] by integrating additional covariate information. Proofs are deferred to Section 7.

2. Framework

Let $(X_1, Y_1, C_1), \ldots, (X_n, Y_n, C_n)$ be n independent copies of a random vector (X, Y, C) taking its values in $E \times (0, \infty) \times (0, \infty)$ where E is a finite-dimensional linear space endowed with a norm $\|\cdot\|$. We assume that for all $x \in E$, given $X = x$, Y and C are independent, possess continuous probability density functions (pdfs) and that the related conditional survival functions (csfs) $\overline{F}_Y(\cdot|x) = 1 - F_Y(\cdot|x)$ of Y given $X = x$ and $\overline{F}_C(\cdot|x) = 1 - F_C(\cdot|x)$ of C given $X = x$ belong to some domain of attraction. Specifically, we shall work in the following setting, where we recall that the left-continuous inverse of a nondecreasing function f is the function $z \mapsto \inf\{y \in \mathbb{R} \mid f(y) \geq z\}$:

(M_1) Y and C are positive random variables and for every $x \in E$, there exist real numbers $\gamma_Y(x)$, $\gamma_C(x)$ and positive functions $a_Y(\cdot|x)$, $a_C(\cdot|x)$ such that the left-continuous inverses $U_Y(\cdot|x)$ of $1/\overline{F}_Y(\cdot|x)$ and $U_C(\cdot|x)$ of $1/\overline{F}_C(\cdot|x)$ satisfy

$$\lim_{t\to\infty} \frac{U_Y(tz|x) - U_Y(t|x)}{a_Y(t|x)} = D_{\gamma_Y(x)}(z) \quad \text{and} \quad \lim_{t\to\infty} \frac{U_C(tz|x) - U_C(t|x)}{a_C(t|x)} = D_{\gamma_C(x)}(z)$$