Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Bayesian estimation of the tail index of a heavy tailed distribution under random censoring

Abdelkader Ameraoui [a,*], Kamal Boukhetala [a], Jean-François Dupuy [b]

[a] Faculty of Mathematics, PoBox 32, Al alia Bab ezzouar, USTHB - Algiers, Algeria
[b] IRMAR-INSA de Rennes, 20 Avenue des Buttes de Coësmes, 35708 Rennes cedex 7, France

ARTICLE INFO

Article history:
Received 31 July 2015
Received in revised form 20 May 2016
Accepted 17 June 2016
Available online xxxx

Keywords:
Extreme value modeling
Random censoring
Maximum a posteriori estimator
Mean posterior estimator
Asymptotic normality of posterior
Simulations

ABSTRACT

Bayesian estimation of the tail index of a heavy-tailed distribution is addressed when data are randomly right-censored. Maximum a posteriori and mean posterior estimators are constructed for various prior distributions of the tail index. Convergence of the posterior distribution of the tail index to a Gaussian distribution is established. Finite-sample properties of the proposed estimators are investigated via simulations. Tail index estimation requires selecting an appropriate threshold for constructing relative excesses. A Monte Carlo procedure is proposed for tackling this issue. Finally, the proposed estimators are illustrated on a medical dataset.

1. Introduction

Tail index estimation is one of the most important issues in extreme value theory. The tail index measures the thickness of the tail of a probability distribution function and thus plays a crucial role for evaluating the risk of occurrence of extreme events. In particular, estimation of the tail index constitutes usually a first step in an extreme value analysis. A vast literature has been dedicated to this topic. Recent overviews can be found in the monographs (Beirlant et al., 2004; Coles, 2001; de Haan and Ferreira, 2006).

Let F be the cumulative distribution function (cdf) of some non-negative random variable Y. We assume that F is heavy-tailed, that is, there exists a constant $\alpha > 0$ such that

$$1 - F(x) = x^{-\alpha}\ell(x), \quad (1)$$

where ℓ is a slowly varying function at infinity:

$$\lim_{x\to\infty} \frac{\ell(tx)}{\ell(x)} = 1 \quad \text{for all } t > 0.$$

If (1) holds, we have:

$$\lim_{x\to\infty} \frac{1 - F(tx)}{1 - F(x)} = t^{-\alpha} \quad \text{for all } t > 0$$

* Corresponding author.
E-mail addresses: aameraoui@usthb.dz (A. Ameraoui), kboukhetala@usthb.dz (K. Boukhetala), Jean-Francois.Dupuy@insa-rennes.fr (J.-F. Dupuy).

http://dx.doi.org/10.1016/j.csda.2016.06.009

Please cite this article in press as: Ameraoui, A., et al., Bayesian estimation of the tail index of a heavy tailed distribution under random censoring. Computational Statistics and Data Analysis (2016), http://dx.doi.org/10.1016/j.csda.2016.06.009

and we say that $\bar{F} = 1 - F$ is regularly varying at infinity with tail index α, which we denote by $\bar{F} \in \mathcal{R}_{-\alpha}$. The positive number $\gamma := \alpha^{-1}$ is called the extreme value index (EVI) of F. The conditions above amount to assuming that the distribution function F is in the max-domain of attraction of a Fréchet distribution. Such distribution functions are useful in practice for investigating phenomena where exceptional values have a significant occurrence frequency. Examples include the number of claims in insurance (Embrechts et al., 1997), transmission times in telecommunications (Resnick, 2007), log-returns of price speculation (Embrechts et al., 1997).

Several estimators have been proposed for the tail index α, or equivalently, for the EVI γ, such as Pickands estimator (Pickands, 1975), Dekkers et al. (or moment) estimator (Dekkers et al., 1989) and Hill estimator (Hill, 1975), which is the most popular estimator of γ in model (1). Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (i.i.d. thereafter) random variables with common cdf F. Let $k \in \{2, \ldots, n\}$ and $X_{n,1} \le X_{n,2} \le \cdots \le X_{n,n}$ be the order statistics of the sample $X_1, X_2, \ldots, X_n$. Hill estimator is defined as

$$H(k) := \frac{1}{k}\sum_{i=1}^{k} \log(X_{n,n-i+1}) - \log(X_{n,n-k}). \tag{2}$$

Consistency of Hill estimator was proved in Mason (1982) under the regular variation condition (1). Its asymptotic normality was further established under an additional condition known as the second-order regular variation condition (see de Haan and Ferreira, 2006).

In this paper, we address the estimation of the tail index α when observations $X_1, X_2, \ldots, X_n$ are randomly right-censored. Censoring commonly occurs in the analysis of event time data. For example, X may represent the duration until the occurrence of some event of interest, such as death of a patient. If a patient is still alive or has dropped out of the study for some reason when the data are collected, the variable of interest X is not available. An appropriate way to model this situation is to introduce a random variable Y (called a censoring random variable) such that observations consist of pairs (Z_i, δ_i), $1 \le i \le n$ where $Z_i = \min(X_i, Y_i)$, $\delta_i = \mathbb{1}_{\{X_i \le Y_i\}}$ and $\mathbb{1}$ is the indicator function. Estimation of the EVI with censored data was considered in Beirlant et al. (2007), Brahimi et al. (2013), Einmahl et al. (2008) and Gomes and Neves (2011). For example, Beirlant et al. (2007) proposed to estimate γ by the following modified version of Hill estimator:

$$H^{C}(k) := \frac{\sum_{i=1}^{k} \log(Z_{n,n-i+1}) - \log(Z_{n,n-k})}{\sum_{i=1}^{k} \delta_{[n-i+1]}}, \tag{3}$$

where $Z_{n,1} \le Z_{n,2} \le \cdots \le Z_{n,n}$ are the order statistics of the censored sample $Z_1, Z_2, \ldots, Z_n$ and $\delta_{[n-i+1]}$ is the concomitant value of δ associated with $Z_{n,n-i+1}$, that is, $\delta_j = \delta_{[n-i+1]}$ if $Z_j = Z_{n,n-i+1}$. More generally, Einmahl et al. (2008) proposed to estimate γ by

$$\hat{\gamma}_Z^{(C)}(k) = \frac{\hat{\gamma}_Z(k)}{\hat{p}}, \tag{4}$$

where $\hat{\gamma}_Z(k)$ is any of the classical EVI estimators calculated on the censored observations $Z_1, \ldots, Z_n$ and $\hat{p} = \frac{1}{k}\sum_{i=1}^{k} \delta_{[n-i+1]}$ is the proportion of uncensored values in the k largest observations of Z. Obviously, the estimator (4) coincides with the adapted Hill estimator (3) when $\hat{\gamma}_Z(k)$ is Hill estimator (2). In this paper, we adopt a completely different approach and investigate Bayesian estimation of the tail index $\alpha := \gamma^{-1}$. Bayesian estimation will allow us to incorporate *a priori* knowledge about the data.

Bayesian estimation provides an alternative to frequentist methods. In the context of extreme values analysis without censoring, Bayesian estimators have been investigated in Cabras and Castellanos (2011), Coles and Powell (1996), Diebolt et al. (2005), do Nascimento et al. (2012), Stephenson and Tawn (2004) and de Zea Bermudez and Kotz (2010). See also Beirlant et al. (2004) (Chapter 11). Applications include operational risk modeling (Ergashev et al., 2013), hydrology (Liang et al., 2011) and market indices modeling (So and Chan, 2014). But to the best of our knowledge, no Bayesian estimator of the tail index α has been proposed when censoring is present. The present work intends to fill in this gap.

We construct several Bayesian estimators of α in model (1). Bayesian estimation requires specifying a prior distribution for the unknown parameter. We investigate various priors (Jeffreys, maximal data information), leading to several maximum a posteriori and mean posterior estimators of α. Convergence of the posterior distribution of α to a Gaussian distribution is established. Finite-sample performance of the proposed estimators is assessed via simulations. Tail index estimation requires choosing an appropriate threshold for defining excesses. We propose a randomization method for tackling this issue. This procedure is evaluated by simulation. Finally, we illustrate our methodology on a set of AIDS survival data.

The paper is organized as follows. In Section 2, we construct our estimators and consider asymptotic results. Proofs are deferred to an Appendix. Section 3 reports the results of a comprehensive simulation study. The randomization procedure for threshold selection is proposed and assessed by simulations in Section 4. Application to AIDS data is carried out in Section 5. A discussion and some perspectives are given in Section 6.

Please cite this article in press as: Ameraoui, A., et al., Bayesian estimation of the tail index of a heavy tailed distribution under random censoring. Computational Statistics and Data Analysis (2016), http://dx.doi.org/10.1016/j.csda.2016.06.009