Journal of Multivariate Analysis 130 (2014) 335–353

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Subsampling extremes: From block maxima to smooth tail estimation

Stefan Wager

Department of Statistics, Stanford University, United States

ARTICLE INFO

Article history:
Received 9 January 2014
Available online 16 June 2014

AMS subject classifications:
62G32
62G09

ABSTRACT

We study a new estimator for the tail index of a distribution in the Fréchet domain of attraction that arises naturally by computing subsample maxima. This estimator is equivalent to taking a U-statistic over a Hill estimator with two order statistics. The estimator presents multiple advantages over the Hill estimator. In particular, it has asymptotically $\mathcal{C}^\infty$ sample paths as a function of the threshold k, making it considerably more stable than the Hill estimator. The estimator also admits a simple and intuitive threshold selection rule that does not require fitting a second-order model.

1. Introduction

Researchers in multiple fields face a growing need to understand the tails of probability distributions, and extreme value theory presents tools which, under certain regularity assumptions, let us build simple yet powerful models for these tails. In the case of heavy tailed distributions, the setting of extreme value theory is as follows: suppose our data is drawn from a distribution F, and assume that there is a constant $\gamma > 0$ and some slowly varying function L such that

$$1 - F(x) = L(x) \cdot x^{-\frac{1}{\gamma}}, \quad \text{with } \lim_{x\to\infty} \frac{L(ax)}{L(x)} = 1 \text{ for all } a > 0. \tag{1}$$

Then, F is in what is called the Fréchet domain of attraction. If F satisfies this property (which most commonly used heavy-tailed distributions do), extreme value theory provides an elegant and concise description of the asymptotic properties of sample maxima of F. A major challenge is that this description relies on knowledge of the parameter γ, called the tail index of the distribution F. And, unfortunately, estimating γ from data is not always straightforward.

The literature on tail index estimation is quite extensive. One of the most widely used estimators is due to Hill [28], who suggests estimating γ with a simple functional of the top $k+1$ order statistics of the empirical distribution:

$$\hat{\gamma}_H := \frac{1}{k}\sum_{j=0}^{k-1} \log\left[\frac{X_{n-j,n}}{X_{n-k,n}}\right]. \tag{2}$$

Here, $X_{1,n} \le \cdots \le X_{n,n}$ denote the order statistics of X, and k must be selected such that $X_{n-k,n} > 0$. Hill showed that $\hat{\gamma}_H$ converges in probability to $\gamma > 0$, provided the threshold sequence $k = k(n)$ is an intermediate sequence that grows to infinity slower than the sample size n. Hill's idea of using a functional of extreme and intermediate order statistics to estimate γ has received considerable attention. Csörgő et al. [8] suggest ways to adaptively weight the order statistics, while Dekkers et al. [11] modify Hill's estimator so that it is also consistent for a generalization of (1) that includes negative γ. There have

E-mail address: swager@stanford.edu.

http://dx.doi.org/10.1016/j.jmva.2014.06.010
0047-259X/

been proposals to eliminate the asymptotic bias of the Hill estimator [2,19,22,32]; recent proposals [6,21,23] show how to do so without increasing asymptotic variance.

Nonetheless, tail index estimation remains quite challenging, especially for smaller samples on the order of a few hundred to a thousand points. Of course, many difficulties are inherent to the subject matter: only a small fraction of any sample will be inside the tail of the underlying distribution, and so even large samples may contain very little information relevant to inference about this tail.

Other challenges, however, seem to arise from specifics of popular estimators. All estimators for γ require choosing a threshold at which the tail area of the distribution begins. Ideally, specifying a good threshold should be easy, and the estimate $\hat{\gamma}$ should not be sensitive to small changes in the threshold. Unfortunately, most commonly used estimators for γ do not reach this ideal. In the case of the Hill estimator – where the parameter k from (2) stands in for the threshold – the choice is far from innocuous:

- Inadequate choice of k can lead to large expected error. Small values of k lead to high variance, while large values of k usually lead to high bias. There is often an intermediate region for k where the estimator has fairly small expected error, but it is not always easy to find this region.
- The Hill estimator is extremely sensitive to small changes in k, even asymptotically: Mason and Turova [31] show that the Hill estimator process converges in law to a modified Brownian motion. Thus, even within the 'good' region with low expected error, a minute change in k can impact the conclusions to be drawn from the model.

The problem of choosing the threshold k has been discussed, among others, by Beirlant et al. [3], Danielsson et al. [9], Drees and Kaufmann [16], and Guillou and Hall [24]. Most existing methods rely on fairly complicated auxiliary models: all but the last of the cited ones require either implicitly or explicitly fitting a difficult-to-fit second-order convergence parameter. As the method due to Guillou and Hall does not require fitting secondary parameters, we use it as our main benchmark in simulation studies. The problem of excessive oscillation of the Hill estimator has been discussed by Resnick and Stărică [36], who recommend smoothing the Hill estimator by integrating it over a moving window. We are not aware of any guidance on how to automatically select k for this smoothed Hill estimator.

In this paper, we study a new estimator for γ that arises from a simple subsampling idea. It is well known that sample maxima from a distribution F satisfying (1) have the following property: if $X_1, \ldots, X_n$ are drawn independently from F, then

$$\lim_{n\to\infty} \mathbb{P}\left[\frac{\max\{X_1, \ldots, X_n\}}{\ell(n)\cdot n^{\gamma}} \leq x\right] = G_\gamma(x),$$

where G_γ is a limiting cumulative distribution function that only depends on γ and $\ell(n)$ is an appropriately chosen slowly varying function. Noting this, we may suspect that when F has positive support,

$$\lim_{s\to\infty} s\cdot(\mathbb{E}[\log\max\{X_1, \ldots, X_s\}] - \mathbb{E}[\log\max\{X_1, \ldots, X_{s-1}\}]) = \gamma. \tag{3}$$

In Theorem 3.3, we show that this relation in fact holds under very mild conditions on F near 0. Our estimator follows directly from this formula. Given a subsample size $1 < s \leq n$, we first estimate the quantities

$$\mathbb{E}[\log\max\{X_1, \ldots, X_s\}] \quad \text{and} \quad \mathbb{E}[\log\max\{X_1, \ldots, X_{s-1}\}]$$

by subsampling our data without replacement, and then use (3) to obtain an estimate for $\hat{\gamma}$. Since this estimator operates by computing the average log maxima of randomly subsampled blocks, we call it the Random Block Maxima (RBM) estimator. Our idea is related to proposals for tail index estimation that study weighted sums of log-ratios of order statistics by, e.g., Drees [14] and Gardes and Girard [20].

The RBM estimator can be understood as belonging to two different frameworks of tail index estimation. The block maxima approach, which was often used in the early days of extreme value theory, aims to directly fit the distribution of the maxima of fixed (e.g., yearly) blocks of data. See Gumbel [25] for a review; Dombry [13] and Ferreira and de Haan [18] provide a modern analysis. In this light, the RBM estimator can be seen as a randomized method of moments estimator in the block maxima framework. Our estimator, however, can also be seen as an outgrowth of the more modern tail estimation paradigm started by the Hill estimator: as we will show, the RBM estimator can be constructed by taking a U-statistic over a Hill estimator with two order statistics. In other words, once we start subsampling the data, the block maxima and Hill estimation frameworks merge and lead to the RBM estimator.

Our estimator behaves much like the Hill estimator; however, it addresses threshold selection much more naturally than the latter:

- The RBM estimator has asymptotically smooth sample paths as a function of its threshold parameter k as defined in (7), and, even in modestly sized samples, does not suffer from small-scale instability in k.
- Thanks to its smoothness properties, the RBM estimator admits a simple and intuitive threshold selection rule that does not require fitting a second-order model.

Fig. 1 shows estimates produced by both the Hill and RBM estimators for the tail index γ of gross proceeds from venture capital backed IPOs in the United States between 1995 and 2011. Both estimates depend on a threshold parameter k. As expected, the RBM sample path is much smoother than the Hill sample path. This makes it easier to select k with the RBM