ARTICLE IN PRESS

Journal of the Korean Statistical Society ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

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

Parameter and quantile estimation for the generalized Pareto distribution in peaks over threshold framework☆

Suyeon Kang, Jongwoo Song *

Department of Statistics, Ewha Womans University, 11-1 Daehyun-dong, Seodaemun-gu, Seoul 120-750, Republic of Korea

ARTICLE INFO

Article history:
Received 16 November 2016
Accepted 23 February 2017
Available online xxxx

AMS 2000 subject classifications:
primary 60G70
secondary 62G32

Keywords:
Generalized Pareto distribution
Peaks over threshold
Maximum likelihood estimator
Nonlinear least squares
Hill estimator

ABSTRACT

In this article, we consider six estimation methods for extreme value modeling and compare their performances, focusing on the generalized Pareto distribution (GPD) in the peaks over threshold (POT) framework. Our goal is to identify the best method in various conditions via a thorough simulation study. In order to compare the estimators in the POT sense, we suggest proper strategies for some estimators originally not developed under the POT framework. The simulation results show that a nonlinear least squares (NLS) based estimator outperforms others in parameter estimation, but there is no clear winner in quantile estimation. For quantile estimation, NLS-based methods perform well even when the sample size is small and the Hill estimator comes to the front when the underlying distribution has a very heavy tail. Applications of EVT cover many different fields and researchers on each field may have their own experimental conditions or practical restrictions. We believe that our results would provide guidance on determining proper estimation method on future analysis.

1. Introduction

Extreme value theory (EVT) is used as a popular tool to model the risk of rare events and deal with extreme values. Applications of extreme value modeling involve the field of meteorology, hydrology, material science, insurance, finance, and survival analysis. There are two types of approaches for extreme value analysis. The first method takes the largest (or smallest) value per certain period of time and generates daily, monthly, or annual maxima series data, leading to the generalized extreme value (GEV) distribution being selected to fitting; see Fisher and Tippett (1928) and Gnedenko (1943). For example, after recording daily rainfall for 10 years, we can take the largest value for each year. The second method extracts the peak values which exceed a certain threshold. It is generally referred to peaks over threshold (POT) method. In this method, the excess values over high threshold are modeled with the generalized Pareto distribution (GPD); see McNeil and Saladin (1997). When given some observations, we can estimate the GPD parameter from the data and estimate extreme values using the estimated GPD parameters. Details for difference between the two approaches can be seen in Caires (2009) and Ferreira and de Haan (2015). We here focus on the second method, POT framework.

The GPD is a three-parameter distribution having location parameter μ, scale parameter σ, and shape parameter ξ. Sometimes the GPD is only specified by σ and ξ. Many references focus on the two parameters and use them to estimate

☆ This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2015S1A5B6036244).

* Corresponding author.
E-mail addresses: korea92721@naver.com (S. Kang), josong@ewha.ac.kr (J. Song).

http://dx.doi.org/10.1016/j.jkss.2017.02.003
1226-3192/© 2017 The Korean Statistical Society. Published by Elsevier B.V. All rights reserved.

Please cite this article in press as: Kang, S., & Song, J., Parameter and quantile estimation for the generalized Pareto distribution in peaks over threshold framework. Journal of the Korean Statistical Society (2017), http://dx.doi.org/10.1016/j.jkss.2017.02.003

extremes. After the GPD is first introduced by Pickands (1975), many researchers have studied the estimation methods for the GPD parameters in the last few decades. Hosking and Wallis (1987) suggested the methods of moments (MOM) and the probability-weighted moment (PWM), but both of the estimators show poor performance when $\xi > 1$ since the mean and variance of the GPD do not exist for the interval. Chen and Balakrishnan (1995) and Zhang (2007) criticized their infeasibility and low asymptotic efficiency. Dupuis and Tsao (1998) tried to improve the estimators, but it was difficult to overcome the drawbacks of the moment-based estimator. Alternatively, Castillo and Hadi (1997) proposed an elemental percentile method (EPM), but its performance was similar with those of MOM and PWM when $-2 < \xi < 2$. Most of the references mentioned above consider the classical maximum likelihood estimator (MLE) and point out its computational difficulties when $\xi < -1$. To solve the problems, del Castillo and Serra (2015) and Zhang (2007) proposed easy-to-compute estimation methods based on the likelihood. After that, Zhang (2010) and Zhang and Stephens (2009) suggested new estimators based on the likelihood and Bayesian approach and showed that the proposed method outperformed its previous versions and classical MLE in terms of bias and efficiency, which is defined as the ratio of the Cramér–Rao lower bound to the mean squared error (MSE). Recently, Song and Song (2012) proposed new estimators using a nonlinear least squares (NLS) method and Park and Kim (2016) suggested a new procedure adapted from the NLS-based approach by Song and Song (2012). More than these methods exist; see Beirlant, Goegebeur, Segers, and Teugels (2006), de Haan and Ferreira (2007) and Embrechts, Klüppelberg, and Mikosch (2013).

When reviewing many references, we are motivated to compare the different estimators under various conditions. The conditions considered in them, such as the value of parameters, sample sizes, and the number of observations used in the estimation, are different from each other. Several articles showed their results only in the restricted condition. For example, some references that include the classical MLE in their performance comparison only consider the case of $\xi > -1$ because the MLE does not exist outside of the interval. Also, in Zhang (2010), no consideration was given to the case when $\xi < -0.5$ since one of their performance metric, estimation efficiency, is not defined because Cramér–Rao lower bound does not exist; see Wang and Chen (2016). Their estimator performed well in the considered interval, however, it was showed that its performance got worse when $\xi < -1$, which is a range not considered in the original analysis; see del Castillo and Serra (2015) for graphical results. In addition, we found that some significant competitors have recently been proposed in the EVT literature. In this context, we think that it is necessary to compare the estimators under various conditions and find out which estimator performs better under which conditions. We here provide results of both of quantile and parameter estimation because these two types of estimations are different matters. We believe that future researchers can use more proper method based on results of this article.

The rest of this paper is organized as follows. Section 2 describes the GPD and POT method and Section 3 introduces six estimation methods considered in our analysis. Moreover, we suggest proper strategies for the estimators in order to compare them under the POT framework. In Section 4, we give the simulation results for estimating the GPD parameters using the estimators both with and without the POT method. And then, Section 5 compares their performance on estimating extreme values with varying experimental conditions, such as sample size, threshold value, and tail-heaviness of the sampling distribution. The last section is devoted to summarize our results and conclude this paper.

2. Extreme value theory

2.1. Generalized Pareto distribution (GPD)

One of the important distribution in extreme value modeling is the generalized Pareto distribution (GPD) with the distribution function being defined as

$$G_{\xi,\sigma}(x) = \begin{cases} 1-(1+\xi x/\sigma)^{-1/\xi}, & \text{if } \xi \neq 0 \\ 1-\exp(-x/\sigma), & \text{if } \xi = 0 \end{cases}$$

where ξ and σ are the shape and scale parameters, respectively. The domain of x is $(0, \infty)$ when $\xi \geq 0$ and $(0, -\sigma/\xi)$ when $\xi < 0$. If the GPD has a location parameter μ, we have $G_{\xi,\mu,\sigma}(x)$ and it is equivalent to $G_{\xi,\sigma}(x-\mu)$ with the support $x > \mu$. The distribution can be classified into three types depending on the shape parameter ξ; $G_{\xi,\sigma}$ is heavy-tailed when $\xi > 0$, medium-tailed when $\xi = 0$, and short-tailed when $\xi < 0$. Note that some authors use the opposite sign of ξ; del Castillo and Serra (2015) and Zhang (2010) are in this case.

2.2. Peaks over threshold (POT)

Let $X_1 > X_2 > \cdots > X_N$ be a sequence of independent and identically distributed random variables with common continuous distribution function F and n be the number of observations over a threshold u. Under the POT framework, our interest is in the n excess values, not all of the observations. For u, the excess distribution over the threshold is defined as

$$F_u(x) = P(X-u \leq x \mid X > u) = \frac{F(x+u)-F(u)}{1-F(u)}, \quad 0 \leq x \leq x_0 - u \tag{1}$$