Computational Statistics and Data Analysis 53 (2009) 3835–3841

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A bootstrap goodness of fit test for the generalized Pareto distribution

José A. Villaseñor-Alva *, Elizabeth González-Estrada

Colegio de Postgraduados, Mexico

ARTICLE INFO

Article history:
Received 30 September 2008
Received in revised form 26 February 2009
Accepted 5 April 2009
Available online 14 April 2009

Keywords:
Intersection-union tests
Parametric bootstrap
Parameter estimation
Asymptotic maximum likelihood estimation
Ozone data

ABSTRACT

This paper proposes a bootstrap goodness of fit test for the Generalized Pareto distribution (GPd) with shape parameter γ. The proposed test is an intersection–union test which tests separately the cases of $\gamma \geq 0$ and $\gamma < 0$ and rejects if both cases are rejected. If the test does not reject, then it is known whether the shape parameter γ is either positive or negative. A Monte Carlo simulation experiment was conducted to assess the power of performance of the intersection–union test. The GPd hypothesis was tested on a data set containing Mexico City's ozone levels.[1]

1. Introduction

The Generalized Pareto distribution (GPd) is the name for a rich family of distributions with distribution function given by

$$F(x;\sigma,\gamma) = 1 - \left(1 + \frac{\gamma}{\sigma}x\right)^{-1/\gamma}, \tag{1}$$

where $\sigma > 0$, and $\gamma \in R$ such that $x > 0$ for $\gamma \geq 0$ and $0 < x < -\sigma/\gamma$ for $\gamma < 0$.

Notice that as $\gamma \to 0$, $F(x;\sigma,\gamma) \to 1 - \exp(-x/\sigma)$, which is the Exponential(σ) distribution. Also notice that when $\gamma = -1$, $F(x;\sigma,\gamma) = x/\sigma$, which is the Uniform$(0,\sigma)$ distribution.

This family of distributions contains heavy tail distributions, the exponential family of distributions as well as a subclass of Beta distributions and others with bounded support.

Due to its properties, the GPd has been used to model probabilities in different fields like finance, environmental sciences and hydrology, among others (Reiss and Thomas, 2001).

For establishing the plausibility of the GPd for fitting a given random sample, there exist arguments in terms of the asymptotic theory of exceedance distributions over a threshold (Pickands, 1975), as well as some data exploratory methods like the use of the sample mean residual life function (Reiss and Thomas, 2001). However, these techniques do not provide a measure of error for accepting the GPd as a model for the random sample when in fact it is not the right model. Therefore, in this case an efficient goodness of fit test is needed.

In applications, due to the Pickands (1975) result, the random sample for testing the GPd hypothesis typically comprises excesses over a given threshold, which are non-negative observations.

There have been some partial solutions for providing such a test. See Meintanis and Bassiakos (2007) and Choulakian and Stephens (2001). These proposals rely on the assumption that either maximum likelihood or moment estimates do exist; however, in this case these estimates do not exist for the whole parameter space.

* Corresponding address: Colegio de Postgraduados, Montecillo, 56230 Texcoco, Mexico. Tel.: +52 5558045900x1419; fax: +52 5558045949.
E-mail addresses: jvillasr@colpos.mx (J.A. Villaseñor-Alva), egonzalez@colpos.mx (E. González-Estrada).

1 The R code for implementing the proposed test can be obtained from the authors.

0167-9473/$ – see front matter
doi:10.1016/j.csda.2009.04.001

Let $\bar{F}(x) = 1 - F(x)$. The definition of a GP distribution given in (1) is equivalent to

$$\left\{\bar{F}(x;\sigma,\gamma)\right\}^{-\gamma} = 1 + \frac{\gamma}{\sigma}x, \quad \sigma > 0, \gamma \in R. \tag{2}$$

Furthermore, adding -1 and taking logarithms on both sides of (2), we have

$$\log\left(\left(\bar{F}(x;\sigma,\gamma)\right)^{-\gamma} - 1\right) = \log\left(\frac{\gamma}{\sigma}\right) + \log(x), \quad \sigma > 0, \gamma \in R. \tag{3}$$

In this article we propose an intersection–union goodness of fit test for the GPd with unknown parameters, which is based on Eqs. (2) and (3) and the sample correlation coefficient as a measure of linearity. A parametric bootstrap is used to obtain the critical values.

This paper is organized as follows. In Section 2, we consider two estimation methods for estimating the shape parameter γ. In Section 3, we present the intersection–union goodness of fit test for GPd. In Section 4, we assess the power of the proposed test against several alternatives by using Monte Carlo simulation. In Section 5, we test the GPd hypothesis on a real data set containing excesses over a threshold of Mexico City's ozone levels.

2. Parameter estimation

The most usual methods for estimating the parameters of a GPd are the maximum likelihood (ML), method of moments, and probability weighted moments (Chaouche and Bacro, 2006) approaches. The ML approach gives the most accurate results (Hosking and Wallis, 1987); however, there are some computational problems as regards obtaining ML estimates and the usual approaches can lead to unreliable estimates (Chaouche and Bacro, 2006).

In this section we propose two methods for estimating the shape parameter γ, depending on whether $\gamma \geq 0$ or $\gamma < 0$.

2.1. Asymptotic maximum likelihood method: Case $\gamma \geq 0$

Let $X_1, X_2, \ldots, X_n$ be iid random variables with distribution function G and let $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ be the corresponding order statistics.

If V is an extreme distribution then we say that G belongs to the attraction domain of V (denoted by $G \in D(V)$) if there exist constants $a_n > 0$ and b_n such that

$$\lim_{n\to\infty} G^n(a_n x + b_n) = V(x),$$

for all continuity points x of V.

Notice that $G^n(a_n x + b_n) = P\left(\frac{X_{(n)} - b_n}{a_n} \leq x\right)$. Therefore, as $n \to \infty$, $X_{(n)}$ is approximately distributed as $V\left(\frac{x - b_n}{a_n}\right)$.

V can be one of the following three types of distributions (Gnedenko, 1943):

$\Lambda(x) = \exp\left\{-e^{-x}\right\}, x \in \Re$.

$\Phi(x) = \exp\left\{-x^{-\alpha}\right\}, x \geq 0, \alpha > 0$.

$\Psi(x) = \exp\left\{-(-x)^{-\alpha}\right\}, x < 0, \alpha > 0$.

There exist characterizations for each $D(V)$ (Galambos, 1987).

When $G(x)$ is a GP distribution $F(x;\sigma,\gamma)$, $\gamma > 0$, $\sigma > 0$, if $X \sim G$ and $W = \log X$ then, as $n \to \infty$,

$$G_W^n(a_n x + b_n) \to \Lambda(x), \quad x \in \Re,$$

where $a_n = \gamma$, $b_n = \log(\sigma n^\gamma/\gamma)$ and G_W denotes the distribution function of W. That is, $G_W \in D(\Lambda)$.

Therefore, it is convenient to estimate the parameters on the basis of the upper k order statistics

$$W_j = \log X_{(j)}, \quad j = n-k+1, n-k+2, \ldots, n.$$

The estimation method that we propose for $\gamma > 0$ is based on the following asymptotic result (Reiss, 1989):

If $G_W \in D(V)$ with constants $a_n > 0$ and b_n, then the k-dimensional vector

$$\left(\frac{W_{n-k+1} - b_n}{a_n}, \frac{W_{n-k+2} - b_n}{a_n}, \ldots, \frac{W_n - b_n}{a_n}\right)$$

converges in distribution to the extreme variable of dimension k, $\eta = (\eta_1, \eta_2, \ldots, \eta_k)$, with density function

$$\nu^{(k)}(x_1, x_2, \ldots, x_k) = V(x_k)\prod_{j=1}^{k}\frac{\nu(x_{k-j+1})}{V(x_{k-j+1})},$$

where ν is the density function of V.

Therefore, the joint density of $(W_{n-k+1}, W_{n-k+2}, \ldots, W_n)$ evaluated at $(x_1, x_2, \ldots, x_k)$ can be approximated by the extreme density $\nu^{(k)}$ at the point $\left(\frac{x_1 - b_n}{a_n}, \frac{x_2 - b_n}{a_n}, \ldots, \frac{x_k - b_n}{a_n}\right)$.