Journal of Statistical Planning and Inference 174 (2016) 11–19

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

A goodness-of-fit test for heavy tailed distributions with unknown parameters and its application to simulated precipitation extremes in the Euro-Mediterranean region

G. Jogesh Babu [a], Andrea Toreti [b,*]

[a] The Pennsylvania State University, United States
[b] European Commission, Joint Research Centre, Italy

ARTICLE INFO

Article history:
Received 9 January 2015
Received in revised form 16 October 2015
Accepted 3 February 2016
Available online 10 February 2016

Keywords:
Anderson–Darling
Extremes
Goodness-of-fit
Bootstrap

ABSTRACT

We establish a general bootstrap procedure combined with a modified Anderson–Darling statistic. This procedure is proved to be valid for heavy tailed generalized Pareto distributions that are commonly used to model excesses over a high threshold in extreme value theory. Then, the method is applied to daily precipitation excesses simulated over the Euro-Mediterranean region in autumn by four regional climate models from the EURO-CORDEX initiative.

1. Introduction

In many fields, e.g. climate sciences, there is an increasing need of modeling extreme values. The natural statistical framework to perform such task is the extreme value theory—EVT (de Haan and Ferreira, 2006; Reiss and Thomas, 2007) that is mainly based on the Fisher–Tippett theorem. Under some regularity conditions, this theorem states that the distribution of the maximum of m i.i.d. random variables converges to a distribution belonging to a specific parametric family: the generalized extreme value (GEV). Based on this result, a similar limiting theorem for excesses over a high threshold holds. In this case, under general regularity conditions, Balkema, de Haan and Pickands (Balkema and de Haan, 1974; Pickands, 1975) established that the limiting distribution belongs to the generalized Pareto (hereafter GP) family composed of three sub-families of distributions: Pareto, Exponential, Beta. A generic distribution belonging to the GP family, can be written as:

$$G_{\sigma,\xi}(x) = \begin{cases} 1 - \left(1 + \dfrac{\xi x}{\sigma}\right)^{-\frac{1}{\xi}} & \xi \neq 0 \\ 1 - \exp(-\dfrac{x}{\sigma}) & \xi = 0 \end{cases} \tag{1}$$

for $\sigma > 0$ and for $x > 0$ when $\xi \geq 0$ and $x \leq -\frac{\sigma}{\xi}$ when $\xi < 0$. Several methods have been developed and proposed to estimate the two parameters controlling the GP distribution, e.g.: maximum likelihood (Smith, 1985), generalized probability weighted moments (Diebolt et al., 2007). Nevertheless, the inference with small samples (especially of ξ) remains difficult as well as testing the convergence condition on which the model relies. Thus, assessing the goodness-of-fit of such

[*] Corresponding author.
E-mail address: andrea.toreti@jrc.ec.europa.eu (A. Toreti).

http://dx.doi.org/10.1016/j.jspi.2016.02.002
0378-3758/

a model in applications to real data can be important. To address this issue, Choulakian and Stephens (2001) proposed tests based on the Cramér–von Mises and the Anderson–Darling statistics both for known and unknown parameters of the GP distribution. However, the former gives equal weight to all observations while the latter gives more weight to both tails. Therefore, when the interest is on heavy tailed distributions (i.e., GP with $\xi > 0$), a modification is needed. With this respect, a modified Anderson–Darling statistic (hereafter MADA) was proposed by Ahmad et al. (1988):

$$A_n = n\int_{-\infty}^{\infty}[F(x) - E_n(x)]^2 \cdot [1 - F(x)]^{-1}dx \tag{2}$$

where n denotes the sample size, F is the theoretical distribution and E_n is the empirical distribution function. However, when the parameters of F are not known and estimated, the asymptotic distribution of A_n (and the critical values for the goodness-of-fit test) is unknown too.

In this paper, we establish a valid general bootstrap procedure for goodness of fit for modified Anderson–Darling statistic under some general conditions on hazard function. The method is also valid for the heavy tailed GP family, as applied in previous studies (Toreti et al., 2013). Then, we apply the test to characterize daily precipitation extremes in autumn over the Euro-Mediterranean region simulated by a set of (recently released) regional climate models in the frame of the EURO-CORDEX initiative (Jacob et al., 2014). The achievement of a better understanding and characterization of precipitation extremes is very important due to the high impacts of these events on human and natural systems (IPCC, 2012), and this is especially true in a climate change context. Furthermore, a potential increase of vulnerability and exposure to climate extremes further enhances this importance. Concerning the Euro-Mediterranean region, its complexity in terms of topography, atmospheric processes, etc. (Lionello et al., 2012) is well reflected in the estimated and observed climate extremes over the region (Ulbrich et al., 2012; Toreti et al., 2010).

In the following section we establish a valid bootstrap procedure for goodness of fit for modified Anderson–Darling statistic under some general conditions on hazard function. The third section is focused on a simulation study, while the fourth one is devoted to the climate analysis and the last one on conclusions.

2. The bootstrap approach

The procedure (and the associated proof) to be combined with MADA builds on the work of Babu and Rao (2004). Let $\mathcal{F} = \{F(\cdot;\theta),\ \theta \in \Theta\}$ be a family of continuous distribution functions with Θ being an open region in a p-dimensional Euclidean space. For instance, the family of GP distributions with positive shape parameter, $\theta = (\sigma, \xi)$ and $\Theta = (0,\infty)\times(0,\infty)$. Then, let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables from a distribution F. The aim is to test $F = F(\cdot;\theta)$ for some $\theta = \theta_0 \in \Theta$ by using the MADA statistics, which is based on the empirical processes $Y_n(x;\theta) = \sqrt{n}\,[F(x) - E_n(x)]$. As soon as an estimator of θ is available (i.e., $\hat\theta_n$), n i.i.d. samples $X_1^\star, X_2^\star, \ldots, X_n^\star$ can be generated according to $F(\cdot;\hat\theta_n)$. Then, the same estimator of the first step can be used to get $\hat\theta_n^\star$ from $X_1^\star, X_2^\star, \ldots, X_n^\star$. Thus, this approach can be applied to obtain the critical levels of the statistic if we show that (under some specific conditions) $\int_{-\infty}^{\infty} Y_n^2(x;\hat\theta_n^\star)[1 - F(x;\hat\theta_n^\star)]^{-1}\,dF(x;\hat\theta_n^\star)$ with $Y_n(x;\hat\theta_n^\star) = \sqrt{n}\,[F(x;\hat\theta_n^\star) - E_n^\star(x)]$ converges for almost all sample sequences to the same limiting distribution of $\int_{-\infty}^{\infty} Y_n^2(x;\hat\theta_n)[1 - F(x;\hat\theta_n)]^{-1}\,dF(x;\hat\theta_n)$ with $Y_n(x;\hat\theta_n) = \sqrt{n}\,[F(x;\hat\theta_n) - E_n(x)]$.

To achieve this objective we need some technical results and the assumptions listed in the Appendix. Given $\theta_0 \in \Theta$ and $\Lambda \subset \Theta$ the closure of a given neighborhood of θ_0, suppose $\{\theta_n\}$ is a sequence in Λ converging to θ_0 as $n \to \infty$. Let $X_{1,n}, \ldots, X_{n,n}$ be i.i.d. random variables from the distribution $F(\cdot;\theta_n)$. Let $\mathbb{P}_{\theta_n}$ denote the probability measure induced by $X_{1,n}, \ldots, X_{n,n}$ and let E_n denote the empirical distribution of these random variables. Suppose $\hat\theta_n$ is an estimator of θ_n, we can just start by stating the following theorem of Babu and Rao (2004). See Appendix for assumptions.

Theorem 2.1 (*Theorem 4.1, in Babu and Rao, 2004*). *Suppose $\theta_n \to \theta_0$, assumption* (A1) *holds, and*

$$\hat\theta_n - \theta_n = \frac{1}{n}\sum_{i=1}^{n}\ell(X_{i,n};\theta_n) + \frac{1}{\sqrt{n}}\epsilon_n, \tag{3}$$

for a score function ℓ satisfying the assumptions (A2)–(A5), *where $\epsilon_n \to 0$ in P_{θ_n}-probability. If $L(\theta_n) \to L(\theta_0)$, then the process Y_n given by*

$$Y_n(x;\hat\theta_n) = \sqrt{n}\big(E_n(x) - F(x;\hat\theta_n)\big)$$

converges weakly to a centered, $\mathbb{E}\{Y(x)\} = 0$, Gaussian process Y, where $L(\theta)$ is defined in the Appendix (see A3*).*

From this theorem and assuming conditions (E) and (P) of Appendix to be valid, it follows that for almost all sample sequences the processes $Y(\cdot,\hat\theta_n^*)$ and $Y(\cdot,\hat\theta_n)$ converge weakly to the same limiting centered Gaussian process Y. Now, let $\lambda(\cdot;\theta)$ denote the hazard function of $F(\cdot;\theta)$ *i.e.*,

$$\lambda(x;\theta) = \frac{f(x;\theta)}{1 - F(x;\theta)},$$

where $f(\cdot;\theta)$ denotes the density function of $F(\cdot;\theta)$.