Computational Statistics and Data Analysis 85 (2015) 84–99

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Bayesian threshold selection for extremal models using measures of surprise

J. Lee [a], Y. Fan [b], S.A. Sisson [b,*]

[a] School of Computing & Mathematical Sciences, Auckland University of Technology, Auckland, New Zealand

[b] School of Mathematics and Statistics, University of New South Wales, Sydney, Australia

ARTICLE INFO

Article history:
Received 13 November 2013
Received in revised form 8 December 2014
Accepted 9 December 2014
Available online 16 December 2014

Keywords:
Bayesian inference
Extremes
Generalised Pareto distribution
Posterior predictive *p*-value
Spectral density function
Surprise
Threshold selection

ABSTRACT

Statistical extreme value theory is concerned with the use of asymptotically motivated models to describe the extreme values of a process. A number of commonly used models are valid for observed data that exceed some high threshold. However, in practice a suitable threshold is unknown and must be determined for each analysis. While there are many threshold selection methods for univariate extremes, there are relatively few that can be applied in the multivariate setting. In addition, there are only a few Bayesian-based methods, which are naturally attractive in the modelling of extremes due to data scarcity. The use of Bayesian measures of surprise to determine suitable thresholds for extreme value models is proposed. Such measures quantify the level of support for the proposed extremal model and threshold, without the need to specify any model alternatives. This approach is easily implemented for both univariate and multivariate extremes.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Extreme value theory is often used for the modelling of rare events in many applied areas, including finance (Embrechts et al., 2003), engineering (Castillo et al., 2004) and the environmental sciences (Coles, 2001). Commonly, a mathematically derived parametric extreme value model is used to describe the tail of the data generation process above some high threshold.

In the univariate case, the generalised Pareto distribution provides a suitable model for the analysis of threshold exceedances, under mild conditions (Pickands, 1975; Balkema and de Haan, 1974; Davison and Smith, 1990). Specifically, if $X_1, X_2, \ldots \in \mathbb{R}$ denote a sequence of independent and identically distributed random variables, then the asymptotic distribution of the exceedances, $Y = X - u | X > u$, of some high threshold u is given by

$$F(y|\xi, \sigma, u) = 1 - \left[1 + \frac{\xi(y-u)}{\sigma}\right]_+^{-1/\xi}, \tag{1}$$

where $[a]_+ = \max\{0, a\}$, and $\sigma > 0$ and $-\infty < \xi < \infty$ denote scale and shape parameters. The generalised Pareto model (1) holds as $u \to \infty$, and so in practice a suitable choice of threshold is the smallest value, u, such that F approximates the tail of the observed data sufficiently well.

* Corresponding author. Tel.: +61 2 9385 7027; fax: +61 2 9385 7123.
E-mail address: Scott.Sisson@unsw.edu.au (S.A. Sisson).

http://dx.doi.org/10.1016/j.csda.2014.12.004
0167-9473/© 2014 Elsevier B.V. All rights reserved.

In the multivariate setting, a standard representation is given in terms of a limiting Poisson process (de Haan, 1985; Resnick, 1987). If $Z_1, \ldots, Z_n \in \mathbb{R}^d$ are an independent and identically distributed sequence of random vectors with unit Fréchet margins (i.e. with distribution function $\exp(-1/z)$, for $z > 0$), then the sequence of point processes $P_n = \{Z_i/n : i = 1, \ldots, n\}$ on $[0, \infty)^d$ converges to a non-homogeneous Poisson process $P_n \to P$ on $[0, \infty)^d \setminus \{0\}$ as $n \to \infty$ (de Haan, 1985). The intensity function of P has the form

$$\nu(dz) = \frac{dr}{r^2} H(dw), \tag{2}$$

where (r, w) denotes the pseudopolar co-ordinates $r = \frac{1}{d}\sum_{i=1}^{d} Z^i$ and $w = Z/r$ (where $Z = (Z^1, \ldots, Z^d) \in \mathbb{R}^d$), and H is a measure function defined on the unit simplex which represents the multivariate dependence structure. As with the univariate case the above Poisson process holds asymptotically, and so in practice (that is, for finite n) it is assumed to hold approximately on regions bounded away from the origin. In this case, the Poisson process with intensity function (2) may be fitted to those observations $(r, w)|r > r_0$ for which r exceeds some high threshold, r_0. As before, a suitable choice of threshold is the smallest value, r_0, such that the above Poisson process approximates the multivariate tails of the observed data sufficiently well (e.g. Coles and Tawn, 1991).

In both univariate and multivariate settings, the choice of a suitable threshold (u or r_0) is problem dependent. As such, a number of approaches have been proposed, primarily for the univariate case, that either offer diagnostics for threshold choice or estimate the threshold as part of the model fitting procedure. A comprehensive review of these methods for the choice of u is given by Scarrott and MacDonald (2012), who loosely characterise the techniques into several categories.

Classical fixed threshold approaches use graphical or other diagnostics to make an assessment of the model fit, in order to make an a priori threshold choice. These include e.g. mean residual life plots, threshold stability plots, Hill plots and general distribution fit diagnostics (e.g. Davison and Smith, 1990; Beirlant et al., 1996; Dupuis, 1998; Drees et al., 2000; Coles, 2001; Choulakian and Stephens, 2001). Disadvantages to these approaches are that graphical diagnostics are sometimes difficult to correctly interpret, and that the uncertainty associated with the threshold, u, is not well accounted for in the frequentist framework although see Cabras and Castellanos (2009b) who develop a Bayesian mean residual life plot. Some methods that have been proposed to informally overcome these problems include tail fraction estimation (Drees et al., 2000; Dekkers, 1993; Feuerverger and Hall, 1999; Goegebeur et al., 2008) and resampling-based approaches (Danielsson et al., 2001; Ferreira et al., 2003; Beirlant et al., 1996; Drees and Kaufmann, 1998).

Rather than making an a priori threshold choice, several Bayesian mixture models have been proposed which treat the threshold as an unknown parameter to be estimated. The mixture components themselves correspond to the generalised Pareto model above u, and parametric or semi-/non-parametric estimators of the bulk of the distribution below the threshold (Frigessi et al., 2003; Behrens et al., 2004; Tancredi et al., 2006; Cabras and Morales, 2007; Cabras and Castellanos, 2011; MacDonald et al., 2011b). These approaches are attractive as they both incorporate threshold uncertainty in the analysis, and also remove the need to make a subjective decision on the value of a fixed threshold. Disadvantages of these approaches are the need to correctly balance the relative influence of the bulk and Pareto mixture components so that neither dominate (MacDonald et al., 2011a), and that it would appear difficult to extend them to the multivariate setting.

All of the above approaches concern the Pareto threshold, u, for univariate extremes. There are virtually no diagnostics to determine the threshold, r_0, for multivariate extremes models. However, noting that the intensity measure (2) is expected to factorise into independent components involving angular (w) and radial (r) components when $r > r_0$, in principle diagnostics may be constructed by determining the smallest value of r_0, such that $r|r > r_0$ and $w|r > r_0$ exhibit independence. Intuitively, this is easiest to achieve for bivariate extremes, so that both w and r are univariate, whereby empirical histograms of $w|r > r_0$ should visually retain the same shape for $r > r_0$ (e.g. Joe et al., 1992; Coles and Tawn, 1994).

In this article we propose a new Bayesian diagnostic for threshold choice for extremal models based on the idea of "surprise" (Meng, 1994; Bayarri and Morales, 2003; Bayarri and Berger, 1998; Cabras and Morales, 2007). Measures of surprise quantify the degree of incompatibility of observed data with a given model, commonly through various (Bayesian) predictive p-values using appropriate test statistics, but without any reference to alternative models. In terms of threshold identification, such measures would enumerate the extent to which observed data exceeding a candidate threshold (u or r_0) are compatible with the asymptotic Pareto or point process model. The smallest threshold values that are not incompatible with the data are then natural candidates for the selected threshold.

Unlike many existing threshold choice methods, as predictive p-values have a natural scale, these measures of surprise allow direct comparison of competing threshold candidates (which have different amounts of data exceeding the threshold), as they do not require any modelling of data below the threshold. Also unlike almost all existing methods, by construction, this approach is equally applicable to both univariate and multivariate extremal models. Ultimately, the proposed surprise-based approach will select a final fixed threshold, u or r_0, for use in a subsequent analysis. As a result, threshold uncertainty is not directly incorporated into this final analysis. However, the threshold selection procedure itself is fully Bayesian, and the final choice of threshold can be made within a full Bayesian decision-theoretic framework.

The remainder of this article is organised as follows: Section 2 provides a brief introduction to measures of surprise and various forms of predictive p-value, before describing the proposed threshold selection procedure. The performance of this approach is evaluated through several simulated examples in Section 3, both univariate and multivariate, and compared to existing approaches for threshold choice. In Section 4 we apply our procedure to several real examples that have been previously analysed in the extremes literature. Finally, we conclude with a discussion.

Download English Version:

https://daneshyari.com/en/article/417435

Download Persian Version:

https://daneshyari.com/article/417435

Daneshyari.com