Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 6252–6268
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

A robust estimator for the tail index of Pareto-type distributions

B. Vandewalle[a, b], J. Beirlant[c,*], A. Christmann[d], M. Hubert[c]

[a]Department of Mathematics, Katholieke Universiteit Leuven, Belgium
[b]Center for Statistics and Applications, Universidade de Lisboa, Portugal
[c]Department of Mathematics and University Center of Statistics, W. de Croylaan 54, 3000-Leuven, Katholieke Universiteit Leuven, Belgium
[d]Department of Mathematics, Vrije Universiteit Brussel, Belgium

Received 16 December 2005; received in revised form 8 January 2007; accepted 8 January 2007
Available online 16 January 2007

Abstract

In extreme value statistics, the extreme value index is a well-known parameter to measure the tail heaviness of a distribution. Pareto-type distributions, with strictly positive extreme value index (or tail index) are considered. The most prominent extreme value methods are constructed on efficient maximum likelihood estimators based on specific parametric models which are fitted to excesses over large thresholds. Maximum likelihood estimators however are often not very robust, which makes them sensitive to few particular observations. Even in extreme value statistics, where the most extreme data usually receive most attention, this can constitute a serious problem. The problem is illustrated on a real data set from geopedology, in which a few abnormal soil measurements highly influence the estimates of the tail index. In order to overcome this problem, a robust estimator of the tail index is proposed, by combining a refinement of the Pareto approximation for the conditional distribution of relative excesses over a large threshold with an integrated squared error approach on partial density component estimation. It is shown that the influence function of this newly proposed estimator is bounded and through several simulations it is illustrated that it performs reasonably well at contaminated as well as uncontaminated data.

Keywords: Extreme value statistics; Relative excesses over a large threshold; Robust tail index estimation; Geopedology

1. Introduction

In extreme value statistics emphasis lies on the modelling of rare events, mostly events with a low frequency but a high impact. Common practice is to characterize the size and frequency of such extreme events mainly by the extreme value index γ. The most prominent estimators of this real-valued parameter are maximum likelihood estimators of specific parametric models which are fitted to excesses over large thresholds (see Smith, 1987). These maximum likelihood estimators often prove to be highly efficient, though non-robust against deviations of the actual distribution from the assumed parametric model. This is for instance the case in the presence of outliers or suspicious data, where the performance of the maximum likelihood estimators and the quality of the corresponding estimates of the tail index are often seriously affected.

* Corresponding author. Tel.: +32 16322831; fax: +32 16322789.
E-mail address: Jan.Beirlant@wis.kuleuven.be (J. Beirlant).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.01.003

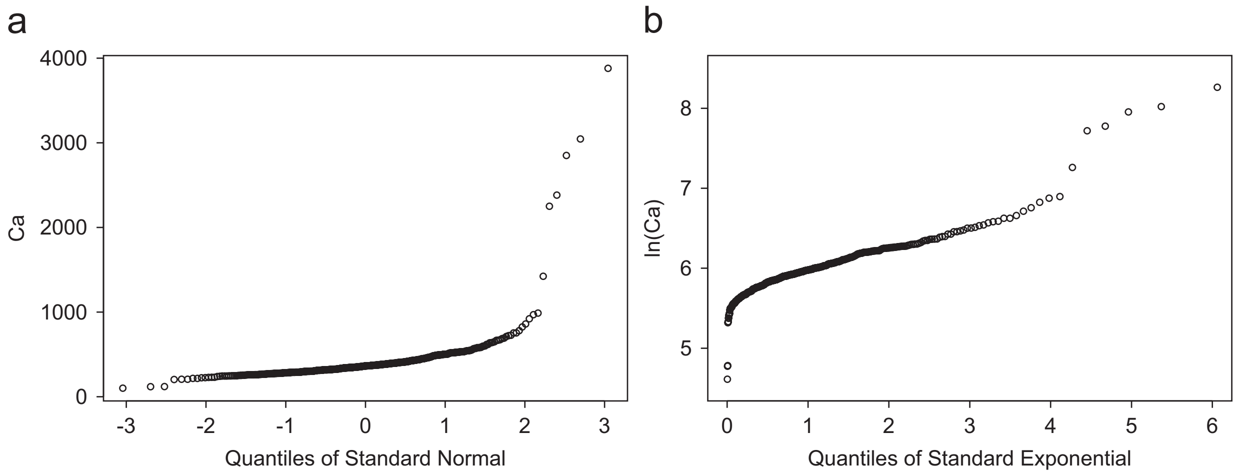

Fig. 1. (a) Normal quantile plot and (b) Pareto quantile plot of Ca for one of the communities of the Condroz region in Belgium.

As shown in Brazauskas and Serfling (2000a), small errors in the estimation of the tail index can already produce large errors in the estimation of quantities based on the tail index γ. Hence, some robust algorithms replacing the maximum likelihood part of this methodology, can be of use, leading to an alternative set of outlier-resistant estimators of the tail index γ and quantities derived from it such as extreme quantiles, small exceedance probabilities and mean excess functions, which are often used in hydrology, meteorology, pedology, structural engineering, economics and actuarial science (e.g. as principles on which to set reinsurance premiums).

In this paper, we study soil data from the Condroz region in Belgium, gathered by the Unit of Geopedology at Gembloux Agricultural University. For more information, see Laroche and Oger (1999) and Goegebeur et al. (2005). In agriculture, a new concept of crop management has emerged, permitting within-field variation of crop techniques as for instance the adjustment of fertilizer inputs on the basis of soil sampling and analysis. The development of this precision farming has drastically increased the demand for soil data and laboratories are now burdened with large data sets, inevitably causing concern about outliers and their influence on the quality of the information. Therefore, robust estimation procedures have become a necessity in the database management of soil data in order to provide high quality information.

Our analysis is limited to calcium (Ca) records from one of the communities in the Condroz region (428 observations, see Goegebeur et al., 2005). The Ca distribution at higher pH-levels appears to be right-skewed and long-tailed, resulting in rather frequent large Ca measurements, as can be seen in the normal quantile plot of Ca given in Fig. 1(a). As a result, robust statistical procedures which assume that the regular data points are sampled from a normal distribution will flag too many large observations as outliers. Such long tailed data can be analyzed more efficiently in the framework of extreme value theory. Nevertheless, as will be seen later on, even in the background of a heavy tailed model some observations will still appear to be suspicious.

Robust methods for extreme values have already been discussed in recent literature. For example, Brazauskas and Serfling (2000b) consider robust estimation in the context of strict Pareto distributions. Dupuis and Field (1998), respectively, Peng and Welsh (2001) and Juárez and Schucany (2004), derived robust estimation methods for the case where the observations follow a generalized extreme value distribution, respectively, a generalized Pareto distribution, light or heavy tailed. In this paper, we propose a new robust tail index estimation procedure for the semi-parametric setting of Pareto-type (or heavy-tailed) distributions. So here, the strict Pareto distribution is assumed to hold only asymptotically, i.e. for excess distributions over high enough threshold values.

In Section 2, we give a brief introduction to univariate extreme value statistics and some well-known estimators of the tail index γ for Pareto-type distributions. In Section 3, a new estimator for γ is introduced, based on a recently developed refinement of the Pareto approximation for the conditional distribution of relative excesses over a large threshold (Beirlant et al., 2004b) in combination with an integrated squared error approach on partial density component (PDC) estimation (Scott, 2004). We investigate its robustness properties in Section 4 by means of the influence function, in the context of heavy-tailed models contaminated with outliers. Throughout, the discussed methods are applied to the