Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2006) 904–917
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Fitting the generalized Pareto distribution to data using maximum goodness-of-fit estimators

Alberto Luceño*

E.T.S. de Ingenieros de Caminos, University of Cantabria, 39005 Santander, Spain

Received 15 January 2005; received in revised form 12 July 2005; accepted 20 September 2005
Available online 12 October 2005

Abstract

Some of the most powerful techniques currently available to test the goodness of fit of a hypothesized continuous cumulative distribution function (CDF) use statistics based on the empirical distribution function (EDF), such as those of Kolmogorov, Cramer–von Mises and Anderson–Darling, among others. The use of EDF statistics was analyzed for estimation purposes. In this approach, maximum goodness-of-fit estimators (also called minimum distance estimators) of the parameters of the CDF can be obtained by minimizing any of the EDF statistics with respect to the unknown parameters. The results showed that there is no unique EDF statistic that can be considered most efficient for all situations. Consequently, the possibility of defining new EDF statistics is entertained; in particular, an Anderson–Darling statistic of degree two and one-sided Anderson–Darling statistics of degree one and two appear to be notable in some situations. The procedure is shown to be able to deal successfully with the estimation of the parameters of homogeneous and heterogeneous generalized Pareto distributions, even when maximum likelihood and other estimation methods fail.

Keywords: Anderson–Darling statistic; Cramer–von Mises statistic; Empirical distribution function; Generalized linear models; Kolmogorov distance; Minimum distance estimator

1. Introduction

The maximum likelihood (ML) estimation method is known to be asymptotically optimal to estimate the parameters of many discrete and continuous distributions. However, there is a considerable number of continuous distributions for which the probability density function—and, subsequently, the likelihood function—can be made arbitrarily large at some point, and hence maximum likelihood estimators (MLEs) of the parameters of these distributions do not generally exist. Moreover, if the range of the distribution depends on unknown parameters, MLEs may not possess their classical statistical properties because Cramer's regularity conditions often fail to hold in this situation. There are also some distributions such that the likelihood function may not have a local maximum for some sample values so that no MLE exists. When the ML method cannot be used, an alternative estimation method is the method of moments (MOM). However, there is also a considerable number of distributions for which some of the first few moments are not finite, with the consequence that moment estimates do not exist. Even when the moments exist, the MOM method may produce inefficient estimators. In addition, the MOM method cannot be used in the context of generalized linear models.

* Tel.: +34 942 201 725; fax: +34 942 201 703.
E-mail address: lucenoa@unican.es.

0167-9473/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2005.09.011

An interesting example of a distribution that poses these difficulties is provided by the generalized Pareto CDF given by

$$F_{\theta,k}(x) = \begin{cases} 1-(1-kx/\theta)^{1/k} & \text{if } k \neq 0, \\ 1-\exp(-x/\theta) & \text{if } k = 0, \end{cases} \tag{1}$$

where $\theta > 0$ is a scale parameter and k is a shape parameter. The range of x is $x \geqslant 0$ for $k \leqslant 0$ and $0 \leqslant x \leqslant \theta/k$ for $k > 0$. When $k > 1$, MLEs do not exist because the probability density corresponding to (1) tends to infinity when x tends to θ/k. Moreover, Cramer's regularity conditions do not hold for $k > \frac{1}{3}$. The mean and variance are $\mu = \theta/(1+k)$ and $\sigma^2 = \theta^2/\left\{(1+k)^2(1+2k)\right\}$, so that μ and σ^2 are finite only for $k > -1$ and $k > -\frac{1}{2}$, respectively. Consequently, moments estimators do not exist for $k \leqslant -\frac{1}{2}$. The problem of fitting the generalized Pareto distribution (GPD) to data has been approached by several authors including Hosking et al. (1985), Hosking and Wallis (1987), Davison and Smith (1990), Walshaw (1990), Grimshaw (1993), Castillo and Hadi (1997), and Castillo et al. (2005), among others. Goodness-of-fit tests for the GPD have been suggested by Choulakian and Stephens (2001). The GPD is also important because it contains the exponential distribution with mean θ as a limiting case when k tends to 0, the uniform distribution in the range $[0, \theta]$ when $k = 1$, and the standard Pareto distribution when $k < 0$. Moreover, its relevance has recently increased considerably (see Appendix A) because—as shown by Pickands (1975)—it can be put in connection with the generalized extreme value distribution (GEVD)having CDF

$$F_{\mu,\psi,k}(x) = \begin{cases} \exp\left\{-(1-k(x-\mu)/\psi)^{1/k}\right\} & \text{if } k \neq 0, \\ \exp[-\exp\{-(x-\mu)/\psi)\}] & \text{if } k = 0. \end{cases} \tag{2}$$

In this paper we analyze a method for estimating the parameters of continuous CDFs, which is based on minimizing empirical distribution function (EDF) statistics and can be used as an alternative or as a complement to other estimation methods. Because EDF statistics are used to test the goodness of fit of continuous distributions, we call this method the maximum goodness of fit (MGF) estimation method. The estimators provided by the MGF method will be called maximum goodness-of-fit estimators (MGFEs).

The origin of the MGF method goes back to Wolfowitz (1953, 1957) and Kac et al. (1955), under the name of minimum distance method. Moreover, Pollard (1980) proved the $\sqrt{n}$-consistency of the minimum distance estimators and found its asymptotic distribution. Because the name "minimum distance method" is often used in other contexts not related to EDF statistics (for example, when minimizing functions of the sample and population autocorrelations of the residuals in time series, or the distance between sample and predicted moments), we prefer to use the name MGF throughout the paper. One important property of the MGF method is that it can be used in situations in which there are no MOM or ML estimators. In contrast with the MOM or ML methods which lead to unique estimators, the MGF method provides several estimators depending on the particular EDF statistic chosen, thus providing a wider inductive basis. For instance, one particular EDF statistic could provide more weight to the left tail of the distribution, whereas a second EDF statistic could assign more weight to the right tail or to the central part of the distribution, and a third statistic could assign equal weight to every part of the distribution. Even though the MGF method seems to have been disregarded as a useful estimation method to fit the GPD to data (see, e.g., Castillo et al., 2005; Coles, 2001; Smith, 2003), we shall show throughout the paper that the MGF method can be successfully used to estimate the parameters of the GPD (and of generalized linear models based on the GPD) even for very extreme values of the shape parameter.

Section 2 compiles the classical EDF statistics used throughout the paper together with some new EDF statistics that are useful for estimation purposes. Section 3 describes the MGF estimation method. The performance of MGF estimators is analyzed for homogeneous GPDs in Section 4 and for generalized linear models based on GPDs in Section 5; a real example of application to ocean engineering is also considered in Section 4. Concluding remarks are given in Section 6.

2. Some EDF statistics useful for estimation

Let $(x_1, \ldots, x_n)$ be a sample of n IID observations on a continuous random variable X with CDF $F(x)$. Let $x_{(1)} \leqslant \cdots \leqslant x_{(n)}$ be the corresponding order statistics and $S_n(x)$ be the empirical distribution function (see Rao, 1973).