Available online at www.sciencedirect.com

Physica A 370 (2006) 49–53

www.elsevier.com/locate/physa

The power-law tail exponent of income distributions

F. Clementi[a,b,*], T. Di Matteo[b], M. Gallegati[c]

[a]*Department of Public Economics, University of Rome "La Sapienza", Via del Castro Laurenziano 9, 00161 Rome, Italy*
[b]*Applied Mathematics, Research School of Physical Sciences and Engineering, The Australian National University, 0200 Canberra, Australia*
[c]*Department of Economics, Università Politecnica delle Marche, Piazzale Martelli 8, 60121 Ancona, Italy*

Available online 12 May 2006

Abstract

In this paper we tackle the problem of estimating the power-law tail exponent of income distributions by using the Hill's estimator. A subsample semi-parametric bootstrap procedure minimizing the mean squared error is used to choose the power-law cutoff value optimally. This technique is applied to personal income data for Australia and Italy.

Keywords: Personal income; Pareto's index; Hill's estimator; Bootstrap

1. Introduction

Since Pareto it has been recognized that a *power-law* provides a good fit for the distribution of high incomes [1]. The Pareto's law asserts that the complementary cumulative distribution $P_>(y) = 1 - \int_{-\infty}^{y} p(\xi)\,\mathrm{d}\xi \to P_>(u)(u/y)^{\alpha}$, with $y \geqslant u$, where $u > 0$ is the threshold value of the distribution and $\alpha > 0$ turns out to be some kind of index of inequality of distribution. The fit of such distribution is usually performed by judging the degree of linearity in a double logarithmic plot involving the empirical and theoretical distribution functions, in such a way that the estimation of u of the distribution does not seem to follow a neutral procedure. Moreover, recent studies have criticized the reliability of this geometrical method by showing that linear-fit based methods for estimating the power-law exponent tend to provide biased estimates, while the maximum likelihood estimation method produces more accurate and robust estimates [2,3]. Hill proposed a conditional maximum likelihood estimator for α based on the k largest order statistics for non-negative data with a Pareto's tail [4]. That is, if $y_{[n]} \geqslant y_{[n-1]} \geqslant \cdots \geqslant y_{[n-k]} \geqslant \cdots \geqslant y_{[1]}$, with $y_{[i]}$ denoting the ith order statistic, are the sample elements put in descending order, then the Hill's estimator is

$$\hat{\alpha}_n(k) = \left[\frac{1}{k}\sum_{i=1}^{k}(\log y_{n-i+1} - \log y_{n-k})\right]^{-1}, \tag{1}$$

[*] Corresponding author. Department of Public Economics, University of Rome "La Sapienza", Via del Castro Laurenziano 9, 00161 Rome, Italy. Tel.: +39 06 49766843; fax: +39 06 4461964.
E-mail address: fabio.clementi@uniroma1.it (F. Clementi).

0378-4371/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2006.04.027

where n is the sample size and k an integer value in $[1, n]$. Unfortunately, the finite-sample properties of the estimator (Eq. (1)) depend crucially on the choice of k: increasing k reduces the variance because more data are used, but it increases the bias because the power-law is assumed to hold only in the extreme tail.

Over the last 20 years, estimation of the Pareto's index has received considerable attention in extreme value statistics [5]. All of the proposed estimators, including the Hill's estimator, are based on the assumption that the number of observations in the upper tail to be included, k, is known. In practice, k is unknown; therefore, the first task is to identify which values are really extreme values. Tools from exploratory data analysis, as the quantile-quantile plot and/or the mean excess plot, might prove helpful in detecting graphically the quantile $y_{[n-k]}$ above which the Pareto's relationship is valid; however, they do not propose any formal computable method and, imposing an arbitrary threshold, they only give very rough estimates of the range of extreme values.

Given the bias-variance *trade-off* for the Hill's estimator, a general and formal approach in determining the best k value is the minimization of the *Mean Squared Error* (*MSE*) between $\hat{\alpha}_n(k)$ and the theoretical value α. Unfortunately, in empirical studies of data the theoretical value of α is not known. Therefore, an attempt to find an approximation to the sampling distribution of the Hill's estimator is required. To this end, a number of innovative techniques in the statistical analysis of extreme values proposes to adopt the powerful bootstrap tool to find the optimal number of order statistics adaptively [6–9]. By capitalizing on these recent advances in the extreme value statistics literature, in this paper we adopt a subsample semi-parametric bootstrap algorithm in order to make a reasonable and more automated selection of the extreme quantiles useful for studying the upper tail of income distributions and to end up at less ambiguous estimates of α. This methodology is described in Section 2 and its application to Australian and Italian income data [10,11] is given in Section 3. Some conclusive remarks are reported in Section 4.

2. Estimation technique for threshold selection

In this section, we consider the problem of finding the optimal threshold u_n^*—or equivalently the optimal number k^* of extreme sample values above that threshold—to be used for estimation of α. In order to achieve this task, we minimize the *MSE* of the Hill's estimator (Eq. (1)) for a series of thresholds $u_n = y_{[n-k]}$, and pick the u_n value at which the *MSE* attains its minimum as u_n^*. Given that different threshold series choices define different sets of possible observations to be included in the upper tail of a specific observed sample $\mathbf{y}_n = \{y_i; i = 1, 2, \ldots, n\}$, only the observations exceeding a certain threshold that are additionally distributed according to a Pareto's cumulative distribution function $PD_{\hat{\alpha}_n(k),u_n}(y)$ are included in the series. In order to check this condition, we perform for each threshold in the original sample a *Kolmogorov–Smirnov* (*K–S*) goodness-of-fit test for the null hypothesis $H_0 : \hat{F}_n(y) = PD_{\hat{\alpha}_n(k),u_n}(y)$ versus the general alternative of the form $H_1 : \hat{F}_n(y) \neq PD_{\hat{\alpha}_n(k),u_n}(y)$, where $\hat{F}_n(y)$ is the empirical distribution function, and $\hat{\alpha}_n(k)$ is a prior estimate for each threshold u_n of the Pareto's tail index obtained through the Hill's statistic. Following the methodology in [12], the formal steps in making a test of H_0 are as follows:

(a) Calculate the original $K–S$ test statistic D by using the formula

$$D = \sup_{-\infty < y < \infty} |\hat{F}_n(y) - PD_{\hat{\alpha}_n(k),u_n}(y)|.$$

(b) Calculate the modified form T^* by using the formula

$$T^* = D\left(\sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}}\right). \tag{2}$$

(c) Reject H_0 if T^* exceeds the cutoff level, z, for the chosen significance level.

To obtain an estimate of finite-sample bias and variance (and thus *MSE*) at each threshold coming from the null hypothesis H_0, a natural criterion is to use the *bootstrap* [13]. In its purest form, the bootstrap involves approximating an unknown distribution function, $F(y)$, by the empirical distribution function, $\hat{F}_n(y)$. However, most times the empirical distribution model from which one resamples in a purely non-parametric bootstrap is not a good approximation of the distribution shape in the tail. Therefore, we initially smooth the tail data by fitting a Pareto's cumulative distribution function

$$PD_{\hat{\alpha}_n(k),u_n}(y) = p = 1 - P_>(u_n)\left(\frac{u_n}{y}\right)^{\hat{\alpha}_n(k)} \tag{3}$$

to the $n_1 \leqslant n$ observations $\mathbf{y}_{n_1} = \{y \in \mathbf{y}_n : T^* \leqslant z\}$, and then use the quantiles $\mathbf{y}_{n_1}^p = \{y \in \mathbf{y}_{n_1} : PD_{\hat{\alpha}_n(k),u_n}(y) \geqslant p\}$ obtained directly from inverting the estimated model (Eq. (3)) to draw the bootstrap samples.

Let us here summarize the adopted methodology:

(1) Evaluate the estimate $\hat{\alpha}_n(k)$ of the Pareto's tail index for each threshold in the original sample $\mathbf{y}_n$ by using the Hill's estimator (Eq. (1)).
(2) For each threshold in the original sample, test the Pareto's approximation by computing the value of the K–S test statistic (Eq. (2)).
(3) Fit the model (Eq. (3)) to the subset of data $\mathbf{y}_{n_1}$ belonging to the null hypothesis H_0.
(4) Select R independent bootstrap samples $\mathbf{y}_1^{\#}, \mathbf{y}_2^{\#}, \ldots, \mathbf{y}_R^{\#}$, each consisting of n_1 values drawn with replacement from the set of quantiles $\mathbf{y}_{n_1}^p$ obtained by inverting the fitted model (Eq. (3)).
(5) For each bootstrap sample $\mathbf{y}_r^{\#}$, $r = 1, 2, \ldots, R$, and for each threshold $u_{n_1}^{\#}$ in the bootstrap sample, evaluate the bootstrap estimate $\hat{\alpha}_{n_1}^{\#}(k_1)$ of the Pareto's tail index by using the Hill's estimator (Eq. (1)).
(6) For each threshold $u_{n_1}^{\#}$, calculate the bias, $B = E[\hat{\alpha}_{n_1}^{\#}(k_1)] - \hat{\alpha}_n(k)$,the variance, $Var = E\{[\hat{\alpha}_{n_1}^{\#}(k_1)]^2\} - \{E[\hat{\alpha}_{n_1}^{\#}(k_1)]\}^2$, and the mean squared error, $MSE = B^2 + Var$, of the Hill's tail index estimates.
(7) Select as the optimal threshold $u_n^* = y_{[n-k^*]}$ that threshold where the MSE attains its minimum.

Minimizing the MSE, thus, amounts to find the MSE minimizing number of order statistics $k^* = \arg\min_k MSE$, from which one infers the optimal estimate of the tail index $\hat{\alpha}_n^*(k^*)$.

3. Empirical application: the Australian and Italian personal income distributions

The data sources we use to illustrate how the methodology proposed in Section 2 can be applied to the analysis of income distributions have been selected from the nationally representative cross-sectional data samples of the Australian and Italian household populations. In particular, we have analyzed the total annual income from all sources in the years 1993–1994 to 1996–1997, and then in 1989–1990, 1998–1999, 1999–2000, and 2001–2002 for Australia, and 1977–2002 for Italy [10,11,14]. Here we report only the results in the year 1999–2000 for Australia and 2000 for Italy.

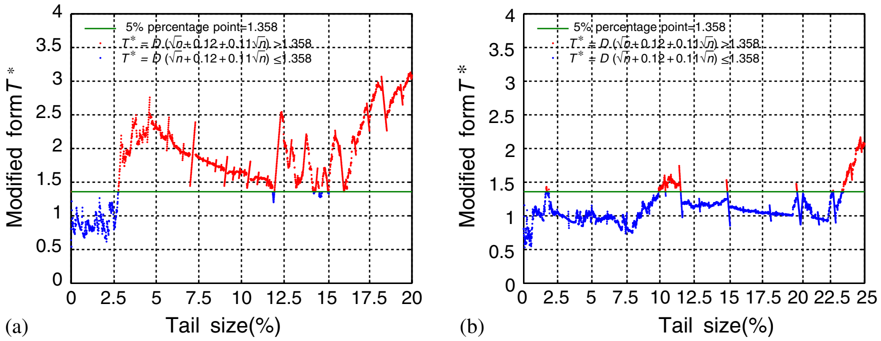

Fig. 1. Modified K–S statistic (Eq. (2)) as a function of the tail size for (a) Australia in 1999–2000 and (b) Italy in 2000.

Figs. 1(a) and (b) depict the outcomes of the complete sequences of K–S test for a selection of tail fractions. Blue points (see on line version) mark all the observations for which the modified K–S statistic (Eq. (2)) does not exceed the 5% cutoff level $z = 1.358$ (solid lines in the figures). The 5% significance point $z = 1.358$ comes from Table *1A* in Ref. [12]. The figures indicate the tail regions that may be tentatively regarded as appropriate for the implementation of the semi-parametric bootstrap technique.

The Hill's estimator (Eq. (1)) is reported in Figs. 2 for Australia (a) and Italy (b), and for tails $\leqslant 20\%$ and $\leqslant 25\%$ of the full sample size, respectively (see solid lines). In these figures, the optimal number of extreme sample values are reported, namely $k^* = 299$ for Australia and $k^* = 3222$ for Italy, providing the following values for the tail power-law exponents: $\hat{\alpha}_n^*(k^*) = 2.3 \pm 0.2$ and $\hat{\alpha}_n^*(k^*) = 2.5 \pm 0.1$, where the errors (with 95% confidence) have been obtained through the *jackknife* method [15]. In these computations, we have used 1000 resamples and the subsample size has been set equal to the number of observations not rejected by the K–S test at the 5% level (see Section 2 and Figs. 1 (a) and (b)). Repeated calculations with a different number of replications produce a spread of tail index estimates with deviations inside the 95% uncertainty band (dashed lines in the figures), showing therefore numerical robustness of our results. We have here obtained more precise values of the power-law tails than the previous one reported in the literature [11].

The use of these $\hat{\alpha}_n^*$ optimal values produces the fits shown by the solid lines in Figs. 3 (a) and (b) for Australia and Italy, where the complementary cumulative distributions are plotted on a log–log scale. The vertical dashed lines indicate the optimal values of the threshold parameter attained by subsample semi-parametric bootstrapping: (a) $u_n^* = \$\,82367$ for Australia in 1999–2000 and (b) $u_n^* = €19655$ for Italy in 2000. As we can see, our procedure succeeds in avoiding deviations from linearity for the largest observations that

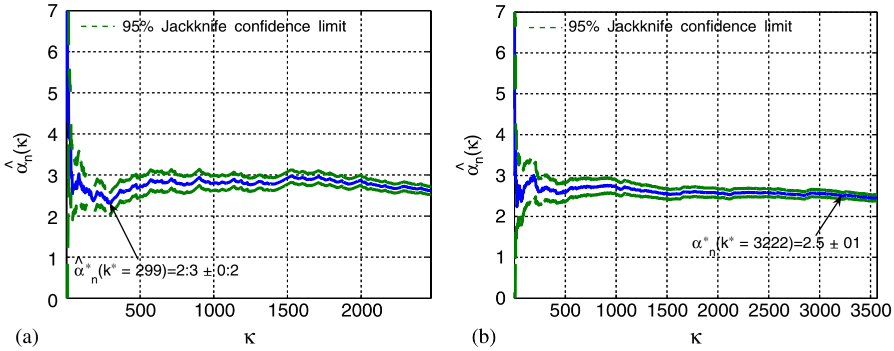

Fig. 2. The Hill's estimator (Eq. (1)) for (a) Australia in 1999–2000 and (b) Italy in 2000. The dashed lines represent the 95% confidence limits of the tail index estimates computed by using the jackknife method. The arrows mark the optimal number of extreme sample values k^*.

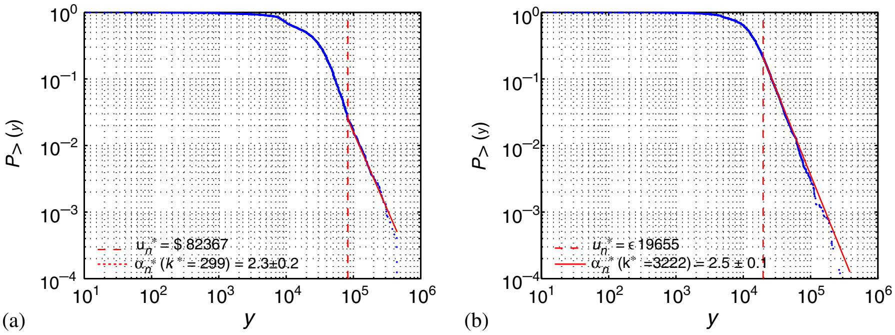

Fig. 3. Complementary cumulative distribution (a) for Australia in 1999–2000 and (b) for Italy in 2000 and power-law fits by using the estimated optimal values for α.

might strongly influence the estimation of α, illustrating therefore the importance of optimally choosing the tail threshold.

4. Concluding remarks

In this paper, we have considered the problem of the estimation of the power-law tail exponent of income distributions and we have adopted a subsample semi-parametric bootstrap procedure in order to arrive at less ambiguous estimates of α. This methodology has been empirically applied to the estimation of personal income distribution data for Australia and Italy. The reliability and robustness of the results have been tested by running different repeated bootstrap replications and comparing the variability of the estimates through a jackknife method.

From the economic point of view, this technique for the estimation of the Pareto's tail index of income distribution is expected to allow a deeper understanding of both the way in which cyclical fluctuations in economic activity affect factor income shares and the channels through which these effects work through the size distribution of income, which are issues of relevance for the modeling of the income process in the high-end tail of the distribution.

Acknowledgments

T. Di Matteo wishes to thank the Australian Social Science Data Archive, ANU, for providing the ABS data and the partial support by ARC Discovery Projects: DP03440044 (2003) and DP0558183 (2005), COST P10 "Physics of Risk" project and M.I.U.R.-F.I.S.R. Project "Ultra-high frequency dynamics of financial markets".

References

[1] V. Pareto, Course d'économie politique, Macmillan, London, 1897.
[2] M.L. Goldstein, S.A. Morris, G.G. Yen, Eur. Phys. J. B 41 (2004) 255–258.
[3] H.F. Coronel-Brizio, A.R. Hernández-Montoya, Physica A 354 (2005) 437–449.
[4] B.M. Hill, Ann. Stat. 3 (1975) 1163–1174.
[5] T. Lux, Appl. Financial Econ. 11 (2001) 299–315.
[6] P. Hall, J. Multivar. Anal. 32 (1990) 177–203.
[7] M.M. Dacorogna, U.A. Müller, O.V. Pictet, C.G. De Vries, The distribution of extremal foreign exchange rate returns in extremely large data sets, Internal document (BPB. 1992-10-22), Olsen & Associates, Zürich, 1992.
[8] J. Danielsson, L. De Haan, L. Peng, C.G. De Vries, J. Multivariate Anal. 76 (2001) 226–248.
[9] T. Lux, Empirical Econ. 25 (2000) 641–652.
[10] T. Di Matteo, T. Aste, S.T. Hyde, Exchanges in complex networks: income and wealth distributions, in: F. Mallamace, H.E. Stanley (Eds.), The Physics of Complex Systems (New Advances and Perspectives), IOS Press, Amsterdam, 2004, pp. 435–442.
[11] F. Clementi, M. Gallegati, Physica A 350 (2005) 427–438.
[12] M.A. Stephens, J. Am. Stat. Assoc. 69 (1974) 730–737.
[13] B. Efron, Ann. Stat. 7 (1979) 1–26.
[14] F. Clementi, T. Di Matteo, M. Gallegati, 2006, in preparation.
[15] O.V. Pictet, M.M. Dacorogna, U.A. Müller, Hill, Bootstrap and Jackknife estimators for heavy tails, Internal document (BPB. 1996-12-10), Olsen & Associates, Zürich, 1996.