View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by CWI's Institutional Repository

Accepted Manuscript

Gini estimation under infinite variance

Andrea Fontanari, Nassim Nicholas Taleb, Pasquale Cirillo

PII: S0378-4371(18)30189-4
DOI: https://doi.org/10.1016/j.physa.2018.02.102
Reference: PHYSA 19222

To appear in: *Physica A*

Received date : 19 July 2017
Revised date : 21 December 2017

Please cite this article as: A. Fontanari, N.N. Taleb, P. Cirillo, Gini estimation under infinite variance, *Physica A* (2018), https://doi.org/10.1016/j.physa.2018.02.102

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Highlights for review:

- We demonstrate that nonparametric methods are not reliable to estimate the Gini index under fat tails. New Lemma and new proofs for the theorems.
- We show that, under infinite variance, it is preferable to use maximum likelihood based techniques.
- We propose a correction for the nonparametric estimator, when parametric methods cannot be applied.
- The paper contributes, methodologically, to the ongoing discussion on wealth inequality and concentration.
- References have been updated and enriched.

ACCEPTED MANUSCRIPT

Manuscript
Click here to view linked References

Gini estimation under infinite variance

Andrea Fontanari[1] - Delft University of Technology and CWI
Nassim Nicholas Taleb - Tandon School of Engineering, NYU
Pasquale Cirillo[1,2] - Delft University of Technology

Abstract

We study the problems related to the estimation of the Gini index in presence of a fat-tailed data generating process, i.e. one in the stable distribution class with finite mean but infinite variance (i.e. with tail index $\alpha \in (1, 2)$). We show that, in such a case, the Gini coefficient cannot be reliably estimated using conventional nonparametric methods, because of a downward bias that emerges under fat tails. This has important implications for the ongoing discussion about economic inequality.

We start by discussing how the nonparametric estimator of the Gini index undergoes a phase transition in the symmetry structure of its asymptotic distribution, as the data distribution shifts from the domain of attraction of a light-tailed distribution to that of a fat-tailed one, especially in the case of infinite variance. We also show how the nonparametric Gini bias increases with lower values of α. We then prove that maximum likelihood estimation outperforms nonparametric methods, requiring a much smaller sample size to reach efficiency.

Finally, for fat-tailed data, we provide a simple correction mechanism to the small sample bias of the nonparametric estimator based on the distance between the mode and the mean of its asymptotic distribution.

Keywords: Gini index; inequality measure; size distribution; extremes; α-stable distribution.

1. Introduction

Wealth inequality studies represent a field of economics, statistics and econophysics exposed to fat-tailed data generating processes, often with infinite variance [2, 19]. This is not at all surprising if we recall that the prototype of fat-tailed distributions, the Pareto, has been proposed for the first time to model household incomes [22]. However, the fat-tailedness of data can be problematic in the context of wealth studies, as the property of efficiency (and, partially,

[1]These authors gladly acknowledge the generous support of the EU H2020 Marie Sklodowska-Curie Grant Agreement No 643045 WakEUpCall. Pasquale Cirillo also acknowledges the support of the EU Marie Sklodowska-Curie Career Integration Grant Multivariate Shocks (PCIG13-GA-2013-618794).

[2]*Corresponding Author:* P.Cirillo@tudelft.nl. *Address:* Applied Probability Group, EEMCS Faculty, Delft University of Technology, Van Mourik Broekmanweg 6, 2628CD Delft, The Netherlands. *Phone:* +31.15.27.82.589.

ACCEPTED MANUSCRIPT

consistency) does not necessarily hold for many estimators of inequality and concentration [13, 19].

The scope of this work is to show how fat tails affect the estimation of one of the most celebrated measures of economic inequality, the Gini index [9, 17, 30], often used (and abused) in the econophysics and economics literature as the main tool for describing the distribution and the concentration of wealth around the world [2, 3, 23].

The literature concerning the estimation of the Gini index is wide and comprehensive (e.g. [9, 30] for a review), however, strangely enough, almost no attention has been paid to its behavior in presence of fat tails, and this is curious if we consider that: 1) fat tails are ubiquitous in the empirical distributions of income and wealth [19, 23], and 2) the Gini index itself can be seen as a measure of variability and fat-tailedness [8, 10, 11, 15].

The standard method for the estimation of the Gini index is nonparametric: one computes the index from the empirical distribution of the available data using Equation (5) below. But, as we show in this paper, this estimator suffers from a downward bias when we deal with fat-tailed observations. Therefore our goal is to close this gap by deriving the limiting distribution of the nonparametric Gini estimator in presence of fat tails, and propose possible strategies to reduce the bias. We show how the maximum likelihood approach, despite the risk of model misspecification, needs much fewer observations to reach efficiency when compared to a nonparametric one[3].

Our results are relevant to the discussion about wealth inequality, recently rekindled by Thomas Piketty in [23, 24], as the estimation of the Gini index under fat tails and infinite variance may cause several economic analyses to be unreliable, if not markedly wrong. Why should one trust a biased estimator?

By fat-tailed data we indicate those observations generated by a positive random variable X with cumulative distribution function (c.d.f.) $F(x)$, which is regularly-varying of order α [16], that is, for $\bar{F}(x) := 1 - F(x)$, one has

$$\lim_{x\to\infty} x^{\alpha}\bar{F}(x) = L(x), \tag{1}$$

where $L(x)$ is a slowly-varying function such that $\lim_{x\to\infty} \frac{L(cx)}{L(x)} = 1$ with $c > 0$, and where $\alpha > 0$ is called the tail exponent.

Regularly-varying distributions define a large class of random variables whose properties have been extensively studied in the context of extreme value theory [7, 13], when dealing with the probabilistic behavior of maxima and minima. As pointed out in [4], regularly-varying and fat-tailed are indeed synonyms. It is known that, if $X_1, ..., X_n$ are i.i.d. observations with a c.d.f. $F(x)$ in the regularly-varying class, as defined in Equation (1), then their data generating process falls into the maximum domain of attraction of a Fréchet distribu-

[3] A similar bias also affects the nonparametric measurement of quantile contributions, i.e. those of the type "the top 1% owns x% of the total wealth" [27]. This paper extends the problem to the more widespread Gini coefficient, and goes deeper by making links with the limit theorems.

ACCEPTED MANUSCRIPT

tion with parameter ρ, in symbols $X \in MDA(\Phi(\rho))$[7]. This means that, for the partial maximum $M_n = \max(X_1, ..., X_n)$, one has

$$P\left(a_n^{-1}\left(M_n - b_n\right) \leq x\right) \xrightarrow{d} \Phi(\rho) = e^{-x^{-\rho}}, \qquad \rho > 0, \tag{2}$$

with $a_n > 0$ and $b_n \in \mathbb{R}$ two normalizing constants. Clearly, the connection between the regularly-varying coefficient α and the Fréchet distribution parameter ρ is given by: $\alpha = \frac{1}{\rho}$ [13].
The Fréchet distribution is one of the limiting distributions for maxima in extreme value theory, together with the Gumbel and the Weibull; it represents the fat-tailed and unbounded limiting case [7]. The relationship between regularly-varying random variables and the Fréchet class thus allows us to deal with a very large family of random variables (and empirical data), and allows us to show how the Gini index is highly influenced by maxima, i.e. extreme wealth, as clearly suggested by intuition [15, 19], especially under infinite variance. Again, this recommends some caution when discussing economic inequality under fat tails.

It is worth remembering that the existence (finiteness) of the moments for a fat-tailed random variable X depends on the tail exponent α, in fact

$$\begin{aligned} &E(X^\delta) < \infty \text{ if } \delta \leq \alpha, \\ &E(X^\delta) = \infty \text{ if } \delta > \alpha. \end{aligned} \tag{3}$$

In this work, we restrict our focus on data generating processes with finite mean and infinite variance, therefore, according to Equation (3), on the class of regularly-varying distributions with tail index $\alpha \in (1, 2)$.

Table 1 and Figure 1 present numerically and graphically our story, already suggesting its conclusion, on the basis of artificial observations sampled from a Pareto distribution (Equation (13) below) with tail parameter α equal to 1.1.

Table 1 compares the nonparametric Gini index of Equation (5) with the maximum likelihood (ML) tail-based one of Section 3. For the different sample sizes in Table 1, we have generated 10^8 samples, averaging the estimators via Monte Carlo. As the first column shows, the convergence of the nonparametric estimator to the true Gini value ($g = 0.8333$) is extremely slow and monotonically increasing; this suggests an issue not only in the tail structure of the distribution of the nonparametric estimator but also in its symmetry.

Figure 1 provides some numerical evidence that the limiting distribution of the nonparametric Gini index loses its properties of normality and symmetry [14], shifting towards a skewed and fatter-tailed limit, when data are characterized by an infinite variance. As we prove in Section 2, when the data generating process is in the domain of attraction of a fat-tailed distribution, the asymptotic distribution of the Gini index becomes a skewed-to-the-right α-stable law. This change of behavior is responsible of the downward bias of the nonparametric Gini under fat tails. However, the knowledge of the new limit allows us to propose a correction for the nonparametric estimator, improving its quality,

ACCEPTED MANUSCRIPT

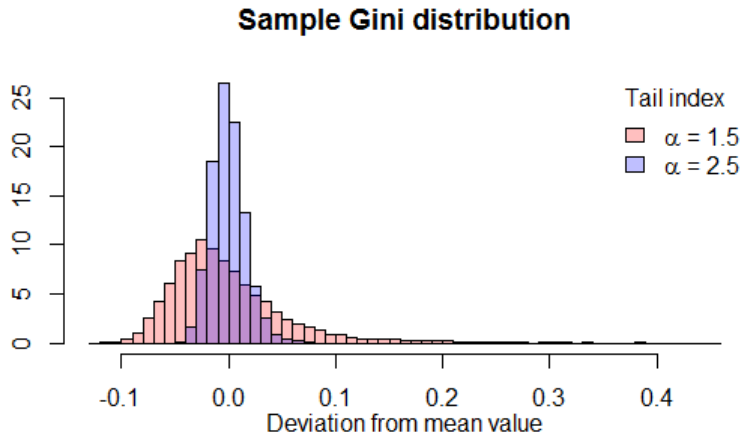


Figure 1: Histograms for the Gini nonparametric estimators for two Paretian (type I) distributions with different tail indices, with finite and infinite variance (plots have been centered to ease comparison). Sample size: 10^3. Number of samples: 10^2 for each distribution.

and thus reducing the risk of badly estimating wealth inequality, with all the possible consequences in terms of economic and social policies [19, 23, 24].

Table 1: Comparison of the Nonparametric (NonPar) and the Maximum Likelihood (ML) Gini estimators, using Paretian data with tail $\alpha = 1.1$ (finite mean, infinite variance) and different sample sizes. Number of Monte Carlo simulations: 10^8. The Error Ratio in the last column is defined as ratio of the mean absolute deviation of the nonparametric estimator over that of the maximum likelihood one.

n	Nonpar		ML		Error Ratio
(number of obs.)	*Mean*	*Bias*	*Mean*	*Bias*	
10^3	*0.711*	*-0.122*	*0.8333*	*0*	1.4
10^4	*0.750*	*-0.083*	*0.8333*	*0*	3
10^5	*0.775*	*-0.058*	*0.8333*	*0*	6.6
10^6	0.790	-0.043	0.8333	0	156
10^7	0.802	-0.031	0.8333	0	10^5+

The rest of the paper is organized as follows: in Section 2 we derive the asymptotic distribution of the sample Gini index when data possess an infinite variance; in Section 3 we deal with the maximum likelihood estimator; in Section 4 we provide an illustration with Paretian observations; in Section 5 we propose a simple correction based on the mode-mean distance of the asymptotic distribution of the nonparametric estimator, to take care of its small-sample bias; finally, Section 6 closes the paper. To ease readability, a technical Appendix contains the longer proofs of the main results in the work.

ACCEPTED MANUSCRIPT

2. Asymptotics of the nonparametric estimator under infinite variance

We now derive the asymptotic distribution for the nonparametric estimator of the Gini index when the data generating process is fat-tailed with finite mean but infinite variance.

The so-called stochastic representation of the Gini g is

$$g = \frac{1}{2}\frac{\mathbb{E}\left(|X' - X''|\right)}{\mu} \in [0, 1], \tag{4}$$

where X' and X'' are i.i.d. copies of a random variable X with c.d.f. $F(x) \in [c, \infty)$, $c > 0$, and with finite mean $\mathbb{E}(X) = \mu$. The quantity $\mathbb{E}\left(|X' - X''|\right)$ is known as the "Gini Mean Difference" (GMD) [30]. For later convenience we also define $g = \frac{\theta}{\mu}$ with $\theta = \frac{\mathbb{E}(|X'-X''|)}{2}$.

The Gini index of a random variable X is thus the mean expected deviation between any two independent realizations of X, scaled by twice the mean [12].

The most common nonparametric estimator of the Gini index for a sample $X_1, ..., X_n$ is defined as

$$G^{NP}(X_n) = \frac{\sum_{1 \le i < j \le n} |X_i - X_j|}{(n-1)\sum_{i=1}^{n} X_i}, \tag{5}$$

which can also be expressed as

$$G^{NP}(X_n) = \frac{\sum_{i=1}^{n}\left(2\frac{i-1}{n-1} - 1\right)X_{(i)}}{\sum_{i=1}^{n} X_{(i)}} = \frac{\frac{1}{n}\sum_{i=1}^{n} Z_{(i)}}{\frac{1}{n}\sum_{i=1}^{n} X_i}, \tag{6}$$

where $X_{(1)}, X_{(2)}, ..., X_{(n)}$ are the ordered statistics of $X_1, ..., X_n$, such that: $X_{(1)} < X_{(2)} < ... < X_{(n)}$ and $Z_{(i)} = \left(2\frac{i-1}{n-1} - 1\right)X_{(i)}$. The asymptotic normality of the estimator in Equation (6) under the hypothesis of finite variance for the data generating process is known [19, 30]. The result directly follows from the properties of the U-statistics and the L-estimators involved in Equation (6)

A standard methodology to prove the limiting distribution of the estimator in Equation (6), and more in general of a linear combination of order statistics, is to show that, in the limit for $n \to \infty$, the sequence of order statistics can be approximated by a sequence of i.i.d random variables [5, 20]. However, this usually requires some sort of L^2 integrability of the data generating process, something we are not assuming here. Lemma 1 (proved in the Appendix) shows how to deal with the case of sequences of order statistics generated by fat-tailed L^1-only integrable random variables.

Lemma 1. *Consider the following sequence* $R_n = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{i}{n} - U_{(i)}\right)F^{-1}(U_{(i)})$ *where* $U_{(i)}$ *are the order statistics of a uniformly distributed i.i.d random sample. As-*

ACCEPTED MANUSCRIPT

sume that $F^{-1}(U) \in L^1$. Then the following results hold:

$$R_n \xrightarrow{L^1} 0, \tag{7}$$

and

$$\frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)} R_n \xrightarrow{L^1} 0, \tag{8}$$

with $\alpha \in (1,2)$ and $L_0(n)$ a slowly-varying function.

2.1. A quick recap on α-stable random variables

We here introduce some notation for α-stable distributions, as we need them to study the asymptotic limit of the Gini index.

A random variable X follows an α-stable distribution, in symbols $X \sim S(\alpha, \beta, \gamma, \delta)$, if its characteristic function is

$$E(e^{itX}) = \begin{cases} e^{-\gamma^\alpha |t|^\alpha (1 - i\beta \operatorname{sign}(t)) \tan(\frac{\pi\alpha}{2}) + i\delta t} & \alpha \neq 1 \\ e^{-\gamma|t|(1 + i\beta \frac{2}{\pi} \operatorname{sign}(t)) \ln|t| + i\delta t} & \alpha = 1 \end{cases},$$

where $\alpha \in (0,2)$ governs the tail, $\beta \in [-1,1]$ is the skewness, $\gamma \in \mathbb{R}^+$ is the scale parameter, and $\delta \in \mathbb{R}$ is the location one. This is known as the $S1$ parametrization of α-stable distributions [21, 25].

Interestingly, there is a correspondence between the α parameter of an α-stable random variable, and the α of a regularly-varying random variable as per Equation (1): as shown in [14, 21], a regularly-varying random variable of order α is α-stable, with the same tail coefficient. This is why we do not make any distinction in the use of the α here. Since we aim at dealing with distributions characterized by finite mean but infinite variance, we restrict our focus to $\alpha \in (1,2)$, as the two α's coincide.

Recall that, for $\alpha \in (1,2]$, the expected value of an α-stable random variable X is equal to the location parameter δ, i.e. $\mathbb{E}(X) = \delta$. For more details, we refer to [21, 25].

The standardized α-stable random variable is expressed as

$$S_{\alpha,\beta} \sim S(\alpha, \beta, 1, 0). \tag{9}$$

We note that α-stable distributions are a subclass of infinitely divisible distributions. Thanks to their closure under convolution, they can be used to describe the limiting behavior of (rescaled) partials sums, $S_n = \sum_{i=1}^n X_i$, in the General Central Limit Theorem (GCLT) setting [14]. For $\alpha = 2$ we obtain the normal distribution as a special case, which is the limit distribution for the classical CLTs, under the hypothesis of finite variance.

In what follows we indicate that a random variable is in the domain of attraction of an α-stable distribution, by writing $X \in DA(S_\alpha)$. Just observe that this condition for the limit of partial sums is equivalent to the one given in Equation (2) for the limit of partial maxima [13, 14].

ACCEPTED MANUSCRIPT

2.2. The α-stable asymptotic limit of the Gini index

Consider a sample $X_1, ..., X_n$ of i.i.d. observations with a continuous c.d.f. $F(x)$ in the regularly-varying class, as defined in Equation (1), with tail index $\alpha \in (1, 2)$. The data generating process for the sample is in the domain of attraction of a Fréchet distribution with $\rho \in (\frac{1}{2}, 1)$, given that $\rho = \frac{1}{\alpha}$.

For the asymptotic distribution of the Gini index estimator, as presented in Equation (6), when the data generating process is characterized by an infinite variance, we can make use of the following two theorems: Theorem 2 deals with the limiting distribution of the Gini Mean Difference (the numerator in Equation (6)), while Theorem 3 extends the result to the complete Gini index. Proofs for both theorems are in the Appendix.

Theorem 2. *Consider a sequence $(X_i)_{1\leq i\leq n}$ of i.i.d random variables from a distribution X on $[c, +\infty)$ with $c > 0$, such that X is in the domain of attraction of an α-stable random variable, $X \in DA(S_\alpha)$, with $\alpha \in (1, 2)$. Then the sample Gini mean deviation (GMD) $\frac{\sum_{i=1}^{n} Z_{(i)}}{n}$ satisfies the following limit in distribution:*

$$\frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)} \left(\frac{1}{n} \sum_{i=1}^{n} Z_{(i)} - \theta \right) \xrightarrow{d} S_{\alpha,1}, \tag{10}$$

where $Z_i = (2F(X_i) - 1)X_i$, $\mathbb{E}(Z_i) = \theta$, $L_0(n)$ is a slowly-varying function such that Equation (37) holds (see the Appendix), and $S_{\alpha,1}$ is a right-skewed standardized α-stable random variable defined as in Equation (9).

Moreover the statistic $\frac{1}{n}\sum_{i=1}^{n} Z_{(i)}$ is an asymptotically consistent estimator for the GMD, i.e. $\frac{1}{n}\sum_{i=1}^{n} Z_{(i)} \xrightarrow{P} \theta$.

Note that Theorem 2 could be restated in terms of the maximum domain of attraction $MDA(\Phi(\rho))$ as defined in Equation (2).

Theorem 3. *Given the same assumptions of Theorem 2, the estimated Gini index $G^{NP}(X_n) = \frac{\sum_{i=1}^{n} Z_{(i)}}{\sum_{i=1}^{n} X_i}$ satisfies the following limit in distribution*

$$\frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)} \left(G^{NP}(X_n) - \frac{\theta}{\mu} \right) \xrightarrow{d} Q, \tag{11}$$

where $\mathbb{E}(Z_i) = \theta$, $\mathbb{E}(X_i) = \mu$, $L_0(n)$ is the same slowly-varying function defined in Theorem 2 and Q is a right-skewed α-stable random variable $S(\alpha, 1, \frac{1}{\mu}, 0)$.

Furthermore the statistic $\frac{\sum_{i=1}^{n} Z_{(i)}}{\sum_{i=1}^{n} X_i}$ is an asymptotically consistent estimator for the Gini index, i.e. $\frac{\sum_{i=1}^{n} Z_{(i)}}{\sum_{i=1}^{n} X_i} \xrightarrow{P} \frac{\theta}{\mu} = g$.

In the case of fat tails with $\alpha \in (1, 2)$, Theorem 3 tells us that the asymptotic distribution of the Gini estimator is always right-skewed notwithstanding the distribution of the underlying data generating process. Therefore heavily

ACCEPTED MANUSCRIPT

fat-tailed data not only induce a fatter-tailed limit for the Gini estimator, but they also change the shape of the limit law, which definitely moves away from the usual symmetric Gaussian. As a consequence, the Gini estimator, whose asymptotic consistency is still guaranteed [20], will approach its true value more slowly, and from below. Some evidence of this was already given in Table 1.

3. The maximum likelihood estimator

Theorem 3 indicates that the usual nonparametric estimator for the Gini index is not the best option when dealing with infinite-variance distributions, due to the skewness and the fatness of its asymptotic limit. The aim is to find estimators that still preserve their asymptotic normality under fat tails, which is not possible with nonparametric methods, as they all fall into the α-stable Central Limit Theorem case [13, 14]. Hence the solution is to use parametric techniques.

Theorem 4 shows how, once a parametric family for the data generating process has been identified, it is possible to estimate the Gini index via MLE. The resulting estimator is not just asymptotically normal, but also asymptotically efficient.

In Theorem 4 we deal with random variables X whose distribution belongs to the large and flexible exponential family, i.e. whose density can be represented as

$$f_\theta(x) = h(x)e^{(\eta(\theta)T(x)-A(\theta))},$$

with $\theta \in \mathbb{R}$, and where $T(x)$, $\eta(\theta)$, $h(x)$, $A(\theta)$ are known functions [26].

Theorem 4. *Let $X \sim F_\theta$ such that F_θ is a distribution belonging to the exponential family. Then the Gini index obtained by plugging-in the maximum likelihood estimator of θ, $G^{ML}(X_n)_\theta$, is asymptotically normal and efficient. Namely:*

$$\sqrt{n}(G^{ML}(X_n)_\theta - g_\theta) \xrightarrow{d} N(0, g_\theta'^2 I^{-1}(\theta)), \tag{12}$$

where $g'_\theta = \frac{dg_\theta}{d\theta}$ and $I(\theta)$ is the Fisher Information.

Proof. The result follows easily from the asymptotic efficiency of the maximum likelihood estimators of the exponential family, and the invariance principle of MLE. In particular, the validity of the invariance principle for the Gini index is granted by the continuity and the monotonicity of g_θ with respect to θ. The asymptotic variance is then obtained by application of the delta-method [26]. □

4. A Paretian illustration

We provide an illustration of the obtained results using some artificial fat-tailed data. We choose a Pareto I [22], with density

$$f(x) = \alpha c^\alpha x^{-\alpha-1}, x \geq c. \tag{13}$$

ACCEPTED MANUSCRIPT

It is easy to verify that the corresponding survival function $\bar{F}(x)$ belongs to the regularly-varying class with tail parameter α and slowly-varying function $L(x) = c^\alpha$. We can therefore apply the results of Section 2 to obtain the following corollaries.

Corollary 1. *Let $X_1, ..., X_n$ be a sequence of i.i.d. observations with Pareto distribution with tail parameter $\alpha \in (1, 2)$. The nonparametric Gini estimator is characterized by the following limit:*

$$D_n^{NP} = G^{NP}(X_n) - g \sim S\left(\alpha, 1, \frac{C_\alpha^{-\frac{1}{\alpha}}}{n^{\frac{\alpha-1}{\alpha}}} \frac{(\alpha-1)}{\alpha}, 0\right). \tag{14}$$

Proof. Without loss of generality we can assume $c = 1$ in Equation (13). The results is a mere application of Theorem 3, remembering that a Pareto distribution is in the domain of attraction of α-stable random variables with slowly-varying function $L(x) = 1$. The sequence c_n to satisfy Equation (37) becomes $c_n = n^{\frac{1}{\alpha}} C_\alpha^{-\frac{1}{\alpha}}$, therefore we have $L_0(n) = C_\alpha^{-\frac{1}{\alpha}}$, which is independent of n. Additionally the mean of the distribution is also a function of α, that is $\mu = \frac{\alpha}{\alpha-1}$. □

Corollary 2. *Let the sample $X_1, ..., X_n$ be distributed as in Corollary 1, let G_θ^{ML} be the maximum likelihood estimator for the Gini index as defined in Theorem 4. Then the MLE Gini estimator, rescaled by its true mean g, has the following limit:*

$$D_n^{ML} = G_\alpha^{ML}(X_n) - g \sim N\left(0, \frac{4\alpha^2}{n(2\alpha-1)^4}\right), \tag{15}$$

where N indicates a Gaussian.

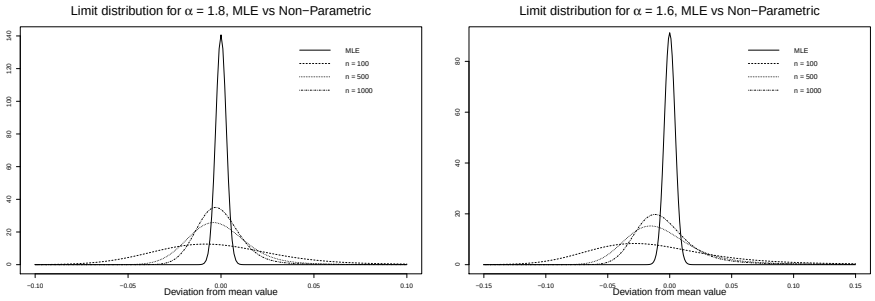
Proof. The functional form of the maximum likelihood estimator for the Gini index is known to be $G_\theta^{ML} = \frac{1}{2\alpha^{ML}-1}$ [19]. The result then follows from the fact that the Pareto distribution (with known minimum value x_m) belongs to an exponential family and therefore satisfies the regularity conditions necessary for the asymptotic normality and efficiency of the maximum likelihood estimator. Also notice that the Fisher information for a Pareto distribution is $\frac{1}{\alpha^2}$. □

Now that we have worked out both asymptotic distributions, we can compare the quality of the convergence for both the MLE and the nonparametric case when dealing with Paretian data, which we use as the prototype for the more general class of fat-tailed observations.

In particular, we can approximate the distributions of the deviations of the estimators from the true value g of the Gini index for finite sample sizes, by using Equations (14) and (15).

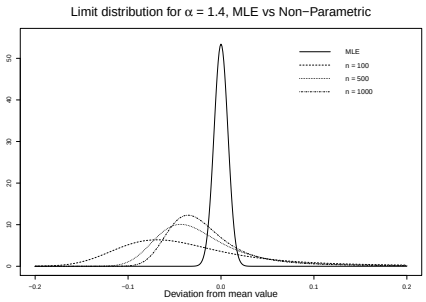
Figure 2 shows how the deviations around the mean of the two different types of estimators are distributed and how these distributions change as the number of observations increases. In particular, to facilitate the comparison

ACCEPTED MANUSCRIPT

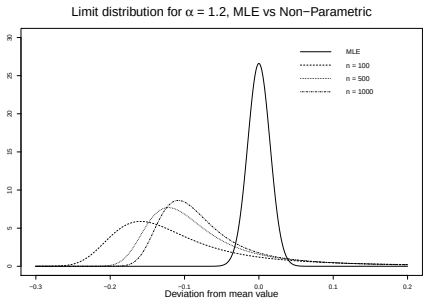


(a) $\alpha = 1.8$

(b) $\alpha = 1.6$



(c) $\alpha = 1.4$



(d) $\alpha = 1.2$

Figure 2: Comparisons between the maximum likelihood and the nonparametric asymptotic distributions for different values of the tail index α. The number of observations for MLE is fixed to $n = 100$. Note that, even if all distributions have mean zero, the mode of the distributions of the nonparametric estimator is different from zero, because of the skewness.

between the maximum likelihood and the nonparametric estimators, we fixed the number of observation in the MLE case, while letting them vary in the nonparametric one. We perform this study for different types of tail indices to show how large the impact is on the consistency of the estimator. It is worth noticing that, as the tail index decreases towards 1 (the threshold value for a infinite mean), the mode of the distribution of the nonparametric estimator moves farther away from the mean of the distribution (centered on 0 by definition, given that we are dealing with deviations from the mean). This effect is responsible for the small sample bias observed in applications. Such a phenomenon is not present in the MLE case, thanks to the the normality of the limit for every value of the tail parameter.

We can make our argument more rigorous by assessing the number of observations $\tilde{n}$ needed for the nonparametric estimator to be as good as the MLE one, under different tail scenarios. Let's consider the likelihood-ratio-type

ACCEPTED MANUSCRIPT

function

$$r(c,n) = \frac{P_S(|D_n^{NP}| > c)}{P_N(|D_{100}^{ML}| > c)}, \tag{16}$$

where $P_S(|D_n^{NP}| > c)$ and $P_N(|D_{100}^{ML}| > c)$ are the probabilities (α-stable and Gaussian respectively) of the centered estimators in the nonparametric, and in the MLE cases, of exceeding the thresholds $\pm c$, as per Equations (14) and (15). In the nonparametric case the number of observations n is allowed to change, while in the MLE case it is fixed to 100. We then look for the value $\tilde{n}$ such that $r(c,\tilde{n}) = 1$ for fixed c. Table 2 displays the results for different thresholds c and tail parameters α. In particular, we can see how the MLE estimator outperforms the nonparametric one, which requires a much larger number of observations to obtain the same tail probability of the MLE with n fixed to 100. For example, we need at least 80×10^6 observations for the nonparametric estimator to obtain the same probability of exceeding the ± 0.02 threshold of the MLE one, when $\alpha = 1.2$.

Table 2: The number of observations $\tilde{n}$ needed for the nonparametric estimator to match the tail probabilities, for different threshold values c and different values of the tail index α, of the maximum likelihood estimator with fixed $n = 100$.

	Threshold c as per Equation (16)*:*			
α	0.005	0.01	0.015	0.02
1.8	27×10^3	12×10^5	12×10^6	63×10^5
1.5	21×10^4	21×10^4	46×10^5	81×10^7
1.2	33×10^8	67×10^7	20×10^7	80×10^6

Interestingly, the number of observations needed to match the tail probabilities in Equation (16) does not vary uniformly with the threshold. This is expected, since as the threshold goes to infinity or to zero, the tail probabilities remain the same for every value of n. Therefore, given the unimodality of the limit distributions, we expect that there will be a threshold maximizing the number of observations needed to match the tail probabilities, while for all the other levels the number of observations will be smaller.

We conclude that, when in presence of fat-tailed data with infinite variance, a plug-in MLE based estimator should be preferred over the nonparametric one.

5. Small sample correction

Theorem 3 can be also used to provide a correction for the bias of the nonparametric estimator for small sample sizes. The key idea is to recognize that, for unimodal distributions, most observations come from around the mode. In symmetric distributions the mode and the mean coincide, thus most observations will be close to the mean value as well, not so for skewed distributions: for right-skewed continuous unimodal distributions the mode is lower than the mean. Therefore, given that the asymptotic distribution of the nonparametric

ACCEPTED MANUSCRIPT

Gini index is right-skewed, we expect that the observed value of the Gini index will be usually lower than the true one (placed at the mean level). We can quantify this difference (i.e. the bias) by looking at the distance between the mode and the mean, and once this distance is known, we can correct our Gini estimate by adding it back[4].

Formally, we aim to derive a corrected nonparametric estimator $G^C(X_n)$ such that

$$G^C(X_n) = G^{NP}(X_n) + ||m(G^{NP}(X_n)) - \mathbb{E}(G^{NP}(X_n))||, \tag{17}$$

where $||m(G^{NP}(X_n)) - \mathbb{E}(G^{NP}(X_n))||$ is the distance between the mode m and the mean of the distribution of the nonparametric Gini estimator $G^{NP}(X_n)$.

Performing the type of correction described in Equation (17) is equivalent to shifting the distribution of $G^{NP}(X_n)$ in order to place its mode on the true value of the Gini index.

Ideally, we would like to measure this mode-mean distance $||m(G^{NP}(X_n)) - \mathbb{E}(G^{NP}(X_n))||$ on the exact distribution of the Gini index to get the most accurate correction. However, the finite distribution is not always easily derivable as it requires assumptions on the parametric structure of the data generating process (which, in most cases, is unknown for fat-tailed data [19]). We therefore propose to use the limiting distribution for the nonparametric Gini obtained in Section 2 to approximate the finite sample distribution, and to estimate the mode-mean distance with it. This procedure allows for more freedom in the modeling assumptions and potentially decreases the number of parameters to be estimated, given that the limiting distribution only depends on the tail index and the mean of the data, which can be usually assumed to be a function of the tail index itself, as in the Paretian case where $\mu = \frac{\alpha}{\alpha-1}$.

By exploiting the location-scale property of α-stable distributions and Equation (11), we approximate the distribution of $G^{NP}(X_n)$ for finite samples by

$$G^{NP}(X_n) \sim S\left(\alpha, 1, \gamma(n), g\right), \tag{18}$$

where $\gamma(n) = \frac{1}{n^{\frac{\alpha-1}{\alpha}}} \frac{L_0(n)}{\mu}$ is the scale parameter of the limiting distribution.

As a consequence, thanks to the linearity of the mode for α-stable distributions, we have

$$||m(G^{NP}(X_n)) - \mathbb{E}(G^{NP}(X_n))|| \approx ||m(\alpha, \gamma(n)) + g - g|| = ||m(\alpha, \gamma(n))||,$$

where $m(\alpha, \gamma(n))$ is the mode function of an α-stable distribution with zero mean.

The implication is that, in order to obtain the correction term, knowledge of the true Gini index is not necessary, given that $m(\alpha, \gamma(n))$ does not depend

[4]Another idea, which we have tested in writing the paper, is to use the distance between the median and the mean; the performances are comparable.

ACCEPTED MANUSCRIPT

on g. We then estimate the correction term as

$$\hat{m}(\alpha, \gamma(n)) = \arg\max_x s(x), \tag{19}$$

where $s(x)$ is the numerical density of the associated α-stable distribution in Equation (18), but centered on 0. This comes from the fact that, for α-stable distributions, the mode is not available in closed form, but it can be easily computed numerically [21], using the unimodality of the law.

The corrected nonparametric estimator is thus

$$G^C(X_n) = G^{NP}(X_n) + \hat{m}(\alpha, \gamma(n)), \tag{20}$$

whose asymptotic distribution is

$$G^C(X_n) \sim S\left(\alpha, 1, \gamma(n), g + \hat{m}(\alpha, \gamma(n))\right). \tag{21}$$

Note that the correction term $\hat{m}\left(\alpha, \gamma(n)\right)$ is a function of the tail index α and is connected to the sample size n by the scale parameter $\gamma(n)$ of the associated limiting distribution. It is important to point out that $\hat{m}(\alpha, \gamma(n))$ is decreasing in n, and that $\lim_{n\to\infty} \hat{m}(\alpha, \gamma(n)) \to 0$. This happens because, as n increases, the distribution described in Equation (18) becomes more and more centered around its mean value, shrinking to zero the distance between the mode and the mean. This ensures the asymptotic equivalence of the corrected estimator and the nonparametric one. Just observe that

$$\begin{aligned} \lim_{n\to\infty} |G(X_n)^C - G^{NP}(X_n)| &= \lim_{n\to\infty} |G^{NP}(X_n) + \hat{m}(\alpha, \gamma(n)) - G^{NP}(X_n)| \\ &= \lim_{n\to\infty} |\hat{m}(\alpha, \gamma(n))| \to 0. \end{aligned}$$

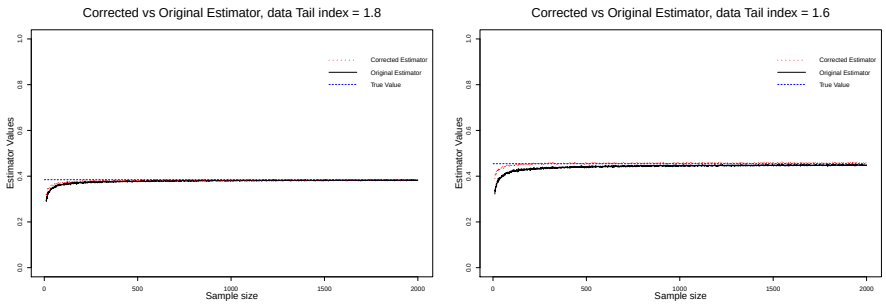
Naturally, thanks to the correction, $G^C(X_n)$ will always behave better in small samples. Consider also that, from Equation (21), the distribution of the corrected estimator has now for mean $g + \hat{m}(\alpha, \gamma(n))$, which converges to the true Gini g as $n \to \infty$.

From a theoretical point of view, the quality of this correction depends on the distance between the exact distribution of $G^{NP}(X_n)$ and its α-stable limit; the closer the two are to each other, the better the approximation. However, given that, in most cases, the exact distribution of $G^{NP}(X_n)$ is unknown, it is not possible to give more details.

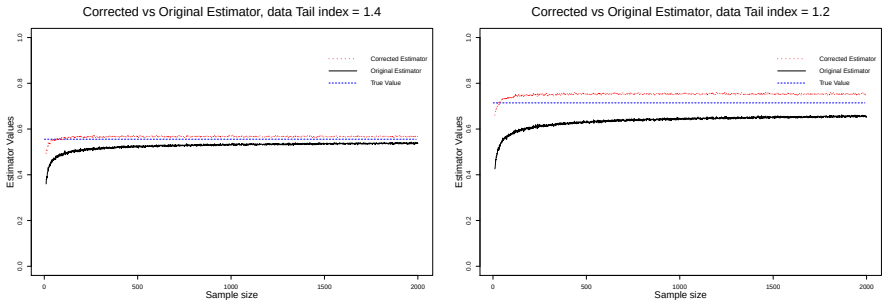
From what we have written so far, it is clear that the correction term depends on the tail index of the data, and possibly also on their mean. These parameters, if not assumed to be known a priori, must be estimated. Therefore the additional uncertainty due to the estimation will reflect also on the quality of the correction.

We conclude this Section with the discussion of the effect of the correction procedure with a simple example. In a Monte Carlo experiment, we simulate 1000 Paretian samples of increasing size, from $n = 10$ to $n = 2000$, and for each

ACCEPTED MANUSCRIPT



(a) $\alpha = 1.8$ (b) $\alpha = 1.6$



(c) $\alpha = 1.4$ (d) $\alpha = 1.2$

Figure 3: Comparisons between the corrected nonparametric estimator (in red, the one on top) and the usual nonparametric estimator (in black, the one below). For small sample sizes the corrected one clearly improves the quality of the estimation.

sample size we compute both the original nonparametric estimator $G^{NP}(X_n)$ and the corrected $G^C(X_n)$. We repeat the experiment for different α's. Figure 3 presents the results.

It is clear that the corrected estimators always perform better than the uncorrected ones in terms of absolute deviation from the true Gini value. In particular, our numerical experiment shows that for small sample sizes with $n \leq 1000$ the gain is quite remarkable for all the different values of $\alpha \in (1, 2)$. However, as expected, the difference between the estimators decreases with the sample size, as the correction term decreases both in n and in the tail index α. Notice that, when the tail index equals 2, we obtain the symmetric Gaussian distribution and the two estimators coincide, given that, thanks to the finiteness of the variance, the nonparametric estimator is no longer biased.

6. Conclusions

In this paper we address the issue of the asymptotic behavior of the nonparametric estimator of the Gini index in presence of a distribution with infi-

ACCEPTED MANUSCRIPT

nite variance, an issue that has been curiously ignored by the literature. The central mistake in the nonparametric methods largely used is to believe that asymptotic consistency translates into equivalent pre-asymptotic properties.

We show that a parametric approach provides better asymptotic results thanks to the properties of maximum likelihood estimation. Hence we strongly suggest that, if the collected data are suspected to be fat-tailed, parametric methods should be preferred.

In situations where a fully parametric approach cannot be used, we propose a simple correction mechanism for the nonparametric estimator based on the distance between the mode and the mean of its asymptotic distribution. Even if the correction works nicely, we suggest caution in its use owing to additional uncertainty from the estimation of the correction term.

Technical Appendix

Proof of Lemma 1

Let $U = F(X)$ be the standard uniformly distributed integral probability transform of the random variable X. For the order statistics, we then have [5]: $X_{(i)} \overset{a.s.}{=} F^{-1}(U_{(i)})$. Hence

$$R_n = \frac{1}{n}\sum_{i=1}^{n}(i/n - U_{(i)})F^{-1}(U_{(i)}). \tag{22}$$

Now by definition of empirical c.d.f it follows that

$$R_n = \frac{1}{n}\sum_{i=1}^{n}(F_n(U_{(i)}) - U_{(i)})F^{-1}(U_{(i)}), \tag{23}$$

where $F_n(u) = \frac{1}{n}\sum_{i=1}^{n} 1_{U_i \leq u}$ is the empirical c.d.f of uniformly distributed random variables.

To show that $R_n \xrightarrow{L^1} 0$, we are going to impose an upper bound that goes to zero. First we notice that

$$\mathbb{E}|R_n| \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}|(F_n(U_{(i)}) - U_{(i)})F^{-1}(U_{(i)})|. \tag{24}$$

To build a bound for the right-hand side (r.h.s) of (24), we can exploit the fact that, while $F^{-1}(U_{(i)})$ might be just L^1-integrable, $F_n(U_{(i)}) - U_{(i)}$ is L^∞ integrable, therefore we can use Hölder's inequality with $q = \infty$ and $p = 1$. It follows that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}|(F_n(U_{(i)}) - U_{(i)})F^{-1}(U_{(i)})| \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\sup_{U_{(i)}}|(F_n(U_{(i)}) - U_{(i)})|\mathbb{E}|F^{-1}(U_{(i)})|. \tag{25}$$

ACCEPTED MANUSCRIPT

Then, thanks to the Cauchy-Schwarz inequality, we get

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\sup_{U_{(i)}}|(F_n(U_{(i)})-U_{(i)})|\mathbb{E}|F^{-1}(U_{(i)})|$$

$$\leq\left(\frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}\sup_{U_{(i)}}|(F_n(U_{(i)})-U_{(i)})|)^2\frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}(F^{-1}(U_{(i)})))^2\right)^{\frac{1}{2}}. \quad (26)$$

Now, first recall that $\sum_{i=1}^{n}F^{-1}(U_{(i)})\overset{a.s.}{=}\sum_{i=1}^{n}F^{-1}(U_i)$ with U_i, $i=1,...,n$, being an i.i.d sequence, then notice that $\mathbb{E}(F^{-1}(U_i))=\mu$, so that the second term of Equation (26) becomes

$$\mu\left(\frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}\sup_{U_{(i)}}|(F_n(U_{(i)})-U_{(i)})|)^2\right)^{\frac{1}{2}}. \quad (27)$$

The final step is to show that Equation (27) goes to zero as $n\to\infty$.

We know that F_n is the empirical c.d.f of uniform random variables. Using the triangular inequality the inner term of Equation (27) can be bounded as

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}\sup_{U_{(i)}}|(F_n(U_{(i)})-U_{(i)})|)^2 \quad (28)$$

$$\leq\frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}\sup_{U_{(i)}}|(F_n(U_{(i)})-F(U_{(i)}))|)^2+\frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}\sup_{U_{(i)}}|(F(U_{(i)})-U_{(i)})|)^2.$$

Since we are dealing with uniforms, we known that $F(U)=u$, and the second term in the r.h.s of (28) vanishes.

We can then bound $\mathbb{E}(\sup_{U_{(i)}}|(F_n(U_{(i)})-F(U_{(i)})|)$ using the so called Vapnik-Chervonenkis (VC) inequality, a uniform bound for empirical processes [1, 6, 28], getting

$$\mathbb{E}\sup_{U_{(i)}}|(F_n(U_{(i)})-F(U_{(i)})|\leq\sqrt{\frac{\log(n+1)+\log(2)}{n}}. \quad (29)$$

Combining Equation (29) with Equation (27) we obtain

$$\mu\left(\frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}\sup_{U_{(i)}}|(F_n(U_{(i)})-U_{(i)})|)^2\right)^{\frac{1}{2}}\leq\mu\sqrt{\frac{\log(n+1)+\log(2)}{n}}, \quad (30)$$

which goes to zero as $n\to\infty$, thus proving the first claim.

For the second claim, it is sufficient to observe that the r.h.s of (30) still goes

ACCEPTED MANUSCRIPT

to zero when multiplied by $\frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)}$ if $\alpha \in (1,2)$.

Proof of Theorem 2

The first part of the proof consists in showing that we can rewrite Equation (10) as a function of i.i.d random variables in place of order statistics, to be able to apply a Central Limit Theorem (CLT) argument.

Let's start by considering the sequence

$$\frac{1}{n}\sum_{i=1}^{n} Z_{(i)} = \frac{1}{n}\sum_{i=1}^{n}\left(2\frac{i-1}{n-1}-1\right)F^{-1}(U_{(i)}). \tag{31}$$

Using the integral probability transform $X \overset{d}{=} F^{-1}(U)$ with U standard uniform, and adding and removing $\frac{1}{n}\sum_{i=1}^{n}\left(2U_{(i)}-1\right)F^{-1}(U_{(i)})$, the r.h.s. in Equation (31) can be rewritten as

$$\frac{1}{n}\sum_{i=1}^{n} Z_{(i)} = \frac{1}{n}\sum_{i=1}^{n}(2U_{(i)}-1)F^{-1}(U_{(i)}) + \frac{1}{n}\sum_{i=1}^{n} 2\left(\frac{i-1}{n-1}-U_{(i)}\right)F^{-1}(U_{(i)}). \tag{32}$$

Then, by using the properties of order statistics [5] we obtain the following almost sure equivalence

$$\frac{1}{n}\sum_{i=1}^{n} Z_{(i)} \overset{a.s.}{=} \frac{1}{n}\sum_{i=1}^{n}(2U_i-1)F^{-1}(U_i) + \frac{1}{n}\sum_{i=1}^{n} 2\left(\frac{i-1}{n-1}-U_{(i)}\right)F^{-1}(U_{(i)}). \tag{33}$$

Note that the first term in the r.h.s of (33) is a function of i.i.d random variables as desired, while the second term is just a reminder, therefore

$$\frac{1}{n}\sum_{i=1}^{n} Z_{(i)} \overset{a.s.}{=} \frac{1}{n}\sum_{i=1}^{n} Z_i + R_n,$$

with $Z_i = (2U_i-1)F^{-1}(U_i)$ and $R_n = \frac{1}{n}\sum_{i=1}^{n} 2\left(\frac{i-1}{n-1}-U_{(i)}\right)F^{-1}(U_{(i)})$.

Given Equation (10) and exploiting the decomposition given in (33) we can rewrite our claim as

$$\frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)}\left(\frac{1}{n}\sum_{i=1}^{n} Z_{(i)}-\theta\right) = \frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i-\theta\right) + \frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)}R_n. \tag{34}$$

From the second claim of the Lemma 1 and Slutsky Theorem, the convergence in Equation (10) can be proven by looking at the behavior of the sequence

$$\frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i-\theta\right), \tag{35}$$

ACCEPTED MANUSCRIPT

where $Z_i = (2U_i - 1)F^{-1}(U_i) = (2F(X_i) - 1)X_i$. This reduces to proving that Z_i is in the fat tails domain of attraction.

Recall that by assumption $X \in DA(S_\alpha)$ with $\alpha \in (1,2)$. This assumption enables us to use a particular type of CLT argument for the convergence of the sum of fat-tailed random variables. However, we first need to prove that $Z \in DA(S_\alpha)$ as well, that is $P(|Z| > z) \sim L(z)z^{-\alpha}$, with $\alpha \in (1,2)$ and $L(z)$ slowly-varying.

Notice that

$$P(|\tilde{Z}| > z) \leq P(|Z| > z) \leq P(2X > z),$$

where $\tilde{Z} = (2U - 1)X$ and $U \perp X$. The first bound holds because of the positive dependence between X and $F(X)$ and it can be proven rigorously by noting that $2UX \leq 2F(X)X$ by the so-called re-arrangement inequality [18]. The upper bound conversely is trivial.

Using the properties of slowly-varying functions, we have $P(2X > z) \sim 2^\alpha L(z)z^{-\alpha}$. To show that $\tilde{Z} \in DA(S_\alpha)$, we use the Breiman's Theorem, which ensure the stability of the α-stable class under product, as long as the second random variable is not too fat-tailed [29].

To apply the Theorem we re-write $P(|\tilde{Z}| > z)$ as

$$P(|\tilde{Z}| > z) = P(\tilde{Z} > z) + P(-\tilde{Z} > z) = P(\tilde{U}X > z) + P(-\tilde{U}X > z),$$

where $\tilde{U}$ is a standard uniform with $\tilde{U} \perp X$.

We focus on $P(\tilde{U}X > z)$ since the procedure is the same for $P(-\tilde{U}X > z)$. We have

$$P(\tilde{U}X > z) = P(\tilde{U}X > z|\tilde{U} > 0)P(\tilde{U} > 0) + P(\tilde{U}X > z|\tilde{U} \leq 0)P(\tilde{U} \leq 0),$$

for $z \to +\infty$.

Now, we have that $P(\tilde{U}X > z|\tilde{U} \leq 0) \to 0$, while, by applying Breiman's Theorem, $P(\tilde{U}X > z|\tilde{U} > 0)$ becomes

$$P(\tilde{U}X > z|\tilde{U} > 0) \to E(\tilde{U}^\alpha|U > 0)P(X > z)P(U > 0).$$

Therefore

$$P(|\tilde{Z}| > z) \to \frac{1}{2}E(\tilde{U}^\alpha|U > 0)P(X > z) + \frac{1}{2}E((-\tilde{U})^\alpha|U \leq 0)P(X > z).$$

From this

$$\begin{aligned} P(|\tilde{Z}| > z) &\to \frac{1}{2}P(X > z)[E(\tilde{U})^\alpha|U > 0) + E((-\tilde{U}^\alpha|U \leq 0)] \\ &= \frac{2^\alpha}{1-\alpha}P(X > z) \sim \frac{2^\alpha}{1-\alpha}L(z)z^{-\alpha}. \end{aligned}$$

ACCEPTED MANUSCRIPT

We can then conclude that, by the squeezing Theorem [14],

$$P(|Z| > z) \sim L(z)z^{-\alpha},$$

as $z \to \infty$. Therefore $Z \in DA(S_\alpha)$.

We are now ready to invoke the Generalized Central Limit Theorem (GCLT)[13] for the sequence Z_i, i.e.

$$nc_n^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}(Z_i)\right) \xrightarrow{d} S_{\alpha,\beta}. \tag{36}$$

with $\mathbb{E}(Z_i) = \theta$, $S_{\alpha,\beta}$ a standardized α-stable random variable, and where c_n is a sequence which must satisfy

$$\lim_{n\to\infty} \frac{nL(c_n)}{c_n^\alpha} = \frac{\Gamma(2-\alpha)|\cos(\frac{\pi\alpha}{2})|}{\alpha - 1} = C_\alpha. \tag{37}$$

Notice that c_n can be represented as $c_n = n^{\frac{1}{\alpha}}L_0(n)$, where $L_0(n)$ is another slowly-varying function possibly different from $L(n)$.

The skewness parameter β is such that

$$\frac{P(Z > z)}{P(|Z| > z)} \to \frac{1+\beta}{2}.$$

Recalling that, by construction, $Z \in [-c, +\infty)$, the above expression reduces to

$$\frac{P(Z > z)}{P(Z > z) + P(-Z > z)} \to \frac{P(Z > z)}{P(Z > z)} = 1 \to \frac{1+\beta}{2}, \tag{38}$$

therefore $\beta = 1$. This, combined with Equation (34), the result for the reminder R_n of Lemma 1 and Slutsky Theorem, allows us to conclude that the same weak limits holds for the ordered sequence of $Z_{(i)}$ in Equation (10) as well.

Proof of Theorem 3

The first step of the proof is to show that the ordered sequence $\frac{\sum_{i=1}^n Z_{(i)}}{\sum_{i=1}^n X_i}$, characterizing the Gini index, is equivalent in distribution to the i.i.d sequence $\frac{\sum_{i=1}^n Z_i}{\sum_{i=1}^n X_i}$. In order to prove this, it is sufficient to apply the factorization in Equation (33) to Equation (11), getting

$$\frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)}\left(\frac{\sum_{i=1}^n Z_i}{\sum_{i=1}^n X_i} - \frac{\theta}{\mu}\right) + \frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)} R_n \frac{n}{\sum_{i=1}^n X_i}. \tag{39}$$

By Lemma 1 and the application of the continuous mapping and Slutsky Theorems, the second term in Equation (39) goes to zero at least in probability.

ACCEPTED MANUSCRIPT

Therefore to prove the claim it is sufficient to derive a weak limit for the following sequence

$$n^{\frac{\alpha-1}{\alpha}} \frac{1}{L_0(n)} \left(\frac{\sum_{i=1}^n Z_i}{\sum_{i=1}^n X_i} - \frac{\theta}{\mu} \right). \tag{40}$$

Expanding Equation (40) and recalling that $Z_i = (2F(X_i) - 1)X_i$, we get

$$\frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)} \frac{n}{\sum_{i=1}^n X_i} \left(\frac{1}{n} \sum_{i=1}^n X_i \left(2F(X_i) - 1 - \frac{\theta}{\mu} \right) \right). \tag{41}$$

The term $\frac{n}{\sum_{i=1}^n X_i}$ in Equation (41) converges in probability to $\frac{1}{\mu}$ by an application of the continuous mapping Theorem, and the fact that we are dealing with positive random variables X. Hence it will contribute to the final limit via Slutsky Theorem.

We first start by focusing on the study of the limit law of the term

$$\frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)} \frac{1}{n} \sum_{i=1}^n X_i \left(2F(X_i) - 1 - \frac{\theta}{\mu} \right). \tag{42}$$

Set $\hat{Z}_i = X_i(2F(X_i) - 1 - \frac{\theta}{\mu})$ and note that $\mathbb{E}(\hat{Z}_i) = 0$, since $\mathbb{E}(Z_i) = \theta$ and $\mathbb{E}(X_i) = \mu$.

In order to apply a GCLT argument to characterize the limit distribution of the sequence $\frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)} \frac{1}{n} \sum_{i=1}^n \hat{Z}_i$ we need to prove that $\hat{Z} \in DA(S_\alpha)$. If so then we can apply GCLT to

$$\frac{n^{\frac{\alpha-1}{\alpha}}}{L_0(n)} \left(\frac{\sum_{i=1}^n \hat{Z}_i}{n} - \mathbb{E}(\hat{Z}_i) \right). \tag{43}$$

Note that, since $\mathbb{E}(\hat{Z}_i) = 0$, Equation (43) equals Equation (42).

To prove that $\hat{Z} \in DA(S_\alpha)$, remember that $\hat{Z}_i = X_i(2F(X_i) - 1 - \frac{\theta}{\mu})$ is just $Z_i = X_i(2F(X_i) - 1)$ shifted by $\frac{\theta}{\mu}$. Therefore the same argument used in Theorem 2 for Z applies here to show that $\hat{Z} \in DA(S_\alpha)$. In particular we can point out that $\hat{Z}$ and Z (therefore also X) share the same α and slowly-varying function $L(n)$.

Notice that by assumption $X \in [c, \infty)$ with $c > 0$ and we are dealing with continuous distributions, therefore $\hat{Z} \in [-c(1 + \frac{\theta}{\mu}), \infty)$. As a consequence the left tail of $\hat{Z}$ does not contribute to changing the limit skewness parameter β, which remains equal to 1 (as for Z) by an application of Equation (38).

Therefore, by applying the GCLT we finally get

$$n^{\frac{\alpha-1}{\alpha}} \frac{1}{L_0(n)} \left(\frac{\sum_{i=1}^n Z_i}{\sum_{i=1}^n X_i} - \frac{\theta}{\mu} \right) \xrightarrow{d} \frac{1}{\mu} S(\alpha, 1, 1, 0). \tag{44}$$

We conclude the proof by noting that, as proven in Equation (39), the weak

ACCEPTED MANUSCRIPT

limit of the Gini index is characterized by the i.i.d sequence of $\frac{\sum_{i=1}^{n} Z_i}{\sum_{i=1}^{n} X_i}$ rather than the ordered one, and that an α-stable random variable is closed under scaling by a constant [25].

References

[1] O. Bousquet, S. Boucheron, G. Lugosi, *Introduction to statistical learning theory*, Springer (2004).

[2] B.K. Chakrabarti, A. Chakraborti, S.R. Chakravarty, A. Chatterjee, *Econophysics of Income and Wealth Distributions*, Cambridge University Press (2013).

[3] D. Chotikapanich, *Modeling income distributions and Lorenz curves*, Springer (2008).

[4] P. Cirillo, *Are your data really Pareto distributed?*, Physica A 392 (2013) 5947-5962.

[5] H. A. David, H. N. Nagaraja, *Order Statistics, Third Edition* , Wiley series in probability and statistics (2003).

[6] A. DasGupta, *Probability for Statistics and Machine Learning*, Springer (2011).

[7] L. De Haan, A. Ferreira, *Extreme value theory: an introduction*, Springer (2007).

[8] I. Eliazar, *Inequality spectra*, Physica A 469 (2017) 824-847.

[9] I. Eliazar, M.H. Cohen, *On social inequality: Analyzing the rich-poor disparity*, Physica A 401 (2014) 148-158.

[10] I. Eliazar, I. M. Sokolov, *Maximization of statistical heterogeneity. From Shannon's entropy to Gini's index* , Physica A 389 (2010) 3023-3038.

[11] I. Eliazar, I.M. Sokolov, *Gini characterization of extreme-value statistics*, Physica A 389 (2010) 4462-4472.

[12] I. Eliazar, I.M. Sokolov, *Measuring statistical evenness: A panoramic overview*, Physica A 391 (2012) 1323-1353.

[13] P. Embrechts, C. Kluppelberg, T. Mikosch, *Modelling Extremal Events for Insurance and Finance*, Springer (2003).

[14] W. Feller, *An introduction to probability theory and its applications*, Vol. 2, Wiley 2008.

[15] A. Fontanari, P. Cirillo, C.W. Oosterlee, *From Concentration Profiles to Concentration Maps. New Tools for the Study of Loss Distributions*, Insurance: Mathematics and Economics 78 (2018) 13-29.

ACCEPTED MANUSCRIPT

[16] A.H. Jesen, T. Mikosch, *Regularly Varying Functions*, Publications de l'Institut Mathématique, Nouvelle série, (2006).

[17] C. Gini, *Variabilità e mutabilità* (1912), Reprinted in: Variabilità e Mutabilità, E. Pizetti and T. Salvemini, Memorie di Metodologica Statistica, Libreria Eredi Virgilio Veschi (1955).

[18] G.H. Hardy, J.E. Littlewood, G. Pólya, *Inequalities*, Cambridge University Press (1952).

[19] C. Kleiber, S.Kotz, *Statistical Size Distributions in Economics and Actuarial Sciences*, Wiley (2003).

[20] D. Li, M.B. Rao, R.J.Tomkins, *The law of the iterated logarithm and central limit Theorem for L-statistics*, Journal of Multivariate Analysis 78 (2001) 191-217.

[21] J.P. Nolan, *Parameterizations and modes of stable distributions*, Statistics and Probability Letters 38.2 (1998) 187-195.

[22] V. Pareto, *La courbe de la répartition de la richesse* (1896), Reprinted in Rivista di Politica Economica 87 (1997) 647-700.

[23] T. Piketty, *Capital in the Twenty-First Century*, Harvard University Press (2014).

[24] T. Piketty, *The Economics of Inequality*, Harvard University Press (2015).

[25] G. Samorodnitsky, M. S. Taqqu, *Stable non-Gaussian random processes: stochastic models with infinite variance*, Vol. 1, CRC Press (1994).

[26] J. Shao, *Mathematical Statistics*, Springer (2003).

[27] N.N. Taleb, R. Douady, *On the super-additivity and estimation biases of quantile contributions*, Physica A: Statistical Mechanics and its Applications 429 (2015) 252-260.

[28] A. W. Van Der Vaart, J. A. Wellner, *Weak convergence and empirical processes*, Springer (1996).

[29] Y. Yang, S. Hu, T. Wu, *The tail probability of the product of depended random variables from max-domains of attraction*, Statistics and Probability Letters 81 (2011) 1876-1882.

[30] S.Yitzhaki, E. Schechtman, *The Gini Methodology: A primer on a statistical methodology*, Springer (2012).