Rapporto n. 220

On the parameters of Zenga distribution

Francesco Porro – Alberto Arcagni

Novembre 2011

Dipartimento di Metodi Quantitativi per le Scienze Economiche ed Aziendali
Università degli Studi di Milano Bicocca
Via Bicocca degli Arcimboldi 8 - 20126 Milano - Italia
Tel +39/02/64483102/3 - Fax +39/2/64483105
Segreteria di redazione: Andrea Bertolini

On the parameters of Zenga distribution*

Francesco Porro, Alberto Arcagni

Abstract

In 2010 Zenga introduced a new three-parameter model for distributions by size which can be used to represent income, wealth, financial and actuarial variables. In this paper a summary of its main properties is proposed. After that the article focuses on the interpretation of the parameters in term of inequality. The scale parameter μ is equal to the expectation, and it does not affect the inequality, while the two shape parameters α and θ are an inverse and a direct inequality indicators respectively. This result is obtained through stochastic orders based on inequality curves. A procedure to generate random sample from Zenga distribution is also proposed. The second part of the article is about the parameter estimation. Analytical solution of method of moments estimators is obtained. This result is used as starting point of numerical procedures to obtain maximum likelihood estimates both on ungrouped and grouped data. In the application, three empirical income distributions are considered and the aforementioned estimates are evaluated.

keywords: mixture, inequality, inequality $I(p)$ curve, income distribution

1 Introduction

Pareto, in 1895 [11], 1896 [12] and 1897 [13], proposed his models for income distributions, and he considered inequality as a major theme. Since then a discussion about inequality in income distributions had origin and different Italian scholars were involved, such as Benini (1897) [4], Amoroso (1925) [1] and D'Addario (1949) [7].

Pareto's first law has a good fitting only for high values of income distributions, therefore several alternative models have been proposed. Gamma and lognormal distributions provide a suitable description of the centre of the income distribution but for the upper and the lower tails are not satisfactory. The aforementioned models are two-parameter depending. Obviously, the greater the number of parameters, more flexible the model can be. Dagum type I [6]

*Although this paper arises from a collaboration between the two authors, the sections 2, 3 and the subsection 4.1 have to be attribuited to F. Porro, while the sections 1, 5 and the subsection 4.2 to A. Arcagni

and Singh - Maddala [18] are examples of three-parameter models for income distributions and they are special cases of the four-parameter GB2 model [10]. It is clear that a high number of parameters creates difficulties in estimation and interpretation, and a number of parameters greater than three could be considered too large. For a complete review of all these models is recommended to refer to Chotikapanich [5] and Kleiber and Kotz [9].

In this paper a new three-parameter model for distributions by size proposed by Zenga [23] is analyzed. It is a mixture of particular truncated Pareto distributions and it has the non-negative real numbers as support.

Dagum [6], introducing his model of income distribution, grouped the existing functional forms by the method used to obtain them:

- functional forms obtained by means of a stochastic process;
- functional forms used to describe income distributions solely by their satisfactory goodness of fit to empirical data, such as the Gamma model and the Weibull model;
- functional forms obtained solving systems of differential equations that define characteristics of regularity and performance observed in the empirical distributions of income, such as Pareto, Singh-Maddala and Dagum itself.

Zenga model does not belong to this classification, as it was not obtained by stochastic processes and differential equations, but it takes its origin by an inequality measure.

As mentioned, Zenga model is a mixture. The conditional densities were derived by Polisicchio [14] as the unique distribution model with expected value finite and positive and with uniform inequality curve $I(p)$ (Zenga, [22]). Defining the *lower mean* as the mean of the values lower or equal to the p-th quantile, and the *upper mean* as the mean of the values higher than the p-th quantile, the condition of uniformity on the $I(p)$ curve implies that the ratio between the lower mean and the upper mean is constant for every $p \in]0,1[$. The obtained distribution is a truncated Pareto distribution with traditional inequality parameter equal to 0.5. The parameters are the scale parameter μ (that can be proved to be the expectation) and the ratio between lower mean and upper mean denoted with k. The support of the Polisicchio distribution depends on the parameters and it is the interval $[\mu k, \mu/k]$.

Zenga distribution is obtained as mixture of Polisicchio distributions with μ constant and $k \in]0,1[$ with mixing function given by the Beta density. The resulting model has useful properties to describe income, wealth, financial and actuarial distributions. In fact, it has positive asymmetry and paretian right-tail but expected value always finite. In literature, it is well-known that income distribution has heavy right-tail but the income of the whole population, and consequentially the average income, is finite. Therefore, if the expected value of a fitted model is not finite, it has not economic interpretation. This are the

cases of Pareto distribution with $\alpha \leq 1$, Dagum distribution with $\delta \leq 1$, Singh - Maddala, and more in general GB2 distribution, with $aq \leq 1$ [9]. In these cases one of the main measures of location is not defined.

To describe distributions by size, location and inequality measures are the most relevant. The new model has three parameters: μ is the scale parameter and is equal to the expected value, α and θ are shape parameters which influence the inequality. Thus Zenga model allows to examine separately location and inequality. This has also implications in parameter estimation, because the restrictions on the expected value and on inequality measure (which are invariant to scale transformations) can be imposed separately. This property can be used to estimate parameters, through D'Addario's *invariants method* [8] or by imposinge restrictions to numerical procedures of optimization, such as minimization of goodness of fit indexes or likelihood maximization. This methods have been applied by Zenga et. al. [24] and Arcagni [2] to estimate the parameters of Zenga ditribution, and through several applications and simulations they observed a good fitting of the model on the whole range of the empirical distribution.

The paretian right-tail allows the model to fit the income distributions for large values. Zenga distribution can assume several shapes and it can be zero-modal and unimodal: this feature allows a good fitting also for small income.

By this short introduction on the properties of Zenga distribution, it can be observed that it meets the requirements proposed by Dagum [6], in particular:

- parsimony, since the distribution function depends only on three parameters;
- economic interpretation of parameters;
- simple and efficient method of parameter estimation;
- model flexibility;
- good fit on the whole range of the distribution.

The paper is organized as follows. Zenga model is presented in section 2 providing conditional densities, mixing function, probability density function, distribution function and some other main features. Section 3 focuses on other features which allow to point out the roles of the shape parameters in terms of inequality, in particular the stochastic orders based on inequality curves are showed. In the same section how to generate random values from Zenga distribution is proposed. In section 4 the estimation methods considered in this paper are described, the analytical solution of the method of moments is obtained and maximum likelihood functions on ungrouped and grouped data are defined. The applications are shown in section 5 with the definition of the intervals and the definition of the goodness of fit indexes. Section 6 is devoted to conclusions and final remarks.

2 Definition and some initial features

Zenga distribution is obtained as a mixture of Polisicchio's truncated Pareto distributions with Beta weights. The conditional densities have probability density function given by:

$$v(x;\mu,k) = \begin{cases} \frac{\sqrt{\mu k}}{2}(1-k)^{-1}x^{-3/2} & x \in [\mu k, \mu/k] \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where μ is the non-negative expectation and $k \in]0,1[$. These densities have the peculiarity that the corresponding inequality $I(p)$ curve is uniform with inequality level equal to $1-k$, and they have been introduced and analyzed by Polisicchio [14].

The mixing function on $k \in]0,1[$ is a Beta distribution (depending on the two positive parameters α and θ) with probability density function given by

$$g(k;\alpha,\theta) = \begin{cases} \frac{k^{\alpha-1}(1-k)^{\theta-1}}{B(\alpha,\theta)} & k \in]0,1[\\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $B(a,b)$ is the Beta function:

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt \qquad a>0, b>0.$$

By definition, therefore, the probability density function of Zenga distribution is

$$\begin{aligned} f(x;\mu,\alpha,\theta) &= \int_0^1 v(x;\mu,k)g(k;\alpha,\theta)\,dk \\ &= \begin{cases} \frac{1}{2\mu B(\alpha;\theta)}\left(\frac{x}{\mu}\right)^{-3/2}\int_0^{\frac{x}{\mu}} k^{\alpha-1/2}(1-k)^{\theta-2}\,dk & \text{if } 0<x<\mu \\ \frac{1}{2\mu B(\alpha;\theta)}\left(\frac{\mu}{x}\right)^{3/2}\int_0^{\frac{\mu}{x}} k^{\alpha-1/2}(1-k)^{\theta-2}\,dk & \text{if } x>\mu \end{cases} \end{aligned} \tag{3}$$

and the distribution function is

$$F(x;\mu,\alpha,\theta) = \begin{cases} \int_0^{\frac{x}{\mu}} \left[1-\sqrt{\frac{\mu k}{x}}\right]\frac{k^{\alpha-1}(1-k)^{\theta-2}}{B(\alpha,\theta)}\,dk & \text{if } 0<x\le\mu \\ 1-\int_0^{\frac{\mu}{x}} \left[\sqrt{\frac{\mu}{xk}}-1\right]\frac{k^{\alpha}(1-k)^{\theta-2}}{B(\alpha,\theta)}\,dk & \text{if } x>\mu. \end{cases} \tag{4}$$

In the following some basic properties of Zenga distribution are briefly presented: further details and more explications can be found in [23], [25] and [24].

The first feature of Zenga distribution is that, since it is a mixture of continuous random variables (each of them with finite expectation μ), it has always finite expected value equal to μ, for any admissible choice of the parameters.

The behaviour of the probability density function in a neighbourhood of 0 and in a neighbourhood of μ needs to be described. It holds that

$$\lim_{x\to 0^+} f(x;\mu,\alpha,\theta) = \begin{cases} +\infty & \text{if } 0<\alpha<1 \\ \dfrac{\theta}{3\mu} & \text{if } \alpha = 1 \\ 0 & \text{if } \alpha > 1 \end{cases}$$

and that

$$\lim_{x\to \mu} f(x;\mu,\alpha,\theta) = \begin{cases} \dfrac{B(\alpha+1/2;\theta-1)}{2\mu B(\alpha,\theta)} & \text{if } \theta > 1 \\ +\infty & \text{if } 0<\theta\leq 1. \end{cases}$$

It is very interesting to note that the parameter α governs the behaviour of the density function as x tends to 0, while the value of the parameter θ regulates the finitess (or not) of the function in a neighbourhood of μ. As it will be shown later, the value of α also controls the finitess of the moments: in other words therefore it can be stated that α affects the tails, while θ affects the behaviour around the mean μ.

Another interesting property of Zenga distribution is the capability to model very different distributions: it is easy to see how many behaviours can have, changing the values of the parameters. In Figures 1 and 2 some probability density functions are showed: in both of them, the value of μ is set to 1, but in the first one, θ is fixed and equal to 4 and 0.5, while in the second one θ changes and α equals 2 and 0.5.

An analysis about the moments of the distribution is important. It has been proved that if X denotes a random variable with Zenga distribution with parameters μ, α and θ, then

$$\mathbb{E}(X^r) = \frac{\mu^r}{(2r-1)B(\alpha,\theta)} \sum_{i=1}^{2r-1} B(\alpha-r+i,\theta) \qquad \text{for } r<\alpha+1.$$

In other words, it is guaranteed the finiteness of the moments of order less then $\alpha+1$. In particular, if $\alpha>1$, then

$$\mathbb{E}(X) = \mu$$

and

$$\mathbb{E}(X^2) = \frac{\mu^2}{3}\left(\frac{\theta}{\alpha-1} + \frac{\alpha}{\alpha+\theta} + 2\right),$$

therefore the variance is:

$$\mathbb{E}(X-\mu)^2 = \frac{\mu^2\theta(\theta+1)}{3(\alpha-1)(\alpha+\theta)}. \tag{5}$$

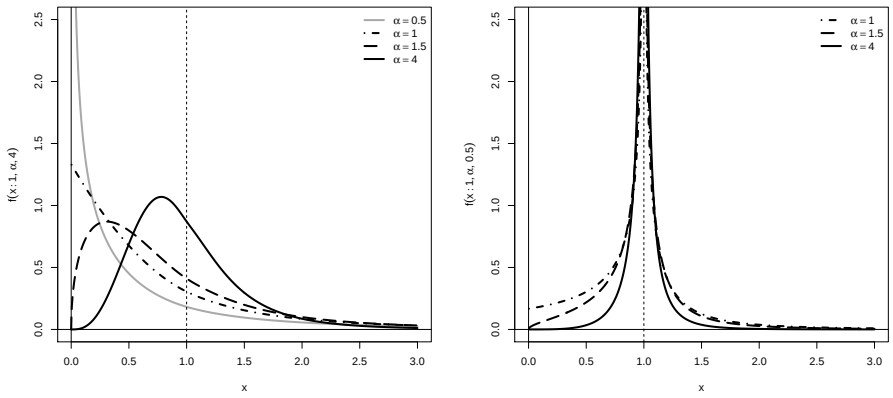

Figure 1: The density functions of Zenga distribution with $\mu = 1$ and $\theta = 4$ (on the left) and $\theta = 0.5$ (on the right)

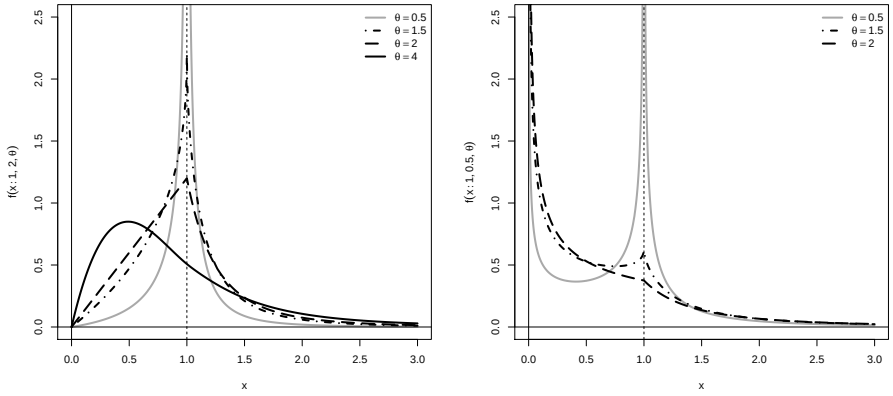

Figure 2: The density functions of Zenga distribution with $\mu = 1$ and $\alpha = 2$ (on the left) and $\alpha = 0.5$ (on the right)

In general, the r^{th} central moment of Zenga distribution is

$$\begin{aligned} \mathbb{E}(X-\mu)^r &= \int_0^{+\infty} (x-\mu)^r f(x;\mu,\alpha,\theta)\, dx \\ &= \int_0^{+\infty} \left[(x-\mu)^r \int_0^1 v(x;\mu,k) g(k;\alpha,\theta)\, dk \right] dx. \end{aligned}$$

Whenever $\mathbb{E}(X-\mu)^r$ is finite, by Fubini theorem:

$$\begin{aligned}\mathbb{E}(X-\mu)^r &= \int_0^1 \left[\int_0^{+\infty} (x-\mu)^r v(x;\mu,k) g(k;\alpha,\theta)\, dx\right] dk \\ &= \int_0^1 \left[g(k;\alpha,\theta)\int_0^{+\infty} (x-\mu)^r v(x;\mu,k)\, dx\right] dk.\end{aligned}$$

As the conditional densities have the same expectation μ, the integral

$$\int_0^{+\infty} (x-\mu)^r v(x;\mu,k)\, dx = \mu_{r,k}$$

coincides with the r^{th} central moment of the conditional density with parameter k. Then:

$$\mathbb{E}(X-\mu)^r = \int_0^1 \mu_{r,k} g(k;\alpha,\theta)\, dk,$$

that is, the r^{th} central moment of Zenga distribution can be calculated via the r^{th} central moments of the conditional densities and the mixing function of the mixture. Because of the third central moment of the conditional density equals $\mu^3(1-k)^4/5k^2$ (see [14]), it follows that the third central moment of Zenga distribution is:

$$\mathbb{E}(X-\mu)^3 = \frac{\mu^3}{5}\cdot\frac{\theta(\theta+1)(\theta+2)(\theta+3)}{(\alpha-1)(\alpha-2)(\alpha+\theta+1)(\alpha+\theta)}. \tag{6}$$

3 Other further features

In this section further properties of Zenga distribution are reported: after an examination on the roles of the parameters, two important indexes are calculated. After that, a remark on the asymmetry is provided. Then a recognition about stochastic orders is approached with particular attention on how the parameters modify the inequality. The final part is devoted to the description of a procedure to obtain samples from a variable with Zenga distribution.

In [23] it is proved that μ is a scale parameter. Starting from this, the parameters can be grouped by their role: μ controls the scale, but it has no influence on inequality, while the remaining two parameters α and θ govern the shape and the inequality. As mentioned in the introduction, this characteristic can be useful also in estimation of the parameters.

For Zenga distribution, the value of Pietra index is

$$P = \frac{\mathbb{E}(|X-\mu|)}{2\mu} = 2F(\mu;\mu,\alpha,\theta) - 1,$$

while, Zenga point inequality measure evaluated in μ coincides with

$$A(\mu) = 1 - \frac{\mathbb{E}(X|X\le\mu)}{\mathbb{E}(X|X\le\mu)} = 1 - \left[\frac{1-F(\mu;\mu,\alpha,\theta)}{F(\mu;\mu,\alpha,\theta)}\right]^2,$$

where

$$F(\mu;\mu,\alpha,\theta)=\frac{1}{\theta-1}\left[(\alpha+\theta-1)-\frac{(\alpha+\theta-1/2)\Gamma(\alpha+1/2)\Gamma(\alpha+\theta)}{\Gamma(\alpha+\theta+1/2)\Gamma(\alpha)}\right].$$

See [22] for further details about Zenga inequality measure. It can be proved that for Zenga distribution $F(\mu;\mu,\alpha,\theta)\geq 1/2$: this implies that the mean is greater than the median and therefore the distribution has positive asymmetry. This characteristic has remarkable importance in the applicative field, since the empirical evidence shows that income distributions usually own positive asymmetry.

About stochastic orders, two results, proved in [16], are summarized in the next theorems: they deal with the convex order, the order based on Lorenz curve and the order based on inequality $I(p)$ curve. It is useful to recall the definition of these three orderings: further details can be found in [17] and [16].

Definition 1. *Let X_1 and X_2 be two continuous non-negative random variables with finite expectations. X_1 is said to be larger (or more unequal) than X_2 in the Lorenz ordering (and it is denoted by $X_1\geq_L X_2$), if*

$$L_{X_1}(p)\leq L_{X_2}(p)\qquad\qquad \forall p\in(0,1)$$

where $L_{X_i}(p)$, $i=1,2$, is the value assumed by the Lorenz curve of X_i in p (with $i=1,2$).

In analogy to the order based on Lorenz curve, Porro in [15] introduced the order based on $I(p)$ curve by the following definition.

Definition 2. *Let X_1 and X_2 be two continuous non-negative random variables with finite expectations. X_1 is said to be larger (or more unequal) than X_2 in the ordering based on $I(p)$ curve (and it is denoted by $X_1\geq_I X_2$), if*

$$I_{X_1}(p)\geq I_{X_2}(p)\qquad\qquad \forall p\in(0,1)$$

where $I_{X_i}(p)$, $i=1,2$, is the value assumed by the inequality $I(p)$ curve of X_i in p (with $i=1,2$).

Definition 3. *Let X_1 and X_2 be two continuous non-negative random variables with finite expectations. X_1 is said to be larger than X_2 in the convex order (and it is denoted by $X_1\geq_{CX} X_2$), if*

$$\mathbb{E}[\phi(X_1)]\geq\mathbb{E}[\phi(X_2)]$$

for all the convex functions $\phi:\mathbb{R}\to\mathbb{R}$, such that the expectations exist.

Now, the two ordering theorems for Zenga distribution can be stated.

Theorem 1. *Let X_1 and X_2 be two continuous random variables such that $X_i\sim$ Zenga (μ,α_i,θ) $i=1,2$, where $\theta>1$, $0<\alpha_1<\alpha_2$ and $\mu>0$. Then it holds that: $X_2\leq_L X_1$, $X_2\leq_I X_1$, and $X_2\leq_{CX} X_1$.*

Theorem 2. *Let X_1 and X_2 be two continuous random variables, such that $X_i \sim$ Zenga (μ, α, θ_i) $i = 1, 2$ where $1 < \theta_1 < \theta_2$, $\alpha > 0$ and $\mu > 0$. Then it holds that: $X_1 \leq_L X_2$, $X_1 \leq_I X_2$, and $X_1 \leq_{CX} X_2$.*

The previous results point out that if one parameter is fixed, the other one is an inequality indicator for Lorenz curve and for inequality $I(p)$ curve: more in detail, the theorems state that α is an inverse inequality indicator (the bigger α, the smaller inequality), while θ is a direct inequality indicator (the bigger θ, the higher inequality). Geometrically, this means that as one parameter changes, Lorenz curves of Zenga distribution are nested. The same holds for the corresponding inequality $I(p)$ curves. It is clear that the parameters α and θ have the same roles for the two inequality indexes related to the two considered cuves: Gini concentration ratio and Zenga inequality index I.

The last issue is about sampling. In Zenga model, the distribution function cannot be inverted analytically. Nevertheless, in order to provide simulations to evaluate the behaviour of the estimation methods, Arcagni [2] proposed to generate random values from such distribution through a two-step sampling. The sampling procedure can be described as follows. First, a value of the parameter k is generated from a Beta distribution with parameters α and θ, then the sampling value is obtained by generating a random value from a variable following a Polisicchio distribution with parameters μ and k.

4 Estimation of the parameters

The subject of the estimation of the parameters of a distribution is fundamental: in this paper two well-known methods are provided. The first one is the classical method of moments, the second one is the maximun likelihood method.

4.1 Method of moments

Because of the features of Zenga distribution, one moment and two central moments will be used in the method of moments: this allows to obtain the estimators in a more manageable analytical form.

Let $(x_1, ..., x_n)$ be a random sample from a Zenga distribution. Let

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2, \qquad m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

be the mean, the variance and the third central moment of the sample, respectively. The first three moments of Zenga distribution are (see (5) and (6)):

$$\begin{cases} \mathbb{E}(X) = \mu \\ \mathbb{E}(X-\mu)^2 = \dfrac{\mu^2}{3} \cdot \dfrac{\theta(\theta+1)}{(\alpha-1)(\alpha+\theta)} \\ \mathbb{E}(X-\mu)^3 = \dfrac{\mu^3}{5} \cdot \dfrac{\theta(\theta+1)(\theta+2)(\theta+3)}{(\alpha-1)(\alpha-2)(\alpha+\theta+1)(\alpha+\theta)}. \end{cases}$$

Because of the particular form of the central moments of such distribution, the second central moment can be replaced in the third central moment, therefore it holds that

$$\begin{cases} \mathbb{E}(X) = \mu \\ \mathbb{E}(X-\mu)^2 = \dfrac{\mu^2}{3} \cdot \dfrac{\theta(\theta+1)}{(\alpha-1)(\alpha+\theta)} \\ \mathbb{E}(X-\mu)^3 = \mathbb{E}(X-\mu)^2 \cdot \dfrac{3\mu(\theta+2)(\theta+3)}{5(\alpha-2)(\alpha+\theta+1)}. \end{cases}$$

Then, according to the method of moments,

$$\begin{cases} \bar{x} = \mu \\ m_2 = \dfrac{\mu^2}{3} \cdot \dfrac{\theta(\theta+1)}{(\alpha-1)(\alpha+\theta)} \\ m_3 = \dfrac{\mu^3}{5} \cdot \dfrac{\theta(\theta+1)(\theta+2)(\theta+3)}{(\alpha-1)(\alpha-2)(\alpha+\theta+1)(\alpha+\theta)}, \end{cases}$$

and equivalently

$$\begin{cases} \bar{x} = \mu \\ m_2 = \dfrac{\bar{x}^2}{3} \cdot \dfrac{\theta(\theta+1)}{(\alpha-1)(\alpha+\theta)} \\ m_3 = m_2 \cdot \dfrac{3\bar{x}(\theta+2)(\theta+3)}{5(\alpha-2)(\alpha+\theta+1)}. \end{cases}$$

After some algebra, it follows that:

$$\begin{cases} \bar{x} = \mu \\ \alpha^2 + \alpha(\theta-1) - \theta - \dfrac{\bar{x}^2\theta}{3m_2}(\theta-1) = 0 \\ \theta^2\left[\dfrac{\bar{x}^2}{3m_2} - \dfrac{3\bar{x}m_2}{5m_3}\right] + \theta\left[\dfrac{\bar{x}^2}{3m_2} - \dfrac{3\bar{x}m_2}{m_3} - 1\right] - \left[\dfrac{18\bar{x}m_2}{5m_3} + 2\right] = 0. \end{cases}$$

It is important to note that thus the third equation becomes a second-degree equation in θ instead of the original fourth-degree one: this largely simplifies the procedure to obtain the solution of the system.

Then the following estimates of the parameters can be achieved:

$$\begin{cases} \hat{\mu} = \bar{x} \\ \hat{\theta} = \dfrac{-\left[\dfrac{1}{3}\dfrac{\bar{x}^2}{m_2} - 3\dfrac{\bar{x}\,m_2}{m_3} - 1\right] + \sqrt{\left[\dfrac{1}{3}\dfrac{\bar{x}^2}{m_2} - 3\dfrac{\bar{x}\,m_2}{m_3} - 1\right]^2 + 4\left[\dfrac{1}{3}\dfrac{\bar{x}^2}{m_2} - \dfrac{3}{5}\bar{x}\dfrac{m_2}{m_3}\right]\left[\dfrac{18}{5}\dfrac{\bar{x}\,m_2}{m_3} + 2\right]}}{2\left[\dfrac{1}{3}\dfrac{\bar{x}^2}{m_2} - \dfrac{3}{5}\bar{x}\dfrac{m_2}{m_3}\right]} \\ \hat{\alpha} = \dfrac{-(\hat{\theta}-1) + \sqrt{\left(\hat{\theta}-1\right)^2 + 4\left[\dfrac{1}{3}\dfrac{\bar{x}^2}{m_2}\hat{\theta}\left(\hat{\theta}+1\right) + \hat{\theta}\right]}}{2}. \end{cases}$$

An important remark is that the analytical solution makes sense only under the restrictions $\hat{\alpha} > 2$, and $\hat{\theta} > 0$ and hence the sample statistics $\bar{x}$, m_2 and m_3 must satisfy the following condition:

$$m_3 > \frac{9m_2^2}{5\bar{x}}.$$

4.2 Maximun likelihood

The maximum likelihood estimates of the parameters are obtained by numerical optimization. Since the procedure on ungrouped data is time consuming, the estimation on grouped data is proposed too. Therefore the definitions of the objective functions are listed here.

Let $(x_1, \ldots, x_n)$ be a random sample from Zenga distribution, then the likelihood function is

$$\mathcal{L}(\mu, \alpha, \theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \mu, \alpha, \theta)$$

and the log-likelihood function is

$$\log \mathcal{L}(\mu, \alpha, \theta; x_1, \ldots, x_n) = \sum_{i=1}^{n} \log f(x_i; \mu, \alpha, \theta). \tag{7}$$

For large values of n it can be time consuming to evaluate the density function for each observation, consequentially it is proposed to group the n observations into s intervals $]x'_0 = 0, x'_1[$, $[x'_{j-1}, x_j[$, for $j = 2, \ldots, s-1$, and $[x_{k-1}, x'_s = \infty[$, such that to cover the entire support of the variable. Let n_j be the empirical frequency of the $j-$th interval, with $j = 1, \ldots, s$. Therefore the maximum likelihood function for grouped data is obtained through the multinomial distribution of parameters n and

$$p_j(\mu, \alpha, \theta) = F(x'_j; \mu, \alpha, \theta) - F(x'_{j-1}; \mu, \alpha, \theta), \quad j = 1, \ldots, s.$$

Likelihood and log-likelihood functions for grouped data are here defined

$$\mathcal{L}_G(\mu, \alpha, \theta; n_1, \ldots, n_s) = \frac{n!}{\prod_{j=1}^{s} n_j!} \prod_{j=1}^{s} p_j(\mu, \alpha, \theta)^{n_j}$$

$$\log \mathcal{L}_G(\mu, \alpha, \theta; n_1, \ldots, n_s) = \log \frac{n!}{\prod_{j=1}^{s} n_j!} + \sum_{j=1}^{s} n_j \log p_j(\mu, \alpha, \theta) \tag{8}$$

which are functions of the Zenga distribution parameters through the theoretical probabilities of inclusion of each interval.

Numerical procedures maximize the log-likelihood objective function through the Nelder and Mead method with starting point set equal to the method of moments estimates.

5 Application to real data

In this section the applications of Zenga distribution on three income distributions are presented. They are selected from those used by Zenga et. al. [24], who provided the estimates of the parameters with the numerical solution of the method of moments, D'Addario's invariants method [8] and through minimization of measures of goodness of fit with and without restrictions. The estimates provided in this paper are obtained through analytical solution of method of moments and maximum likelihood method. Maximum likelihood estimates are obtained for ungrouped and grouped data.

5.1 Goodness of fit indexes

To compare the fitting of different estimation methods three indexes are provided.

Assume that the n observations are grouped into s intervals $]x'_{j-1}, x'_j]$, for $j = 1, \ldots, s$. Let be n_j the empirical frequency and

$$\hat{n}_j = n\left[F(x'_j; \hat{\mu}, \hat{\alpha}, \hat{\theta}) - F(x'_{j-1}; \hat{\mu}, \hat{\alpha}, \hat{\theta})\right]$$

the estimated frequency of the j^{th} interval. Therefore the Mortara index A_1, the quadratic K. Pearson index A_2 and the modified quadratic index A'_2 are defined as follow to evaluate the goodness of fit

$$\begin{aligned}
A_1 &= \frac{1}{n}\sum_{j=1}^{s}\frac{|n_j - \hat{n}_j|}{\hat{n}_j}\hat{n}_j = \frac{1}{n}\sum_{j=1}^{s}\frac{|n_j - \hat{n}_j|}{n_j}n_j = \frac{1}{n}\sum_{j=1}^{s}|n_j - \hat{n}_j| \\
A_2 &= \left\{\frac{1}{n}\sum_{j=1}^{s}\left|\frac{n_j - \hat{n}_j}{\hat{n}_j}\right|^2 \hat{n}_j\right\}^{1/2} = \left\{\frac{1}{n}\sum_{j=1}^{s}\frac{|n_j - \hat{n}_j|^2}{\hat{n}_j}\right\}^{1/2} \\
A'_2 &= \left\{\frac{1}{n}\sum_{j=1}^{s}\left|\frac{n_j - \hat{n}_j}{n_j}\right|^2 n_j\right\}^{1/2} = \left\{\frac{1}{n}\sum_{j=1}^{s}\frac{|n_j - \hat{n}_j|^2}{n_j}\right\}^{1/2}.
\end{aligned}$$

Defining the absolute relative frequency deviations

$$\begin{aligned}
a_j &= \frac{|n_j - \hat{n}_j|}{\hat{n}_j} \\
a'_j &= \frac{|n_j - \hat{n}_j|}{n_j}
\end{aligned}$$

it can be observed that

$$\begin{aligned}
A_1 &= M_1(a_j; \hat{n}_j) = M_1(a'_j; n_j) \\
A_2 &= M_2(a_j; \hat{n}_j) \\
A'_2 &= M_2(a'_j; n_j)
\end{aligned}$$

that is, Mortara index A_1 is both the arithmetic mean of a_j with weights $\hat{n}_j$ and the arithmetic mean of a'_j with weights n_j, A_2 index is the quadratic mean of a_j with weights $\hat{n}_j$ and A'_2 index is the quadratic mean of a'_j with weights n_j. Therefore the variance of a_j with weights $\hat{n}_j$ and the variance of a'_j with weights n_j can be obtained as follow

$$Var(a_j; \hat{n}_j) = {A_2}^2 - {A_1}^2$$
$$Var(a'_j; n_j) = {A'_2}^2 - {A_1}^2$$

and small differences between A_1 index and quadratic indexes (A_2 or A'_2) mean low variability of absolute relative frequency deviations, and a uniform fitting of the model on the whole range of the empirical distribution.

5.2 Intervals

The same intervals used to evaluate the A_1, A_2 and A'_2 indexes are used for maximum likelihood estimation on grouped data. As in [24] the n observations are grouped into $s = 25$ intervals starting from prefixed values of the cumulative relative frequencies p'_j:

j	1	2	3	4	5
p'_j	0.010	0.020	0.035	0.050	0.100
j	6	7	8	9	10
p'_j	0.150	0.200	0.250	0.300	0.400
j	11	12	13	14	15
p'_j	0.500	0.600	0.700	0.750	0.800
j	16	17	18	19	20
p'_j	0.850	0.900	0.920	0.935	0.950
j	21	22	23	24	25
p'_j	0.960	0.970	0.980	0.990	1.000

Then each np'_j is approximated with its nearest integer np_j. Consequentially the (integer) frequencies n_j are given by

$$n_j = n(p_j - p_{j-1}); \quad j = 1, \ldots, s; \quad p_0 = 0.$$

The upper bounds of the s intervals are:

$$x'_j = x_{(np_j)}; \quad j = 1, \ldots, s-1; \quad x'_s = \infty$$

where $x_{(np_j)}$ is the np_j^{th} order statistic of n empirical individual observation. Since x'_s is not finite the last interval is open on both sides.

5.3 Empirical distributions and parameter estimation

The empirical distribution used in this paper are:

- Italy 2006 Household income [3], 7,762 observations;
- Swiss 2005 Household income [19], 3,071 observations;
- USA 2008 Household income [21], 2,899,458 observations.

In figures 3, 4 and 5 fitted models overlapping the histogram are reported.

Table 1 shows the results obtained with different estimation methods.

	estimates			**goodness of fit indexes**		
	$\hat{\mu}$	$\hat{\alpha}$	$\hat{\theta}$	A_1	A_2	A_2'
method of moments						
Italy 2006	31918.93	2.4447	4.0653	0.1233	0.1657	0.2025
Swiss 2005	6783.75	4.0210	4.8195	0.1359	0.1702	0.1658
USA 2008	82460.21	3.9922	10.7071	0.2677	0.5930	0.3232
max. likelihood						
Italy 2006	31701.73	2.9090	4.2534	0.0781	0.0893	0.0892
Swiss 2005	6921.68	2.9383	3.4698	0.1068	0.1403	0.1506
USA 2008	84674.75	1.5255	3.1434	0.0454	0.0654	0.0690
max. likelihood on grouped data						
Italy 2006	31453.77	3.3967	4.9909	0.0639	0.0715	0.0723
Swiss 2005	6915.71	3.0915	3.6654	0.1080	0.1392	0.1489
USA 2008	84278.03	1.5702	3.2413	0.0462	0.0645	0.0670

Table 1: results of the estimation methods: estimates and goodness of fit measures

By rows are reported estimation methods and empirical distributions.

From first to third column there are the parameters estimates. Method of moments estimates obtained with analytical solutions are close to ones obtained with the numerical procedure by Zenga et. al. [24]. It is important to note that maximum likelihood estimate of parameter α, on ungrouped and grouped data from USA distribution, is lower than 2. This result cannot be obtained with the method of moments, because the third central moment is not finite and the third equation is not defined.

Last three columns provide values of goodness of fit indexes. The fourth column shows values of Mortara A_1 index. The highest index value is obtained with method of moments on USA distribution and it can be attributed to the restriction on the parametric space. Maximum likelihood method, both on ungrouped and grouped data, provides good results on Italy and USA distributions. For instance, in case of maximum likelihood on USA ungrouped data, the empirical frequencies, n_j, differ on average from estimated ones, $\hat{n}_j$, by 4.54%. Considering only the maximum likelihood method, indexes of goodness of fit do not change significantly if the fitted model is evaluated on ungrouped or grouped data. The same considerations can be done for the other goodness of fit indexes too.

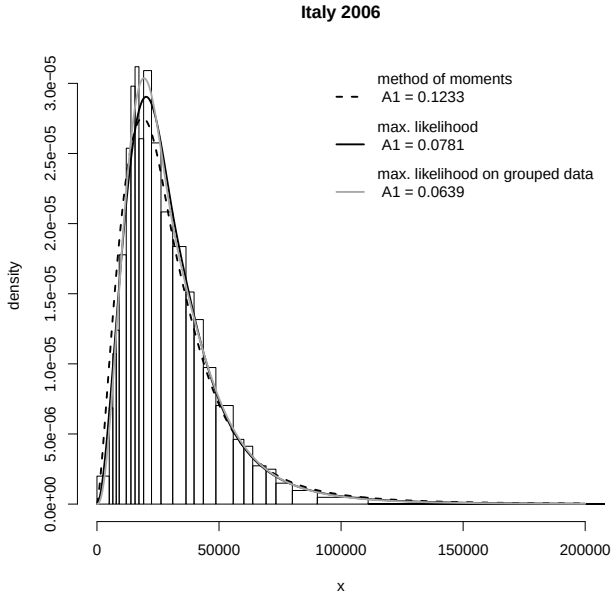

Figure 3: Zenga densities fitted to empirical Italy 2006 Household income distribution

The comparison between A_1 index and quadratic indexes allows to evaluate the fitting uniformity. For instance, it has been observed that the model with parameters estimated by method of moments does not fit good on USA distribution. Note that, in this case, there are also wide differences between A_1 index and quadratic indexes, which mean an irregular fitting. In Figure 5 it can be observed that the fitted model with method of moments overestimates the empirical distribution for low and high income values, and underestimates the middle of the distribution. In the other hand maximum likelihood methods provide estimated models with a uniform fitting to USA income distribution, that can be observed in the same figure and it is confirmed by small differences between index A_1 and quadratic indexes.

By goodness of fit indexes and graphs it can be observed that the method of moments estimated model sometimes does not fit good, but the analytical solution obtained in this paper is an important result because it is an easy way to obtain parameter estimates and thay can be used as starting point for numerical procedures. By the small differences of the goodness of fit index between maximum likelihood estimates obtained on ungrouped data and grouped data, the second ones could be preferred because numerical procedures are much faster.

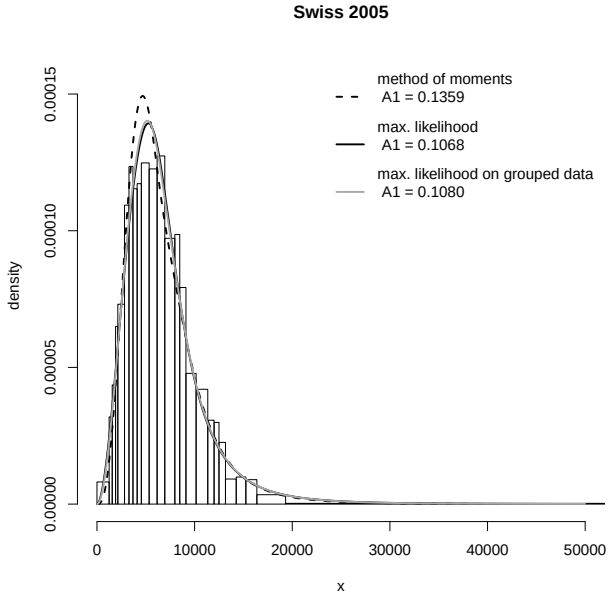

Figure 4: Zenga densities fitted to empirical Swiss 2005 Household income distribution

6 Conclusions and final remarks

In this paper Zenga distribution and the role of its parameters are presented. This model has positive asymmetry, it has paretian right tail, and it can be used to describe economic distributions by size. It can assume several shapes; it can be zero modal or unimodal and it seems to have a good fitting for income distributions either at low and high values. The distribution depends on three parameters: in particular μ is a scale parameter, α is an inverse inequality indicator and it controls the tails of the distribution, while θ is a direct inequality indicator and it controls the distribution around the expected value μ.

The estimation of the parameters is also presented in the paper. Several estimation methods are presented by Zenga et al. [24], here method of moments and maximum likelihood method were presented and applied. The fitting to the data of the estimated distribution with method of moments it not so good, this may be due to the restriction on parametric space required for the existence of the third moment. However the estimates achieved by the method of moments can be used as starting point of numerical procedures to obtain maximum likelihood estimates. Maximum likelihood method is applied on ungrouped and grouped data. Since the values of Mortara goodness of fit index for grouped and ungrouped data are not very different, it seems to be preferable the maximum

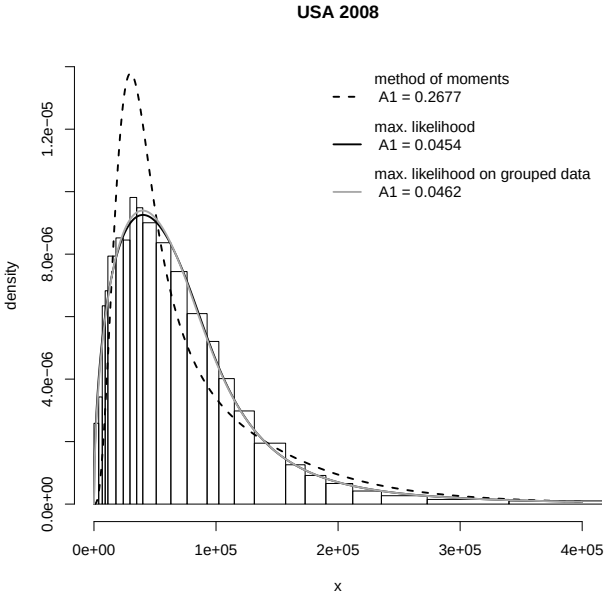

Figure 5: Zenga densities fitted to empirical USA 2008 Household income distribution

likelihood estimation for grouped data because of the smaller computational time.

References

[1] Amoroso, L., Ricerche intorno alla curva dei redditi, Annali di Matematica Pura ed Applicata, Serie 4-21, II, 123-157 (1925)

[2] Arcagni, A., La determinazione dei parametri di un nuovo modello distributivo per variabili non negative: aspetti metodologici e applicazioni, PhD thesis: Università degli Studi di Milano Bicocca (2011)

[3] Banca d'Italia, The 2006 Bank of Italy sample Survey on Household income and wealth, Supplements to the Statistical Bulletin Sample Surveys, XVIII(7), Available at http://www.bancaditalia.it (2008)

[4] Benini, R. Di alcune curve descritte da fenomeni economici aventi relazione colla curve del reddito o con quella del patrimonio, Giornale degli Economisti, 14, 177-214 (1897)

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

[6] Dagum, C., A new model of personal income distribution: specification and estimation, Economie Appliquee (1977)

[7] D'Addario, R., Ricerche sulla curva dei redditi, Giornale degli Economisti e Annali di Economia, 8, 91-114 (1949)

[8] D'Addario, R., La curva dei redditi: sulla determinazione numerica dei parametri della seconda equazione paretiana, Annali dell'Istituto di Statistica dell'UniversitÃ di Bari, XII (1939)

[9] Kleiber, C. and Kotz, S., Statistical size distributions in economics and actuarial sciences, (381), Wiley-Interscience (2003)

[10] McDonald, J., Some generalized functions for the size distribution of income, Econometrica (1984)

[11] Pareto, V., La legge della domanda, Giornale degli economisti, (10), 59–68 (1895)

[12] Pareto, V., Escrits sur la coubre de la répartition de la richesse, in G. Busino (Ed.), Complete works of V. Pareto, Librarie Droz, Genève, 1965 (1896)

[13] Pareto, V., Cours d'economie politique, new edition by G. H. and G. Busino, Librarie Droz, Genève (1897)

[14] Polisicchio, M., The continuous random variable with uniform point inequality measure $I(p)$, Statistica & Applicazioni, VI(2), 137-151 (2008)

[15] Porro, F., Equivalence between partial order based on curve L(p) and partial order based on curve I(p), Proceedings of SIS 2008, CLUEP, Padova (2008)

[16] Porro, F., Inequality order for Zenga distribution, Technical Report 215, Dipartimento di Metodi Quantitativi per le Scienze Economiche ed Aziendali, Università degli Studi di Milano-Bicocca, Available at http://boa.unimib.it (2011)

[17] Shaked, M. and Shanthikumar, J.G., Stochastic orders and their applications,Springer New York, NY (2007)

[18] Singh, S. K. and Maddala, G. S., A Function for Size Distribution of Incomes, Econometrica: Journal of the Econometric Society (1976)

[19] Swiss Federal Statistical Office, Income and Consumption Survey (2005)

[20] Thisted, R. A., Elements of statistical computing: numerical computation, Chapman & Hall London (1988)

[21] U.S. Census Bureau, Current Population Survey (2008)

[22] Zenga, M. M., Inequality curve and inequality index based on the ratios between lower and upper arithmetic means, Statistica & Applicazioni, V(1), 3–27 (2007)

[23] Zenga, M. M., Mixture of Polisicchio's truncated Pareto distributions with beta weights, Statistica & Applicazioni, VIII(1), 3-25 (2010)

[24] Zenga, M. M., Pasquazzi, L., Zenga, Ma., First Applications of a New Three Parameter Distribution for Non-Negative Variables, Technical Report 187, Dipartimento di Metodi Quantitativi per le Scienze Economiche ed Aziendali, Università degli Studi di Milano-Bicocca, Available at http://boa.unimib.it (2010)

[25] Zenga, M. M., Polisicchio, M., Zenga, Ma., Pasquazzi, L., More on M. M. Zenga's new three-parameter distribution for non-negative variables, Statistica & Applicazioni, IX(1), 5–33 (2011)