Annals of the University of Craiova, Mathematics and Computer Science Series
Volume 50(1), 2023, Pages 29–41, DOI: 10.52846/ami.v50i1.1602
ISSN: 1223-6934

Varying bandwidth parameter method on Kernel Gini index estimation

Komi Agbokou and Yaogan Mensah

Abstract. Most of measures of income inequality are derived from the Lorenz curve and many authors state that the Gini index is the best single measure of inequality. The present paper reviews some of theoretical properties of the Lorenz curve and provides a non-parametric estimate of the Gini index and the almost sure convergence of this estimate. And to confirm the performance of the estimator, a simulation on real data was carried out.

2010 Mathematics Subject Classification. 62G05, 62G07 ,62G30 , 62P20, 65D30.
Key words and phrases. Lorenz curve, Gini index, Nonparametric estimation, Variable bandwidth.

1. Introduction

It is well known that the cumulative income distribution is graphically represented by Lorenz curve (see Figure 1). On the latter, the percentage of households is plotted on the x-axis while the percentage of income on the y-axis. It shows for the bottom $p_1\%$ of households, what percentage $p_2\%$ of the total income they possess. This theory was initiated by Max O. Lorenz in 1905 in order to represents inequality in wealth distribution (see Cowell [7]). If $p_1 = p_2$, then the Lorenz curve is the upward diagonal line which means, for instance, that 50% of the households possess 50% of the total income. Thus the straight line represents perfect equality. Any case in which the Lorenz curve is not a straight line implies income inequality. The standard definition of the Lorenz curve is defined in two equivalent ways. Firstly, one has to determine a particular quantile, which means solving for z the equation:

$$L(F(z)) = \frac{1}{\mu}\int_0^z tf(t)dt$$

where

$$F(z) = \int_0^z f(t)dt \quad \text{and} \quad \mu = \int_0^\infty tf(t)dt.$$

Secondly, using a notation popularized by Gastwirth [8], $z = F^{-1}(p)$, one may write the Lorenz curve in a direct way :

$$L(p) = \frac{1}{\mu}\int_0^p F^{-1}(t)dt.$$

If everybody had the same income, the cumulative percentage of total income held by any bottom proportion p of the population would also be p. The Lorenz curve would then be $L(p) = p$: population shares and shares of total income would

Received January 8, 2022. May 22, 2022.

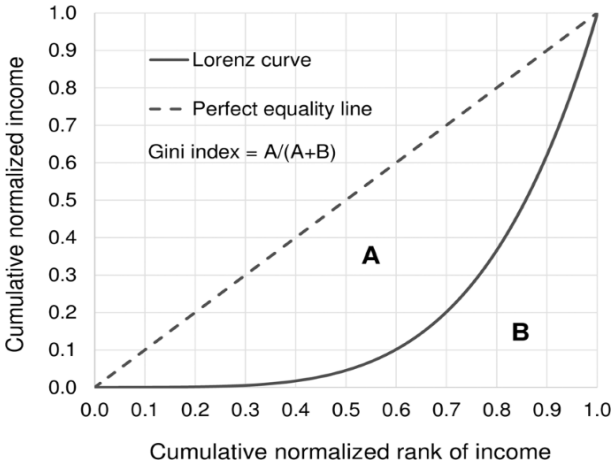

FIGURE 1. The area between the equality line and the Lorenz curve.

be identical. A useful informational content of a Lorenz curve is thus its distance, $p - L(p)$, from the line of perfect equality in income. Compared to perfect equality, inequality removes a proportion $p - L(p)$ of total income from the bottom $100 \times p\%$ of the population. The larger that " deficit ", the larger the inequality of income. There is thus an interest in computing the average distance between these two curves or the surface between the diagonal p and the Lorenz curve L(p). We know that the Lorenz curve is contained in the unit square having a normalized surface of 1. The surface of the lower triangle is 1/2. If we want to obtain a coefficient at values between 0 and 1, we must take twice the integral of $p - L(p)$ given by Lubrano [16]:

$$G = 2\int_0^1 (p - L(p))dp = 1 - 2\int_0^1 L(p)dp.$$

which is nothing but the usual Gini coefficient. Xu [21] gives a good account of the algebra of the Gini index. This definition above is an interpretation of the Gini index as a surface. The alternative definition of Gini index is in form of a mean of absolute differences. There are other formula too. All of these formula are equivalent. So the alternative formula for Gini index G, is based on the mean difference Δ, of the underlying distribution fonction $F(x)$ and is given by Kendall and al. [13]. The Gini index of the Lorenz curve $L(p)$ generated by a distribution fonction $F(x)$ is $G = \dfrac{\Delta}{2\mu}$, were:

$$\Delta = \int_{\mathbb{R}}\int_{\mathbb{R}} |x - y| dF(x)dF(y) \tag{1}$$

where y and x are two random variables of the same distribution F. As $F(x)$ and $1 - F(x)$ are simply the proportions of individuals with incomes below and above x, integrating the product of these proportions across all possible values of x gives again the Gini coefficient, in its forms given by Δ by (see Gastwirth [11]):

$$\Delta = 2\int_{\mathbb{R}} F(x)[1 - F(x)]dx = 4\int_{\mathbb{R}} x\left[F(x) - \frac{1}{2}\right]dF(x)$$

The formula G shows that the Gini index measures relatives inequality as it is the ratio of a mesure of dispersion, the mean difference to the average value (μ).

Inequality measures in general and Gini index in particular, have been used from a descriptive point of view. However, data available from statistical agencies frenquently come from sample surveys; inequality indices turn out to be computed on the basis of sample data. Therefore, it is necessary to use them not only as descriptive tools, but also as tools for formal statistical inference. The approch to statistical inference can be either nonparametric or parametric. A comprehensive survey of the main results in the estimation of G according to these two approaches is in Giorgi [9]. Conti and Giorgi [6] investigated the strong consistency of an estimator of the kernel Gini index. The Gini coefficient can be obtained from a simple ordinary least square regression based approach: see for instance Lerman and Yitzhaki [15], Shalit [19], Ogwang [17], Giles [10]. Shahryar and al. [20] investigated on the Gini Coefficient estimators based on the linearization and U-statistics methods. Also, some authors have proposed the resampling techniques to estimate the standard error of the Gini concentration index (see Berger [5] and Yitzhaki and Schechtman [22]).

The present paper provides a non-parametric estimator of the Gini coefficient based on the kernel method with varying bandwith parameter. Firstly, this varying bandwith parameter will vary according to the random variables and secondly, according to the variable x. The document is organized as follows; apart from the introduction and the conclusion, we will first go through the construction of the estimator and the study of its strong consistency, then a simulation study to conclude.

2. Gini index non parametric estimator

Let $X_1, \cdots, X_n$ be a random sample of size n from a population X with density function f. The X_i, for $i = 1, \cdots n$; $n \in \mathbb{N}$, are independent and identically distributed (i.i.d.) observations. The main of nonparametric density estimation is to estimate f with as few assumptions about f as possible. One of the well known estimators of f is the classical kernel density estimator, which we will denote by

$$f_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \quad x \in \mathbb{R}, \tag{2}$$

where $h = h_n$ is the bandwidth sequence satisfying:

$$h \longrightarrow 0 \quad \text{and} \quad nh \longrightarrow +\infty \text{ for } n \text{ large enough} \tag{3}$$

and K is a kernel function and is assumed to be a continuous density, symmetric with respect to 0. The kernel K satisfies the following conditions to get the order of the bias and variance of the classical kernel density estimator:

$$\int K(x)dx = 1, \quad \int xK(x)dx = 0, \quad \text{and} \quad \int x^2K(x)dx = \sigma_K^2 > 0. \tag{4}$$

One had remarked in the literature that the estimate (2) is more local in nature but is scarcely a reasonable estimate of a smooth density. Note that the problem of additive form (2) is that it require the preservation of the continuity and differentiability properties of K. For example, the uniform density is discontinuous, so the kernel density estimate based on a uniform kernel function is discontinuous. Thus, a smoother kernel function will thus lead to a smoother kernel density estimate. The

ordinary estimate (2) does not allow for different levels of smoothing at different parts of the density, as it is controlled by the single bandwidth h. Therefore is obviously not optimal. Jeffrey [12] shows that the mean squared error (MSE) of $f(x)$ at any point x is directly related to $f(x)/h$ and $[f''(x)]^2h^4$. In other words, in order to reduce MSE, h should increase with $f(x)$ (to reduce variance) and should decrease with $f''(x)$ (to reduce bias).

For filling this void of inadaptation, there is a way to vary h in the kernel estimator to try to improve performance, is to choose $h(X_i)$ as a function of the evaluation point X_i for $i \in \{1, \cdots, n\}$. From a practical point of view, the usual kernel density estimator (2) is susceptible to bumpiness in the tails, since it does not adapt to local variations in smoothness. The estimator can be generalized to allow this, by using broader windows for the contribution of values associated with regions of low density and narrower windows for values associated with regions of high density. The general formula for one such estimator, *the variable-bandwidth kernel estimator*, is defined in Jeffrey [12] by:

$$\hat{f}_n(x) = \frac{1}{nh(X_i)} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h(X_i)}\right), \quad x \in \mathbb{R}, \tag{5}$$

where $h(X_i) = \dfrac{h_n}{f^{1/2}(X_i)}$ vary inversely with the underlying density, since the goal is to smooth less where there is more structure (and more where there is less structure). The choice of (5) is particularly advantageous, since it results in the bias of (5) being $\mathcal{O}(h^4)$, rather than the usual (2) $\mathcal{O}(h^2)$, while leaving the variance $\mathcal{O}(n^{-1}h^{-1})$. Clearly, we observe that the choice of this variable h has a great influence on the speed of convergence of (5), which induces the rapid convergence of the estimator (6). We can also remark that if the density function f is uniform, then the formula (5) is reduced to (2).

From the estimator (5) of the density function, it is obvious that we obtain an estimators of the Gini index, which are given by:

$$\hat{G} = \frac{\hat{\Delta}_n}{2\hat{\mu}_n}. \tag{6}$$

where

$$\hat{\Delta}_n = \int_{\mathbb{R}} \int_{\mathbb{R}} |x - y| \hat{f}_n(x) \hat{f}_n(y) dxdy,$$

or

$$\hat{\Delta}_n = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \int_{\mathbb{R}} \int_{\mathbb{R}} |x - y| \frac{f^{1/2}(X_i)}{h_n} K\left[\frac{(x - X_i) f^{1/2}(X_i)}{h_n}\right] \frac{f^{1/2}(X_j)}{h_n} \times K\left[\frac{(y - X_j) f^{1/2}(X_j)}{h_n}\right] dxdy. \tag{7}$$

and

$$\hat{\mu}_n = \int_{\mathbb{R}} x \hat{f}_n(x) dx = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

In the next paragraph, we study the almost sure convergence or the strong consistency of our estimator (6).

3. Convergence of the Gini index estimator

For the following, apart from the classical regularity conditions on the kernel function K in (4) and on h (3), we adopt the following hypotheses.

3.1. Hypotheses.

(H.1) The density function $f : \mathbb{R} \longrightarrow \mathbb{R}_+$ is:

a. a bounded function: $\exists \quad 0 < m < M$ such that $m \leqslant f(x) \leqslant M \quad \forall x \in \mathbb{R}$,

b. a κ-Lipschitzian function: $|f(x) - f(y)| \leqslant \kappa|x - y| \quad \forall x, y \in \mathbb{R}$.

c. a function such that $\int_{\mathbb{R}} f^{1/2}(x)dx = \Theta_f > 0$.

(H.2) The Kernel $K : \mathbb{R} \longrightarrow \mathbb{R}_+$ satisfies the conditions below:

a. $\int_{\mathbb{R}} |u|K(u)du = \Omega_K > 0, \quad \forall u \in \mathbb{R}$,

b. $\int_{\mathbb{R}} \int_{\mathbb{R}} |u - v|K(u)K(v)dudv = \Delta_K > 0, \quad \forall u, v \in \mathbb{R}$,

(H.3) The variable bandwidth parameter h satisfies: $h_n\kappa\sqrt{M}\Omega_K \leqslant 2m^2$, $\forall n \in \mathbb{N}$.

These pre-enumerated hypotheses make it possible to study the strong consistency of the estimator $\hat{\Delta}$ (7), which induces that of $\hat{G}$ (6).

3.2. Almost sure convergence. Suppose $X_1 \cdots X_n$ be i.i.d random variables of a distribution F. Consider a parametric function θ for which there is an unbiased estimator. The parametric function θ may be represented as

$$\theta = \mathbb{E}[\phi(X_1, \cdots, X_m)] = \int \cdots \int \phi(x_1, \cdots, x_m)dF(x_1) \cdots dF(x_m),$$

where $\phi = \phi(x_1, \cdots, x_m)$ is function of $m(m \leqslant n)$ i.i.d random variables, called the kernel for θ. For any kernel ϕ, the corresponding U-statistic for estimating of θ on the basis of a random sample of size n is obtained by averaging the kernel ϕ symmetrically over the observations

$$U_n = U(X_1, \cdots, X_m) = \frac{1}{C_n^m} \sum_c \phi(X_{i_1} \cdots X_{i_1}),$$

where $\sum_c$ denotes summation over the C_n^m combinations of m distinct elements $\{i_1 \cdots i_m\}$ from $\{1, \cdots, n\}$. In particular cases, we have:

Lemma 3.1. *For n large enough, we have*

- $$\hat{\mu}_n = U_1 = \frac{1}{n} \sum_{i=1}^{n} X_i \longrightarrow \mu = \mathbb{E}(X) = \int xdF(x) \quad p.s., \tag{8}$$

where $\phi(x) = x$.

•

$$U_2 = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} |X_i - X_j| \longrightarrow \Delta = \mathbb{E}(|X_1 - X_2|) = \int \int |x_1 - x_2| dF(x_1) dF(x_2) \quad p.s., \tag{9}$$

where $\phi(x_1, x_2) = |x_1 - x_2|$.

Proof. This proof can easily be found in Lehmann [14] and Pranab [18]. □

The following lemma plays an important role in the convergence of the estimator.

Lemma 3.2. *Under hypothesis (H.1), the function* $f^{-1/2}$ *is* δ_κ*-Lipschitzian, that is to say:*

$$\left| \frac{1}{f^{1/2}(x)} - \frac{1}{f^{1/2}(y)} \right| \leqslant \delta_\kappa |x - y| \quad \forall x, y \in \mathbb{R}, \quad \textit{where} \quad \delta_\kappa = \frac{\kappa \sqrt{M}}{2m^2}.$$

Proof. Indeed, it suffices to note that:

$$\left| \frac{\sqrt{x} - \sqrt{y}}{x - y} \right| = \frac{1}{\sqrt{x} + \sqrt{y}} \leqslant \frac{1}{2\sqrt{\epsilon}} \quad \forall x, y \geqslant \epsilon > 0, \tag{10}$$

and

$$\left| \frac{1}{x} - \frac{1}{y} \right| = \left| \frac{1}{xy} \right| |x - y| \leqslant \frac{1}{\epsilon^2} |x - y| \quad \forall x, y \geqslant \epsilon > 0. \tag{11}$$

From the inequalities (10), (11) and under the hypothesis (H.1), we can write

$$\begin{aligned} \left| \frac{1}{f^{1/2}(x)} - \frac{1}{f^{1/2}(x)} \right| &\leqslant \frac{\sqrt{M}}{2} \left| \frac{1}{f(x)} - \frac{1}{f(y)} \right| \\ &\leqslant \frac{\sqrt{M}}{2m^2} |f(x) - f(y)| \\ &\leqslant \frac{\kappa \sqrt{M}}{2m^2} |x - y|. \end{aligned}$$

□

Theorem 3.3. *Under the assumptions (H.1), (H.2) and (H.3), we have*

$$|\hat{\Delta}_n - \Delta| \longrightarrow 0 \quad p.s, \quad \textit{when} \quad n \longrightarrow +\infty.$$

Proof.

$$\begin{aligned} \hat{\Delta}_n &= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \int_{\mathbb{R}} \int_{\mathbb{R}} |x - y| \frac{f^{1/2}(X_i)}{h_n} K \left[\frac{(x - X_i) f^{1/2}(X_i)}{h_n} \right] \frac{f^{1/2}(X_j)}{h_n} K \left[\frac{(y - X_j) f^{1/2}(X_j)}{h_n} \right] dxdy \\ &= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \int_{\mathbb{R}} \int_{\mathbb{R}} \left| \frac{u h_n}{f^{1/2}(X_i)} - \frac{v h_n}{f^{1/2}(X_j)} + (X_i - X_j) \right| K(u) K(v) du dv \\ &= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \int_{\mathbb{R}} \int_{\mathbb{R}} \left| (X_i - X_j) + \frac{h_n}{f^{1/2}(X_i)} (u - v) + \left(\frac{1}{f^{1/2}(X_i)} - \frac{1}{f^{1/2}(X_j)} \right) v h_n \right| K(u) K(v) du dv \\ &\leqslant \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \int_{\mathbb{R}} \int_{\mathbb{R}} \left[|X_i - X_j| + \frac{h_n}{f^{1/2}(X_i)} |u - v| + \left| \frac{1}{f^{1/2}(X_i)} - \frac{1}{f^{1/2}(X_j)} \right| |v| h_n \right] K(u) K(v) du dv. \end{aligned}$$

From the Lemma 3.2 and assumptions (H.1) - (H.2) we deduce

$$\begin{aligned}\hat{\Delta}_n &\leqslant \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int_{\mathbb{R}}\int_{\mathbb{R}}\left[|X_i - X_j| + \frac{h_n}{f^{1/2}(X_i)}|u-v| + \delta_\kappa|X_i - X_j||v|h_n\right]K(u)K(v)dudv \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[|X_i - X_j| + \frac{h_n}{f^{1/2}(X_i)}\int_{\mathbb{R}}\int_{\mathbb{R}}|u-v|K(u)K(v)dudv\right. \\ &\quad \left. + h_n\delta_\kappa|X_i - X_j|\int_{\mathbb{R}}|v|K(v)dv\right] \\ &\leqslant \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[|X_i - X_j| + \frac{h_n}{\sqrt{m}}\Delta_K + h_n\delta_\kappa\Omega_K|X_i - X_j|\right] \\ &\leqslant \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}(1 + h_n\delta_\kappa\Omega_K)|X_i - X_j| + \frac{h_n}{\sqrt{m}}\Delta_K\end{aligned}$$

Using hypothesis (H.3) and noticing that $|a-b| \leqslant |\,|a| - |b|\,|$, this leads to

$$|\hat{\Delta}_n - U_2| \leqslant h_n\frac{\Delta_K}{\sqrt{m}}. \tag{12}$$

Thus, considering the first limit of equality (3), we get $|\hat{\Delta}_n - U_2| \longrightarrow 0 \quad p.s.$

Moreover, the triangular inequality allows us to write

$$|\hat{\Delta}_n - \Delta| \leqslant |\hat{\Delta}_n - U_2| + |U_2 - \Delta|.$$

Relation (9) of Lemma 3.1 completes the proof of the theorem. □

Corollary 3.4. *Under the assumptions of Theorem 3.3, we have*

$$|\hat{G}_n - G| \longrightarrow 0 \quad p.s, \quad when \quad n \longrightarrow +\infty.$$

Proof.

$$|\hat{G}_n - G| = \frac{1}{2}\left|\frac{\hat{\Delta}_n}{\hat{\mu}_n} - \frac{\Delta}{\mu}\right| \leqslant \frac{1}{2}\left[\frac{1}{\hat{\mu}_n}|\hat{\Delta}_n - \Delta| + \left|\frac{1}{\mu} - \frac{1}{\hat{\mu}_n}\right|\Delta\right]$$

From relation (8) of Lemma 3.1 we get $\left|\frac{1}{\mu} - \frac{1}{\hat{\mu}_n}\right| \longrightarrow 0 \quad p.s.$

And finally Theorem 3.3 completes the proof of this corollary. □

Corollary 3.5. *Under the assumptions of Theorem 3.3, $\hat{G}_n$ is an asymptotically unbiased estimator i.e.,*

$$\mathbb{E}(\hat{G}_n) = G \quad a.s. \quad as\ the\ sample\ size\ tends\ to\ infinity.$$

Proof. From the expression

$$\hat{\Delta}_n = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int_{\mathbb{R}}\int_{\mathbb{R}}\left|h_n\left(\frac{u}{f^{1/2}(X_i)} - \frac{v}{f^{1/2}(X_j)}\right) + (X_i - X_j)\right|K(u)K(v)dudv,$$

using Fubini's theorem, we can write,

$$\mathbb{E}(\hat{\Delta}_n) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int_{\mathbb{R}}\int_{\mathbb{R}}\mathbb{E}\left[\left|h_n\left(\frac{u}{f^{1/2}(X_i)} - \frac{v}{f^{1/2}(X_j)}\right) + (X_i - X_j)\right|\right]K(u)K(v)dudv,$$

by noting that $|a+b| \leqslant |a|+|b|$, $|a-b| \leqslant |a|+|b|$ and by considering the fact that the random variables are i.i.d., we have

$$\mathbb{E}(\hat{\Delta}_n) \leqslant \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int_{\mathbb{R}}\int_{\mathbb{R}} \mathbb{E}\left[h_n(|u|+|v|)\mathbb{E}\left[f^{-1/2}(X)\right]+\mathbb{E}|X_i-X_j|\right]K(u)K(v)dudv,$$

hypothesis (H.1) c, leads to

$$\mathbb{E}(\hat{\Delta}_n) \leqslant \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}|X_i-X_j| + \Theta_f h_n \int_{\mathbb{R}}\int_{\mathbb{R}} (|u|+|v|)\,K(u)K(v)dudv.$$

Taking into account that $\mathbb{E}|X_i-X_j| = \Delta$, using Jensen's inequality expectation and Fubini's theorem, hypothesis (H.2) a, lead to

$$\mathbb{E}|\hat{\Delta}_n - \Delta| \leqslant 2\Theta_f \Omega_K h_n,$$

or

$$\mathbb{E}|\hat{\Delta}_n - \Delta| = \mathcal{O}(h_n). \tag{13}$$

By applying linearity of the expectation, the last equality (13) and Lemma 3.1 to the following relation

$$|\hat{G}_n - G| = \frac{1}{\mu}|\hat{\Delta}_n - \Delta| + \left|\frac{1}{\hat{\mu}_n} - \frac{1}{\mu}\right|\hat{\Delta}_n,$$

we have

$$\mathbb{E}|\hat{G}_n - G| \longrightarrow 0 \quad \text{a.s.} \quad \text{when} \quad n \longrightarrow \infty.$$

In particular,

$$\mathbb{E}(\hat{G}_n) \longrightarrow \mathbb{E}(G) = G \quad \text{a.s.} \quad \text{when} \quad n \longrightarrow \infty.$$

□

4. Applications

4.1. Gini index and Lorenz curves generated by some common distributions. The table below proposes theoretical expressions of the Lorenz curve and of the Gini index for certain usual probability laws.

Distribution	**Formula of $F(x)$**	**Lorenz curve $L(p)$**	**Gini index G**
Equal	$1_{[\mu,+\infty[}(x)$	p	0
Exponential	$1-e^{-\lambda x},\ x>0$	$p+(1-p)\ln(1-p)$	$\frac{1}{2}$
Shifted Exponential	$1-e^{-\theta(x-a)},\ x \geqslant a>0$	$p+\frac{1}{1+\theta a}(1-p)\ln(1-p)$	$\frac{1}{2(\theta a+1)}$
General Uniform	$\frac{x-a}{\theta},\ a<x<a+\theta$	$\frac{ap+\frac{\theta p^2}{2}}{a+\frac{\theta}{2}}$	$\frac{\theta}{3(2a+\theta)}$
Pareto	$1-\left(\frac{\beta}{x}\right)^{\alpha-1},\ x \geqslant \beta>0,\ \alpha>2$	$1-(1-p)^{\frac{\alpha-2}{\alpha-1}}$	$\frac{1}{2\alpha-3}$

Table 1: Some Gini index and Lorenz curves.

Among the above probability laws, only the Pareto and exponential (shifted) distributions satisfy the hypotheses (H.1), (H.2) and (H.3). In addition, we can notice that the exponential distributions (shifted) give a theoretical Gini index strictly lower (for the shifted exponential distribution) or equal (for the exponential distribution) to 0.5, that is to say that the Gini index theoretical result from the two above-mentioned distributions, gives an almost equal distribution of income (because its value is closer to 0). Thus exponential (shifted) distributions do not allow us to make a good judgment on the real data we have. However, the Pareto distribution gives a Gini index between 0 and 1 and it is also a distribution that is very useful in the literature for studies on inequalities.

4.2. Application to real data. The data we have was collected on the World Bank website [23]. We collected data on adjusted net national income from 45 countries in Africa (sub-Saharan Africa) and 185 countries around the world. The data collected is the only data that exists on the site since not all countries have it. The probability law considered is the Pareto distribution which has two parameters α and β. The two parameters of the Pareto distribution being unknown, in order not to choose these two parameters at random for our study, we chose to estimate the two parameters by the maximum likelihood method and these two unbiased and convergent estimators will represent the parameters α and β for the Pareto distribution in our study. The following table provides the expressions of the two parametric estimators and their characteristics.

Parameters ρ	**MLE estimators** $\hat{\rho}_n$	$E(\hat{\rho}_n)$	$\sigma^2(\hat{\rho}_n)$
α	$\hat{\alpha}_n = \dfrac{(n-1)\left[1+\dfrac{n}{\sum_{i=1}^{n}\log\left(\frac{X_i}{\beta}\right)}\right]+1}{n}$	α	$\dfrac{(\alpha-1)^2}{n-2}$
β	$\hat{\beta}_n = \left[1-\dfrac{1}{n(\alpha-1)}\right]X_{(1)}$	β	$\dfrac{\beta^2}{n(\alpha-1)(n(\alpha-1)-2)}$

Table 2: Maximum Likehood Estimators of Pareto distribution parameters.

Since in the expressions of $\hat{\alpha}_n$ and $\hat{\beta}_n$ the parameters α and β are unknown, in practice, we choose as estimators of α and β the quantities $\hat{\alpha}_n^*$ and $\hat{\beta}_n^*$ defined by:

$$\hat{\alpha}_n^* = \frac{(n-1)\left[1+\dfrac{n}{\sum_{i=1}^{n}\log\left(\frac{X_i}{X_{(1)}}\right)}\right]+1}{n} \quad \text{and} \quad \hat{\beta}_n^* = \left[1-\frac{1}{n(\hat{\alpha}_n^*-1)}\right]X_{(1)}.$$

For the simulations, we have selected four kernel functions, two of which have compact support and the others defined on $\mathbb{R}$. The smoothing parameter $h(\bullet)$ is chosen such that $h_n = \left[\dfrac{R(K)}{\sigma_K^4 R(\hat{f}_n'')}\right]^{1/9} n^{-1/9}$ or more simply $h_n = \mathcal{O}(n^{-1/9})$ where

the notation follows the convention $R(g) = \int g^2(x)dx$ for appropriate functions g. For the determination of h_n, the estimator $\hat{f}_n$ used is based on a different bandwidth from the one that is appropriate for the estimation and is estimated from the data (see Jeffrey[12]).

The different kernel functions used are grouped together in the following table:

Kernel name	**Formula of $K(u)$**
Epanechnikov	$\frac{3}{4}\left(1-u^2\right) 1_{\lvert u\rvert\leqslant 1}$
Cosine	$\frac{\pi}{4}\cos\left(\frac{\pi}{2}u\right) 1_{\lvert u\rvert\leqslant 1}$
Gaussian	$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^2\right)$
Logistic	$\frac{1}{\exp(-u)+2+\exp(u)}$

Table 3: Some usefull Kernels.

Using the Matlab software, we obtain the following results, shown in Table 4 below. Recall that the values of the parameters α and β come from the samples collected and are such that $\alpha \simeq \hat{\alpha}_n^*$ and $\beta \simeq \hat{\beta}_n^*$. For these two values obtained, we tried to construct the Lorenz curve given by Figure 2 for the two types of samples ($n = 45$ and $n = 185$). Then, we calculated the theoretical Gini index G given by the formula $G = \frac{1}{2\alpha - 3}$, in order to compare it with the estimated Gini index, noted $\hat{G}_n$.

n	$K(u)$	α	β	$\hat{\mu}_n$	$\hat{\Delta}_n$	$\hat{G}_n$	G	Bias$(\hat{G}_n^*, G)$
45	Epanechnikov	2.1500	1.8352e+05	2.9138e+07	4.2175e+07	0.7237	0.7692	0.0455
	Cosine				4.3574e+07	0.7477		0.0215
	Gaussian				4.4001e+07	0.7550		0.0142
	Logistic				4.4164e+07	0.7578		0.0114
185	Epanechnikov	2.0599	3.8915e+06	7.4388e+10	1.3266e+11	0.8916	0.8930	0.0013
	Cosine				1.3265e+11	0.8915		0.0014
	Gaussian				1.3283e+11	0.8928		0.0002
	Logistic				1.3279e+11	0.8926		0.0004

Table 4: Simulation results.

4.3. Interpretation of results and commentary. Table 4 shows that for $n = 45$ (sample africa), convergence is slow while with $n = 185$ (sample world) almost four times the size of sample africa, convergence is rapid. Moreover, we can observe in both situations that the choice of the kernel has a slight influence on the convergence of the Gini estimator, since one can notice that the kernels with compact support, provide a slower convergence compared to the others. In other words, we can summarize that

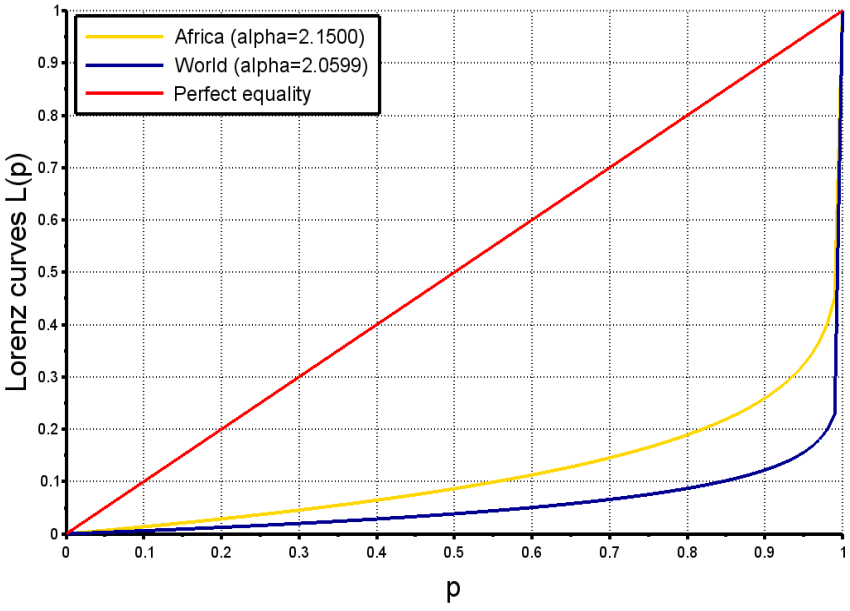

FIGURE 2. Theorical Lorenz curves for Africa and World.

the size of the sample and the choice of the kernel contribute to a faster convergence without forgetting the choice of the smoothing parameter which is of great importance. In view of the results obtained, we can notice that the adjusted net national incomes are distributed unequally in the two cases but the inequality obtained with the 185 countries of the world is more severe. This result is predictable because only about 20 countries hold most of the income in the world.

5. Conclusion

The nonparametric estimator of the Gini index studied in this manuscript allowed us to see the importance of the variable smoothing parameter which played a primordial role in this study, especially with regard to the speed of convergence. Choosing the variable smoothing parameter makes it possible to take into account the variability of the data collected, even if the theoretical calculations and especially the computer programming are heavier with this choice (variable smoothing parameter), nevertheless it contributes to very good convergence and faster of the estimator compared to the constant smoothing parameter which is much more used in the literature (see Agbokou and al.[1, 2, 3, 4]). Regarding the simulations, we planned to take several samples but for lack of a calculator, we decided to select only the two samples (Africa

and the world) to evaluate the performance of our estimator and in view of the results obtained, we conclude that this Gini estimator is efficient even if it can still be improved so that with $n = 50$, convergence is faster. In our next study, we will compare the estimator of the Gini index with variable smoothing parameter to the Zenga index with a constant smoothing parameter. In a brief way, let us recall that the main objective of this work is to highlight the rapid convergence using a variable bandwidth in terms of simulations and to see if there is a conformity with the theory. Future work will focus on asymptotic normality (Central Limit Theorem) to get an idea of the behavior of the estimator in terms of bias and variance and in addition to study the adequacy with simulations.

References

[1] K. Agbokou and K.E. Gneyou, On the strong convergence of the hazard rate and its maximum risk point estimators in presence of censorship and functional explanatory covariate, *Afrika Statistika* **12** (2017) (3), no.3,1397–1415.

[2] K. Agbokou, K.E. Gneyou, and E. Deme, Almost sure representations of the conditional hazard function and its maximum estimation under right-censoring and left-truncation, *Far East Journal of Theorical Statistics* **54** (2018), no 2, 141–173.

[3] K. Agbokou and K.E. Gneyou, On the asymptotic distribution of non-parametric conditional quantile estimator under random censorship, *Far East Journal of Theorical Statistics* **56** (2019), no. 1, 1–34.

[4] K. Agbokou and K.E. Gneyou, Asymptotic properties of the conditional hazard function and its maximum estimation under right-censoring and left-truncation, *Jordan Journal of Mathematics and statistics* **12** (2019), no.3, 351–374.

[5] Y.G. Berger, A note on the asymptotic equivalence of jackknife and linearization variance estimation for the Gini coefficient, *Journal of Statist.* **24** (2008), no. 1, 541–555.

[6] P.L. Conti and G.M. Giorgi, Distribution - free estimation of the Gini inequality index: the kernel method approach, *Statistica* **LXI** (2001), no. 1.

[7] F. Cowell, *Measuring Inequality. LSE Handbooks on Economics Series*, Prentice Hall, London, 1995.

[8] J.L. Gastwirth, A general definition of the lorenz curve, *Econometrica* **39** (1971), no. 6, 1037–1039.

[9] G.M. Giorgi, Income inequality measurement: the statistical approach, In: (J. Silber ed.) *Handbook of income inequality measurement*, Kluwer Academic Publishers, Boston, 1999, 245–260.

[10] D.E. Giles, A cautionary note on estimating the standard error of the Gini index of inequality: comment, *Oxford Bulletin of Economics and Statistics* **68** (2006), no. 1, 395–396.

[11] J.L. Gastwirth, The estimation of the Lorentz curve and Gini index, *The Review of Economics and Statistics* **54** (1972), no. 3, 306–316.

[12] S. Jeffrey, *Smoothing Methods in Statistics*, Springer, New York University, 1996.

[13] M.G. Kendall and A. Stuart, *The Advanced Theory of Statistics 1*, 2nd edition,Charles Griffen and Company, London, 1963.

[14] E.L. Lehmann, *Elements of Large-Sample Theory*, Springer Texts in Statistics, 1998.

[15] R.I. Lerman and S. Yitzhaki, A note on the calculation and interpretation of the Gini index, *Economics Letters de Estadistica* **15** (1984), 363–368.

[16] M. Lubrano, The econometrics of inequality and poverty measurement, lecture notes, 2017. https://perso.amse-aixmarseille.fr/lubrano/cours/Lecture-9.pdf

[17] T. Ogwang, A convenient method of computing the Gini index and its standard error, *Oxford Bulletin of Economics and Statistics* **47** (2000), 123–129.

[18] K.S. Pranab and M.S. Julio, *Large sample methods in statistics: an introduction with applications*, Springer-Science+Business Media, B. V., 1993.

[19] H. Shalit, Calculating the Gini index of inequality for individual data, *Oxford Bulletin of Economics and Statistics* **47** (1985), 185–189.

[20] M. Shahryar, M.B. Gholam Reza, and A. Mohammad, A Comparative Study of the Gini Coefficient Estimators Based on the Linearization and U-Statistics Methods, *Rev. Col. de Estad.* **40** (2017), no. 2, 205–221.
[21] K. Xu, *How has the literature on Gini's index evolved in the past 80 years?* Economics working paper, Dalhousie University, 2003.
[22] S. Yitzhaki and E. Schechtman, *The Gini Methodology: A primer on a Statistical Methodology*, Springer, New York, 2013.
[23] https://data.worldbank.org/indicator/NY.ADJ.NNTY.KD?locations=1W&most_recent_value_desc=true

(Komi Agbokou) Department of Mathematics, University of Lome, Togo
E-mail address: ffomestein@gmail.com

(Yaogan Mensah) Department of Mathematics, University of Lome, Togo
E-mail address: mensahyaogan2@gmail.com