Working Paper Series

A 1-1-1 relationship for World Bank Income Data and the Gini

Franz Josef Radermacher
Estelle Herlyn

ECINEQ WP 2018 - 473

ECINEQ 2018 - 473
July 2018

www.ecineq.org

A 1-1-1 relationship for World Bank Income Data and the Gini*

Franz Josef Radermacher†
Institute for Applied Knowledge Processing (FAW/n)

Estelle Herlyn
FOM University of Applied Science

Abstract

The paper provides insights of significant practical relevance into the nature of empirical income distri-bution data, provided by World Bank (and by EU-SILC as well). The insight covers mature states and their economic and societal system. Proceeding from a Gini value .., as published by the World Bank, it is possible to derive, as close approximation, the actual income distribution, with standardised total income 1, and a mathematical representation. We call this the standard Lorenz curve L_G. L_G is of type $L_G = 0.6 \cdot Pareto(\varepsilon) + 0.4 \cdot Polynomial(\varepsilon)$, where $Pareto(\varepsilon)$ and $Polynomial(\varepsilon)$ are the Pareto and polynomial Lorenz curves for a parameter ε with $\varepsilon = \frac{1-G}{1+G}$ and $G = \frac{1-\varepsilon}{1+\varepsilon}$. If the total income level of the considered distribution is known, then the distribution of absolute income can also be derived. If, in addition, one knows the number of income earners, then one also knows the distribution of the absolute income within a population. All together our summarizing statement is: "For mature economies, analysing World Bank and EU-SILC income data, there is essentially a cross-country and cross-year 1-1-1 correspondence between the GINI, the corresponding decile resp. quintile information and the respective standard Lorenz curve described above." Some interesting mathematics is involved to reach the main result. The insights obtained will hopefully enable economists and social scientists to further develop their work in the field of income inequality and associated social phenomena.

Keywords: Approximation, income inequality, Lorenz curves, World Bank data.

*In memoriam Holger M. Hinkel and Thomas Kämpke.

†**Contact details:** Prof. Dr. Dr. Dr. h.c. Franz Josef Radermacher (corresponding author), Institute for Applied Knowledge Processing (FAW/n), Lise-Meitner-Str. 9, D-89081 Ulm, Phone +49 731 50 39-100, Fax +49 731 50 39-111, E-Mail: radermacher@faw-neu-ulm.de; Prof. Dr. Estelle Herlyn, FOM University of Applied Science, Toulouser Allee 53, D-40476 Düsseldorf, Phone +49 211 3611799-936, E-mail: estelle.herlyn@fom.de.

Introduction

The development of income distributions and the "widening gap" over time are important topics of the current political-social debate, especially in the light of the Brexit decision in Great Britain and the most recent presidential election in the USA. The consequences of these developments threaten the open world market and prerequisites for growing prosperity. These developments reflect the fact, that globalisation generates losers as well as winners. A strong negative attitude is building up against the current system by considerable parts of the population. It is directed against the economic winners of the status quo, and against the political actors viewed as responsible for this development. Either by promoting it or by letting it happen. At the World Economic Forum in Davos in early 2017 and 2018, social polarisation was identified and articulated as the greatest current threat to the existing international order, followed by environmental risks.

The issue of social polarisation is a difficult one. There is a political, but also scientific debate concerning basic concepts, relevant data and the availability of data. The present paper deals with income distributions, Lorenz curves and the so-called Gini value. The main insight consists of the following: For all World Bank data concerning empirical income distributions (and for EU-SILC data as well), there is, cross-country and cross-year, essentially a **1-1-1 relationship** between the respective Gini values and the other data. E.g., this concerns the Gini and six quintiles / deciles in the World Bank case and the Gini and 9 deciles in the EU-SILC case. It further more concerns a so-called standard Lorenz Curve that we discuss below. This is all surprising, as usually quite different income distributions may have the same Gini. We analyse the special behaviour observed for the World Bank data (and EU-SILC data). We develop mathematical formulas for approximating Lorenz curves (so-called standard Lorenz curve) of the respective income distributions and point at interesting applications which now become tractable while they could not be treated before because of missing data.

This paper builds upon previous publications on the topic with participation of the authors from the last 15 years (cf. e.g. Kämpke et al. (2003)). These address, for example, the so-called **efficient inequality range** (Cornia and Court 2001): societies should aspire to have Gini values between 0.25 and 0.35. Gini values ≤ 0.25 and Gini values ≥ 0.35 generally have no positive effect on societies. Too much "equality" of income apparently hinders social dynamism; too little hinders the full development of the human potential of a society.

The paper gives references to related work but essentially **concentrates** on establishing the mentioned link between the Gini and the World Bank data (and EU-SILC data). This special insight can be understood quite in itself, e.g. it does not need much input from the literature. This may make it easier for many readers to understand the observations made.

The paper is organised as follows. In section 1, we deal with **'Concepts and data sources**'. In Section 2 **'From the Gini to the Lorenz curve – the main findings'**, we give our main result which combines mathematics with empirical findings. Section 3 **'Limitations of applicability'** shows where limits of the main empirical insights have to be taken into account. This means that in any case, when going from a Gini to a Lorenz curve, even in the World Bank (and EU-SILC) case and mature states, one has to be careful. Finally, we also refer to a software tool under development for the general public.

1. Concepts and data sources

Income distributions represent a complex subject area, both in terms of their mathematical description and in terms of empiricism. The complexity begins right at the point of definition. As a rule, income distributions are derived in a standardised manner from the (adjusted) household incomes collected in the context of taxation, or alternatively (or additionally), from consumer surveys. The tax and consumption data are apportioned to the members of households using different methods.

Details of income distribution are usually not recorded statistically – especially the details on very high incomes (1% quantile, 1 ‰ quantile). For a long time, there was a tendency to keep details of top income out of the discussion, perhaps so as to avoid a so-called 'envy debate'. Instead, there was often a broad focus on low income and the subject of 'poverty', although only a comparatively small share of the total income is allocated there as a whole. [3]

Clarity about the very high income segment, which is important for the coherence of societies, was first provided by research papers by Thomas Piketty (2016), Anthony Atkinson (2015), Facundo Alvaredo and Emmanuel Saez, together with other researchers (2013). With painstaking detail, and using a variety of methodological approaches, they have compiled a corresponding data collection (World Top Incomes Database, since 2015 then World Wealth and Income Database, WID, http://wid.world/).

With regard to the existing standardised information on income distributions in the above sense, there are a limited number of data sources, with partially differing values. In Galbraith et al. (2015), 5 such important sources are discussed. The World Bank data (WDI) considered here and the European SILC data are included. Our main insight considers World Bank data and – to some extent – EU-SILC data. Whether similar observations are true for the other data sources looked at in Galbraith et al. (2015) is not known to us. This is a topic for future research.

We also do not know, whether taking into account average income within quantiles would change the picture. Krieger (1984), in a review of the literature on grouped data, demonstrates that grouping leads to a major loss of information, which is no surprise. Lyon et al. (2016) demonstrate that having averages available increases accuracy for US income data. Our problem is that data on average income within quantiles (as e.g. given by the World Bank) is not generally available. Therefore, this is not the focus of our research. Our focus is World Bank data and, to some extent, EU-SILC data, as it is.

---

[3] Inequality is dealt with in the sense of social stratification, not inequality between regions of a country. In Milanovic (2012), it is shown, that for the USA, the differences in averaged income between the US federal states are much smaller than those between the individual EU states. On the other hand, socially conditioned differences are significantly higher in each of the US states than in the EU states. The resulting total inequality in the USA and in Europe is similar according to Milanovic (2012), at a Gini value of around 0.4 (2007).

For that reason, we also did not look into approximations of income data by lognormal distributions. More on this can be found in Ultsch and Lösch (2017). There, one can also find a data based standardized Gini index. Again, that is not our topic of interest.

Our interest is the 'normal' Gini value. We just deal with the aforementioned data sets (e.g. World Bank data and EU-SILC data) and are happy to derive a major observation for this case. The result is of a type that other approximations for the same data sets, which of course exist (e.g. in particular the generalized Tukey's lambda distribution, see below), will not add much insight in this special case.

Why is World Bank data of such high interest from our point of view? When looking back for many years or even decades, there is only limited data material available in all the sources – mostly only information about quantiles. We think it is fair to say that the **most important international reference database**, with information on states of all five continents, is provided by the World Bank (WDI) [http://databank.worldbank.org]. This lists the 10%, 20%, 40%, 60%, 80% and 90% quantiles of income distributions for numerous states and years, and in addition (as a quite independent further factor), the respective Gini coefficient. The EU-SILC data contains all deciles and the Gini as well.

The World Bank data combines tax data and data from consumer surveys according to a developed methodology. The World Bank has the habit of delivering data on a per capita basis when measuring consumption or inequality or poverty. Eurostat distributes survey data within SILC with no adjustment, as well. However, when dealing with general income issues and poverty, it uses **equivalised** income according to the new OECD-modified equivalence scale. This can lead to major differences between the World Bank data and the EU-SILC data. Equivalising means, that the first adult person is weighted by a factor of 1, further adults and children over 14 years by a factor of 0.5, and children under 14 years by a factor of 0.3 (Eurostat Statistics Explained 2017). This sometimes makes families significantly 'richer' than when just dividing by the number of family members. We give hints to the effects of the differences between World Bank data and EU-SILC data later.

The most important groundwork for the present text, that we use, is a monograph (Kämpke and Radermacher 2015), which deals in particular with a **mathematical foundation of the theory of Lorenz**

curves on the basis of the theory of measurable functions and probability distributions. There, in line with Gastwirth's approach (1971), the Lorenz curve is defined using the generalised inverse of the respective distribution function. The monograph (Kämpke and Radermacher 2015) contains all the mathematical foundations required for this text. These are also laid out in shortened form in a recent publication (Radermacher 2016). For a number of topics addressed, references can be found in (Herlyn 2012). Note in particular, that in order to obtain a Lorenz curve from an income distribution (understood as a probability distribution, given by its distribution function F), we perform the following steps. Whether, on this way, the Lorenz curve can be obtained in closed form depends on the generalized inverse and the integrals over it being explicitly computable.

$$\boxed{\text{Distribution function } F(x)}$$
$$\downarrow$$
$$\boxed{\text{Generalized inverse } F^{-1}(u)}$$
$$\downarrow$$
$$\boxed{\text{Lorenz density } l(u) = \frac{F^{-1}(u)}{\int_0^1 F^{-1}(v)\,dv}}$$
$$\downarrow$$
$$\boxed{\text{Lorenz curve } L(u) = \frac{\int_0^u F^{-1}(v)\,dv}{\int_0^1 F^{-1}(v)\,dv}}$$

The aforementioned Gini coefficient $G(F)$ is the most important and most commonly used single parameter for describing inequality with regard to an income distribution F. Its value equates to 2 times the size of the differential area between the so-called Lorenz curve $L = L(F)$ (of income distribution F) and the (principle) diagonal of the unit square. The value $L(x) \in [0,1]$ of the Lorenz curve describes for all $x \in [0,1]$ the cumulative share $L(x)$ of the total income belonging to the x% of the population with the lower incomes. The notation $G(F) = G(L)$ is also used for the Gini coefficient.

Obviously, a Lorenz curve L always lies below the (principle) diagonal of the unit square. In the case of a uniform distribution (identical income for all), both curves coincide; the Gini value $G(L)$ is then zero (see Appendix).

At maximum inequality, the Lorenz curve coincides with the x-axis (value 0) and then, at the point $x = 1$, jumps to the value $L(1) = 1$. The associated Gini value $G(L)$ is then one.

Lorenz curves have a multitude of interesting mathematical properties (see for instance Sarabia 1997 as reported in Sarabia (2008)). In Kämpke and Radermacher (2015) some of those properties are listed

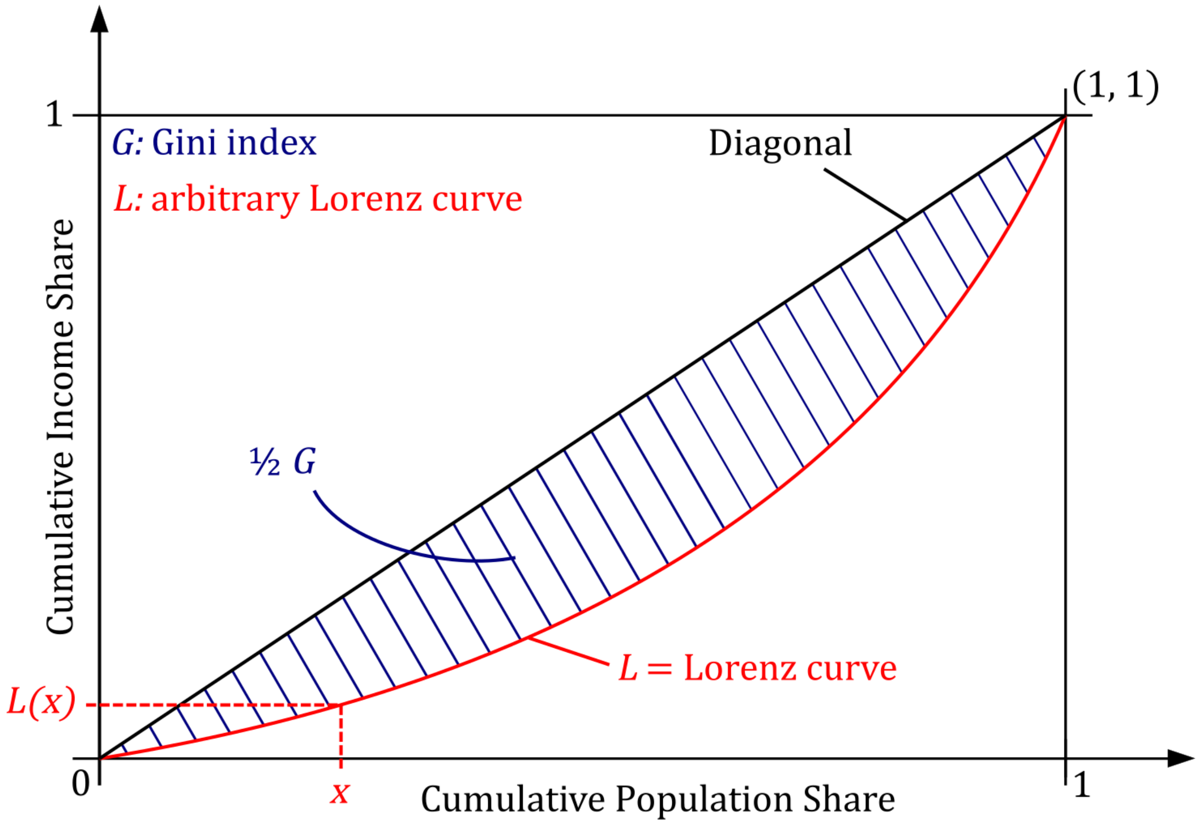

Figure 1: Lorenz curve ***L*** and its associated Gini index.

with reference to the literature. Here, we mention the following: They are monotonically increasing, continuous, and convex over [0,1], and both $L(0) = 0$ and $L(1) = 1$ are true. The inverse is also true, i.e., every function with the properties mentioned is the Lorenz curve of a suitable distribution. In addition, it holds that if distributions have the same Lorenz curve, they are identical except for a multiplicative factor.

In Kämpke and Radermacher (2015), one can also find that (1) the reflection L^{ref} of a Lorenz curve L on the secondary diagonal of the unit square is also a Lorenz curve. As straightforwardly verified, for the case of certain incomes being '0', the Lorenz curve is not everywhere strictly monotonically increasing. The definition of L^{ref} then needs a minor modification. All results mentioned below, remain however valid. $L^{ref}(u) = 1 - L^{-1}(1 - u)$ for $u \in [0,1]$. One finds (2) that the Gini coefficients of both

Lorenz curves are the same (by comparison of the two surfaces or by analysis, see Appendix). Note, that Taguchi (1968) deals with reflections as well. He is interested in the case that a Lorenz curve and its reflection are identical (he calls this a self-symmetrical Lorenz curve).

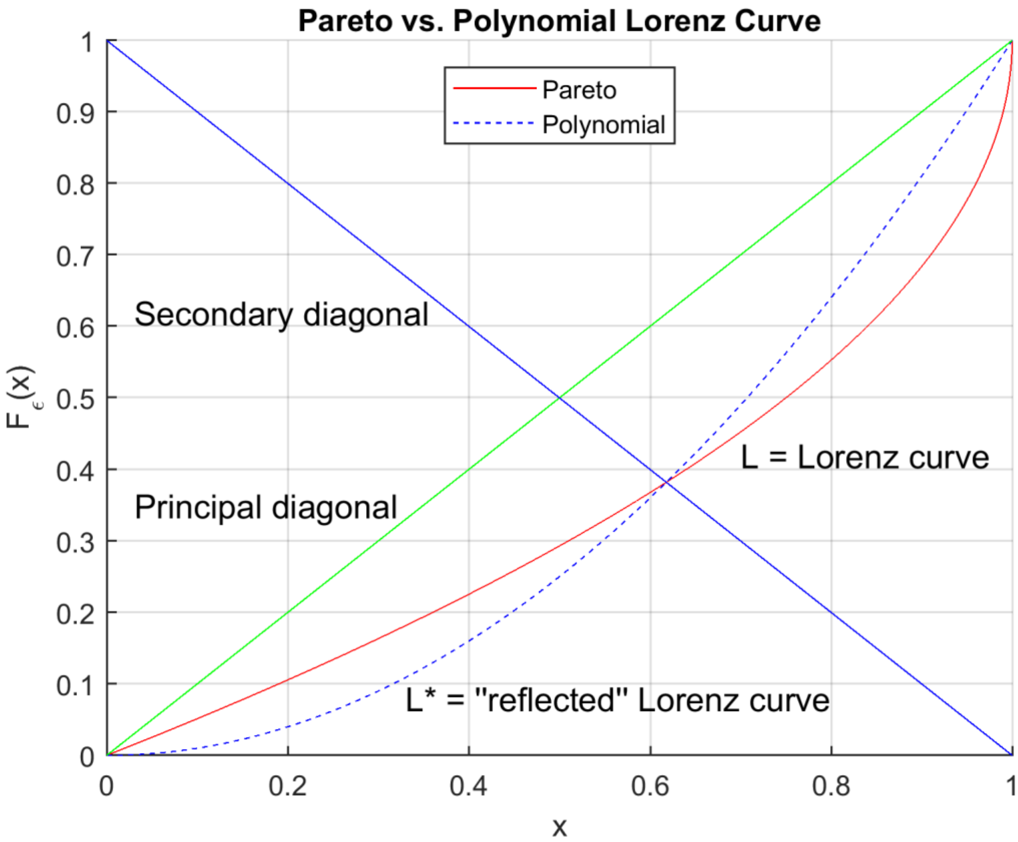

Figure 2: Lorenz curve $\boldsymbol{L}$ (red) and its the reflection $\boldsymbol{L}^*$ (dashed blue).

In our approach, two well-known types of Lorenz curves play a major role, namely the **Pareto curves** and the **Polynomial curves** which are dealt with in more detail below. Figure 2 shows the Pareto and the Polynomial type, respectively. It can be shown that the one type of curve can always be generated by reflecting the other type of curve at the secondary diagonal as in Figure 2 (see Appendix). In cases of reflection, as mentioned above, the Gini values are identical.

Using a similar argument, each convex combination of two Lorenz curves L and L^* of the form $\alpha \cdot L + (1-\alpha) \cdot L^*, 0 \leq \alpha \leq 1$ proves to be a Lorenz curve. If L and L^* have the same Gini coefficient, then this is true for every convex combination.

The insights, reported in this paper, can be embedded into the study of a class of Lorenz curves, defined by three parameters as convex combinations of Pareto and Polynomial curves, see below. Doing so, one can vary 3 parameters, where eventually one will do. This 3-parametric resulting class of Lorenz curves is a subclass of the generalized Tukey's lambda distribution. More on this is given below. The 3 parameters include the ones defining the Pareto- and the Polynomial-type Lorenz curves and a mixture parameter is used. More on this is given below.

Figure 2 shows for L and L^*, as it is well known, that very different Lorenz curves can have the same Gini value. Therefore, in principle, it cannot be expected that one Lorenz curve can be unambiguously inferred from one Gini value. The question of unambiguousness therefore does not arise for many fields of application for Lorenz curves outside the world of income distribution, because it is obvious, that this is not the case there. Regarding income distributions, according to the current state of the literature, this has likewise not been studied until now. **Exactly this is the main issue of this paper.** We want to understand, whether for income data for mature states, the Gini alone can essentially lead to the information, given in the World Bank data base (and EU-SILC data base, as well). This can help in empirical work in the fields of social and economic sciences. Because often, one knows one point value (e.g. the Gini value of the considered income distribution), but would like to know other details about the income distribution as well. The data situation would be substantially improved if, from the value of the Gini index, one was able in certain cases of income data to infer an essentially unambiguous Lorenz curve (with relative precision). However, to begin with, there has been no such claim up to now. Even if there was one, it would not be clear what the corresponding Lorenz curve should look like. In Section 2, we are now going to answer the questions posed.

2. From the Gini to the Lorenz curve – the main finding

In what follows, we switch to using the World Bank data bank (EU-SILC data works similarly). This is data for nation states, i.e., we concentrate on the mature cases. Internationally, the World Bank data base is the most comprehensive database in the field, and it is also largely standardised. From corresponding analyses at the Research Institute for Applied Knowledge Processing (FAW/n) in Ulm (Orthen 2017), we also know that these data for European states is very similar to those of EU statistics on income and living conditions (http://ec.europa.eu/euro-stat/data/database), apart from differences resulting from equalisation of income data in the EU-SILC case.

The following five examples in Figure 8 deal with 3 states each, having the same Gini values within the World Bank data. The examples show the broad empirical fit of income distribution quantiles. In

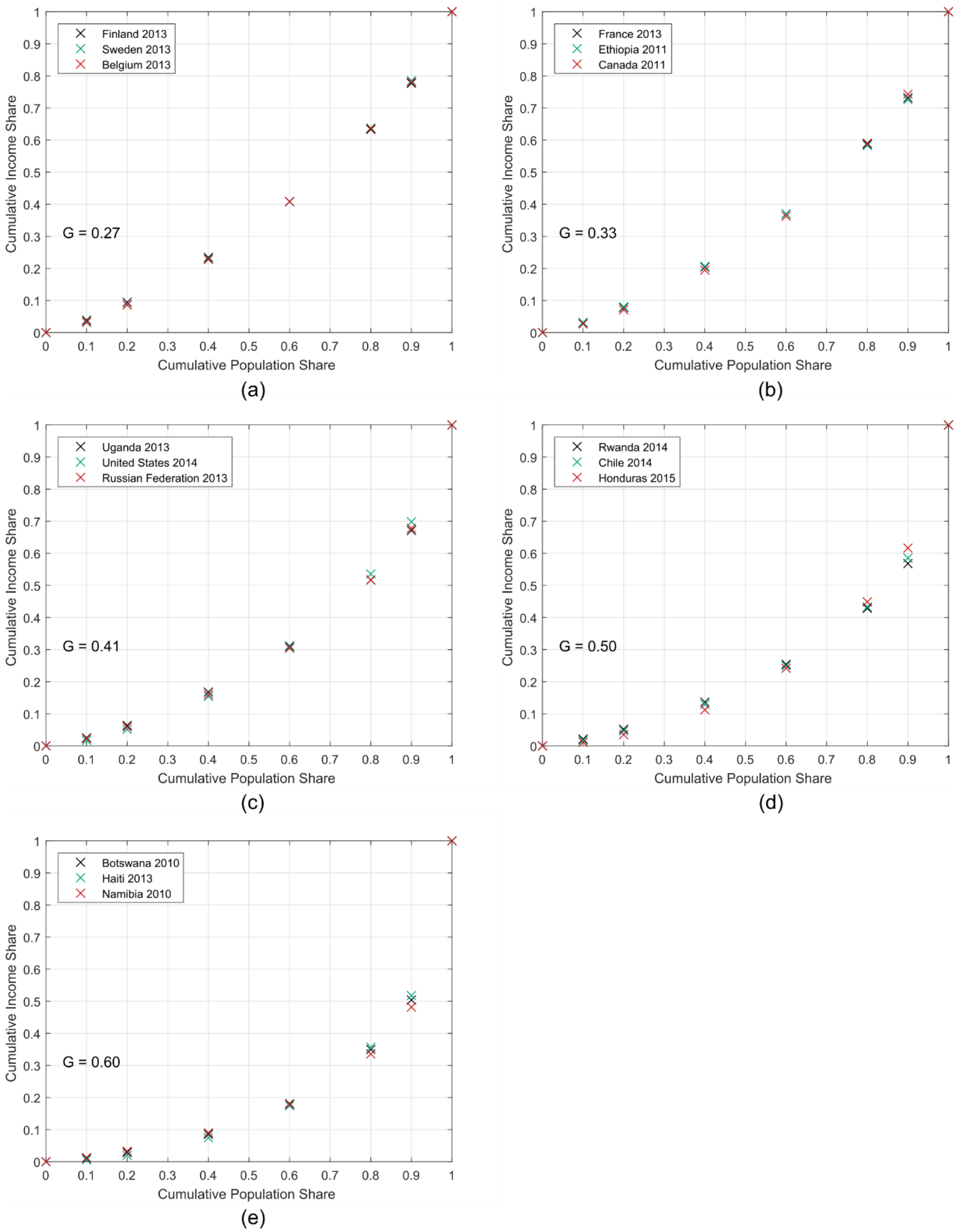

Figure 3: Five examples of World Bank data for three countries each with the same Gini value. The data points show a good fit between the corresponding deciles.

general, we see that the data points coincide with each other. There are certain derivations for countries with high inequality ($G \geq 0.5$), especially in the high-income segment. It seems, that these might be linked to issues of data quality. The fit is particularly good in the case of highly developed countries (OECD), where good data quality can generally be assumed. This empirical finding motivates the hypothesis of this text that, when taking the World Bank data as a basis, (similar EU-SILC data) there is a (universal) unambiguous relationship between Gini value and a standard Lorenz curve, at least, concerning the data points given in the World Bank data base. The same is true for EU-SILC data, with somewhat less approximation quality. Note here, that the behaviour for the five examples chosen with three different countries, each shows the level of approximation quality that we observe for all states (Orthen 2017). The minimized root mean squared error (RMSE) values differ from 0.0014 to 0.014 in the examples given. This is the general picture for all states.

We now ask, what **standard Lorenz curve** should be associated to a Gini value for a mature state?

One can approach this issue by considering different combinations of Pareto and Polynomial Lorenz curves observing three free parameters. Trying an optimal fit with World Bank and EU-SILC data will not give much information (because of overfitting). Similar effects would occur for the more general Tukey's lambda distribution with 4 parameters. So, statistics will not tell us so much. If one, however, starts from the right guess, data studies show that the results are good in the sense that they are close to the optimal result for all parameters chosen. Our approach was, as often, quite different.

When starting our work, we tried the assumption that mature societies with market structures exhibit certain typical patterns regarding income distribution that differ from arbitrary Lorenz curves and also from Lorenz curves for conglomerations of states that are ad hoc combined to form larger units. Assuming there is essentially a one-to-one relationship between the Gini and the Lorenz curve, we need an idea of what a standard Lorenz curve could look like for this case. We come close to an answer by looking at a certain type of distribution structure within societies that we refer to as **self-similar**. Note, that self-similarity is different from self-symmetrical behaviour as discussed by Taguchi (1968).

In what follows, we describe the property of **self-similarity** of an income distribution or Lorenz curve **going upwards** or alternatively **going downwards**. This property in essence means that the higher (lower) income segments (viewed in isolation) behave like the whole society in respect of their distribution. Mathematically, this leads to **differential equations** that have unique solutions. These solutions are the (known) Lorenz curves of the Pareto and polynomial type to a given parameter ε, where for ε, the relationship $\varepsilon = \frac{1-G}{1+G}$ applies, if G denotes the Gini value of the corresponding distribution.

The combination of $0.6 \cdot Pareto(\varepsilon)$ and $0.4 \cdot Polynomial(\varepsilon)$ proves to be a suitable candidate for a standard Lorenz curve belonging to G. With this, in the case of income distributions such as the World Bank (or EU-SILC) describes them, a good empirical approximation of the given quantiles is reached. We will show this **as an empirical finding below**.

Walkthrough

Starting from a given Gini value G from the World Bank, a **standard income Lorenz distribution** L_G can be derived such that L_G has the Gini value G, and (almost) identically reproduces the decile values of income distribution of the World Bank database (The same holds for the EU-SILC case). The standard Lorenz curve has only one free parameter, the corresponding Gini value. The situation is as follows:

(1) To G belongs the Lorenz curve L_G, with Gini value G.

(2) L_G reproduces G identically, and it reproduces the decile values of the World Bank database (or the EU-SILC data base) relatively well.

(3) L_G can be chosen as $L = L_\varepsilon = 0.6 \cdot Pareto(\varepsilon) + 0.4 \cdot Polynomial(\varepsilon)$, where $Pareto(\varepsilon)$ and $Polynomial(\varepsilon)$ are special Lorenz curves with special properties (e.g. self-similarity), and $\varepsilon = \frac{1-G}{1+G}$ and $G = \frac{1-\varepsilon}{1+\varepsilon}$ are true.

(4) The link between G and L_G is (largely) unambiguous; that is, from G, one can reliably infer L_G, or, at least, the corresponding data points from the World Bank (or EU-SILC) data base. This

applies to income distributions of mature countries, in terms of the World Bank (or EU-SILC) data, and represents an empirical finding. **This is generally not the case for other Lorenz curves.**

Described in more detail below are the Pareto Lorenz curves appearing in (3) and the polynomial Lorenz curves, which are a function of a parameter ε and are defined as follows (cf. also Kämpke and Radermacher (2015)):

Definition: The **Pareto Lorenz curves** as a function of $x \in [0,1]$ are defined as the class of curves $L_{\varepsilon}(x) = Pareto(\varepsilon)(x) = 1 - (1-x)^{\varepsilon}$, with $0 \leq \varepsilon \leq 1$.

Note, that in literature often other formulations of the Pareto curves are used (see, e.g. Arnold 2016). We use the parametrisation of Pareto curves from Kämpke, Radermacher (2015), which has only one parameter. Often, people work with a 2-parametric version of the Pareto curve, using a shape parameter $\alpha > 0$ and a range parameter $x_m > 0$. The Lorenz curve is, however, independent of the range parameter $x_m > 0$ and only depending on the shape parameter. This allows us to make things straighter and also helps with proving results on self-similarity. Here, $\alpha = \frac{1}{1-\varepsilon}$, for $\varepsilon = 0.6, \alpha = \frac{1}{0.4} = 2.5$.

Figure 4 shows (for $\varepsilon = 0.6$) the associated Pareto Lorenz curve (left) and the Pareto Lorenz density (right), where the Lorenz density is the first derivative of the Lorenz curve. It can be shown that this always exists (almost everywhere). Sometimes this density is called the Pen's parade. Here the following is true: the Lorenz curve depicts the cumulative proportions of income where income is ordered by increasing size, while the Lorenz density depicts the relative absolute income level. In the case of the Pareto Lorenz density, ε is the value of the lowest income, while the highest incomes tend towards infinity. The mathematical formula of the Lorenz density of the Pareto distribution is $L'_{\varepsilon}(x) = f_{\varepsilon}(x) = \varepsilon(1-x)^{\varepsilon-1}$. In the case $\varepsilon = 0.6$, the associated Gini index has the value $G = 0.25$.

Definition: The **Polynomial Lorenz curves**, as a function of $x \in [0,1]$,

are defined as the class of curves $L_{\varepsilon}(x) = x^{1/\varepsilon}$, where $0 \leq \varepsilon \leq 1$.

Here, ε now appears as an exponent with the value $1/\varepsilon$. The value $1/\varepsilon$ here corresponds to the highest incidence of income, see Lorenz density. In comparison, on the low-income side, the lowest income is 0. Figure 5, analogous to Figure 4, shows the polynomial Lorenz curve (left) and its density (right), again for the case $\varepsilon = 0.6$ and $G = 0.25$. The Lorenz density of the polynomial distribution is $L'_{\varepsilon}(x) = f_{\varepsilon}(x) = (1/\varepsilon) \cdot x^{(1/\varepsilon)-1}$.

Note, that $Pol(\varepsilon) = x^{1/\varepsilon}$ (Polynomial curve) is the **reflection** of $Pa(\varepsilon)$ (Pareto curve), i.e. $Pol(\varepsilon)(u) = 1 - Pa(\varepsilon)^{-1}(1-u)$ for all $u \in [0,1]$. For this, take $u = 1 - (1 - (1-x)^{\varepsilon}) = (1-x)^{\varepsilon}$. The detailed calculation can be found in the appendix. The result also holds the other way around.

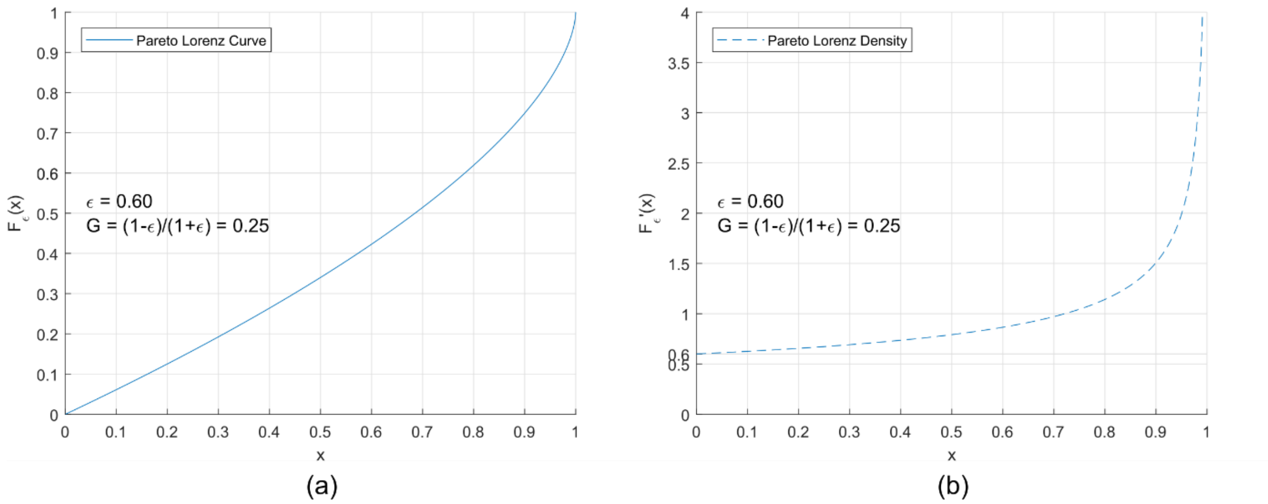

Figure 4: Pareto Lorenz curve (left) and corresponding Pareto Lorenz density (right) for $\boldsymbol{G = 0.25}$, where $\boldsymbol{\varepsilon = 0.6}$.

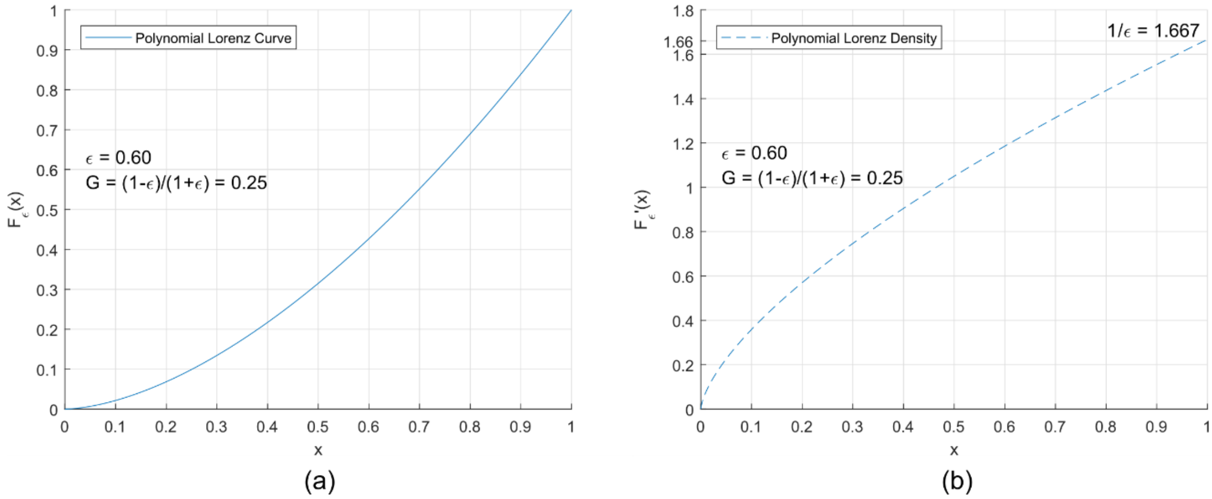

Figure 5: Polynomial Lorenz curve (left) and corresponding Lorenz density (right) for $\boldsymbol{G} = \mathbf{0.25}$, where $\boldsymbol{\varepsilon} = \mathbf{0.6}$.

Definition: Self-similarity

Self-similarity is an interesting property of Lorenz curves. If there is self-similarity, the income distribution in certain income segments corresponds to the distribution for the whole population. This is also an interesting property in terms of its practical societal interpretation. Self-similarity with respect to income distributions also translates to self-similarity with respect to Gini values or (under additional assumptions) to self-similarity with respect to the median values of the income distributions considered (Herlyn 2012; Kämpke and Radermacher 2015). Self-similarity is to be distinguished in terms of 'self-similarity upwards' and 'self-similarity downwards'. Note again, that self-similarity is different from 'self-symmetrical' as discussed in Taguchi (1968). In our example mentioned above the Pareto and Polynomial Lorenz are reflections. As a result of reflection, self-similarity upwards changes to self-similarity downwards. The same is true the other way around. This is given below.

Definition: **Self-similarity of Lorenz curves upwards** means that the distribution situation among the δ percent of individuals with the highest income, when viewed as a whole, is identical to the Lorenz curve for the total population and this holds for all $\delta \in [0,1]$.

Figure 6 illustrates this property. The Lorenz curve is here regarded in the interval $[\delta, 1]$ (left figure). The rectangle $[\delta, 1] \times [L(\delta), 1]$ is then enlarged onto the unit square $[0, 1]^2$ (right figure). The Lorenz curve that arises after stretching corresponds with self-similarity upwards to the total Lorenz curve over the interval $[0, 1]$.

Proposition: The Pareto Lorenz curves are exactly the upwardly self-similar Lorenz curves. If, therefore, a Lorenz curve is upwardly self-similar, there exists an $\varepsilon \in [0,1]$, such that $L = L_\varepsilon = 1 - (1 - x)^{\varepsilon}$ is true.

Symmetrical to the self-similarity upwards, one can define **self-similarity downwards** by considering the Lorenz curve in the interval $[0, \delta]$, that is, for the δ percent of the population with the lowest income. Correspondingly, the square $[0, \delta] \times [0, L(\delta)]$ is enlarged to the unit square $[0, 1]^2$, and it is claimed that for every $\delta \in [0, 1]$, the resulting Lorenz curve corresponds to the total Lorenz curve over the interval $[0, 1]$. As with the Pareto case, a solvable differential equation results from the defining requirement. This allows the formulation of the following proposition:

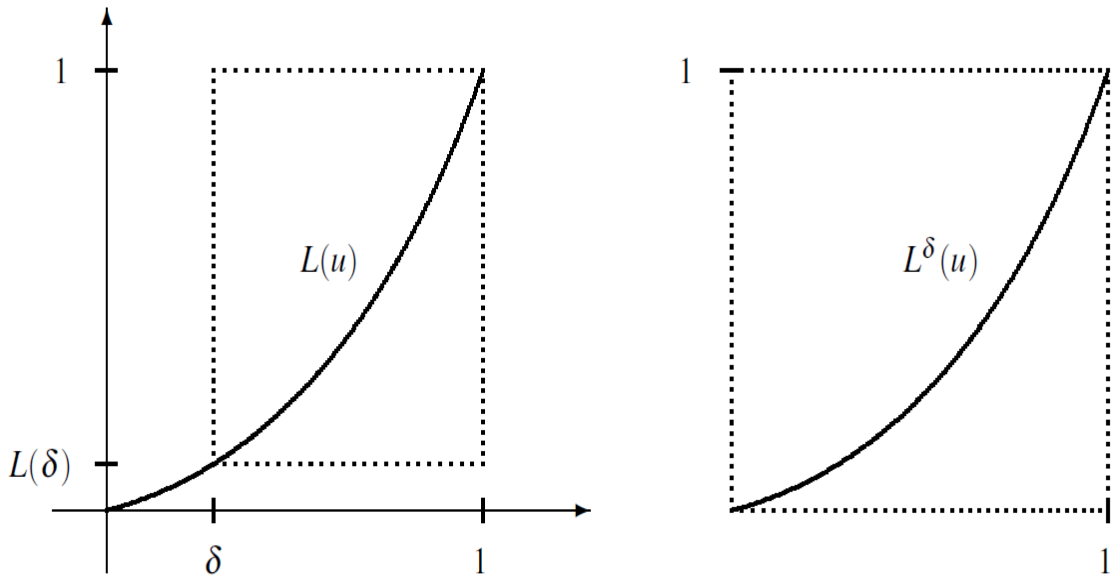

Figure 6: Lorenz curve with a cut-off point δ (left) and the cut-out Lorenz curve (right). The rectangle [δ, 1] x [L(δ), 1] (left) is stretched to the unit square $[0, 1]^2$ (right).

Proposition: The polynomial Lorenz curves are exactly the downwardly self-similar Lorenz curves. If, therefore, a Lorenz curve is downwardly self-similar, there is a $\varepsilon \in [0, 1]$, such that $L = L_\varepsilon = Polynomial(\varepsilon)(x) = x^{1/\varepsilon}$ is true.

Proposition: It can thus be shown (Kämpke and Radermacher 2015) that

1. $Pareto(\varepsilon)$ are the only upwardly self-similar Lorenz curves
2. $Polynomial(\varepsilon)$ are the only downwardly self-similar Lorenz curves
3. $Pareto(\varepsilon)$ is the reflection of $Polynomial(\varepsilon)$ at the counter-diagonal, and vice versa (see Figure 2 and proof in the Appendix).
4. As a consequence, $Pareto(\varepsilon)$ and $Polynomial(\varepsilon)$ have the same Gini value G. This is also true for any linear combination of the two in the form $\alpha \cdot Pareto(\varepsilon) + (1-\alpha) \cdot Polynomial(\varepsilon)$, where $\alpha \in [0,1]$,
5. The Gini G of the corresponding Lorenz curves has the value $\frac{(1-\varepsilon)}{(1+\varepsilon)}$; ε has the value $\frac{(1-G)}{(1+G)}$.

What analytical options do now exist, based on the World Bank data (or the EU-SILC data) that include the Gini value G? One can determine the associated ε value using $\varepsilon = \frac{1-G}{1+G}$. Thus, one has at one's disposal $Pareto\ (\varepsilon)$, $Polynomial(\varepsilon)$, and any linear combination of these two Lorenz curves. All these Lorenz curves reproduce the Gini value identically.

An alternative approach is, in relation to the three parameters ε_1 (Pareto), ε_2 (Polynomial) and α, to implement the best alignments to the data material using the least squares method. This approach has also been investigated – and for different weights of the Gini value, too (Orthen 2017). It is shown that for this approach, too, good alignments to the World Bank data (and the EU-SILC data) are possible for $\varepsilon = \varepsilon_1 = \varepsilon_2$, and, indeed, in such a way that $\varepsilon = \frac{1-G}{1+G}$, or equivalently $G = \frac{1-\varepsilon}{1+\varepsilon}$, is also true. General alignments using the least squares method with potentially different ε_1 and ε_2 values and free mixing parameter $\alpha \in [0,1]$ thus result, with good approximation to the respective decile values, in the **one-parametric** combination $\varepsilon_1 = \varepsilon_2 = \frac{1-G}{1+G}$ and $\alpha = 0.6$. This supports the decision for the standard Lorenz curve L_G, pertaining to G, that we will define next.

Definition: The Standard Lorenz curve pertaining to a Gini G is defined as $L_G = 0.6 \cdot Pareto(\varepsilon) + 0.4 \cdot Polynomial(\varepsilon)$.

Note: The Standard Lorenz curve is **not self-similar.** Detailed empirical studies show that in the case of World Bank data [http://databank.worldbank.org], the standard Lorenz curve represents a good approximation to the given data for 156 countries from various years in the period from 1995 to 2013. The respective Gini value from the World Bank table is reproduced identically. We refer to the examples in Figure 8 and Figure 9. A similar empirical finding is true for EU-SILC data (Orthen, 2017).

Finding: In the field of income distributions, there seems to be an (empirically sufficient) 1:1 relationship between the Gini index and the Lorenz curve depicting it. Therefore, if one knows the Gini value, one knows (to a very great extent) the associated Lorenz curve, i.e. with the Gini value, a lot more information is available with reasonable quality than the pure point information in the form of the Gini value.

Derivation

Definition: Mixed Lorenz curve

We consider all Lorenz curves of type $L_{\varepsilon,\alpha}(x) = \alpha(1-(1-x)^{\varepsilon}) + (1-\alpha)x^{1/\varepsilon}$. They are all convex compositions with a parameter $\alpha \in [0,1]$ from the two curve types Pareto Lorenz curve $F_\varepsilon(x) = 1-(1-x)^{\varepsilon}$ and polynomial Lorenz curve $F_\varepsilon(x) = x^{1/\varepsilon}$, both with the same ε and with proportional weighting by the factor $\alpha \in [0,1]$. For an improved alignment to the data compared to the Pareto or polynomial curves, which are not satisfactory in their alignment (see below), the additional parameter α is added, as discussed above. Figure 7 first shows the (not satisfactory) approximation of the World Bank data for each of Germany, India and the USA, using pure Pareto or pure polynomial Lorenz curves.

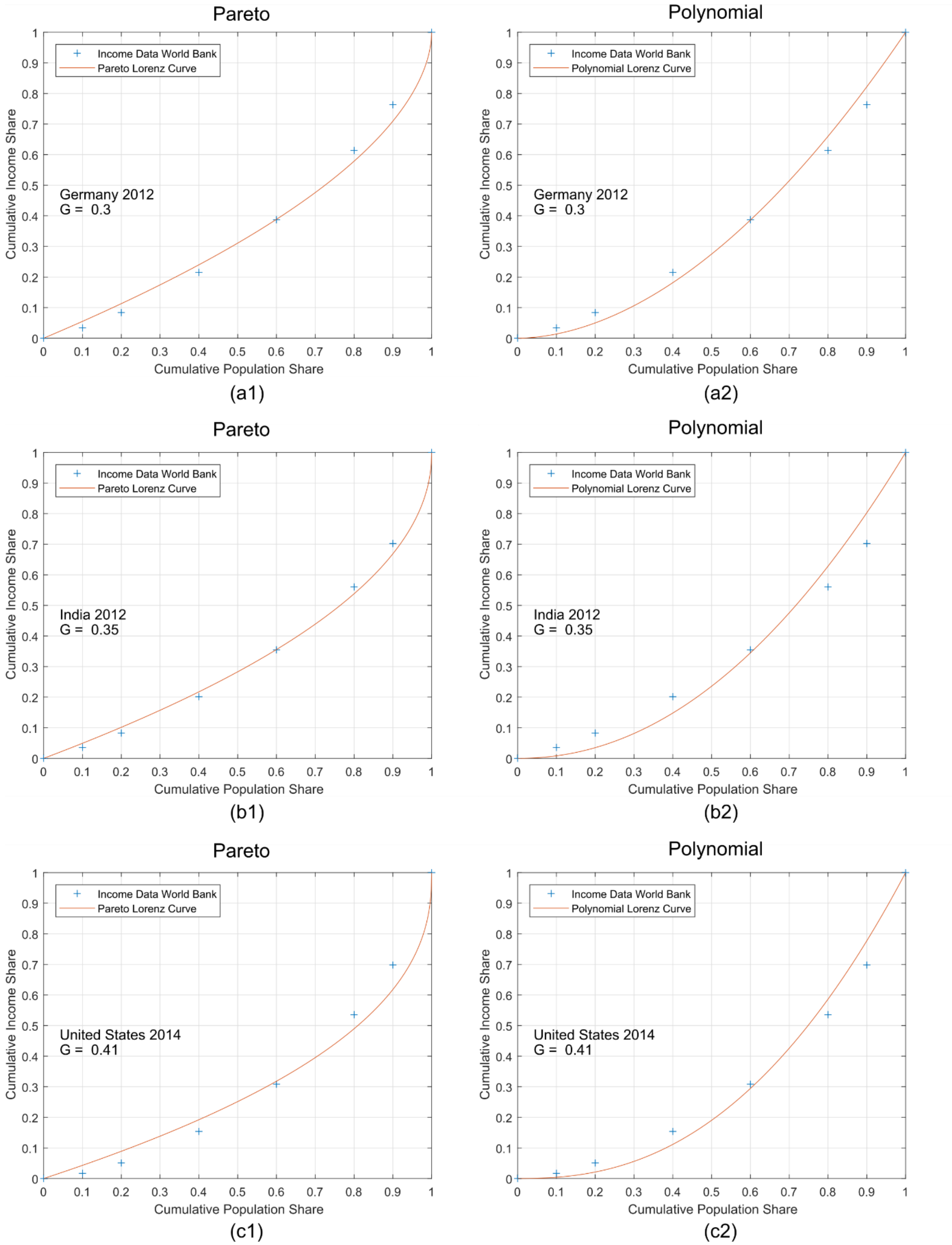

Figure 7: Approximation of World Bank data for Germany, India and the USA with Pareto (left column) and polynomial (right column) Lorenz curves, with an identical Gini value taken from the World Bank data.

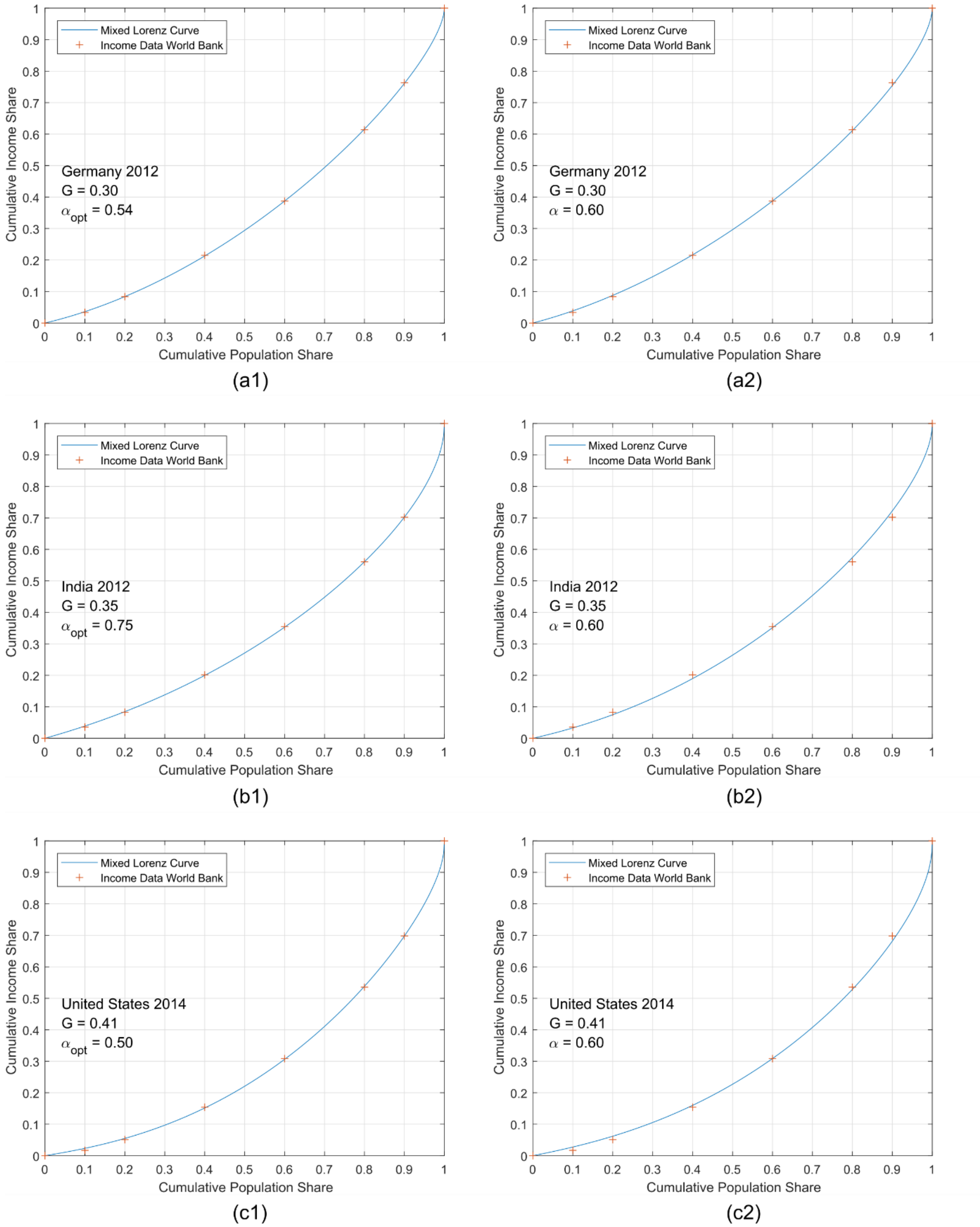

Figure 8: Approximation of World Bank data for Germany, India and the USA with mixed Lorenz curves with optimal value α_{opt} (left column), which is different for each country, and the value $\alpha = 0.6$, the standard value, which is the same for all countries (right column).

Empirical results (extract)

Approximations of World Bank data for Germany, India and USA for the two distribution types considered; namely, Pareto and Polynomial. For the following examples, $\varepsilon = \frac{1-G}{1+G}$ is always chosen.

In addition, we show mixed approximations for the respective $\boldsymbol{\varepsilon}$ values, with optimal[4] α_{opt} **and the selected default value $\boldsymbol{\alpha = 0.6}$** (Figure 8). This case leads to results that are satisfactory. The resulting minimal root mean squared errors (RMSE) range from 0.0014 to 0.014, given above.

Note: Error function for the mixed value $\boldsymbol{\alpha}$

As is shown in Herlyn and Radermacher (2017), the influence of the factor α when considering the quantile values of the World Bank data is comparatively low in the vicinity of the optimum. From this basically arises the chance that a fixed value α could be sufficiently good for all income distributions. This is, however, not true at the boundaries, as e.g. the value of the Lorenz density at point zero is $\alpha \cdot \varepsilon$. Obviously, α has a significant influence there.

No question, our claim needs to be empirically proven. Namely, it could also be the case that the optimal α values are obviously correlated with G. The authors have thus pursued the thesis that a low G, i.e., a high level of balance, tends to also result in a large α, that is, a large proportion of the Pareto Lorenz curve. There were reasons to expect this. The result is that there is no such connection. This is elucidated below in Figure 9:

Evidently, there is no link between the form of the optimal α and the respective Gini value. The countries of the World are sorted in descending order in terms of inequality (i.e., falling Gini values), that is, starting with countries of high inequality (large Gini), on the left, to countries with ever increasing balance (small Gini). Given each time is the optimised α (optimized using the least squares method to the

[4] Determined by the least squares method.

6 quantiles of the World Bank data). As one can see, there is obviously no clear regression relationship between falling G and the respective α_{opt}. One can also recognise that with the value $= 0.6$, one very frequently, regardless of how big G is, lies relatively close to the respective optimal α. Even if there is a greater distance, this does not translate to huge differences to the World Bank data, as already mentioned. Here, differences in the RMSE are measured.

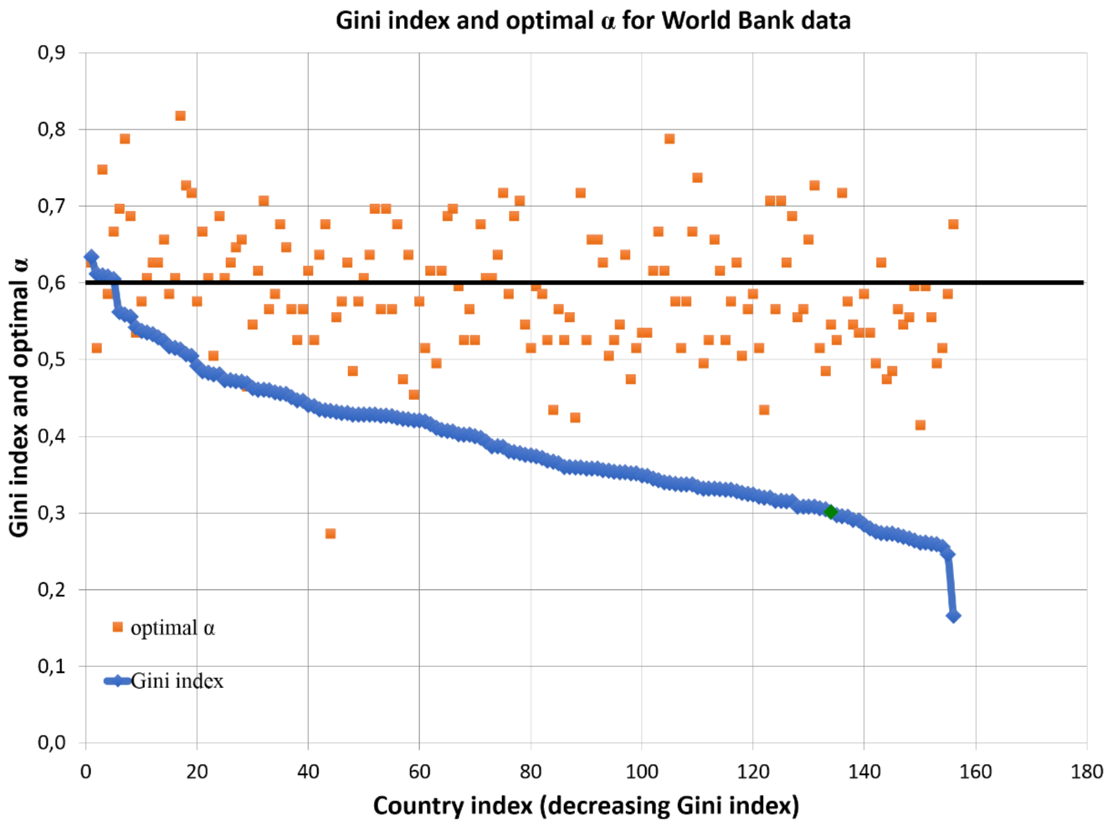

Figure 9: Countries sorted by descending Gini values (blue curve, country index). The red dots show the optimal $\boldsymbol{\alpha}$. The values for $\boldsymbol{\alpha}$ and $\boldsymbol{G}$ can be read on the y-axis.

Decision

At this point, the authors have decided to essentially approximate the empirical income distribution data of the World Bank (and similarly EU-SILC data) by means of a Lorenz curve (called the **standard Lorenz curve**) that is a function of ε only, and therefore of the Gini index, only. The Gini-ε-value is

$\varepsilon = \frac{1-G}{1+G}$. The combination $(L_G =)\ 0.6 \cdot Pareto(\varepsilon) + 0.4 \cdot Polynomial(\varepsilon)$ generally leads to a good approximation of the data provided by the World Bank (and the EU-SILC data).

This result has an interesting consequence. In the world of empirical income distribution data (for mature states), based on the World Bank (and EU-SILC) data, there is, as previously announced, a **relatively stable 1:1 relationship between Gini index and income distribution function**. Not only is a given Gini value associated with a unique Lorenz curve of the mixed type where $\varepsilon = \frac{1-G}{1+G}$. Conversely, a standard income distribution with parameter ε has the (unique) Gini value $G = \frac{1-\varepsilon}{1+\varepsilon}$. A complete overview of the data situation can be found in Herlyn and Radermacher (2017).

These findings represent a great advancement, because the (surprising) result means that in the income case, when one has a Gini value for a country, one is always, and empirically reliably, equipped with much more detailed information than the Gini usually represents; namely, the actual income distribution (except for a constant factor that characterises the general level of prosperity, e.g. GDP/person). If we also know the average income of a country, one can even infer from the Gini value the concrete distribution of the absolute income and much more.

Reference to literature

There is a broad literature on approximation of distribution data. The standard Lorenz curve proposed here is no reasonable solution in general cases. We deal here only with income data of (mature) national states on the basis of the World Bank data (and EU-SILC data). One interesting general class in this regard is the generalized Tukey's Lambda distribution described by Sarabia (1997). It has 4 parameters, namely η_1, η_2, α_1 and α_1. Putting $\eta_1 = 0$ and $\alpha_1 = 1/\alpha_2$ leads to the class of mixed curves we study. In this case, the two parts of the Lorenz curve (Pareto and Polynomial) have the same Gini index equal to $(1-\alpha)/(1+\alpha)$ and the complete Lorenz curve also has the same Gini (which is not the case for generalized Turkey's Lambda distributions in general). Our mixture curves have really nice properties and we eventually got even one step further, namely to $\boldsymbol{\alpha_2 = 0.4}$, leading to a **1-parameter** situation.

The Gini may be chosen to be this single parameter. Going beyond nation states, e.g. for Europe as a whole or for the world as a whole, the standard Lorenz curve is not sufficient to achieve a good approximation, as shown in Figures 12, 13 and 14. That may also be the case for other income data, that we did not investigate. And is obviously true for all kind of other distributional data for which one would like to have a Lorenz curve and needs a good approximation. Also, our standard curve is only a special case of more general cases discussed in the literature, as was already mentioned.

3. Limitations of applicability

As a general rule, it is not possible to infer an income distribution from a Gini value. This is well-know. We give examples for this, that are informative and helpful in our context. Note, that in the case of real world data, the examples do not concern mature states but instead conglomeration of states (EU and the world as a whole). In these examples we thus look at income distributions of territories, that have not long existed as open, economic-political systems. They lack the coherence of mature countries that result from long-active common markets, freedom of travel and settlement, freedom of individual economic activities, integrated financial markets, etc. This means limited character of self-similarity of structural elements. The EU as a whole is a borderline case in this respect.

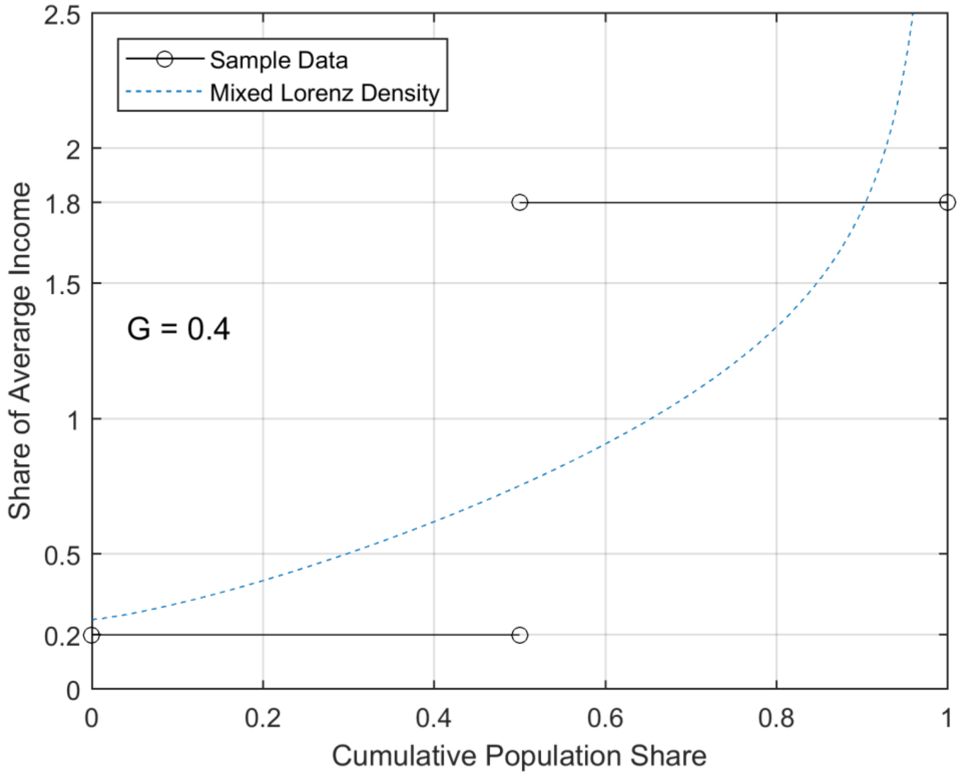

Figure 10: Two very different Lorenz densities having the same Gini value of 0.4.

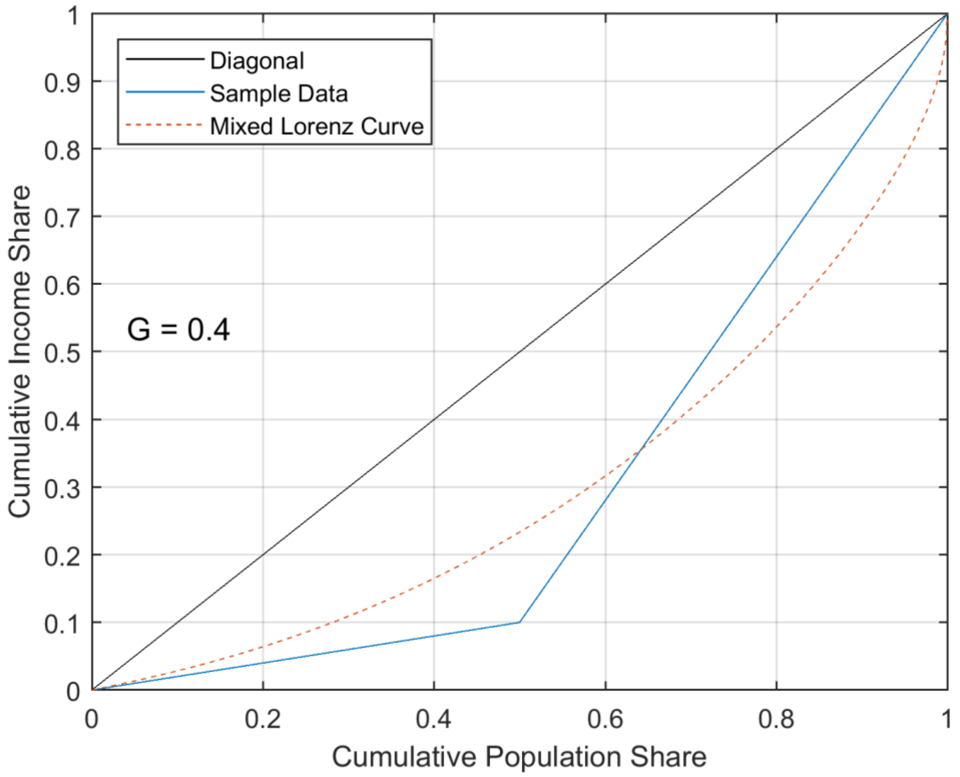

Figure 11: Lorenz curves for the Lorenz densities in Figure 10.

Example 1

We first consider an "artificial example" of two states with the same population size and equally distributed income (identical income for all individuals) and with the situation where the income of the one group is higher than that of the other by a factor of 9. The distribution is therefore completely homogeneous within the groups but varies extremely between the groups. In Figure 10, we show the Lorenz density, and in Figure 11, the Lorenz curve for this distribution. The Gini value is $G = 0.4$. The standard income Lorenz curve developed in this text L_G, with Gini $G = 0.4$ and the associated Lorenz density L'_G, are also given.

It can be seen that the distributions are very different, although the Gini value is the same. Note that this is not surprising. The literature has many such examples. However, it is the core message of this text, that for income distributions of (mature) states using the World Bank data or the EU-SILC data, such examples do **not** exist.

Example 2

Represented in Figure 12 is the income distribution of the EU as an aggregated system of states: the EU in 2002 with 15 member states, and then in 2016, with those same member states. It can be seen that the income distribution within the union of the original 15 member states has not changed in the period 2002-2016 (Figure 13). The Gini value for the EU as a whole stagnated at 0.33; i.e., the inequality may have changed in certain states but stayed the same for the whole EU. Also, the case for all present EU states (28) on the basis of 2016 World Bank data is shown (Figure 12). For both cases, the associated standard Lorenz curves are drawn, which approximate the data points comparatively well but not as good as we would like.

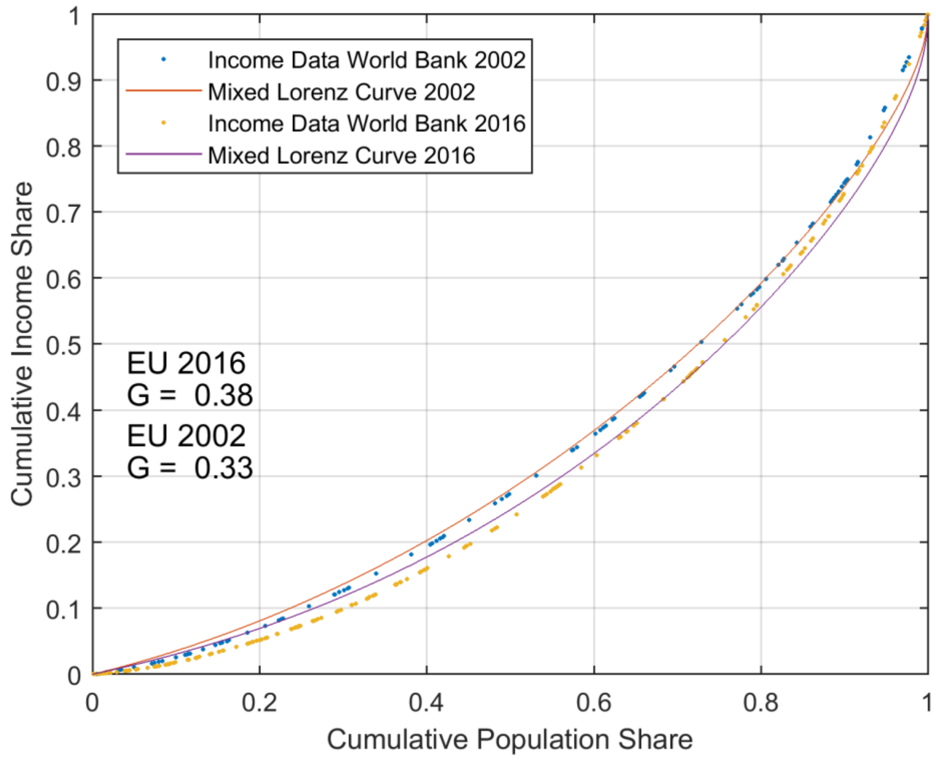

Figure 12: Income distribution of the EU as an aggregated system of states.

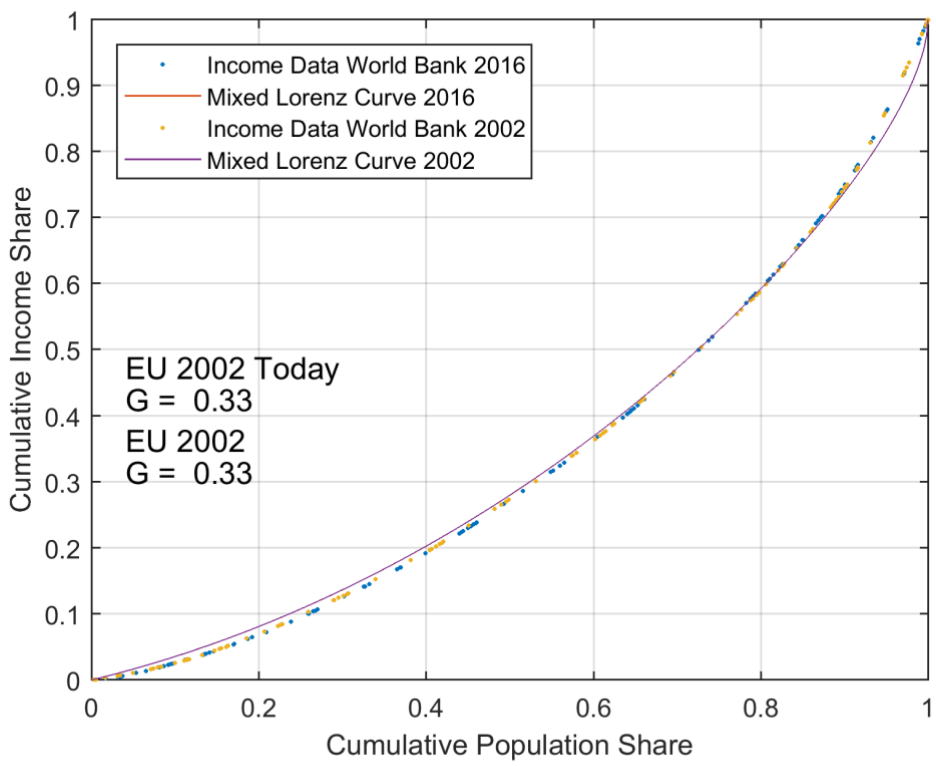

Figure 13: Income distribution of the EU in 2002, with 15 states (yellow) and in 2016, with 27 states (blue, Malta is missing because no data is available).

Note:

The selected procedure for determining the EU income distribution based on World Bank data for the participating EU countries results in many new data points. An aggregation procedure is necessary, because the World Bank does not provide any income distribution data for the European Union as an aggregated construct. Further information on the method used can be found in Orthen (2017).

Note that the alignment of our approximation with the income distribution at an EU level, using the standard Lorenz curves, is not good. But we have no claim in this direction. Approximations for the individual European states are much better.

The following statements refer to two different stages of the EU's development: a time when it had 15 states, in 2002, and a time when it had 28 states, in 2016. The Gini coefficient w.r.t. World Bank data increased from 0.33 to 0.38, i.e., inequality has increased significantly. The approximation using the standard Lorenz curve is less good in 2016 than in 2002, see Figure 12 and Figure 13. The inequality in the extended European Union is higher than in the ‘old’ European Union. This should come as no surprise, given the much lower income levels in the new member states. Only a short time has passed since 2002; this is not much time for the necessary adjustment processes for achieving a typical income distribution.

Note: We make no claim concerning the quality of the standard Lorenz curve approximation which is obviously not good. The level of inequality may interest readers. As mentioned above, for the EU with 28 member states in 2016, using EU-SILC data, the Gini value is 0.36. This is lower than for World Bank data. Looking into aggregation of EU-SILC data, the Gini index of the original 15 EU member states changed from 0.29 to 0.31. That is a slight increase of inequality. This systematic difference between the data sets is probably a consequence of calculating the equivalised income for families (EU-SILC) versus the per capita income (World Bank), as mentioned above.

As mentioned below, the Lorenz curves of the individual European countries (not of Europe as a whole) are nevertheless almost identical for World Bank data and Eurostat (Orthen 2017), while the level of inequality is somewhat greater on the World Bank side. The reason for this is probably a relative homogeneity of family size in the lower and upper ranges of the income pyramid, with a somewhat smaller family size in the middle. The resulting effects more or less seem to cancel out.

Figure 14 shows the Lorenz curves of **world income distribution** from 2002 and 2016, derived from aggregation. As with the EU, many data points result from this approach. It is clear that inequality has declined as a result of increasing global economic cooperation. The Gini coefficient has decreased from 0.78 to 0.70. Alignment with the standard Lorenz curve has also improved. However, it is of course not satisfactory, and we make no claim that the standard Lorenz curve is appropriate to approximate the world income distribution. International homogeneity is still very low in comparison to long existing nation states. The respective curve pairs show different Lorenz curves and the same Gini value.

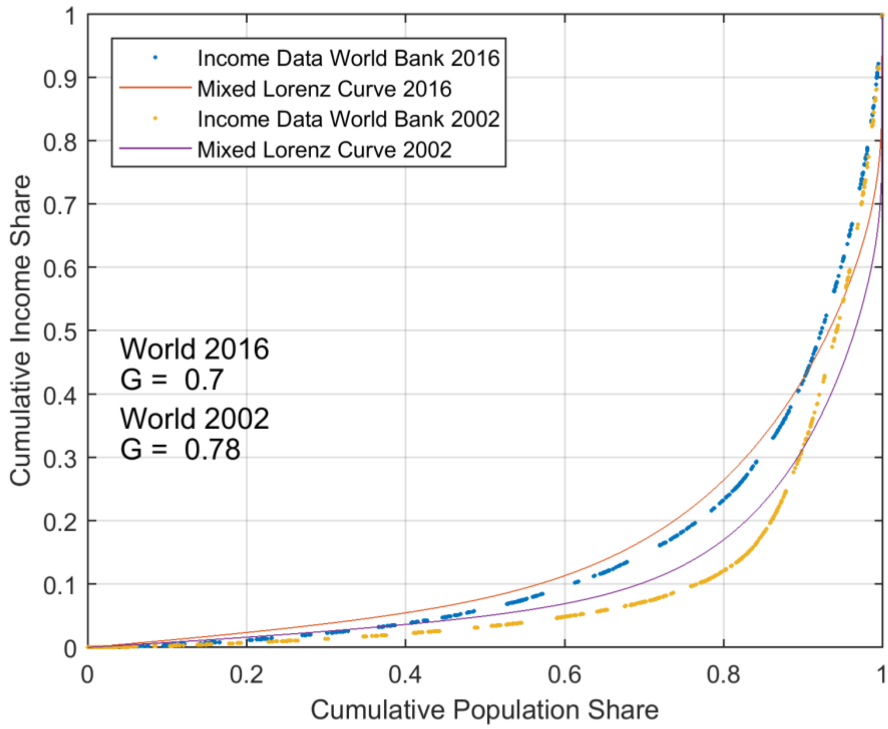

Figure 14: Global income distribution according to World Bank data for the years 2002 and 2016.

4. Using the new options available

Knowledge of income distributions in the form of a Lorenz curve in the sense of a closed mathematical function has many advantages for applications, even if the quality of the inferred data has always to be considered separately. As already mentioned, it is generally possible to show that the derivatives of Lorenz curves exist almost everywhere. In the case of the standard Lorenz curve, they are of the form $L'_{\varepsilon}(x) = \alpha \cdot \varepsilon(1-x)^{\varepsilon-1} + (1-\alpha) \cdot (1/\varepsilon) \cdot x^{(1/\varepsilon)-1}$. For every income position x, these Lorenz densities give the relative level of income. If one changes the level of balance of a society from a Gini value of G_1 to a Gini value of G_2, and thus from an equity value of ε_1 to ε_2, the relative incomes change. Winners and losers (in the sense of social position, not people) can be precisely identified (though in reality, this may mean different people), as can be the size of the relative losses and profits. **Distribution issues** can be discussed in a much more nuanced and factual manner than is currently the case where corresponding mathematical instruments for analysis are not available. With the approach described, the total volume of redistribution can also be determined, and the distribution effects for selected sub-segments of the population can also be calculated.

The approach given, provides access to new data material, which can be used for dealing with a multitude of interesting questions, such as for a deepened study **of different poverty parameters**.

In this context, it is desirable to develop a software application that provides derivable information from a Gini value. Some of this can be found in (Herlyn 2012; Herlyn and Radermacher 2017; Kämpke and Radermacher 2015); more will be elaborated in further research at the FAW/n in Ulm, with the support of the **Vector Foundation**. Corresponding publications are in preparation. This includes, in particular, an accurate analysis of the effects on winners and losers when changing the Gini value: who wins how much and who loses how much if the Gini value of a country changes? These changes, and the winners-losers effects, of course, have repercussions on democratic processes and possible formations of majorities in politics.

Final remarks: Making an analytical tool available for public use

FAW/n will make available the mathematical analysis possibilities that result from the findings presented here in a wide range of possible applications as **open source software** for the general public, in particular, for the scientific community, and, what's more, under a "GNU General Public License" (GNU GPL). For interested users, this opens up the possibility of individually developing the software further. An according system is in preparation. Status updates will be posted at www.faw-neu-ulm.de.

Acknowledgements

The authors would like to thank Tobias Orthen, Maximilian Kertel and Masoud Mansouryar for their support in drafting this text within the framework of the existing cooperation at the University of Ulm and the Research Institute for Applied Knowledge Processing (FAW/n) in Ulm. We would also like to thank the rest of the FAW/n team in Ulm for carrying out a lot of work, required for the results reported here. We thank Wolfgang Eichhorn and Manfred Krtscha for some valuable input. The same is true for two anonymous referees who helped us with their reports.

Funding

The present work was supported by FAW/n in Ulm and the Vector Foundation in Stuttgart as part of the project "Growing income inequality – mathematical-analytical investigations into social reality and its perspective".

Conflict of Interest

The authors declare that they have no conflict of interest.

References

Alvaredo, F., Atkinson, A. B., Piketty, T., Saez, E.: The top 1 percent in international and historical perspective. Journal of Economic Perspectives 27 (3), 3-20 (2013)

Atkinson, A. B.: Inequality – What can be done? Harvard University Press (2015)

Cornia, G. A., Court, J.: Inequality, growth and poverty in the era of liberalization and globalization. WIDER Policy Brief 2001/004 Helsinki, UNU-WIDER (2001)

Eurostat Statistics Explained, Glossary: Equivalized disposable income, http://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Equivalised_disposable_income (23 Aug 2017), German: http://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Equivalised_disposable_income/de (23 Aug 2017)

Galbraith, J. K., Choi, J., Halbach, B., Malinowska, A., Zhang, W.: A comparison of major world inequality data sets: LIS, OECD, SILC, WDI, EHII. The University of Texas Inequality Project, UTIP Working Paper No. 69 (2015)

Gastwirth, J. L.: A general definition of the Lorenz curve. Econometrica 39, 1037-1039 (1971)

Herlyn, E.: Einkommensverteilungsbasierte Präferenz- und Koalitionsanalysen auf der Basis selbstähnlicher Equity-Lorenzkurven - Ein Beitrag zu Quantifizierung sozialer Nachhaltigkeit. Springer Gabler (2012)

Herlyn, E., Radermacher, F. J.: Lorenzkurven und Gini-Index im Bereich der Einkommensverteilung – Neue inhaltliche Einsichten. FAW/n Report an die Vector Stiftung (2017)

Kämpke, T., Pestel, R., Radermacher, F. J.: A computational concept for normative equity. European Journal of Law and Economics No. 15, 129-163 (2003)

Kämpke, T., Radermacher, F. J.: Income modeling and balancing. A rigorous treatment of distribution patterns. Lecture notes in economics and mathematical systems 679 (Fandel, G. and Trockel, W., eds.). Springer International Publishing Switzerland (2015)

Krieger, A. M., Gastwirth, J. L.: Interpolation from Grouped Data for Unimodal Densities. Econometrica, Vol. 52, No. 2, pp. 419-426 (1984)

Lyon, M., Cheung, L. C., Gastwirth, J. L.: The Advantage of Using Group Means in Estimating Lorenz Curve and Gini Index from Grouped Data. The American Statistician (2015)

Milanovic, B.: Income inequality in Europa and the U.S.: regional vs. social-class inequality, Inequality in Focus, vol. 1, no. 2. The World Bank (2012)

Orthen, T.: Einkommensverteilungen und Lorenzkurven von Staatengruppen auf Basis von Datensätzen der Weltbank und der Europäischen Union. FAW/n Technical Report (2017)

Piketty, T.: Capital in the Twenty-First Century. The Belknap Press of Harvard University Press (2014)

Radermacher, F. J.: Lorenz Curves and Atkinson Theorem – Some Recent Insights (In memoriam Thomas Kämpke), Nachhaltiges Entscheiden: Beiträge zum multiperspektivischen Performancemanagement von Wertschöpfungsprozessen. Festschrift zum 65. Geburtstag von Prof. Dr. Harald Dyckhoff (eds. Ahn, H., Clermont, M., Souren, R.). Springer Gabler, 49-72 (2016)

Sarabia, J. M.: A Hierarchy of Lorenz Curves Based on the Generalized Tukey's Lambda Distribution, Econometric Reviews, 16, pp. 305-320 (1997)

Sarabia, J. M.: Parametric Lorenz Curves: Models and Application in Modeling Income Distributions and Lorenz Curves (ed. Chotikapanich, D.), Chapter 9, pp.167-190, Springer, New-York (2008)

Ultsch, A., Lötsch, J.: A data science based standardized Gini index as a Lorenz dominance preserving measure of the inequality of distributions. PLoS ONE 12(8): e0181572 (2017)

Figure legends

Figure 1: Lorenz curve $\boldsymbol{L}$ and its associated Gini index.

Figure 2: Lorenz curve $\boldsymbol{L}$ (red) and its reflection $\boldsymbol{L}^*$ (dashed blue).

Figure 3: Five examples of World Bank data for three countries each with the same Gini value. The data points show a good fit between the corresponding deciles.

Figure 415: Pareto Lorenz curve (left) and corresponding Pareto Lorenz density (right) for $\boldsymbol{G} = \mathbf{0.25}$, where $\boldsymbol{\varepsilon} = \mathbf{0.6}$.

Figure 5: Polynomial Lorenz curve (left) and corresponding Lorenz density (right) for $\boldsymbol{G} = \mathbf{0.25}$, where $\boldsymbol{\varepsilon} = \mathbf{0.6}$.

Figure 6: Lorenz curve with a cut-off point δ (left) and the cut-out Lorenz curve (right). The rectangle [δ, 1] x [L(δ), 1] (left) is stretched to the unit square $[0, 1]^2$ (right).

Figure 7: Approximation of World Bank data for Germany, India and the USA with Pareto (left column) and polynomial (right column) Lorenz curves, with an identical Gini value taken from the World Bank data.

Figure 8: Approximation of World Bank data for Germany, India and the USA with mixed Lorenz curves with optimal value α_{opt} (left column), which is different for each country, and the value $\alpha = 0.6$, the standard value, which is the same for all countries (right column).

Figure 9: Countries sorted by descending Gini values (blue curve, country index). The red dots show the optimal $\boldsymbol{\alpha}$. The values for $\boldsymbol{\alpha}$ and $\boldsymbol{G}$ can be read on the y-axis.

Figure 10: Two very different Lorenz densities having the same Gini value of 0.4.

Figure 11: Lorenz curves for the Lorenz densities in Figure 10.

Figure 12: Income distribution of the EU as an aggregated system of states.

Figure 13: Income distribution of the EU in 2002, with 15 states (blue) and in 2016, with 28 states (yellow).

Figure 14: Global income distribution according to World Bank data for the years 2002 and 2016.

Appendix

1. **The Gini is unchanged under reflection.** The reflection of a Lorenz curve $L(x), x \in [0,1]$, which means mirroring at the contra diagonal of the unit square, is $L^{ref}(u) = 1 - L^{-1}(1-u)$.

The Gini $G(L)$ of L is $2 \cdot \int_0^1 \big(u - L(u)\big)\, du$. The Gini $G^*(L^*)$ of L^* ist thus

$$2 \cdot \int_0^1 \big(u - L^*(u)\big)\, du = 2 \cdot \int_0^1 \Big(u - \big(1 - L^{-1}(1-u)\big)\Big)\, du$$

Using the integral formula and observing $L(0) = 0$ and $L(1) = 1$, we look at

$$\int_{f(a)}^{f(b)} f^{-1}(y)dy = [xf(x)]_a^b - \int_a^b f(x)dx$$

$$2 \cdot \int_0^1 \big(u - 1 + L^{-1}(1-u)\big)du$$

setting $v = 1 - u$ leads to

$$= \ -2 \cdot \int_1^0 \big(-v + L^{-1}(v)\big)\, dv = 2 \cdot \int_0^1 \big(-v + L^{-1}(v)\big)dv$$

$$= -1 + 2 \cdot [v \cdot L(v)]_0^1 - 2 \cdot \int_0^1 L(v)dv = -1 + 2 - 2 \cdot \int_0^1 L(x)dx$$

$$= 1 - 2\int_0^1 L(u)du = 2 \cdot \int_0^1 \big(u - L(u)\big)du = G(L)$$

2. **The Reflection of the Pareto Lorenz curve is the Polynomial Lorenz curve**

$y = L(u) = 1 - (1-u)^{\varepsilon}$ (Pareto)

$$\Leftrightarrow 1 - y = (1-u)^{\varepsilon} \Leftrightarrow (1-y)^{\frac{1}{\varepsilon}} = 1 - u \Leftrightarrow u = 1 - (1-y)^{\frac{1}{\varepsilon}} = L^{-1}(y)$$

$$L^{-1}(1-x) = 1 - \big(1 - (1-x)\big)^{\frac{1}{\varepsilon}} = 1 - x^{1/\varepsilon}$$

$$L^{ref}(x) = 1 - L^{-1}(1-x) = 1 - \left(1 - x^{\frac{1}{\varepsilon}}\right) = x^{1/\varepsilon}$$

3. **The Lorenz curve of a single income level, is the main diagonal of the unit square.**

What is the Lorenz density and the Lorenz curve for a one-point distribution with single income level x_0. The distribution function in this case is

$$F(x) = \begin{cases} 0 & if\ x < x_0 \\ 1 & if\ x_0 \leq x \end{cases} .$$

The **generalized inverse** F^{-1} for any increasing and right-continuous function $F: \mathbb{R} \to [0,1]$ with $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$ is defined for $x \in [0,1]$ by

$$F^{-1}(u) = \inf\{x \mid F(x) \geq u\}.$$

Intuitively, the generalized inverse F^{-1} thus indicates the minimum income level such that the probabilities of all incomes up to that level accumulate to a given probability value u, or more. In our case, $F^{-1}(u) = x_0$ for all $u \in (0,1)$. $F^{-1}(u)$ is thus the uniform density of value x_0 on the interval $(0,1)$ with expectation $\int_0^1 F^{-1}(v)dv = x_0$. Normalisation by dividing by x_0 gives the uniform density of value 1 on the interval $(0,1)$. $L(u)$ then is the main diagonal of the unit square, resulting in the Gini value $'0'$.