Supplementary Information

Measuring Inequality Beyond the Gini Coefficient May Clarify Conflicting Findings

Kristin Blesch, Oliver P. Hauser, Jon M. Jachimowicz

E-Mail Addresses: blesch@leibniz-bips.de,o.hauser@exeter.ac.uk, jjachimowicz@hbs.edu

This PDF file includes:

Supplementary Figures 1 to 24
Supplementary Tables 1 to 11
SI References

Notation Preface

- Gamma function: $\Gamma(\nu) = \int_0^\infty \exp^{-t} t^{\nu-1} dt$
- Lower incomplete gamma function ratio: $G(x,\nu) = \int_0^x t^{\nu-1}\exp(-t)dt/\Gamma(\nu)$
- Lower incomplete beta function ratio: $B(x;a,b) = \frac{\int_0^x t^{a-1}(1-t)^{b-1}dt}{\int_0^1 t^{a-1}(1-t)^{b-1}dt}$

1. Functional forms of Lorenz curve models

Properties. To ensure that the proposed functional form can serve as a Lorenz curve model, certain properties of Lorenz curves should be satisfied. As described in (1–3), general properties of the Lorenz curve L with respect to the cumulative percentages of the population p are the following:

1. $L(u)$ is monotone increasing
2. $L(u) \leq p$
3. $L(u)$ is convex
4. $L(0) = 0$ and $L(1) = 1$

More formally, the following theorem (cited by (4, 5) but attributed to Pakes 1981) determines what functions qualify as Lorenz curves:

Theorem 1 (Lorenz curve)
A function $L(u)$, continuous on $[0,1]$ and with second derivative $L''(u)$ is a Lorenz curve if and only if $L(0) = 0, L(1) = 1, L'(0^+) \geq 0, L''(u) \geq 0$

Supplementary Table 1. 1.-9. Lorenz curve models from distributional origin. 10.-17. Functional forms proposed to model Lorenz curves. Model 14 is recognized as a family of Lorenz curves but not proposed as a Lorenz curve specifically. As this family is the most general form of the specific Lorenz curve that Sarabia proposes, we use it as a four-parameter Lorenz curve (see (4, 6–9)). η denotes the cumulative percentage of income, u denotes the cumulative percentage of the population. $\Phi()$ is the cumulative distribution fucntion of the standard normal distribution, $G()$ is the incomplete gamma function ratio, $B()$ is the lower incomplete beta function ratios as defined in SI Section 1.

Originates from	Lorenz curve $\eta(u)$	# Par.	Parameter restrictions
1. Pareto distribution	$1-(1-u)^{1-1/\alpha}$	1	$\alpha > 1$
2. Lognormal distribution	$\Phi(\Phi^{-1}(u)-\sigma)$	1	$\sigma > 0$
3. Gamma distribution	$G(G^{-1}(u;\sigma);\sigma+1)$	1	$\alpha, \sigma > 0$
4. Weibull distribution	$G(-\log(1-u); \frac{1}{\alpha}+1)$	1	$\alpha > 0$
5. Gen. Gamma distr.	$G\left(G^{-1}(u;p); p+\frac{1}{a}\right)$	2	$a, p > 0$
6. Dagum distribution	$B(u^{1/q}; q+\frac{1}{a}, 1-\frac{1}{a})$	2	$q > 0; a > 1$
7. Singh-Maddala distr.	$B(1-(1-u)^{1/q}; 1+\frac{1}{a}, q-\frac{1}{a})$	2	$q, a > 0,\ q > \frac{1}{a}$
8. GB1 distribution	$B(B^{-1}(u;p,q); p+\frac{1}{a}, q)$	3	$p, q, a > 0$
9. GB2 distribution	$B\left(B^{-1}(u;p,q); p+\frac{1}{a}, q-\frac{1}{a}\right)$	3	$p, q, a > 0;\ q > \frac{1}{a}$
10. Kakwani/Podder [1973] (10)	$ue^{-\beta(1-u)}$	1	$\beta > 0$
11. Rasche et al. [1980] (11)	$(1-(1-u)^\alpha)^{1/\beta}$	2	$0 < (\alpha,\beta) \leq 1$
12. Ortega et al. [1991] (12)	$u^\alpha(1-(1-u)^\beta)$	2	$\alpha \geq 0; 0 < \beta \leq 1$
13. Chotikapanich [1993] (13)	$\frac{e^{ku}-1}{e^k-1}$	1	$k > 0$
14. Sarabia et al. [1999] (14)*	$u^{\alpha+\gamma}[1-a(1-u)^\beta]^\gamma$	4	$0 \leq a \leq 1; 0 < \beta \leq 1;$ $0 \leq \alpha; \gamma \geq 1$
15. Abdalla/Hassan [2004] (15)	$u^\alpha(1-(1-u)^\delta e^{\beta u})$	3	$\alpha \geq 0; 0 \leq \beta \leq \delta \leq 1$
16. Rhode [2009] (16)	$u \cdot \frac{\beta-1}{\beta-u}$	1	$\beta > 1$
17. Wang et al. [2011] (17)	$\delta u^\alpha[1-(1-u)^\beta]$ $+(1-\delta)[1-(1-u)^{\beta_1}]^\nu$	5	$\alpha \geq 0; \nu \geq 0; \alpha+\nu \geq 1;$ $0 < (\delta,\beta,\beta_1) \leq 1$

2. Detailed Description of Data Cleaning

General Procedure to Match the Datasets. Data from both sources (American Community Survey (ACS) 2011-2015 (18), Economic Policy Institute (EPI) (19)) were collected at the US county level, which allows us to calculate the Lorenz curve representation of the income distribution using the following procedure: recall that the Lorenz curve is depicted through the cumulative share of population on the x-axis and cumulative share of income on the y-axis. We therefore construct a dataset that contains the share of population (from low-income to high-income) who own a certain percentage of total income, such that we can draw a Lorenz curve using the cumulative sum of these data points.

While the EPI report already presented the high-income earner data in such a way, further processing had to be undertaken for the ACS data: the data were given as headcounts per income bucket, which required transformation to income shares for the Lorenz curve representation. For this transformation, we assumed that people within income buckets were distributed symmetrically around the mean of the respective bucket. For example, a uniform distribution of people within an income bucket seems plausible in that people's income is likely to be equidistantly spread between the narrow boundaries of 45 000 USD and 49 999 USD per year. We could then calculate the volume of income held by the people belonging to that bucket by multiplying the number of people in the respective income bucket with the mean value of the bucket range, and then dividing this number by the aggregate income in that county, giving us the share of total income. Based on this transformation, a Lorenz curve could be constructed for each US county. To verify that our approximated Lorenz curve data are in line with the true income share percentiles of that ACS dataset (the 20^{th}, 40^{th}, 60^{th}, 80^{th} and 95^{th} income share percentiles are provided), we evaluated deviations between our approximated Lorenz curve and true income share data from ACS. We found good agreement between the approximated Lorenz curves with the ACS income shares, which we detail in Section 3.

Matching the ACS and EPI datasets revealed that, on average, the EPI data implied a higher level of inequality than the ACS data. This may arise in part because the EPI data are based on actual tax records at the taxpayer level, whereas the ACS data are from a self-reported survey at the household level, the latter of which is already an aggregate that typically underestimates the inequality suggested by the according Lorenz curve (20). For both ACS and EPI data, the exact 95^{th} percentile was available, which enabled us to perform an exact scaling, i.e., adjusting the ACS household-level data to the EPI taxpayer-level data, using this data point as a link between datasets, see section 3 detailing this procedure. We adjust to the taxpayer level because it reflects the true level of income inequality in that individuals earn income, not households as a unit itself. We further believe that the EPI data are closer to reality, as tax reports are more difficult to manipulate and do not rely on self-reports that might be inaccurate, falsely remembered, or strategically misreported.

Merging Source Tables. This subsection =describes the code data_cleaning_merge_b6_nhigs.R which was used to merge the raw data tables provided by ACS and EPI.

We merge Tables B6 and B4[*] from https://www.epi.org/publication/income-inequality-in-the-us/#epi-toc-20 and Tables NHGIS A and NHGIS B from https://data2.nhgis.org/main that are from the American Community Survey 2011-2015. Source Table NHGIS A is taken from the dataset with NHGIS code 2011_2015_ACSa, and the source codes of the variables are B19001, B19013, B19025. Source Table NHGIS B is taken from the dataset with NHGIS code 2011_2015_ACSb, and the source codes of the variables are B19080, B19081, B19082, B19083. As additional information, a file with abbreviations and full names of US states (e.g. AK = Alaska) is taken from https://developers.google.com/public-data/docs/canonical/states_csv.

The procedure to merge the source tables is as follows:

- Load data and exclude Puerto Rico and the District of Columbia
- Merge ACS data NHGIS A and NHGIS B by county name such that all data from the survey are in a single dataset
- Adjust county names to prepare for the match: let the B6 county names (format: "San Francisco, CA") look like NHGIS county names (format: "San Francisco County, California"). To do so, the B6 county data is split at "," to separate the county name and state name. With the additional file on state abbreviations and names, the county state abbreviations are transformed into their actual name (e.g. from CA to California). Not only does the state name abbreviation differ in the B6 from the NHGIS format; it also says "San Francisco *County*, California". Therefore, to create a new B6 column that looks like the NHGIS county name, the county name (San Francisco), the word "County", ",", and the full state name "California" are pasted into a single column such that we end up with a column in B6 of the county name format "San Francisco County, California" to match with NHGIS

  - For the special cases Census Areas or Cities: don't paste "County" after "Census Area" or "City"

  - For the special case Alaska: Alaska is not divided into counties but into cities, boroughs, or census areas. NHGIS names them as City/Borough/Census Areas, but B6 does not, so we omit everything after the first word (which is a unique determinant of the actual area) in both datasets to derive a matching name for the corresponding area in Alaska

  - For the special case Louisiana: Louisiana is not divided into counties but into Parishes, so we paste "Parish" instead of "County" after county names in Louisiana

- Transform encoding of NHGIS data from 'ISO-8859-1' to = 'UTF-8'

[*] Note that B4 is relevant not for the present study but for other (future) studies that intend using this dataset.

- Use a fuzzy string matching algorithm to merge B4/B6 and NHGIS data by county name: Fuzzy string matching has to be double checked by visual inspection of the county names to ensure that only correct merges have taken place. Iterative procedure to minimize the amount of counties that have to be inspected and matched by hand: From all the imperfect matches (distance > 0), which exhibit a very similar pattern, e.g. "St." instead of "St", transform "St." to "St" such that all of these cases are now perfect matches (distance $= 0$). Fuzzy match again and repeat procedure. When most of the common structures like "St." -> "St" are cured, we can inspect the resulting imperfect matches for counties that we need to match by hand. For some counties, different names exist, e.g. Shannon County, South Dakota, is another name for Oglala Lakota County, South Dakota

- Write a single .csv file for the merged tables

3. Calculation of the Lorenz Curves

This subsection describes the code create_lorenz_curves.R to calculate Lorenz curve values for each county. The goal is to calculate the share of income held by shares of the population (from low-income to high-income). A quick recap of the information that the ACS and EPI source tables give us:

- Table B6: Income share held by 90^{th},95^{th} and 99^{th} percentile of the population $\rightarrow$ no further transformation needed

- NHGIS B: Income share held by 20^{th}, 40^{th}, 60^{th}, 80^{th} and 95^{th} percentile of the population $\rightarrow$ no further transformation needed

- NHGIS A: Aggregated income per county, people per county, count of people that fall into a certain income bucket, e.g., have an income between $45,000$ USD and $49,999$ USD a year (see codebook in zip file for details) $\rightarrow$ need to transform this information, procedure:

  - Assume that people are symmetrically distributed around the mean value of the income bucket range within each closed income bucket, i.e., we do not use the top income bucket $> 200,000$ USD.

  - Use mean value of the income bucket range multiplied by the number of people that fall into that bucket as estimate of the income held by people belonging to the corresponding income bucket.

  - Divide this number by the income aggregate for the respective county, such that we end up with the share of total income held by the income bucket

  - Divide the number of people belonging to that income bucket by the total number of people in that county to get the share of people belonging to that income bucket

- Check for consistency in the ACS dataset: Inspect whether the estimated income shares per bucket are coherent with the information on the (true) income shares held by the 20^{th}, 40^{th}, 60^{th}, 80^{th}, and 95^{th} percentile of the population $\rightarrow$ found to be consistent; see related Supplementary Figure 1.

- Merge Lorenz curve data from ACS and EPI: Table B6 systematically suggests a higher level of inequality than the ACS data. This is a well-known phenomenon (20), as the ACS is at the household level (already an aggregate, e.g., two income earners living together in a household) whereas the B6 data are at the taxpayer level). We favor B6 data to depict a more realistic picture of the true inequality and hence decided to scale the ACS data to match the B6 data at the 95^{th} percentile:

  - We have exact information on the 95^{th} percentile, so we can use the 95^{th} percentile as the anchor point for scaling to account for the difference in the data induced by the fact that B6 is at the taxpayer level and NHGIS at the household level. This means that we multiply the NHGIS percentile data by the 95^{th} percentile of the B6 data and then divide it by the 95^{th} percentile of the NHGIS data. To ensure convexity, we use solely ACS data below the 95^{th} percentile and solely EPI data above the 95^{th} percentile.

  - Check for data consistency prior and post scaling: Visually, most of the scaled data are close to the non-scaled data. However, as an example of an extreme case, which also illustrates that Table B6 delivers valuable information, we can look at Teton County, WY, further described in 3.

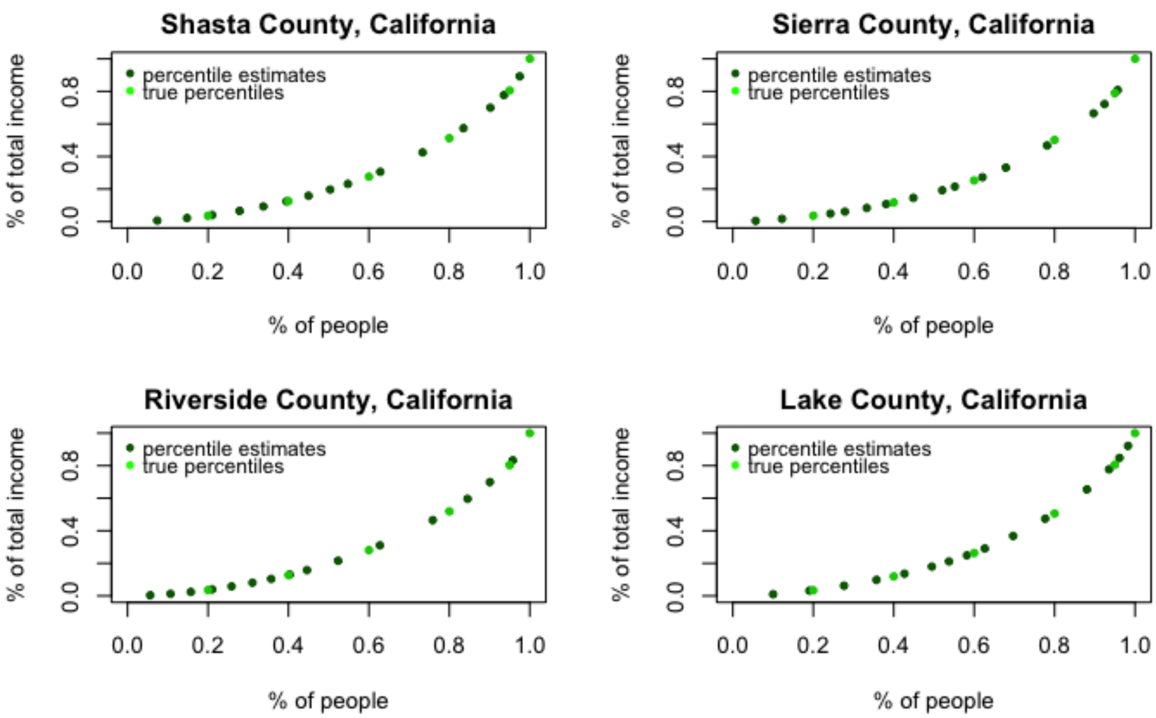
Systematic Evaluation of Constructed Lorenz Curves. We have already performed a brief cross-check for data consistency of the ACS dataset; i.e., we checked whether our approximation of income shares using the income buckets is close to the few true income share percentiles provided by the ACS. Now, we check the consistency of the ACS data more systematically.

We estimated the share of total income held by each income bucket (for all closed income buckets; i.e., we omit the top income bucket $> 200,000$ USD) under the assumption of symmetrically distributed incomes around the mean income within each income bucket. As we have true income share percentiles for some percentiles of the population, namely, the $20^{th}, 40^{th}, 60^{th}, 80^{th}$, and 95^{th} population percentiles, we can evaluate our estimated income shares by adding the true percentiles to our estimated Lorenz curves and for their fit. Remember that empirical Lorenz curves are defined by data points that are then linearly interpolated. Hence, we also linearly interpolate between our estimated income percentiles and calculate the estimated income percentile at the $20^{th}, 40^{th}, 60^{th}$, and 80^{th} population percentile for which the ACS provides exact data. This allows us to calculate the residual sum of squares (RSS) between the estimated income percentile at the $20^{th}, 40^{th}, 60^{th}$ and 80^{th} population percentiles and the true $20^{th}, 40^{th}, 60^{th}$ and 80^{th} percentiles.[†]

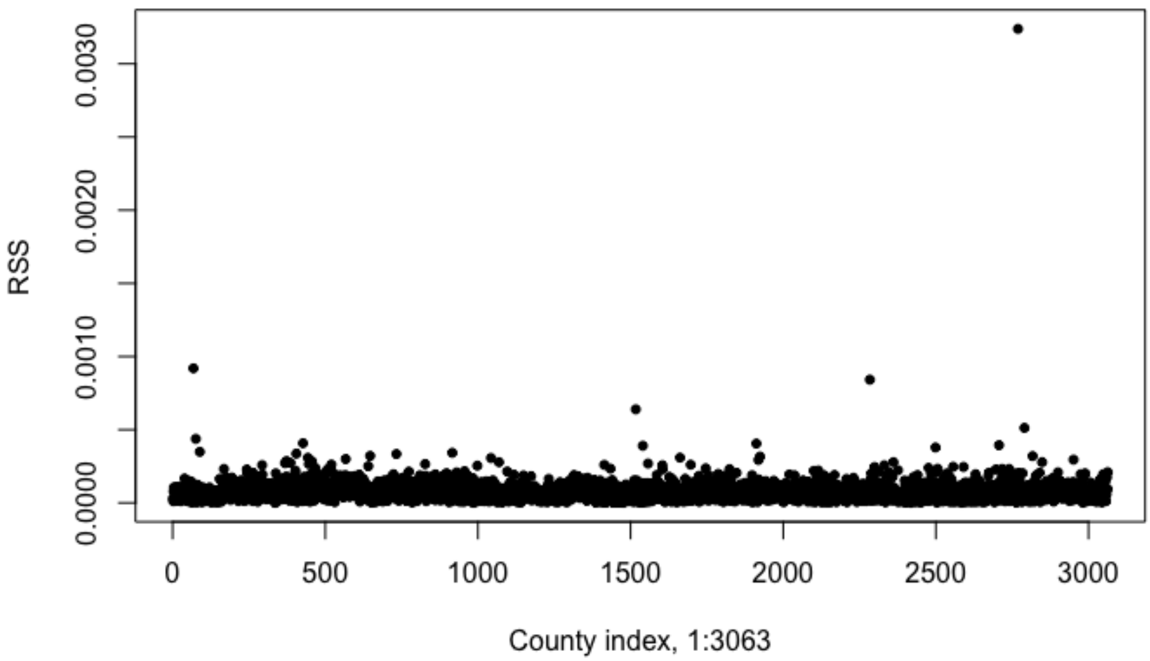
While Supplementary Figure 1 already suggested that the estimated income percentiles from the income buckets seem to fit very well to the true income percentiles, we aim to quantify the fit more formally and calculate the RSS as described above. In Supplementary Figure 2, we can see one clear outlier, and potentially three more. Hence, we take a closer look at the counties with the top four RSS scores, which turn out to be [1] Falls Church city, Virginia, [2] Monroe County, Alabama, [3] Allendale County, South Carolina, and [4] Holmes County, Mississippi.

The Lorenz curve plots of these counties, depicted in Supplementary Figure 3, reveal the following: for the county with the highest RSS score, Falls Church, we can clearly see that this high RSS score results from the fact that a significant fraction of its population falls into the top income bucket, $> 200\,000$ USD. This forces a linear interpolation straight from a 0.73 percentile to the boundary of (1,1). We know this interpolation is not trustworthy, which is why we enrich the data at the top percentiles with EPI data and hence a comparably large deviation from the true 80^{th} percentile should not worry us too much. For the remaining counties, the percentiles still seem to fit the Lorenz curve reasonably well. Therefore, we can conclude that there is no need to exclude any outliers from further analyses.

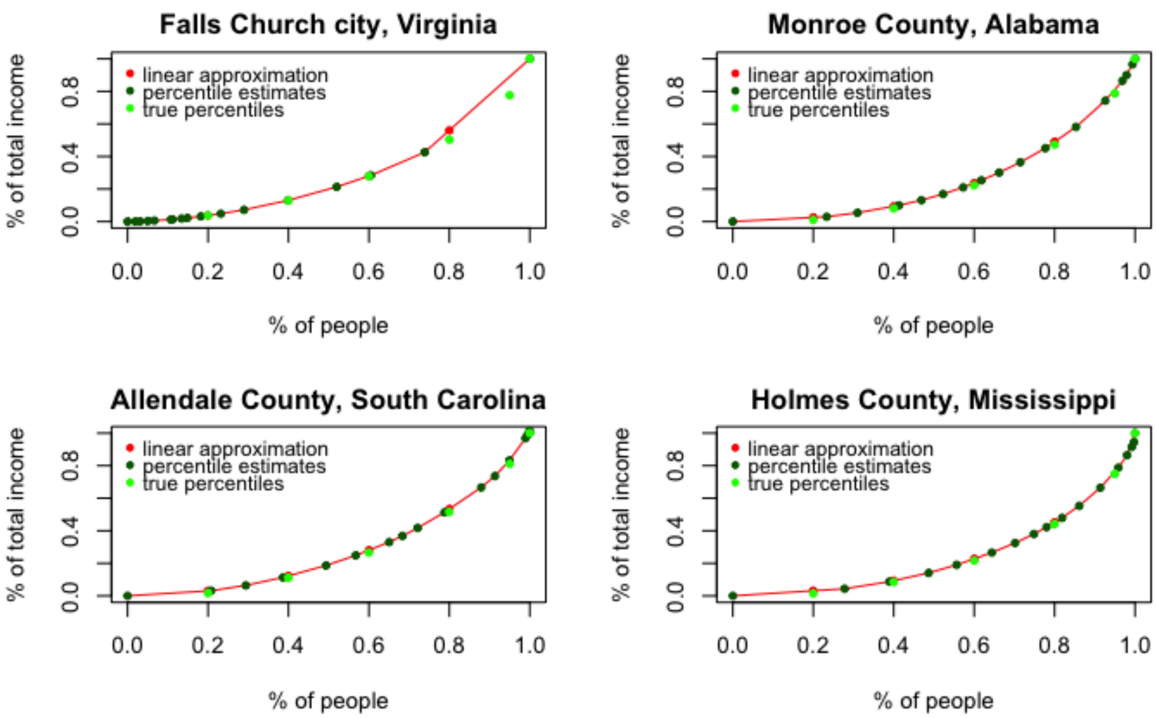
[†] We omit the 95^{th} from the analyses here because we know that linear approximation is not a good approximation for top income shares, which is why we use EPI data from B6 for the 95^{th} percentile and above.



Supplementary Figure 1. Estimated and true income percentiles for some exemplary counties



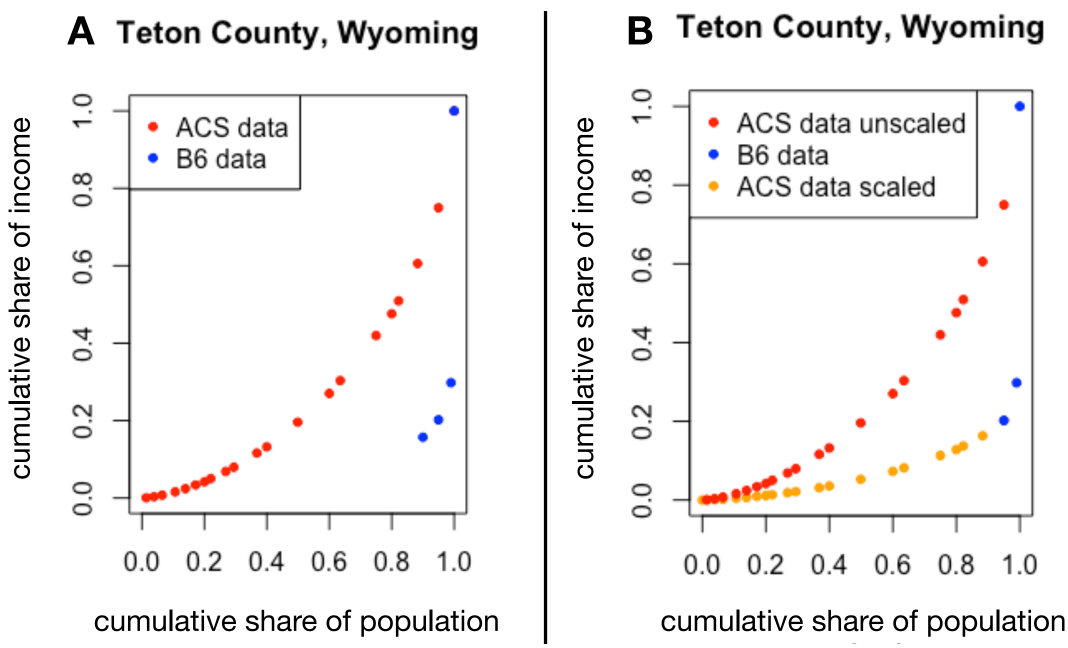
Supplementary Figure 2. RSS for estimated percentiles of income shares. Here, we refer to residuals as the difference between the true income share and estimated income share. Residuals are then squared and summed over all available data points. This was performed for each county out of all 3063 counties.



Supplementary Figure 3. Interpolated Lorenz curves from estimated income shares for the four counties with highest RSS scores in Supplementary Figure 2.

Scaling the Data in an Exemplary County: Teton, Wyoming. Teton, WY, is an example of a county that exhibits a special distribution of income that we could not have guessed with the ACS data alone. The data of the American Community Survey alone are fine-grained for low and medium income levels, yet the ACS data alone might lead to unrealistic approximations of the top populations' income shares, as the top income bucket $> 200\,000$ USD is an open interval that does not provide any information on how people are distributed within that interval. Table B6, however, gives us detailed information on the income shares of the top-income percentages of the population on the taxpayer level.

Apparently, there are a few people living in Teton, WY, that have an income far above the threshold 200 000 USD. In Supplementary Figure 4 Panel A, we can clearly see that the income share of the top 5% and top 1% percent of income earners far exceeds what we would have expected from the American Community Survey data. Now looking at the scaled data presented in Panel B of Supplementary Figure 4, i.e., taking into consideration the information from the EPI dataset, we can clearly see the immense difference. This example highlights the importance of considering Table B6 as an additional data resource for the construction of close-to-reality Lorenz curves.



Supplementary Figure 4. Panel A provides raw Lorenz curve data from ACS and Table B6; Panel B depicts scaled data for Teton County, Wyoming.

4. Maximum Likelihood Estimation (MLE) via Dirichlet Distribution

An approach to estimate Lorenz curves based on maximum likelihood estimation (MLE) was proposed by Chotikapanich et al. (2002) (21). They assume the income shares from grouped data to follow a Dirichlet distribution. Chang et al. (2018) (22) agree with this perspective and argue that the Dirichlet distribution "naturally accommodates the proportional nature of income share data and the dependence structure between the shares" (22, p. 2), which is a major advantage compared with the NLS estimation procedure (15). Chotikapanich et al. (2002) (21) demonstrate analytically that it is possible to relate desired functional forms of the Lorenz curve to the Dirichlet parameters; i.e., parameters of the Dirichlet distribution are set so that they incorporate the proposed functional form of the Lorenz curve with its parameters. The density of the Dirichlet distribution (with newly defined parameters that consist of the Lorenz curve parameters) is then used to construct the likelihood that is maximized later on. In detail, the procedure to model the Lorenz curve models with maximum likelihood estimation using the Dirichlet distribution described in (21) is as follows:

Let $\eta_i = L(u_i; \boldsymbol{\theta})$ be the cumulative income share held by the cumulative share of the population u_i. Then, $q = (q_1, \ldots, q_M)$ with $q_i = \eta_i - \eta_{i-1}$ are assumed to be random variables that follow a Dirichlet distribution. The probability density function of the Dirichlet distribution is given by

$$f(q|\alpha) = \frac{\Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_M)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_M)} \cdot q_1^{\alpha_1 - 1} q_2^{\alpha_2 - 1} \ldots q_M^{\alpha_M - 1}$$

where the gamma function is defined as $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \exp^{-x} dx$. The method is now to relate the parameters α of the Dirichlet distribution to the functional form of the Lorenz curve that we want to estimate. This can be conveniently be done by setting

$$\alpha_i = \lambda[L(u_i; \boldsymbol{\theta}) - L(u_{i-1}; \boldsymbol{\theta})]$$

where λ is an additional unknown parameter. Now we can write the probability density function as

$$f(q|\lambda, \boldsymbol{\theta}) = \Gamma(\lambda) \prod_{i=1}^{M} \frac{q_i^{\lambda[L(u_i;\boldsymbol{\theta}) - L(u_{i-1};\boldsymbol{\theta})]-1}}{\Gamma(\lambda[L(u_i; \boldsymbol{\theta}) - L(u_{i-1}; \boldsymbol{\theta})])}$$

To now estimate the parameters, we simply have to maximize the log-likelihood that takes the form

$$\log[f(q|\lambda, \boldsymbol{\theta})] = \log\Gamma(\lambda) + \sum_{i=1}^{M}(\lambda[L(u_i; \boldsymbol{\theta}) - L(u_{i-1}; \boldsymbol{\theta})] - 1) \cdot q_i - \sum_{i=1}^{M} \log\Gamma(\lambda[L(u_i; \boldsymbol{\theta}) - L(u_{i-1}; \boldsymbol{\theta})])$$

This maximum likelihood based estimation of Lorenz curve parameters is, however, not widely used. The original study of (21) was replicated and advanced by (22) and (15), finding mixed results. In detail, (22) find that the MLE estimation via the Dirichlet distribution provides a better fit to empirical data, and (15) find that NLS provides a "better and more reliable fit compared to the maximum likelihood estimation" (15, p. 117)). Moreover, (21) find that most Lorenz curve parameter estimates are not sensitive to the estimation method; i.e., they compared parameters estimated by NLS and MLE and found them yielding very similar point estimates for the parameters for most Lorenz curves proposed (but not all of them, which they attribute to estimation instability). (15) find similar point estimates of NLS and MLE as well, but report, as (21), much larger standard errors of the estimated parameters of the MLE method.

5. Akaike Information Criterion (AIC) and AIC$_c$ Simulation Study

While the AIC measure of goodness-of-fit is well known as a tool for model selection in many fields of applied statistics, such as ecology (23) or astrophysics (24), it has not previously been used to systematically analyze the optimal number of parameters needed to adequately represent empirical Lorenz curves. One reason the AIC has not been used in prior literature may be the more common use of nonlinear least squares (NLS) approaches as an estimation procedure for Lorenz curves, which does not allow for the use of AIC. The NLS approach is widespread because it does not impose distributional assumptions on the data, which is a requirement for MLE. However, within the NLS framework, researchers typically rely on the residual sum of squares as a measure of goodness-of-fit. Residual sum of squares does not trade-off fit for model complexity, which commonly results in the most complicated Lorenz model as the winner. For our research question—determining how many parameters are necessary to capture relevant information—we therefore focus on the MLE/AIC framework in order to balance complexity and model fit. As mentioned in the paper, we use the small-sample bias adjusted version of the criterion, namely AIC$_c$.

Our key question we want to answer with our simulation study is: Will AIC$_c$ suggest that we use the correct model? To answer this question, we will simulate Lorenz curve data points according to a certain model. Based on these data points, we will estimate the parameters of all 17 models we analyzed in the previous chapters and then let AIC$_c$ choose the best model. If AIC$_c$ actually picks the correct model that was used for data generation sufficiently often, the reliability of AIC$_c$ as a criterion for model selection is supported for our setting. However, if AIC$_c$ fails to pick the correct model, we have to question our previous results and take them with a (big) grain (rock) of salt. We will vary the sample size, i.e., the number of data points used for model estimation, to get a clearer picture of where our setting stands with respect to the extent to which we trust in AIC$_c$ picking the correct model. Only then we can judge whether AIC$_c$ can be used as an indicator of the number of parameters needed to describe income-inequality Lorenz curves.

Through an AIC_c-based ranking and Borda voting procedure, we found the Ortega Lorenz curve model (2 parameters), the GB2 Lorenz curve model (3 parameters), and the Wang Lorenz curve model (5 parameters) to be among the most suitable models. To verify that our judgment, especially between those three most promising models, is trustworthy, we will focus on those three models for income share generation. In detail, we will run three simulations, with the only difference being the model used to generate the income shares. One might wonder why we run the simulation not only with one exemplary income-generating model but with three models. The reason is that we then can cross-compare results between the income-generating routines. For example, we could detect whether a certain model is preferred by AIC_c regardless of the true data-generating process. In other words, AIC_c might always choose the same model.

Simulation Setup. For ease of comprehensibility, we will describe the simulation procedure in a numbered list. The structure of the simulation study is as follows:

1. Generate a vector that imitates population shares: $\pi = (0, \pi_1, \ldots, \pi_n, 1)$ with $\pi_i \sim \text{Unif}(0{,}1)$.
2. Generate a vector of cumulative income shares $\eta = L(\pi, \theta)$, where $L(\theta)$ is a known Lorenz curve model of either type Ortega, GB2, or Wang with known parameters θ[‡] and population shares π that were generated in the previous step. For each Lorenz curve model used for income-share generation, we run a separate simulation.
3. Use MLE to fit all 17 Lorenz curve models[§] to the data generated above and store the model name with minimum AIC_c value.
4. Evaluate whether AIC_c has chosen the model that was used to generate the cumulative income shares or not.
5. Repeat this procedure for $sim = 1\,000$ population share vectors generated. Then vary the length of the population share vector and apply the same procedure.
6. Evaluate the percentage of instances where AIC_c was able to detect the model that was used for income-share generation for each vector length and each of the the Lorenz curve models that are used to generate income.

Simulation Results. Results show that we observed a high true-model detection rate even for small sample sizes, see Tables Supplementary Table 2, Supplementary Table 3, Supplementary Table 4, and Figures Supplementary Figure 5, Supplementary Figure 6, Supplementary Figure 7. For our sample size range of 19-23 data points—and assuming that the two-parameter Ortega truly was the Lorenz curve generating model—the true discovery rate would be ≥ 0.97 (lower bound of 95% confidence interval), see Supplementary Table 4 and Supplementary Figure 7). This result provides additional confidence in the reliability of AIC_c given our specific setting.

[‡] To find reasonable parameters, we used the mean value across the US county parameter estimates.

[§] See Table 1 in the paper

Supplementary Table 2. Rate of bias corrected AIC picking the true data-generating model for varying sample sizes out of $1\,000$ simulation runs. A sample size of 102 means we have 100 data points generated between 0 and 1, plus 0 and 1 as boundary values. Lower and upper bounds correspond to the 95% confidence interval, based on a binomial test. True model: GB2

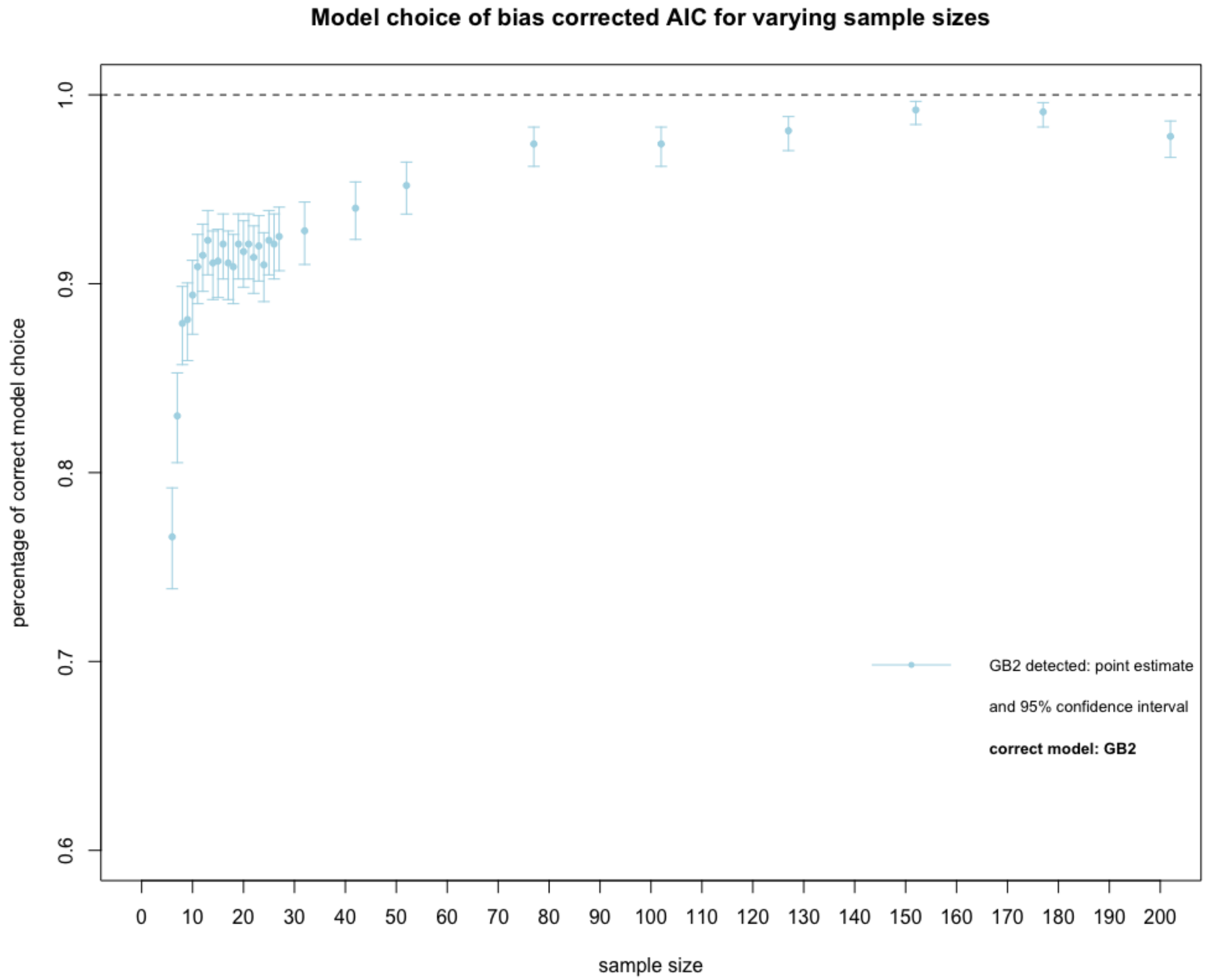
sample size	rate	lower bound	upper bound
6	0.766	0.738	0.792
7	0.830	0.805	0.853
8	0.879	0.857	0.899
9	0.881	0.859	0.900
10	0.894	0.873	0.912
11	0.909	0.889	0.926
12	0.915	0.896	0.932
13	0.923	0.905	0.939
14	0.911	0.892	0.928
15	0.912	0.893	0.929
16	0.921	0.903	0.937
17	0.911	0.892	0.928
18	0.909	0.889	0.926
19	0.921	0.903	0.937
20	0.917	0.898	0.933
21	0.921	0.903	0.937
22	0.914	0.895	0.931
23	0.920	0.901	0.936
24	0.910	0.891	0.927
25	0.923	0.905	0.939
26	0.921	0.903	0.937
27	0.925	0.907	0.941
32	0.928	0.910	0.943
42	0.940	0.923	0.954
52	0.952	0.937	0.964
77	0.974	0.962	0.983
102	0.974	0.962	0.983
127	0.981	0.970	0.989
152	0.992	0.984	0.997
177	0.991	0.983	0.996
202	0.978	0.967	0.986

Supplementary Table 3. Rate of bias corrected AIC picking the true data-generating model for varying sample sizes out of 1000 simulation runs. A sample size of 102 means we have 100 data points generated between 0 and 1, plus 0 and 1 as boundary values. Lower and upper bounds correspond to the 95% confidence interval, based on a binomial test. True model: Wang

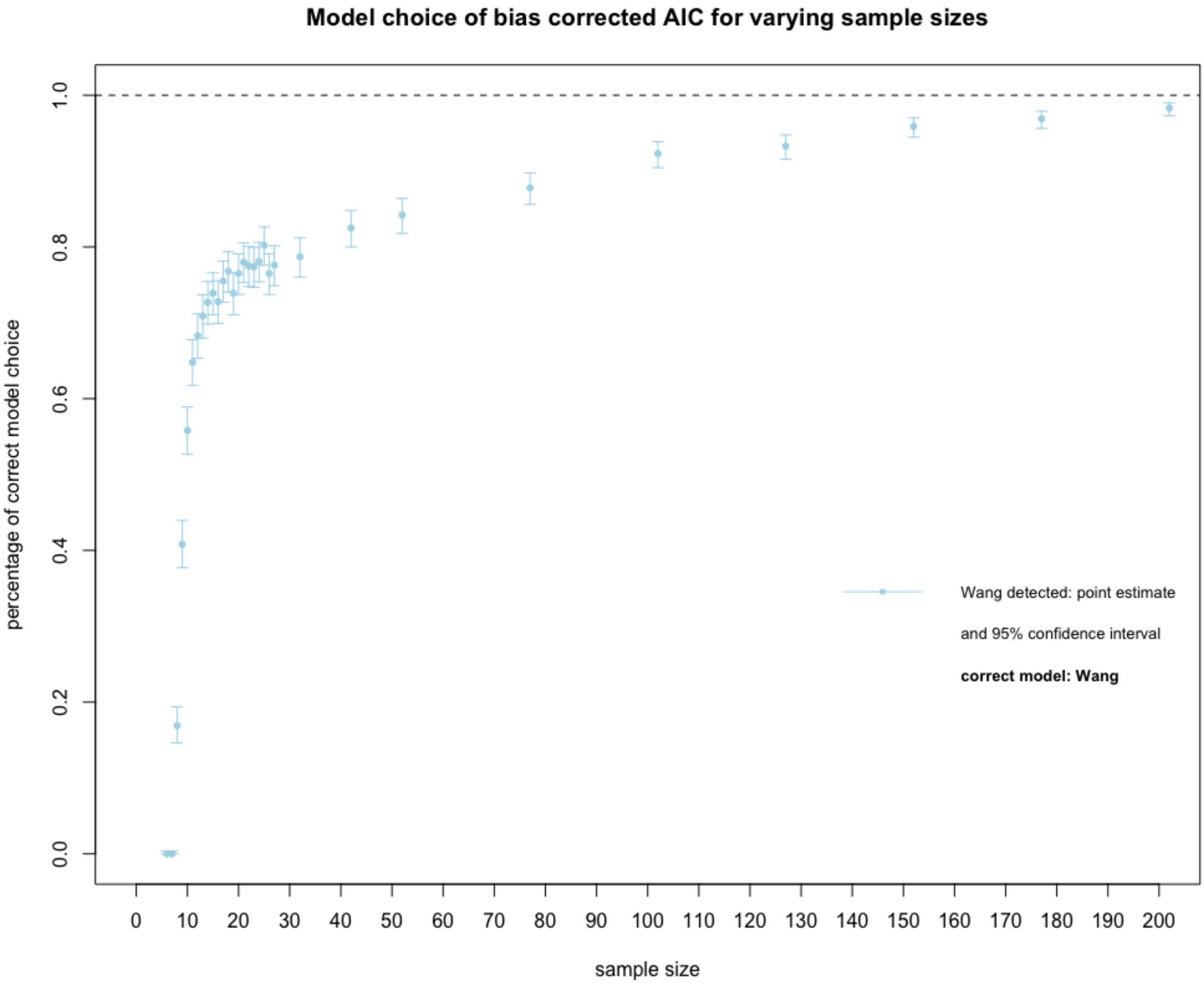
sample size	rate	lower bound	upper bound
6	0.000	0.000	0.004
7	0.000	0.000	0.004
8	0.169	0.146	0.194
9	0.408	0.377	0.439
10	0.558	0.527	0.589
11	0.648	0.617	0.678
12	0.683	0.653	0.712
13	0.709	0.680	0.737
14	0.727	0.698	0.754
15	0.739	0.711	0.766
16	0.728	0.699	0.755
17	0.755	0.727	0.781
18	0.768	0.741	0.794
19	0.739	0.711	0.766
20	0.765	0.737	0.791
21	0.780	0.753	0.805
22	0.775	0.748	0.801
23	0.774	0.747	0.800
24	0.781	0.754	0.806
25	0.802	0.776	0.826
26	0.765	0.737	0.791
27	0.776	0.749	0.801
32	0.787	0.760	0.812
42	0.825	0.800	0.848
52	0.842	0.818	0.864
77	0.878	0.856	0.898
102	0.923	0.905	0.939
127	0.933	0.916	0.948
152	0.959	0.945	0.970
177	0.969	0.956	0.979
202	0.983	0.973	0.990

Supplementary Table 4. Rate of bias corrected AIC picking the true data generating model for varying sample sizes out of $1\,000$ simulation runs. A sample size of 102 means we have 100 data points generated between 0 and 1, plus 0 and 1 as boundary values. Lower and upper bounds correspond to the 95% confidence interval, based on a binomial test. True model: Ortega

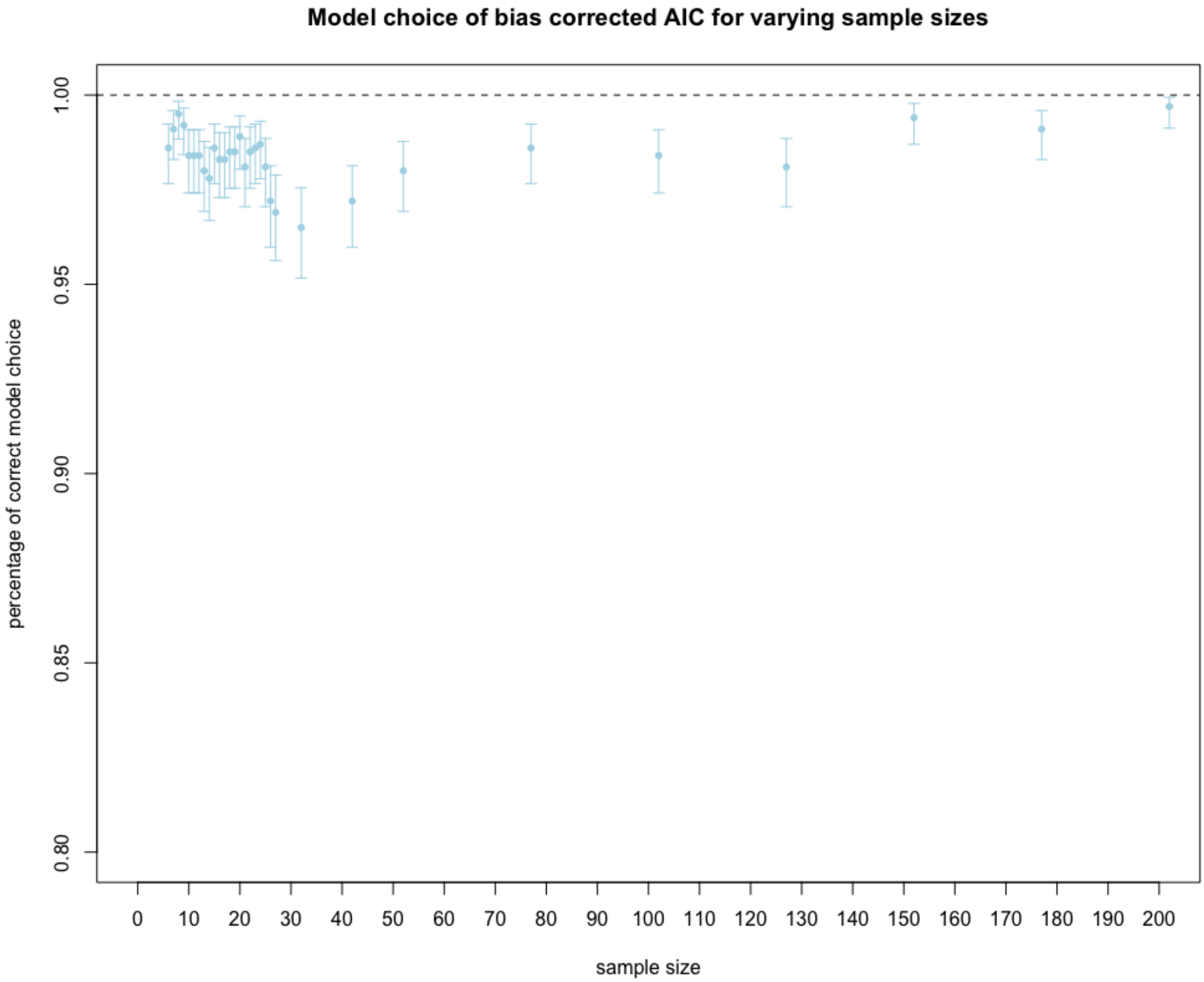
sample size	rate	lower bound	upper bound
6	0.986	0.977	0.992
7	0.991	0.983	0.996
8	0.995	0.988	0.998
9	0.992	0.984	0.997
10	0.984	0.974	0.991
11	0.984	0.974	0.991
12	0.984	0.974	0.991
13	0.980	0.969	0.988
14	0.978	0.967	0.986
15	0.986	0.977	0.992
16	0.983	0.973	0.990
17	0.983	0.973	0.990
18	0.985	0.975	0.992
19	0.985	0.975	0.992
20	0.989	0.980	0.994
21	0.981	0.970	0.989
22	0.985	0.975	0.992
23	0.986	0.977	0.992
24	0.987	0.978	0.993
25	0.981	0.970	0.989
26	0.972	0.960	0.981
27	0.969	0.956	0.979
32	0.965	0.952	0.976
42	0.972	0.960	0.981
52	0.980	0.969	0.988
77	0.986	0.977	0.992
102	0.984	0.974	0.991
127	0.981	0.970	0.989
152	0.994	0.987	0.998
177	0.991	0.983	0.996
202	0.997	0.991	0.999



Supplementary Figure 5. Simulation results for GB2 being the true income share generating model out of a selection of 17 possible models. Point estimates of the percentage of correct model detection are reported together with confidence bounds of the 95% confidence interval.



Supplementary Figure 6. Simulation results for Wang being the true income share generating model out of a selection of 17 possible models.Point estimates of the percentage of correct model detection are reported together with confidence bounds of the 95% confidence interval.



Supplementary Figure 7. Simulation results for Ortega being the true income share generating model out of a selection of 17 possible models.Point estimates of the percentage of correct model detection are reported together with confidence bounds of the 95% confidence interval.

6. Voting

For interested readers, we recommend the literature of the *Handbook of Social Choice and Welfare* (25), which describes the voting procedures in more depth. This section is based on this handbook as well and aims to present voting procedures relevant for our study in a comprehensive way.

According to Arrow's impossibility theorem, there exists no single best voting procedure across the board (26). As a result, researchers have to choose the voting procedure that best fits the problem at hand. We suggest that the Borda count is particularly well suited for our context as it provides insight into which fitted model has good performance across all counties instead of a great fit in some counties but an inferior fit in other counties. We note that others arrive at a different conclusion and prefer a different voting procedure; in that case, we encourage interested readers to use our comprehensive voting results given in the subsection below.

Relying on the principle 'the winner takes all,' *plurality voting* is a simple and intuitive voting procedure. Each individual has one vote, and the candidate receiving the most votes wins. Of course, this reveals only a fraction of the voters' preferences, namely their top choice, but it neglects any remaining preference ordering behind the top choice. In our case, plurality voting corresponds to evaluating which Lorenz curve model was ranked first the most.

A procedure that does not only take the first choice into consideration but performs pairwise comparisons between options is the so-called *Condorcet procedure*. In detail, each option is compared with any other option, and a winner between those options is determined. A quick example illustrates the procedure: Imagine that there are three possible options, A, B, and C, to choose from. Individual 1 has the preference ordering A $>$ B $>$ C[¶] while the preference of individual 2 is B $>$ C $>$ A. To aggregate the preferences of both individuals, we can now compare how often an option was ranked ahead of another option. In this case, option A was preferred over B once (by individual 1), B was preferred over C twice (by individual 1 and 2), and C was preferred over A once (by individual 2); so in this case, the winner of the Condorcet procedure is option B. As we have an AIC_c-based ranking between Lorenz curve models for each county, we can perform such pairwise comparisons across counties. Note that the dominance matrix introduced above depicts these pairwise comparisons, i.e., displays how often a certain model was preferred over the remaining Lorenz curve models.

However, the Condorcet procedure can result in circular preferences and compares the options only in a pairwise fashion. A voting procedure that fully takes into account the ranking of the options is the so-called *Borda count*. This procedure scores the different options according to their ranks. In detail, if there are n options to choose from, the option ranking first receives n points, the option ranking second $n-1$ points, ..., the least favored option receives 0 points. The points received are summed for all individuals, and the option receiving the most points wins the Borda count. Thus, options with a consistently high ranking have a greater chance to win than options that are brilliant for some individuals but heavily undesirable for others. This is exactly the behavior we desire for our Lorenz curve model comparison: we want to detect the model that overall achieves good performance across counties. Therefore, the Borda count is the most relevant voting procedure for our purpose.

Voting results. It is important to again emphasize that the Borda count winner is not the only choice one could make. Other Lorenz curve models winning other voting procedures might be legitimate models as well. The crucial point is that one has to decide which aspects to focus on. By design, different voting mechanisms will lead to different model winners, as they–purposely–emphasize different aspects. Where researchers want to emphasize other aspects, another Lorenz curve model might be more useful. As Arrow's impossibility theorem states, the aggregation of preferences cannot be performed using a single best selection procedure but with different procedures for different kinds of problems and suitable outcomes. For our setting, we find the Borda count procedure superior. However, we do not want to discourage researchers from concluding that other Lorenz curve models might be superior if faced with a different scenario. We therefore provide various voting results below.

In our application, the results are as follows: in plurality voting, the Wang Lorenz curve model wins; applying the Condorcet procedure, the winner is the GB2 Lorenz curve model; and the Borda count winner is the Ortega Lorenz curve model. As the Borda count voting procedure depends on the goodness-of-fit criterion used to judge the models, we cross-check whether those results are driven by AIC_c or whether they are robust to the use of another information criterion. Therefore, we rerun the Borda voting procedure using the Bayesian information criterion (BIC) as indicator to rank the models. The BIC is defined as

$$\text{BIC} = -2 \cdot \ell(\hat{\boldsymbol{\theta}}) + 2p \cdot ln(n)$$

Voting results are similar to the AIC_c-based Borda count; see 6. This result shows that these three models (Wang, GB2, and Ortega) are the most promising.

[¶] In words: Individual 1 prefers option A over B over C, so individual 1 ranks A first, B second, and C third.

	ABDALLA_HASSAN	CHOTIKAPANICH	DAGUM	GAMMA	GB1	GB2	GENERALIZED_GAMMA	KAKWANI_PODDER	LOGNORMAL	ORTEGA	PARETO	RASCHE	RHODE	SARABIA	SINGH_MADDALA	WANG	WEIBULL
ABDALLA_HASSAN	0	3038	1243	3039	2585	161	2800	3037	2894	69	3053	2156	3055	2192	1642	1344	3046
CHOTIKAPANICH	18	0	17	9	1	18	0	2992	18	17	796	17	99	19	17	20	13
DAGUM	1813	3039	0	3042	2594	1154	2840	3038	2984	773	3051	2838	3055	2079	2084	1361	3046
GAMMA	17	3047	14	0	65	15	25	3046	32	12	1919	11	3000	38	11	41	2633
GB1	471	3055	462	2991	0	439	359	3053	467	477	2965	443	3055	603	439	550	3023
GB2	2895	3038	1902	3041	2617	0	2836	3037	2938	1619	3054	2401	3055	2587	2255	1586	3045
GENERALIZED_GAMMA	256	3056	216	3031	2697	220	0	3055	233	237	2808	186	3055	493	179	412	3044
KAKWANI_PODDER	19	64	18	10	3	19	1	0	18	18	753	18	97	19	18	21	14
LOGNORMAL	162	3038	72	3024	2589	118	2823	3038	0	101	2738	45	3055	507	44	360	3035
ORTEGA	2987	3039	2283	3044	2579	1437	2819	3038	2955	0	3055	2684	3055	2442	2599	1533	3048
PARETO	3	2260	5	1137	91	2	248	2303	318	1	0	4	1887	4	0	4	1255
RASCHE	900	3039	218	3045	2613	655	2870	3038	3011	372	3052	0	3055	1682	446	1085	3050
RHODE	1	2957	1	56	1	1	1	2959	1	1	1169	1	0	2	1	4	140
SARABIA	864	3037	977	3018	2453	469	2563	3037	2549	614	3052	1374	3054	0	1079	1102	3030
SINGH_MADDALA	1414	3039	972	3045	2617	801	2877	3038	3012	457	3056	2610	3055	1977	0	1256	3048
WANG	1712	3036	1695	3015	2506	1470	2644	3035	2696	1523	3052	1971	3052	1954	1800	0	3026
WEIBULL	10	3043	10	423	33	11	12	3042	21	8	1801	6	2916	26	8	30	0

Supplementary Figure 8. Condorcet matrix on a county level. Count of how often models in the rows achieve a higher AIC$_c$ rank than models in the columns, out of all $3\,056$ counties.

Supplementary Table 5. Plurality voting results. In each county, the Lorenz curve model with the lowest AIC$_c$ value gets one vote.The model with the highest number of total votes wins.

Num. of Parameters	Model	Votes
5	Wang	998
2	Ortega	546
2	Dagum	399
3	GB2	364
3	GB1	355
4	Sarabia	153
2	Generalized Gamma	80
2	Rasche	70
2	Singh-Maddala	53
1	Lognormal	28
1	Gamma	6
1	Weibull	2
3	Abdalla-Hassan	1
1	Kakwani-Podder	1

Supplementary Table 6. Borda count result using AIC$_c$ as information criterion. In each county, the Lorenz curve models were scored using the Borda count procedure. The model with the highest Borda score wins.

Num. of Parameters	Model	Borda Score
2	Ortega	42597
3	GB2	41906
2	Dagum	38791
5	Wang	38187
2	Singh-Maddala	36274
3	Abdalla-Hassan	35354
4	Sarabia	32272
2	Rasche	32131
1	Lognormal	24749
2	Generalized Gamma	23178
3	GB1	22852
1	Gamma	13926
1	Weibull	11400
1	Pareto	9522
1	Rhode	7296
1	Chotikapanich	4071
1	Kakwani-Podder	1110

Supplementary Table 7. Borda count result using BIC as information criterion. In each county, the Lorenz curve models were scored using the Borda count procedure. The model with the highest Borda score wins.

Num. of Parameters	Model	Borda Score
2	Ortega	42595
3	GB2	41760
5	Wang	38861
2	Dagum	38806
2	Singh-Maddala	36297
3	Abdalla-Hassan	35153
4	Sarabia	32208
2	Rasche	32109
1	Lognormal	24830
2	Generalized Gamma	23084
3	GB1	22779
1	Gamma	13931
1	Weibull	11420
1	Pareto	9594
1	Rhode	7310
1	Chotikapanich	3909
1	Kakwani-Podder	970

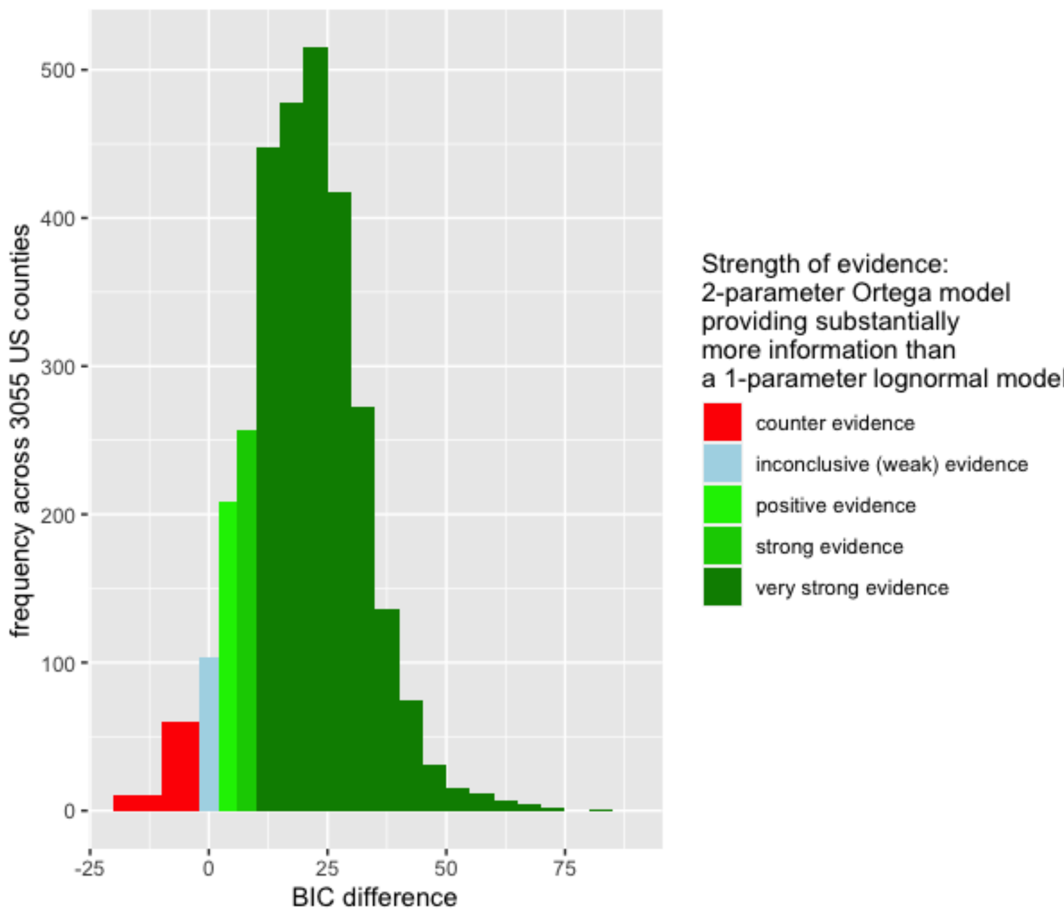
7. Analysis of BIC differences

In order to rule out that the choice of information criterion (AIC_c) influenced the results of our analysis, we reran the Δ analysis while using the Bayesian information criterion (BIC). The differences in BIC are defined in analogy to the AIC_c differences (Δ) as

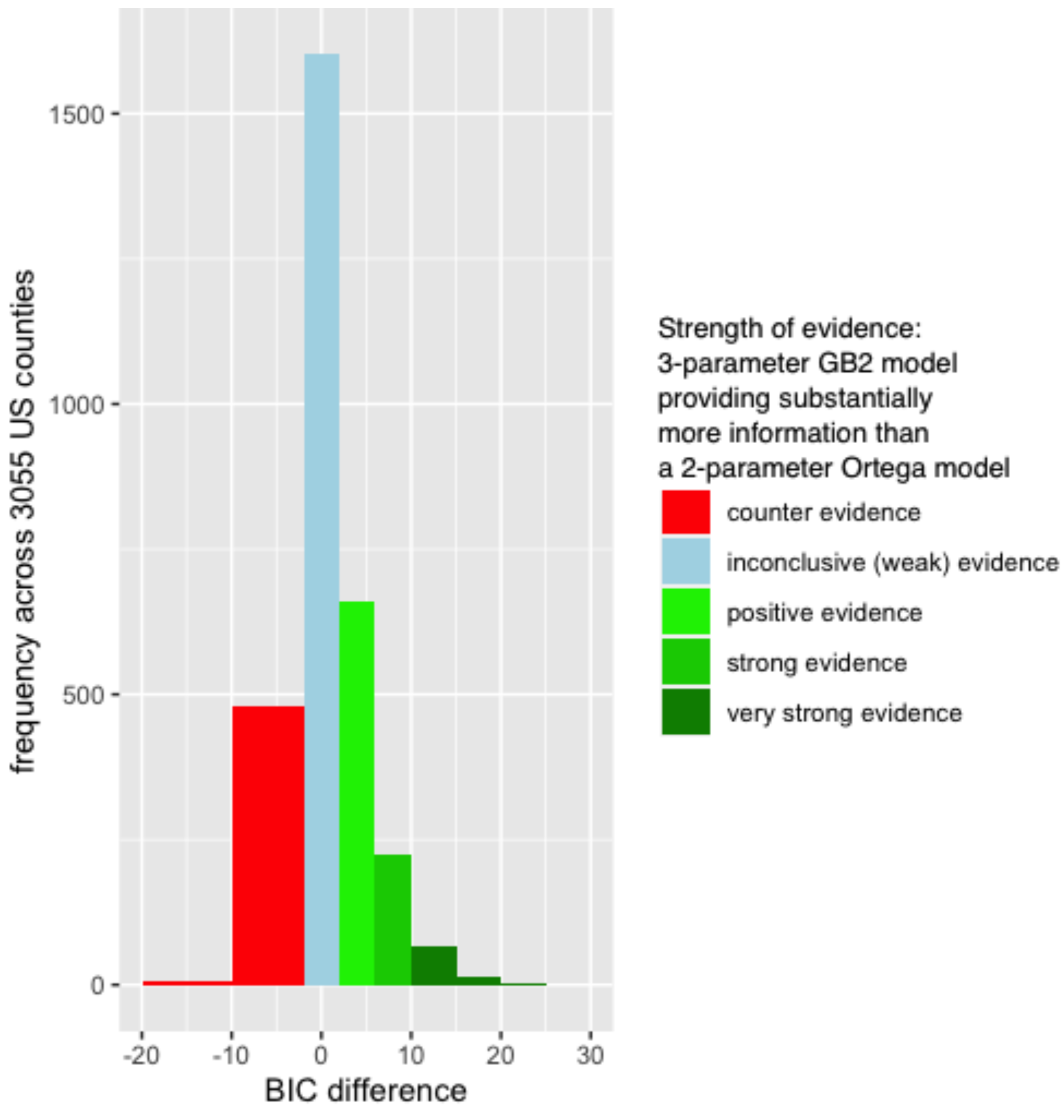
$$\text{BIC difference} = BIC_i - BIC_j \quad [1]$$

For BIC, the analysis of differences is also applied in the literature, yet with a slightly differing usage of wording and boundaries. While the interpretation of the differences is the same for both differences in AIC_c and BIC (namely, the larger the difference between the values, the less support there is for the competing model's ability to provide as good an approximation of the data as the other one), the boundaries are shifted. (27) sets the boundaries of BIC differences as described in 7. Respecting those boundaries, we arrive at similar histograms as with the analysis of AIC_c differences; see Figures Supplementary Figure 9, Supplementary Figure 10, and Supplementary Figure 11. Hence, we conclude that the superiority of Ortega compared with single-parameter models is irrespective of the chosen information criterion.

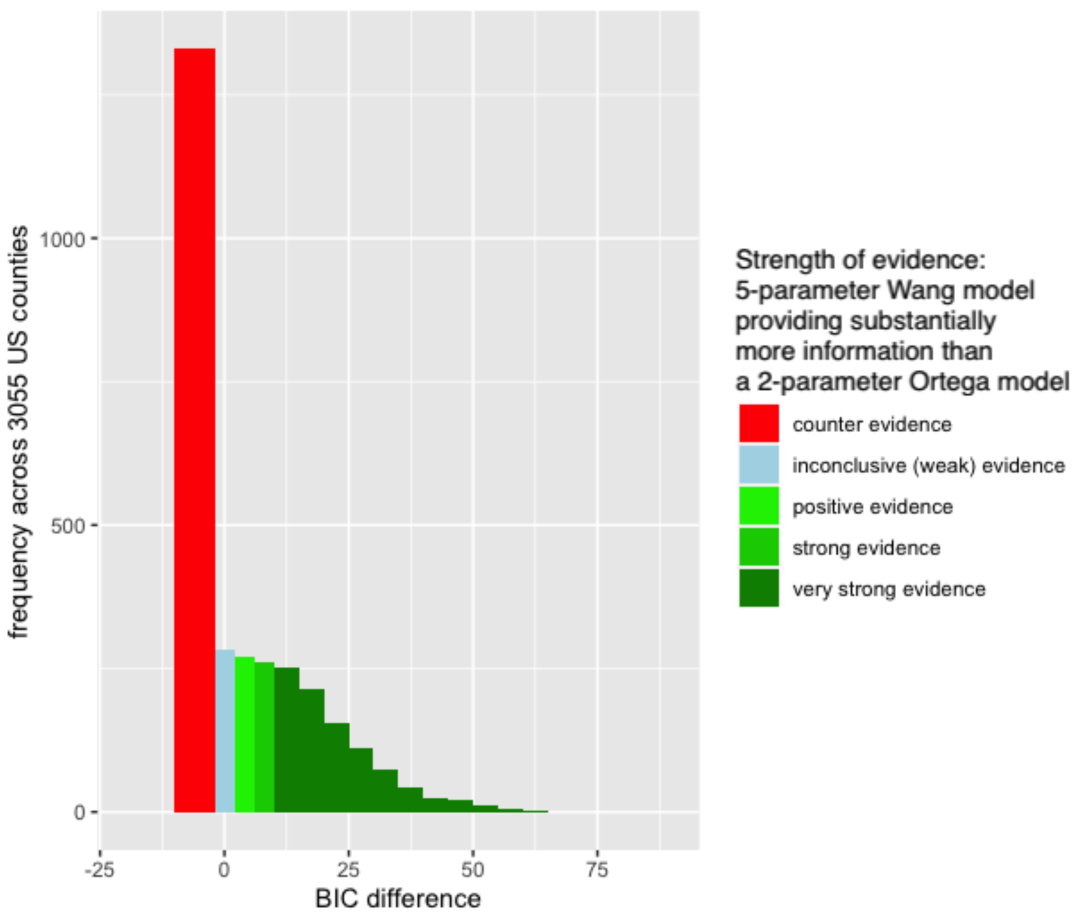
BIC difference	Evidence
0-2	weak
2-6	positive
6-10	strong
>10	very strong



Supplementary Figure 9. Histogram of BIC differences between the one-parameter lognormal model i and the two-parameter Ortega and j.



Supplementary Figure 10. Histogram of BIC differences between the three-parameter GB2 model i and the two-parameter Ortega and j.



Supplementary Figure 11. Histogram of BIC differences between the five-parameter Wang model i and the two-parameter Ortega and j.

8. Δ-AIC analysis of Ortega vs. GB2 and Ortega vs. Wang model

Using the Borda count voting procedure, we have determined the two-parameter Ortega Lorenz curve to be the winning model. However, the GB2 model using three parameters tightly comes second in the Borda count, and the Wang five-parameter model also performs well and wins the majority voting procedure. So do the three- and five-parameter models potentially provide substantially more information for some counties than a two-parameter model? To investigate this question, we calculated the AIC_c differences between Ortega and GB2 as well as Ortega and the Wang model.

We draw on prior literature, namely the guidelines given by Burnham and Anderson (28), to set up an evaluation strategy tied to the specific problem at hand of investigating the extent to which a certain model fits the data better than other models.

Burnham and Anderson (28) acknowledge that an interpretation of absolute AIC values, and hence a comparison between competing models, is hindered because of arbitrary constants. Instead, (28) propose using differences in AIC values, $\Delta_i = AIC_i - AIC_{min}$, that represent the information loss experienced when using model i rather than the best model which exhibits the minimum AIC value AIC_{min}. The severity of information loss can be characterized by defining intervals for Δ_i values, with larger values representing a higher amount of information loss. Burnham and Anderson (28) provide some rules of thumb: Models i with $\Delta_{i,j} \leq 2$ have substantial support; for $4 \leq \Delta_{i,j} \leq 7$ considerably less support and for $\Delta_{i,j} > 10$ no support for being the best approximating model in the candidate set. In other words, the higher the Δ_i value, the less support there is for the hypothesis that the two models of comparison provide an equally well characterization of the empirical data. This information can then be used to evaluate the strength-of-evidence in favor of the minimum AIC model (28), i.e., to get a sense of whether the minimum AIC model is substantially better.

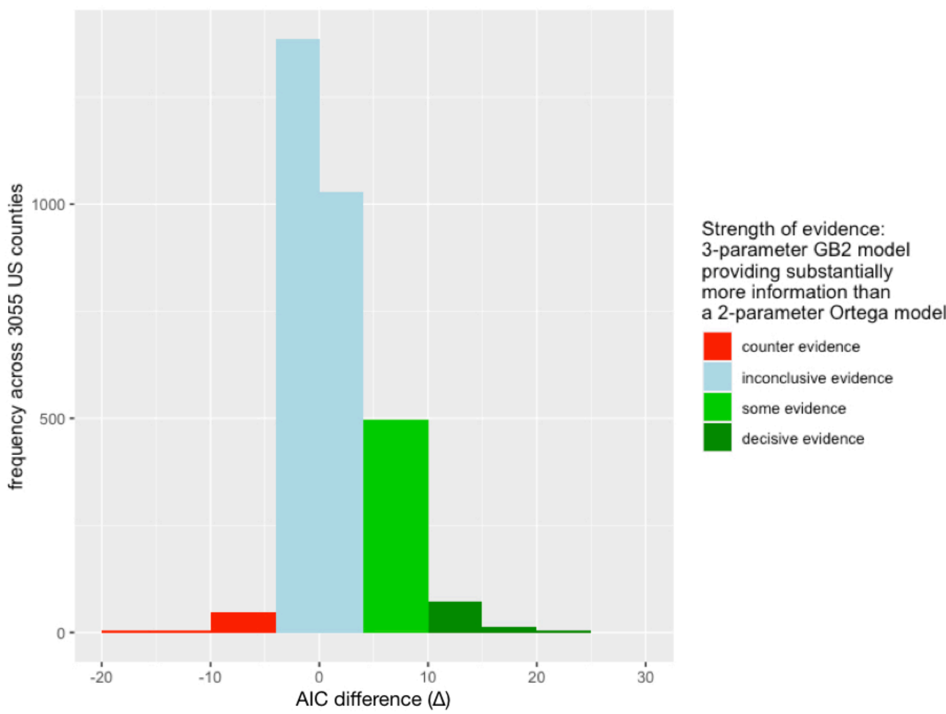
For the setting of a Lorenz curve comparison as outlined in this paper, we generalize the evaluation strategy of (28) and fine-tune the interpretation in order to provide a more intuitive understanding. First, let us note that we work with the small-sample bias corrected version of AIC values (AIC_c values), which does not affect the evaluation strategy, but changes the name of the strategy to evaluating AIC_c differences instead of AIC differences. Second, we do not necessarily compare the model of interest to the minimum AIC_c model in the respective US county, but fixed models, e.g., Ortega versus lognormal model. Hence, instead of $\Delta_i = AIC_i - AIC_{min}$, we introduce a more general version $\Delta_{i,j} := AIC_i - AIC_j$. To enhance ease of interpretation, we do not take on the perspective of (28) that focus on characterising the support of various models in being the best approximation of the data, but propose a slightly different perspective on the values: Starting off with the interpretation of (28) that $\Delta_{i,j}$ represents the information loss experienced when using model i rather than model j, we frame the $\Delta_{i,j}$ values directly as strength-of-evidence in favor of model j. This means that higher values of $\Delta_{i,j}$ provide evidence in favor of model j capturing the information given by the empirical data more aptly. With this general setup of $\Delta_{i,j}$ values, we might now encounter the situation of negative values in AIC_c differences, which is not possible with the AIC difference values defined in (28) as they set model j to the model with minimum AIC value. However, negative values of AIC_c values simply correspond to the case where i and j are reversed, hence gathering evidence for model i or, in other words, evidence for counter model j. Finally, we are forced to redefine the value intervals: (28) leave out interpretation guidelines for Δ_i in the intervals $[2, 4]$ and $[7, 10]$, and we therefore extend their intervals in a conservative manner.

In summary, our strength-of-evidence classification in terms of AIC_c differences is as follows: We find inconclusive evidence on whether model j, e.g., the Ortega model, is superior in modeling relevant information compared to model i, e.g., the lognormal model, if the AIC_c difference $\Delta_{i,j}$ is $\in$ [-4,4], some evidence that model j is superior if $\Delta_{i,j} \in$ [4,10] and decisive evidence that model j is superior to model i if $\Delta_{i,j} > 10$. If $\Delta_{i,j} \in [-4, -10]$, we find some evidence *against* model j's superiority, and decisive evidence *against* model j's superiority for $\Delta_{i,j} < -10$. With histograms of AIC_c differences ($\Delta_{i,j}$), we can see how often, i.e., in how many US counties, we find supporting evidence for whether one model indeed provides more substantial information about the data.

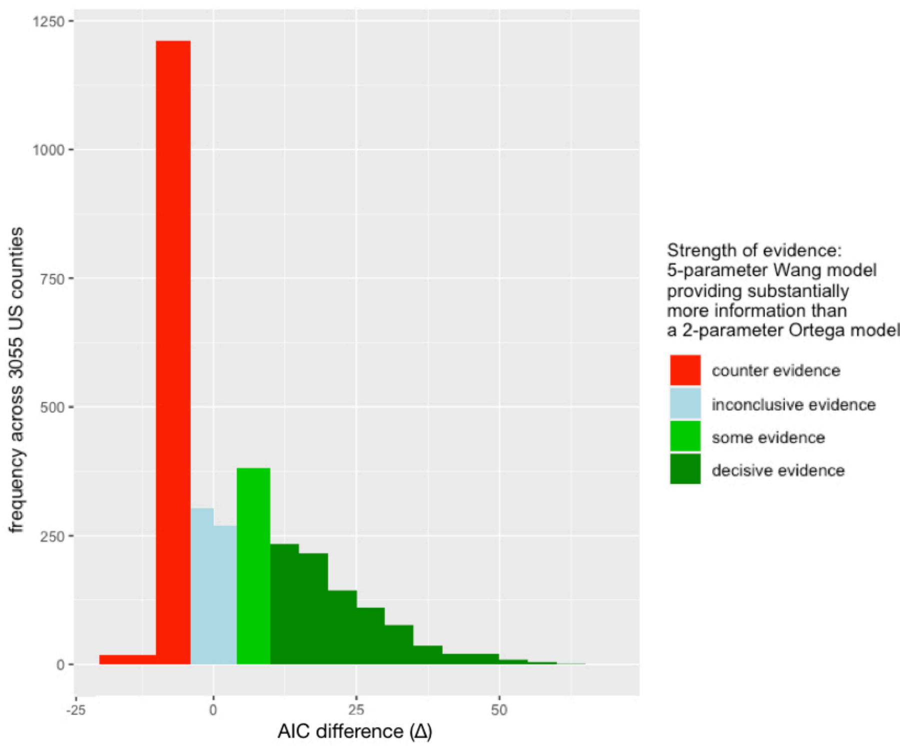
As a recap, for the comparison between an Ortega two-parameter model and the single-parameter lognormal model, we find a clear picture in support of the two-parameter model; see Figure 2 in the main text.

Now evaluating Ortega versus GB2, we see a much more inconclusive picture; see Supplementary Figure 12. For most of the counties, there is inconclusive evidence; i.e., there is substantial support that both models perform similarly well in modeling the information given in the empirical data. This indicates that the three- and two-parameter models are somewhat comparable. Given this information, it is debatable which model to prefer, but as Ortega is the simpler model, we clearly favor it over GB2.

In comparing the Ortega model and the five-parameter Wang model, we get a more distinct histogram; see Supplementary Figure 13. On the one hand, we can clearly see that for many counties, we have evidence that the five-parameter model captures relevant information better than the two-parameter Ortega model. On the other hand, we find counter-evidence in many counties as well: i.e., that the two-parameter model performs that task better. This result is unsurprising given which aspects the various voting procedures emphasize: the Borda count values good performance across counties (Ortega won), whereas majority voting honors how often a model performs best in a county (Wang won). That is, the Wang five-parameter model is excellent many times but also inferior many times compared with the two-parameter Ortega model. As we seek a model that performs well across all US counties, we prefer Ortega for that purpose.



Supplementary Figure 12. Histogram of AIC$_c$ differences ($\Delta_{i,j}$) between the three-parameter GB2 model i and the two-parameter Ortega and j.



Supplementary Figure 13. Histogram of AIC$_c$ differences ($\Delta_{i,j}$) between the five-parameter Wang model i and the two-parameter Ortega and j.

9. Nonlinear Least Squares (NLS)

In terms of Lorenz curves, we are dealing with functions that are nonlinear in their parameters, which is why we call the framework in this case nonlinear least squares (NLS). The NLS approach is a widely used method for estimating the parameters of functional forms of the Lorenz curve, e.g., in (8, 15, 29–32). The objective we are trying to minimize is the sum of squared residuals. We recognize the estimation task as

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{K} \left(L(u_i, \boldsymbol{\theta}) - \eta_i\right)^2 \qquad [2]$$

where $\boldsymbol{\theta}$ is the parameter vector of the Lorenz curve model and η_i the cumulative empirical income share observed for the cumulative population share u_i.

Using the NLS procedure, we get consistent estimates. However, they are not efficient, as least squares estimation in the Lorenz curve setting exhibits auto-correlated and heteroskedastic residuals (5, 8). Krause (2014) used the approach of minimizing the MSE in their recent study and mentions that other procedures to gain efficiency, e.g., proposed by (10), hardly change results given their setting.

A main disadvantage of NLS stems from ignoring the proportional nature of the data (33) and "overlook[ing] the fact that the sum of the income shares is, by definition, equal to one" (8, p. 11). Both features of the data are neglected by NLS and hence fruitful opportunities in using this special structure of the data are missed.

Apart from that, the NLS estimation method is still widely used for estimating Lorenz curves and does not provide efficient, but more importantly, consistent, estimates.

NLS estimates for each county are provided for the present study and will be evaluated as a robustness check.

10. Comparison of MLE and NLS Estimates

We explore whether potential estimation method artifacts account for our results by comparing the estimated parameters for the 17 Lorenz curve models using both NLS and MLE. We find similar point estimates for most model parameters. The median relative difference between the MLE and NLS estimates across counties is depicted in Supplementary Table 8 below. An exception is the GB1 model, for which differences were large: for the generalized gamma and GB1 Lorenz curve model, the differences between NLS and MLE estimates were large, e.g., 84.1659 for the second GB1 parameter. This observation is not surprising, as those two Lorenz curve models exhibited severe estimation instabilities, which we take as indicating their unsuitability as a basis for deriving inequality measures. For this reason, we classify the GB1 model as unsuitable and exclude it from further analysis.

The remaining models exhibit small relative differences between both estimation methods. For example, the median relative difference between MLE and NLS point estimates of the Ortega parameters was 0.0234 for Ortega parameter α and 0.0165 for Ortega parameter β. Hence, we have no reason to believe that the estimation technique has a systematic influence on the model parameters estimated.

We refer to the relative difference as given by

$$\text{relative difference} = \frac{|\hat{\theta}_{MLE} - \hat{\theta}_{NLS}|}{|\hat{\theta}_{NLS}|}$$

The median of the relative difference of parameter estimates across all $N = 3\,056$ US counties included in our study is given in Supplementary Table 8.

Supplementary Table 8. Median relative difference between MLE and NLS estimates across all counties.

Model	Param. 1	Param. 2	Param. 3	Param. 4	Param. 5
Abdalla-Hassan	0.0315	0.9410	0.0168	-	-
Chotikapanich	0.1498	-	-	-	-
Dagum	0.0605	0.0186	-	-	-
Gamma	0.2135	-	-	-	-
GB1	83.3003	84.1659	0.8906	-	-
GB2	0.2329	0.1884	0.1548	-	-
Generalized Gamma	80.8858	0.8900	-	-	-
Kakwani-Podder	0.2040	-	-	-	-
Lognormal	0.0165	-	-	-	-
Ortega	0.0234	0.0165	-	-	-
Pareto	0.0751	-	-	-	-
Rasche	0.0218	0.0145	-	-	-
Rhode	0.0250	-	-	-	-
Sarabia	0.3857	0.0292	0.0890	0.0680	-
Singh-Maddala	0.0718	0.0342	-	-	-
Wang	0.2122	0.1616	0.0894	0.4698	0.9285
Weibull	0.0810	-	-	-	-

11. Relationship between Ortega parameters and Pareto index

Sarabia et al. (1999) (34) introduced a general method to build ordered families of Lorenz curves, noting that one of the Pareto Lorenz curve families coincides with the Ortega Lorenz curve. We draw on this work in advancing the correspondence between the Pareto distribution parameter and one of the Ortega parameters.

To derive the relationship between Ortega parameter β and the Pareto index, let us first introduce some definitions. The Ortega Lorenz curve is given by (12):

$$L_{Ortega}(u) = u^{\alpha} \cdot (1 - (1-u)^{\beta}) \qquad [3]$$

where $\alpha \leq 0, 0 < \beta \leq 1$.

The cumulative distribution function of the classical Pareto distribution is given by

$$F(x) = 1 - \left(\frac{\sigma}{x}\right)^{a} \qquad [4]$$

where $\sigma, a > 0$. Following this notation, we can recognize σ as a scale parameter and a as a shape parameter. The Pareto index equals the shape parameter of the classical Pareto distribution (e.g., used in (35)). Being consistent with our notation, we can therefore define

$$\text{Pareto index } := a \qquad [5]$$

To show that there is a relationship between β and a, it is useful to calculate the Lorenz curve for the classical Pareto distribution first. The general definition of a Lorenz curve is given by (36):

$$L(u) = \mu^{-1} \int_0^u F^{-1}(t)dt \qquad [6]$$

where μ is the finite mean and $F^{-1}(t)$ the inverse of the cumulative distribution function. For the classical Pareto case with $\mu = \frac{a\sigma}{a-1}$ and $F^{-1}(t) = \sigma(1-t)^{-\frac{1}{a}}$, we get

$$\begin{aligned} L_{Pareto}(u) &= \frac{a-1}{a\sigma} \int_0^u \sigma(1-t)^{-\frac{1}{a}} dt && [7] \\ &= \frac{a-1}{a\sigma} \left[\frac{-\sigma}{1-\frac{1}{a}} \cdot (1-t)^{1-\frac{1}{a}} \right]_0^u && [8] \\ &= \frac{a-1}{a\sigma} \left[\left(\frac{-\sigma}{1-\frac{1}{a}} \cdot (1-u)^{1-\frac{1}{a}} \right) - \left(\frac{-\sigma}{1-\frac{1}{a}} \right) \right] && [9] \\ &= \left(1 - \frac{1}{a}\right) \cdot \left[\frac{-1}{1-\frac{1}{a}} (1-u)^{1-\frac{1}{a}} + \frac{1}{1-\frac{1}{a}} \right] && [10] \\ &= 1 - (1-u)^{1-\frac{1}{a}} && [11] \end{aligned}$$

We can see that the Pareto Lorenz curve depends on the Pareto index a only. If we are able to relate the Pareto Lorenz curve to the Ortega Lorenz curve and demonstrate that the Pareto index is linked to one of the two Pareto parameters only, we know that we can transform that parameter into the Pareto index. (34) actually introduced a family of Lorenz curves that helps explain the relationship between the Pareto Lorenz curve and the Ortega Lorenz curve. In detail, their second theorem states:

Theorem 2 ((34)) *Let $L(p)$ be a Lorenz curve and consider the transformation $L_\alpha(p) = p^\alpha \cdot L(p)$, where $\alpha \leq 0$. Then, if $\alpha \geq 1$, $L_\alpha(p)$ is a Lorenz curve too. In addition, if $0 \leq \alpha < 1$ and $L'''(p) \geq 0$, $L_\alpha(p)$ is also a Lorenz curve.*

(34) further show that the condition $L'''(p)$ is satisfied for the Pareto Lorenz curve such that for $\alpha \geq 0$, we can transform the Pareto Lorenz curve using theorem 2, which yields

$$\begin{aligned} L_\alpha(u) &= u^\alpha \cdot L_{Pareto}(u) && [12] \\ &= u^\alpha \cdot \left(1 - (1-u)^{1-\frac{1}{a}}\right) && [13] \end{aligned}$$

Now looking at the Ortega Lorenz curve as defined in 3, we can clearly see that the Ortega Lorenz curve is nothing other than the Pareto Lorenz curve, extended by a newly introduced parameter α through the use of theorem 2 and a redefined parameter

$$\beta := 1 - \frac{1}{a} \qquad [14]$$

In other words, we can see the Ortega Lorenz curve as an extension to the Pareto Lorenz curve. Having established this close link between the two Lorenz curves, we can think of Ortega parameter β as being in close relation to the Pareto index a, using the relationship defined in 14. If the true income distribution were to follow a Pareto distribution, Ortega parameter α would be zero and the Ortega parameter β would be an exact monotonic transformation of the Pareto index. However, in cases where the true income distribution was not generated by a Pareto distribution, of course, the additional estimation of Ortega parameter α might capture aspects that are also correlated to β, such that the exact monotonic transformation given in 14 is rather an approximate relationship, depending on the data. Although this is a weaker statement, it is still useful for our purpose: we want to know which aspects of the income distribution the Ortega parameters capture. We know that the lower the Pareto index, the larger the proportion of very-high-income people. And we derived above that the higher the Pareto index associated with the income distribution, the higher the Ortega parameter β. Having demonstrated the close relationship between β and the Pareto index a in the above section, we see this as evidence of a capturing the occurrence of very top incomes. We therefore conclude that Ortega β has the following interpretation: the lower the Pareto index, the larger the proportion of very-high-income people. We therefore propose it as a measure of top-concentrated income inequality.

12. Interpreting the Ortega Lorenz curve

Visual inspection of Ortega parameters. To visually inspect how a change in parameters affects the Ortega Lorenz curve, we simulate Ortega Lorenz curves while varying α and γ. The R code *simulation_ortega_lorenz_curves.R* replicates this simulation and is available in the GitHub repository we provide for this paper (see www.measuringinequality.com). In detail, first we plot the Ortega Lorenz curves varying α between 0.01 and 1.5 while keeping γ fixed at 0.5 (for α, the side constraint is ≥ 0; the upper limit 1.5 is chosen as an extension to the empirical values that valued 1.23 at max). In our empirical estimation of US county-level Ortega Lorenz curves, for α a typical value was 0.5 and for γ 0.5, which is why we fix the respective values at that level. Then, we plot Ortega Lorenz curves with $\alpha = 0.5$ and vary γ between 0.01 and 0.99 (side constraint $0 \leq \gamma < 1$).

Our simulation results are generally in line with prior theory, i.e., that Ortega parameter γ is associated with top-concentrated inequality. The asymmetry line in Figure 3 in the main text of the paper, Panel B, facilitates comprehension whereby we observe a disproportionate change in the Lorenz curve on the right side (i.e., at higher incomes). Note that we observe top-concentrated inequality arising when there is a step increase in the Lorenz curve shortly ahead of the cumulative share of population reaching 100%. Further, our observations are in accordance with the direction of change we expected through the relationship between γ and the Pareto index, i.e., a higher value of γ indicating a higher level of top-concentrated inequality.

In sum, our simulation study suggests that α is a reflection of bottom0concentrated inequality whereas γ is a reflection of top-concentrated inequality.

When varying α while keeping γ a fixed constant, we can see that an increase in α stretches the left side of the Lorenz curve toward the x-axis (i.e., at lower incomes). The higher α, the more this is the case, as seen in Figure 3A in the main text. This effect can again be acknowledged when adding the asymmetry line to the plot, which helps in identifying the disproportionate change in the curves. With a more intense change on the left side, one can conclude that α captures specificities on the left tail of the income distribution.[‖] Therefore, we conclude that α is a measure of bottom-concentrated inequality.

Determining the relationship between Ortega parameters and other measures of inequality. To further investigate the interpretation of the Ortega parameters, we relate them to income ratios, as they are more intuitive and used in some prior research to measure inequality. First, we explore the dependency between Ortega parameters and common percentile measures (95/50 and 50/10 ratios). Then, we move on to evaluate which percentile ratios might reflect the information captured by the Ortega parameters more precisely.

A common measure of top-concentrated income inequality is the fraction of income held by the 95^{th} percentile divided by the median income share (also known as a 95/50 ratio), whereas bottom-concentrated income is often measured using a 50/10 ratio; see (37–39). We have argued that Ortega parameter γ is related to top-concentrated inequality and should increase with higher levels of inequality. The 95/50 ratio also aims at capturing the phenomenon of top-concentrated income inequality, which is why we suspect the quantities to be highly positively correlated. We also hypothesized that Ortega parameter α is related to bottom-concentrated income inequality and should increase with higher levels of inequality. Another measure that aims at capturing bottom inequality is the 50/10 ratio, i.e., the income share held by the lower 50% of the population divided by the income share held by the lower 10% of the population. We suspect that both quantities, i.e., α and the 50/10 ratio, should be highly positively correlated because they should measure the same underlying phenomenon (bottom-concentrated inequality).

To test whether our suggested correlational dependencies hold true, we first simulate Ortega Lorenz curves with varying parameters α and γ, then calculate the income percentile ratios 95/50 and 50/10 for those Lorenz curves, and consequently analyze the correlation between Ortega and percentile ratio quantities. In detail, we simulate a total of 10 000 Ortega Lorenz curves with varying parameter values. We vary α from 0.01 to 1 with a step size of 0.01 and γ from 0 to 0.99 with the same step size of 0.01. Subsequently, we calculate partial correlations between the quantities. Doing so, we control for all other variables included in this analysis; i.e., we correlate α with the 50/10 ratio controlling for γ and the 95/50 ratio.

Our results, depicted in Supplementary Table 9, show that α indeed highly correlates with the bottom-concentration ratio 50/10 while γ highly correlates with the top-concentration ratio 95/50. However, it is worth pointing out that this correlational dependency only becomes apparent when focusing on the full parameter space of γ ($0 \leq \gamma < 1$) while limiting the parameter space of α for the same range as γ. For the empirical US county-level Lorenz curves, we encountered a parameter range of 0.12 to 1.23 for α and 0.3 to 0.93 for beta. In this range of parameters, the correlation between α and the 50/10 ratio, and γ and the 95/50 ratio, gets distorted, which indicates high sensitivity of the correlational structure regarding the parameter range.

This gives us reason to believe that those ratios might not reflect the type of top- and bottom-concentrated inequality that is measured by the Ortega parameters. Revising Figure 3 in the main text, we can see γ affecting rather the very top of the distribution. Exploring the dependency structure percentile ratios and the Ortega parameters, it indeed becomes clear that

[‖] A high level of bottom-concentrated inequality can be recognized from the Lorenz curve if the curve is rather flat near the bottom percentiles but exhibits a sharp increase before reaching the median population.

Ortega γ is instead measuring inequality in the very top percentiles and that α captures a broader range of the distribution. We find the correlational dependency between the 99/90 ratio with γ and 90/10 ratio with α very robust to the parameter range. Also, the strength of correlational dependency is more distinct; see Supplementary Table 10, which depicts the correlations within the same parameter range used for Lorenz curve generation as in Supplementary Table 9.

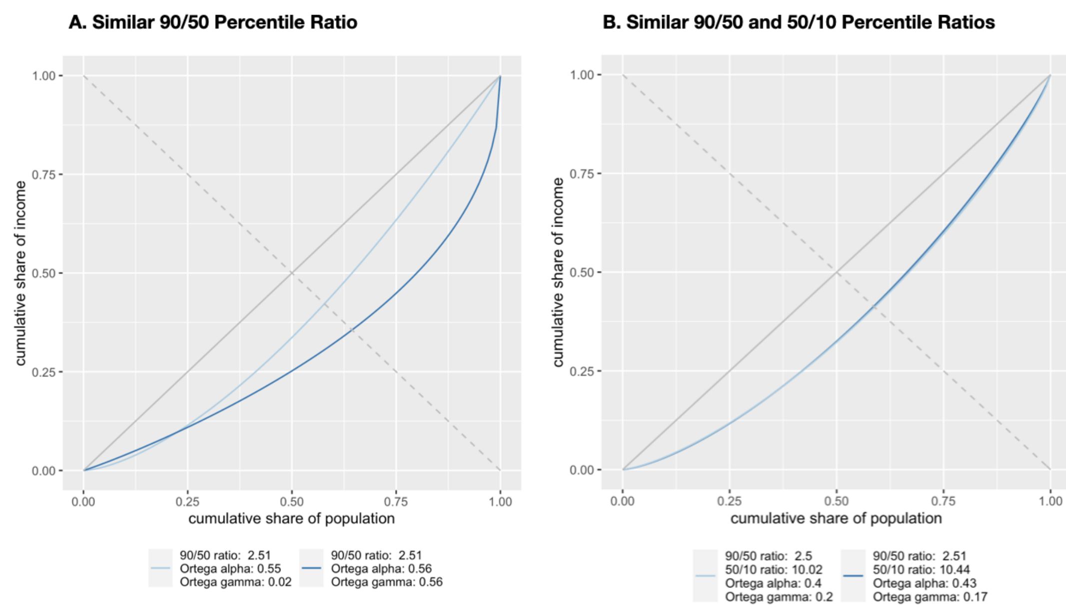
We therefore conclude that our suggested interpretation of the Ortega parameters should not directly be linked to current measures of top- and bottom-concentrated inequality, i.e., the 95/50 and 50/10 ratios, but to measures of inequality at the very top (99/90 ratio) and most of the remainder of the distribution (90/10 ratio).

Supplementary Table 9. Partial correlations between Ortega parameters and percentile ratios, controlling for all other quantities; e.g., the partial correlation between γ and the 95/50 ratio is 0.940 after controlling for α and the 50/10 ratio.

	50/10 ratio	95/50 ratio
Ortega α	0.786	0.137
Ortega γ	-0.259	0.940

Supplementary Table 10. Partial correlations between Ortega parameters and percentile ratios, controlling for all other quantities; e.g., the partial correlation between γ and the 99/90 ratio is 0.9088 after controlling for α and the 90/10 ratio.

	90/10 ratio	99/90 ratio
Ortega α	0.9081	-0.0408
Ortega γ	-0.0620	0.9088



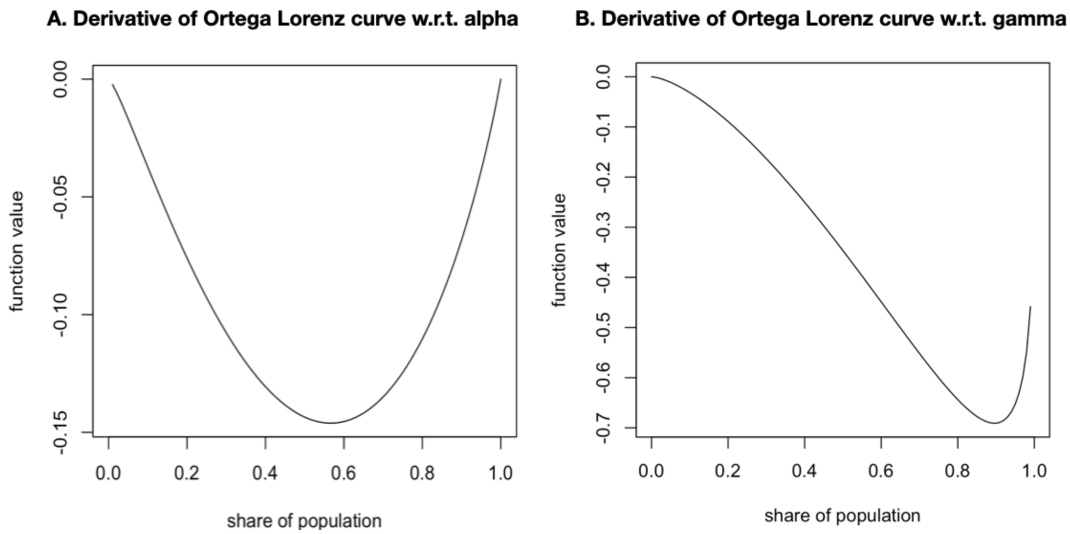
Supplementary Figure 14. Panel A illustrates two very different Lorenz curves exhibiting the same 90/50 percentile ratio. In Panel B we can notice that when fixing both the 90/50 *and* the 50/10 percentile ratios into a similar range, the resulting Lorenz curves must have a similar shape. This indicates that (at least) two parameters should be provided to limit the potential volatility of the resulting Lorenz curves.

Analytical investigation of the Ortega Lorenz curve: Derivatives. A natural way to investigate how a function is affected by its parameters is to inspect the (partial) derivatives. For the Ortega Lorenz curve, the partial derivatives with respect to α and γ are

$$\frac{\delta}{\delta\alpha}\left(u^{\alpha}(1-(1-u)^{1-\gamma})\right) = (u^{\alpha}(1-(1-u)^{1-\gamma})\log(u) \quad [15]$$

$$\frac{\delta}{\delta\gamma}\left(u^{\alpha}(1-(1-u)^{1-\gamma})\right) = (u^{\alpha}(1-u)^{1-\gamma}\log(1-u) \quad [16]$$

From this it is not immediately obvious how the Ortega Lorenz curve is affected by the parameters. However, we can note that both derivatives are ≤ 0 within the allowed parameter space. What we are especially interested in is whether the interpretation of the parameters suggested by the simulation study (α more intensely emphasizing bottom-concentrated inequality and γ highlighting top-concentrated inequality) can be seen analytically as well. To test this, we take a closer look at the rate of change, i.e., the partial derivatives, at certain regions along the x-axis. In other words, if the Lorenz curve function is more intensely affected by a parameter in a certain region of the population, we could conclude that this parameter is more sensitive to this area of the population: e.g., the top or bottom. Supplementary Figure 15 visualizes the derivatives of the Ortega Lorenz curve with respect to α and γ along the x-axis (i.e., cumulative share of population) while keeping the parameters themselves fixed at $\alpha = 0.5, \gamma = 0.5$, just as when simulating Ortega Lorenz curves in the above section. Note that we need to evaluate the absolute values of rate of change for the respective parameters, i.e., the absolute values of the partial derivatives. From Supplementary Figure 15, we can clearly see that a variation in α most intensely affects the Lorenz curve around the middle of the population (the absolute value of the derivative with respect to α is largest around the percentiles $\sim$ 0.45-0.65). In contrast, a variation in γ has the highest rate of change within the top percentile of the population (the absolute value of the derivative with respect to γ is largest around the top percentiles $\sim$ 0.80-0.95.

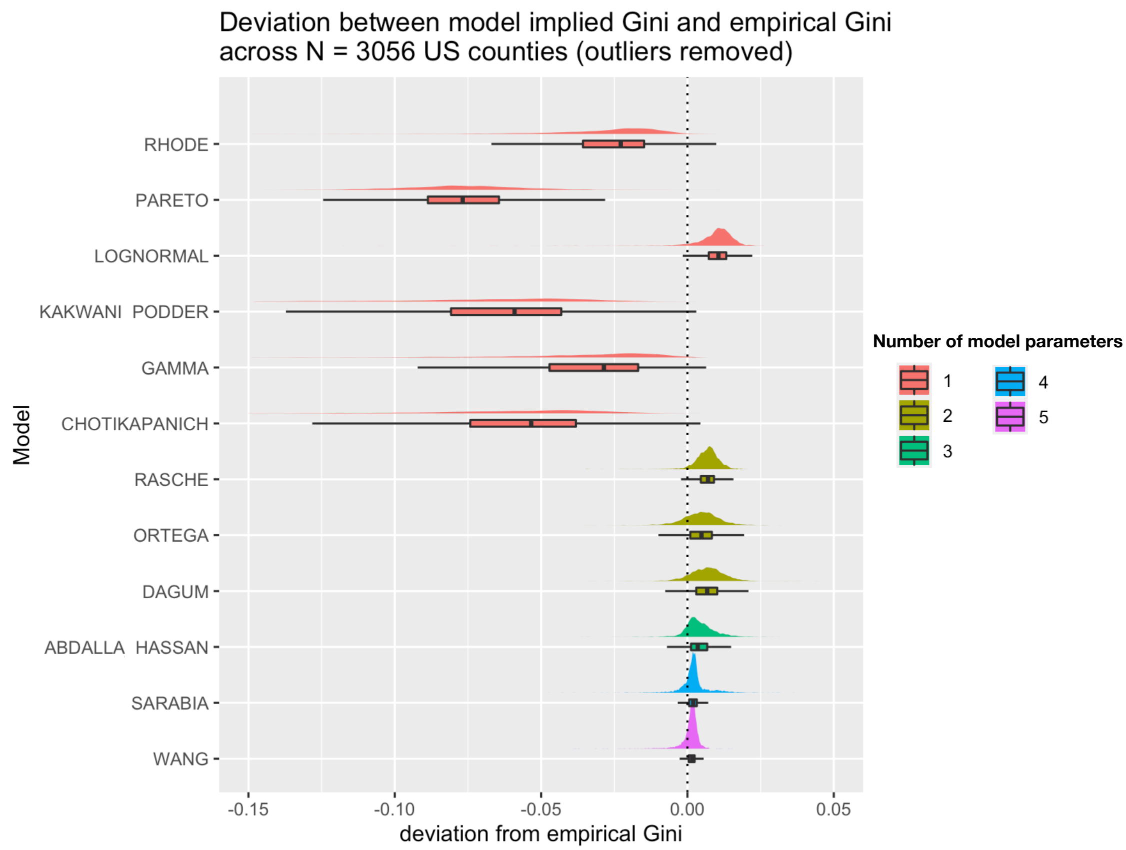


Supplementary Figure 15. Value of the derivatives of the Ortega Lorenz curve function $L(u) = u^{0.5}\dot{(}1 - (1-u)^{0.5}$ across the cumulative share of population.

13. Approximating the empirical Gini coefficient

To assess how well the different models approximate the main distributional statistics related to inequality, we compare the Gini coefficients implied by the model parameters with those Gini coefficients calculated nonparametrically on the US county data. The nonparametric Gini coefficients are calculated using the given data points of the empirical income distribution with linear interpolation, whereas the Gini coefficients implied by the models utilize integral calculus[**] for determining the area between the Lorenz curve and the line of perfect inequality.

These analyses, visualized in Supplementary Figure 16, reveal that when taking into account the number of parameters included in the model—ideally as few as possible—we can see that the Ortega model provides a reasonable trade-off between deviation from the nonparametric Gini and the number of parameters needed. Most notably, one-parameter models (red distributions in the figure) substantially deviate from the ideal average deviation of zero, while two-parameter models (brown) are a major improvement. Across the two-parameter models, the Ortega model is the one closest to the deviation of zero (dotted line) with a substantial number of data points (see boxplot touching the dotted line). While with more parameters (green, blue, and purple boxplots), precision further increases, the improvements are much smaller than those between one- and two-parameter models. This analysis demonstrates that using more than one parameter improves the approximation of empirical distributional statistics such as the Gini coefficient, and that further improvement in precision with more parameters is possible but is much smaller.



Supplementary Figure 16. Comparison across various parametric Lorenz curve models in approximating the empirical (nonparametric) Gini coefficient. Note that in order to prevent a masking effect of severe outliers, we omitted them in the plot. The boxes depict the $25^{th}, 50^{th}$ and 75^{th} percentiles of the deviations from the empirical Gini. The whiskers extend from the hinge to the smallest value at most (or largest value and no further, respectively) 1.5 times the inter-quartile range of the hinge. Minimum and maximum values as well as the center of the distributions are visualized by plotting the actual distribution of deviations above the boxes.

[**] For the Lorenz curve models based on the generalized beta distribution (GB1, GB2), we faced difficulties in calculating the integrals necessary for parametric Lorenz curve derivation, which is why these models are missing in our analysis.

14. Exploratory correlational study

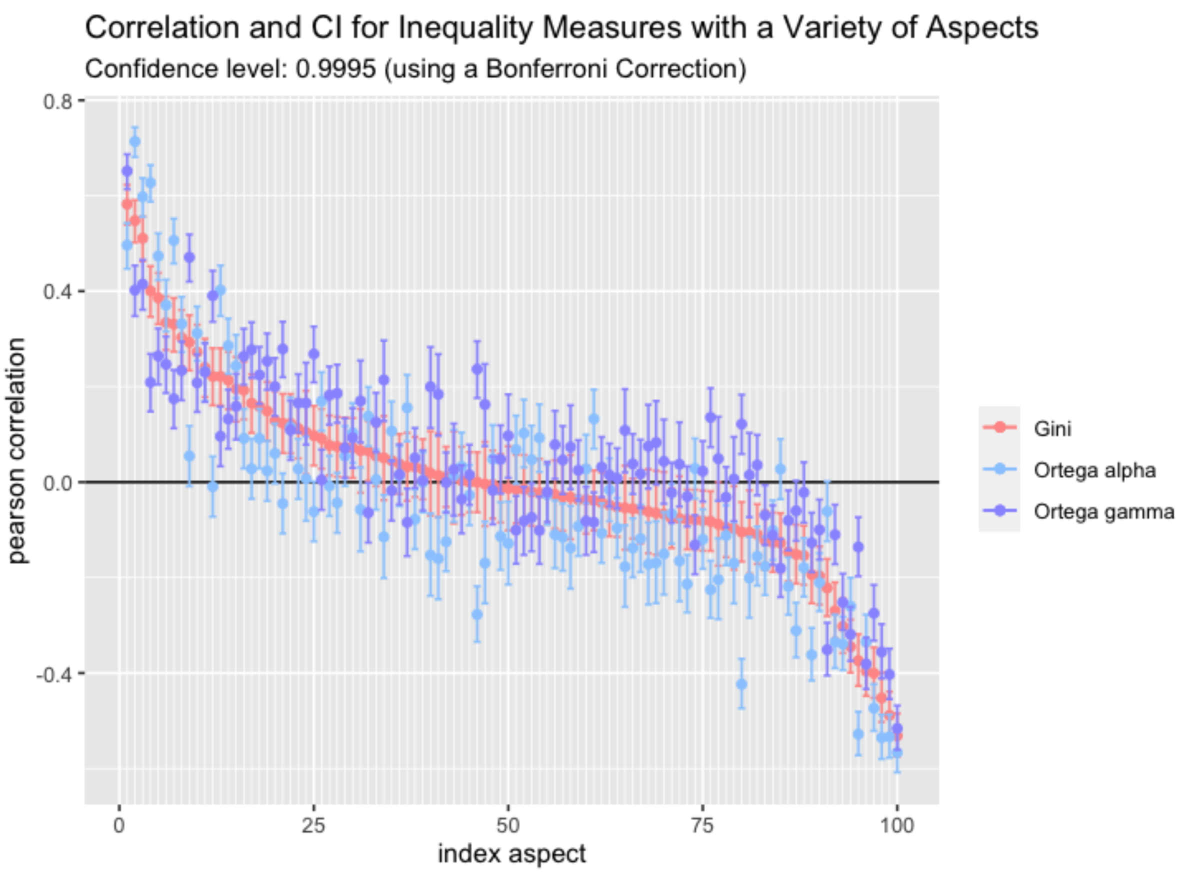
In our exploratory correlational study, for which we provide results below, we correlate 100 variables from policy-relevant fields to inequality measures. Our source of data is the ACS Survey 2011-2015, from which we pulled relevant source tables directly from https://data2.nhgis.org/main, and the data from (40) and (41) are publicly available at https://opportunityinsights.org. Code to replicate the study, as well as detailed information on the data used—i.e., a codebook—is available at www.measuringinequality.com.

We propose the use of both the Ortega parameters simultaneously (i.e., in a regression setting, researchers should include both Ortega parameters as independent variables within the regression model equation), which is why we calculate partial Pearson correlations between covariates and Ortega parameters. For the Gini coefficient, simple Pearson correlations are sufficient, as this is a single-parameter inequality measurement approach. We use the Gini coefficient provided by the ACS dataset. One might argue that we should have used the Gini index implied by the empirical Lorenz curves we used in the Ortega parameter estimation. However, the US Census Bureau, which conducts the ACS, has more fine-grained data (inaccessible to the public) available to calculate the Gini index for each county highly accurately, which makes their Gini indices more reliable.

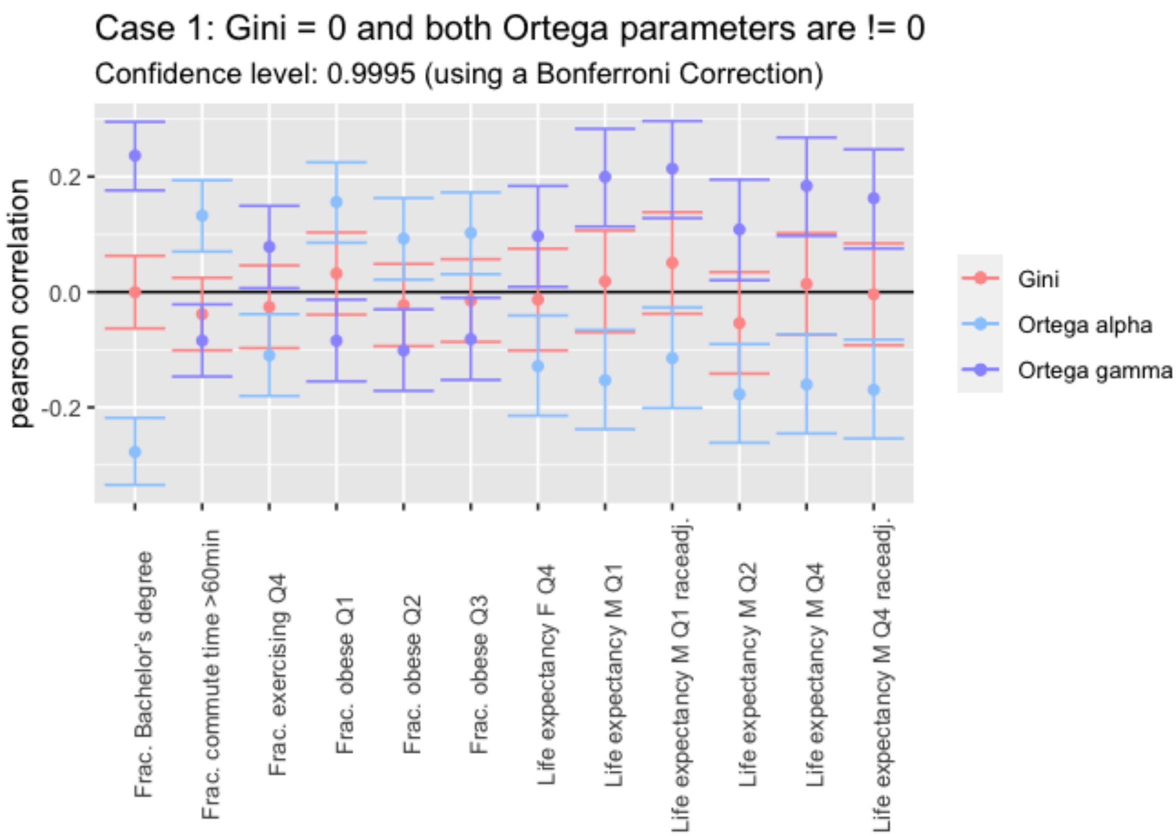
In Supplementary Table 11, we provide an overview of potential outcomes and the frequency of their occurrence across our analysis. Case ID 1 can be interpreted as Ortega's ability to disentangle (probably counteracting) effects related to inequality present in different parts of the income distribution, and case ID 2 might also shed light on a specific region of the income distribution being correlated to policy outcomes. For case ID 3, i.e., that neither Gini nor Ortega parameters show significant correlations, we have a coherent suggestion from both inequality measures that there is no association between inequality and the correlated variable. We also find coherent guidance on whether inequality is associated with a variable for case IDs 4 and 5. However, these cases show that use of the Ortega parameter might refine the insights we can obtain: while the Gini only reveals that there is an association between overall inequality and the variable, using the Ortega parameters, we can differentiate which part of the income distribution drives the significant correlation, including the magnitude. For case ID 6, i.e., that Gini is significant but none of the Ortega parameters are, the interpretation of such cases is rather puzzling. A potential interpretation is that in such cases, the association between inequality and the variable is driven by a feature of inequality that is captured through the Gini coefficient measuring overall inequality but is not explained by the concentration of income in different parts of the income distribution.

Supplementary Table 11. Cases occurring in our exploratory study correlating 100 covariates with the Gini index and calculating partial correlations between covariates and Ortega parameters.

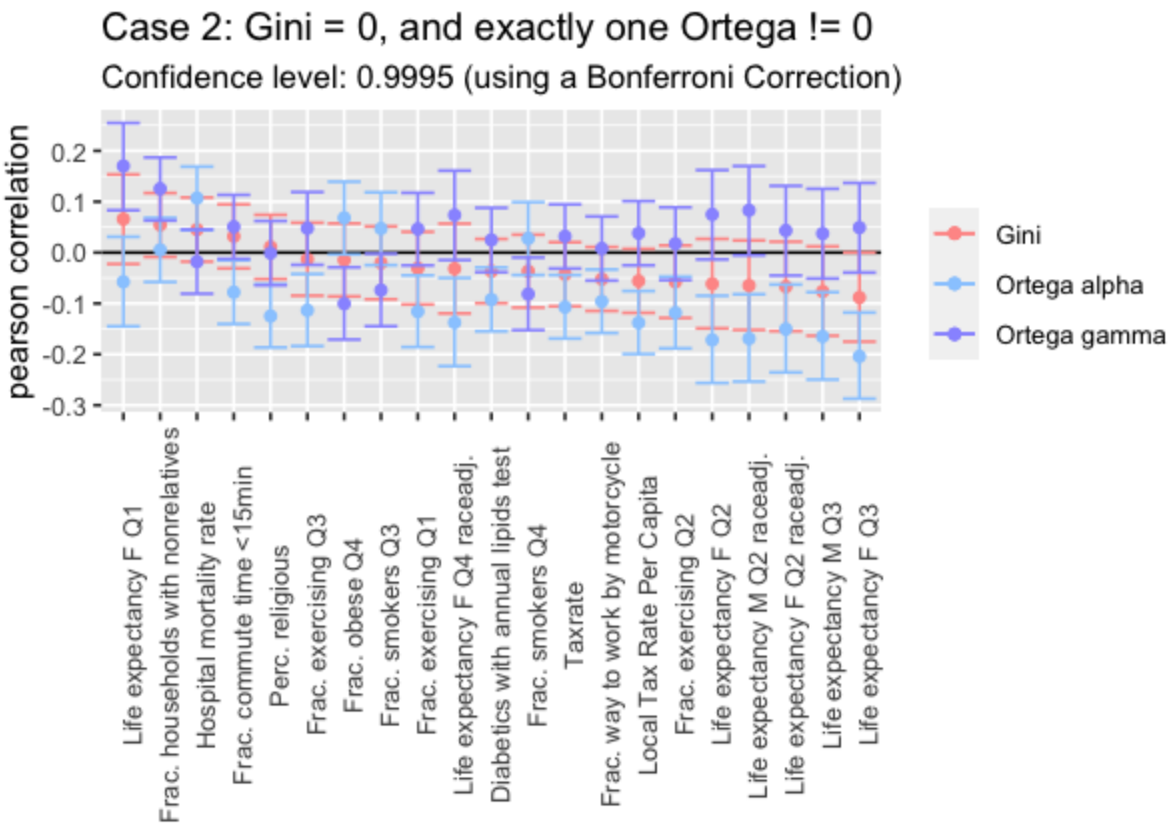
Case ID	Correlation with Gini coefficient $\neq 0$	Correlation with ... Ortega parameters $\neq 0$	Number of occurrences
1	no	2	12
2	no	1	21
3	no	0	8
4	yes	1	25
5	yes	2	34
6	yes	0	0
			100 = total number of covariates



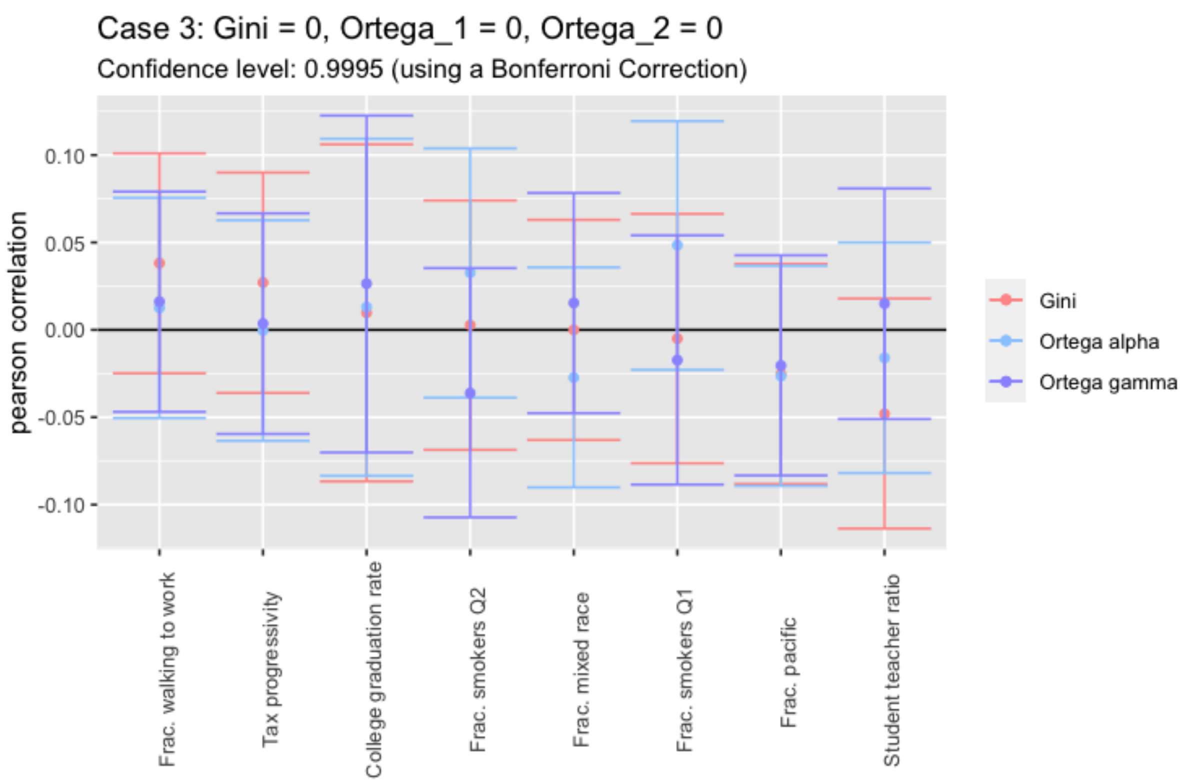
Supplementary Figure 17. Pearson correlations between inequality measures and county-level covariates. The plot shows Pearson correlations with the Gini index and partial Pearson correlations with the Ortega parameters, i.e., the correlation between one Ortega parameter and the covariate while controlling for the other Ortega parameter across $N = 3\,049$ US counties. Pearson correlation point estimates are visualized within confidence bounds of the Bonferroni corrected confidence interval.



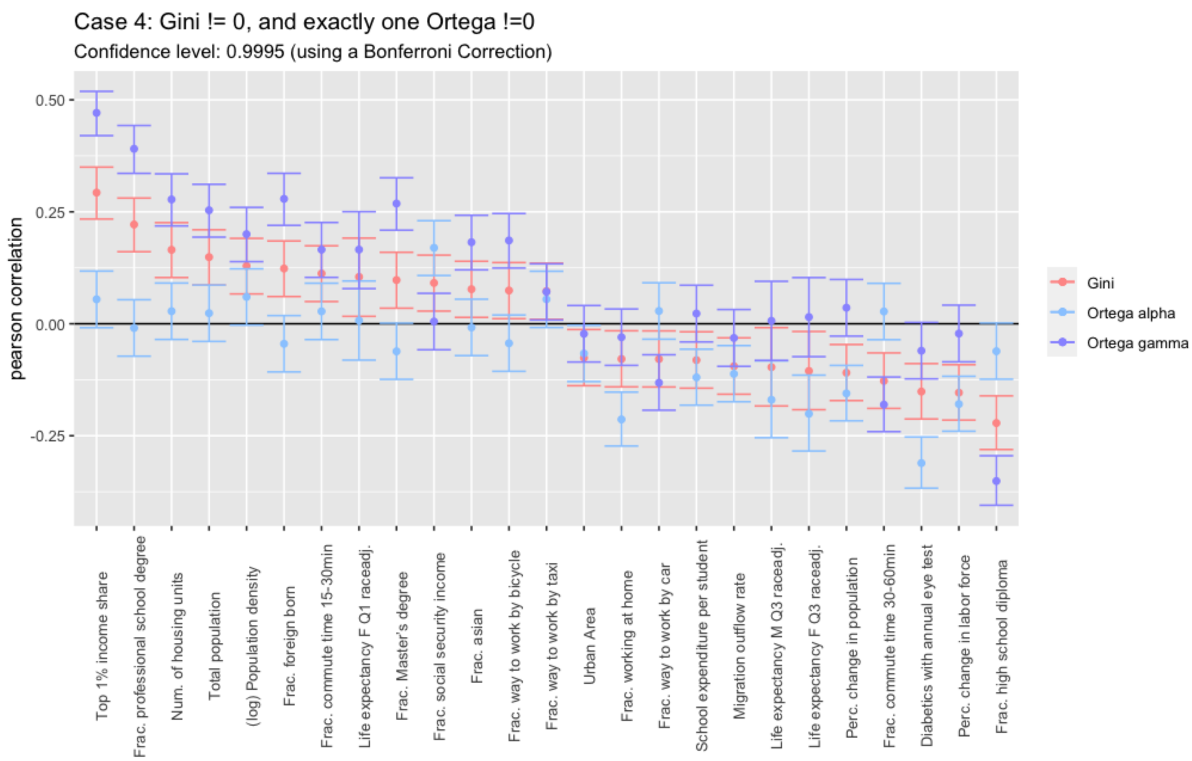
Supplementary Figure 18. The plot shows Pearson correlations for instances of case ID 1 (see Supplementary Table 11) with the Gini index and partial Pearson correlations with the Ortega parameters, i.e., the correlation between one Ortega parameter and the covariate while controlling for the other Ortega parameter across $N = 3\,049$ US counties. Pearson correlation point estimates are visualized within confidence bounds of the Bonferroni corrected confidence interval.



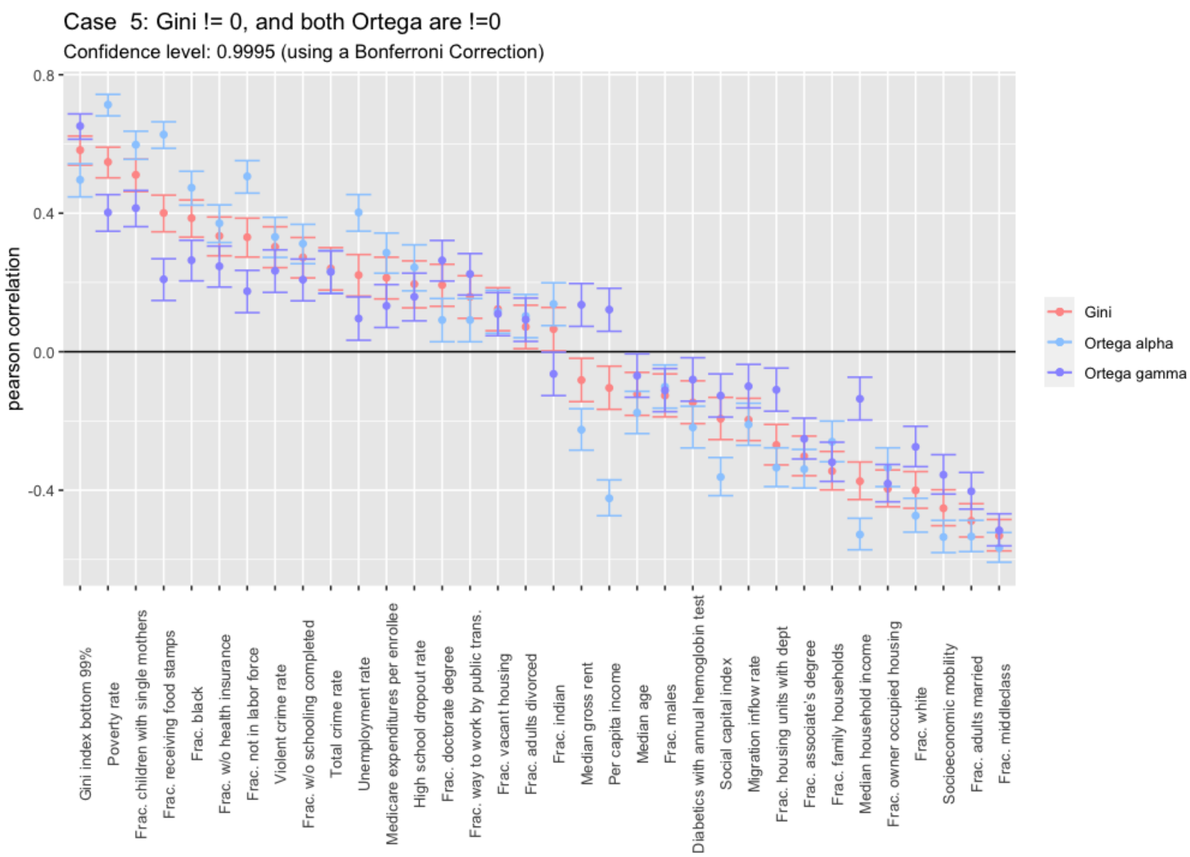
Supplementary Figure 19. The plot shows Pearson correlations for instances of case ID 2 (see Supplementary Table 11) with the Gini index and partial Pearson correlations with the Ortega parameters, i.e., the correlation between one Ortega parameter and the covariate while controlling for the other Ortega parameter across $N = 3\,049$ US counties. Pearson correlation point estimates are visualized within confidence bounds of the Bonferroni corrected confidence interval.



Supplementary Figure 20. The plot shows Pearson correlations for instances of case ID 3 (see Supplementary Table 11) with the Gini index and partial Pearson correlations with the Ortega parameters, i.e., the correlation between one Ortega parameter and the covariate while controlling for the other Ortega parameter across $N = 3\,049$ US counties. Pearson correlation point estimates are visualized within confidence bounds of the Bonferroni corrected confidence interval.



Supplementary Figure 21. The plot shows Pearson correlations for instances of case ID 4 (see Supplementary Table 11) with the Gini index and partial Pearson correlations with the Ortega parameters, i.e., the correlation between one Ortega parameter and the covariate while controlling for the other Ortega parameter across $N = 3\,049$ US counties. Pearson correlation point estimates are visualized within confidence bounds of the Bonferroni corrected confidence interval.



Supplementary Figure 22. The plot shows Pearson correlations for instances of case ID 5 (see Supplementary Table 11) with the Gini index and partial Pearson correlations with the Ortega parameters, i.e., the correlation between one Ortega parameter and the covariate while controlling for the other Ortega parameter across $N = 3\,049$ US counties. Pearson correlation point estimates are visualized within confidence bounds of the Bonferroni corrected confidence interval.

15. Simulation Study: Minimum Dataset Requirements

We introduce and evaluate **three key criteria** that datasets for inequality estimation need to possess in order for us to include them in this systematic "tournament-style" comparison to identify the best-fitting inequality measure given empirical income distributions. We find that such datasets need to contain **(1)** at least 15 or more data points per Lorenz curve; **(2)** at least two data points on top income shares above the 90th percentile of the income distribution; and **(3)** at least 60 Lorenz curves—and ideally, many more. We conducted numerous simulation studies, outlined in this section, to estimate these requirements.

In the simulation study on data granularity in the SI, Section 10, we found that for a sufficient granularity (15+ data points), and in the absence of noise, the MLE procedure will detect the correct model in almost every case if it was generated by an Ortega model (>98% of cases; see Supplementary Table 4). However, empirical observations contain observational noise. Is the AICc procedure for a given granularity of, say, 20 data points—in the presence of observational noise—still able to detect Ortega? In this case, the number of Lorenz curves available becomes crucial; i.e., if the number of Lorenz curves is too small, the reduced certainty in detecting Ortega via AICc for each Lorenz curve could lead to a false overall conclusion. But how many Lorenz curves are necessary to reduce uncertainty to reasonable amounts?

We quantify uncertainty in deciding the correct model for a given number of Lorenz curves (N) by considering each of the N Lorenz curves as independent draws from some Ortega Lorenz curve. Mathematically speaking, we can see AICc's chance of success for detecting Ortega in each of the N Lorenz curves in terms of a Bernoulli distributed variable, i.e., AICc either detects Ortega (success = 1) or not (no success = 0). From this perspective, we can interpret the Bernoulli parameter p (probability of success) as the expected percentage of Ortega detections. For N Lorenz curves, we would expect to detect $p \cdot N$ Lorenz curves as Ortega. Note that for simplicity, we assume the researcher decides for Ortega if it is detected in the majority of cases; hence we require $p>0.5$.

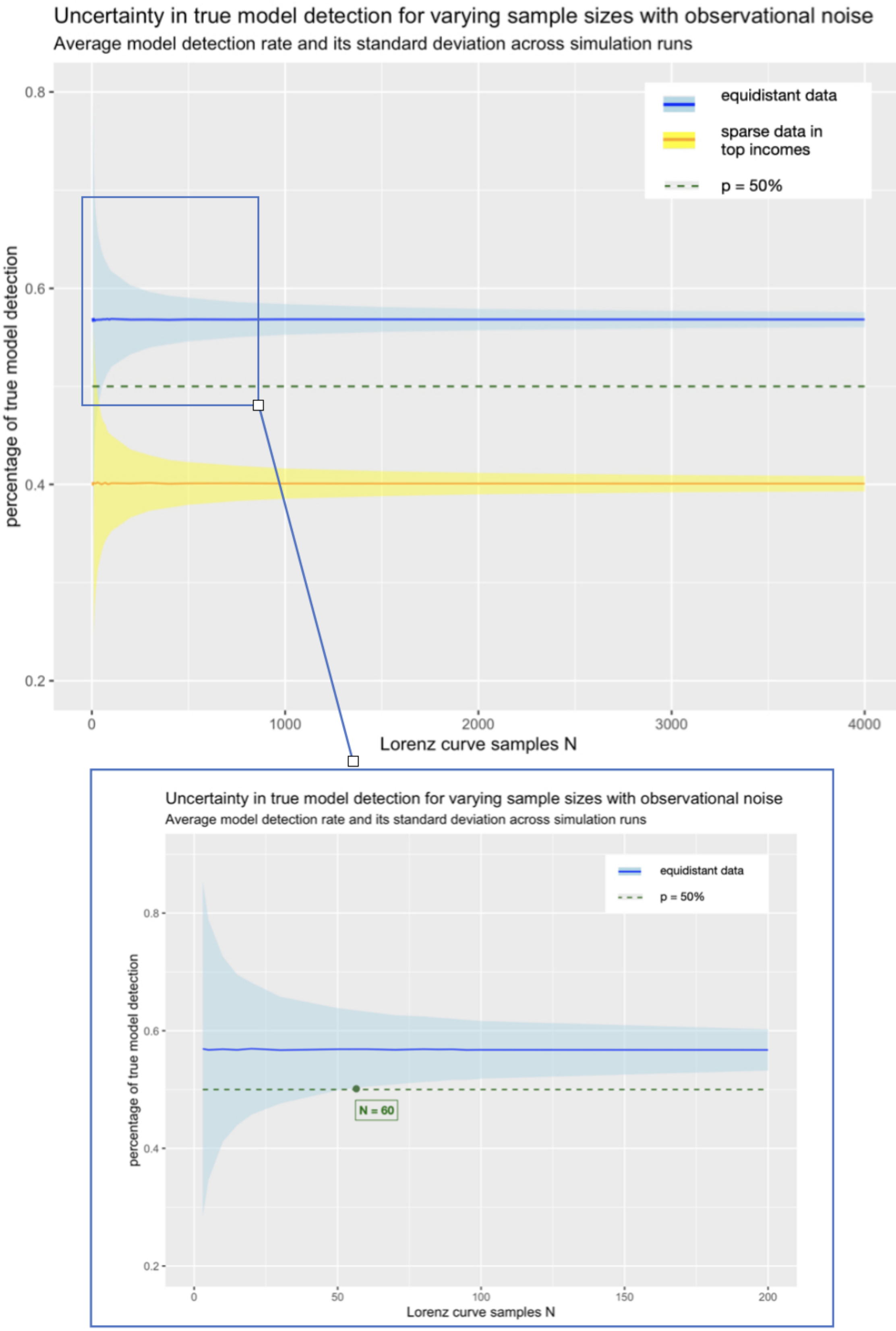
The crucial point of N is that the percentage of Ortega detections, which corresponds to the maximum likelihood estimate of Bernoulli parameter p, will approximate the true value of p more accurately with increasing N: variation in estimated p across sample sizes N is the actual quantity we are interested in when quantifying the uncertainty of determining the correct model overall. We can derive the variance of this estimator analytically; i.e.,

$$\mathrm{Var}(\hat{p}) = \frac{p(1-p)}{N} \qquad [17]$$

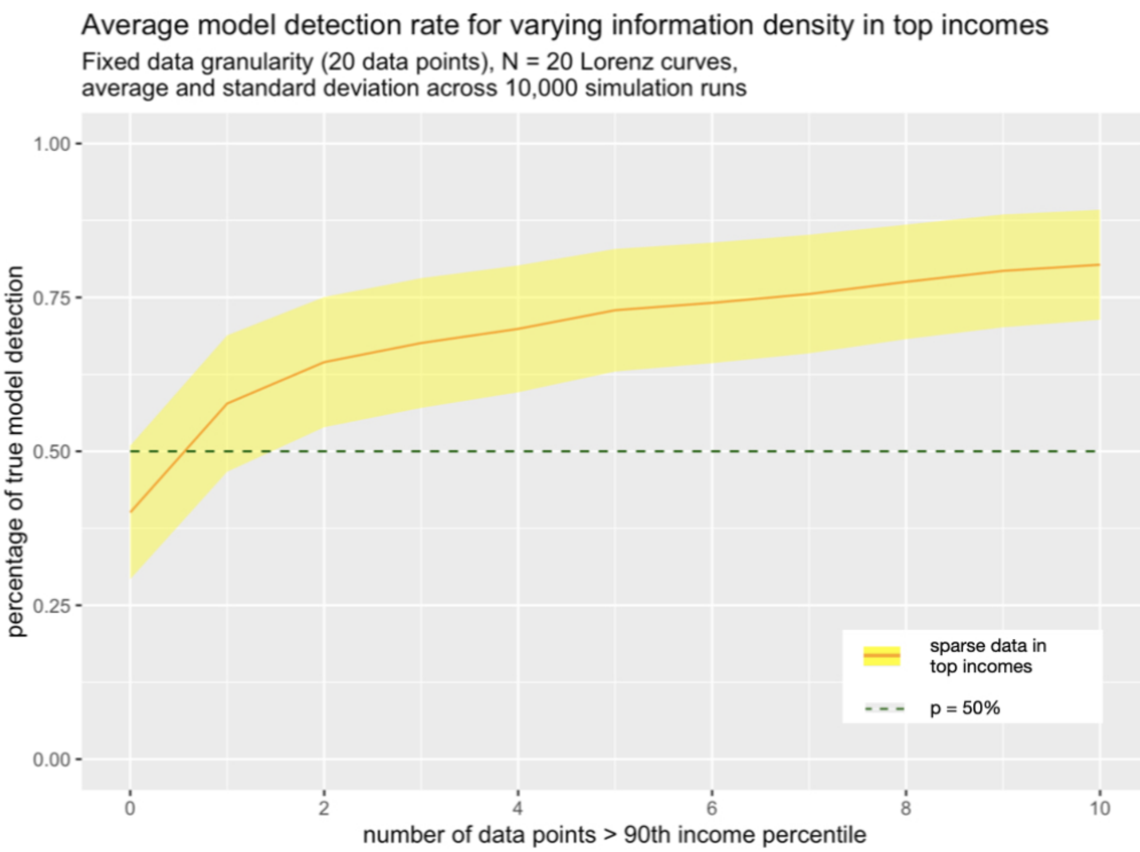
For the simulation, we vary the number of N Lorenz curves to be generated from some underlying Ortega Lorenz curve model, allowing for each of the N samples to exhibit different Ortega parameters, and a small normally distributed random noise term (mean = 0, sd = 0.002) to reflect observational noise. We then use our MLE procedure to fit various Lorenz curve models, let AICc determine the optimum model, and divide the number of detected Ortega models by N to get an estimate for p. Repeating this procedure 10 000 times gives us an estimate for the empirical standard deviation of estimated p, i.e., the standard deviation in the percentage of correctly classified Lorenz curves.

Our results show that with increased sample size N, the standard deviation of the percentage of correct model detections decreases; critically, we show that at least 60 Lorenz curves are necessary to ensure that the share of correctly classified Lorenz curves is above 50%; see Supplementary Figure 23. When fewer than 60 Lorenz curves are available, the identification of the correct model is below 50%, reflecting the challenges of using datasets that contain fewer Lorenz curves, in line with criterion #3.

In this simulation setup, we can further analyze the effects of sparse top-income data. In the base setting, we use equidistant population data shares with fixed granularity level (20 data points including population levels 0 and 1), i.e., a case where we have as much information on top-income shares as on any other parts of the income distribution. We compare this with a case where we have sparser information on top-income shares: we use the same granularity of 20 data points, but now these data points are shifted on the x-axis of the Lorenz curve toward the bottom of the income distribution, resulting in a lack of information on the top income percentiles. For example, if 1 out of the 20 data points is above the 90th percentile, this means that we have information on the bottom 90% of income earners and the 95th percentile, whereas in the case of 3 out of 20 data points being above the 90th percentile, we would have information on the bottom 90% of income earners and the 92.5th, 95th, and 97.5th percentiles. We see a considerable increase in the average percentage of true model detection as more information on top income earners is available; see Supplementary Figure 24. When fewer than two data points on top-income earners above the 90th percentile are available, the share of correctly identified models again drops below 50%, in line with criterion #2. Note that the number of Lorenz curves becomes irrelevant in this case: a higher number of Lorenz curves that do not contain top-income information do not improve our selection of the overall best-fitting model, given that $p = 0.4 < 0.5$ even when the estimated p converges with a large N. This analysis additionally reveals that our three criteria can not be treated separately but must be considered jointly.



Supplementary Figure 23. Uncertainty in true model detection: Variation in sample size



Supplementary Figure 24. Uncertainty in true model detection: Variation in information density within top incomes

References

1. J Fellman, Income inequality measures. *Theor. Econ. Lett.* **08**, 557–574 (2018).
2. N Kakwani, On a class of poverty measures. *Econometrica* **48**, 437–446 (1980).
3. R Basmann, K Hayes, D Slottje, J Johnson, A general functional form for approximating the lorenz curve. *J. Econom.* **43**, 77 – 90 (1990).
4. BC Arnold, JM Sarabia, *Majorization and the Lorenz Order with Applications in Applied Mathematics and Economics.* (Springer International Publishing), 1 edition, (2018).
5. M Krause, Parametric lorenz curves and the modality of the income density function. *Rev. Income Wealth* **60**, 905–929 (2014).
6. C Kleiber, S Kotz, A characterization of income distributions in terms of generalized gini coefficients. *Soc. Choice Welf.* **19**, 789–794 (2002).
7. C Dagum, Wealth distribution models: Analysis and applications. *Statistica* **3** (2006).
8. V Jorda, JM Sarabia, M Jantti, Estimation of income inequality from grouped data. arxiv working paper 1808.09831 (2018).
9. J McDonald, Some generalized functions for the size distribution of income. *Econometrica* **52**, 647–63 (1984).
10. NC Kakwani, N Podder, On the estimation of lorenz curves from grouped observations. *Int. Econ. Rev.* **14**, 278–292 (1973).
11. RH Rasche, J Gaffney, AYC Koo, N Obst, Functional forms for estimating the lorenz curve. *Econometrica* **48**, 1061–1062 (1980).
12. P Ortega, G Martín, A Fernández, M Ladoux, A García, A new functional form for estimating lorenz curves. *Rev. Income Wealth* **37**, 447–452 (1991).
13. D Chotikapanich, A comparison of alternative functional forms for the lorenz curve. *Econ. Lett.* **41**, 129 – 138 (1993).
14. JM Sarabia, E Castillo, DJ Slottje, An ordered family of lorenz curves. *J. Econom.* **91**, 43 – 60 (1999).
15. IM Abdalla, MY Hassan, Maximum likelihood estimation of lorenz curves using alternative parametric model. *Metodoloski Zvezki* **1**, 109–118 (2004).
16. N Rohde, An alternative functional form for estimating the lorenz curve. *Econ. Lett.* **105**, 61 – 63 (2009).
17. Z Wang, YK Ng, R Smyth, A general method for creating lorenz curves. *Rev. Income Wealth* **57**, 561–582 (2011).
18. U.S. Census Bureau, American community survey, 2011-2015: 5-year period estimates in *https://data2.nhgis.org/main.* (United States of America), (2016).
19. E Sommeiller, M Price, E Wazeter, Income inequality in the us by state, metropolitan area, and county, Technical report (2016).
20. FA Farris, The gini index and measures of inequality. *The Am. Math. Mon.* **117**, 851–864 (2010).
21. D Chotikapanich, WE Griffiths, Estimating lorenz curves using a dirichlet distribution. *J. Bus. & Econ. Stat.* **20**, 290–295 (2002).
22. AC Chang, P Li, SM Martin, Comparing cross-country estimates of lorenz curves using a dirichlet distribution across estimators and datasets. *J. Appl. Econom.* **33**, 473–478 (2018).
23. K Aho, D Derryberry, T Peterson, Model selection for ecologists: the worldviews of aic and bic. *Ecology* **95**, 631–636 (2014).
24. AR Liddle, Information criteria for astrophysical model selection. *Mon. Notices Royal Astron. Soc. Lett.* **377**, L74–L78 (2007).
25. SJ Brams, PC Fishburn, Chapter 4 voting procedures in *Handbook of Social Choice and Welfare*, Handbook of Social Choice and Welfare. (Elsevier) Vol. 1, pp. 173 – 236 (2002).
26. KJ Arrow, A difficulty in the concept of social welfare. *J. political economy* **58**, 328–346 (1950).
27. AE Raftery, Bayesian model selection in social research. *Sociol. Methodol.* **25**, 111–163 (1995).
28. KP Burnham, DR Anderson, Multimodel inference: Understanding aic and bic in model selection. *Sociol. Methods & Res.* **33**, 261–304 (2004).
29. E Belz, Estimating Inequality Measures from Quantile Data, (Center for Research in Economics and Management (CREM), University of Rennes 1, University of Caen and CNRS), Economics Working Paper Archive (University of Rennes 1 & University of Caen) 2019-09 (2019).
30. S Md, C Saeki, A new functional form for estimating lorenz curves. *J. Bus. Econ. Res.* **1** (2011).
31. KS Cheong, An empirical comparison of alternative functional forms for the lorenz curve. *Appl. Econ. Lett.* **9**, 171–176 (2002).
32. JM Sarabia, V Jorda, C Trueba, The lamé class of lorenz curves. *Commun. Stat. - Theory Methods* **46**, 5311–5326 (2017).
33. D Chotikapanich, W Griffiths, Averaging Lorenz curves. *The J. Econ. Inequal.* **3**, 1–19 (2005).
34. JM Sarabia, E Castillo, D Slottje, An ordered family of lorenz curves. *J. Econom.* **91**, 43–60 (1999).
35. BC Arnold, *Pareto Distribution.* (American Cancer Society), pp. 1–10 (2015).
36. JL Gastwirth, A general definition of the lorenz curve. *Econometrica* **39**, 1037–1039 (1971).
37. RJ Gordon, I Dew-Becker, Selected Issues in the Rise of Income Inequality. *Brookings Pap. on Econ. Activity* **38**, 169–192 (2007).
38. S Voitchovsky, Does the profile of income inequality matter for economic growth? *J. Econ. growth* **10**, 273–296 (2005).
39. SF Reardon, K Bischoff, Income inequality and income segregation. *Am. journal sociology* **116**, 1092–1153 (2011).

40. R Chetty, et al., The association between income and life expectancy in the united states, 2001-2014. *Jama* **315**, 1750–1766 (2016).
41. R Chetty, N Hendren, The Impacts of Neighborhoods on Intergenerational Mobility II: County-Level Estimates. *The Q. J. Econ.* **133**, 1163–1228 (2018).