README to Replicate:

Multidimensional Inequality Measurement via Optimal Transport

Forthcoming at the Review of Economics and Statistics

Yanqin Fan, Marc Henry, Brendan Pass and Jorge A. Rivero*

October 26, 2024

Contents

1 Requirements **2**

1.1 Installation 3

2 Quick Guide **3**

3 Replication Folders & Files **4**

3.1 main.R 4

3.2 Functions Folder 4

3.2.1 package_installer.R 4

3.2.2 scaling.min.max.R 4

3.2.3 vquantile.R 5

*Contact: Jorge A. Rivero (riverj11@rpi.edu)

3.2.4 Lmap.R . . . 6
3.2.5 eCDF.R . . . 7
3.2.6 ILF.R . . . 7
3.2.7 create_dataset.R . . . 8
3.2.8 separate_imputes.R . . . 8
3.2.9 aggregate_inequality.R . . . 8
3.2.10 grouped_inequality.R . . . 8
3.3 Data Folder . . . 9
3.4 Figure Folder . . . 9
3.4.1 fig1.R . . . 9
3.4.2 fig2.R . . . 9
3.4.3 fig3.R . . . 9
3.4.4 fig4.R . . . 10
3.4.5 fig5.R . . . 10
3.4.6 fig6.R . . . 10
3.4.7 fig7.R . . . 10
3.4.8 fig8.R . . . 11
3.5 Results Folder . . . 11

4 Data **11**

References **11**

1 Requirements

- Windows 10 or 11 operating systems.

- A minimum of 8 GB of RAM (16 GB is strongly recommended).
- R version 4.0.5 (`https://cran.r-project.org/bin/windows/base/old/`).
- Rtools 4.0 (`https://cran.r-project.org/bin/windows/Rtools/rtools40.html`).
- R Studios is strongly recommended (Point 2 in `https://posit.co/download/rstudio-desktop/`).

1.1 Installation

The user may consult with the basic installation directions for each of these programs. During installation, the default options should be chosen. After R 4.0.5, Rtools 4.0 and R Studios have been installed, verify that R Studios has detected R 4.0.5 by typing `R.version` in the console. If another version of R is indicated, navigate to Tools → Global Options → General and under R Sessions choose R 4.0.5.

2 Quick Guide

After installation, set the working directory to where `main.R` is located. Running `main.R` will run `package_installer.R`, which will prompt the user to update or install packages. To proceed, the prompts should be ignored by pushing enter on the keyboard. Afterwards, the program will continue to run and execute all the scripts to generate figures. Occasionally, minor warnings will come up; see the `vquantile.R` and `aggregate_inequality.R` below for a description of them.

The total run time was approximately 17 hours using a desktop with an AMD Ryzen 7 9700X 8-Core CPU and approximately a week with an Intel i5 8600k CPU.

3 Replication Folders & Files

3.1 main.R

This script automatically installs and loads all necessary packages, load all functions and generates all figures for the user.

3.2 Functions Folder

3.2.1 package_installer.R

This script installs and loads all the necessary packages and functions in order to replicate the paper. Since the required version of R is n ot the current version, binary versions of packages are used since they are compatible and provide the quickest installation. The exception is the need to install version $0.14 - 7$ of `transport` and to install `Rgeogram` from Github. There is an error as of the date of this document that makes `Rgeogram` incompatible with the current release of R, hence the need for R version 4.0.5. When this issue is resolved, this document will be revised to ensure replication is reliable as updates are released for R. See the .R file for descriptions of the packages in use.

Note, the user should ignore all updates and recommended installations when prompted. Simply press enter on your keyboard to skip.

3.2.2 scaling.min.max.R

This function shifts and scales the bivariate allocation to fit in $[0, 1] \times [0, 1]$. This is a required step to use the data as input for the `otm2d` function from the `Rgeogram` package, used in the following `vquantile.R` function.

3.2.3 vquantile.R

This function solves for the vector quantile Q_X for a discrete bivariate allocation X using the (continuous) uniform distribution over $[0,1] \times [0,1]$ as a reference distribution. This problem is known as semi discrete optimal transport, which amounts to an unconstrained convex minimization routine to obtain a vector of weights that form a power diagram; see Aurenhammer [1987]. This function uses the `otm2d` function from the `Rgeogram` package[1] to calculate these weights and the `power_diagram` function from the `transport` package is used to obtain the cells of the power diagram implied by the weights. The output of this function is a crucial input for ILF.R.

The `otm2d` function utilizes the multi-scale approach by Mérigot [2011] that improves stability– the solution weights should correspond to a power diagram with no empty cells. The empty cell issue may arise if convergence is not met and the algorithm is halted prematurely. The multi-scale approach does not guarantee stability, so `vquantile.R` also repeats this process until the issue is resolved. In the code, the function checks if the first-order conditions for the weights are satisfied by considering the worst-case error. If this error passes a threshold, the algorithm is repeated.

A known issue within input data is the presence of duplicate points. Duplicate points in the allocation will cause empty cells, essentially because the point has been assigned to an existing cell and no two points can be assigned to the same cell. The beginning portion of the .R file aggregates duplicates by adjusting the sample weight of that point. This is done first to avoid unnecessary repeats of the algorithm.

Please see the commented-out section in the .R file for information on all inputs and outputs.

[1] Github repository: `https://github.com/TraME-Project/Rgeogram`

3.2.4 Lmap.R

Please refer to the computation section in the paper for relevant formulas and terminology.

This function calculates the vector cumulative share of resources possessed by the bottom vector rank specified by the user. The power diagram output from `vquantile.R` function is a crucial input for this function. The remainder details calculating the weights found in the paper, which are the areas of intersections of the lower rectangle formed by the specified vector rank and cells associated to each data point.

We can leverage inequality conditions on the vertices of the power diagram cells in relation to the selected vector ranks to quickly determine if intersections are empty or trivial, where the cell is completely contained in the rectangle formed by selected vector ranks. In the case of empty intersections, the weight is zero. In the case of trivial intersections, the weight is simply the sample weight associated to the point, which is given by the first-order conditions. If the intersection is neither, we use the `sf` package to form polygon objects out of the intersections and calculate their areas. This process reduces many calculations of intersections.

Finally, a weighted sum is calculated using the areas of intersections as weights and is returned as the Lorenz map vector value to be primarily used in the ILF.R function. See the .R file for details on outputs and inputs.

Please see the commented-out section in the .R file for information on all inputs and outputs.

3.2.5 eCDF.R

This function calculates the empirical cumulative distribution function in a vectorized and simple fashion, returning only a scalar output.

3.2.6 ILF.R

This function generates a pseudo sample of Lorenz map vector values of a bivariate allocation using many generated vector ranks from a uniform distribution over $[0,1] \times [0,1]$ and the `Lmap.R` function. The sample of Lorenz map values are computed in parallel.

With the Lorenz map vector values, values of the the Inverse Lorenz Function (ecdf) are calculated using `eCDF.R` and returned as a matrix with rows and columns representing a grid on $[0,1] \times [0,1]$ and values are the ecdf values associated to the grid point. An estimate of the Gini index for the bivariate allocation is obtained via plug-in. In addition, if a matrix with rows comprising of vector ranks of interest is supplied, `ILF.R` returns a matrix of vector cumulative shares with respect to those ranks via `Lmap.R`.

With the ecdf output, the user can extract α-Lorenz curves as the contours associated to the $0 < \alpha < 1$ level of the ecdf. We use standard graphical tools, such as `ggplot2`, to plot these contours.

Please see the commented-out section in the .R file for information on all inputs and outputs.

3.2.7 create_dataset.R

This script collects the variables used in the empirical illustration (sample weight, income, assets, age and race) and stacks each year release into one large data set.

3.2.8 separate_imputes.R

Using the data set output of `create_dataset.R`, the script creates a list that separates each year-impute combination to then be used to calculate all relevant quantities for each year-impute pair.

3.2.9 aggregate_inequality.R

This script calculates the ILF, Gini index and resource share estimates for each year-impute combinations used in figures 5,6 and 7 and their respective .R files. In some years, data points can be extremely close in the lower strata of income-wealth earners and can result in stability issues; see the `vquantile.R` section. To counteract this, a very small amount of noise relative to the magnitude of income/wealth is added to the allocation. In addition, the error message: "*warning from vquantile: In constraint - W : longer object length is not a multiple of shorter object length*" appears frequently when running this script. This is due to the constraint being met when checking for empty cells and can be safely ignored.

3.2.10 grouped_inequality.R

This script is identical to `aggregate_inequality.R`, with the exception that the data is filtered to the groupings required for figure 8 (white/black and working age/retired age).

3.3 Data Folder

This folder contains csv files for the public 1989 – 2022 triennial releases of the Survey of Consumer Finances; see the Data section for a description. The format is "`SCFP<year>.csv`".

3.4 Figure Folder

This folder contains .R files that are aptly named `fig<#>.R` to signify which figure from the paper they pertain to.

3.4.1 fig1.R

This script generates figure 1. Data is generated from a bivariate standard normal distribution and the vector quantile is calculated using `otm2D`. Then, the power diagram is calculated and a plot with the sample juxtaposed on the power diagram is generated.

3.4.2 fig2.R

This script generates figure 2. A list is generated where each entry is a sample from a different multivariate lognormal distribution indexed by different correlation parameter as given in the paper. Then, the ILF is calculated across list elements using `ILF.R`. The contours are plotted using `ggplot2` and `geom_contour`.

3.4.3 fig3.R

This script generates figure 3. Similarly to `fig2.R`, a list is generated where each entry is a sample from a different multivariate lognormal according to the parameters

given in the paper. The plots are generated similarly to `fig2.R` after using `ILF.R` to estimate each ILF.

3.4.4 fig4.R

This script generates figure 4. The two samples are generated from multivariate lognormal distributions, their ILFs are estimated using `ILF.R` and then contours plotted.

3.4.5 fig5.R

This script generates figure 5 using the output of `aggregate_inequality.R`. For each year, the ILF estimates of each impute are averaged. The paper limits the curves to 4 years: 1992, 2007, 2010 and 2022, however this file can be modified to show other years.

3.4.6 fig6.R

This script generates figure 6 using the output of `aggregate_inequality.R`. For each year, the resource share estimates for both share thresholds of each impute are averaged. The result is plotted over all the years in the data set.

3.4.7 fig7.R

This script generates figure 7 using the output of `aggregate_inequality.R`. For each year, the multivariate Gini index estimates of each impute are averaged. The same is done for computing Kendall's tau and the univariate income and wealth Gini indices using the `Gini` function from the `DescTools` package. The result is plotted

over all the years in the data set.

3.4.8 fig8.R

This script generates figure 8 using the output of `grouped_inequality.R`. For each year, the multivariate Gini index estimates for each group of each impute are averaged. The result is plotted over all the years in the data set.

3.5 Results Folder

The results folder contains .png files of all figures that are generated by the replication files.

4 Data

The data is sourced from the public version of the triennial Survey of Consumer Finances from 1989–2022, which is publicly available. You can find the 2022 release and the 1989–2019 releases by following `https://www.federalreserve.gov/econres/scfindex.htm` and `https://www.federalreserve.gov/econres/scf-previous-surveys.htm`, respectively. We've also included the data as csv files separated by year.

Data points are sometimes duplicated in the data files and are aggregated by sample weight to ensure the vector quantile is computed correctly; see the section on `vquantile.R` for issues related to duplicate points.

References

F. Aurenhammer. Power diagrams: properties, algorithms and applications. *SIAM Journal on Computing*, 16:78–96, 1987.

Q. Mérigot. A multiscale approach to optimal transport. *Computer Graphics Forum*, 30:1583–1592, 2011.