POLITECNICO DI MILANO
School of Industrial and Information engineering
Master degree in Mathematical Engineering

POLITECNICO
MILANO 1863

A Functional Approach to the Study of Income Inequality

Supervisor: **Prof. Simone Vantini**
Co-supervisor: **Dr. Matteo Fontana**
Prof. Alessandra Menafoglio

Candidate:
Lorenzo Boletti
Matr. 904188

AY 2018-2019

Table of contents

List of figures

List of tables

Abstract

Economic inequality is an ever-present topic in the economic debate due to the consequences on the modern society. The most used statistical measure to represent economic inequality is the Gini index, defined as the ratio between the area between the Lorenz curve and perfect equality line, and the total area under the perfect equality line. Although it lends itself to comparing the inequality between countries and the evolution over time, the Gini index is unable to describe the inequality composition in details.

The aim of this thesis is to give a complete and realistic description of income distribution, able to catch the inequality composition inside the population. To do this, we proved an alternative methodology to Gini index in the analyses of inequality. In this regard, we use of an infinite-dimensional approach based on functional objects and the Functional Data Analysis. The functions used to describe income inequality are Lorenz curves and density distributions.

We conduct a first exploratory analysis to identify the presence of significant differences in the domain of functional objects with respect to geographical and economic factors. To do this, we introduce the Interval-Wise Testing Procedure, a non-parametric inferential methodology. To identify the variables affecting the inequality profile, we compute a regression model with functional response and scalar variables. Finally, we describe agglomerative clustering techniques that, defined an index of dissimilarity, divide different countries into clusters according to their inequality profile.

Keywords: Functional Data Analysis, Lorenz curves, Income inequality

Sommario

La disuguaglianza economica è un argomento molto attuale nel dibattito economico per le conseguenze che ne conseguono sulla società moderna. La misura statistica più usata per quantificare la disuguaglianza economica è l'indice di Gini, definito come il rapporto dell'area tra la curva di Lorenz e la linea di egualità, e l'area totale sotto tale linea. Nonostante si presti a confrontare la disuguaglianza fra diversi stati e come questa evolve nel tempo nel tempo, l'indice di Gini non è in grado di descriverne la composizione nei dettagli.

L'obbiettivo di questa tesi è quello di fornire una descrizione più completa e veritiera della distribuzione del reddito, capace di cogliere la composizione della disuguaglianza all'interno della società. Per fare questo, proponiamo un'alternativa metodologica all'indice di Gini nell'analisi della disuguaglianza. A tal proposito ci avvarremo di un approccio infinito-variato basato sugli oggetti funzionali e sull'analisi di dati funzionali. Le funzioni usate sono le curve di Lorenz e le distribuzioni di densità.

All'interno della tesi viene condotta una prima analisi esplorativa per individuare la presenza di differenze significative nel dominio degli oggetti funzionali rispetto a fattori geografici ed economici. Per fare questo, introduciamo la Interval-Wise Testing Procedure, una metodologia inferenziale non-parametrica. Per identificare quali variabili incidano sulla conformazione della disuguaglianza, viene impostato un modello di regressione con risposta funzionale e variabili scalari. Infine, vengono descritte alcune tecniche di clustering agglomerativo che consentono, definito un indice di dissimilarità, di suddividere in clusters le diverse nazioni secondo il proprio profilo di disuguaglianza.

Parole chiave: Analisi di dati funzionali, Curve di Lorenz, Disuguaglianza di reddito

Chapter 1

Introduction and Motivations

Inequality is the term used to indicate a situation where different people or groups of people have a differentiated access to services and resources. In a world dominated by money, the access to tangible (i.e. necessary, luxury) and intangible (i.e. health, education) assets depends heavily on economic resources. It is not surprising that, among the various typologies of inequalities, the economic one has a particular interest in the public and academic debate, especially for all for the consequences to which it leads. Extreme economic inequality can represent a barrier to poverty reduction, weaken economic growth, lead to unequal access to services such as health and education and lead to differences in opportunities. Poverty and inequality interact, making poor people increasingly poorer [4].

The two main types of economic inequality are wealth inequality, which represents how assets are distributed, and income inequality, which indicates new earnings added to the wealth. The latter is the subject of our study.

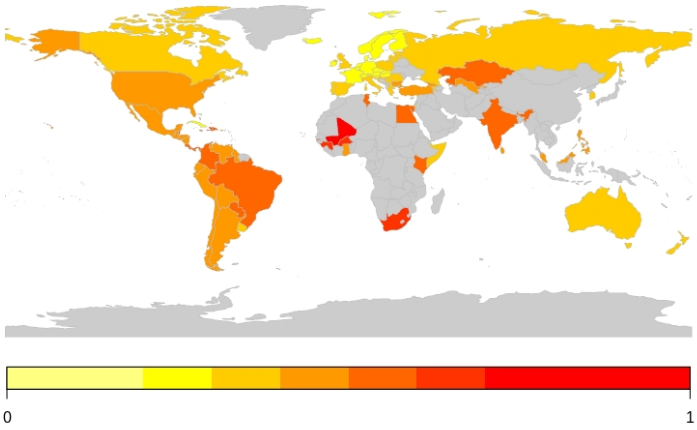


Figure 1.1: Income Gini index on World map

The most used measure in describing and analysing economic inequality is the Gini index. This is mainly due to its dual characteristic of quantifying the society internal inequality and comparing it between different countries as a relative measure. It is defined as the ratio between the area between the Lorenz curve and the egalitarian line, the extreme case where all the population has the same income, and the total area under the perfect equality line.

$$\text{Gini} = 1 - 2\int_0^1 L(z)dz$$

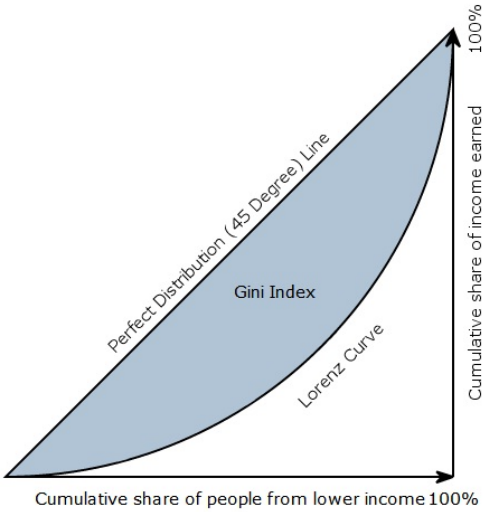


Figure 1.2: Gini index from Lorenz curve

Gini index is typically used in large-scale analyses, as the comparison between multiple inequality profiles, or in quantitative analyses where we want to study its evolution and composition through statistical tools [15]. Although Gini index is widely used, an increasing number of economists, among which stands out Thomas Piketty [9], are sceptical that a single statistical measure can capture the entire inequality of a country. As constructed, Gini index puts equal weight on the entire income distribution, without showing anything about inequality composition in the population ranges. In fact, a Gini index could decrease even when absolute poverty increases.

The aim of this thesis is to provide a methodological alternative to Gini index, able to give an inequality description and analysis as complete as possible, filling the gaps that this statistical measure has. As complete, we want to detect the inequality nested in the population, without generalize any aspect. Gini index, as mentioned, implies an income that varies constantly between the population, a situation that often does not respect reality. We want to detect the real composition of income

inequality, in and between which ranges this is more or less accentuated. In this regard, we introduce an infinite-dimensional approach, based on functional objects and Functional Data Analysis as statistical tools to compute the analyses.

The first functional objects we analyse are the Lorenz curves, constrained functions showing the $p_2\%$ of income explained by the poorest $p_1\%$ of population. Since Gini index is computed by Lorenz curves as summary measure, we want to point out the inequality aspects not perceptible from Gini index. Another functional object used to analyse income distribution, is the Probability Density Function. We can detect different aspects in inequality profile with respect to Lorenz curves, and this approach give an easier visual interpretation to the distribution. As constrained functions, both the functional objects require to be mapped in Hilbert spaces to compute consistent analyses [3] [7].

In a first exploratory analysis we use Interval-Wise Testing (IWT) procedure [10], a non-parametric inferential technique developed in the frameworks of Functional Data Analysis. We apply it to identify to which extent the inequality profiles differ from each other's with respect to a geographical and wealth factors. We can also detect the portion of the domain in which these differences are more significant. In a following descriptive analysis we compute a regression model with functional response and scalar variables to visualize how the regressors affect the inequality profiles of functional objects. Finally, we describe agglomerative clustering techniques that, defined an index of dissimilarity, divide different counties into clusters according to their inequality profile.

This thesis is organized as follow:

- In chapter 2, we provide a review of the methodology about Functional Data Analysis and transformation of embedded spaces into Hilbert spaces.
- In chapter 3, we present the inequality dataset and the variables used in the regression. Then we introduce the functional objects used in the analysis.
- In chapter 4, we present the results of the analyses computed with statistical tools, described in the previous chapter. Results for different functional objects (Lorenz curves and density functions) are provided in different sections.
- In chapter 5, we summarize the conclusions and the results obtained, discussing some possible improvements and developments.

Chapter 2

Methodology

The use of infinite dimensional elements for the description of inequality requires ad-hoc methods for analysis, different from those used in univariate analysis. The branch of statistics that analyses curves, treated as functions, is called Functional Data Analysis (FDA).
In this chapter, first, we describe the smoothing process to obtain functional data from row data [14]. We provide two data transformation to deal and analyse embedded functions [3] [7]. Then, we explain how different type of functional regression work and we introduce Interval-Wise Testing procedure [10] [1] to identify the significant intervals for the variable coefficients. Finally, we provide an overview on hierarchical clustering.

2.1 Smoothing

In this section the smoothing process of functional data is introduced. Data raw consist in n pairs (t_j, y_j) for each realization, and it is assumed that all the realizations have the same number of points. The smoothing procedure estimates the functions from which discrete data derive, and this is the process of going from data points to functions.

In Functional Data Analysis, it is assumed that raw data are observations of functions affected by noise. The signal plus noise model notation is the following:

$$\mathbf{y(t) = x(t) + e(t)}$$

where **x** is to be considered as a fixed effect and **e** the error with zero mean and finite variance.

Smoothness allows to build functions differentiable one and more times. So, if we work with derivative, the smoothing process is a necessary step [14]. The method used to smooth the data is a penalized and constrained basis expansion. The request that the bases are constrained is due to the structure of the functional data which have conditions to satisfy. The method combines basis functions and projects discrete data onto them.

2.1.1 Spline basis system

A basis function system is a set of known functions ϕ_k that are mathematically independent of each other and that have the property to approximate arbitrarily well any function by taking a weighted sum or linear combination of a sufficiently large number K of these functions [14].
A function $x(t)$ can be represented as a linear combination of K known basis functions ϕ_k:

$$x_k(t) = \sum_{k=1}^{K} c_k \phi_k(t)$$

Or in matrix notation:

$$x = \boldsymbol{\Phi} \mathbf{c}$$

Where $\mathbf{c}$ is the vector of coefficients c_k and $\boldsymbol{\Phi}$ the functional vector of basis functions ϕ_k. If $K = n$ the vector $\mathbf{c}$ can be chosen to achieve a perfect interpolation. Usually, increasing the parameter K from 1 to n, the basis fits the observed data always better, but there is the possibility of overfitting. The observed data are assumed to be a combination of informative data and noise. In addition, high frequency oscillations are a negative consequence of overfitting, that lead to derivatives with high variability [14].

Two are the most useful basis adopted in applications depending on the nature of data: Fourier basis for periodic data and B-spline basis for non- periodic data. The structure of the data points clearly shows that the quantiles are not periodic. So, B-spline basis functions will be used.

2.1.2 B-spline basis

First of all, a spline function is defined over an interval dividing it in L sub-intervals separated by values τ_l $l = 1, ..., L-1$ called breakpoints. Including the two endpoints, $L + 1$ breakpoints are defined. The way to have a better fitting is to increase the number of breakpoints, and consequentially the number of basis functions.

On each interval the spline is defined as a polynomial of order m, remembering that the order of a polynomial is the number of parameters required to define it. Moreover, there exists the property that adjacent polynomials must join smoothly at the breakpoints, with derivatives up to order $m - 2$.
Summing up, a spline function is determined by:

- the order of the polynomial segments;
- the breakpoints sequence τ;

The most popular spline system is the B-spline basis system. B-splines have some further properties:

- Any linear combination of B-spline basis functions is a spline function;
- An order m B-spline is positive over no more than m adjacent intervals.

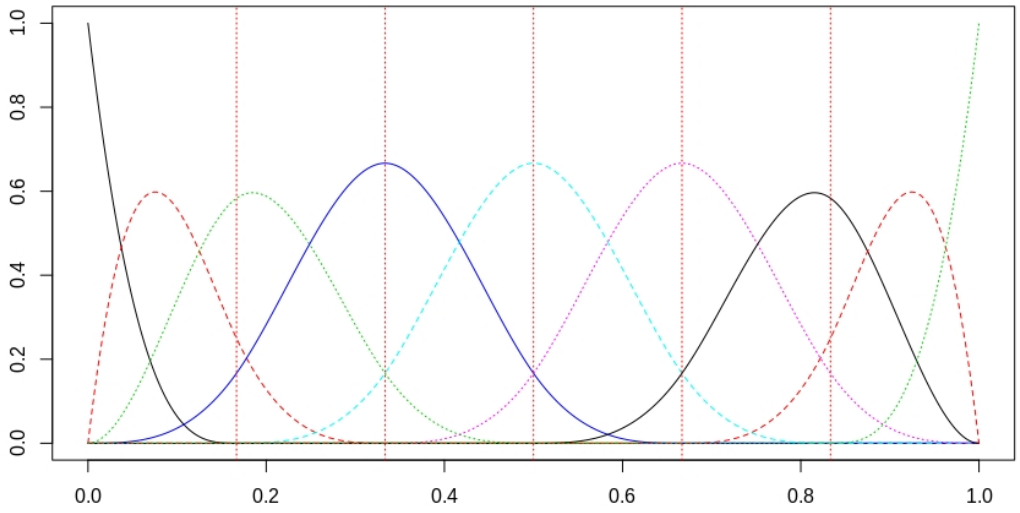


Figure 2.1: B-spline basis

A system of B-spline basis of order 4 with 9 breakpoints (7 internal + 2 endpoints) is represented in figure (2.1). Notice that the number of basis functions is equal to $m + L - 1$, i.e. the order of polynomials plus the number of internal breakpoints.

To approximate a given function with this basis system, least squares estimate procedure is used. This involves minimizing the least squares criterion:

$$\mathrm{SMSSE}(\mathbf{y} \mid \mathbf{c}) = \sum_{j=1}^{n} [y_i - \sum_{k=1}^{K} c_k \phi_k(t_j)]^2$$

Defining the matrix $\boldsymbol{\Phi}$ as:

$$\boldsymbol{\Phi} = \begin{bmatrix} \phi_1(t_1) & \phi_2(t_1) & \phi_3(t_1) & \dots & \phi_K(t_1) \\ \phi_1(t_2) & \phi_2(t_2) & \phi_3(t_2) & \dots & \phi_K(t_2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_1(t_n) & \phi_2(t_n) & \phi_3(t_n) & \dots & \phi_K(t_n) \end{bmatrix}$$

The equation can be written in matrix notation:

$$SMSSE(\mathbf{y} \mid \mathbf{c}) = (\mathbf{y} - \boldsymbol{\Phi}\mathbf{c})'\mathbf{W}(\mathbf{y} - \boldsymbol{\Phi}\mathbf{c}) \tag{2.1}$$

Where $\mathbf{W}$ is a symmetric positive definite matrix that allows a weighting of residuals. The estimate of the coefficient vector $\mathbf{c}$ is obtained taking the derivative of (2.1) with respect to $\mathbf{c}$, and setting the equation to zero and then solve it for $\mathbf{c}$. This gives the following estimator:

$$\hat{\mathbf{c}} = (\boldsymbol{\Phi}^{\mathsf{T}}\mathbf{W}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathsf{T}}\mathbf{W}\mathbf{y}$$

2.1.3 Roughness penalty

With the described process, the amount of smoothness of the functions depends on the number of basis and consequently on the number of breakpoints, for which a decision about where to position them is to be made. A valid and much used alternative is instead the so-called smoothing spline approach, which involves a roughness penalty and a breakpoint set equal to the structure of data.

The roughness penalty approach is often preferred to basis function or local smoothing techniques in most applications because it has the same advantages but also some further properties such as producing better results in the estimation of derivatives [14]. The roughness penalty method varies the optimization of the fitting criterion (2.1) adding a penalty term:

$$PEMSSE_{\lambda}(\mathbf{y} \mid \mathbf{c}) = (\mathbf{y} - \boldsymbol{\Phi}\mathbf{c})'\mathbf{W}(\mathbf{y} - \boldsymbol{\Phi}\mathbf{c}) + \lambda PEN_{m}(\boldsymbol{\Phi}\mathbf{c}) \tag{2.2}$$

Where λ is a smoothing parameter.
Since the general idea is to control the curvature of function, the penalizing term usually considered is the L^2 norm of the second derivative of the B-spline function x:

$$PEN_2(x) = ||D^2x||^2 = \int [D^2x(s)]ds$$

2.1.4 Constrained smoothing

Sometimes curves must satisfy constrains. If data are values that cannot be negative, the estimated curve has to be positive, even over regions where the values are close to 0. If growth curves are estimated, it is probably the case that negative slopes are implausible [14].

Positive smoothing

A positive function x can be defined as the exponential of an unconstrained function W:

$$x(t) = e^{W(t)}$$

Often base e is used to transform the data, but other bases are always appropriate. Since $W(t)$ is an unconstrained function, it is reasonable expand it in terms of basis function:

$$W(t) = \sum_k c_k \phi_k(t)$$

It is possible use different basis in terms of the data type.

Monotone smoothing

Asking monotonicity is equal to the positivity of the first derivative Dx.
Define the derivative as an exponential function:

$$Dx(t) = e^{W(t)}$$

By integrating both sides of the equation, it is obtained:

$$x(t) = C + \int_0^t e^{W(z)} dz$$

Convex smoothing

In the same way, asking for the convexity constrain require applying a transformation to the second derivative D^2x.

$$D^2x(t) = e^{W(t)}$$

Getting the general solution:

$$x(t) = c_1 + c_2 p + \int_0^t \int_0^z e^{W(x)} dx dz$$

2.2 Negative centred log-ratio transformation

Constrained functional data cannot be analysed with the classical methods of functional data analysis, because of their complex structure, embedded in a simplex of L^2. The functions require a map that takes them in a Hilbert space [3].

2.2.1 Lorenz curves as functional data

Let the Lorenz curve family be defined as:

$$L_{or} = \{L : [0,1] \to [0,1] : L(0) = 0,\ L(1) = 1,\ L \in C^2[0,1],\ L'(p) > 0, L''(p) > 0\}$$

This $\forall p \in (0,1)$. Where L_{or} is a subset of $L^2_{[0,1]}$, the Hilbert space of square integrable real function on $[0,1]$.

Considering a family $\mathbb{X}$ composed by different populations in form of random variable X and a map $\mathbb{L} : X \to L(*)$, it is possible to define the mean curve and the covariance operator. $\forall p \in [0,1]$ and taking $v \in L^2_{[0,1]}$:

$$l(p) = E\{L(p)\}, \quad \boldsymbol{\Sigma}(v) = E\{\langle L - l, v\rangle(L - l)\}$$

Meanwhile, considering the empirical counterpart, let's deal with a sample $X_1, ..., X_n$ and the corresponding empirical Lorenz curves $\hat{L}_i$.
The empirical mean and covariance operator are defined as follow;

$$\hat{l}_n(p) = \frac{1}{n}\sum_{i=1}^{n}\hat{L}_i(p), \quad \hat{\boldsymbol{\Sigma}}_n(v) = \frac{1}{n}\sum_{i=1}^{n}\langle\hat{L}_i - \hat{l}_n, v\rangle(\hat{L}_i - \hat{l}_n)$$

2.2.2 L''(p)

Dealing with constrained data, a possible approach is to express them as solutions of differential equations and work directly with derivatives if they have a physical meaning [3]. The second derivative of the Lorenz curve, L''(p), can be represented in the following way:

$$L''(p) = \frac{s(p)}{\mu}$$

s(p) represents the sparsity function. It measures the data sparseness around the quantile q(p). L''(p) is interpretable as the local measure of inequality of individuals close to the p-quantile.

The relation between Lorenz curve and its second derivative is summarized as follow:

$$L_{or} \xrightarrow{D^2} D^2 L_{or}$$
$$L_{or} \xleftarrow{BVP} D^2 L_{or}$$

Where $D^2 L_{or} = \{L'' : L \in L_{or}\}$.

BVP is an operator that, given a second derivative and two boundary conditions, returns the Lorenz curve.

$$\begin{cases} L''(p) = f(p) \quad \forall p \in (0,1) \\ L(0) = 0 \quad L(1) = 1 \end{cases}$$

The general solution is:

$$L(p) = c_1 + c_2 p + \int_0^1 \int_0^z f(t)dtdz$$

Adding the boundary conditions, leads to $c_1 = 0$ and $c_2 = 1 - \int_0^1 \int_0^z f(t)dtdz$, equivalent to $c_2 = 1 + \int_0^1 (z-1)f(z)dz$.

$$L(p) = p\{1 + \int_0^1 (z-1)f(z)dz\} + \int_0^1 \int_0^z f(t)dtdz$$

By integrating by parts:

$$L(p) = p + p\int_0^1 (z-1)f(z)dz + \int_0^1 (p-z)f(z)dz$$

And by straightforward calculation:

$$L(p) = p + (p-1)\int_0^p zf(z)dz + p\int_p^1 (z-1)f(z)dz$$

So the final map from $L''(p)$ and $L(p)$, is given by:

$$L(p) = p + (p-1)\int_0^p zL''(z)dz + p\int_p^1 (z-1)L''(z)dz$$

2.2.3 nclr transformation

Working directly on L_{or} or L^2_{or}, suffers from drawbacks that restrict their applicability and their interpretability as consequence of the fact that both are not vector spaces in $L^2_{[0,1]}$ [3]. In fact, D^2L_{or} is not a vector space because it contains only non-negative function.

The negative centred log-ratio transformation is a isomorphic mapping that takes a function in the embedded space D^2L_{or}, bringing it into the vector space $L^2_c = \{g : g \in L^2_{[0,1]}, \int g = 0\}$.

$$nclr : D^2L_{or} \to L^2_c : h \to -ln(h) + \int_0^1 ln(h(t))dt$$

$$nclr^{-1} : L^2_c \to D^2L_{or} : g \to exp(-g)/k_g$$

with $k_g = \int_0^1 \int_0^p exp(-g(z))dzdp$.

Defining k_g in this way, requires a strong assumption: $L'(0) = 0$. This condition is not always true but, working in a normalized domain, it can be reasonable.

The relation is summarized by:

$$D^2Lor \xrightarrow{nclr} L_c^2$$

$$D^2Lor \xleftarrow{nclr^{-1}} L_c^2$$

2.2.4 Conclusion

Combining the above maps, the relationship $\Psi(L) = nclr(L'')$ associates each Lorenz curve to an element in the Hilbert space L_c^2. The inverse map is:

$$\Psi^{-1}(g) = BVP(exp(-g)/k_g)$$

2.3 Centred log-ratio transformation

Density functions are particular cases of functional data, positive on a generic support T and with a integral unity constraint. For this type of function, the L^2 metric is not appropriate both because of the unitary constrain of the integral, and because of the positivity constrain all over the domain. In order to perform a statistical Functional Data Analysis on probability functions, it is necessary to apply an isomorphic mapping from the Bayes space to a standard L^2 space [7].

Considering a finite support $I = [a, b]$, on which the Lebesgue measure is applicable, a possible isomorphism is the centred log-ratio transformation (clr), which represents the simplest isomorphism to pass from a Bayes space to an L_2 space.

Taking a density function $f(x)$:

$$clr[f(x)] = f_c(x) = lnf(x) - \frac{1}{\eta}\int_I ln(f(x))dx \tag{2.3}$$

where $\eta = b - a$ is the length of the interval I. clr transformation is a one-to-one mapping, and so, the inverse clr transformation is obtained as

$$clr^{-1}[f_c(x)] = \frac{exp(f_c(x))}{\int_I exp(f_c(x))dx} \tag{2.4}$$

The constant denominator is made to guarantee the unity constraint of the integral to the resulting density. Functions (2.3) and (2.4) represent the functional version of the log-ratio transformation for multivariate compositional data [7].

A further constraint on the clr is given by:

$$\int_T clr[f(x)]dx = 0 \tag{2.5}$$

Working with a function evaluation on a discrete domain, the transformation is given by:

$$z_i = ln\frac{y_i}{g(y_1, ..., y_n)}$$

Where g indicates the geometric mean. In this way, the condition $\sum z_i = 0$, equivalent to $\int_T clr[f(x)] = 0$, is true.

While the inverse transformation to obtain the density estimation, is:

$$y_i = \exp(z_i)\frac{\sum z_i}{\sum \exp(z_i)}$$

Where the second term represents the correction on the unity integral constraint.

2.4 Functional Principal Component Analysis

Functional Principal Component Analysis (FPCA) identifies the complexity of data, in the sense of how many types of curves and characteristics are to be found.

In the fist step, we chose the principal component function $\epsilon_1(t)$ in order to maximize:

$$\frac{1}{N}\sum_i \int (\epsilon_1(t)x_i(t))^2 dt$$

Subject to $\int \epsilon_1(t)^2 dt$, the continuous analogue of the unite sum constraint. Step by step, the principal component function ϵ_n is also required to satisfy the orthogonality constraints $\int \epsilon_k(t)\epsilon_n(t)dt$ for $k < n$. Each principal component has to determine the most important model of variation in the curve, subject to orthogonal constraint to all the modes defined on previous steps [14].

Visualizing the result

interpreting the components is not always intuitive, and an entirely straightforward matter for most functional PCA problems. In particular working with embedded functions. A helpful method is to examine plots of the overall mean function and functions obtained by adding and subtracting a suitable multiple of each principal component function. This approach is called $mode\ of\ variation$.

Starting from functions $\hat{\Psi}_i(p), ..., \hat{\Psi}_n(p)$, the empirical mean $\hat{\Psi}_n(p)$, the covariance operator $\hat{\Sigma}_{\Psi,n}$ and so its eigenvalues and vectors $(\hat{\alpha}_{j,n}, \hat{v}_{j,n})$ are computed.

The j-th mode of variation is:

$$m_{j,n}(k,p) = \hat{\Psi}_n(p) + k * \hat{v}_{j,n}(p) * (\hat{\alpha}_{j,n})^{\frac{1}{2}}$$

The modes of variation are weight with respect to the corresponding eigenvalue. k is an integer value used to show how the functions deviate from the mean with respect to different scores related to the eigenvector.

Plotting principal component scores

An important aspect of PCA is the examination of the scores of each curve on each component. Interpreting the effects of the main components and plotting the scores, the composition of the functions is described both individually and grouping.

2.5 Functional regression

Functional regression can be computed in many ways to point out different features of data. 1-way functional ANOVA shows the significant differences in the structure of two or more functional data groups. A regression with functional response and scalar variables explores how much variability of the functional response is explainable by other variables. The coefficient $beta_i(t)$ shows how each variable influences the output and an easier interpretation is provided by plotting how a function deviates from its mean value when the variable values change [14].

2.5.1 1-way functional ANOVA

Data functions are divided into G groups and is tested if at least one group has a significant difference with respect to the others in a interval domain.
For the $m-th$ element of the $g-th$ group, the functional data is defined as:

$$y_{mg}(t) = \mu(t) + \alpha_g(t) + \epsilon_{mg}(t) \quad \forall t \in T$$

where:

- $\mu(t)$ is the mean function;
- $\alpha_g(t)$ is the deviation for the $g-th$ group;
- $\epsilon_{mg}(t)$ is the residual for the $m-th$ data;

To identify the $\alpha_g(t)$ uniquely, the following constrain is required:

$$\sum_g \alpha_g(t) = 0 \quad \forall t \in T$$

Defining a set of regression functions β_j such that $\beta_1 = \mu, ..., \beta_{G+1} = \alpha_G$, the model has an equivalent formulation:

$$y_{mg}(t) = \sum_{j=1}^{G+1} z_{mg,j} * \beta_j(t) + \epsilon_{mg}(t) \tag{2.6}$$

And a matrix formulation:

$$\mathbf{y} = \mathbf{Z}\beta + \boldsymbol{\epsilon}$$

Parameters β are chosen to minimize the residual sum of square:

$$\text{LMDRR}(\beta) = \sum_{g}^{G} \sum_{m}^{N_g} \int [y_{mg}(t) - \sum_{j}^{q} z_{mg,j}\beta_j(t)]^2 dt \tag{2.7}$$

Subject to the constrain $\sum_2^G \beta_j = 0$, gives the least square estimated $\hat{\beta}$.

2.5.2 Functional response with scalar independent variable

The structure of the regression is the same of the previous case:

$$y = \mathbf{Z}\beta + \boldsymbol{\epsilon}$$

The $\beta_0(t)$ represent the main function overall the data function. $\beta_i(t)$ are the functional parameters related to every variable contained in the model. The least squares estimation of the beta parameters $\hat{\beta}$ is obtain to minimise (2.7). A constrain of the form $\sum_2^G \beta_j = 0$ is still valid only if there are variables related to groups to avoid over-fitting. Otherwise it is not valid anymore.

2.5.3 Statistical test on functional regression parameters

To evaluate the overall significance of the function regression model, it is used the standard F-Test:

$$\begin{cases} H_0 : \beta_g(t) = 0 \quad \forall g \in \{1, ..., G\} \quad \forall t \in T \\ H_1 : \beta_g(t) \neq 0 \quad \text{for some} \quad g \in \{1, ..., G\} \quad \text{and some} \quad t \in T \end{cases}$$

A possible statistic test to use is the following:

$$T_F = \int_I \sum_{g=1}^{G} \left(\frac{\hat{\beta}_g(t)}{\hat{se}(\hat{\beta_g})} \right)^2 dt$$

Meanwhile the test on single functional parameter β_g is the functional version of the classical t-test hypotheses:

$$\begin{cases} H_0 : \beta_g(t) = 0 \quad \forall t \in T \\ H_1 : \beta_g(t) \neq 0 \quad \text{for some} \quad t \in T \end{cases}$$

And the following statistical test is used:

$$T_g = \int_I \left(\frac{\hat{\beta}_g(t)}{\hat{se}(\hat{\beta_g})} \right)^2 dt$$

Where $\hat{\beta}_g$ is the OLS estimate of β_g and $\hat{se}(\hat{\beta_g})$ its estimated standard error.

Both tests are computed directly by the Interval-Wise Testing procedure. These tests can be fitted not only to evaluate the significance of the model or a parameter on the whole domain, but also to select significant intervals of the domain taking $I \subset T$.

2.6 Interval-Wise testing

In this section the Interval-wise Testing procedure is described. It is an inferential procedure for functional data, able to select the intervals of the domain imputable of rejecting a functional null hypothesis [10]. In the exploratory analysis, IWT (Interval-Wise Testing) is applied to to select intervals of domain where there exists a significant difference between factors.

2.6.1 Interval-Wise Testing Procedure

The Interval-Wise Testing procedure (IWT) is a purely nonparametric inferential method able to detect the portion of the domain in which a significant difference between functions is guaranteed by the rejection of a null hypothesis test. Since the first approach relies on a basis expansion of data, conclusions could change depending on the basis chosen to project data. In the second case the choice of the initial partition in sub-intervals can affect the conclusions. Instead the IWT is purely non-parametric and it can detect the portion of the domain with a rejection of the null hypothesis for functional data embedded in L^2.

The procedure

- **Interval-wise testing:** For any interval $I \subset T$ a p-value functional test is performed.

- **Definition of the p-value functions:** An unadjusted and an adjusted p-value function are defined.

- **Domain selection:** The domain selection is achieved by thresholding the unadjusted or the adjusted p-value function.

The first step of the procedure involves any functional test of a null hypothesis H_0 against an alternative hypothesis H_1.

In the second step, using the family of interval-wise tests defined in the first point, two definitions of p-value functions are introduced. The p-value p^I is obtained by performing a test on every interval $I \subset T = (a, b) \subset \mathbb{R}$. The unadjusted p-value function $p(t)$ is defined as the superior limit of the p-values p^I when both extremes of I converge to t. The adjusted p-value function $\tilde{p}(t)$ is defined as the supremum over all p-values $p(t)$ with $t \in I$.

$$p(t) = \limsup_{I \to t} p^I \quad \forall \quad t \in T$$

$$\tilde{p}(t) = \sup_{t \in I} p^I \quad \forall \quad t \in T$$

The third step is the domain selection. The domain is selected by thresholding the p-values functions. H_0 is rejected with level α if $p(t) < \alpha$ ($\tilde{p}(t) < \alpha$ considering the adjusted p-value), and the selected domain is composed by every $t \in T$ for which $p(t) < \alpha$.

Properties

- The unadjusted p-value function $p(t)$ provides a control of the pointwise error rate. $\forall\ \alpha \in (0, 1)$:

$$\forall t \in T \quad \text{s.t.} \quad \exists I \ni t : H_0^I \quad \text{is} \quad \text{true} \Rightarrow \mathbb{P}[p(t) \leq \alpha] \leq \alpha$$

- The adjusted p-value function $\tilde{p}(t)$ provides a control of the interval-wise error rate(IWER).$\forall\ \alpha \in (0, 1)$:

$$\forall I \subseteq T \quad \text{s.t.} \quad H_0^I \quad \text{is} \quad \text{true} \Rightarrow \mathbb{P}[\forall t \in I, \tilde{p}(t) \leq \alpha] \leq \alpha$$

- The unadjusted p-value function $p(t)$ and the adjusted p-value function $\tilde{p}(t)$ are consistent. $\forall\ \alpha \in (0, 1)$:

$$\forall t \in T \quad \text{s.t.} \quad \nexists I \ni t : H_0^I \quad \text{is} \quad \text{true} \Rightarrow \mathbb{P}[p(t) \leq \alpha] \xrightarrow{n \to \infty} 1$$

$$\forall I \subseteq T \quad \text{s.t} \quad \nexists J \subseteq I : H_0^J \quad \text{is} \quad \text{true} \Rightarrow \mathbb{P}[\forall t \in I, \tilde{p}(t) \leq \alpha] \xrightarrow{n \to \infty} 1$$

The unadjusted p-value functions detect the **pointwise** probability of type-1 error, i.e. the false rejections. Meanwhile the adjusted p-value function detects the **interval-wise** probability of type-I error. The adjusted p-value function is to be preferred for inferential purposes. Indeed, the unadjusted p-value function may lead to too many false rejections since there can be a high correlation between the pointwise p-values.

2.6.2 Interval-Wise Testing for functional regression models

In this section, the domain selection IWT procedure is extended to functional-on-scalar linear models [1]. The result of the procedure is an adjusted p-value function for testing:

$$\begin{cases} H^t_{0,C} : C\boldsymbol{\beta}(t) = \mathbf{c}_0(t) \\ H^t_{1,C} : C\boldsymbol{\beta}(t) \neq \mathbf{c}_0(t) \end{cases} \tag{2.8}$$

The adjusted p-value function can be thresholded at level α to select the portions of the domain imputable for the rejection of the null hypothesis $H_{0,C}$. This type of control guarantees that, for every interval of the domain where $H_{0,C}$ is true, the probability that $H_{0,C}$ is rejected at least at one point of the interval is less or equal to α. It is important to notice that (2.8) is the general case of the functional version of the classical F-test hypotheses for the overall model and of the functional version of the classical t-test hypotheses for the significance of single β parameter. Both tests can be obtained through a specific choice of C and $\mathbf{c}_0$.

The domain selection procedure, proposed by [1], is based on three steps.

Interval-wise testing

Given any closed interval $I \subset T$, performing:

$$\begin{cases} H^I_{0,C} : C\boldsymbol{\beta}(t) = \mathbf{c}_0(t) \quad \forall t \in I \\ H^I_{1,C} : C\boldsymbol{\beta}(t) \neq \mathbf{c}_0(t) \quad \text{for some } t \in I \end{cases} \tag{2.9}$$

The linear hypotheses for the overall model and the gth functional regression parameter on I are respectively:

$$\begin{cases} H^I_{0,F} : \beta_g(t) = 0 \quad \forall g \in \{1, ..., G\} \quad \forall t \in I \\ H^I_{1,F} : \beta_g(t) \neq 0 \quad \text{for some } g \in \{1, ..., G\} \text{ and some } t \in I \end{cases} \tag{2.10}$$

$$\begin{cases} H^I_{0,g} : \beta_g(t) = 0 \quad \forall t \in I \\ H^I_{1,g} : \beta_g(t) \neq 0 \quad \text{for some } t \in I \end{cases} \tag{2.11}$$

p-value function

In order to identify significant intervals in the domain, adjusted p-value functions are used. Denoting p_C^I as the p-value of the test (2.9), the adjusted p-value function $\tilde{p}_C(t)$ at point t for testing general linear hypothesis with constraint C, is defined as the supremum p value of all interval wise tests on intervals containing t, as follows:

$$\tilde{p}_C(t) = \sup_{I:t\in I} p_C^I \quad t \in [a, b]$$

Analogously, denoting by , and the p-values from testing (2.10) and (2.11) , the adjusted p-value functions for testing hypotheses on the overall model and on the gth functional parameter are defined as:

$$\tilde{p}_F(t) = \sup_{I:t\in I} p_F^I \quad \tilde{p}_g(t) = \sup_{I:t\in I} p_g^I \quad t \in [a, b]$$

Domain selection

The intervals of the domain presenting a rejection of any of the null hypotheses are obtained by thresholding the corresponding adjusted p-value functions at level α. The intervals presenting at least one significant effect by thresholding $\tilde{p}_F(t)$, are selected. For the single variable β_g, are selected the intervals presenting a significant effect of the $g-th$ covariate by thresholding $\tilde{p}_g(t)$. The introduced domain selection procedure is provided with a (asymptotic) control of the IWER. This type of control implies that the probability of detecting false positive intervals is (asymptotically) controlled at level α.

2.7 Clustering

Clustering is a set of data analysis techniques, aimed to select and group homogeneous elements in such a way that objects in the same group are more similar, in some sense, to each other than to those in other groups. Clustering techniques are based on measures related to the similarity between the elements. Hierarchical clustering requires the definition of a linkage criteria to determine the distance between sets of observations as a function of the pairwise distances between observations.

2.7.1 Hierarchical clustering

Hierarchical clustering is a clustering technique able to produce a hierarchy of nested clusters, representing it with a dendrogram.

The aim is to group data in clusters based on a selected definition of distance and linkage. The agglomerative approach of hierarchical clustering, first, groups together the two units that are most similar, and then the most similar groups or units in an iterative way. This approach exploits a bottom up strategy to join different

units and clusters together. In order to decide which clusters should be combined, a measure of dissimilarity between units and a method to compute dissimilarity between clusters need to be defined. This is usually achieved by a metric and a linkage.

Metric

The metric is used to build the distance matrix between the functional data.
A distance function $d(f, g)$ is defined as the integral on the overall domain of the square difference function:

$$d(f, g) = \int_T (f(t) - g(t))^2 dt$$

Linkage

The Linkage criterion defines the distance between sets as a function of distances between elements in the sets.
Given Q the number of units x_i and $D = \{d_{ij}\}_{i,j \in \{1,\dots,Q\}}$ the dissimilarity matrix, two disjoint sets of units, $U = \{x_i\}$ and $V = \{x_j\}$, are taken. The dissimilarity $d(U, V)$ can be defined with different linkage criteria:

- **Single linkage**: $d(U, V) = \min\{d_{ij} : x_i \in U, x_j \in V\}$
- **Complete linkage**: $d(U, V) = \max\{d_{ij} : x_i \in U, x_j \in V\}$
- **Average linkage**: $d(U, V) = \frac{1}{|U||V|} \sum_{i:x_i \in U} \sum_{j:x_j \in V} d_{ij}$
- **Ward linkage**: $d(U, V) = SSE(U \cup V) - (SSE(U) + SSE(V))$
  Where SSE is the sum of squares in clusters.

Clustering evaluation

A common way to evaluate the results of a clustering analysis is to look at the dendrogram. A dendrogram is a diagram which shows the distance which the clusters have been merged at. The clusters can be obtained from just thresholding the dendrogram at a certain height. It is important to specify that this is a choice of the statistician and not an outcome of the algorithm.

A way to measure how much the dendrogram represents the clustering structure is the Cophenetic Correlation Coefficient (CPCC). In the clustering framework, the distance matrix is a matrix containing the cophenetic distances between statistical units. This distance between two units corresponds to the height of the dendrogram where the two branches that include the two objects, merge into a single branch.

The CPCC compares the two distances, the original distance given by the dissimilarity matrix D and the cophenetic distance given by the cophenetic matrix C, by computing the correlation between the two matrices:

$$CP\,CC = cor(D, C)$$

A higher value means that the two structures are more similar, i.e. the original distances between units and the way the dendrogram links them. At this point different types of linkage can be used and the one, achieving the maximum value of the CPCC coefficient, is usually selected.

Pseudo-Algorithm

1. Let each data point be a cluster
2. Compute the dissimilarity matrix
3. Repeat
    - Merge the two closet clusters
    - Update the distance matrix
4. Until only a single cluster remains

Chapter 3

Data and pre-processing

In this chapter, we make an overview on data describing inequality [17], and the process to obtain functional objects. Lorenz curves are computed directly from the quantiles of income distribution. Density functions can be obtained from Lorenz curves through a bijection [6].
Then we introduce the variables used to explain the inequality variability [15], and the imputation methods to deal with missing values. Due to the high correlation between variables, we compute the Principal Component Analysis (PCA) in order to make a regression with fewer independent variables.

3.1 Inequality dataset

The data used in this work to describe the inequality are from the World Income Inequality Database (WIID), provided by the United Nations University. This dataset combines information coming from many sources, including Eurostat, The World Bank and The Organisation for Economic Co-operation and Development (OECD).

The inequality values are about the income explained by population groups. The values $d_1 - d_{10}$ indicate the percentage of a given resource sharing by the 10%'s of population, from the poorest to the richest with respect to a resource.

There are two main types of resources treated: Income and Consumption.

The income definition is the one recommended by the Canberra Group and considers of different items as:

- Employee income
- Income from self-employment
- Income less expenses from rentals
- Property income
- Current transfers received

The consumption definition follows the guidelines of "Living Standard Measurement Study" by Deaton and Zaidi. The items considered for the welfare measurement are:

- Food consumption
- Non-Food consumption
- Durable goods
- Housing

More information about the compositions of the items, is contained in [17].

For each country, we extract the most recent available d1-d10 values related to the income inequality, with some extra information as Gini index, Gross Domestic Product (GDP), continent, year of the sampling and Gross Domestic Product per capita (GDP PPP). This type of data are very difficult to be sampled, especially in poorest countries. The lack of information prevents the constructions of data time series, and so the impossibility to analyse the time variation of inequality inside the countries. For the least developed countries, where the rural agriculture is large and it is difficult to gather accurate income data, any information about the income inequality is not available, not even for remote years. In these cases, consumption inequality is used [17]. This replacement could be a good approximation of the real values by observing that the differences between the income and consumption data, are not so relevant in the countries with both observations for the same years. However, literature about the relationship between income and consumption, has established that the consumption is smoother and less variable than income [17]. In the following analyses, a corrective factor between the two types of inequality will be introduced.

Raw data objects consist in a vector of 10 evaluations $d_1 - d_{10}$, one for every country in the dataset, related to different years. The sampling is about income inequality if it is available, otherwise about consumption. For the exploitative analysis only data about income inequality are used. In functional regression all data available are used and an explanatory variable is introduced to distinguish the income inequality and the consumption inequality.

3.2 Lorenz curve constriction

The Lorenz curve is a graphical tool to visualize and analyse the inequality distribution related to the income, with respect to the cumulative distribution.
It shows p_2% of income explained by the poorest p_1% of population.

Properties

- x-axis and y-axis are percentage. So both p and L(p) are defined over [0,1]

$$L : [0, 1] \rightarrow [0, 1]$$

- L(p) is increasing
  The value of L(p) grows as p increases. The income owned by an increasingly population group, that incorporates the previous one, increases.

- L(p) is convex
  Increasing p, time by time, a section of population, with a higher income respect to the ones already considered, is added. This implies that the slope of the curve must be increasing.

- L(0) = 0
  The 0% of the population owns the 0% of the income.

- L(1) = 1
  The whole population owns all the income.

The Lorenz Curve has two extreme cases:

- LC as the egalitarian line
  Every single person has the same income as any other person in the population. So, the contribution to growth of cumulative income, is equal and constant. LC is a line passing for $(0, 0)$ and $(1, 1)$

- LC as the line of perfect inequality
  This is the case in which all the income is owned by one single person. LC remains 0 adding every person to the cumulative population and passes from 0 to 1 when the person, owning all the income, is added to the cumulative population at the end.

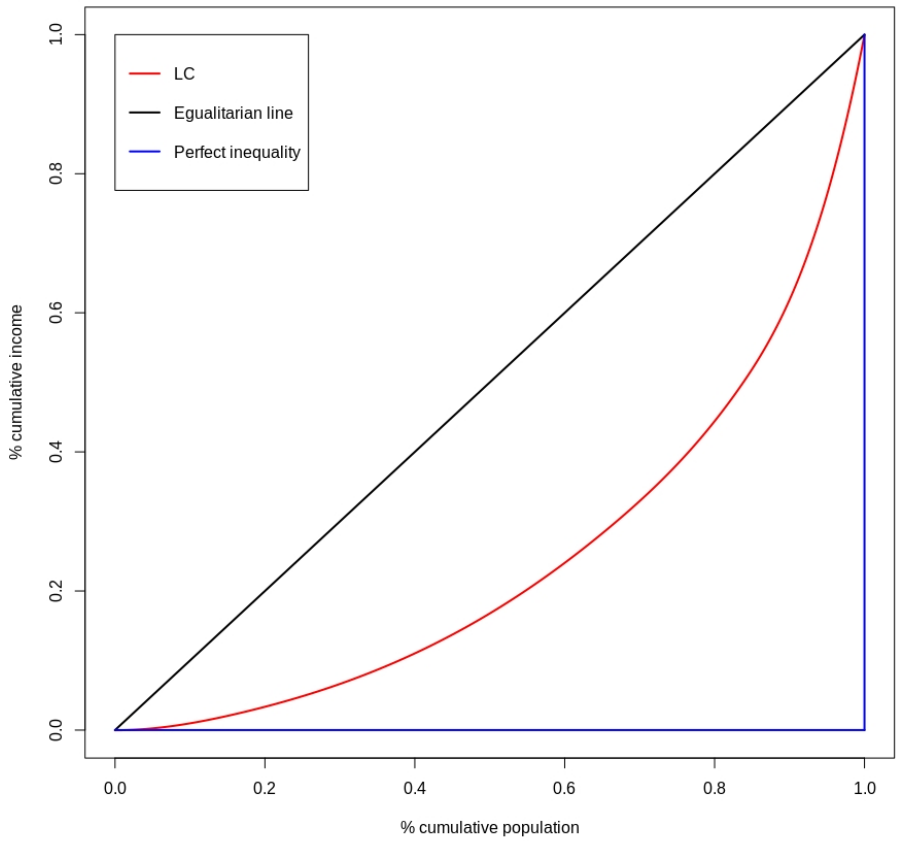


Figure 3.1: Lorenz Curve & borderline cases

An interpretative Lorenz curve representation is obtained as :

$$L(p) = \frac{\int_0^p q(t)dt}{\int_0^1 q(t)dt}$$

The numerator sums the income of the poorest p percentage of population. The denominator sums the incomes of all the population.

Construction

The available data are the portion of income coming from each 10% of population, increasing from the poorest one to the richest one.

d1	1.8%
d2	4.5%
d3	5.8%
d4	7%
d5	8.3%
d6	9.6%
d7	10.9%
d8	12.6%
d9	15.1%
d10	24.4%

Table 3.1: Portion of income explained for Italy in 2016

From these, the quantiles $q_1 - q_{10}$ can be obtained summarize the d_i step by step:

$$q_i = \sum_{j=1}^{i} d_j$$

With the extra information $q_0 = 0$.

q0	0%
q1	1.8%
q2	6.3%
q3	12.1%
q4	19.1%
q5	27.4%
q6	37%
q7	47.9%
q8	60.5%
q9	75.6%
q10	100%

Table 3.2: Quantiles for Italy in 2016

Quantiles represent the percentage of income owned by different groups of cumulative population. These points are used to build the Lorenz curves.

Lorenz Curves are functional data, defined as a combination of cubic b-spline basis, with modified coefficients to ensure the monotonicity and convexity constraints. Roughness penalty term is not applied to guarantee a perfect fit on the quantiles.

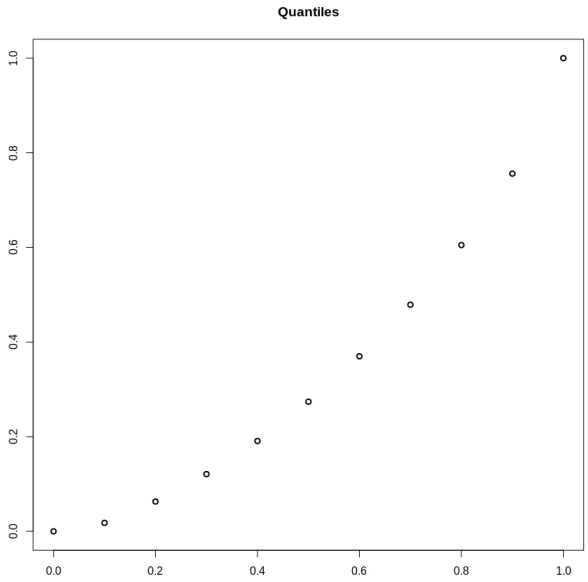


Figure 3.2: Quantiles plot for Italy in 2016

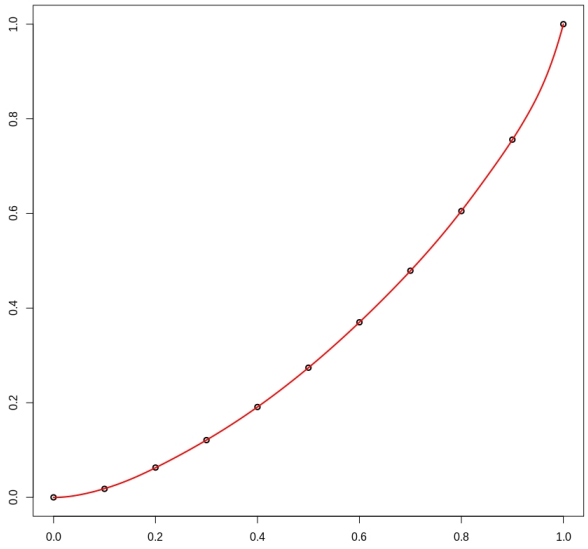


Figure 3.3: Lorenz Curve plot for Italy in 2016

A way to understand how much these Lorenz curves truthfully reflect the inequality of the respective country, is to calculate the Gini index from them and compare it with the real value, built on the entire population and available in the dataset.

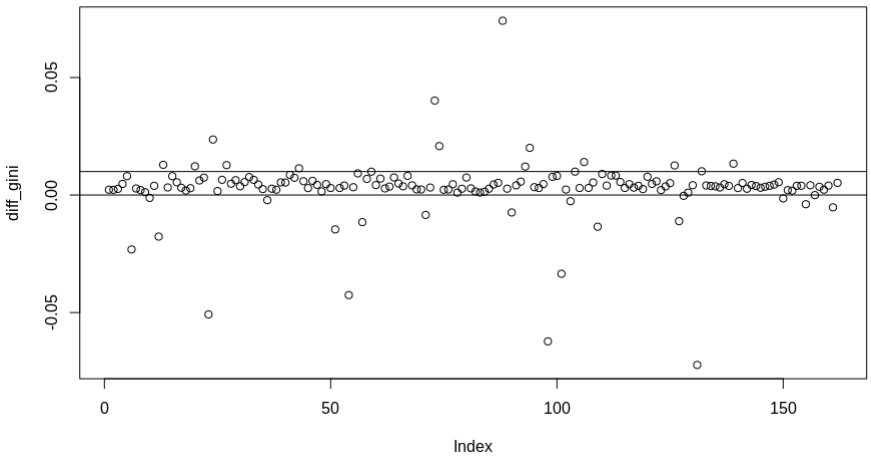


Figure 3.4: Real Gini minus estimated Gini

$$\text{Diff} = \text{Gini}_{Real} - \text{Gini}_{LC}$$

All the estimated Gini indices do not differ so much from the real values. Most of them differ in the order of 1% and only a few rare cases in the order of 5%. This is a very satisfactory result.

Notice that all the curves that produce an estimate of the Gini index very close to the real value, tend to underestimate it. In fact, most of the differences in the order of 0-1% are positive.

3.3 Density function

Density function $f(x)$ shows the income distribution in the population. Supposing $F(x)$, the cumulative distribution function, continuous and differentiable, there exists a density function defined as:

$$f(x) = F'(x)$$

A Normalization is computed on the domain of the density function $f : [0, 1] \rightarrow \mathbb{R}^+$. The values of domain 0 and 1 correspond respectively to the lowest and highest income available in each country.

Density function $f(x)$, defined on an interval T, has 2 main properties to satisfy:

- $f(x) \geq 0 \quad \forall x \in T$
- $\int_T f(x)dx = 1$

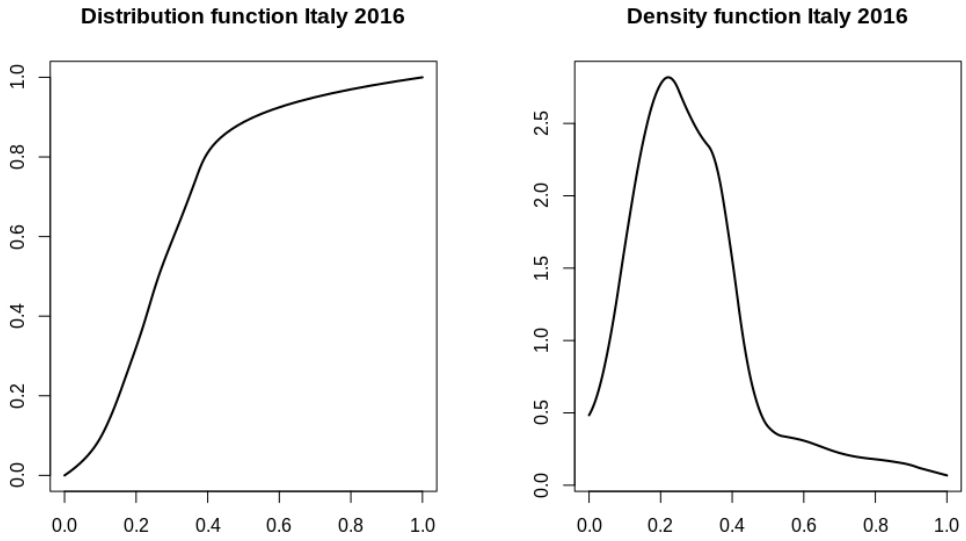


Figure 3.5: Cumulative Distribution Function & Probability Density Function

Relationship between Lorenz curve and Density function

Theorem 3.1. *Suppose L(p) is defined and continuous on [0,1], with second derivative L″(p). The function L(p) is a Lorenz curve iff L(0)=0, L(1)=1, L′(0⁺) ≥ 0 and L″(p) ≥ 0 ∀ p ∈ (0,1).*

Take the definition of a Lorenz curve and express it first with the standard representation, and then with a change of variables:

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

$$\mu * L(F(x)) = \int_0^x y * dF(y)$$

Deriving one time:

$$\mu * L'(F(x)) * f(x) = x * f(x)$$

Simplifying by $f(x)$ and taking the derivative:

$$\mu * L''(F(x)) * f(x) = 1$$

Theorem 3.2. *If L″(p) exists and it is positive everywhere in an interval* (x_1, x_2)*, then* F_x *has a finite positive density in the interval* $(\mu L'(x_1^+), \mu L'(x_2^-))$ *given by:*

$$f(x) = \frac{1}{\mu L''(F(x))}$$

3.4 Variables

It is difficult to understand which factors can be considered significant to explain income inequality in a functional regression. Generally, in regression analyses, inequality is expressed as a scalar index, and not as a function. However, it is reasonable to expect that the factors that affect income inequality when it is represented by a single parameter, do the same, in a different way, even on an infinite-dimensional representation. The factors affecting inequality, are related to education, innovation technology, public spending at different levels, political situation and quality of life [15].

1	Primary Education Rate
2	Secondary Education Rate
3	Tertiary Education Rate
4	TFP
5	Public spending on education
6	Public spending on health
7	Urban population rate
8	HDI

Table 3.3: List of variables

Primary, Secondary and tertiary education rate represent the gross percentage of people attending schools, for each of the 3 levels. Gross percentage implies that there exists the possibility of percentages greater than 100 %, because it indicates the rate between all the ones who attend a given degree of education, and the number of people in age to attend it.

TFP is the Total Factor Productivity. It can be defined as the residual part of output, i.e. GDP, exceeding labour and capital inputs. Total factor productivity measures residual growth in total output of a national economy that cannot be explained by the accumulation of traditional inputs such as labour and capital. The method to calculate TFP gives it a meaning that can be interpreted only as variation over the years. So, it makes sense to consider it only in studies that take cares of the variation of inequality over time. Furthermore, this index is only available for some countries, generally the more developed ones. TFP is often considered the primary contribution to GDP Growth Rate, so a possible solution is to replace TFP with the variation of GDP PPP per capita. Not dealing with time series, this approach results useless, but suggests working with another variable: GDP PPP per capita. In support of this thesis, [8] show that exists a relationship between GDP per capita and income per capita and so GDP PPP per capita may also influence the income distribution.

The public spending on education and health is the percentage of spending for the corresponding item with respect to the GDP.

Urban population rate is the percentage of people living in the city area.

HDI is the Human Development Index. It is an index that statistically contains multiple items related to the welfare of a country as life expectancy, quality of life, efficiency of the education system and others. Its value belongs to a range from 0 to 1 where values near to 0 represent a very low standard of living and values near to 1 represent a high standard of living. As composite index, there exists the possibility of collinearity with variables already present in the dataset and, at the same time, used to build this index. The three main items used in the calculation of the HDI index, are the Gross national income (GNI), the Mean years of schooling and Expected years of schooling and the life expectancy at birth. The first two are redundant with variables already available in the dataset. Life expectancy will be added to the variables.

No index about political situation is taken because all the ones available are superficially built.

3.5 Implementation for variable values

As previously reported in section 2, the variables considered to compute the regression, are the following:

- Primary Education Rate
- Secondary Education Rate
- Tertiary Education Rate
- Public spending on education
- Public spending on health
- GDP PPP per capita
- Urban population rate
- Life expectancy

Information related to GDP PPP per capita is provided directly in the inequality dataset. For the other variables, the values come from the World Bank datasets. All the World Bank datasets have the same structure. Given a variable, the values are in a grid table, in which the rows represent the different states and the columns the different years in which the sampling is detected. Datasets have a lot of missing values, especially dealing with data coming from underdeveloped counties and with

remote years.

A first method to solve the problem is to implement the missing values by exploring other datasets. This method is used every time it is possible.
There are two mainly sources used:

- UIS Unesco Dataset
- Knoema Dataset

The second method consists in imputation. Wherever there is a missing value for a given variable and a given country, all the values over time are used as a time series. In each time series, the number and the distribution of data not available is different, so it is not possible to apply a universal computation of missing values because this may not be realistic.

Two different computation methods are used.

Back Fill Forward Fill methods

The missing value to compute has a 1-step or a 2-step known value away.

2006	2007	2008	2009	2010	2011	2012	2013	2014
-	-	-	-	-	-	-	0.84	-

2006	2007	2008	2009	2010	2011	2012	2013	2014
0.15	-	-	-	-	-	0.36	-	-

Back fill method assigns a known value to the previous missing values.
Forward fill method assigns a known value to the subsequent missing values.

2006	2007	2008	2009	2010	2011	2012	2013	2014
-	-	-	-	-	0.84	-	0.84	-

Table 3.4: Back fill method

2006	2007	2008	2009	2010	2011	2012	2013	2014
0.15	0.15	-	-	-	-	0.36	-	-

Table 3.5: Forward fill method

These methods are used only if, around the missing one, only one value is known.

Weighted average

The weighted average method is used if there are both previous and subsequent values in the neighbourhood of the missing value to compute.

2006	2007	2008	2009	2010	2011	2012	2013	2014
0.15	-	-	-	-	-	0.36	-	-

2006	2007	2008	2009	2010	2011	2012	2013	2014
0.15	-	-	-	-	0.25	-	-	0.66

The weights applied to the average are inversely proportional to the distance from the missing value. If two points have the same distance, the weighted average is equal to an arithmetic average. If one of the points has a distance twice with respect to the other, the weights are respectively $\frac{1}{3}$ for the further value and $\frac{2}{3}$ for the closer one, and so on for all the possible cases.

$$Value_1(2009) = \frac{0.15 + 0.36}{2} = 0.225$$

$$Value_2(2013) = \frac{1}{3} * 0.25 + \frac{2}{3} * 0.66 = 0.523$$

2006	2007	2008	2009	2010	2011	2012	2013	2014
0.15	-	-	0.255	-	-	0.36	-	-

Table 3.6: known values with same distance

2006	2007	2008	2009	2010	2011	2012	2013	2014
0.15	-	-	-	-	0.25	-	0.523	0.66

Table 3.7: known values with different distance

Not computing

In all other cases, no type of assignment is made to the missing values, and these remain unknown.

3.6 Variables PCA

Principal Component Analysis (PCA) is a statistical tool, using an orthogonal transformation to convert a set of possibly correlated variables into a set of linearly uncorrelated variables, called principal components. This transformation is defined to compute the first principal component with the largest variance possible, and so on, under the constraint that each principal component is orthogonal to the previous ones. To obtain consistent results, the original data have to be centred in 0 and scaled with respect to the variance. In this thesis, Principal Component Analysis is performed to avoid correlated regressors in the model. Correlation between variables is evaluated by Pearson coefficients.

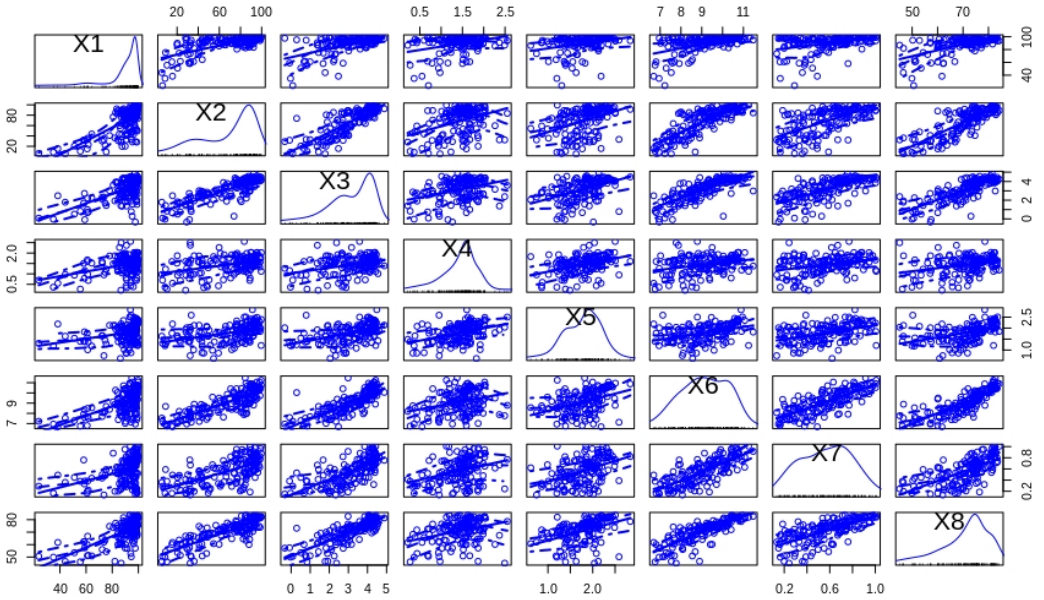


Figure 3.6: Variables scatter plot

-	x1	x2	x3	x4	x5	x6	x7	x8
x1	1	0.68	0.58	0.34	0.29	0.54	0.36	0.65
x2	-	1	0.84	0.38	0.43	0.81	0.62	0.83
x3	-	-	1	0.34	0.46	0.79	0.69	0.81
x4	-	-	-	1	0.4	0.29	0.33	0.32
x5	-	-	-	-	1	0.40	0.44	0.43
x6	-	-	-	-	-	1	0.77	0.82
x7	-	-	-	-	-	-	1	0.67
x8	-	-	-	-	-	-	-	1

Table 3.8: Pearson Correlation matrix

Considering the meaning of each original variables, they are all indices of a welfare measure, so all positive Pearson coefficients are coherent with that we expect. We also observe that correlation coefficients related to x4 (education spending) and x5 (health spending) are lower. This is since, being percentages weighed on the GDP PPP, they are relative and not absolute welfare measures.

When we compute a regression model with high correlated variables, the coefficients are unstable and difficult to interpret. So we will use the Principal Component Regression (PCR) where the regressors are the principal components.

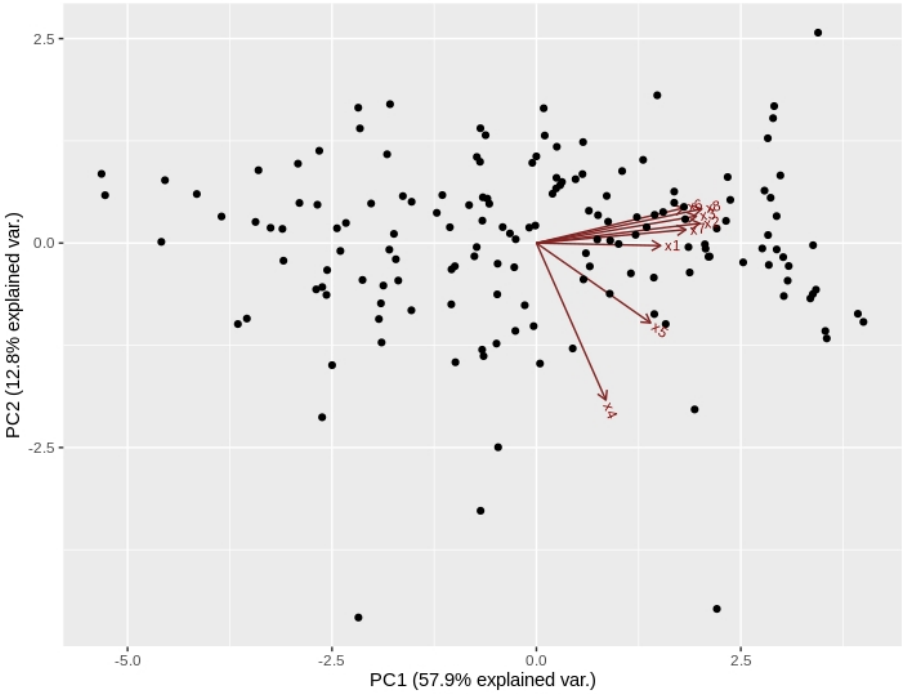


Figure 3.7: First and second Principal components

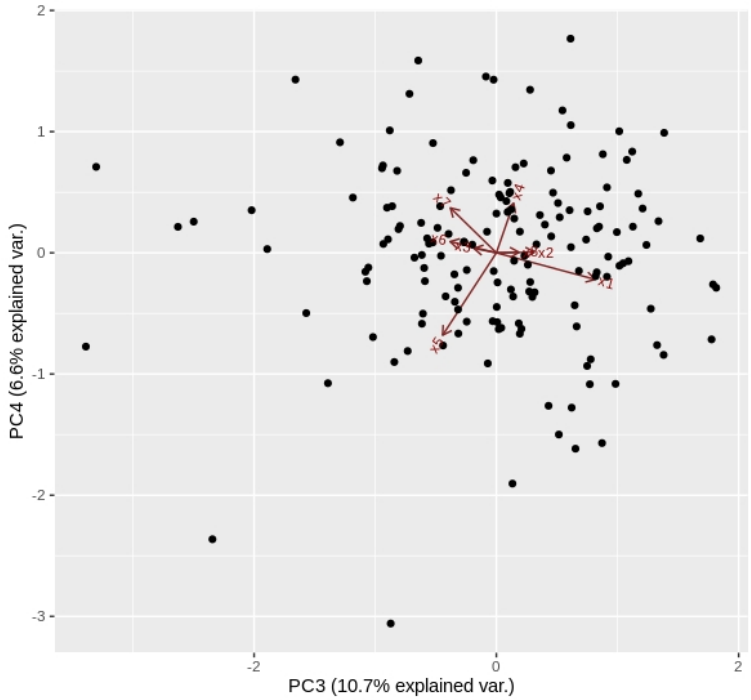


Figure 3.8: Third and fourth Principal components

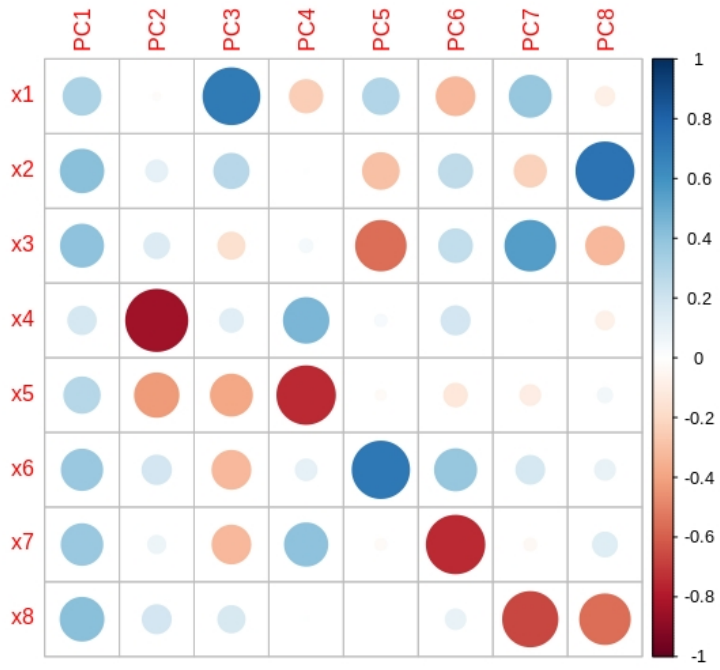


Figure 3.9: PCA composition

The first component explains the 57.9 % of the data variability, the second one 12.8 %, the third 10.7 % and the fourth 6.6 %.
Positive values for the first component correspond to projections with positive values for all the original variables, and vice versa for negative values. We can easily interpret its meaning. It is an index for wealth and quality of life. Second component is strongly influenced by health and education spending. It indicates how much countries spend in their social policy with respect to GDP PPP, independently by the fact that they are developed nations or not. Give an interpretation to the third component is not easy. It seems to divide counties with low standards of living but with a high primary and secondary education rates, from countries where education is not growing up as standard of life.

Chapter 4

Results

In this chapter, we present the results obtained from the analyses on functional data. We have to map the data in a Hilbert space to make the analyses consistent. After that, the results are mapped into the originals spaces to be interpreted.

In the exploitative analyses, we compute regression model with functional response and categorical independent variables to identify the structure of income inequality between groups. There are many interesting patterns which this approach can be used on. In this thesis, continents and GDP PPP per capita range are chosen. Functional Principal Component Analysis is a specific tool of functional data and detects the intervals where the data variability is nested. Functional regression identifies the most significant variables in explaining inequality and show how their variations affect the functional inequality profile. The aim of clustering analysis is to group countries with similar profiles for each inequality representation, and to plot them on the World map.

4.1 Analysis of Lorenz Curves

The Lorenz curves are embedded functions, so we have to work with πclr transformations of their second derivative (4.1), belonging to a Hilbert spaces, to make consistent analyses. As consequence of the smoothing process with a b-spline basis, the functions present regular oscillations on the abscissa. In particular, the first oscillation is more emphasized as condition $L'(0) = 0$ is required. However, even if the oscillation positions and amplitudes are due to the choice of the base and the number of knots, we can't interpret them as noise, because indicate the change of slope necessary to fit, as good as possible, the quantiles in the Lorenz curves.

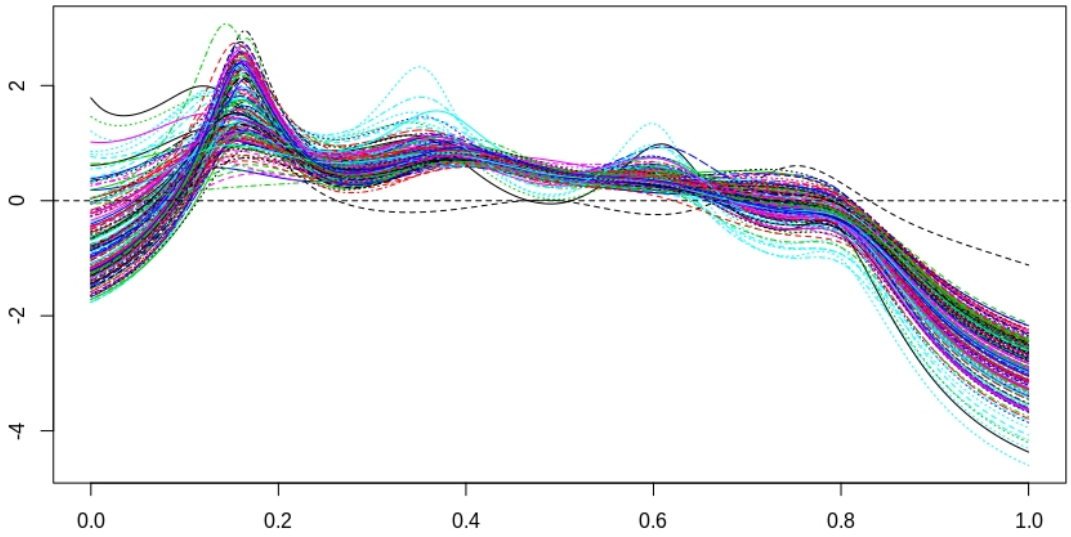


Figure 4.1: Lorenz curves second derivative transformations

4.1.1 Exploratory analyses

Computing exploitative analyses on Lorenz curves, we visualize how inequality is nested in different ways in population strata for each group. Interval-Wise Testing procedure detects also the intervals where there exists significant difference between the groups.

The choice of using continents and GDP PPP per capita range, as features, is interesting because we can give to the inequality an interpretation with a geographical and economic meanings.

Continents

Functional ANOVA requires that every continent is identified by a categorical variable. To avoid over-fitting, the design matrix is defined as follow:

- Europe : $x_1 = 1, x_2 = 0, x_3 = 0, x_4 = 0$
- Africa : $x_1 = 0, x_2 = 1, x_3 = 0, x_4 = 0$
- America : $x_1 = 0, x_2 = 0, x_3 = 1, x_4 = 0$
- Asia : $x_1 = 0, x_2 = 0, x_3 = 0, x_4 = 1$
- Oceania : $x_1 = -1, x_2 = -1, x_3 = -1, x_4 = -1$

The summary of the functional regression model is:

```
$call
IWTlmFoF(formula = y ~ x1 + x2 + x3 + x4)

$ttest
            Minimum p-value
(Intercept)           0.000 ***
x1                    0.000 ***
x2                    0.000 ***
x3                    0.774
x4                    0.113

$R2
              Range of functional R-squared
Min R-squared                    0.01723829
Max R-squared                    0.67798616

$ftest
  Minimum p-value
1               0 ***
```

Figure 4.2: Continents model summary

The p-value of the F-test is low enough (approximated to 0) to assert that there exists an interval on domain where the continent factor is significant. By consulting the summaries in the Appendix B, we have a better representation of the different inequality profile for all the continent combinations. European coefficient is significant different from all the others. The same for the African one. On the other way, there no exists significance in considering the American, Asian and Oceanic coefficients different from each other.

Computing the $\beta_i(t)$ coefficients, we estimate the curves for each continent and then we remap them as Lorenz curves.

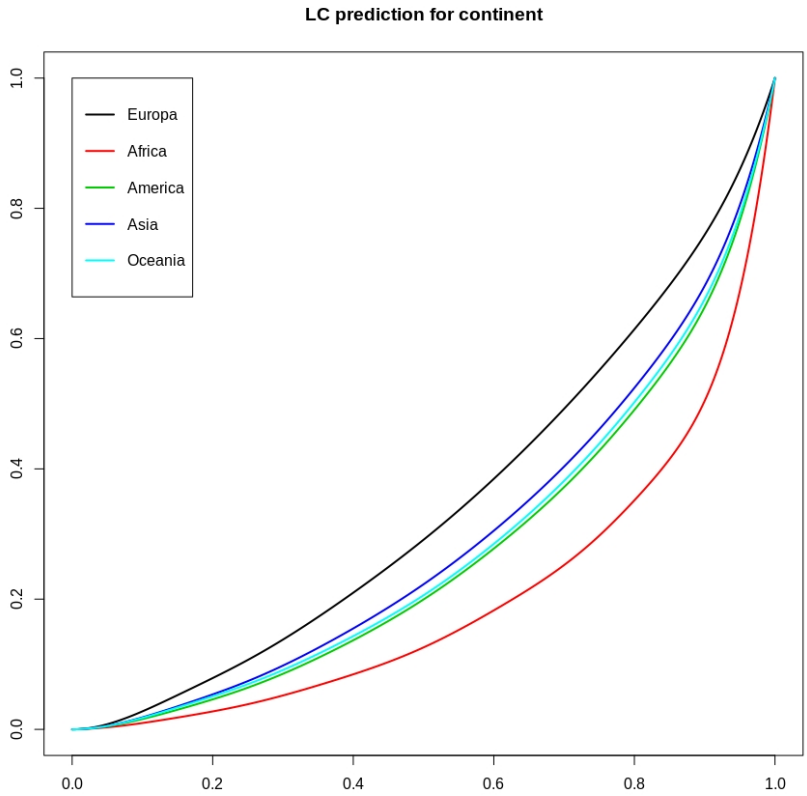


Figure 4.3: Continents LC predicted comparison

As previously said, the Lorenz curves of Asia, Oceania and America are not significant different. The European curve shows a greater global equality. African countries have a greater inequality that is not uniformly distributed on the domain, but more accentuated after the eightieth quantile, indicating a very rich elite. The plots in appendix E, show, for each variable, which intervals are significant different from the average. These results are very useful to detect how local inequality changes between continents. We focus on coefficients related to x_1 and x_2. For the middle range of domain, all the coefficients seem to be similar. The main differences are in the tails. In the left tail, European countries show a local inequality below average, while African ones above average, pointing out a larger group of absolute poor people. In the right tail, European counties have a lower local inequality. As already shown in the plot, Africans counties not only have an elite that owns most of the income, but also there exists a grater inequality inside this elite of people.

GDP PPP range

First the available data are clustered in 3 groups with respect to the GDP PPP per capita values.

We compute Hierarchical clustering with Euclidean distance and Ward linkage and cut the dendrogram to obtain 3 clusters. This method is preferred to k-means because it is less sensitive to outliers.

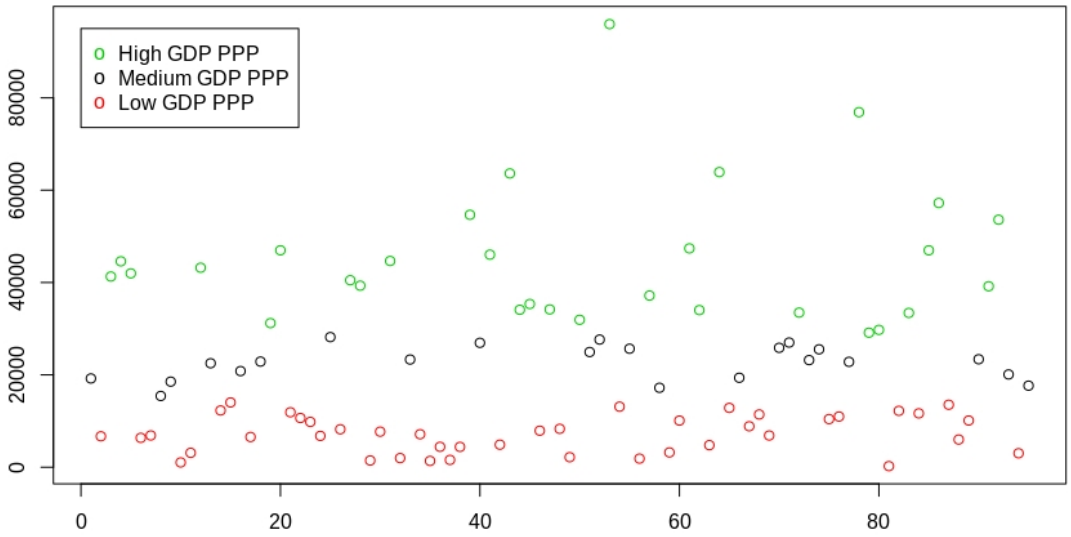


Figure 4.4: GDP PPP per capita clustering

As in the previous functional ANOVA, the design matrix is defined as follow to avoid overfitting:

- High GDP PPP : $x_1 = -1, x_2 = -1$
- Medium GDP PPP : $x_1 = 1, x_2 = 0$
- Low GDP PPP : $x_1 = 0, x_2 = 1$

The summary of the functional regression model is:

```
$call
IWTlmFoF(formula = y ~ x1 + x2)

$ttest
            Minimum p-value
(Intercept)               0 ***
x1                        0 ***
x2                        0 ***

$R2
              Range of functional R-squared
Min R-squared                    0.0005870462
Max R-squared                    0.5019822166

$ftest
  Minimum p-value
1               0 ***
```

Figure 4.5: GDP PPP per capita model summary

There exists an interval where the model is better than the null model, so the income range is a significant factor in describing inequality.

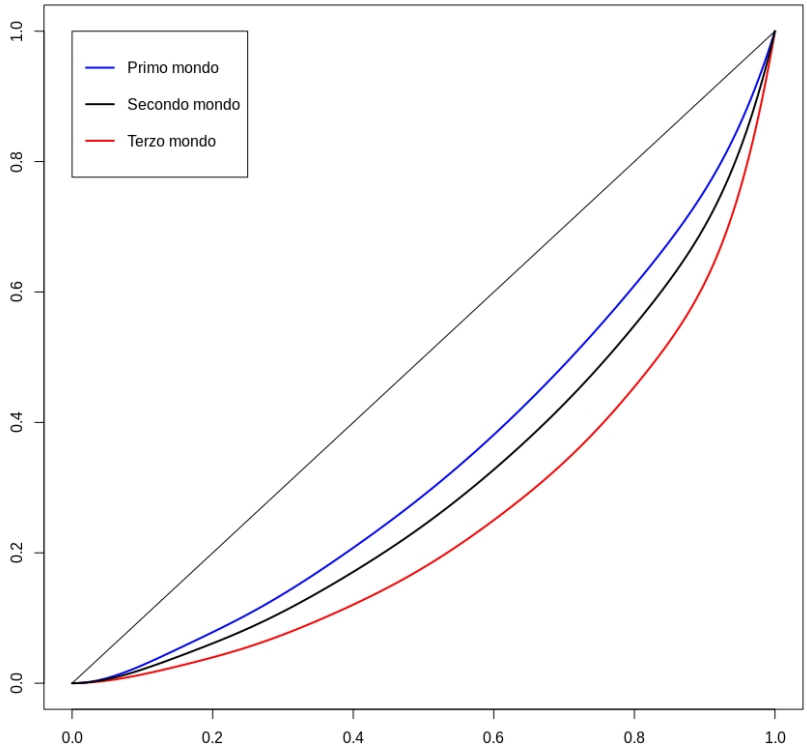


Figure 4.6: GDP PPP per capita predicted comparison

Switching to an ever-higher GDP PPP per capita range, the inequality decreases. A significant difference between groups, is given by the percentage of income owned by the richest population. The plot clearly shows how, for least developed countries, this percentage is very high. The domain intervals performed the most significant differences in local inequality, are close to 0 and 1.

4.1.2 Functional PCA & Scores

Functional Principal Component Analysis

Functional Principal Component Analysis is a strong graphical tool to visualize Lorenz curve intervals with a larger data variability and how this is expressed. Before computing the principal components, we subtract the mean function from each function.

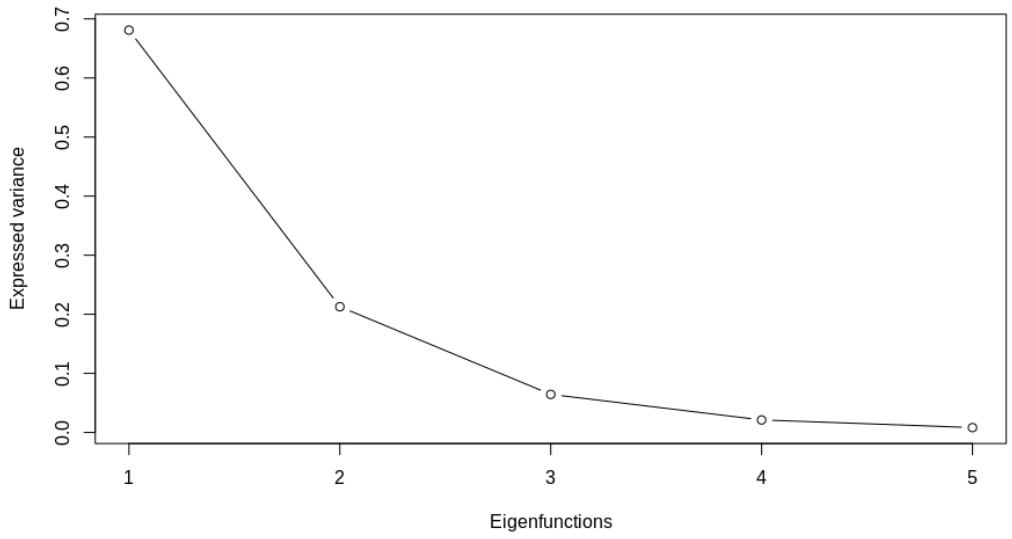


Figure 4.7: Scree plot of variance for Lorenz curves eigenfunctions

- 1st principal component : explained variance = 68 %
- 2nd principal component : explained variance = 21.8 %

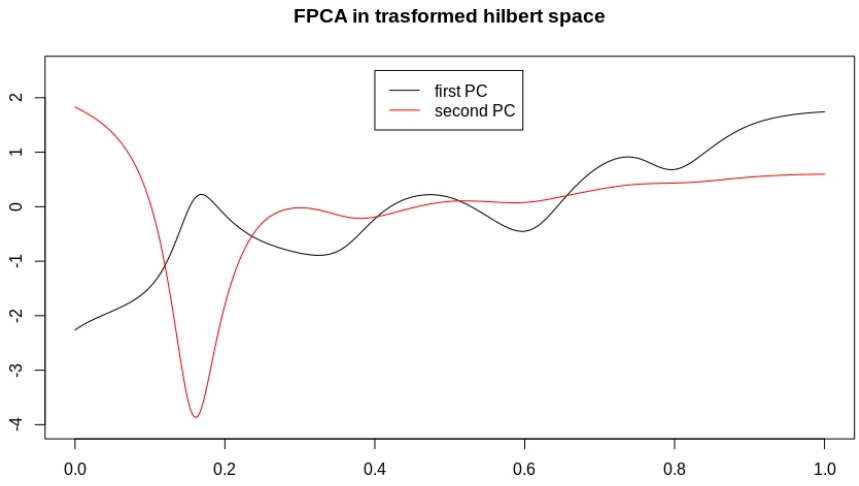


Figure 4.8: First and Second Lorenz curves eigenfunctions in Hilbert space

By the scree plot, we know that the first two principal components explain about 90% of the variance, so a clear representation of the variance in the data is provided without considering other principal components. Since the analysis is performed in a Hilbert space and giving an interpretation in sense of Lorenz curve variations is difficult and counter-intuitive, modes of variation are used to visualize the results.

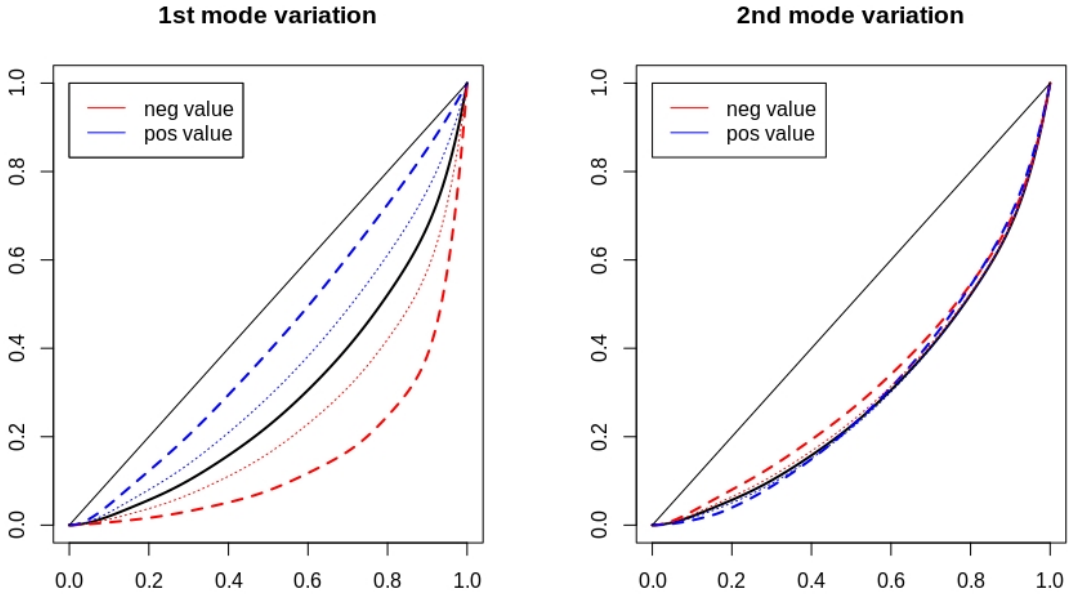


Figure 4.9: Variation modes for first and second principal components

Modes of variation show how positive and negative values, for the first and the second principal component, affect the Lorenz curve shape.

Observing the first mode of variation, the interpretation of the coefficient is quite simple. Positive values make the Lorenz curve tend to the egalitarian line, and so to a inequality reduction. Negative values determine a strong income inequality, pointing out an evident change of slope in the richest group of population. This is equivalent to a strong increase in inequality between rich and poor people and indicates the presence of a small part that owns most of the income.

Observing the second mode of variation, the difference between a positive and negative component is not so evident. A negative component defines a greater equality in the poor-medium range, while a positive one indicates a larger poor population group, shown by a flatter interval near the 0.

Scores

An important aspect is the examination of scores. We can compare the inequality profile of each state with respect the others by looking at its principal component scores.

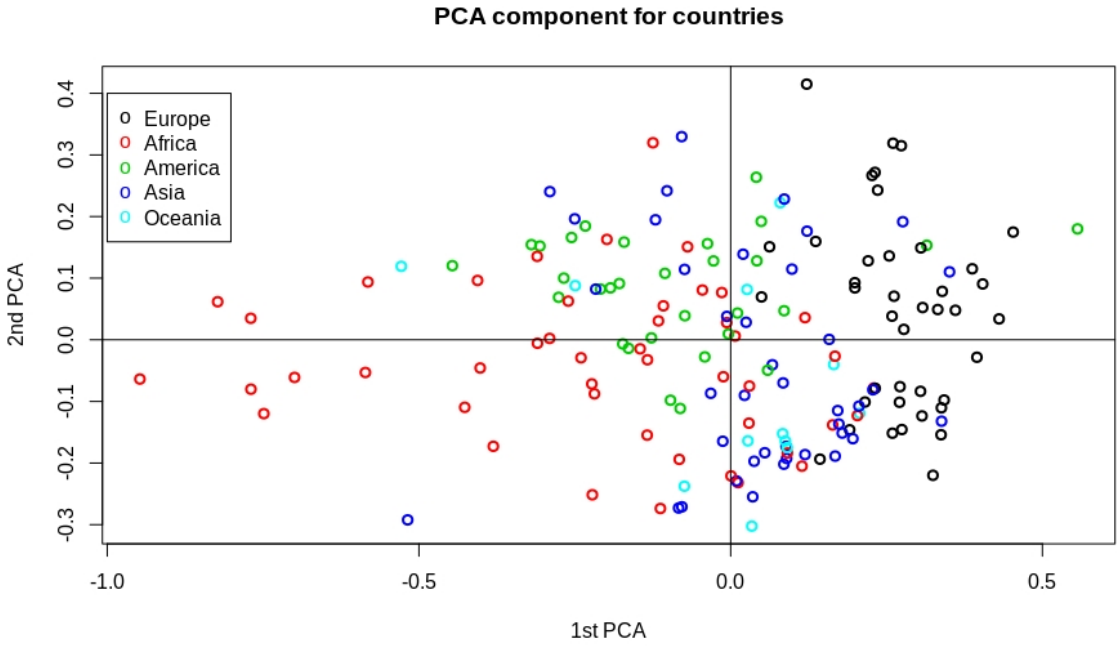
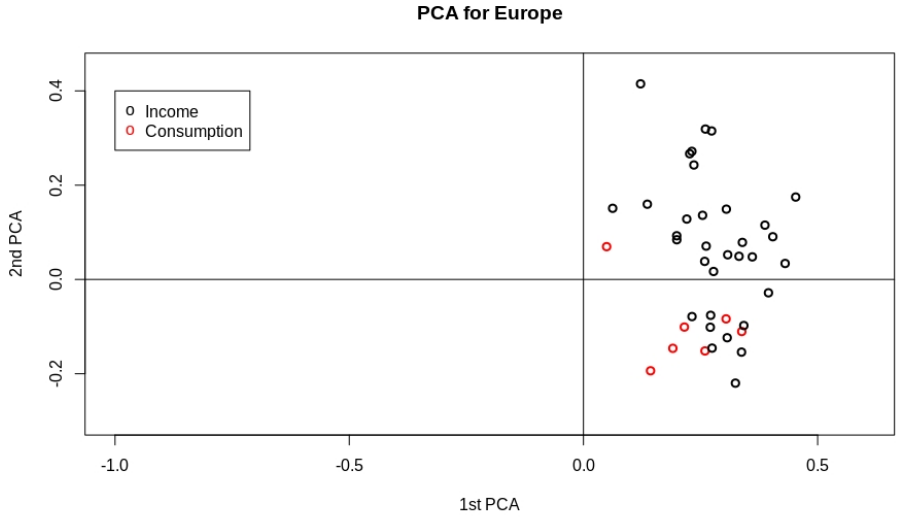
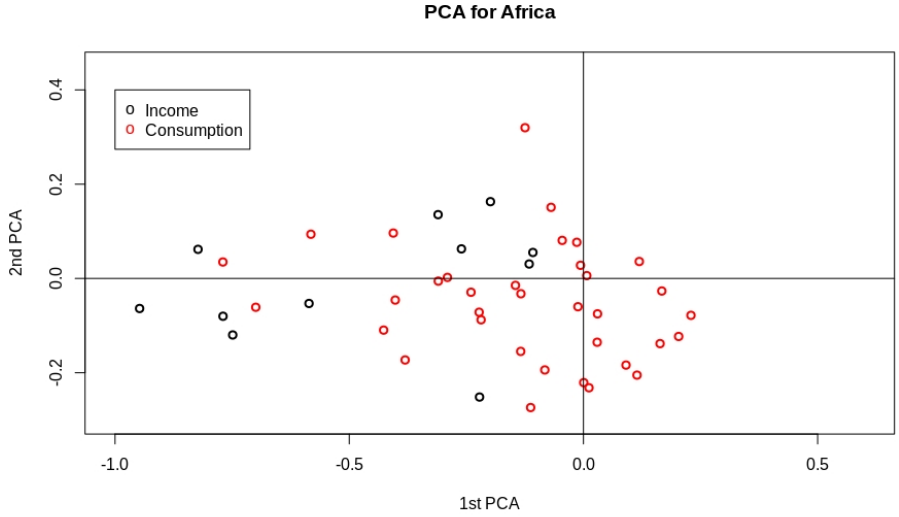


Figure 4.10: Function PCA scores for Lorenz curves

Focusing on continents, we have a further proof of the results obtained by the functional ANOVA model. All the European countries have positive score in the first component, while the Africans have negative scores, especially for the curves describing income inequality. The other continents show a homogeneous distribution around the origin, except for Cuba which has a very particular inequality profile, as described in section 4.3.





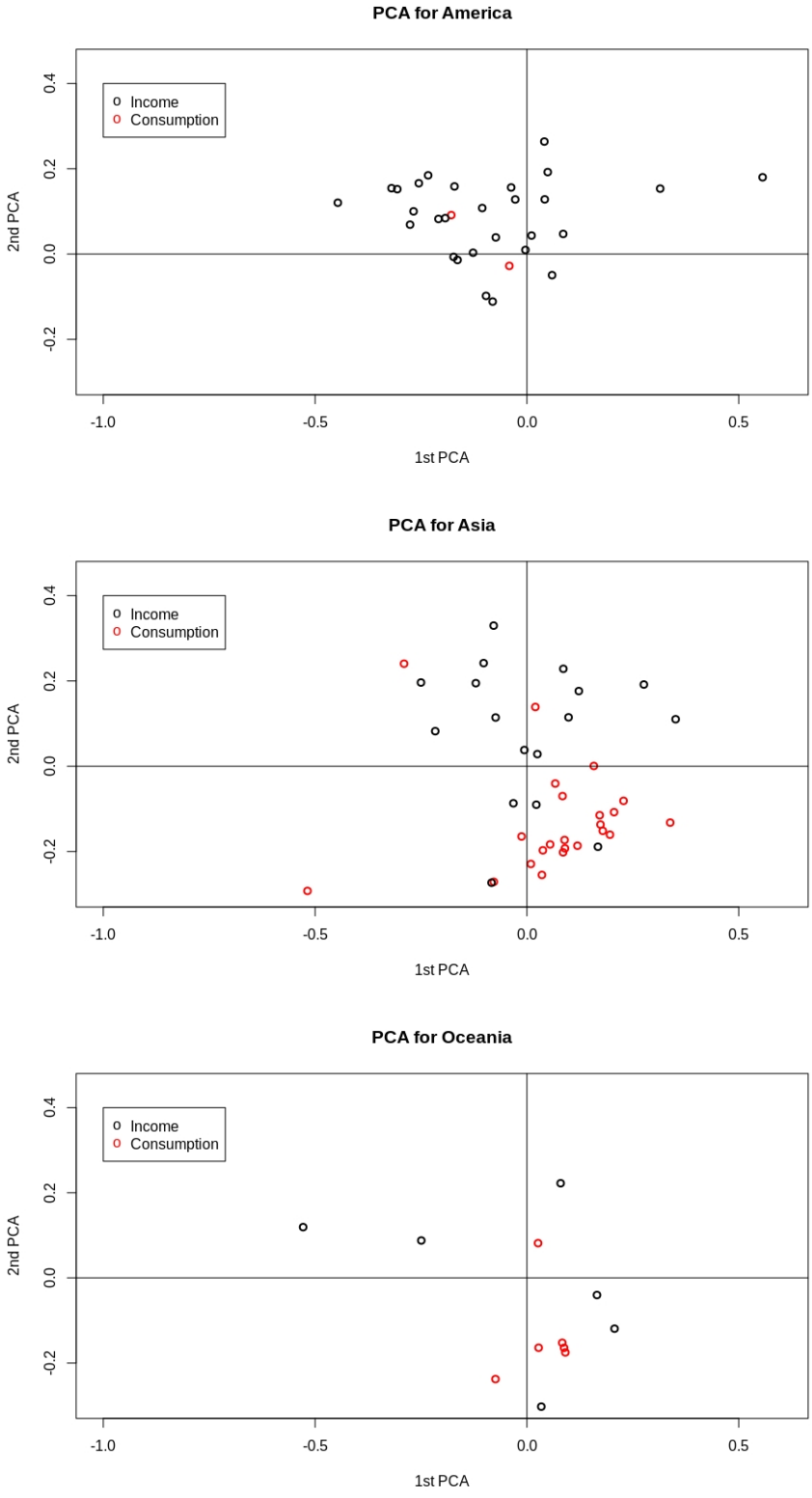


Figure 4.11: PCA score for continent

We consider 4 countries with all the possible sign combinations for the scores. The Lorenz Curves are compared with the mean curve, pointing out the principal component effects.

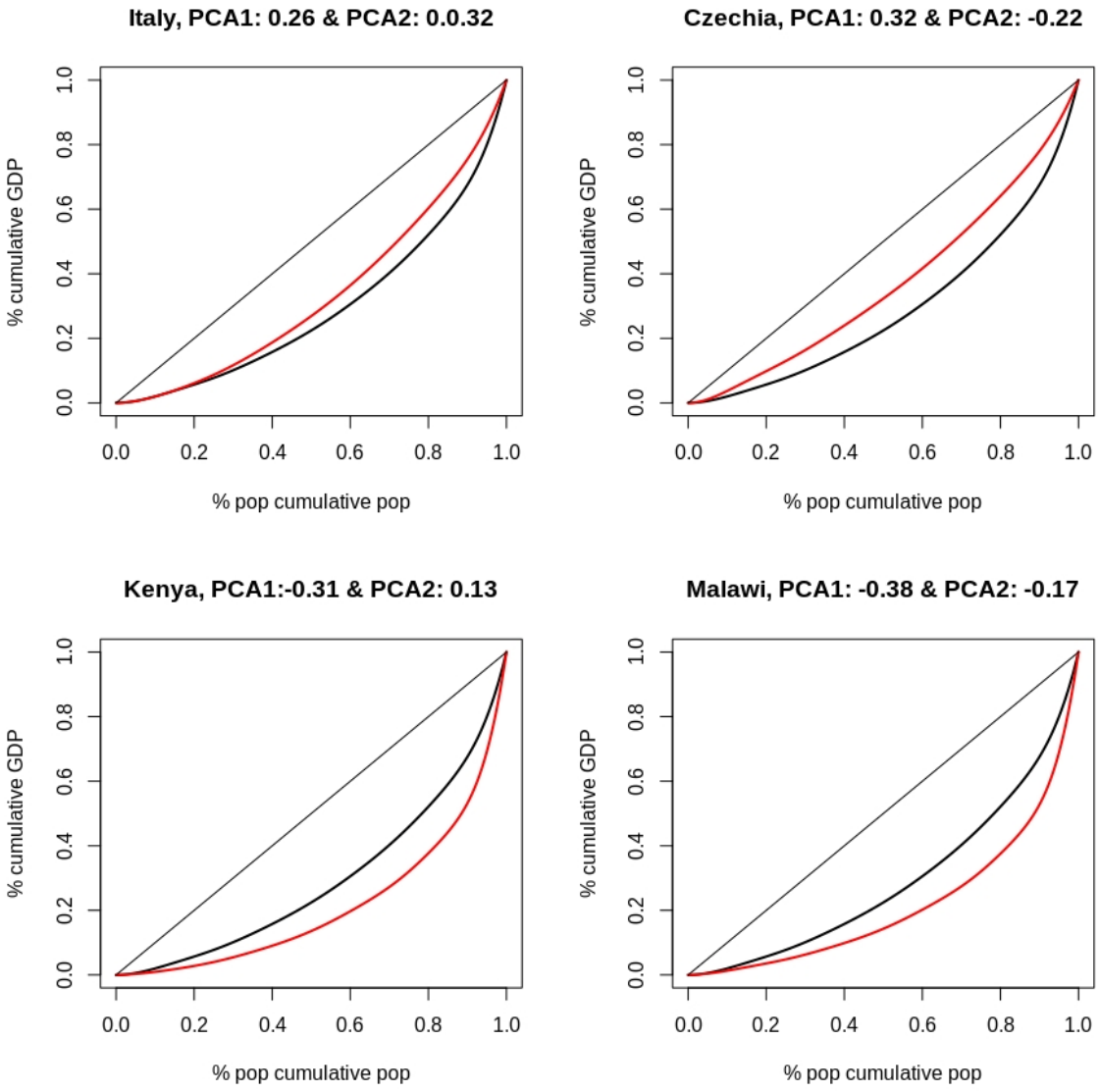


Figure 4.12: LC for different scores signs

Now we compare all the Lorenz curves in the same plot.

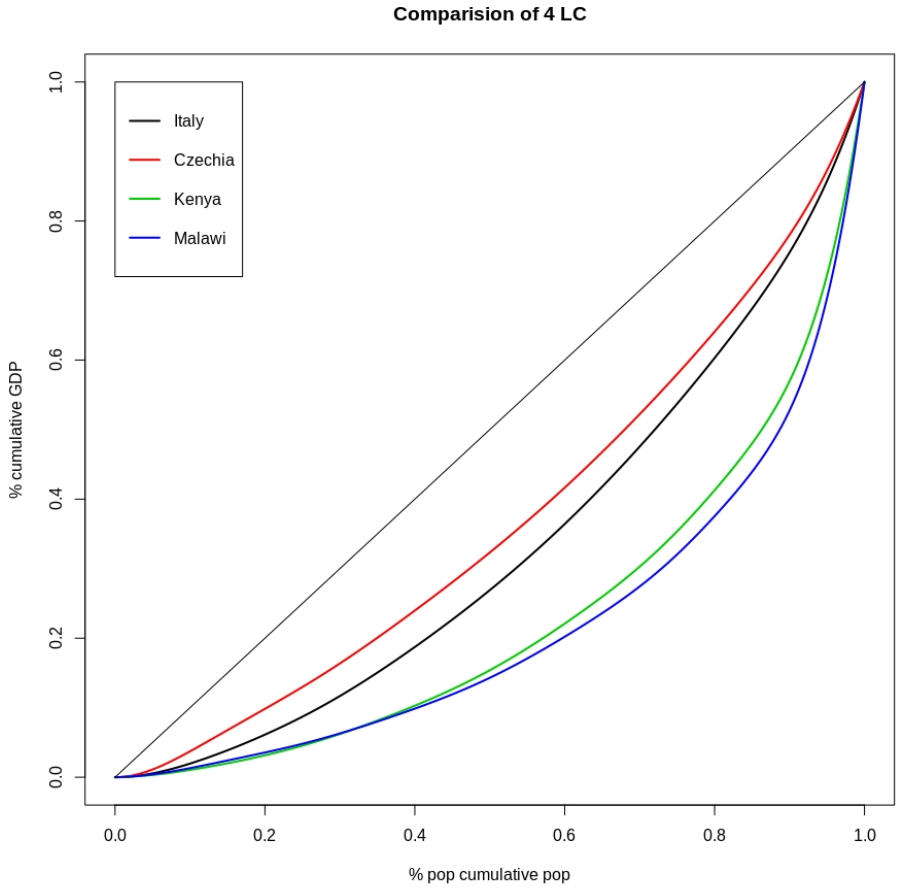


Figure 4.13: Lorenz curves comparison

There exists a clear difference between the curves with positive and negative first principal component in sense of global inequality. The Czech Republic is closer to the egalitarian line than Italy for a greater value in the first principal component. Furthermore, the two curves have the maximum distance in the initial part, as consequence of second component different sign. This is also evident observing the other two countries. In the initial part, Malawi has higher Lorenz curve value than Kenya, despite a lower first principal component value. This as consequence of the second component.

4.1.3 Functional regression

A scalar regression model with Gini index as response, only defines if a given variable determines an increase or decrease in inequality globally. Through a functional regression model with functional response, we can determine how the value of each regressor affects the profile of the Lorenz curves, catching the inequality variations

at each level of the population. We use the principal components computed in section 3.6 to avoid high correlated variables in the regression model.

```
$ttest
            Minimum p-value
(Intercept)           0.000 ***
x_cons                0.000 ***
pca1                  0.000 ***
pca2                  0.016   *
pca3                  0.828
pca4                  0.414
x_cons:x6             0.015   *

$R2
              Range of functional R-squared
Min R-squared                    0.08000069
Max R-squared                    0.51070416

$ftest
  Minimum p-value
1               0 ***
```

Figure 4.14: Summary of complete regression model

After fitting the complete model, backward feature eliminations method is used to remove the non-significant regressors. Start training on n input features, then one input feature at a time is removed and we train the same model on $n-1$ input. The input feature, whose removal has produced the smallest increase in the error rate, is removed.

```
$call
IWTlmFoF(formula = y1 ~ x_cons + pca1 + pca2 + x_cons:x6)

$ttest
            Minimum p-value
(Intercept)           0.000 ***
x_cons                0.000 ***
pca1                  0.000 ***
pca2                  0.012   *
x_cons:x6             0.009  **

$R2
              Range of functional R-squared
Min R-squared                    0.06362989
Max R-squared                    0.50878813

$ftest
  Minimum p-value
1               0 ***
```

Figure 4.15: Summary of reduced regression model

Income vs Consumption

x_cons is a dummy variable, showing how the consumption curve, on average, differs from the income curve. It's not the only effect that describes how inequality differs between consumption and income. Another significant variable explains GDP PPP per capita effect on income/consumption inequality. The comparison of the two Lorenz curves shows the fixed effect of marginal propensity to save, which makes consumption more equal than income.

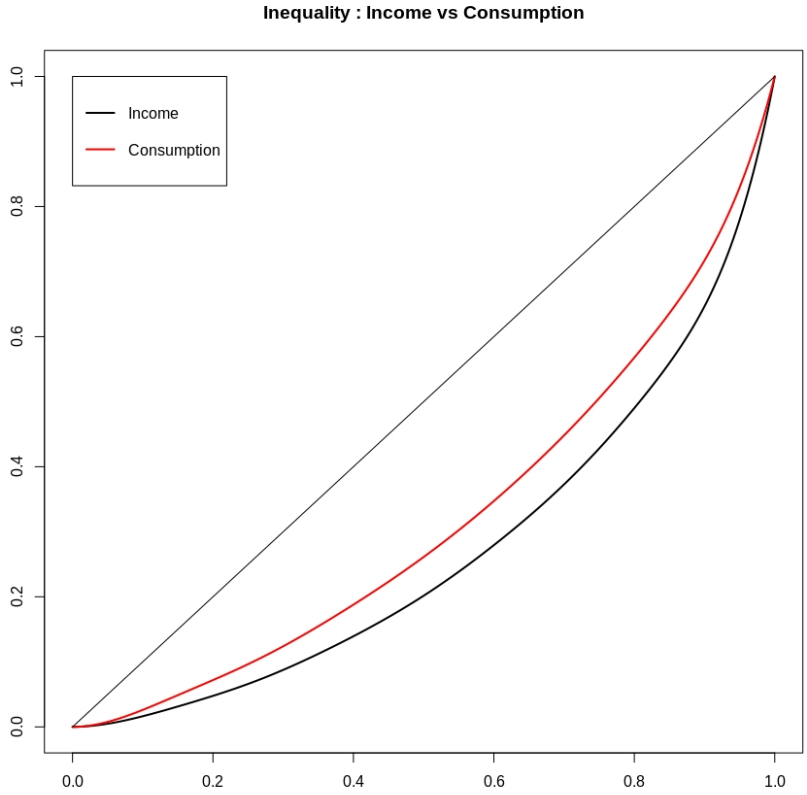


Figure 4.16: Consumption fixed effect

Consumption effect related to GDP PPP per capita

Income can be defined as follow:

$$Income = Consumption + Saving = (1 - \alpha) * Income + \alpha * Income$$

α is called "marginal propensity to save" and indicates the portion of income that is not consumed in the year in which it is produced, but it is saved [2]. It is reasonable to expect that α is not constant between different population ranges or different

states. For this reason, in addition to a dummy variable income/consumption, a GDP PPP-consumption variable is introduced to verify if and how income and consumption inequality differ with respect to a GDP PPP per capita values.

By introducing a GDP PPP-consumption variable and already considering the fixed effect, when the GDP PPP per capita value increases, the difference between the consumption and income Lorenz curves decreases.

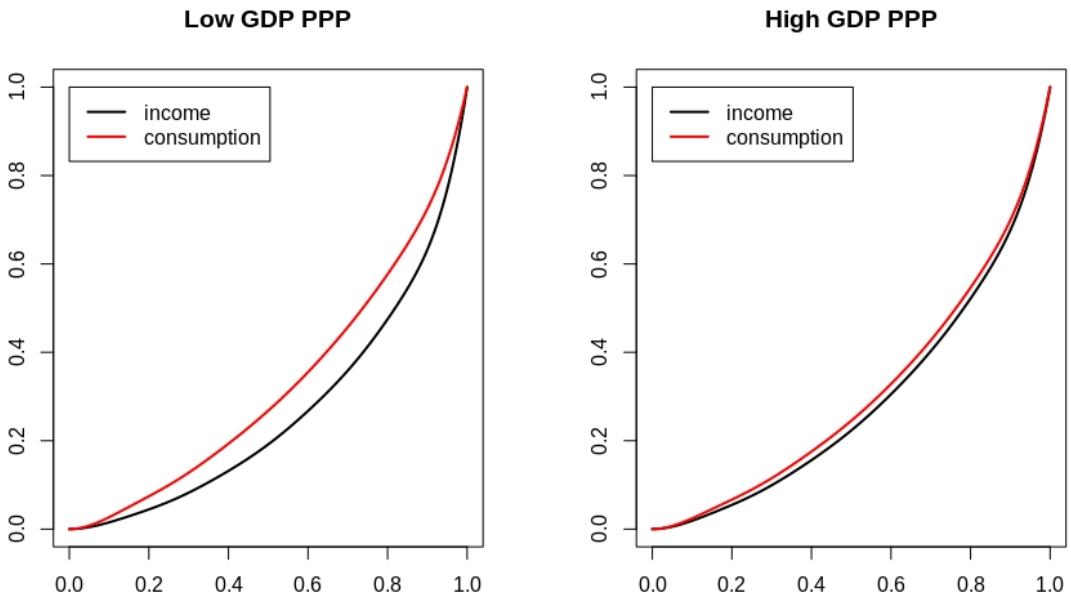


Figure 4.17: Income vs Consumption with respect to GDP PPP per capita

The GDP PPP per capita values considered in the figure, are two extreme values within the range of countries with inequality information related to consumption. Comparing the curve predictions for countries with GDP PPP per capita values outside this range could give inconsistent results.

Principal component effects

After interpreting the principal components, we observe how the Lorenz curve change with respect to each component taken individually.
As easily understood by the composition, the first component determines a variation of the inequality that includes all the population groups, from the poorest to the richest. Positive values indicate a more homogeneous income distribution. For extreme negative values, we see that more than half of the income is owned by a low percentage of people. The second component is not very relevant, except for extreme values.

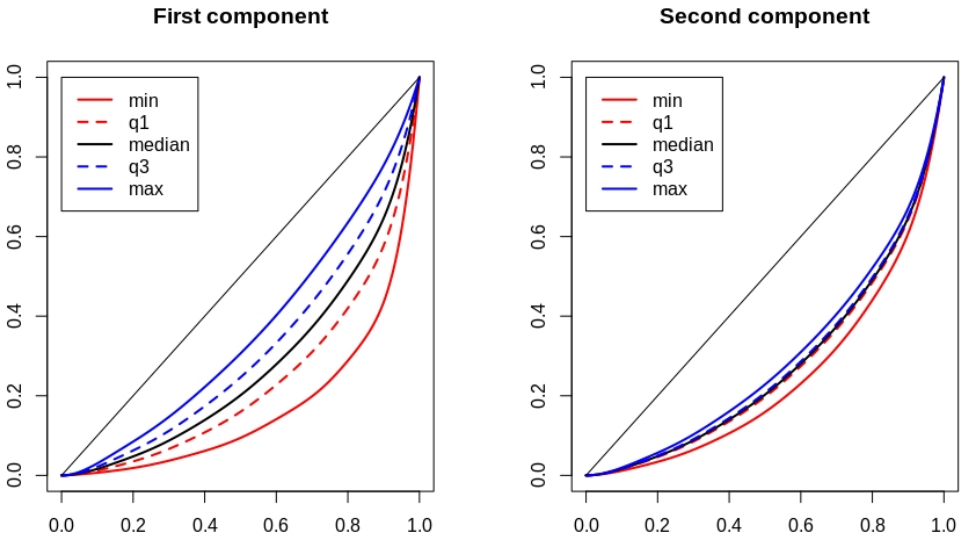


Figure 4.18: First and second principal component effects

Attribute effects

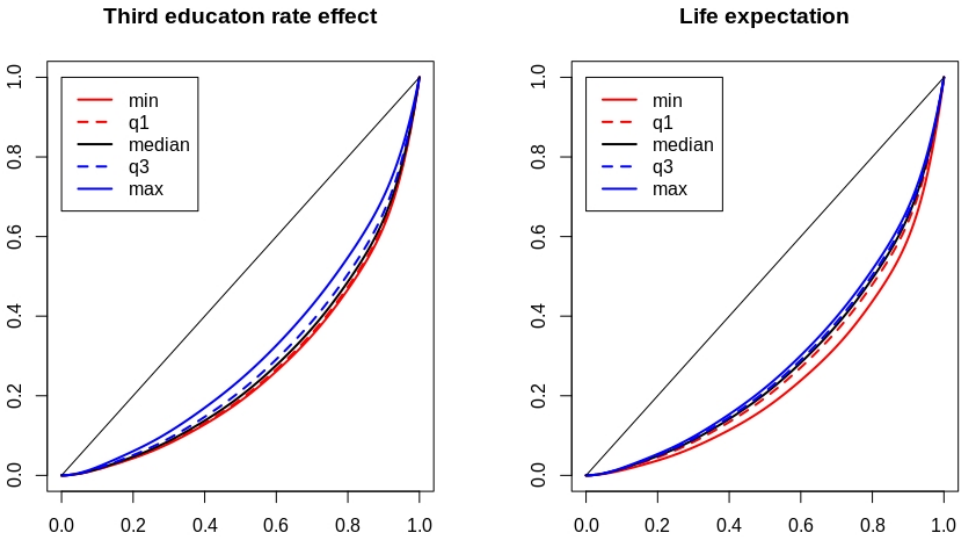


Figure 4.19: tertiary education rate and life expectancy effects

Going back to the original variables, we look for the ones affecting mostly the Lorenz curves. Tertiary education rate and life expectancy have significant effects.

4.2 Analysis of density functions

Density functions are computed from the Lorenz curves with a bijection.
The introduction of density distribution to describe inequality has the purpose of showing the inequality profile from another point of view, pointing out some aspects hardly perceptible from the analysis of the Lorenz curves. The functions are mapped with the clr transformation in a Hilbert space to make consistent analyses.

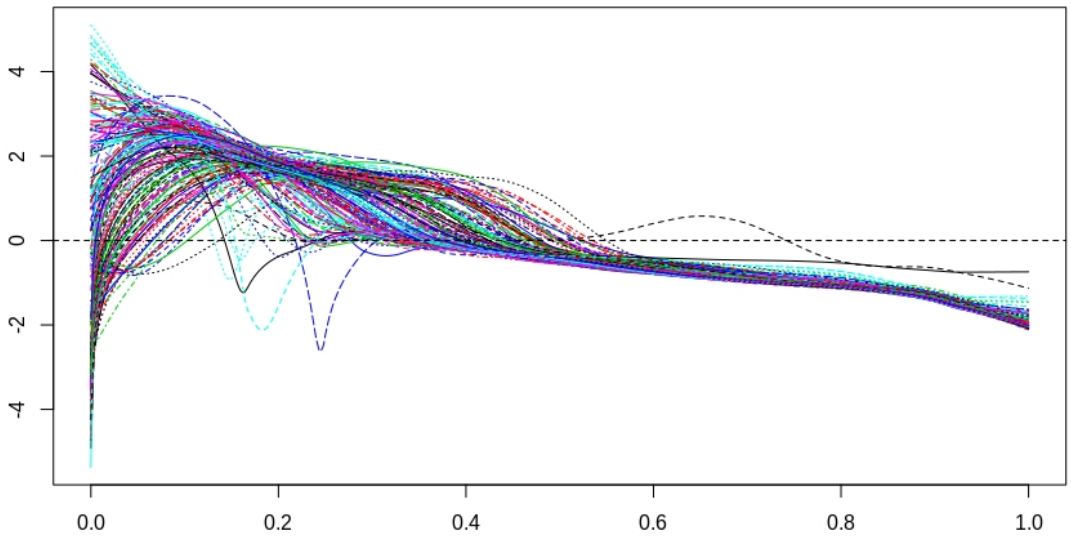


Figure 4.20: Density functions transformations

4.2.1 Exploratory analyses

The introduction of density functions shows how the density of every population range changes, from the poorest to the richest, in relation to different factors.

Continents

Functional ANOVA requires that every continent is identified by a categorical variable. To avoid over-fitting, the design matrix is defined as follow:

- Europe : $x_1 = 1, x_2 = 0, x_3 = 0, x_4 = 0$
- Africa : $x_1 = 0, x_2 = 1, x_3 = 0, x_4 = 0$
- America : $x_1 = 0, x_2 = 0, x_3 = 1, x_4 = 0$
- Asia : $x_1 = 0, x_2 = 0, x_3 = 0, x_4 = 1$
- Oceania : $x_1 = -1, x_2 = -1, x_3 = -1, x_4 = -1$

The summary of the functional regression model is:

```
$call
IWTlmFoF(formula = y ~ x1 + x2 + x3 + x4)

$ttest
            Minimum p-value
(Intercept)           0.000 ***
x1                    0.000 ***
x2                    0.001  **
x3                    0.031   *
x4                    0.105

$R2
              Range of functional R-squared
Min R-squared                    0.07990323
Max R-squared                    0.69933285

$ftest
  Minimum p-value
1               0 ***
```

Figure 4.21: Continents model summary

As parameters and continents significant differences, we obtain the same results as the Lorenz curves functional ANOVA. This is reasonable by the fact that we are

describing the same inequality, even if a different point of view. By this analysis, we are interested in seeing how the differences are expressed graphically and in detecting some different details.

Computing the $\beta_i(t)$ coefficients, the clr transformed functions are computed for each continent and, then, remapped as densities.

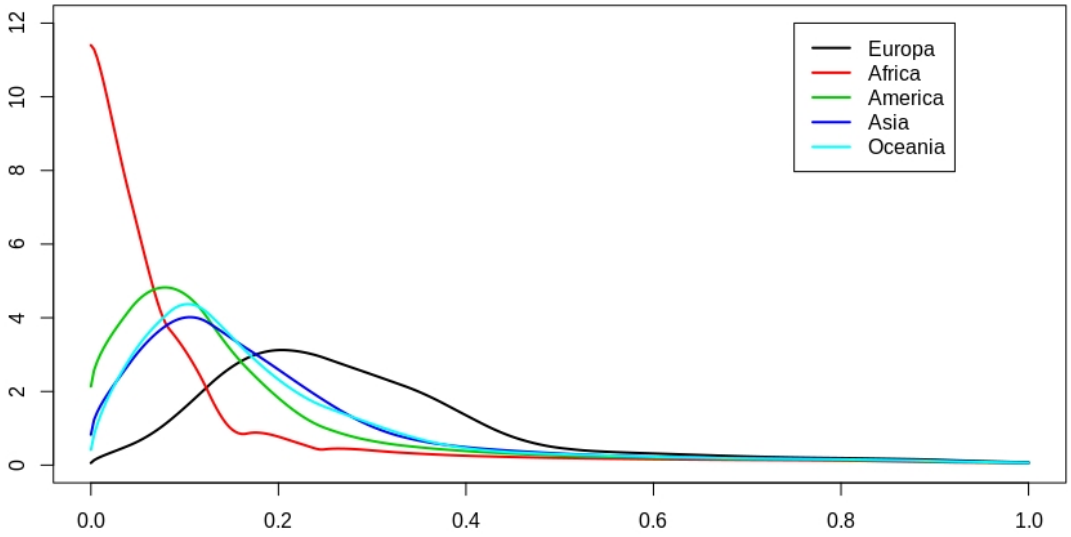


Figure 4.22: Continents density functions predicted comparison

Considering Oceania, America and Asia three continents with the same income distribution, the factors differ mainly in term of peak position and variability around it. African countries have a high peak near 0 value, with a quickly decrease. European counties have a peak toward the centre of the domain and a large variance, with a small density close to the very poor range of population. This is a clear indication of a greater equality with respect to the other continents. Considering the intervals, we observe that there is not a difference in density for the richest part of the domain, but it is all detected in the middle-poor ranges.

GDP PPP range

We use the same clustering introduced in the Lorenz curve analysis.

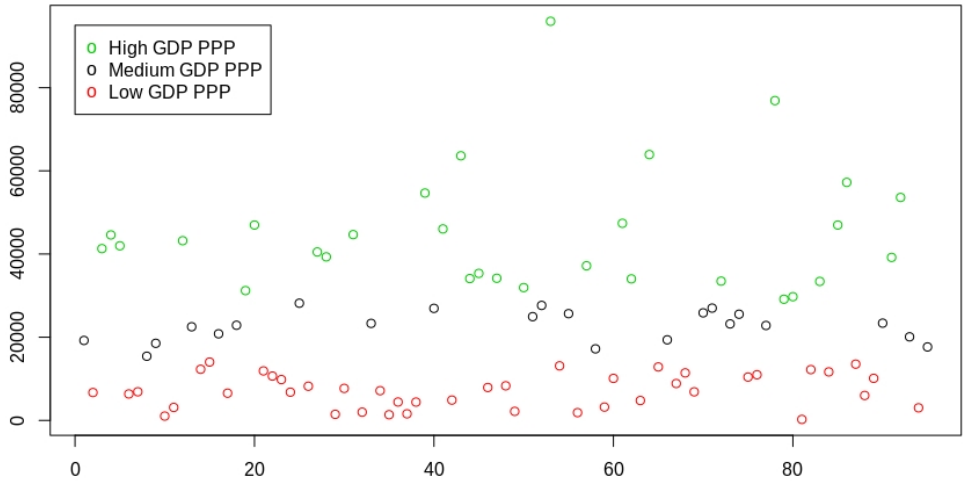


Figure 4.23: GDP PPP per capita clustering (2)

The design matrix is defined as follow to avoid over-fitting:

- High GDP PPP : $x_1 = -1, x_2 = -1$
- Medium GDP PPP : $x_1 = 1, x_2 = 0$
- Low GDP PPP : $x_1 = 0, x_2 = 1$

The summary of the functional regression model is:

```
$call
IWTlmFoF(formula = y ~ x1 + x2)

$ttest
            Minimum p-value
(Intercept)            0.00 ***
x1                     0.53
x2                     0.00 ***

$R2
              Range of functional R-squared
Min R-squared                   0.008218663
Max R-squared                   0.611528465

$ftest
  Minimum p-value
1               0 ***
```

Figure 4.24: GDP PPP per capita model summary

Summaries at appendix B show significant differences in term of inequality between the factors.

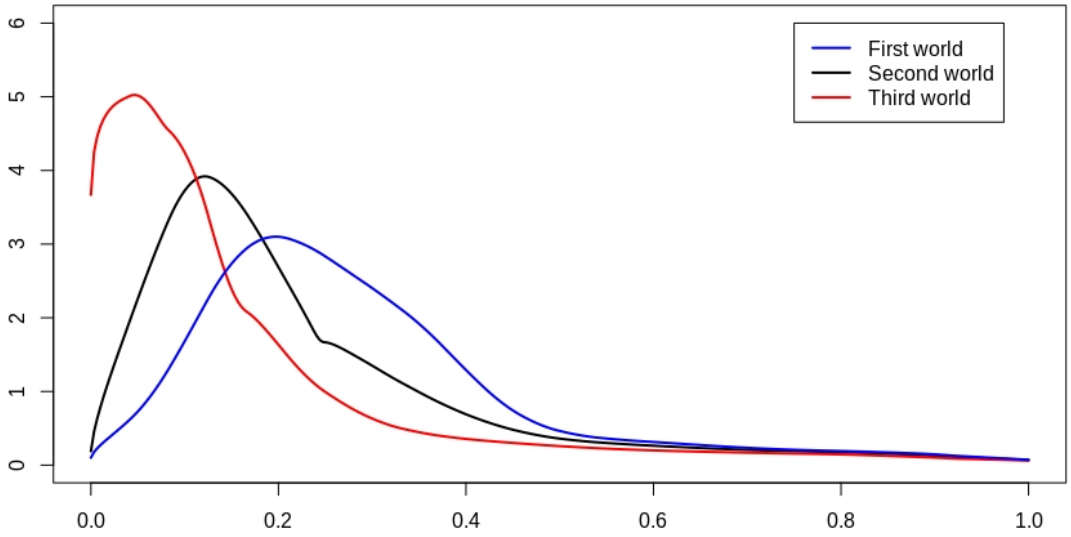


Figure 4.25: GDP PPP per capita density functions predicted comparison

The peak of density shifts towards higher wealth groups as we pass to the clusters with an higher GDP PPP per capita. The poorest counties with respect to GDP PPP per capita values, have a right tail going to 0 very quickly, index of a clear split of the population between few rich and many poor men.

4.2.2 Functional PCA & Scores

Functional Principal Component Analysis

Computing functional Principal Component Analysis for density functions, we provide a further representation of the data variability. On a graphic level, density functions express different aspects of inequality profile with respect to Lorenz curves, and we expect to see this also in the principal component eigenfunctions and scores. After subtracting the mean function from each function, we compute the analysis.

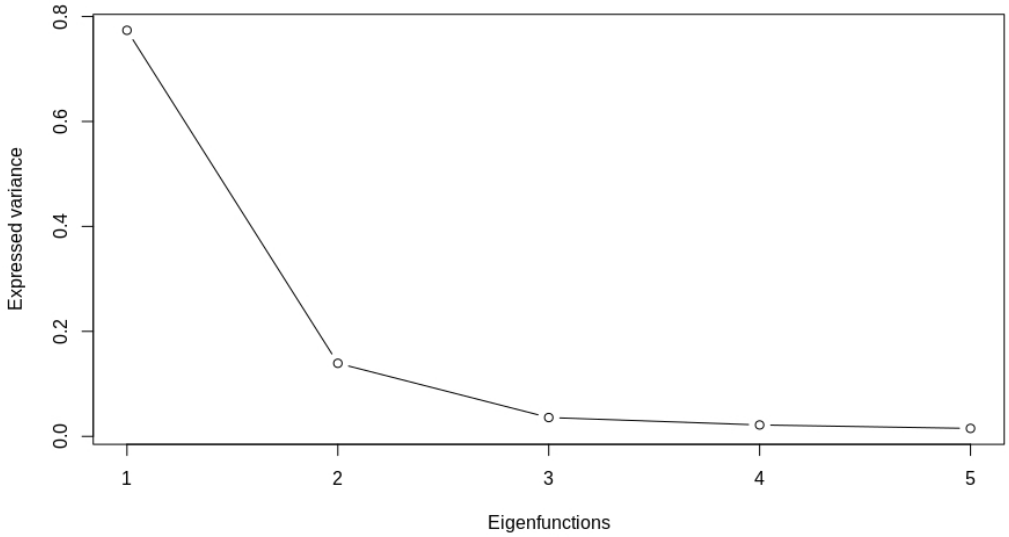


Figure 4.26: Scree plot of variance for densities eigenfunctions

- 1st principal component : explained variance = 77 %
- 2nd principal component : explained variance = 14 %

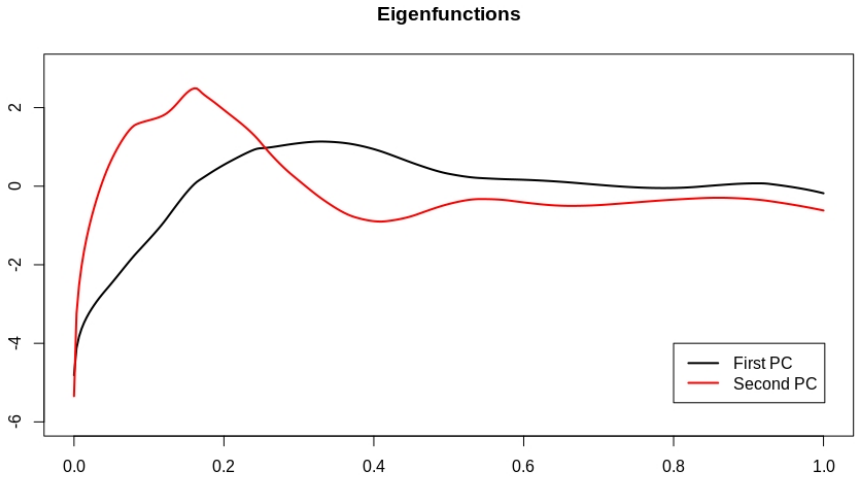


Figure 4.27: First and Second density eigenfunctions in the transformed space

As for Lorenz curves, the first two principal component explained the most of variance, about the 91%. To interpret the effects of the principal components, we use the mode of variations.

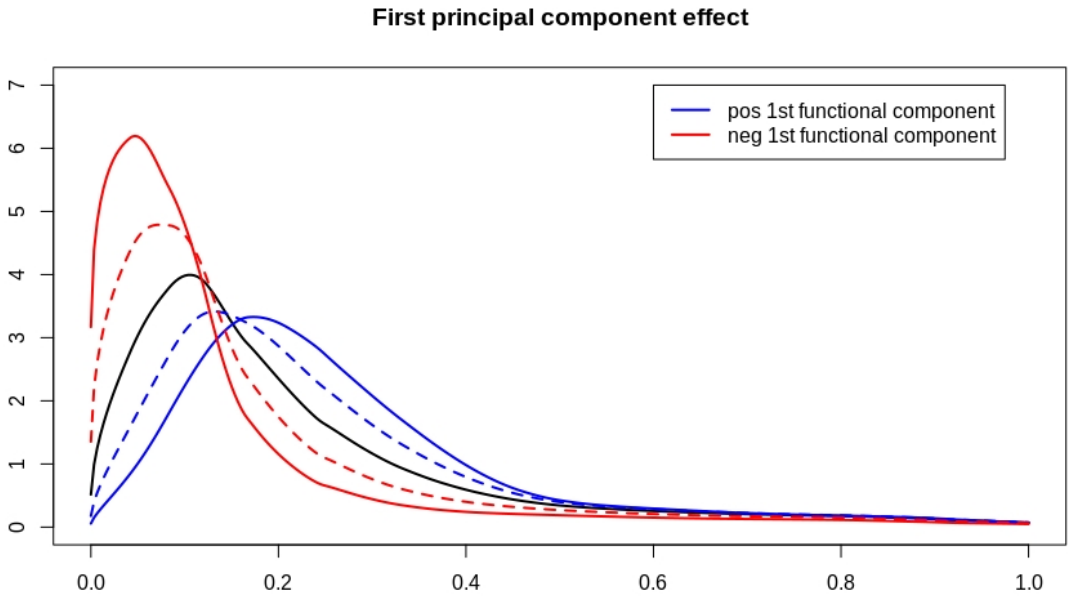


Figure 4.28: Variation mode for first eigenfunction

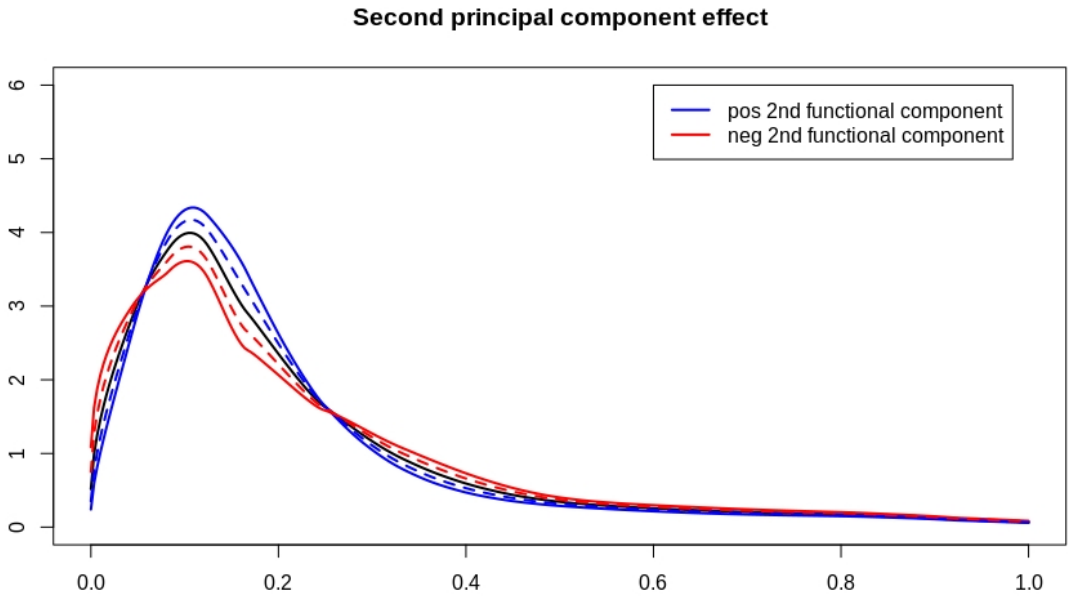


Figure 4.29: Variation modes for second eigenfunction

Observing the first mode of variation, moving from negative to positive values, we observe a shift of the peak from the extremely poor population to a wealthier

range with more variance around it and an almost zero number of absolute poor. The fist component indicates the income range where the most of population is.

Observing the second mode of variation, the effect of negative scores is to make the density distribution in the medium-low range more homogeneous, as shown by a lower peak and a right tail less decreasing.

Scores

The scores distribution is not uniform around the axes origin, as for the Lorenz curves, but it forms a semicircle without countries with null first and negative second components. This confirms that there not exists a perfect one-to-one correspondence between the principal components for the functional objects.

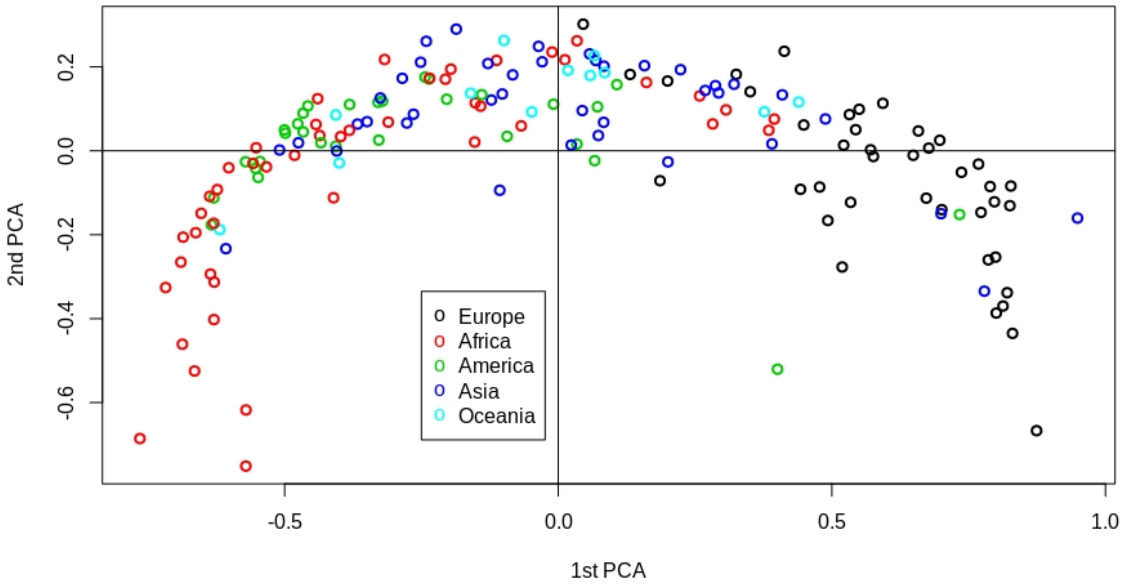
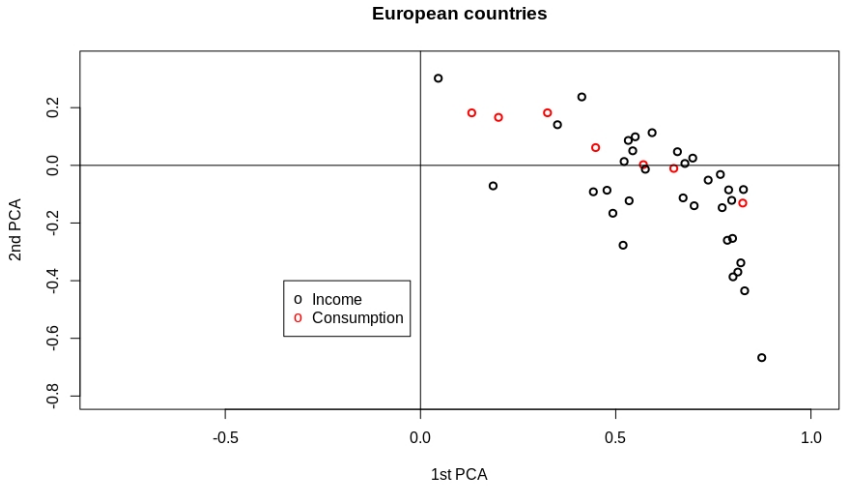
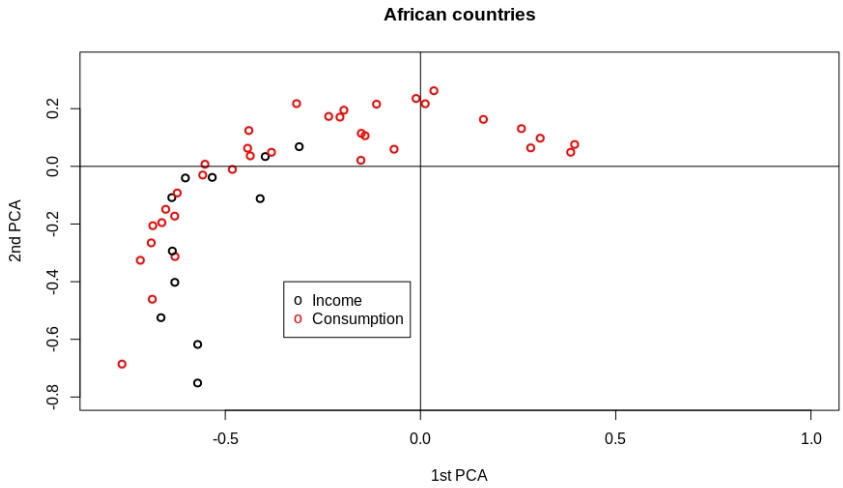
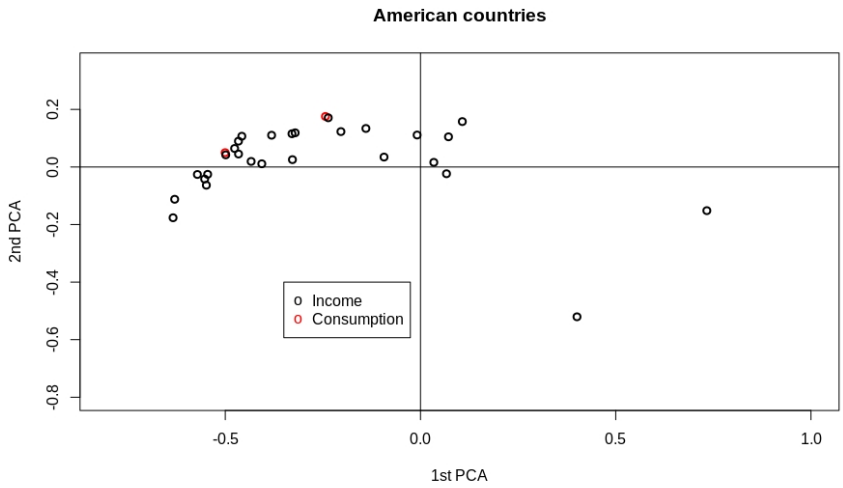


Figure 4.30: Function PCA scores for density functions

Focusing on continents, we obtain consistent results with the functional ANOVA and with the Lorenz curves results.


European countries
2nd PCA
1st PCA
Income
Consumption
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-0.5
0.0
0.5
1.0


African countries
2nd PCA
1st PCA
Income
Consumption
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-0.5
0.0
0.5
1.0


American countries
2nd PCA
1st PCA
Income
Consumption
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-0.5
0.0
0.5
1.0

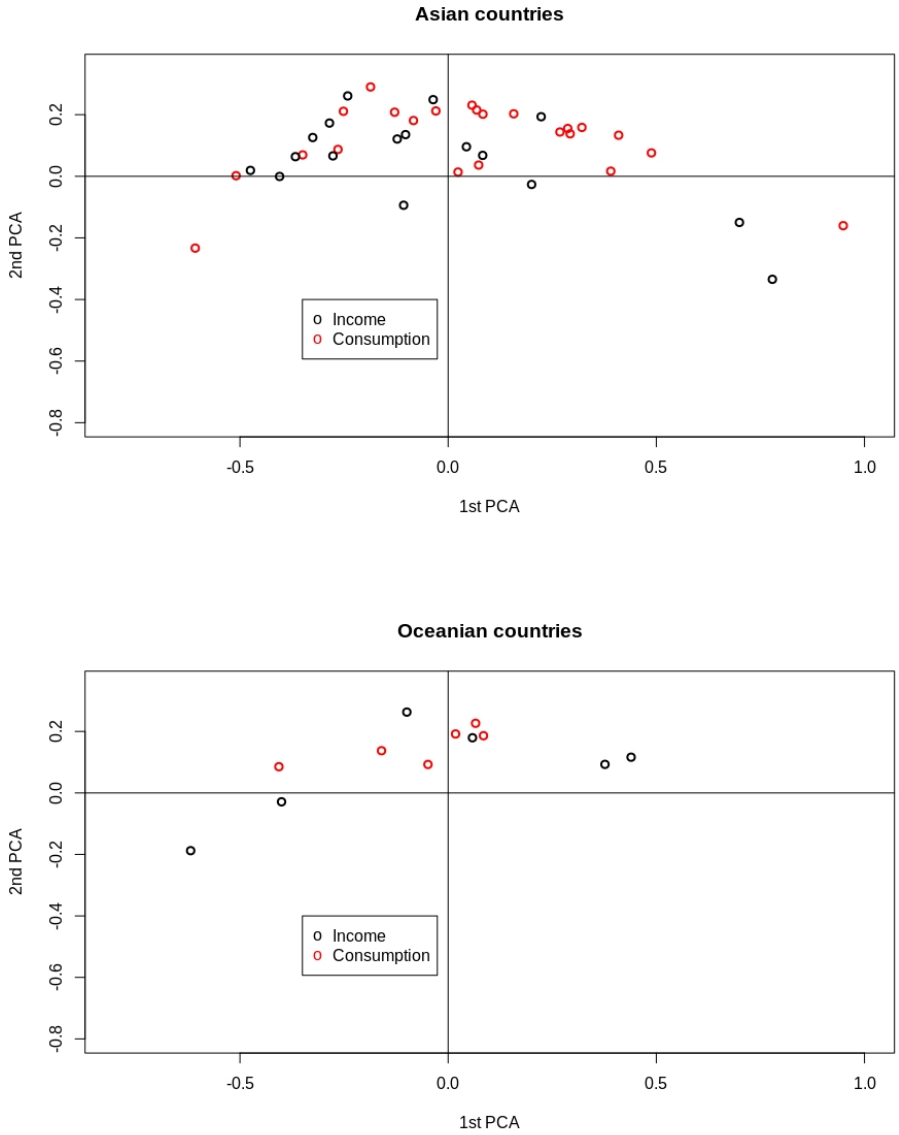


Figure 4.31: PCA score for continent

As for the Lorenz curves, we consider 4 density, combining the principal component signs, and compare with the mean function. After that, we compare each other's in the same plot.

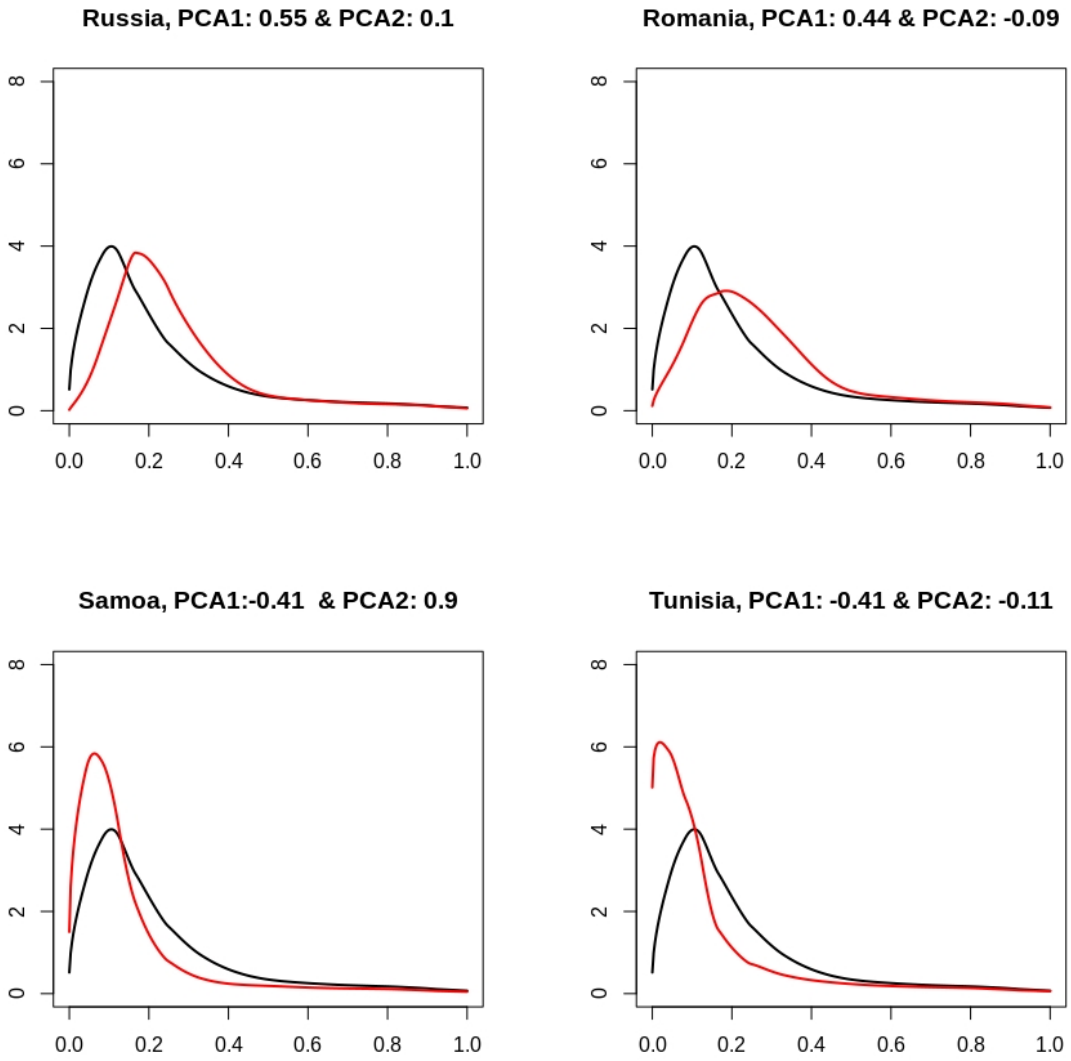


Figure 4.32: Density functions for different scores signs

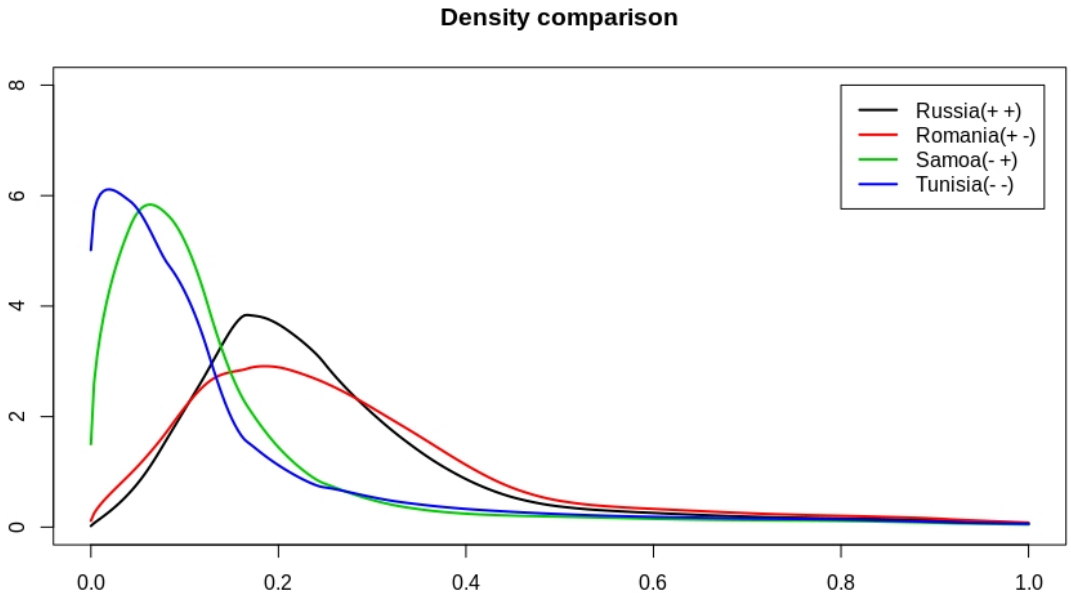


Figure 4.33: Density functions comparison

There exists a clear difference between the densities with positive and negative first principal component. Examining densities with a positive first component, the effect of the second principal component is very clear. Romania, with a negative value, has a lower peak and a more distributed density around it, with a right tail that decreases slowly. Considering the two densities with a negative first principal component, Tunisia has a peak shifted towards left and then a less accentuated descent. The translation of the peak is a direct consequence of the change in sign of the second component. A negative component implies a greater density of extremely poor people, but also a slower decrease to the right.

4.2.3 Functional regression

As shown in the previous sections, density distribution captures different aspects on income inequality with respect to the Lorenz curve. Computing a functional regression model, we detect if and how the variables affect density functions.

```
$call
IWTlmFoF(formula = y1 ~ x_cons + pca1 + pca2 + pca3 + pca4
    x_cons:x6)

$ttest
            Minimum p-value
(Intercept)           0.000 ***
x_cons                0.011   *
pca1                  0.000 ***
pca2                  0.011   *
pca3                  0.000 ***
pca4                  0.985
x_cons:x6             0.158

$R2
              Range of functional R-squared
Min R-squared                     0.1580336
Max R-squared                     0.5040576

$ftest
  Minimum p-value
1               0 ***
```

Figure 4.34: Summary of complete regression model

To delete the non-significant regressors, we use backward features eliminations method.

```
$call
IWTlmFoF(formula = y1 ~ x_cons + pca1 + pca2 + pca3)

$ttest
            Minimum p-value
(Intercept)           0.000 ***
x_cons                0.043   *
pca1                  0.000 ***
pca2                  0.015   *
pca3                  0.001  **

$R2
              Range of functional R-squared
Min R-squared                     0.1521216
Max R-squared                     0.4833209

$ftest
  Minimum p-value
1               0 ***
```

Figure 4.35: Summary of reduced regression model

By comparing this summary with the Lorenz curve one, we observe some differences. The third principal component becomes significant, while the dummy variable consumption-GDP PPP per capita is not significant. This means that the variation from consumption density to income density is constant, and no more affected by the GDP PPP per capita values.

Income vs Consumption

The income-consumption relationship is constant, and it is represented by a density shift towards the right.

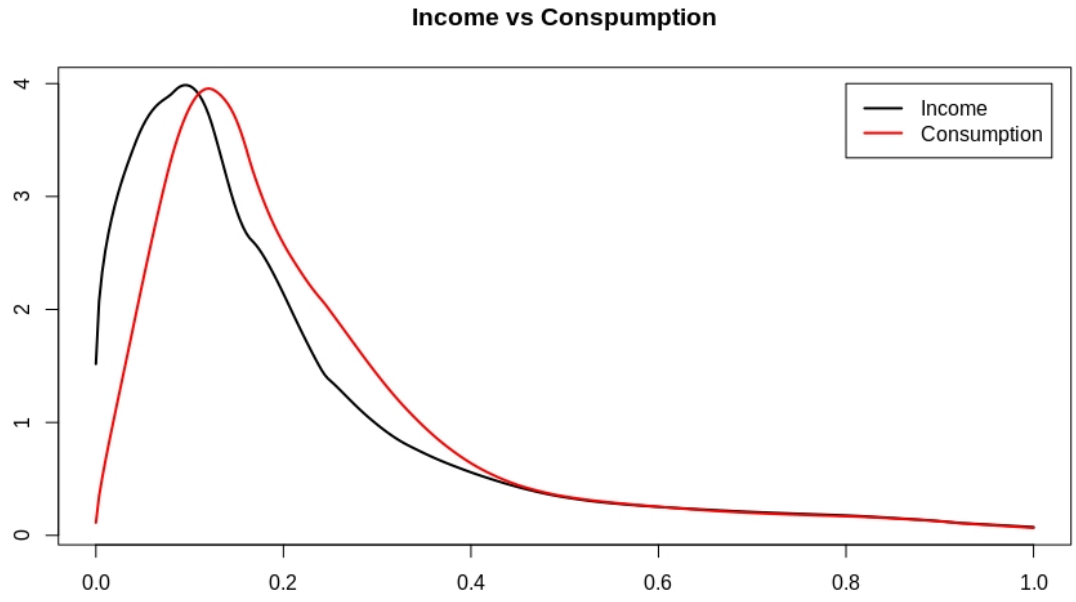


Figure 4.36: Consumption effect

Principal component effects

If, with Lorenz curve representation, it was difficult to interpret the regressor effects on income distribution in the population strata, with density functions it is much easier to give a clear interpretation.

As first component effects, we notice that, passing from negative to positive values, the most of population density moves from the absolute poverty band to an intermediate income range. For positive values of the third component, the estimate density has a right tail slowly decreasing. While the effect of the second component has not a strong intractability, even if globally significant.

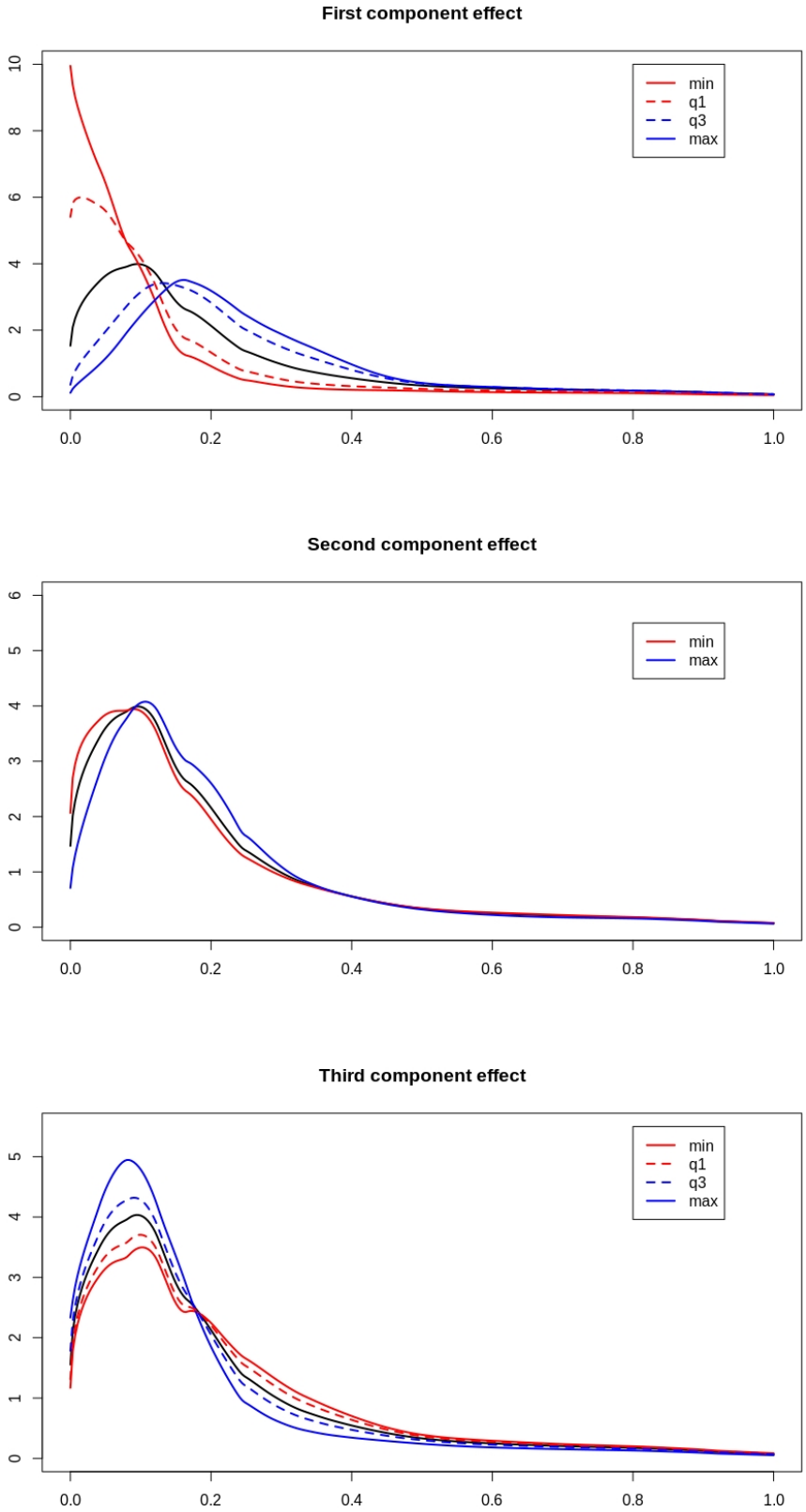


Figure 4.37: First, Second & Third principal component effects

Attribute effects

As previously mentioned, the variations expressed through the density functions, are visually more perceptible. This is also observed in the representation with the original variables.

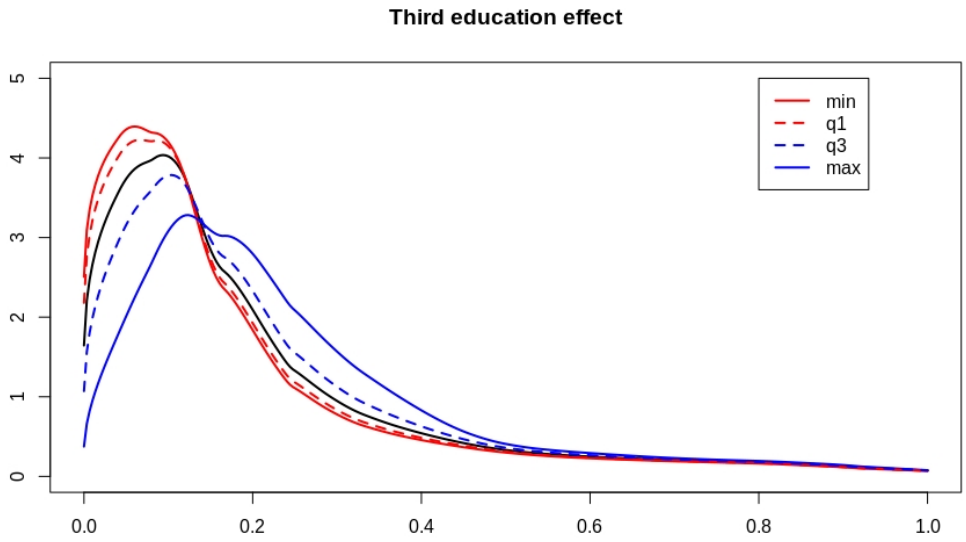


Figure 4.38: tertiary education rate effect

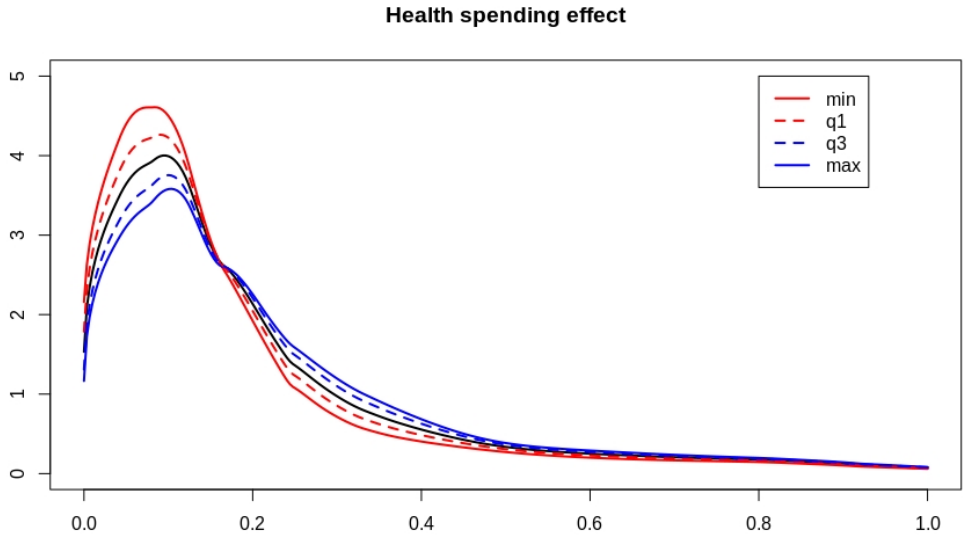


Figure 4.39: Health spending effect

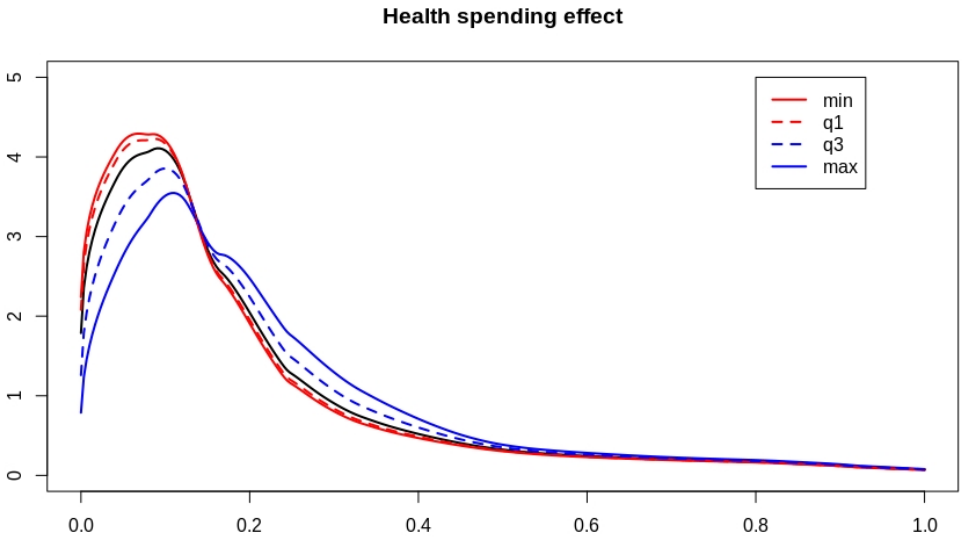


Figure 4.40: GDP PPP per capita effect

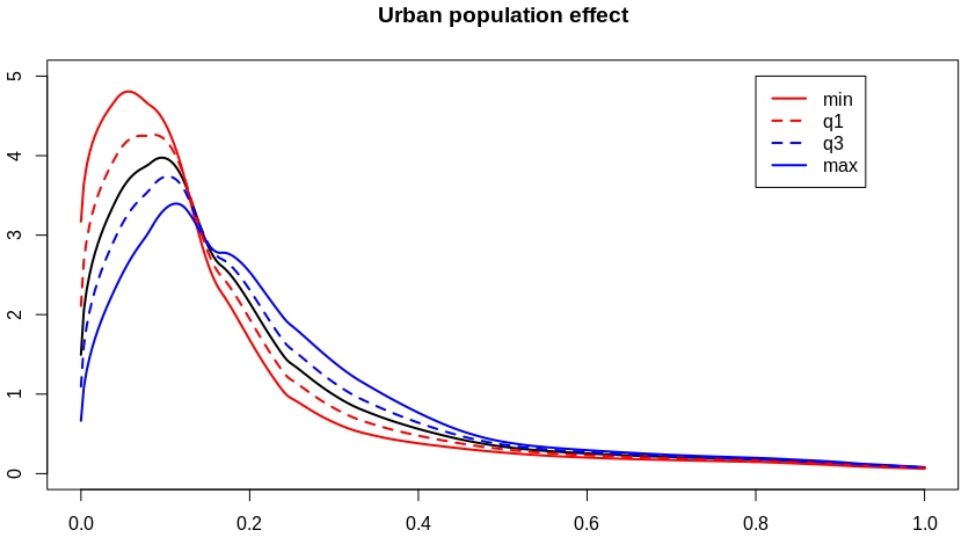


Figure 4.41: Urban population rate effect

4.3 Clustering analysis

Clustering analysis is performed to detect similar inequality profile among the counties, and to compare how different data types lead to different clusters. For both Lorenz curves and density functions, we use the transformed functions to evaluate the distances between the units. In this way, we visualized non-trivial and hardly perceptible differences in inequality. For each inequality measure, we define a distance and a linkage method, and we construct the relative dendrogram. Finally, the results are compared to see how different types of data lead to different cluster compositions, according to different inequality features that tend to emphasize. The different types of linkage methods are compared to evaluate the best one. CPCC coefficient measures how much the dendrogram capture the clustering structure of the data. As shown in the following table, the average linkage is the best possible choice, so it is used to build the dendrograms for the cluster analysis.

CPCC scores			
Linkage	Gini indices	LC″ transformations	Densities transformations
Single	0.70	0.63	0.41
Complete	0.79	0.58	0.75
Average	**0.81**	**0.82**	**0.83**
Ward	0.66	0.56	0.78

Figure 4.42: CPCC scores

4.3.1 Clustering by income data

In this section, clustering analysis is computed only on countries with income inequality data. No consumption inequality data are included.

Gini index

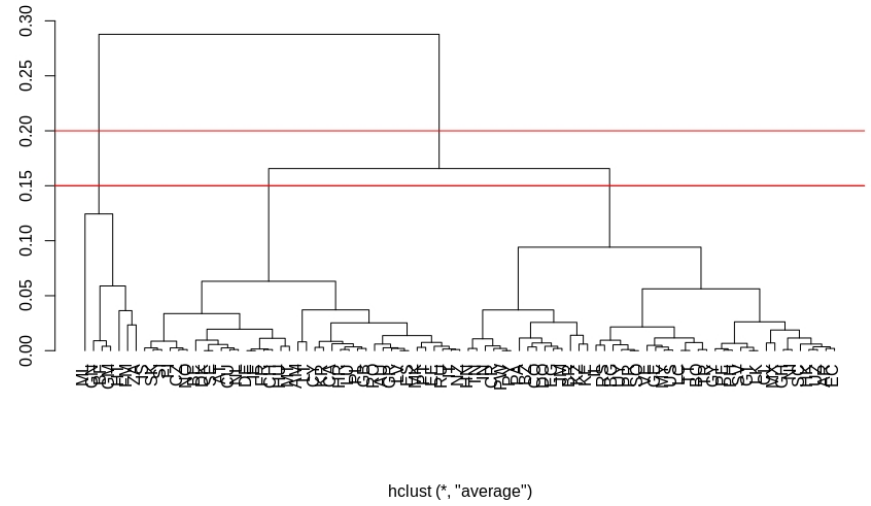


Figure 4.43: Income Gini index Dendogram with Ward linkage

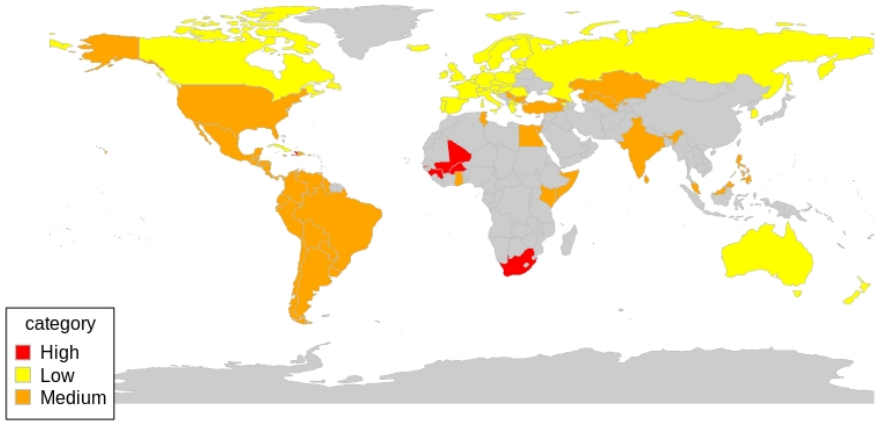


Figure 4.44: World clustering for Gini index with average linkage (3)

LC″ transformation

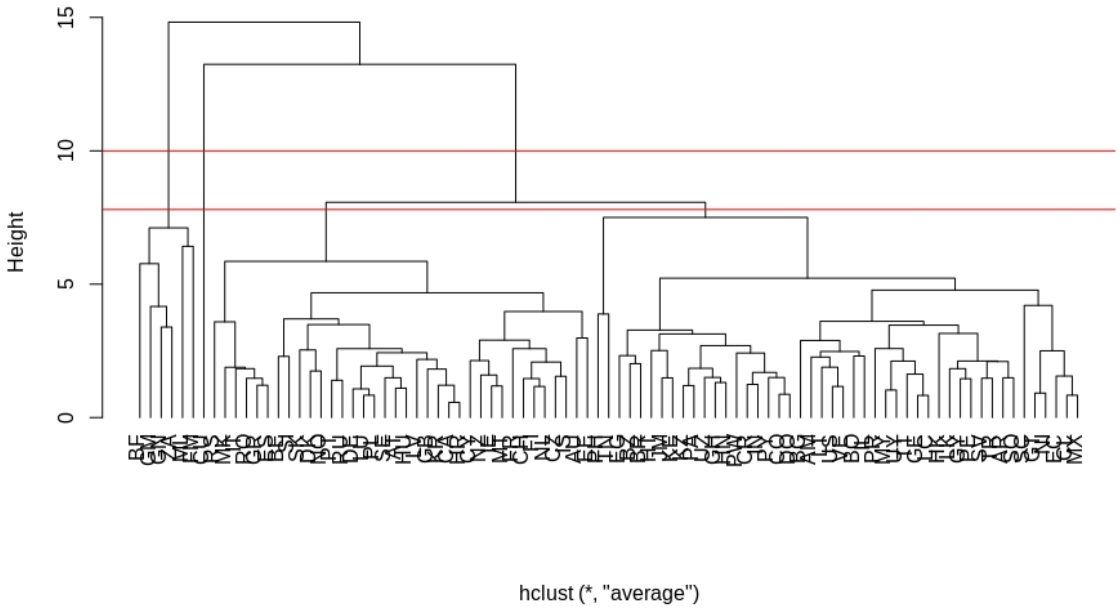


Figure 4.45: Income LC″ transformation Dendogram with average linkage

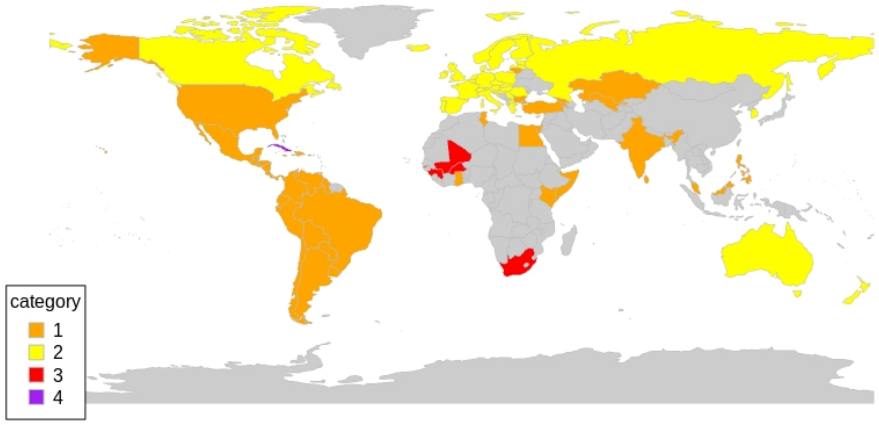


Figure 4.46: World clustering for LC″ transformation with average linkage (4)

Density transformation

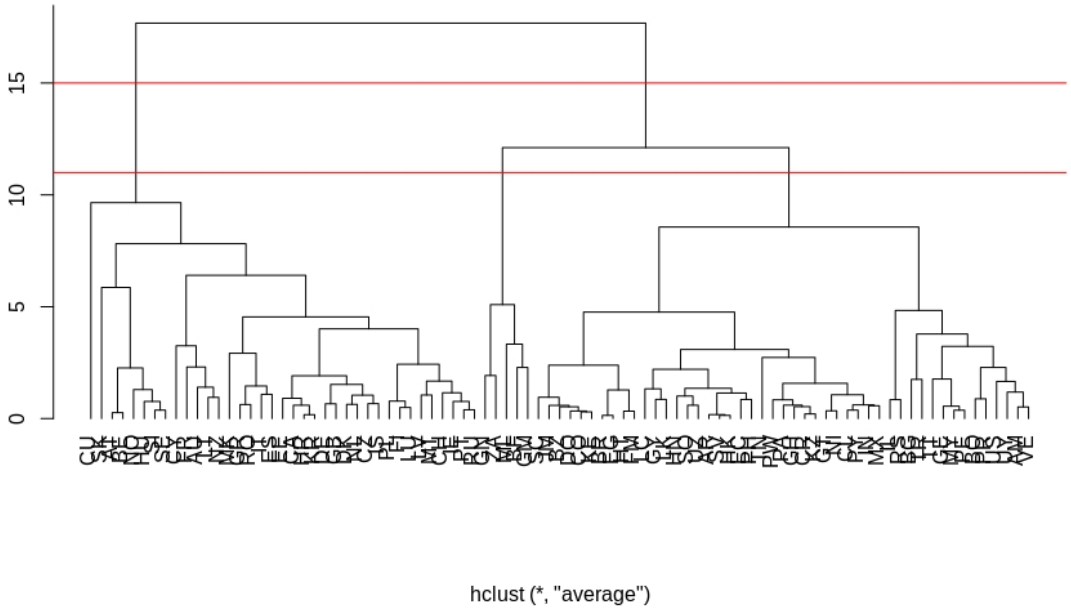


Figure 4.47: Income density transformation Dendogram with average linkage

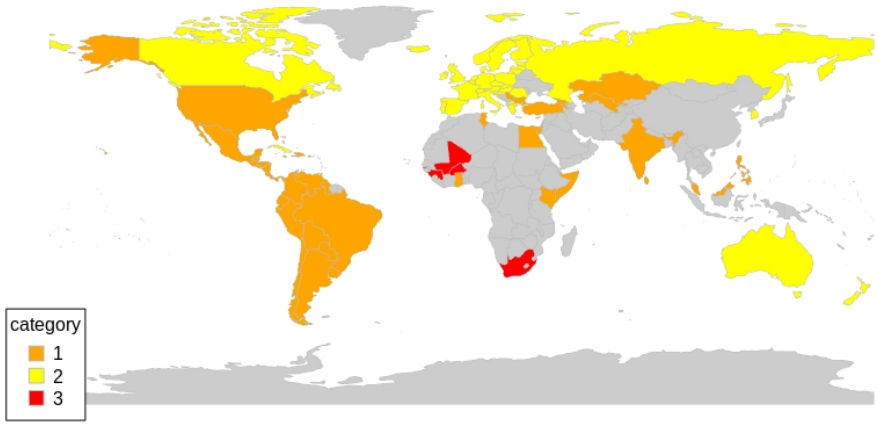


Figure 4.48: World clustering for density transformation with average linkage (3)

Comments

The available income data cover half of the countries. Furthermore, the distribution is not uniform on the world map. Almost all of Europe and America are covered, while there is a lack of data for continents as Africa and Asia.

Comparing the World maps with 3 main groups, the clustering seems to be very similar, differing only for few cases.

An interesting observation can be done switching to a two main groups representation. The Gini indices and the Lorenz curves tend to split the African states from the rest of the world, while, the density function split Europe, Canada and Australia from American, African and Asian countries.

At last, Cuba represents an interesting case. By analysing the Lorenz curves, it is a cluster of its own. Also looking at the density dendrogram, Cuba appears an isolated group, but with less evidence than Lorenz curves.

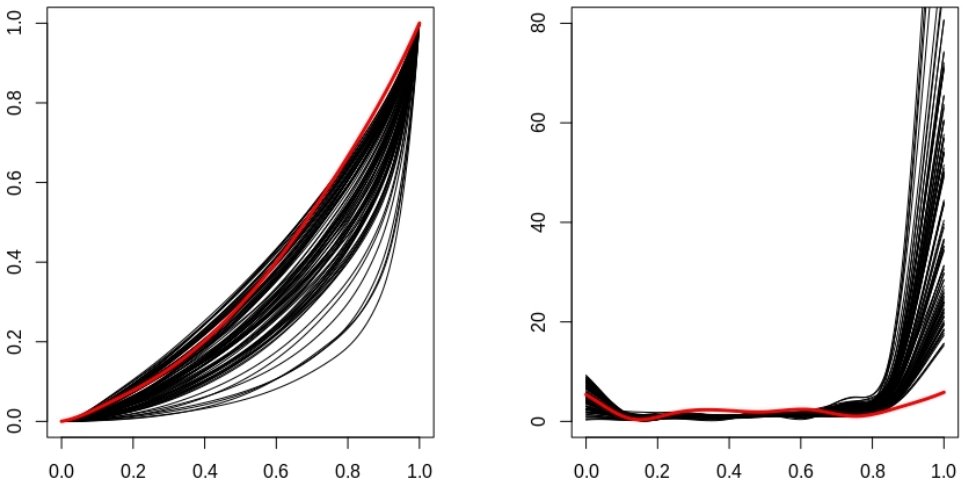


Figure 4.49: LC and LC" for Cuba

As can be seen from the plot of the Lorenz curve second derivatives, Cuba has a very low right tail that clearly distinguishes it from all the other countries. This can also be seen in the Lorenz curve. This means that, in Cuba, even among the richest population, there is no high local inequality.

4.3.2 Clustering by regression model prediction

In this section, clustering analysis is computed on the regression model predictions. For all the cases, the assumption that the dummy variable about income/consumption is equal to 0, is taken. So, income inequality predictions are considered.

Gini index

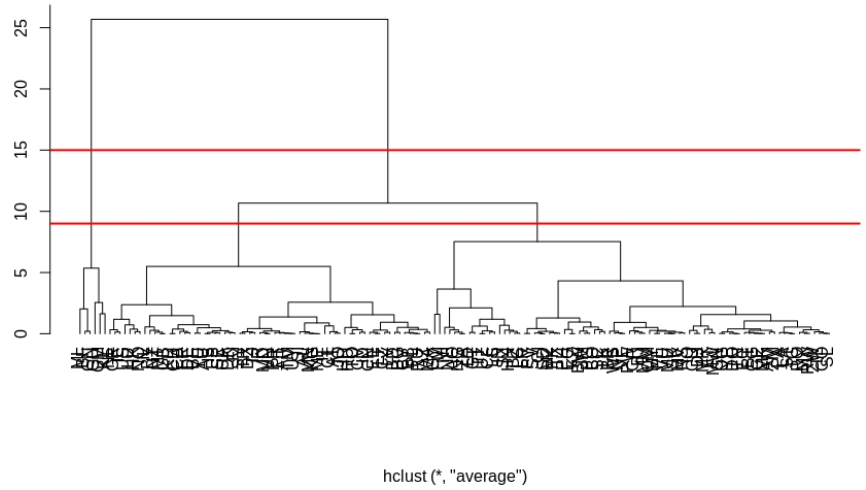


Figure 4.50: Dendogram for Gini indices predictions with average linkage

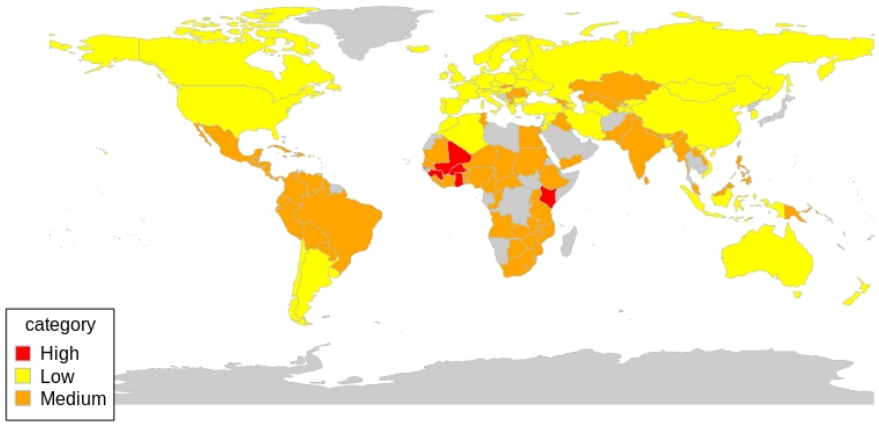


Figure 4.51: World clustering for Gini indices predictions with average linkage (3)

LC″ transformation

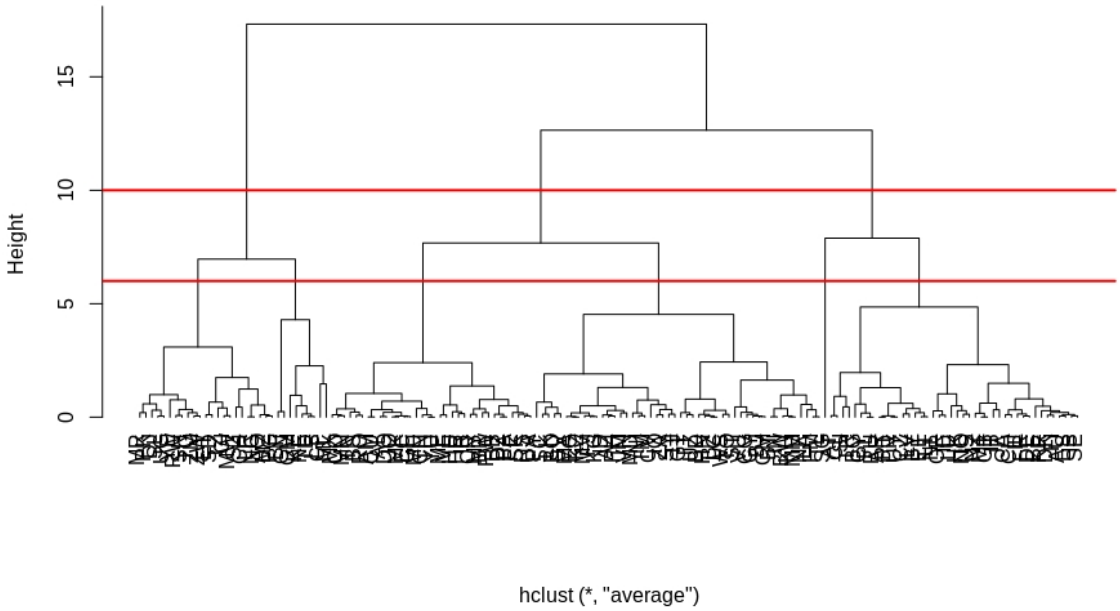


Figure 4.52: Dendogram for LC″ transformation predictions with average linkage

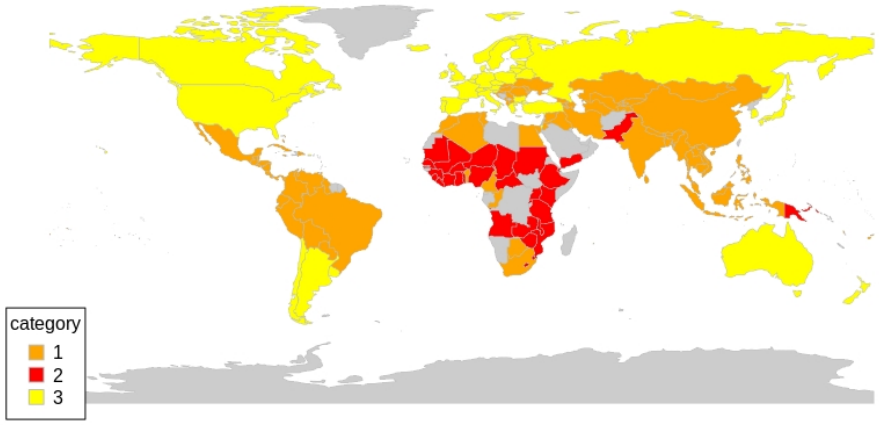


Figure 4.53: World clustering for LC″ transformation predictions with average linkage (3)

Density transformation

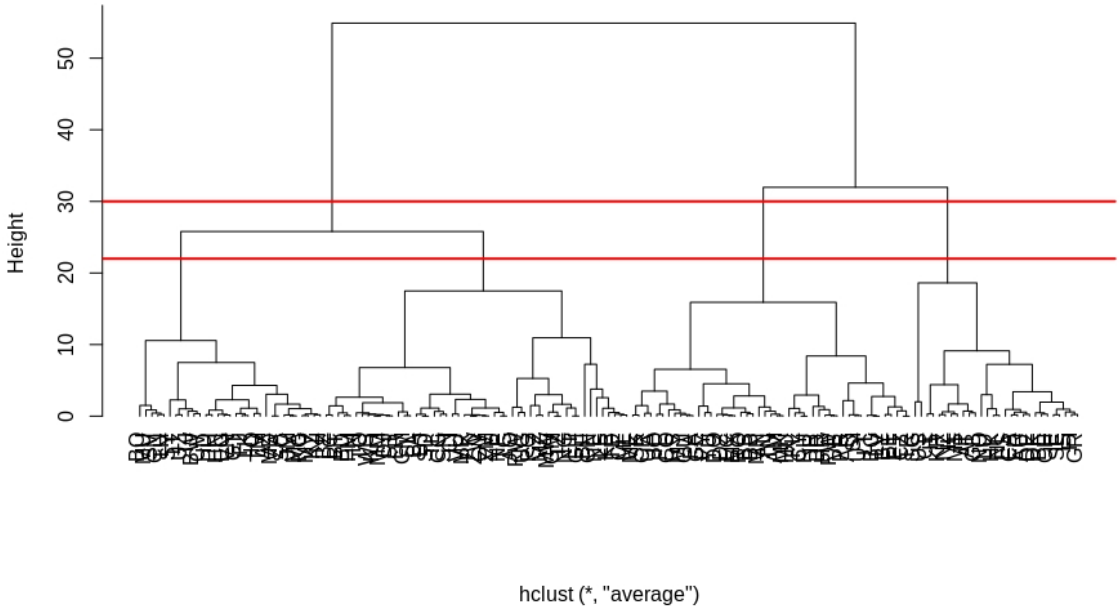


Figure 4.54: Dendogram for density transformation predictions with average linkage

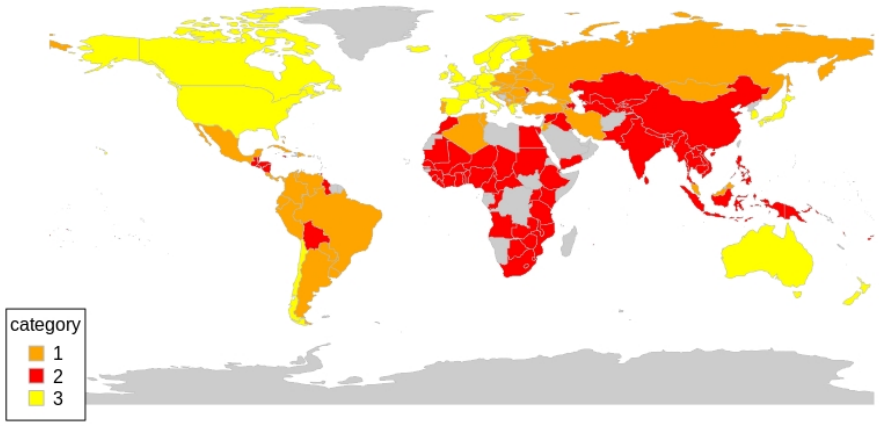


Figure 4.55: World clustering for density transformation predictions with average linkage (3)

Comments

With a first look at dendrograms and world plots, some details can be noticed. Unlike the previous cases, the 3 methods lead to clusters quite different. By focusing on mini clusters, the three dendrograms merge states in the same way. The main difference between the methods is how these groups are clustered cluster together to identify the most significant ones.

In Gini index clustering, there exists a mini cluster of 6 Central African countries separate from the rest of the predictions. The same group appears in the Lorenz curve dendrogram, although with less relevance and less different from other African countries. In fact, this is only detected on a deeper level. The reason of this isolated group can be explained by the different temporal detection of the sampling. For some African states, such as Mali, sampling was made in the late 1990s, when primary and secondary education indices, but also others variables, were significantly lower than the current ones. This can lead to predictions very influenced by this aspect.

Density functions lead to a completely different clustering. This is very interesting because confirms the fact that both Lorenz curves, through their second derivatives, and density functions capture inequality, but pointing out different aspects. A 3-cluster representation merges more countries with a high inequality into a single cluster than the Lorenz curves, and it tends to separate better countries with less inequality. Getting down of one level, a first subdivision is made between countries with a greater inequality.

Chapter 5

Conclusions

In this section we draw conclusions on the results obtained and the improvements compared to the univariate analysis of income inequality. Then, we show the main limitations of the approach used in the thesis and propose possible developments.

5.1 Conclusions

The study clearly shows how the introduction of functional elements gives a more correct and complete inequality representation than the Gini index which tends to generalize many aspects.

This thesis work has first entailed the application of smoothing procedure to pass from the quantiles to the Lorenz curves. To satisfy positivity, monotonicity and convexity constraints we use b-spline basis with modified coefficients. Since both Lorenz curves and density functions are embedded functions, we map the data in Hilbert spaces through `clr` and `nclr` transformations to make consistent analyses.

The analyses on the Lorenz curves clearly show that inequality is not distributed equally in population, as instead implicit in Gini Index, but it is differently distributed within the population. Infinite-dimensional approach avoids a loss of information due to the description of income inequality through an univariate measure. In the exploratory analyses, we found that the income profiles and local inequality among different factors, differ only in some population ranges. In this regard, the use of ITW procedure precisely detects the intervals with significant differences. Even in the functional regression model, the use of functional objects captures the effects of regressors on the local income distribution. The functional Principal Component Analysis highlights that not only the inequality is not homogeneous among all the strata, but also its variation between countries affects mainly few population ranges. Gini index is inadequate in describing this aspect because an increasing indicates a greater inequality within and between all the ranges. With the introduction of the density function, we have an additional tool to represent inequality, much more visually interpretable than Lorenz curves. This choice takes value because densities and Lorenz curves point out different aspects in inequality.

Clustering analysis is an extra. Due to the low predictive level of the regression models, clustering on predictions is not consistent, as well as the obtained clustering. Instead, we get more relevant and interesting results from the comparison of the different clustering on real data, but it is related to a limited number of countries and related to different years.

5.2 Limitations and further developments

The main limitation in using relative measures to describe inequality, is the inability to catch changes in income at the absolute level. These representations violate Pareto principle, according to which, income inequalities can increase with an increase in all incomes in a society, without that the inequality measure changes. Both the Gini index, Lorenz curve and the density function have this limit.

Some suggestions are reported to anyone who wants to get ideas from this thesis to perform further insights or related research.

- Find a basis for the smoothing process which minimize the oscillations in Lorenz curve second derivatives. Another possible way is to have more points available in the smoothing.
- Introduce more significant variables in the regression to have a better fitting.
- For all the countries, use data related to the same year or close years. This would give clustering analysis more significance and interpretability.
- Use higher quality data. Today, it is difficult to find many data relating to income inequality quantiles, especially for least development countries. In the following years, it is plausible that this type of data will be available for all countries and therefore there will be the possibility to make analyses using the income inequality indices for each country. If the samplings are constantly made over the years, a possible development of this project can be to study the evolution of the inequality curves over the years with a time series approach.
- Analyze in detail the difference between income and consumption inequality and find all the factors that determine it.

Appendices

Appendix A

Extra clustering with Average linkage

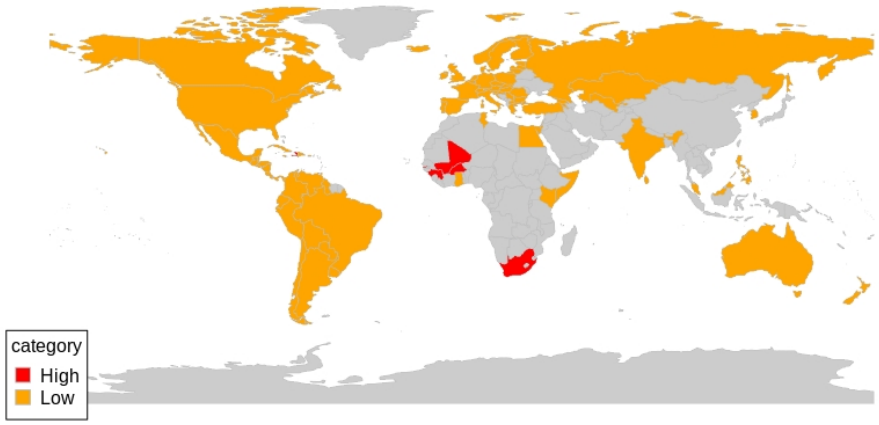


Figure A.1: World clustering for Gini index with average linkage (2)

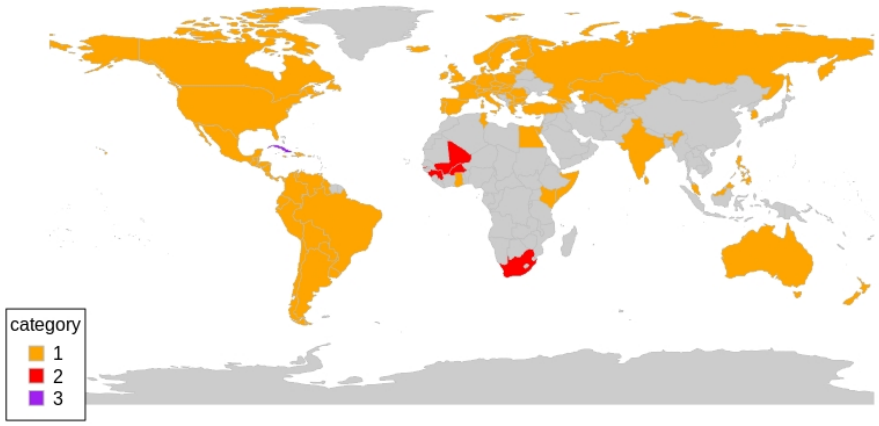


Figure A.2: World clustering for LC″ transformation with average linkage (3)

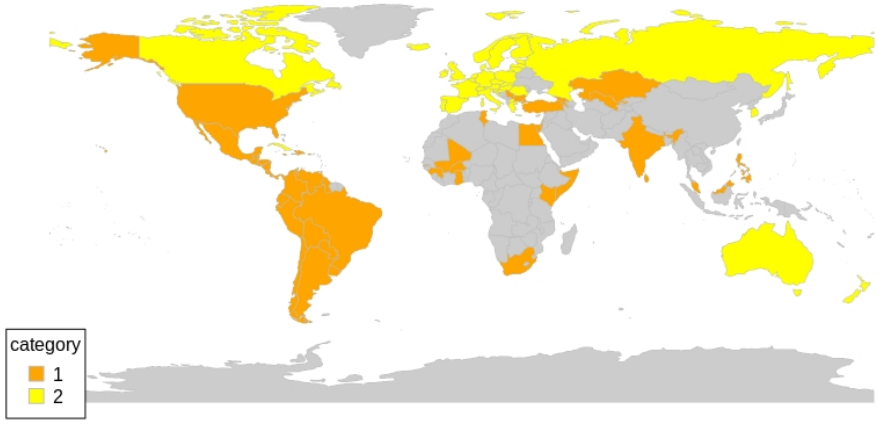


Figure A.3: World clustering for density transformation with average linkage (2)

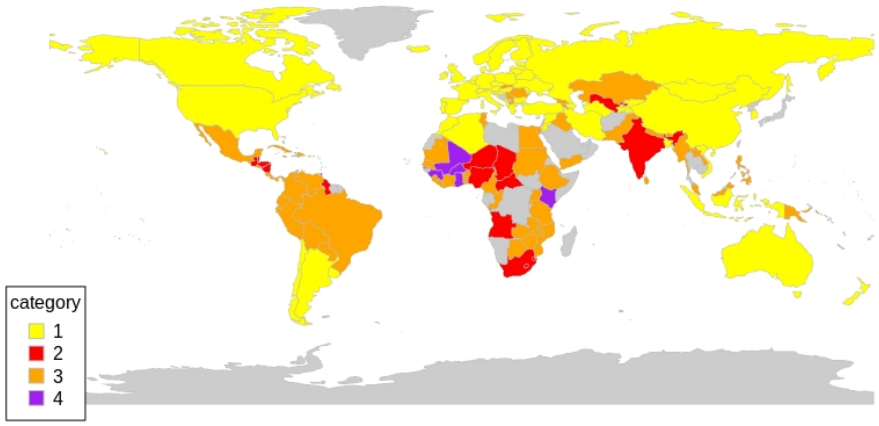


Figure A.4: World clustering for Gini indices predictions with average linkage (4)

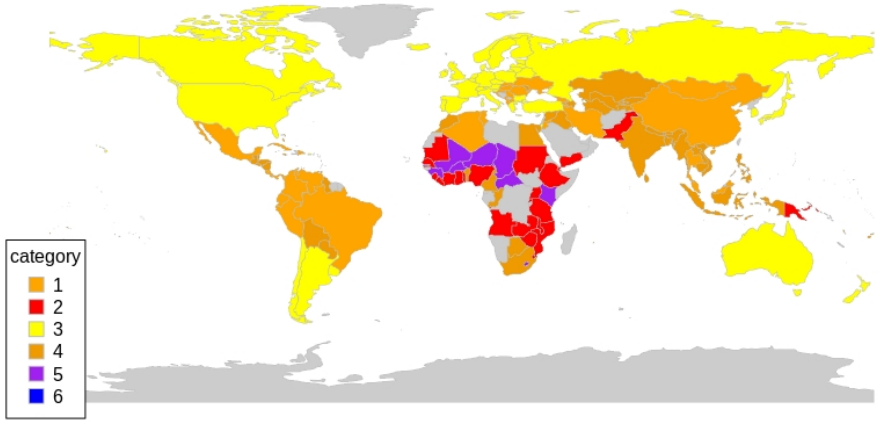


Figure A.5: World clustering for LC″ transformation predictions with average linkage (6)

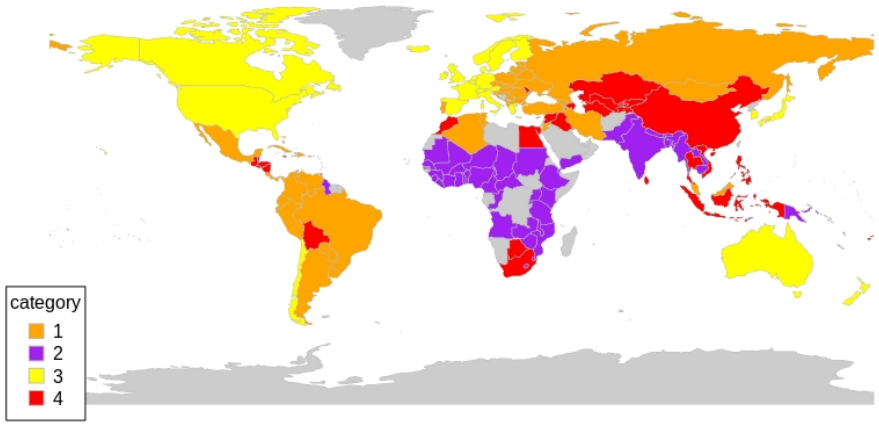


Figure A.6: World clustering for density transformation predictions with average linkage (4)

Appendix B

Extra summaries for exploratory analyses

ANOVA by continents for Lorenz curves

```
$call
IWTlmFoF(formula = y1 ~ Africa + America + Asia + Oceania)

$ttest
            Minimum p-value
(Intercept)           0.000 ***
Africa                0.000 ***
America               0.000 ***
Asia                  0.000 ***
Oceania               0.003  **

$R2
              Range of functional R-squared
Min R-squared                    0.01934693
Max R-squared                    0.67839005

$ftest
  Minimum p-value
1               0 ***
```

```
$call
IWTlmFoF(formula = y1 ~ Europa + America + Asia + Oceania)

$ttest
            Minimum p-value
(Intercept)           0.000 ***
Europa                0.000 ***
America               0.000 ***
Asia                  0.000 ***
Oceania               0.001  **

$R2
              Range of functional R-squared
Min R-squared                    0.01934693
Max R-squared                    0.67839005

$ftest
  Minimum p-value
1               0 ***
```

Figure B.1: LC ANOVA by continents wrt Europe and Africa

```
$call
IWTlmFoF(formula = y1 ~ Europa + Africa + Asia + Oceania)

$ttest
            Minimum p-value
(Intercept)           0.000 ***
Europa                0.000 ***
Africa                0.000 ***
Asia                  0.327
Oceania               0.602

$R2
              Range of functional R-squared
Min R-squared                    0.01934693
Max R-squared                    0.67839005

$ftest
  Minimum p-value
1               0 ***
```

```
$call
IWTlmFoF(formula = y1 ~ Europa + Africa + America + Oceania)

$ttest
            Minimum p-value
(Intercept)           0.000 ***
Europa                0.000 ***
Africa                0.000 ***
America               0.298
Oceania               0.348

$R2
              Range of functional R-squared
Min R-squared                    0.01934693
Max R-squared                    0.67839005

$ftest
  Minimum p-value
1               0 ***
```

Figure B.2: LC ANOVA by continents wrt America and Asia

ANOVA by GDP PPP ranges for Lorenz curves

```
$call
IWTlmFoF(formula = y1 ~ MediumGDP + LowGDP)

$ttest
            Minimum p-value
(Intercept)           0.000 ***
MediumGDP             0.019   *
LowGDP                0.000 ***

$R2
             Range of functional R-squared
Min R-squared                    0.001584836
Max R-squared                    0.502001566

$ftest
  Minimum p-value
1               0 ***
```

```
$call
IWTlmFoF(formula = y1 ~ HighGDP + LowGDP)

$ttest
            Minimum p-value
(Intercept)           0.000 ***
HighGDP               0.010   *
LowGDP                0.001  **

$R2
             Range of functional R-squared
Min R-squared                    0.001584836
Max R-squared                    0.502001566

$ftest
  Minimum p-value
1               0 ***
```

Figure B.3: LC ANOVA by GDP PPP wrt high and medium values

ANOVA by continents for densities

```
$call
IWTlmFoF(formula = y1 ~ Africa + America + Asia + Oceania)

$ttest
            Minimum p-value
(Intercept)           0.000 ***
Africa                0.000 ***
America               0.000 ***
Asia                  0.000 ***
Oceania               0.028   *

$R2
             Range of functional R-squared
Min R-squared                   0.08007998
Max R-squared                   0.69899307

$ftest
  Minimum p-value
1               0 ***
```

```
$call
IWTlmFoF(formula = y1 ~ Europe + America + Asia + Oceania)

$ttest
            Minimum p-value
(Intercept)           0.000 ***
Europe                0.000 ***
America               0.022   *
Asia                  0.001  **
Oceania               0.041   *

$R2
             Range of functional R-squared
Min R-squared                   0.08007998
Max R-squared                   0.69899307

$ftest
  Minimum p-value
1               0 ***
```

Figure B.4: Density ANOVA by continents wrt Europe and Africa

```
$call
IWTlmFoF(formula = y1 ~ Europe + Africa + Asia + Oceania)

$ttest
           Minimum p-value
(Intercept)           0.000 ***
Europe                0.000 ***
Africa                0.021   *
Asia                  0.190
Oceania               0.542

$R2
             Range of functional R-squared
Min R-squared                    0.08007998
Max R-squared                    0.69899307

$ftest
  Minimum p-value
1               0 ***
```

```
> summary(model1)
$call
IWTlmFoF(formula = y1 ~ Europe + Africa + America + Oceania)

$ttest
           Minimum p-value
(Intercept)           0.000 ***
Europe                0.000 ***
Africa                0.002  **
America               0.181
Oceania               0.966

$R2
             Range of functional R-squared
Min R-squared                    0.08007998
Max R-squared                    0.69899307

$ftest
  Minimum p-value
1               0 ***
```

Figure B.5: Density ANOVA by continents wrt America and Asia

ANOVA by GDP PPP ranges for densities

```
$call
IWTlmFoF(formula = y1 ~ MediumGDP + LowGDP)

$ttest
             Minimum p-value
(Intercept)            0.000 ***
MediumGDP              0.008  **
LowGDP                 0.000 ***

$R2
               Range of functional R-squared
Min R-squared                    0.008357288
Max R-squared                    0.611349152

$ftest
  Minimum p-value
1               0 ***
```

```
$call
IWTlmFoF(formula = y1 ~ HighGDP + LowGDP)

$ttest
             Minimum p-value
(Intercept)            0.000 ***
HighGDP                0.007  **
LowGDP                 0.022   *

$R2
               Range of functional R-squared
Min R-squared                    0.008357288
Max R-squared                    0.611349152

$ftest
  Minimum p-value
1               0 ***
```

Figure B.6: Density ANOVA by GDP PPP wrt high and medium values

Appendix C

Analysis of Gini indices

In this appendix, all the analyses are computed to Gini index.

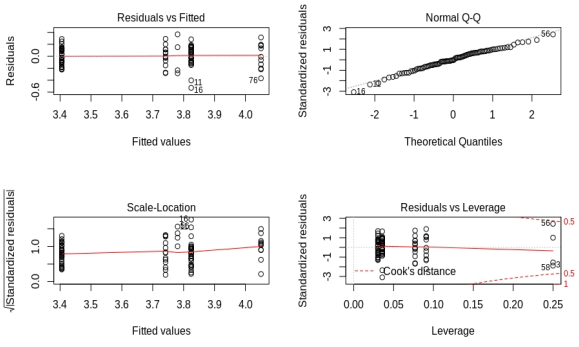
1-way ANOVA on Continent factor

In order to respect the basic hypothesis of ANOVA model and linear regression, the Gini index logarithmic transformation is considered. After checking the normality of the data in the different groups and assuming a homogeneous variance between them (Bartlett-test p-value: 0.15), the 1-way ANOVA is applied:

- p-value ANOVA $< 2.3 * 10^{-16}$

There exists a significant difference between different continents.
In order to have information about the coefficients, a linear model with dummy variables is computed:

```
lm(formula = log(y1) ~ region_un)

Residuals:
     Min       1Q   Median       3Q      Max
-0.52879 -0.10862 -0.00628  0.12971  0.36667

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        4.05084    0.05244  77.241  < 2e-16 ***
region_unAmericas -0.22621    0.06189  -3.655 0.000447 ***
region_unAsia     -0.30927    0.07126  -4.340 3.95e-05 ***
region_unEurope   -0.64541    0.06056 -10.658  < 2e-16 ***
region_unOceania  -0.27026    0.10156  -2.661 0.009329 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1739 on 84 degrees of freedom
Multiple R-squared:  0.6478,    Adjusted R-squared:  0.6311
F-statistic: 38.63 on 4 and 84 DF,  p-value: < 2.2e-16
```



To validate the model, in addition to the plots, the following tests are done:

- Shapiro-Wilk normality test: p-value = 0.44.

- Breusch-Pagan test (homoscedasticity of residues): p-value = 0.11.

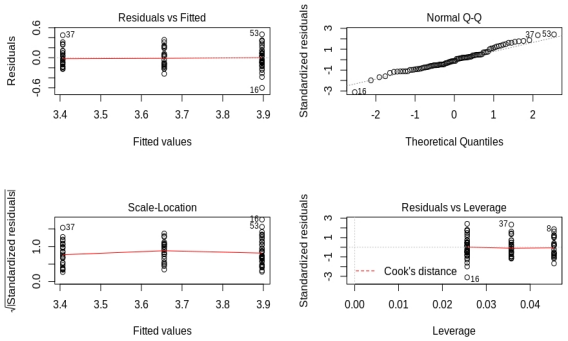
1-way ANOVA on GDP PPP range

As the previous case, a Gini index logarithmic transformation is applied. The 3 GDP PPP per capita groups are the same used in functional analyses.
After checking the data normality in the different groups and assuming a homogeneous variance between them (Bartlett-test p-value: 0.62), the 1-way ANOVA is computed:

- p-value ANOVA = $1.92 * 10^{-15}$

This result indicates a significant difference between the three groups.
In order to have a information about the coefficients, a linear model with dummy variables is computed:

```
Call:
lm(formula = log(y2) ~ gruppi)

Residuals:
     Min       1Q   Median       3Q      Max
-0.60083 -0.13272 -0.02225  0.09417  0.46771

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.40664    0.03692  92.280  < 2e-16 ***
gruppiLow GDP     0.49003    0.04839  10.127 2.51e-16 ***
gruppiMedium GDP  0.24950    0.05565   4.483 2.26e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1953 on 86 degrees of freedom
Multiple R-squared:  0.5453,    Adjusted R-squared:  0.5347
F-statistic: 51.56 on 2 and 86 DF,  p-value: 1.922e-15
```



To validate the model, the following tests are done:

- Shapiro-Wilk normality test: p-value = 0.19.

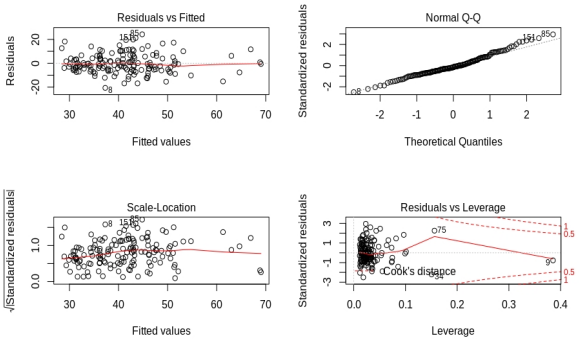
- Breusch-Pagan test (homoscedasticity of residues): p-value = 0.52.

Linear regression model

Linear regression does not need data transformation, so the model response is Gini index.
Starting from a complete model, step-by-step, the least significant variable is dropped

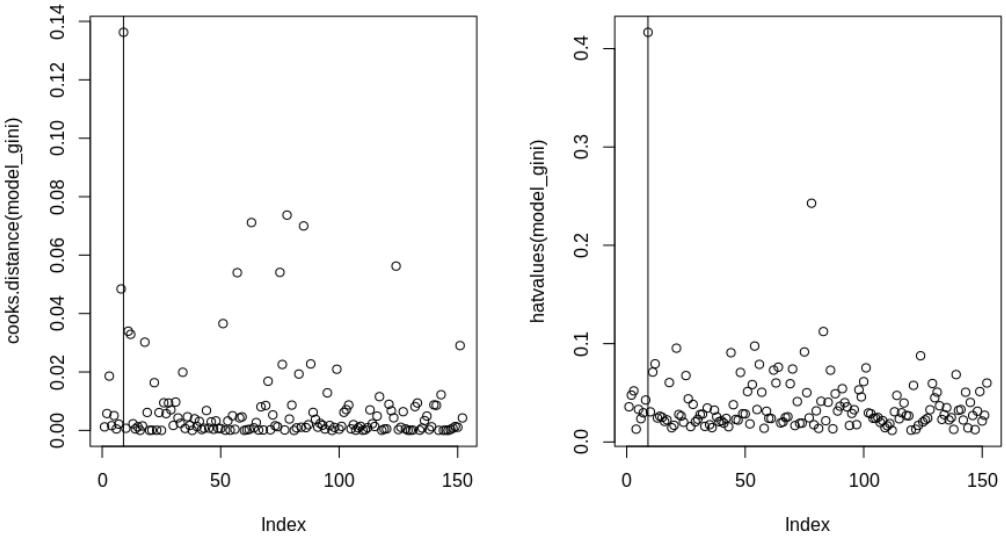
by the model until only significant ones are considered.
The final model is:

```
Call:
lm(formula = y3 ~ x_cons + pca1 + pca2 + x_cons:x6)

Residuals:
    Min      1Q  Median      3Q     Max
-20.649  -5.341  -1.317   4.632  24.278

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.601e+01  9.795e-01  46.970  < 2e-16 ***
x_cons      -1.481e+01  2.265e+00  -6.539 9.59e-10 ***
pca1        -4.635e+00  3.968e-01 -11.681  < 2e-16 ***
pca2        -1.919e+00  7.017e-01  -2.734  0.00702 **
x_cons:x6    6.017e-04  1.979e-04   3.040  0.00280 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.307 on 147 degrees of freedom
Multiple R-squared:  0.4876,    Adjusted R-squared:  0.4737
F-statistic: 34.97 on 4 and 147 DF,  p-value: < 2.2e-16
```



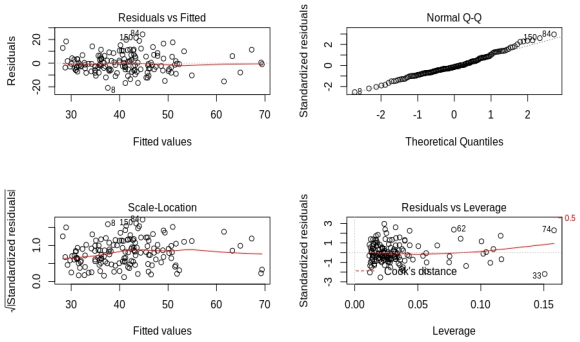
xcons is a dummy variable to identify consumption Gini indices.

Focusing on leverage and cook's distance plot, there is a observation with an high leverage value and a significant cook's distance.



Fitting a linear model without the observation, the following model is obtained:

```
Call:
lm(formula = y3 ~ x_cons + pca1 + pca2 + x_cons:x6)

Residuals:
    Min      1Q  Median      3Q     Max
-20.923  -5.358  -1.193   5.000  24.368

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.606e+01  9.833e-01  46.841  < 2e-16 ***
x_cons      -1.552e+01  2.446e+00  -6.343 2.67e-09 ***
pca1        -4.689e+00  4.035e-01 -11.621  < 2e-16 ***
pca2        -1.915e+00  7.027e-01  -2.725  0.00721 **
x_cons:x6    7.162e-04  2.476e-04   2.893  0.00440 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.318 on 146 degrees of freedom
Multiple R-squared:  0.4896,    Adjusted R-squared:  0.4756
F-statistic: 35.02 on 4 and 146 DF,  p-value: < 2.2e-16
```



To validate the model, the following tests are done:

- Shapiro-Wilk normality test: p-value = 0.06.

- Breusch-Pagan test (homoscedasticity of residues): p-value = 0.23.

Appendix D

Lorenz curve and density eigenfunctions

As already mentioned, Lorenz curves and density functions point out different aspects of inequality.
In this appendix, the variation modes of the density functions are mapped in the Lorenz curve space to compare the eigenfunctions in different spaces.

For every density mode of variation, the Lorenz curve representation is obtained through the inverse transformation.

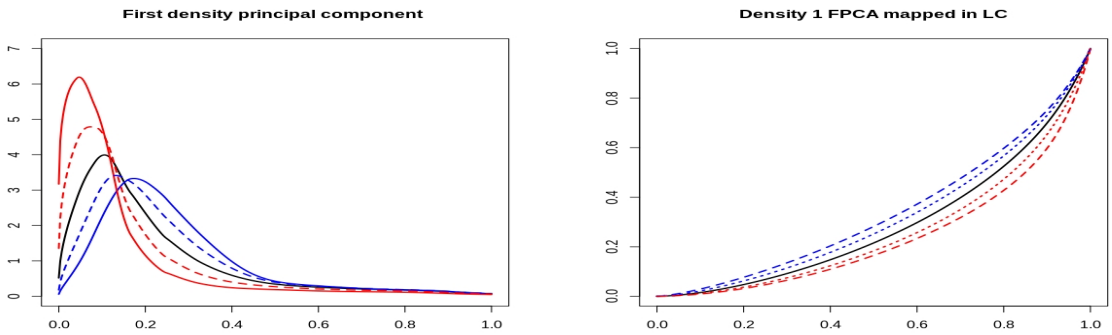


Figure D.1: 1PC mapped from density space to Lorenz curve space

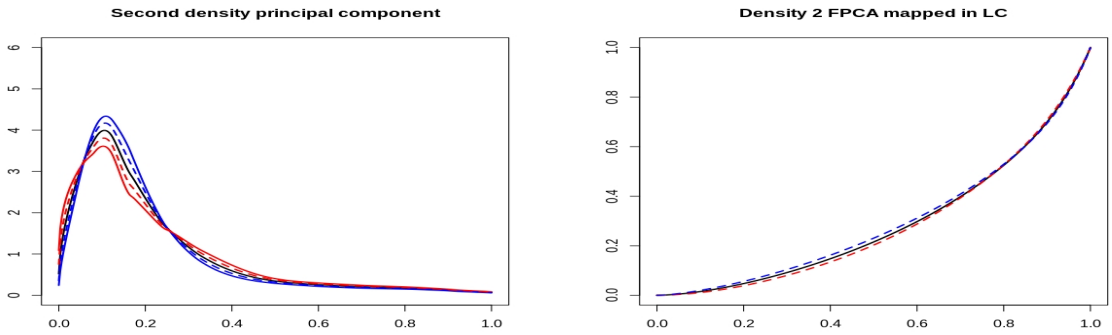


Figure D.2: 2PC mapped from density space to Lorenz curve space

Then, these modes of variation are compared with the modes computed on Lorenz curves.

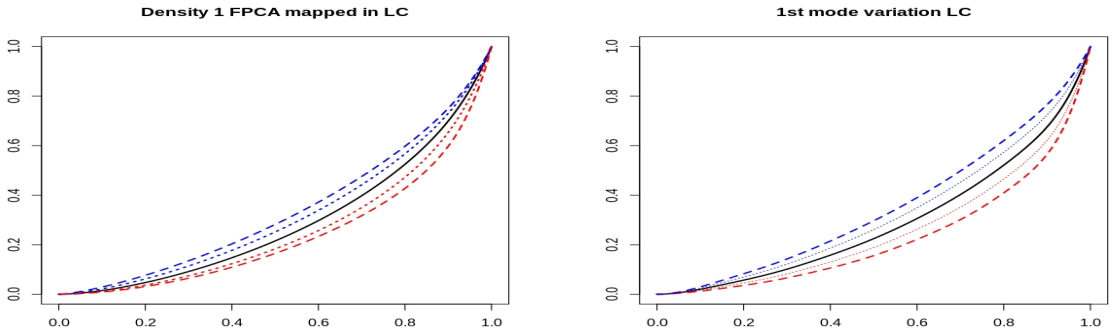


Figure D.3: comparison of first principal component between map and computation

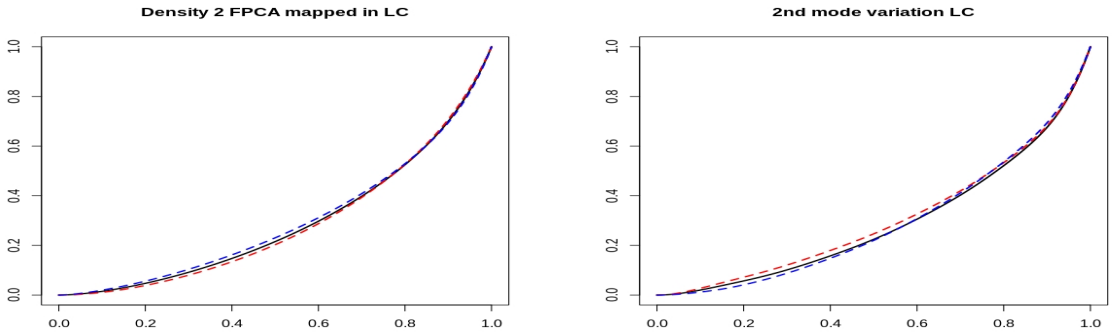


Figure D.4: comparison of second principal component between map and computation

In the Lorenz curves variation modes, the parameter k is modified in order to make the curves comparable. The variability explained appears to be quite similar.

Appendix E

Coefficients plot

GDP PPP coefficients for Lorenz curves

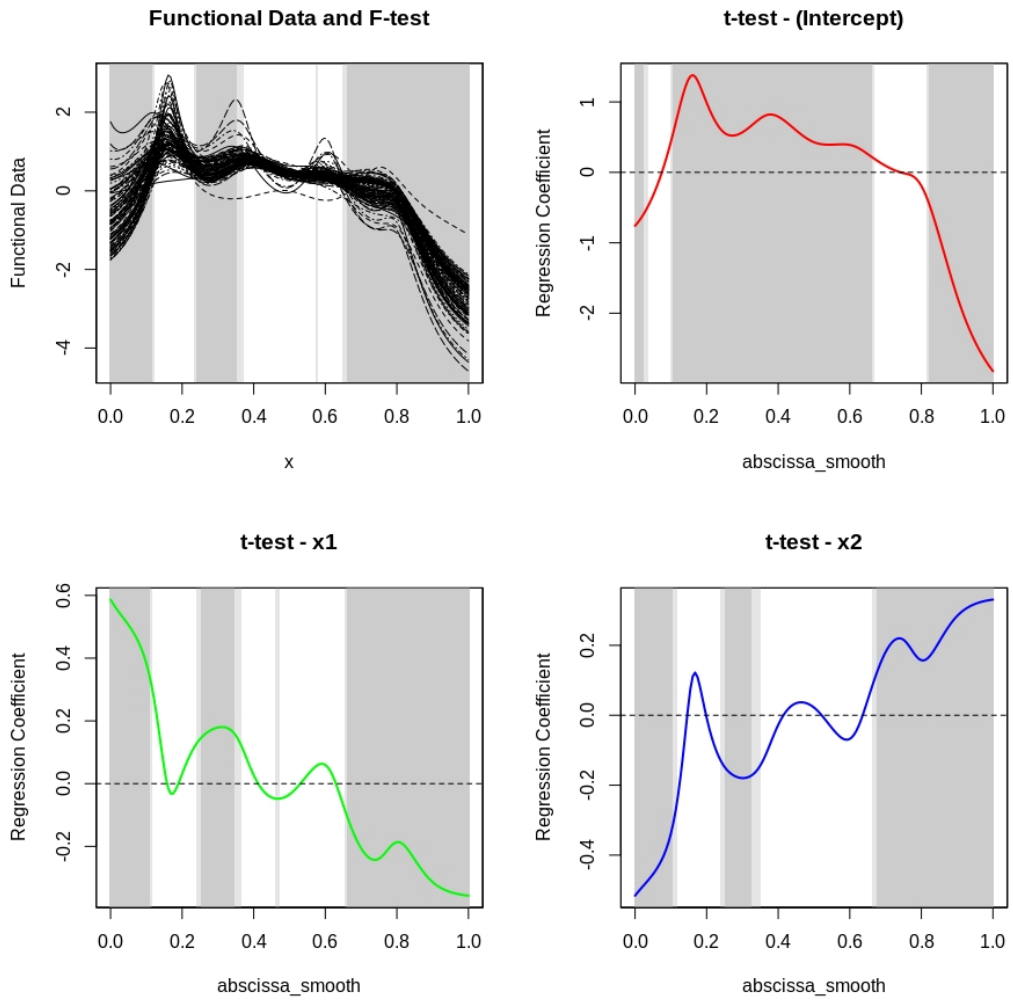


Figure E.1: GDP PPP coefficients for Lorenz curves

Continent coefficients for Lorenz curves

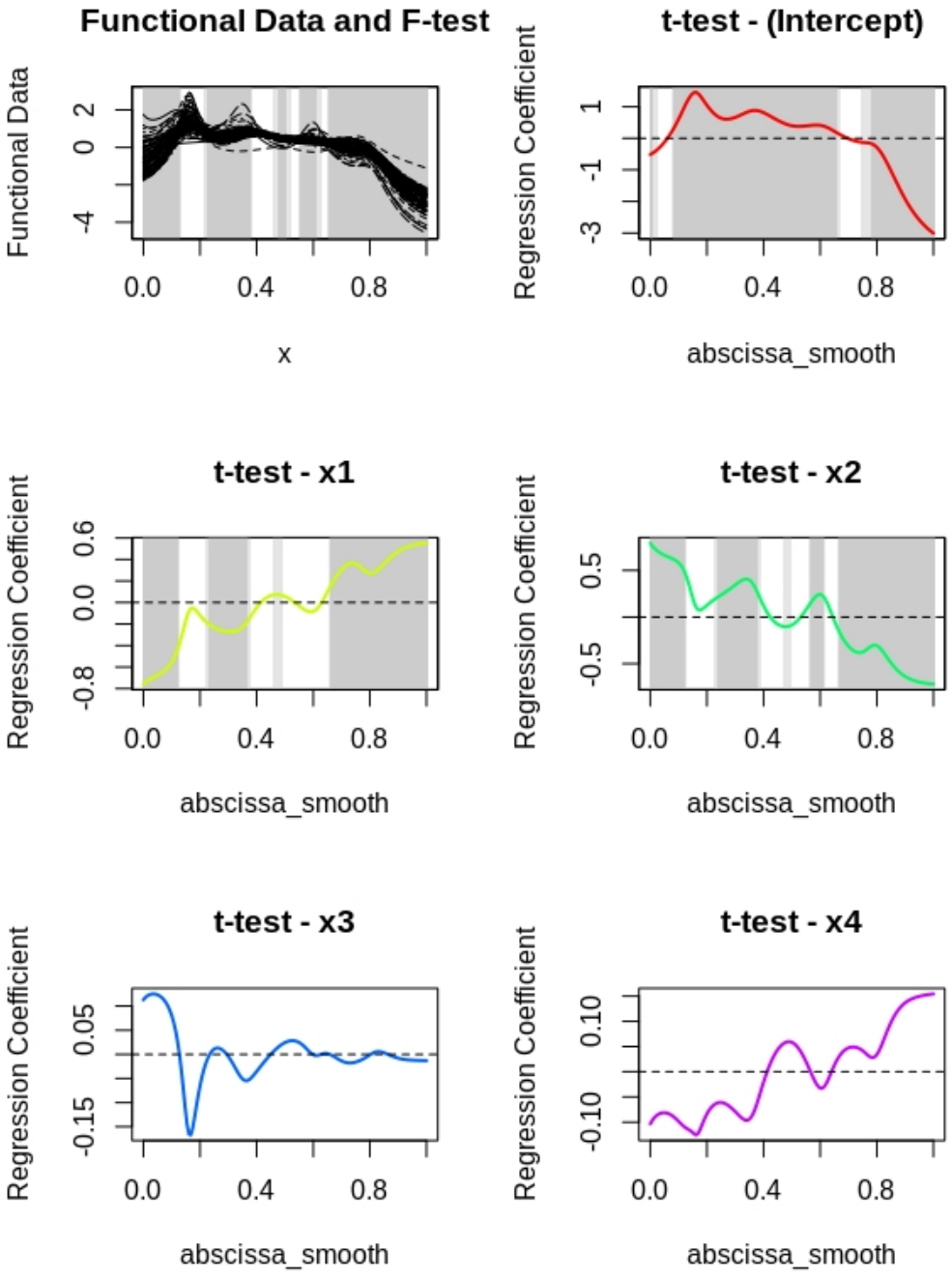


Figure E.2: Continent coefficients for Lorenz curves

GDP PPP coefficients for densities

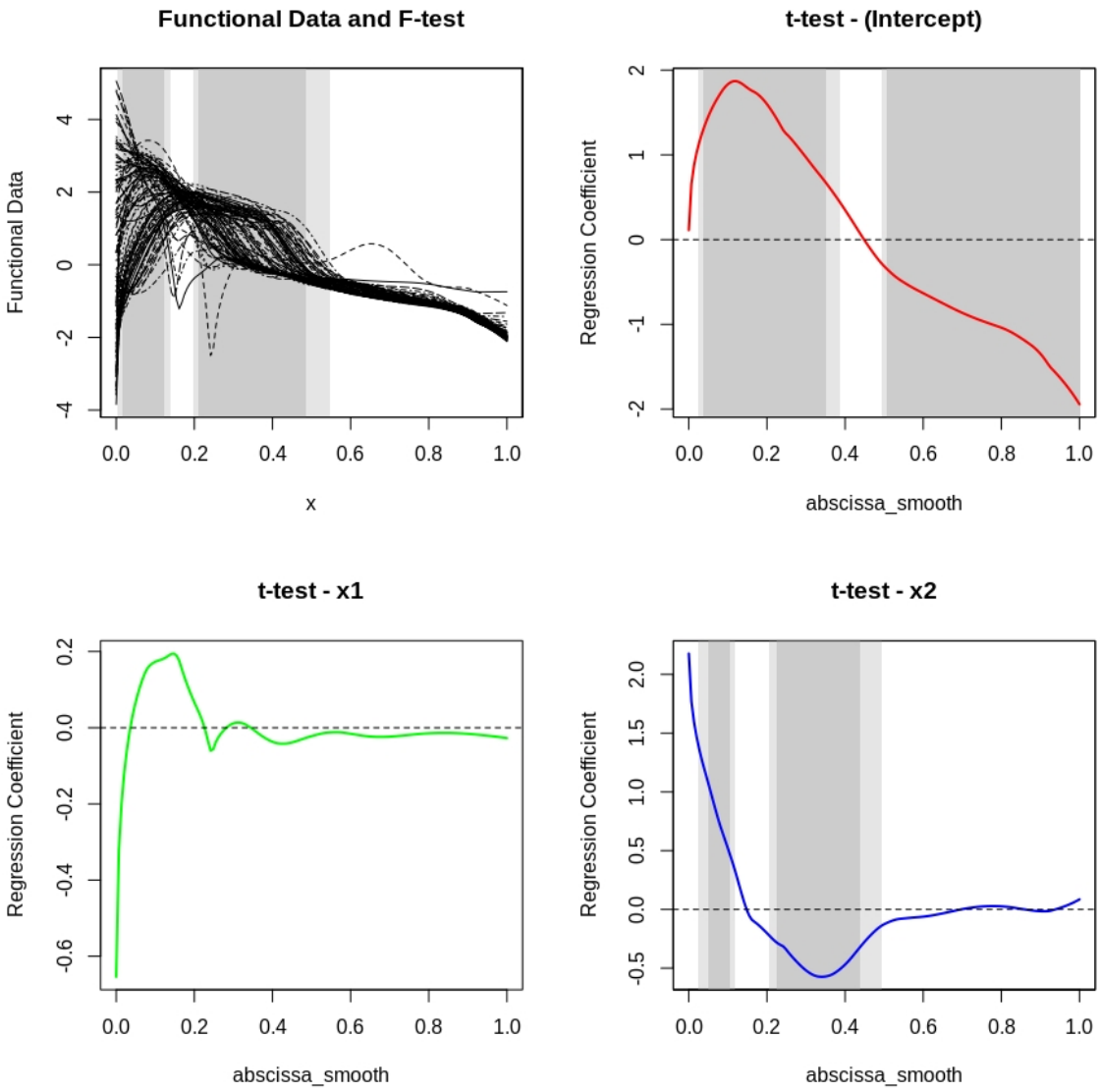


Figure E.3: GDP PPP coefficients for densities

Continent coefficients for densities

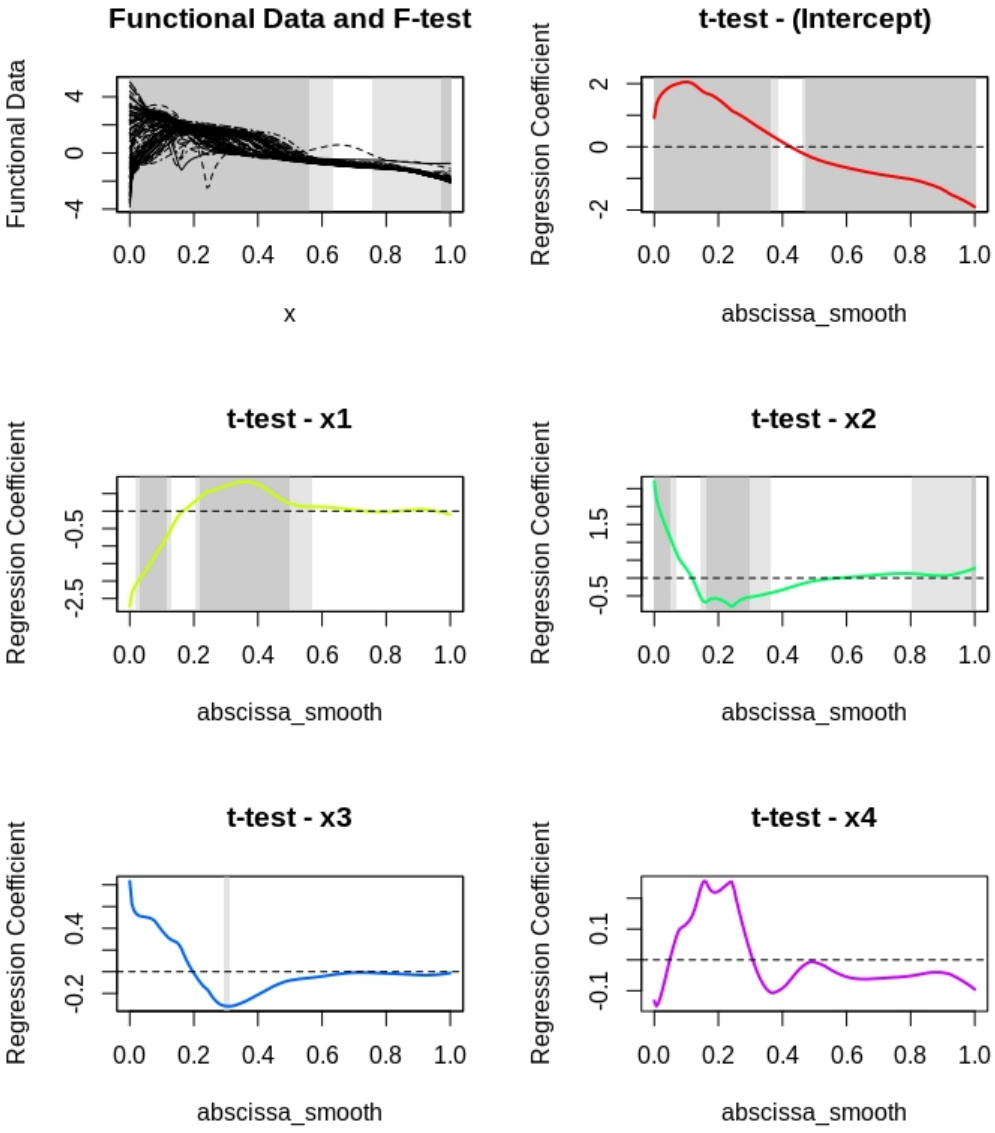


Figure E.4: Continent coefficients for densities

Ringraziamenti

Un sentito ringraziamento va al Professor Vatini, alla Professoressa Menafoglio e al Dottor Fontana per essersi messi a disposizione nel coadiuvare al meglio il mio lavoro. Un ringraziamento sentito anche a Johannes Emmerling che si è mostrato molto disponibile nel fornirmi gli spunti e i dati senza i quali probabilmente oggi non presenterei questa tesi.
Un grandissimo ringraziamento alla mia famiglia che mi ha sempre appoggiato in tutta la mia carriera studentesca, guidandomi, senza mai imporsi, nelle scelte che mi hanno portato ad arrivare dove sono oggi.
Grazie ad Albi, Ale, Cigo, Dave, Detta, Edo, Halit, Paggia e Picchia, che mi hanno accompagnato in questi anni al Politecncio, tra calcetti, fantacalcio e non solo.
Grazie a Sampa e Aldo, con i quali ho condiviso una fantastica esperienza, più che di Erasmus, di vita a Barcellona.
Grazie a tutte le fantastiche persone che ho avuto modo di conoscere in questi anni univeristari.
Grazie al gruppo "Gli gnari", compagni speciali di mille avventure con cui ho condiviso serate, cene, capodanni, compleanni, vacanze e mille altre cose.
Grazie agli amici di Villa, con i quali la statistica non era l'argomento all'ordine del giorno, ma con cui passo dei bellissimi momenti da tutta una vita.
Grazie a Re, Massi, Fius, Solaz, Alessia, Agaz e Dani che hanno condiviso con me la bellissima esperienza di vivere Corsica 33.
Grazie ai compagni del liceo, che, nonostante le poche occasioni di vederci, non hanno mai fatto mancare il loro contributo molto "caloroso" nelle nostre "Cene e ...".
GRAZIE

Bibliography

[1] Konrad Abramowicz, Charlotte Hager, Alessia Pini, Lina Schelin, Sara Luna, and Simone Vantini. "Nonparametric inference for functional-on-scalar linear models applied to knee kinematic hop data after injury of the anterior cruciate ligament". In: *Scandinavian Journal of Statistics* (July 2016) (cit. on pp. 4, 17).

[2] Olivier Blanchard. *Marcoeconomics*. Pearson, 2017 (cit. on p. 51).

[3] Enea Bongiorno and Aldo Goia. "Describing the concentration of income populations by functional principal component analysis on Lorenz curves". In: *Journal of Multivariate Analysis* 170 (Sept. 2018) (cit. on pp. 3, 4, 8–10).

[4] Ricardo Fuentes-Nieva and Nicholas Galasso. "Working for the Few". In: *Oxfam Internatinal* (2014) (cit. on p. 1).

[5] *https://en.wikipedia.org/wiki/Economic_inequality*.

[6] Michel Lubrano. "The econometrics of inequality and poverty". In: (2017) (cit. on p. 21).

[7] J. Machalová, Karel Karel Hron, and Gianna Serafina Monti. "Preprocessing of centred logratio transformed density functions using smoothing splines". In: *Journal of Applied Statistics* 43 (Dec. 2015) (cit. on pp. 3, 4, 11).

[8] Brian Nolan, Max Roser, and Stefan Thewissen. "GDP Per Capita versus median household income: what gives rise to the divergence over time and how does this vary across OECD countries?" In: *The reviw of income and wealth* (2019), pp. 465–494 (cit. on p. 29).

[9] Thomas Piketty. *Capital in the Twenty-First Century*. Harvard University Press, 2017 (cit. on p. 2).

[10] Alessia Pini and Simone Vantini. "Interval-Wise Testing for Functional Data". In: *Journal of Nonparametric Statistics* 29 (Mar. 2017) (cit. on pp. 3, 4, 15).

[11] Alessia Pini and Simone Vantini. "The interval testing procedure: A general framework for inference in functional data analysis". In: *Biometrics* 72.3 (2016), pp. 835–845.

[12] Steven Pressman. "Why inequality is the most important economic challenge facing the next president". In: (2016).

[13] J.O. Ramsey, Giles Hooker, and Spencer Graves. *Functional Data Analysis with R and MATLAB*. Springer, 2009.

[14] J.O. Ramsey and B.W. Silverman. *Functional Data Analysis*. Springer, 2005 (cit. on pp. 4, 5, 7, 12, 13).

[15] Narasimha D. Rao, Petra Sauer, and Matthew Gidden. "Income inequality projections for the Shared Socioeconomic Pathways (SSPs)". In: *Futures* (2019), pp. 27–39 (cit. on pp. 2, 21, 29).

[16] Jesper Roine, Jonas Vlachos, and Daniel Waldenström. "The long-run determinants of inequality: What can we learn from top income data?" In: *Journal of Public Economics* (2009), pp. 974–988.

[17] United Nations University. "World Income Inequality Database (WIID), Guide and Data Sources". In: (2019) (cit. on pp. 21, 22).