ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN

ECONOMICS

Ciclo XXXV

Settore Concorsuale: 13/A5

Settore Scientifico Disciplinare: SECS-P/01

From Methodological Challenges to Empirical Insights: Reassessing Income Inequality and Instability

Presentata da: FEDERICO ATTILI

Coordinatore Dottorato
Andrea Mattozzi

Supervisore
Michele Costa

Esame finale anno 2024

From Methodological Challenges to Empirical Insights: Reassessing Income Inequality and Instability

Federico Attili

May 14, 2024

Contents

Nomenclature

Chapter 1-2

(μ, sd) Average and standard deviation of the correlation between replicates in the Monte Carlo experiment

$\bar{(\mu)}$ Average (average) correlation in simulations with similar parameters

$\bar{F}_k$ Average rank of the members of group k in the overall population

$\bar{sd}$ Average sd of the correlation in simulations with similar parameters

$\bar{X}^m$ Mean of the mixture between the pre- and post-shock distributions

Δx_q^{kh} Relative difference between the $q-th$ quantiles of group k and h

$\Delta\mu_{kh}$ Relative difference between the averages of group k and h

$\mathbb{1}$ Indicator function

$\mathbb{E}(\cdot)$ Expected value

$\mathbf{n}$ Vector of group sizes

$\mathbf{p}$ Vector of group popolation shares

$\mathbf{x}^k$ Income (or consumption) vector of group k

μ Average income (or consumption) of the population

μ_k Average income (or consumption) of the group k

μ_q Average income (or consumption) of the individuals ranking in the quantile q in their group

G_b^{BM}, or $GGini$ Between component of the Bhattacharya and Mahalanobis [1967a] decomposition

$G_k(\mathbf{x}^k)$, or G_k Gini index of group k

ρ Probability of random (symmetric) shock

$\rho(a,b)$ Correlation between the vectors a and b

ρ^k Probability of (asymmetric) shock for group k

ρ_{kq} Probability of (asymmetric) shock for quantile q in group k

ρ_q Probability of (asymmetric) shock for quantile q

A Gini index decomposition proposed in Chapter 1

BM Bhattacharya and Mahalanobis [1967a] decomposition

c Variation induced by the shock

C_b^{kh} Gross contribution of the interaction between groups k and h to between-group inequality

c_b^{kh} Numerator of the gross contribution of the interaction between groups k and h to between-group inequality

c_i

C_w^k Gross contribution of group k to within-group inequality

c_w^k Numerator of the gross contribution of group k to within-group inequality

$cov(a,b)$ Covariance between the vectors a and b

$CV[\mathbb{E}[\mu_k]]$ Expected Coefficient of Variation between the group averages

d_Q^{kh} Measure of distance between the mean-scaled income distributions of group k and h as defined in Ebert [1984]

Eb_{kh} Measure of distance between the income distributions of group k and h as defined in Ebert [1984]

$F(\cdot)$ Cumulative distribution function

$G(\mathbf{x};\mathbf{w})$ Gini index of the vector $\mathbf{x}$, weighted by the vector $\mathbf{w}$

$g(\mathbf{x};\mathbf{w})$ Numerator of the Gini index of the vector $\mathbf{x}$, weighted by the vector $\mathbf{w}$

G^k Group k contribution to the Gini index

g^k Numerator of the group k contribution to the Gini index

g_b Numerator of the between component of the Gini index decomposition proposed in Chapter 1

$G_b(\mathbf{x};\mathbf{w})$ Between component of the Gini index of the vector $\mathbf{x}$, weighted by the vector $\mathbf{w}$

$g_b(\mathbf{x};\mathbf{w})$ Numerator of the between component of the Gini index of the vector $\mathbf{x}$, weighted by the vector $\mathbf{w}$

G_b^k Group k standardised contribution to the between component of the Gini index

G_b^A Between component of the decomposition proposed in this thesis

$g_b^{k_h}$ Group k standardised contribution to between inequality arising from the interac- tion with group h

G_b^{kh} Standardised contribution of the interaction between groups k and h to between-group inequality

g_b^{kh} Numerator of the standardised contribution of the interaction between groups k and h to between-group inequality

G_b^{YL} Between component of the Yitzhaki and Lerman [1991a] decomposition

g_w Numerator of the within component of the Gini index decomposition proposed in Chapter 1

$G_w(\mathbf{x};\mathbf{w})$ Within component of the Gini index of the vector $\mathbf{x}$, weighted by the vector $\mathbf{w}$

$g_w(\mathbf{x};\mathbf{w})$ Numerator of the within component of the Gini index of the vector $\mathbf{x}$, weighted by the vector $\mathbf{w}$

G_w^A Within component of the decomposition proposed in this thesis

G_w^{BM} Within component of the Bhattacharya and Mahalanobis [1967a] decomposition

G_w^k Standardised contribution of group k to within-group inequality

g_w^k Numerator of the standardised contribution of group k to within-group inequality

G_w^{YL} Within component of the Yitzhaki and Lerman [1991a] decomposition

$GGini_q$ Gini index between individuals belonging to different groups and ranking in the $j-th$ position in their group income vector

HI Horizontal inequality benchmark

N Population size

n Size of groups when groups are equal sized

n_k Size of group k

p_k Share of population belonging to group k

Q Number of quantiles employed to assess the between group inequality

r Parameter defining the maximum of the uniform distribution for simulation purposes in Section 1.6

R^{BM} Residual term of the Bhattacharya and Mahalanobis [1967a] decomposition

R^{YL} Residual term of the Yitzhaki and Lerman [1991a] decomposition

S_b Between component share

t $= [0,1]$ time while the shock produces its effects

T_j This specifies the option type=j in the function *quantile()* in software R. The different options are in accordance with the definitions presented in Hyndman and Fan [1996]

w_{ij}^k Decomposition weight of the difference between individuals i and j of group k

w_{ij}^{kh} Scaling factor to decompose the contribution to the Gini index coming from the difference between the individual i of group k and the individual j of group h

w_j^{kh} Decomposition weight of the difference between the individuals ranking in the $j-th$ position of groups k and h

X Income distribution after the shock

x Variable of interest, income or consumption vector of the population

x_q^k Quantile q of the group k income vector

X^m Mixture between the pre- and post-shock distributions

x_i Income (or consumption) of individual i

x_i^k Income (or consumption) of the individual ranking in the $i-th$ position of the non decreasing income vector of group k

Y Income distribution before the shock

y_i^k Pre-shock income (or consumption) of the individual ranking in the $i-th$ position of the non decreasing income vector of group k

y_i Pre-shock income (or consumption) of individual i

YL Yitzhaki and Lerman [1991a] decomposition

$G(\mathbf{x})$, or G	Gini index of vector $\mathbf{x}$
$g(\mathbf{x})$, or g	Numerator of the Gini index
HBS	Household Budget Survey
ISTAT	Italian National Statistics Institute
lcm	Least common multiple
MAE	Mean Absolute Error
MC	Monte Carlo
MSE	Mean Squared Error
OECD	Organisation for Economic Co-operation and Development
SIPP	Survey of Income and Program Participation

Chapter 3

Δ	Arc-percentage change
$\delta_{i,\,t}$	Income shock of individual i at time t
$\Delta_{i;\,(t,\,t+1)}$	Arc-percentage change of the individual i income between t and $t+1$
$\overline{Y_i}$	Average income of individual i
$\overline{\ln Y_i}$	Average of the logarithm of income of individual i
σ_i^2	Variance of the logarithm of income of individual i
σ_{pop}^2	Population average of the individual variance of the logarithm of income
$\sigma_{t,\,t+1}(\Delta)$	Standard deviation (between N individuals) of the arc-percentage change between two periods t and $t+1$
G_i	Gini index of individual i income over time
G_pop	Population average of the individual Gini index over time
$G_{i,\,b}$	Between component of the individual i Gini index over time
$G_{i,\,w}$	Within component of the individual i Gini index over time

$G_{pop,\,b}$ Population average of the individual between component of the Gini index over time

$G_{pop,\,w}$ Population average of the individual within component of the Gini index over time

M Number of subperiods in a year (e.g., $M = 12$ if the analysis is monthly)

N Population size

SCV squared coefficient of variation

SCV_i squared coefficient of variation of individual i

$SCV_{i,\,infra,\,y}$ Infra-annual component, relative to year y, of the individual squared coefficient of variation

$SCV_{i,\,infra}$ Infra-annual component of the individual i squared coefficient of variation

$SCV_{i,\,inter}$ Inter-annual component of the individual i squared coefficient of variation

$SCV_{i,\,net-infra}$ Infra-annual component net of seasonality of the individual i squared coefficient of variation

$SCV_{i,\,seasonal}$ Seasonal component of the individual i squared coefficient of variation

$SCV_{pop,\,infra}$ Population average of the infra-annual component of the individual squared coefficient of variation

$SCV_{pop,\,inter}$ Population average of the inter-annual component of the individual squared coefficient of variation

$SCV_{pop,\,net-infra}$ Population average of the infra-annual component net of seasonality of the individual squared coefficient of variation

$SCV_{pop,\,seasonal}$ Population average of the seasonal component of the individual squared coefficient of variation

T Number of periods (generally months)

t Time index

Y Number of years

$Y_{i,\cdot,m}$ Average income of individual i in month m

$Y_{i,t}$ Income of individual i in period t

$Y_{i,y,\cdot}$ Average income of individual i in year y

$Y_{i,m,y}$ Income of individual i in month m of year y

MC Monte Carlo

SIPP U.S. Survey on Income and Programme Participation

Introduction

In the current epoch of globalisation, technological upheaval, and socio-economic transformations, the study of income inequality and instability is paramount. As the world changes and confronts monumental challenges, from financial crises to pandemics and inflation, understanding the nuances and intricacies of income distribution and inequality dynamics helps to effectively identify and mitigate the difficulties that people face while world and national economies evolve.

The title of this manuscript, "From Methodological Challenges to Empirical Insights: Reassessing Income Inequality and Instability", encapsulates a journey from grappling with methodological challenges to deriving profound empirical insights into the dynamics of inequality and instability. This thesis is the result of research work that I carried out during my Ph.D. in Economics at the University of Bologna and during my visiting at the Organisation for Economic Co-operation and Development (OECD). It provides a methodological discussion to guide and improve the assessment of income inequality and income instability. I briefly outline below the contents of the three chapters.

Chapter 1 consists of a methodological contribution to the assessment of horizontal inequality, which concerns itself with the disparities between groups defined, for example, by factors such as race, gender, or geography. Traditional methods to assess horizontal inequality partition the population of interest into groups based on a singular factor and decompose inequality measures, say the Gini index, into within- and between-group components. Notably, the between-component assesses horizontal inequality. While starting with decompositions offers a direct way to represent horizontal inequality as a proportion of overall inequality, emphasising the relevance of the examined factor, it also imposes limitations on the mathematical expressions available to capture the multifaceted nature of horizontal inequality. The between-components of the conventional decompositions assess horizontal inequality simply by comparing the group means. Contrasting them with an axiomatically derived benchmark of horizontal inequality, they underestimate horizontal inequality when there is no stochastic dominance among the group distributions. This chapter introduces a two-component decomposition of the Gini index that provides a more nuanced understanding of group differences and solves the risk of underestimation. A Monte Carlo experiment compares the axiomatically derived benchmark with the alternative between-components and shows that the proposed between-component has the highest correlation with the benchmark, exceeding 0.95 for all plausible parameter specifications. Analysis of real data confirms this evidence and highlights the risk of underestimating horizontal inequality by using conventional decompositions. In conclusion, this chapter reinforces the decomposition approach, providing a precise tool for assessing horizontal inequality.

In Chapter 2, we shift our gaze to the aftermath of asymmetric shocks on horizontal inequality. The world has witnessed profound economic perturbations over the past 15 years, from financial crises to the COVID-19 pandemic and the high inflation rate. Although there is a plethora of research on the overarching impact of these shocks on inequality, there is a conspicuous gap when it comes to discerning their

effects on the disparities between socio-economic groups. The biggest shocks usually impact the population with heterogeneous intensity between groups and between quantiles. This chapter shows that conventional (mean-based) methods for assessing inequality between groups do not adequately capture heterogeneous and asymmetric modifications of the group distributions. It discusses alternative indicators and suggests the use of the Gini decomposition introduced in Chapter 1. Empirical analysis focusses on COVID-19 and uses consumption microdata from Italian households to explore the main factors of inequality such as age, sex, and geography. The revelations about the age and gender gap, in particular, challenge conventional wisdom and underscore the value of our methodological approach. What emerges is that the increase in the age and gender gap after Covid is measured as significantly smaller than it would be using the conventional methodology. Before Covid, there is no stochastic dominance between distributions, then the asymmetry of the shock creates stochastic dominance. For this reason, the conventional decompositions underestimate the inequality between groups before Covid, whereas the underestimation disappears after Covid because of stochastic dominance. The increase in conventional indicators is mainly explained by this mechanism, while the increase in my between-component is representative of the dynamic of the distributions. The analysis also provides evidence, with rigorous simulations, warning about the weakness of mean-based indicators in the presence of asymmetric shocks.

Lastly, Chapter 3 pivots towards the subject of income instability. The predominant empirical narrative of income instability focusses on year-on-year fluctuations in annual income, overlooking the intricacies of short-term variations. This chapter underscores the importance and challenges of measuring high-frequency individual instability and proposes the average squared coefficient of variation as a potent tool for its assessment. It possesses important properties, being decomposable into infra- and inter-annual components, and allowing the possibility to disentangle the share of instability due to seasonality and upward mobility. An easy-to-interpret model of monthly earnings dynamic feeds a Monte Carlo experiment, which compares the assessment of instability using monthly data or annual averages. The underestimation due to annual data is particularly high when the parameters of the model represent periods of recessions or individual with unstable conditions in the labor market, which is when instability matters the most . Leveraging monthly income data from the U.S. Survey of Income and Program Participation (SIPP), our analysis confirms the significance of high-frequency data for the proper assessment of income instability.

Together, these chapters weave a tapestry of advanced methodologies and empirical analyses that aim to redefine our understanding of income inequality and instability in a rapidly changing world. Through rigorous theoretical expositions, innovative decomposition techniques, and empirical explorations, this thesis aims to make a significant contribution to the study of economic inequality.

Chapter 1

Uncovering Complexities in Horizontal Inequality: A novel decomposition of the Gini index

Keywords: Horizontal Inequality, Inequality decomposition, Gini index, Regional inequality

1.1 Preamble

This chapter provides a novel tool to study how inequality factors, such as geography, race, or gender, contribute to overall inequality, and explains when it should be preferred to existing methods. Partitioning a population into groups according to one factor and additively decomposing an inequality measure in within- and between-group components is the usual approach to evaluate the share of inequality due to the differences between groups. After discussing the theoretical impossibility of additively decomposing the Gini index into within- and between-components, in fact, we propose a Gini decomposition into two highly informative components. They measure within- and between-group inequality, with substantial improvement upon the usual assessment of inequality between groups. The between-group component of conventional decompositions only compares the means of the groups, failing to consider the complexity of the differences between groups. The proposed method overcomes this limitation: our between-group component accurately captures the complexity of group differences and is a convenient measure of horizontal inequality. Through rigorous simulations and empirical analysis on the Organisation for Economic Co-operation and Development (OECD) Income Distribution Database, we discuss the validity and usefulness of our method for evaluating and understanding inequality, finally providing a vademecum to choose the most appropriate

measure of horizontal inequality.

In this chapter, while we primarily discuss the application of our new decomposition method to income data, it is important to note that the framework is equally applicable to other cardinal and discrete variables such as educational attainment, health outcomes, or wealth. This versatility makes it a valuable tool for a broad spectrum of socioeconomic research.

This chapter represents a profound extension of my master's degree thesis. I am deeply thankful to Professors G. Pellegrini and R. Zelli for their invaluable guidance during the initial phase of this work. Special gratitude is owed to Prof. M. Costa, whose insights were pivotal in shaping the final manuscript. I am also thankful to Professors C. D'Ippoliti, G. Cavaliere, G. Pignataro, M. Kobus, G. De Marzo, F. Subioli and my friends for their constructive feedback. I extend my sincere thanks for the data provided by the Centre on Well-being, Inclusion, Sustainability, and Equal Opportunity (WISE) of the OECD, used under specific licence agreements. A preliminary version of this project was presented to the Department of Economics of Bologna. An improved version of this work was presented at the Tenth Meeting of the Society for the Study of Economic Inequality (Aix-en-Provence 2023); at the Fourth PhD Workshop (Manciano 2023), jointly organised by the University of Bologna, the Marco Fanno Association and the UniCredit Foundation; and at the Seventeenth Winter School on Inequality And Social Welfare Theory (Canazei 2024). They all offered further valuable refinement opportunities. A derivative paper of this chapter [Attili, 2024] has been recently published in Social Indicators Research, and I am grateful for the insights provided by the Editor Filomena Maggino and two anonymous referees. An earlier version of this work was also featured in the *Quaderni - Working Paper DSE* series.

The analytical methodologies developed for this chapter, involving the R codes to calculate the proposed decomposition, are detailed in the Appendix.

1.2 Introduction

Assessing inequality in a population can benefit from the analysis of how subgroups differ from each other. It is crucial for the study of populations characterised by strong gender or territorial divisions, as well as disparities associated with age, ethnicity, or religion. All these sources of horizontal inequality hamper well-being and development. Their negative impact is widely recognised, to the extent that their reduction is the focus of Sustainable Development Goals 5 and 10 of the United Nations Development Programme.

What are the main factors (geography, ethnicity, gender, etc.) associated with divides? Are these gaps actually narrowing? What are the most effective treatments to bridge the gap between population subgroups? The usual strategy for answering these policy questions is to decompose inequality by factors, obtaining two components that measure the contributions to inequality from within- and between-group disparities, with the latter containing much of the information on horizontal inequality. This approach has been extensively

used in the literature on horizontal inequality (see, e.g., Gachet et al., 2019; McDoom et al., 2019; Canelas and Gisselquist, 2019).

Given the wide range of possible inequality measures and decompositions, Josa and Aguado [2020] recently proposed an excellent review of the available methodologies and their implications, providing a practical framework for choosing the most appropriate measure. Among the various measures of inequality, the Gini coefficient has been particularly noted for its broad applicability and ease of interpretation. However, differently from other inequality measures such as the Generalized Entropy indices, that is decomposable into within and between inequality components, the Gini index is not additively decomposable in conventional terms. Despite the Gini coefficient lack of (conventional) additive decomposability into within- and between-group components, which poses challenges in understanding the sources of horizontal inequality, the Gini index remains widely used in such studies. The interest in proposing and interpreting new decomposition persists in recent efforts (Heikkuri and Schief, 2022), especially in the literature focusing on the distinction between fair inequality and inequality of opportunity (Moramarco, 2023;Sarkar, 2023).

In this thesis we propose a further decomposition of the Gini index, which contributes to the existing literature because its between-component is well suited to measure horizontal inequality. It stands out from several decompositions currently employed by researchers for the same purpose, solving a critical issue that we refer to as *oversimplification*.

The methods for decomposing inequality indicators (see Deutsch and Silber, 1999 for a review) share a common approach, pursuing the *additive decomposability* property (Bourguignon, 1979; Shorrocks, 1980). An inequality measure is additively decomposable if it can be expressed as the sum of two components observing the following constraints: the within-group component has to be the average inequality within subgroups weighted by population size, while the between-group component has to depend explicitly on the distance between the group means and on the group sizes.
As discussed by Ebert [2010], the constraint on the between-group component leads to oversimplification when analysing horizontal inequality between groups with similar means but different distributions, or studying inequality dynamics. In some cases, mean-based indicators may suggest that the groups are moving closer together, while, in reality, the distributions of the groups may be moving farther apart. While this phenomenon is occasional, it is common for the distance between means to differ from the distance between distributions, or for the two distances to have different dynamics.

The Gini index, a commonly used measure of inequality, is not additively decomposable. Its conventional decompositions require an additional term if the within- and between-components observe the constraints of additive decomposability. In this chapter we show that, by relaxing these constraints, it is possible both to obtain a two-component decomposition of the Gini index and to solve oversimplification. Indeed, our proposal possesses two components measuring the inequality within and between groups, with a novelty on the between-group component. It solves oversimplification by averaging the inequality between individuals ranked in the same quantiles of different subgroups.

We generally refer to income data in this chapter, but the introduced decomposition method is versatile and can be applied also to other variables like education, health, and wealth, making it valuable for diverse socioeconomic research.

The chapter unfolds as follows. Section 1.3 introduces the notation that we use throughout the paper; then it discusses the conventional subgroup decompositions of the Gini index and introduces a benchmark measure for horizontal inequality, before presenting the intuition leading to our decomposition. Section 1.4 formalises the new decomposition in the case of equal-sized groups and shows its properties. Section 1.5 extends the decomposition to the case of groups with different sizes. Section 1.6 presents the Monte Carlo experiment that studies the correlations between alternative measures, highlighting the informativeness of the proposed decomposition. Section 1.7 uses the OECD Income Distribution Database to analyse the income inequality of EU countries, providing striking evidence in favour of our between-component. Section 1.9 provides conclusive remarks and a vademecum for choosing the most appropriate measure of horizontal inequality.

1.3 The Gini index and horizontal inequality

The Gini coefficient, introduced by Corrado Gini in 1912, has been a central tool in the analysis of economic disparities. Renowned for its straightforward interpretation and ease of calculation, the Gini coefficient quantifies the extent of inequality in a distribution by measuring the average income disparities among all pairs of individuals. Consider a population of N units. We denote by x_i the $i-th$ element of the population income vector $\mathbf{x}$, by $\mu = \sum_1^N x_i/N$ the average income of the population, and by $G(\mathbf{x})$ its Gini index. When considering a partition of the population into K groups, denote the vector of their sizes with $\mathbf{n} = (n_1, \dots n_K)$, where $\sum_{k=1}^K n_k = N$. Let x_i^k be the i-th element of the group $k = 1, \dots, K$ (non decreasing) vector of incomes $\mathbf{x}^k = (x_1^k, \dots, x_{n_k}^k)$. Furthermore, we denote by μ_k the mean of group k, and by $G_k(\mathbf{x}^k)$ its Gini index.

Among the many different formulations of the Gini index (see Giorgi et al., 2005; Giorgi, 2011; Ceriani and Verme, 2015 and Ceriani and Verme, 2012), we consider the following:

$$G = \frac{1}{2\mu N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} |x_i - x_j| = \frac{g}{2\mu N^2} \tag{1.1}$$

where the numerator g is the sum of all the pairwise absolute differences between individual income. It is normalised by the factor $(2\mu N^2)^{-1}$, so that G is scale invariant, meaning it does not change with the scale of the data - namely, the currency used -. Consequently, $G \in [0,1]$ if all $x_i \geq 0$.

1.3.1 Subgroup decomposition of the Gini index

There is a wide variety of Gini index decompositions. They originate by alternative formulations of the index and by different approaches. We mainly focus on the most widespread and intuitive decomposition, whose between-component, often called GGini, is widely used to measure horizontal inequality. It was presented for the first time in Bhattacharya and Mahalanobis [1967a]. This decomposition consists of two components that measure inequality within and between groups, plus a third term, which by construction is the residual of the decomposition. We can express the general structure of the Bhattacharya and Mahalanobis decomposition as follows:

$$G = G_w^{BM} + G_b^{BM} + R^{BM} \tag{1.2}$$

The within component G_w^{BM} measures the inequality within groups by a weighted average of the Gini index of each group G_k. It reads:

$$G_w^{BM} = \sum_{k=1}^{K} \frac{\mu_k}{\mu} \frac{n_k^2}{N^2} G_k \tag{1.3}$$

and each weight is the product between the income and population shares of the group k.
As for the between component in Eq. (1.2), it reads:

$$GGini = G_b^{BM} = \frac{1}{2\mu N^2} \sum_{k=1}^{K} \sum_{h=1}^{K} n_k n_h |\mu_k - \mu_h| \tag{1.4}$$

The GGini has a nice feature: the weight of each mean difference is the product of the group sizes, hence the GGini measures the inequality between groups applying the Gini index to the case with no inequality within groups, i.e. when each observation has the average income of its group. It is the inequality between the weighted means. Unfortunately, due to oversimplification, this is not always fully representative of horizontal inequality. As a compelling example, $GGini = 0$ when groups have the same mean but their distributions differ in terms of variability, skewness, or higher moments, indicating the presence of horizontal inequality. Oversimplification also arises when the averages are different, being not the predominant source of the differences between the distributions of the groups. With the possibility of rare exceptions, the oversimplification of horizontal inequality manifests itself as underestimation of the inequality between groups.

Relation 1.4 represents the reference for further decompositions of the Gini index proposed over time; a contribution that partially deviates from the approach of Bhattacharya and Mahalanobis is attributed to Yitzhaki and Lerman [1991a]. Its general structure is the same as Eq. (1.2). We identify the components of this decomposition by replacing the superscripts of the three components with YL. Regarding G_w^{YL}, it only

differs from Eq. (1.3) in the structure of the weights, weighting each G_k by the income share of group k:

$$G_w^{YL} = \sum_{k=1}^{K} \frac{\mu_k}{\mu} \frac{n_k}{N} G_k \tag{1.5}$$

Given the structure of the weights, the correlation $\rho(G_w^{BM}, G_w^{YL}) = 1$ if the groups have the same size, but it decreases with the size variability.
Regarding G_b^{YL}, it reads:

$$G_b^{YL} = 2cov(\mu_k, \bar{F}_k)/\mu \tag{1.6}$$

It is the covariance between the group averages and the average rank of the groups members in the overall population ($\bar{F}_k$). The two between-components in Eq. (1.4)-(1.6) appear very different, but, like any between component of a subgroup decomposition of inequality, G_b^{YL} is also based on the comparison of the means of the groups and suffers from oversimplification.

Despite its widespread use, the Gini index is criticized because it does not observe the conventional definition of additive decomposability and includes the residual term — a limitation that the Generalized Entropy (GE) indices address effectively. GE indices, including Theil's T index and the mean log deviation, are characterized by their additive decomposability, allowing researchers to separate total inequality into within and between group components ($T = T_w + T_b$). This property makes GE measures particularly useful for understanding the sources of inequality, especially in complex, heterogeneous populations. However, as G_b^{BM} and G_b^{YL}, the between components of the GE indices decomposition also suffer of oversimplification due to their dependence on the means of the groups. The decomposition we present in this chapter is not additive decomposable in conventional terms but provides an additive formula for the Gini index ($G = G_w + G_b$). The decomposition metodology proposed herein solves the problem of oversimplification of between inequality and provides an additive formula to decompose the Gini index in within and between components. This justifies the continued focus on the Gini index within this thesis and the comparison with other Gini decomposition. The comparison of our decomposition with the GE decompositions would be less interesting since most of the difference between the components and their evolution would result due to the difference in scale and sensitivity to transfers of the alternative measures.

1.3.2 A benchmark measure of horizontal inequality

Having two vectors of Q quantiles representing two income distributions, we consider the following as a measure of their diversity:

$$Eb_{kh} = \frac{1}{Q} \sum_{q=1}^{Q} |x_q^k - x_q^h|$$

It is proposed by Ebert [1984] and is the simple average difference between quantiles. In his paper, Ebert

proposes a more general class of measures based on a parameter. Eb_{kh} corresponds to the case in which this parameter is equal to 1. A previous proposal by Dagum [1980] had already developed a measure of *economic distance* between two income distributions, but it has been criticised by Shorrocks [1982] due to its asymmetric nature. Ebert proposal, instead, presents all the properties of a distance and observes a general axiomatic approach. Furthermore, it perfectly reflects our idea that a measure of horizontal inequality between groups must compare their overall distributions. We generalise this measure to the case of K groups by using the same weighting structure of Eq. (1.4):

$$HI = \frac{1}{2\mu N^2} \sum_{k=1}^{K} \sum_{h=1}^{K} n_k n_h Eb_{kh}$$

so that HI is scale invariant and the weight of each Eb_{kh} depends on the number of pairs between the two groups.

We provide an additional reason why HI is a suitable benchmark for horizontal inequality. We define the GGini for quantile q as

$$GGini_q = \frac{1}{2\mu_q N^2} \sum_{k=1}^{K} \sum_{h=1}^{K} n_k n_h |x_j^k - x_j^h| \tag{1.7}$$

where $\mu_q = \sum_{k=1}^{K} x_q^k / K$ is the average, across groups, of observations ranking in the $q-th$ position of their group. Using Eq. (1.7), we can rewrite HI as follows:

$$HI = \frac{1}{Q} \sum_{q=1}^{Q} \frac{1}{2\mu N^2} \sum_{k=1}^{K} \sum_{h=1}^{K} n_k n_h |x_q^k - x_q^h| = \sum_{q=1}^{Q} \frac{\mu_q}{Q\mu} GGini_q \tag{1.8}$$

This benchmark evaluates the horizontal inequality of each quantile using the Gini index, then averages the results weighting each $GGini_q$ by the income share of quantile q. The advantage of HI over $GGini$ in measuring horizontal inequality is twofold. First, HI allows us to consider the differences between groups that are not captured by the mean. Second, decomposing HI in its addenda by k and/or by q produces informative indicators that allow one to know which groups and which parts of their distributions struggle the most. One can study the contribution of the bottom quartile to horizontal inequality and discover that the poor in one group suffer relatively more inequality than the poor in the other group, or that even if the groups are not equal on average, the poor are similar in the two groups. Horizontal inequality between two groups has different implications if it originates at the top, middle, or bottom of the distribution. For example, knowing the sources of horizontal inequality could be crucial when relating it to conflicts. Horizontal inequality triggers the start of conflicts, while within-group inequality shapes their intensity (Cederman et al., 2011; Esteban and Ray, 2011; Huber and Mayoral, 2019). The inequality-conflict literature clearly states that the presence of poor people experiencing bad living conditions and rich people, who can finance the conflict, is an essential engine for civil war. Explicitly considering horizontal differences at the top, middle, or bottom of

the distribution helps to study which aspects of group distributions and group differences shape the incentive to fight, allowing one to test refined hypotheses about the drivers of conflicts and their intensity.

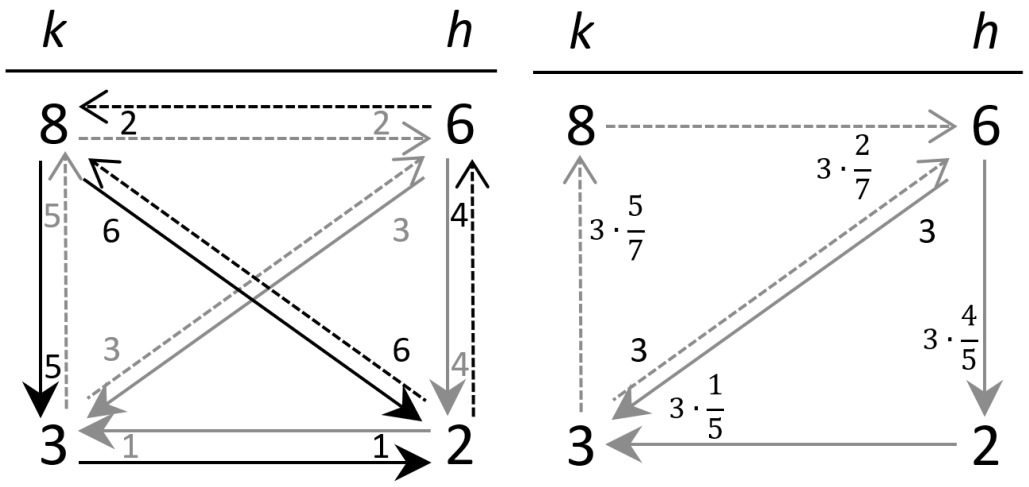
To conclude this section, we describe the intuition to derive, from the Gini index, a measure with the peculiarities of *HI* and such that its complement to the Gini measures the inequality within groups.

1.3.3 A new insight

Consider a population partitioned into K equal-sized groups and define n as their size. The numerator of the Gini index in Eq. (1.1) can be written as:

$$g = \sum_{i=1}^{N}\sum_{j=1}^{N}|x_i - x_j| = \sum_{k=1}^{K}\sum_{h=1}^{K}\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i^k - x_j^h| \tag{1.9}$$

Figure 1.1 provides an innovative insight into the structure of the Gini index. It illustrates a two-group-two-individual situation, with group $k = \{8,3\}$ and group $h = \{6,2\}$. Figure 1.1a highlights all pairwise differences between units, considered twice so that their sum constitutes g.
As the scheme suggests, we can distinguish three kinds of difference: vertical, horizontal, and diagonal. Vertical differences involve same-group pairs. Horizontal differences involve same-rank (same-quantiles) pairs from different groups. Diagonal differences involve different-rank pairs from different groups. We as-



(a) Pairwise differences composing g

(b) Decomposition in the non-trivial case

Figure 1.1: A two-group-two-individual illustration, with group $k = \{8,3\}$ and group $h = \{6,2\}$. The left panel highlights all the pairwise differences between units, considered twice so that their sum constitutes g. The vertical, horizontal, and black diagonal differences have intuitive decomposition. The right panel illustrates the decomposition of the grey diagonals

sign vertical and horizontal differences to within and between components, respectively. Although diagonal differences involve pairs of different groups, they also reflect vertical (same-group) differences and are not entirely attributable to inequality between groups. For example, imagine replacing the values in the scheme so that the groups are identical: pose $x_1^h = x_1^k = 8$ and $x_2^h = x_2^k = 3$. The values of the diagonal differences - 5 - are equal to the vertical ones and should not contribute to the absent horizontal inequality.

At this stage, diagonal differences can be instinctively thought of as the addenda of a residual term arising from the decomposition, and it seems natural to associate their sum with the conventional residuals R^{BM} and R^{YL}. These residuals - which are non-negative and disappear if the distributions of the groups do not overlap - are interpretable in terms of overlapping, stratification, and transvariation between the distributions of the groups (Yitzhaki and Lerman, 1991a; Lambert and Aronson, 1993; Yitzhaki, 1994; Dagum, 1997 and Costa, 2021). This interpretation does not apply to the sum of diagonal differences, which is positive even when the groups do not overlap. The sum of diagonals is zero only if there is perfect equality, since the diagonals contain information on both within- and between-group inequality. Going back to Figure 1.1, we propose a strategy to disentangle each diagonal difference in two informative contributions to within- and between-group inequality.

The two black diagonals in Figure 1.1a are intuitively decomposable. For example, looking at the solid black diagonal line and moving along the legs of the solid black triangle, the difference between the richest member of group k and the poorest member of group h is 6 since the former is 5 points richer than the poorest individual in her group, who is 1 unit richer than her counterpart in group h ($6 = 5 + 1$). A similar argument holds from the opposite point of view, which is looking at the dashed black diagonal line, representing the difference between the poorest member of group h and the richest member of group k ($6 = 4 + 2$). The two black diagonal differences are predominantly due to and reflect the inequality within the two groups. Consequently, we suggest splitting their contribution to g ($6 + 6 = 12$) assigning $5 + 4 = 9$ to the within component and $1 + 2 = 3$ to the between one.

This strategy is not feasible in the case of grey diagonals, which are the focus of Figure 1.1b. Here, the three values involved in the path along the grey legs do not increase or decrease monotonically as for the black lines, namely the product between the horizontal and the vertical signed differences is negative. In such cases, we should subtract the horizontal value from the vertical value to obtain the value of the diagonal difference. However, it would be paradoxical to decrease the between component by the horizontal value, i.e. by 1 in the case of the solid grey lines[1].

As Figure 1.1b illustrates, we suggest splitting each diagonal difference proportionally to the vertical and horizontal ones and assigning these two (positive) values to the within and between components, respectively. Using the proportional scaling to decompose the diagonals ensures that adding all their contributions

[1]To see the paradox, imagine replacing the poorest individual of group h with a poorer one. Subtracting $3 - (2 - \varepsilon) > 1$ would produce a lower value of the between component, although intuition suggests that the between inequality is now higher because the poor group is poorer.

to the within and between components preserves the proportion between the sum of vertical and horizontal differences. This is the key for the informativeness of the final components.

We have just presented the intuition that underlies the decomposition. In the next section, we formalise the decomposition under the hypothesis of equal-sized groups, which is relaxed in Section 1.5.

1.4 The decomposition proposal

Starting from Eq. (1.9), we propose decomposing each non-zero difference as follows[2]:

$$
\begin{aligned}
|x_i^k - x_j^h| &= |x_i^k - x_j^h| \frac{|x_i^k - x_j^k| + |x_j^k - x_j^h|}{|x_i^k - x_j^k| + |x_j^k - x_j^h|} = \\
&= |x_i^k - x_j^k| \frac{|x_i^k - x_j^h|}{|x_i^k - x_j^k| + |x_j^k - x_j^h|} + |x_j^k - x_j^h| \frac{|x_i^k - x_j^h|}{|x_i^k - x_j^k| + |x_j^k - x_j^h|} = \\
&= |x_i^k - x_j^k| w_{ij}^{kh} + |x_j^k - x_j^h| w_{ij}^{kh}
\end{aligned}
\tag{1.10}
$$

where the two addenda are, respectively, contributions to the within and between components, while $w_{ij}^{kh} = |x_i^k - x_j^h| / (|x_i^k - x_j^k| + |x_j^k - x_j^h|)$ is the scaling factor. It is $w_{ij}^{kh} = 1$ for vertical differences ($k = h$), horizontal differences ($i = j$) and differences such as black diagonals ($k \neq h$, $i \neq j$ and $(x_i^k - x_j^k) \cdot (x_j^k - x_j^h) \geq 0$). For these differences, Eq. (1.10) reduces to $|x_i^k - x_j^h| = |x_i^k - x_j^k| + |x_j^k - x_j^h|$. Differences such as grey diagonals ($k \neq h$, $i \neq j$ and $(x_i^k - x_j^k) \cdot (x_j^k - x_j^h) < 0$) are associated with $w_{ij}^{kh} \in [0, 1)$. In this case, $w_{ij}^{kh} < 1$ because $|x_i^k - x_j^h| < |x_i^k - x_j^k| + |x_j^k - x_j^h|$: the scaling factor reduces the vertical and horizontal differences so that the contributions to the within and between components add up to $|x_i^k - x_j^h|$.

The decomposition of the Gini index follows by substituting Eq. (1.10) into Eq. (1.9). Denoting $\sum_{h=1}^{K} w_{ij}^{kh} = w_{ij}^{k}$ and $\sum_{i=1}^{n} w_{ij}^{kh} = w_{j}^{kh}$, we obtain the numerators of the two components of the decomposition:

$$
g = g_w + g_b = \sum_{k=1}^{K} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}^{k} |x_i^k - x_j^k| + \sum_{k=1}^{K} \sum_{h=1}^{K} \sum_{j=1}^{n} w_{j}^{kh} |x_j^k - x_j^h| \tag{1.11}
$$

and we can write

[2] We underline that the equal-sized groups hypothesis guarantees that, given the couple (x_i^k, x_j^h), the element x_j^k always exists. In the following, we consider x_j^k without loss of generality. It is equivalent to consider x_i^h because the Gini index sums each difference twice by inverting the indices of the summations.

$$G = G_w^A + G_b^A = \frac{g_w}{2\mu N^2} + \frac{g_b}{2\mu N^2} \tag{1.12}$$

The Gini index consists of two terms. We interpret G_w^A and G_b^A as the within and between components of inequality because, respectively, they depend on the contributions from same-group and same-rank pairwise differences. Clearly, G_b^A does not explicitly depend on the group means, therefore it does not observe the (conventional) additive decomposability property. Pursuing a between-component that measures horizontal inequality while comparing the entire distributions of the groups inevitably leads to contrast with the definition of additive decomposability. However, while additive decomposability is desirable when the goal is to understand how resources are unequally distributed between groups, we believe that relaxing the constraint it imposes on the between component is essential to accurately capture horizontal inequality.

The within and between components involve, respectively, w_{ij}^k and w_j^{kh}. These weights ensure that each same-group (same-rank) difference contributes to within (between) inequality according to how much it affects the diagonal ones. For example, if a vertical difference increases, thus enlarging some of the grey-like diagonal differences, then the related scaling factors consistently increase and inflate the weight w_{ij}^k. The structure of the weights follows from that of the scaling factors, which is not necessarily unique. In Eq. (1.10) we multiply $|x_i^k - x_j^h|$ by 1, expressed as the ratio of $|x_i^k - x_j^k| + |x_j^k - x_j^h|$ to itself. Instead, we could multiply by the ratio of $f(x_i^k - x_j^k) + f(x_j^k - x_j^h)$ to itself, being $f(\cdot)$ monotonic and continuous function.

$$\begin{aligned}
|x_i^k - x_j^h| &= |x_i^k - x_j^h| \frac{f(|x_i^k - x_j^k|) + f(|x_j^k - x_j^h|)}{f(|x_i^k - x_j^k|) + f(|x_j^k - x_j^h|)} = \\
&= f(|x_i^k - x_j^k|) \frac{|x_i^k - x_j^h|}{f(|x_i^k - x_j^k|) + f(|x_j^k - x_j^h|)} + f(|x_j^k - x_j^h|) \frac{|x_i^k - x_j^h|}{f(|x_i^k - x_j^k|) + f(|x_j^k - x_j^h|)} = \\
&= f(|x_i^k - x_j^k|) \omega_{ij}^{kh} + f(|x_j^k - x_j^h|) \omega_{ij}^{kh}
\end{aligned} \tag{1.13}$$

This produces a more general class of decompositions; depending on the choice of $f(\cdot)$, decomposition puts more emphasis on large or small differences (e.g., $f(x) = x^2$ or $f(x) = \sqrt{|x|}$). The ω_{ij}^{kh} are the scaling factors of this class; they equal w_{ij}^{kh} when $f(\cdot)$ is the identity function. Since we are decomposing a linear measure of inequality [Mehran, 1976], we think that the natural choice is to take $f(\cdot)$ as the identity function, which is the only one that preserves the linearity of the Gini index in the components of its decomposition.

1.4.1 Properties of the decomposition

Our decomposition enjoys relevant properties, both in the within- and the between-group components. Given $w_{ij}^{kk} = 1$, $w_{jj}^{kh} = 1$ and $w_{ij}^{kh} \geq 0$, we have $w_{ij}^{k} \geq 1$ and $w_{j}^{kh} \geq 1$. Therefore, the following properties hold:

$$G_w^A = 0 \iff |x_i^k - x_j^k| = 0 \quad \forall\ i,j,k \tag{i}$$

$$G_b^A = 0 \iff |x_j^k - x_j^h| = 0 \quad \forall\ j,k,h \tag{ii}$$

The first relation ensures that the within component is zero iff all the same-group differences are zero, i.e. there are no differences within groups. The second condition guarantees that the between-component is zero iff all the same-rank differences are zero, i.e. the groups have the same distribution.
Properties (i)-(ii) are conceptually analogous. All the most widespread decompositions of inequality have a within component that, like ours, observes property (i). As for property (ii), G_b^{BM} and G_b^{YL} satisfy its sufficiency - they are zero if the groups have the same distribution - while they do not satisfy its necessity - they are zero even if the groups have different distributions. Our between-component observes both the sufficiency and necessity of property (ii), being zero iff the groups have the same distribution.

Additional important properties concern the high correlation of G_w^A with Eq. (1.3)-(1.5), and of G_b^A with Eq. (1.8), which follows from algebraic similarity of those equations. Regarding G_w^A, we can rewrite it as

$$\begin{aligned} G_w^A &= \frac{1}{2\mu N^2} \sum_{k=1}^{K} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}^k |x_i^k - x_j^k| = \\ &= \sum_{k=1}^{K} \frac{\mu_k}{\mu} \frac{n^2}{N^2} \cdot \frac{1}{2\mu_k n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}^k |x_i^k - x_j^k| \end{aligned} \tag{1.14}$$

The term $\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}^k |x_i^k - x_j^k| / 2\mu_k n^2$ would equal G_k if all the weights $w_{ij}^k = 1$, resulting in $G_w^A = G_w^{BM}$. This never happens but, as we discussed, each w_{ij}^k preserves the information of the vertical differences that it multiplies. Therefore, the correlation between G_w^A and G_w^{BM} is naturally high. Being $n/N = 1/K$, if all the weights are $w_{ij}^k = K$ then $G_w^A = G_w^{YL}$. Actually, $1 \leq w_{ij}^k \leq K$, therefore the weighting structure of G_w^A is the middle ground between those of G_w^{BM} and G_w^{YL}. This is why G_w^A has the good property of being highly correlated with both G_w^{BM} and G_w^{YL}.
A similar discussion holds by comparing G_b^A with the horizontal inequality benchmark defined in Eq. (1.8). Our between component reads:

$$\begin{aligned} G_b^A &= \frac{1}{2\mu N^2} \sum_{k=1}^{K} \sum_{h=1}^{K} \sum_{j=1}^{n} |x_j^k - x_j^h| w_j^{kh} = \\ &= \sum_{j=1}^{n} \frac{\mu_j}{n\mu} \frac{1}{2\mu_j N^2} \sum_{k=1}^{K} \sum_{h=1}^{K} n w_j^{kh} |x_j^k - x_j^h| \end{aligned} \tag{1.15}$$

which has a comparable structure to Eq. (1.8). G_b^A and HI would be equivalent (in the case $n = Q$) if all the weights were $w_j^{kh} = n$. Again, it never happens and since $1 \leq w_j^{kh} \leq n$ then G_b^A is usually lower than HI. It is important to avoid confusing this relation with an underestimation of horizontal inequality. We argue that the weights w_j^{kh} guarantee such a strong correlation between G_b^A and HI that we can consider G_b^A lower than HI simply due to a scaling transformation. We confirm the high correlation between G_b^A and HI, and between G_w^A, G_w^{BM} and G_w^{YL}, using a Monte Carlo simulation. We present the experiment and its results in Section 1.6.

Concluding this section, we observe that, as discussed for HI, isolating the addenda of G_b^A by j and k provides indicators which allow one to understand which parts of the distributions differ the most between the groups, and which group differs the most from the others. We believe that these indicators are another tool from which several fields of the inequality literature can benefit.

1.5 The different-sized groups extension

In this section we show that the equal-sized groups hypothesis is not binding. It was necessary to understand the decomposition arguments, but the proposal can be easily extended to cope with more general situations in which the K groups have different sizes. In this case, Eq. (1.9) becomes:

$$g(\mathbf{x}) = \sum_{i=1}^{N}\sum_{j=1}^{N}|x_i - x_j| = \sum_{k=1}^{K}\sum_{h=1}^{K}\sum_{i=1}^{n_k}\sum_{j=1}^{n_h}|x_i^k - x_j^h|$$

and a way to ensure the existence of the element x_j^k is necessary for the implementation of our decomposition proposal. We propose two distinct solutions. The first evaluates the two components without approximation, but necessitates potentially unaffordable computations. The second drastically reduces computational requirements by paying the cost of a negligible approximation.

1.5.1 The *exact* approach

Consider a new common size $n = lcm(\mathbf{n})$ and the vector of the groups population shares $\mathbf{p} = (p_1, \ldots, p_K)$, where $p_k = n_k/n$ is the population share of group k, so to build the *repopulated* vectors $z^k = (z_1^k, \ldots, z_n^k) = (\underbrace{x_1^k \ldots x_1^k}_{p_k^{-1}}, \ldots \underbrace{x_{n_k}^k \ldots x_{n_k}^k}_{p_k^{-1}})$. It is easy to show that:

$$g(\mathbf{x}) = \sum_{k=1}^{K}\sum_{h=1}^{K}\sum_{i=1}^{n_k}\sum_{j=1}^{n_h}|x_i^k - x_j^h| = \sum_{k=1}^{K}\sum_{h=1}^{K}\sum_{i=1}^{n}\sum_{j=1}^{n} p_k p_h |z_i^k - z_j^h| = g(\mathbf{z};\mathbf{p}) \tag{1.16}$$

where $g(\mathbf{z};\mathbf{p})$ is the numerator of the weighted Gini index. Figure 1.2 provides the intuition of Eq. (1.16). Imagine two groups composed, respectively, of two and three individuals, as reported in the left rectangle.

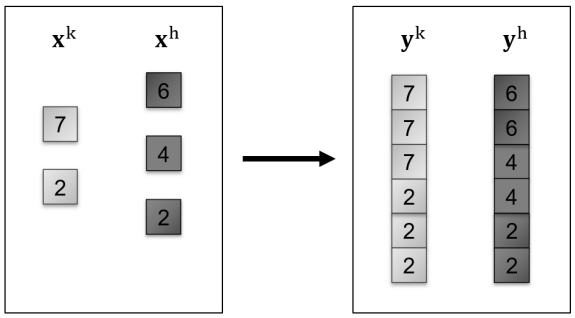


Figure 1.2: A two-group illustration of repopulation in the exact approach. The left panel represents two groups composed, respectively, of two and three individuals. The exact approach replace them with the equal-sized groups in the right rectangle and introduces the weights p_k and p_h in Eq. (1.16) to preserve the correspondence with the Gini index

Replace them with those in the right rectangle. According to the principle of population, $\mathbf{x}^k$ and $\mathbf{z}^k$ (as well as $\mathbf{x}^h$ and $\mathbf{z}^h$) have the same within-group inequality. In addition, the cumulative distribution functions of the two groups are the same before and after the replacement, and hence the distance between the two groups is also unvaried. However, each difference between couples in the left scheme appears, in the right scheme, 9 times if the couple belongs to $\mathbf{z}^k$, 4 times if it belongs to $\mathbf{z}^h$ and 6 times if the two units belong to different groups. The p_k and p_h in Eq. (1.16) adjust by multiplying the differences, respectively, by $1/9$, $1/4$ and $1/6$. In this way, equal-sized groups are obtained preserving the correspondence with the Gini index and with the original distributions of the groups.

We decompose the Gini index using a technique that is analogous to the one used to derive Eq. (1.11):

$$G(\mathbf{x}) = G(\mathbf{z};\mathbf{p}) = G_w(\mathbf{z};\mathbf{p}) + G_b(\mathbf{z};\mathbf{p})$$

The only difference is in the structure of the new weights $w_{ij}^k = \sum_{h=1}^K p_k p_h w_{ij}^{kh}$ and $w_j^{kh} = \sum_{i=1}^n p_k p_h w_{ij}^{kh}$. This adjustment integrates essential information to maintain the original significance of each pair, akin to employing a sample population with corresponding sample weights.

Unfortunately, in most cases, this approach requires an unaffordable computational effort because of the potentially huge magnitude of the least common multiple. To reduce computational requirements, we present an alternative procedure, which we refer to as *quantilisation*.

1.5.2 *Quantilisation*

Differently from the exact approach, we propose to consider a lower value of n and to calculate differently each $\mathbf{z}^k$: for each group, the vector $\mathbf{z}^k$ contains the n quantiles from the income vector of the group. As for p_k, their calculation is the same employed in the exact approach, but now nothing constrains $n \geq n_k$, thus it can be $p_k > 1$. The decomposition is the same, but in this case the value of the Gini index and the components

of the decomposition incur in some minimal approximation:

$$G(\mathbf{x}) \approx G(\mathbf{z};\mathbf{p}) = G_w(\mathbf{z};\mathbf{p}) + G_b(\mathbf{z};\mathbf{p}) \tag{1.17}$$

To employ this method and minimise the approximation, there are the definition of quantile and the value of n to be selected. For the former, we advise the Definition 7 reported in Hyndman and Fan [1996], which is the default definition adopted by the *quantile* function in various statistical software. Given each vector $\mathbf{x}^k$, accordingly to this definition and in order to minimise the approximation, we suggest first to interpolate linearly the n_k vertices $\left((i-1)/(n_k-1), x_i^k\right)$, and then to estimate the n quantiles by the values associated with the probabilities

$$prob_j = \frac{j-1}{n-1} \quad j = 1,\ldots,n \tag{1.18}$$

on the resulting piecewise linear curve.
Regarding the value of n, we define $p_k = n_k / \sum_{k=1}^{K} n_k$ and advise the value:

$$n = \sum_{k=1}^{K} p_k n_k = \frac{\sum_{k=1}^{K} n_k^2}{\sum_{k=1}^{K} n_k} \tag{1.19}$$

which determines n as the average of the n_k, each weighted by its own share of population p_k.
The decisions proposed for both the quantile definition and for the value of n are motivated in the appendix. Here, we only inform that, if they are employed, the approximation that the quantilisation procedure copes with is minimal and negligible. To obtain two estimates of the exact components, which are consistent and sum up to the Gini index of the original data, it is sufficient to multiply the shares of the components, obtained by quantilisation, by the value of the index:

$$G(\mathbf{x}) = G_w(\mathbf{x}) + G_b(\mathbf{x}) \approx G_w(\mathbf{z};\mathbf{p}) \cdot \frac{G(\mathbf{x})}{G(\mathbf{z};\mathbf{p})} + G_b(\mathbf{z};\mathbf{p}) \cdot \frac{G(\mathbf{x})}{G(\mathbf{z};\mathbf{p})}$$

1.6 Monte Carlo experiment: comparison of alternative decompositions

In this section we study, using a Monte Carlo simulation, the correlation of G_w^A and G_w^{BM} with G_w^{YL}; and the correlation of G_b^A, *GGini* and G_b^{YL} with *HI*. We also carried out the experiment using G_w^{BM} instead of G_w^{YL} as the reference point for the inequality within the groups. This additional simulation confirms the discussion after Eq. (1.14), which stresses that the weighting structure of G_w^A is the balancing between those of G_w^{BM} and G_w^{YL}.

The experiment works with three predetermined parameters: the number of groups, the parameter(s) of the distribution of **n** and the coefficient of variation between the averages of the groups ($CV[\mathbb{E}[\mu_k]]$). The latter is an indirect parameter, which derives from imposing credible conditions on the parameters of the lognormal distribution that is used to sample incomes. More details about the income simulation procedure and its theoretical foundations can be found in the Appendix. Here, we only stress that the parameters of the lognormal distribution are micro-founded. Indeed, as detailed in the second part of the Appendix, they are chosen sampling from the parameters estimated in Bandourian et al. [2002] using real data along different countries and periods. This guarantees robust results with respect to real income distributions.

The procedure is schematised in Figure 1.3 and can be summarised as follows:

Step I. Fixing K and (n,r), generate the vector **n**: each n_k is drawn from a uniform $[n,(1+r)\cdot n]$, where $100\cdot r$ is the maximum percentage deviation from the minimum n.

Step II. Fixing $CV[\mathbb{E}[\mu_k]]$, generate the income vectors from the lognormal distribution 50 times, each time evaluating all the statistics involved (G_w^A, G_w^{BM}, G_w^{YL}, G_b^A, G_b^{BM}(=*GGini*), G_b^{YL} and HI). The 50 points allow us to estimate the following triples of correlation estimates:

$$\begin{bmatrix} \rho\left(G_w^A, G_w^{YL}\right) \\ \rho\left(G_w^{BM}, G_w^{YL}\right) \\ \rho\left(G_w^{YL}, G_w^{YL}\right) \end{bmatrix}, \quad \begin{bmatrix} \rho\left(G_b^A, HI\right) \\ \rho\left(G_b^{BM}, HI\right) \\ \rho\left(G_b^{YL}, HI\right) \end{bmatrix}$$

Step III. To assess the correlation in scenarios where the variability of the group means increases, repeat *Step II* for eight distinct values of $CV[\mathbb{E}[\mu_k]]$. .

The simulation runs *Step I-III* 20 times and delivers, for each value of $CV[\mathbb{E}[\mu_k]]$, 20 replicates of the triples defined in *Step II*. For each position in the triple, we summarise its 20 replicates by their average and standard deviation (μ, sd). For each position in the triple, the eight pairs (μ, sd) corresponding to the eight values

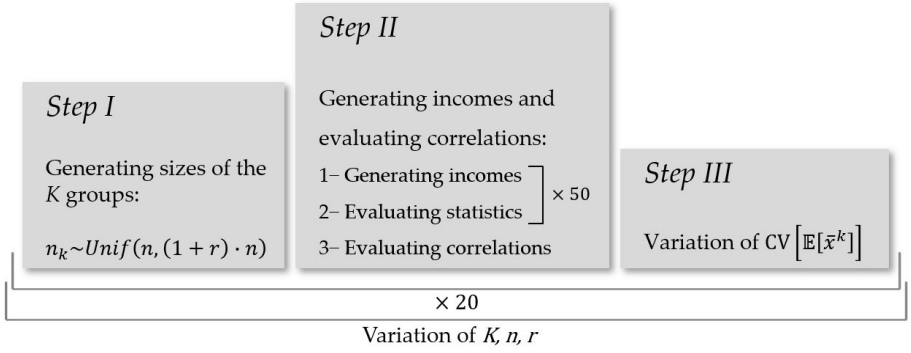


Figure 1.3: The Monte Carlo experiment in a scheme

of $CV[\mathbb{E}[\mu_k]]$ are averaged pairwise, determining four pairs $\left(\overline{\mu}, \overline{sd}\right)$. They correspond to low, medium-low, medium-high, and high (L, M-L, M-H, H) $CV[\mathbb{E}[\mu_k]]$. The experiment evaluates multiple scenarios by also varying the number of groups K and the parameters (n, r).

Results. We present in Table 1.1 the four pairs $(\overline{\mu}, \overline{sd})$ for representative parameters ($K = 3, 30$; $n = 10, 100$; $r = 1, 4$).

As for the within component, both $\rho(G_w^A, G_w^{YL})$ and $\rho(G_w^{BM}, G_w^{YL})$ are rarely below 0.9, with the first being generally higher and less volatile. The only exception occurs when the variability in $\mathbf{n}$ is low (i.e. when r is low): in this case $\rho(G_w^{BM}, G_w^{YL})$ is sometime higher. This is because $\rho(G_w^{BM}, G_w^{YL}) = 1$ when the groups are equal-sized, decreasing with the variability of $\mathbf{n}$.

The correlations depend marginally on the specification of the parameters. Higher values of K negatively influence $\rho(G_w^A, G_w^{YL})$, but increasing the values in $\mathbf{n}$ absorbs this small effect; $\rho(G_w^A, G_w^{YL})$ also decreases for higher values of $CV[\mathbb{E}[\mu_k]]$, while its variability increases. In any case, all the $\overline{\mu}$ referred to our within component are never below 0.92 and the highest $\overline{sd}$ is $2.8 \cdot 10^{-2}$.

When performing the Monte Carlo experiment using G_w^{BM} as the reference instead of G_w^{YL}, we obtain results that substantially mirror those presented. As expected, thanks to its weighting structure, our within component is the middle ground between G_w^{YL} and G_w^{BM}.

Regarding the comparison of G_b^A, G_b^{BM} and G_b^{YL} with HI, regardless of K and $\mathbf{n}$, $\rho(G_b^A, HI)$ is always the highest (≈ 1) and least volatile, showing striking advantages in situations where the variability of the means is not large. It slightly decreases and shows higher $\overline{sd}$ when the variability in $\mathbf{n}$ increases and the values in $\mathbf{n}$ and K are small. When the variability of the means increases, explaining most of the differences between groups and reducing their overlap, $\rho(G_b^{BM}, HI)$ increases and narrows the gap with $\rho(G_b^A, HI)$; also $\rho(G_b^{YL}, HI)$ increases with the variability of the means, but remains much smaller and the most volatile. In conclusion, according to the Monte Carlo experiment, G_b^A is the most suitable between-component to capture the complexity of horizontal inequality when the groups have similar means. It provides richer information than the conventional between-components unless the group averages are so far apart as to drastically reduce overlapping between the distributions and explain most of the difference between the groups.

1.7 Horizontal inequality between the EU country pairs

This section points out the importance of choosing the appropriate measure to assess horizontal inequality and its evolution. Our analysis makes use of the OECD Income Distribution Database, which provides the average household income in each decile of all EU countries, adequately transformed in PPP, from 2004 to 2018.

			Within components								Between components							
			$K=3$				$K=30$				$K=3$				$K=30$			
Magnitude of $CV[\mathbb{E}[\mu_k]]$			L	M-L	M-H	H	L	M-L	M-H	H	L	M-L	M-H	H	L	M-L	M-H	H
$\mathbf{n} \sim U\left([10,20]^K\right)$	$\overline{\mu}$	A	0.972	0.966	0.964	0.961	0.943	0.942	0.944	0.941	0.988	0.986	0.984	0.978	0.990	0.989	0.983	0.977
		BM	0.981	0.977	0.968	0.957	0.963	0.960	0.949	0.932	0.771	0.840	0.903	0.952	0.829	0.881	0.908	0.944
		YL	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.386	0.465	0.704	0.807	0.117	0.313	0.468	0.668
	$\overline{sd}$	A	0.005	0.013	0.010	0.012	0.018	0.016	0.015	0.018	0.006	0.007	0.008	0.006	0.003	0.003	0.005	0.007
		BM	0.013	0.017	0.024	0.031	0.013	0.012	0.018	0.029	0.073	0.050	0.053	0.019	0.060	0.050	0.037	0.015
		YL	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.199	0.214	0.131	0.085	0.197	0.215	0.202	0.100
$\mathbf{n} \sim U\left([10,50]^K\right)$	$\overline{\mu}$	A	0.974	0.970	0.959	0.943	0.946	0.945	0.936	0.921	0.964	0.966	0.965	0.972	0.984	0.982	0.976	0.972
		BM	0.939	0.923	0.899	0.841	0.937	0.925	0.894	0.857	0.683	0.766	0.881	0.943	0.737	0.796	0.891	0.937
		YL	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.238	0.445	0.692	0.851	0.033	0.234	0.475	0.734
	$\overline{sd}$	A	0.009	0.010	0.012	0.020	0.012	0.018	0.020	0.028	0.028	0.031	0.030	0.025	0.006	0.006	0.008	0.009
		BM	0.034	0.043	0.052	0.080	0.024	0.033	0.049	0.047	0.124	0.093	0.057	0.041	0.080	0.069	0.031	0.019
		YL	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.249	0.180	0.137	0.066	0.162	0.241	0.232	0.138
$\mathbf{n} \sim U\left([100,200]^K\right)$	$\overline{\mu}$	A	0.968	0.962	0.949	0.938	0.947	0.953	0.940	0.921	0.996	0.995	0.994	0.992	0.995	0.994	0.993	0.992
		BM	0.989	0.985	0.974	0.953	0.979	0.971	0.952	0.932	0.425	0.748	0.926	0.974	0.519	0.718	0.906	0.963
		YL	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	-0.019	0.525	0.797	0.912	-0.120	0.369	0.694	0.858
	$\overline{sd}$	A	0.011	0.012	0.014	0.019	0.017	0.013	0.016	0.020	0.003	0.003	0.003	0.003	0.002	0.003	0.003	0.003
		BM	0.011	0.011	0.020	0.044	0.009	0.012	0.017	0.017	0.155	0.087	0.033	0.009	0.123	0.082	0.027	0.012
		YL	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.265	0.173	0.072	0.029	0.202	0.175	0.114	0.049
$\mathbf{n} \sim U\left([100,500]^K\right)$	$\overline{\mu}$	A	0.970	0.966	0.952	0.930	0.950	0.960	0.942	0.921	0.996	0.995	0.994	0.993	0.996	0.993	0.993	0.992
		BM	0.980	0.968	0.937	0.892	0.958	0.950	0.903	0.853	0.336	0.700	0.924	0.974	0.401	0.661	0.896	0.966
		YL	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	-0.040	0.437	0.812	0.908	-0.086	0.357	0.716	0.879
	$\overline{sd}$	A	0.009	0.014	0.011	0.017	0.015	0.010	0.017	0.025	0.003	0.003	0.003	0.002	0.002	0.002	0.002	0.003
		BM	0.013	0.022	0.039	0.073	0.017	0.015	0.041	0.041	0.184	0.109	0.032	0.012	0.125	0.092	0.034	0.011
		YL	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.233	0.207	0.071	0.034	0.208	0.162	0.075	0.038

Table 1.1: Correlations between the benchmarks and the components from three Gini index decompositions: A, BM, and YL in the third column identify the decomposition proposed in this work, in Bhattacharya and Mahalanobis [1967a] and in Yitzhaki and Lerman [1991a], respectively. The correlations are evaluated by the algorithm presented in Section 1.6 for different values of K, $\mathbf{n}$ and $CV[\mathbb{E}[\mu_k]]$. In this table, the eight vectors of correlation replicates for each component are summarised by their average and standard deviation, which are averaged pairwise in four pairs ($\overline{\mu}$, $\overline{sd}$), corresponding to low, medium-low, medium-high and high (L, M-L, M-H, H) $CV[\mathbb{E}[\mu_k]]$

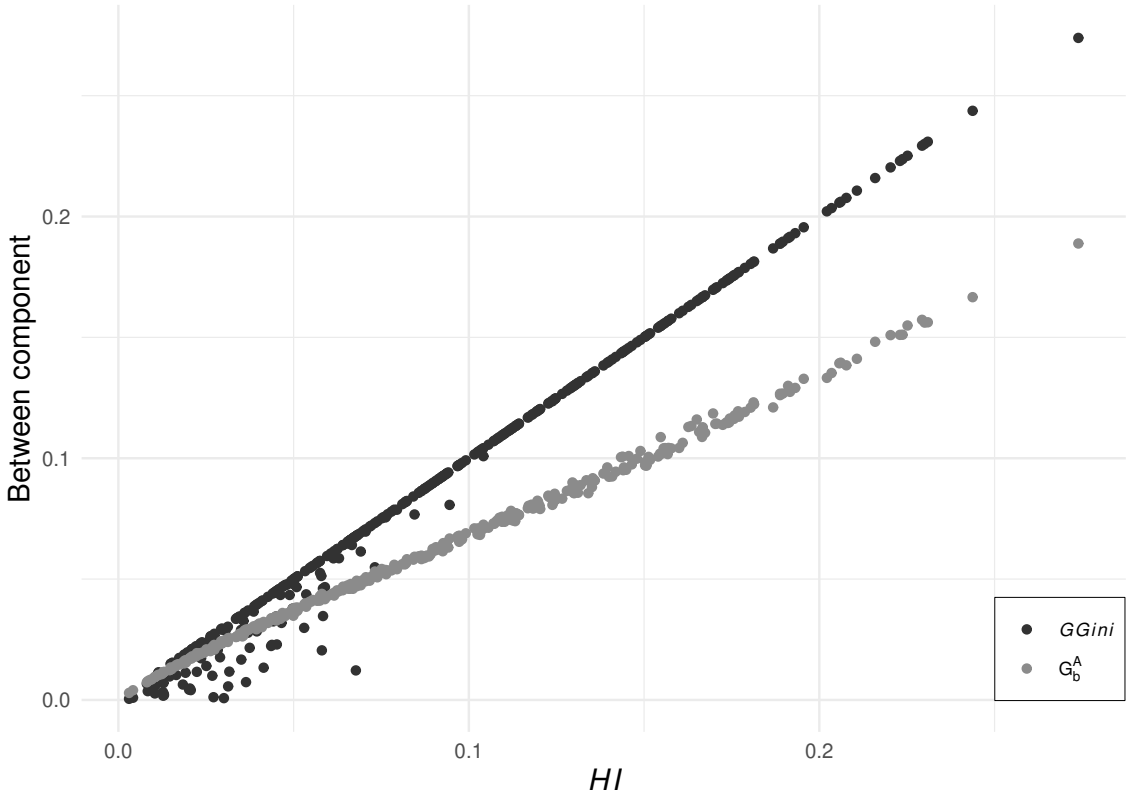


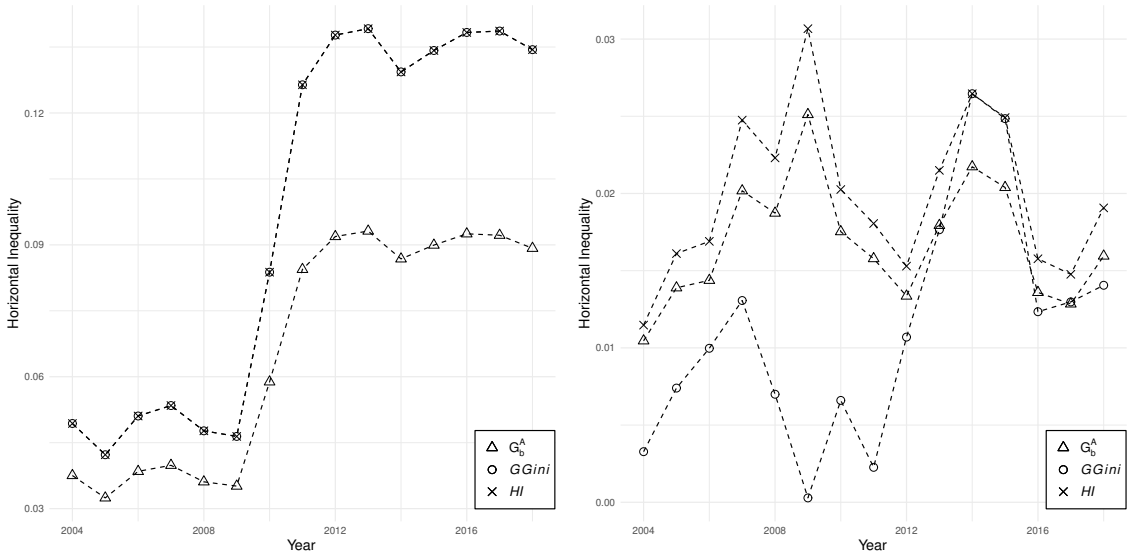
Figure 1.4: Horizontal inequality of the European Union country pairs in 2018, as measured by HI, GGini and G_b^A. Comparison of HI (x-axes) with the GGini and G_b^A (y-axes). Oversimplification of the GGini takes the form of underestimation of horizontal inequality and is not rare when groups have similar mean. The scatter relating HI and G_b^A has no outliers and confirms $\rho(G_b^A, HI) \approx 1$

First, we measure the income inequality of all EU country pairs in 2018. For each country pair, we decompose the inequality between their income deciles by Eq. (1.2) and Eq. (1.31). We also evaluate HI between the two countries. Finally, we compare the inequality between countries as measured by G_b^A, $GGini$ and HI to study the differences between alternative measures of horizontal inequality. In Figure 1.4, each point represents the horizontal inequality of a country pair in the space (HI, G_b). Black and grey points relate HI, respectively, to $GGini$ and G_b^A. Comparing $GGini$ and HI highlights that they are perfectly correlated over several country pairs, especially the most dissimilar couples. However, there are couples of countries that significantly reduce $\rho\left(G_b^{BM}, HI\right)$. Indeed, when the inequality between two countries is medium to low, GGini often underestimates the inequality between countries, sometimes being significantly lower than the value we would expect based on HI. The oversimplification issue does not involve our between-group

component. Comparing *HI* with G_b^A highlights their strong correlation over all the country pairs, even those with similar means and different distributions.

Focusing on the evolution of horizontal inequality over time, we consider two country pairs representing two opposite situations: the pairs over which the GGini and *HI* are correlated, and the pairs suffering of oversimplification.

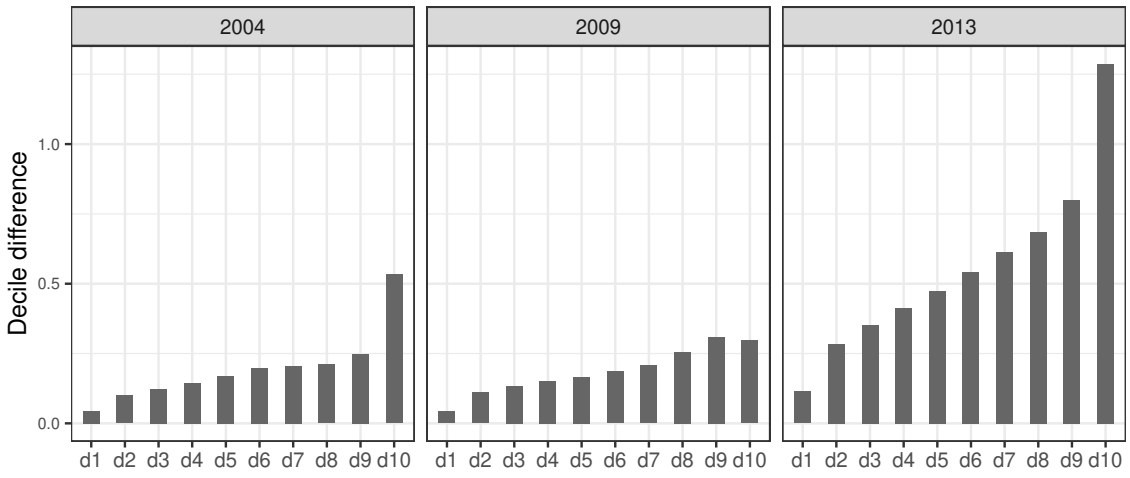
The first scenario, which involves Italy and Greece, is represented in Fig. 1.5a. All three measures report the same evolution, as long as the correlation between *HI* and *GGini* is perfect and that between *HI* and G_b^A is above 0.99. The inequality between the two countries decreases slightly from 2004 to 2009, while it more than doubles between 2009 and 2013. In order to investigate the reasons explaining the evolution of horizontal inequality, in Fig. 1.6a we focus on three peculiar years (2004, 2009, 2013) and plot the difference (as a fraction of the mean) between the income deciles of the two countries. The reduction of horizontal inequality between Italy and Greece from 2004 to 2009 is primarily due to the decrease in the gap in the richest decile. Between 2009 and 2013, all (positive) differences between deciles are at least double. Consequently, also the difference between the Italian and Greek means more than doubles, which explains the perfect correlation between *GGini* and *HI* in Fig. 1.5a. Comparing the United Kingdom



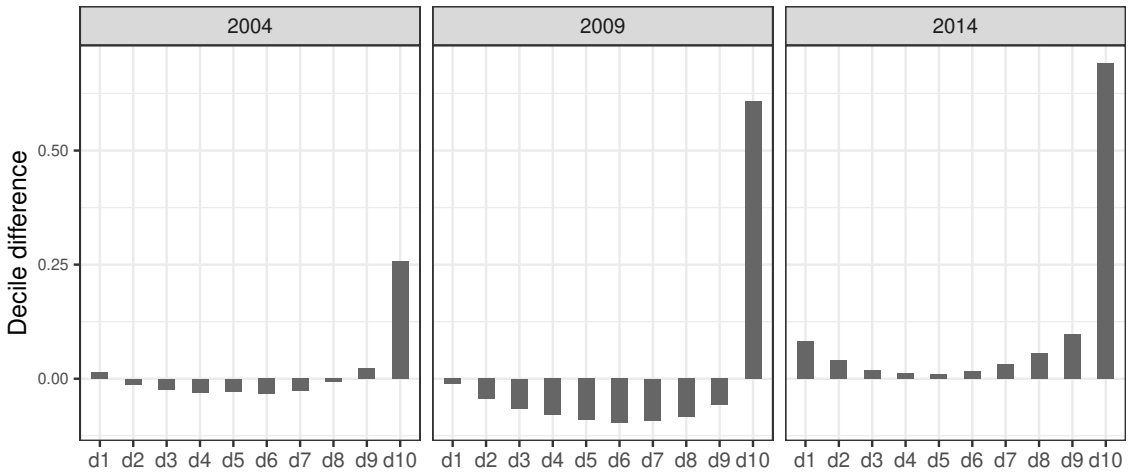
(a) Italy vs. Greece

(b) United Kingdom vs. Italy

Figure 1.5: Horizontal inequality between Italy and Greece (left panel) and between the UK and Italy (right panel) from 2004 to 2018, as measured by *HI*, *GGini* and G_b^A. All three measures in the left panel depict the same evolution of horizontal inequality between Italy and Greece, while *HI* and G_b^A strongly differ from *GGini* when comparing the United Kingdom and Italy (right panel)



(a) Italy vs. Greece



(b) United Kingdom vs. Italy

Figure 1.6: Differences between income deciles of Italy and Greece (top panel) and of the UK and Italy (bottom panel). The difference is reported as a fraction of the overall mean. The Italian deciles always dominate the Greek ones. Differently, there is stochastic dominance between income deciles of the UK and Italy only in the most recent years. This explains the conflicting trajectories of Figure Fig. 1.5b.

and Italy (Fig. 1.5b and Fig. 1.6b), we obtain a completely different picture. In this case, GGini evolves differently from HI - the correlation of the two series is 0.27. On the contrary, the correlation between G_b^A and HI is still greater than 0.99. We use Fig. 1.6b to explain this contrasting evidence. We note that all

the deciles of the two countries are quite similar in 2004, except for the last one. The distance between the deciles (but the first) increases considerably in 2009. Accordingly, both HI and G_b^A reach their maximum in 2009, while GGini reaches its minimum. This happens because the 10^{th} decile is higher in the UK while the other deciles are higher in Italy; therefore, the differences between the deciles compensate and produce a small difference between the means. We believe that it is not appropriate to consider the two distributions closer in 2009 than in 2004, and we argue that this is a relevant case in which oversimplification is at work. As further confirmation of our argument, the GGini massively increases between 2009 and 2014, despite no evidence of such a large increase in the distance between the two distributions. Again, the fast increase of the GGini is motivated by the sign of the differences between the deciles, rather than their magnitude. Going back to Figure 1.4, all perfectly correlated black dots correspond to pairs of countries such that the income distribution of one country dominates the other over all quantiles (stochastic dominance), as in the case of Italy and Greece. When there is no stochastic dominance between distributions, it is our conviction that the ability of G_b^A to measure horizontal inequality is considerably superior to that of the GGini.

1.8 Pairwise Differences and Group Contributions: A Topic to Explore

The significance of starting from Eq. (1.1) and directly addressing the pairwise differences in Gini index decomposition strategies has been emphasized by Rey and Smith [2013], Ceriani and Verme [2015], and Modalsli [2017].

Rey and Smith [2013] challenge the traditional notion of additive decomposability, introducing a within-between decomposition of the Gini index ($G = G_w + G_b$). Specifically, they decompose the Gini index using a matrix that defines pairs of neighbors and non-neighbors. The disparities among neighboring pairs form the within component, while the rest contribute to the between component. Notably, their between component, which hinges on pairwise differences rather than group averages, addresses the issue of oversimplification. Nevertheless, it is not suitable for mutually exclusive groups, such as geographical partitions, where it tends to overestimate the inequality between groups. For instance, their between-component yields a positive value even when groups have identical distributions, exemplifying the overestimation problem.

Ceriani and Verme [2015] introduce a group-based decomposition, focusing on each group's contribution to overall inequality. While their approach does not follow a strict within-between decomposition, it is still relevant to our discussion. They define the Gini index as the sum of individual contributions. They show that these contributions, which are based on the summation of pairwise differences, adhere to a set of desirable properties to measure individual diversity. This is intended to describe the degree of diversity of each individual from all the other members of the population. Interestingly, the authors showed that individual contributions can be aggregated, delivering group-wise contributions to global inequality.

The idea of group contributions is enhanced by Modalsli [2017]. His proposal is definitely comparable to the Rey and Smith decomposition, but the author built it to deal with a partition and the between term encounters the drawback that we have already discussed. To overcome this, he introduces a "mean between" term and a residual component, without any drastic innovation w.r.t. the decompositions considered at the beginning of this section. However, its work has the merit of having isolated "the contributions from country and country pairs as well as regions and region pairs to global inequality".

We believe that our decomposition proposal could combine and enhance the results of these three papers. Indeed, the procedure that calculates the two components can also deliver informative contributions from groups (or part of groups) and group pairs to global inequality, while also ensuring a robust assessment of horizontal inequality.

We can rewrite the numerator of Eq. (1.1) as

$$g = \sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}^{k}\left|y_i^k - y_j^k\right| + \sum_{k=1}^{K}\sum_{h=k+1}^{K}\sum_{j=1}^{n}\left(w_j^{kh} + w_j^{hk}\right)\left|y_j^k - y_j^h\right| \tag{1.20}$$

Defining the terms,

$$c_w^k = p_k g_w^k, \tag{1.21}$$

$$g_w^k = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}^{k}\left|y_i^k - y_j^k\right| / p_k \tag{1.22}$$

$$c_b^{kh} = p_k p_h g_b^{kh} \tag{1.23}$$

$$g_b^{kh} = \sum_{j=1}^{n}\left(w_j^{kh} + w_j^{hk}\right)\left|y_j^k - y_j^h\right| / p_k p_h \tag{1.24}$$

we obtain

$$g = \sum_{k=1}^{K} c_w^k + \sum_{k=1}^{K}\sum_{h=k+1}^{K} c_b^{kh} = \sum_{k=1}^{K} p_k g_w^k + \sum_{k=1}^{K}\sum_{h=k+1}^{K} p_k p_h g_b^{kh} \tag{1.25}$$

The terms c_w^k and g_w^k represent the gross and standardized numerator of the contributions of group k to within-group inequality. The gross contribution, c_w^k, scales linearly with the size of the group, while the standardized contribution, g_w^k, normalizes this contribution, allowing for direct comparison of within-group inequalities across different groups. Similarly, c_b^{kh} and g_b^{kh} denote the gross and standardized contributions, respectively, of the interaction between groups k and h to the between-group inequality. To obtain the gross and standardized contributions C_w^k, G_w^k, G_b^{kh} and C_b^{kh} one can divide c_w^k, g_w^k, c_b^{kh} and g_b^{kh} by the denominator of the Gini index. Thus, the Gini index is comprised of the sum of within-group contributions across the K groups and the $K(K-1)/2$ pairwise group contributions to between-group inequality.

We allocate half of each g_b^{kh} to the two groups k and h, defining

$$G_b^{k_h} = G_b^{h_k} = \frac{1}{2} G_b^{kh}$$

$G_b^{k_h}$ is the group k standardised contribution to between inequality arising from the interaction with group h. At this point, defining $G_b^k = \sum_{h=1}^{K} p_h G_b^{k_h}$ as the group k standardised contribution to between inequality, we can rewrite

$$G_b = \sum_{k=1}^{K} p_k G_b^k \tag{1.26}$$

Hence, we can also write the Gini index as

$$G = \sum_{k=1}^{K} p_k G_w^k + \sum_{k=1}^{K} p_k G_b^k \tag{1.27}$$

Finally, defining $G^k = G_w^k + G_b^k$ as the group k contribution to inequality, we can write

$$G = \sum_{k=1}^{K} p_k g^k \tag{1.28}$$

Summing up, the decomposition strategy proposed in this Chapter allows to define all the following decompositions of the Gini index:

$$G = G_w + G_b \tag{1.29}$$

$$G = \sum_{k=1}^{K} p_k G_w^k + \sum_{k=1}^{K} \sum_{h=k+1}^{K} p_k p_h G_b^{kh} \tag{1.30}$$

$$G = \sum_{k=1}^{K} p_k G_w^k + \sum_{k=1}^{K} p_k G_b^k \tag{1.31}$$

$$G = \sum_{k=1}^{K} p_k G^k \tag{1.32}$$

The decomposition in Eq. (1.29) is the one introduced in Section 1.4 and delivers the two within and between-group components of inequality. Eq (1.30) proposes a further decomposition of these two components, respectively, in the summations of the within contribution of each group and of the pairwise group contributions to between inequality. Eq. (1.31) expresses the Gini index as the summation of the group contributions to within and between inequality. Aggregating by groups these contributions we obtain the last decomposition - Eq. (1.32) - which sums up the contribution of each group to overall inequality in the spirit of Ceriani and Verme [2015] and Modalsli [2017].

Combining the information supplied by the components of the defined decompositions it is possible to fully characterise the relevance and the role that each group plays in the formation of overall, within and between inequality.

1.9 Conclusions

Our study proposes a new decomposition technique for the Gini index, which addresses the limitations of conventional methods in assessing horizontal inequality. The most widespread subgroup decompositions of inequality deliver within- and between-group components. The latter, which is based on the comparison of the means of the groups, is commonly used to assess horizontal inequality. However, further characteristics of the distributions are relevant for measuring horizontal inequality. Our decomposition fills this gap, proposing a between-group component that compares the entire distributions of the groups, rather than just their means. This makes it particularly appropriate to measure horizontal inequality, as confirmed by a Monte Carlo experiment and empirical analysis, both assessing the strong correlation of our component with a benchmark measure of horizontal inequality.

Our decomposition has another advantage. Conventional decompositions of the Gini index present a residual term in addition to the within- and the between-components. This residual disappears only if the distributions of the groups do not overlap. Exploiting a new insight into the Gini index, we disentangle each addendum of the index into two informative contributions to within- and between-group inequality, obtaining a two-component decomposition without the need to include a residual term. Hence, in our decomposition, inequality within groups explains a share of the Gini index, while excess inequality only depends on inequality between groups.

Empirical analysis confirms the relevance of our decomposition to support both cross-sectional and time-series inequality analysis. Studying the cross-sectional inequality between the European country pairs in 2018, we compare our between component (G_b^A), and that of the most widespread Gini decomposition (*GGini*), with the horizontal inequality benchmark (*HI*). *GGini* and *HI* are strongly correlated over several country pairs, but *GGini* underestimates horizontal inequality for countries with similar means and different distributions. On the contrary, G_b^A and *HI* have a strong correlation over all the country pairs. Studying inequality over time, our analysis spans between 2004 and 2018 and involves the Italy-Greece and the United Kingdom-Italy country pairs. It highlights that, differently from G_b^A and *HI*, the GGini is not always able to accurately capture the complexity of the differences between groups and their evolution.

Our discussion points out that both *GGini* and G_b^A accurately measure horizontal inequality when there is stochastic dominance between the distributions. In this case, the information provided by the two measures is the same of *HI*, and the advantage of our decomposition is to avoid the residual. This advantage disappears when the distributions of the groups do not overlap, because the residual of the conventional decompositions of the Gini index vanishes. When, instead, there is no distribution that dominates the other over all quantiles, $\rho(G_b^A, HI)$ remains high while *GGini* underestimates horizontal inequality and fails to assess its evolution. In such cases, we argue that our decomposition provides a more nuanced understanding of inequality between groups. The advantage of our decomposition comes at the cost of higher computational cost, but an efficient R code, which is publicly available on GitHub as detailed in the Appendix, mitigates this drawback.

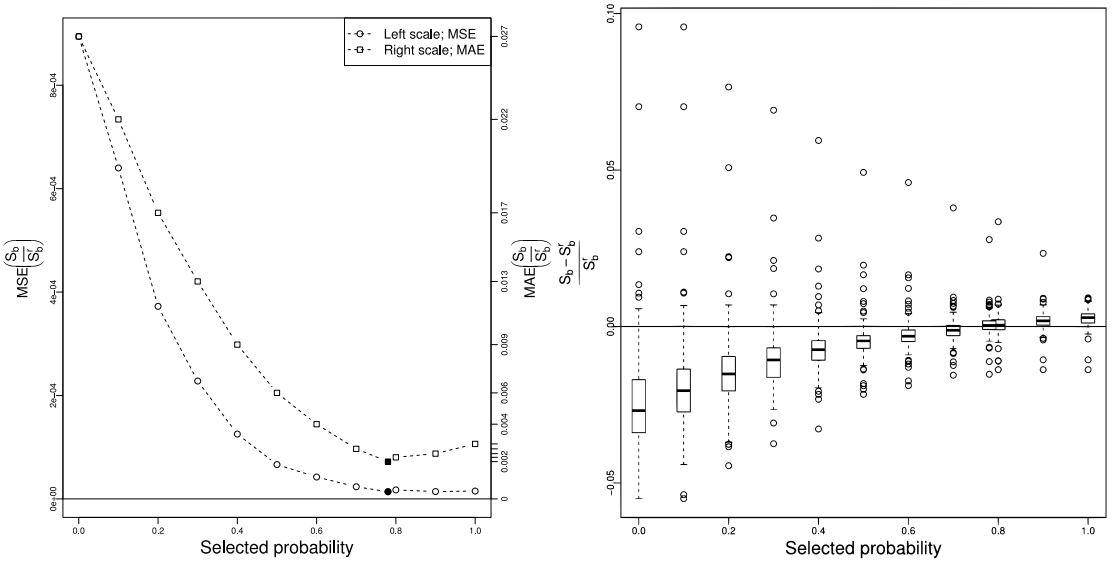
In conclusion, we firmly believe that the decomposition technique outlined in this paper greatly enhances the applicability of the Gini index to the study of inequality, as it opens new avenues for research into horizontal inequality and related areas.

Appendix

1.A On the quantilisation procedure

This part of the appendix discusses the suggested value of n and of the definition of quantiles in the quantilisation procedure, explaining their optimality and quantifying the magnitude of the approximation incurred.

Defining $\mathbf{p} = (p_1, \dots p_k)$ we can rewrite the suggested value of n as $n = \mathbf{p}\mathbf{n}^\intercal$. This expression determines



(a) MSE (left scale) and MAE (right scale) for different choices of n

(b) Boxplots of the 150 relative differences in S_b for different choices of n

Figure 1.A.1: Approximation of the between component share for different choices of n. The approximation is evaluated by comparing the 150 shares obtained by the quantilisation procedure with those by the exact approach

n as the average of the n_k, each weighted by the share of population p_k. We study the performance of this value in Figure 1.A.1, where we evaluate the approximation by looking at the relative discrepancy between

the two shares G_b/G from the exact approach and the quantilisation method. Precisely, we define $S_b = G_b/G$ as the between component share obtained by the quantilisation method and $S_b^r = G_b^r/G^r$ as the reference share obtained by the exact approach. The relative discrepancy is measured by the Mean and the Absolute Squared Error of S_b/S_b^r w.r.t. $1 = S_b^r/S_b^r$. They are calculated running 150 simulations and evaluating the empirical counterpart of $\text{MSE}(S_b/S_b^r) = \mathbb{E}[(S_b/S_b^r - 1)^2]$ and $\text{MAE}(S_b/S_b^r) = \mathbb{E}[|S_b/S_b^r - 1|]$.

At each iteration, we draw lognormal income vectors with sizes $\mathbf{n}$, as described in the second section of this Appendix. We compare the approximation with alternative choices of n, which are the minimum and the maximum of $\mathbf{n}$, its deciles (expressed in the plot as probabilities of the inverse distribution function) and the value obtained by Eq. (1.19). Generating the vector $\mathbf{n}$, we impose constraints on its elements to ensure affordable values for $lcm(\mathbf{n})$. To be specific, the algorithm firstly specifies K (= 5, 10 or 20). Then it builds a vector **mul** composed by the divisors of $2^4 3^3 5$ that belong to an interval $[min, max]$. The min (= 36 or 72) and the max (= 360 or 720) are both included in $\mathbf{n}$. The other $K-2$ values are sampled with repetition from **mul**. With this choice the lcm cannot exceed the value 2160 and the computations are affordable. Figure 1.A.1 represents the results for $K = 20$, $min = 72$ and $max = 720$.

As we show in Figure 1.A.1a, the proposed value of n - represented by the solid indicators - minimizes (or reach a value very close to the minimum of) the approximation that this method copes with, both for the MSE (left scale) and the MAE (right scale). This result is achieved thanks to vanished distortion and variance reduction, as we show in Figure 1.A.1b. We stress the irrelevance of the approximation when the suggested n is employed: according to the MAE, which is interpretable as average absolute percentage error, the error is 0.22%.

The magnitude of the percentage approximation changes with the simulation parameters, as Table 1.A.1 points out. It reports the percentage MAE of the between component share for different choices of n, K and of the interval $[min, max]$. Results are really encouraging. The values of the MAE are below the percentage point in half of the parameter specifications, and they are always below 1% when the suggested choice of n is employed.

For each choice of n, when the ratio max/min decreases, the approximation reduces, too. If that ratio stays constant, the MAE informs about better performance for higher min and max. Results are enhanced when n is selected by Eq. (1.19) and the number of groups is high. The described dependence of the MAE on the values of n, K and of the interval $[min, max]$ proves the consistency of our procedure, and may ensure even lower approximation in many realistic scenarios where the parameters are presumably more conducive.

The suggested choice of n almost always guarantees a relevant reduction in the computational cost associated to $n = max(\mathbf{n})$. This reduction is not negligible in our simulations: $\bar{p}$ is the average value, in the 150 simulations, associated to our choice of n in the inverse distribution function of $\mathbf{n}$. It is reported in the last column of the table and range from 0.69 to 0.84.

As supported by the values in the third column of the table - which decrease when $min(\mathbf{n})$ increases - it could be also acceptable to choose a value $n << min(\mathbf{n})$ if $min(\mathbf{n})$ is high and a computational saving choice is

		Percentage approximation for different deciles											$n = \mathbf{pn}^{\intercal}$	
K	$[min, max]$	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	n	$\bar{p}$
5	$[36, 360]$	6.12	4.88	4.14	3.20	2.53	2.17	1.42	1.20	1.00	0.92	0.90	0.94	0.78
	$[36, 720]$	8.28	6.41	5.45	4.13	3.26	2.79	1.71	1.44	1.05	0.91	0.86	0.91	0.83
	$[72, 360]$	2.66	2.19	1.93	1.52	1.18	1.01	0.76	0.69	0.61	0.57	0.58	0.60	0.73
	$[72, 720]$	4.07	3.24	2.79	2.14	1.70	1.45	1.00	0.86	0.69	0.61	0.59	0.64	0.79
10	$[36, 360]$	4.72	3.90	2.97	2.23	1.61	1.14	0.86	0.65	0.55	0.56	0.60	0.53	0.77
	$[36, 720]$	5.44	4.42	3.35	2.56	1.94	1.43	0.99	0.69	0.52	0.45	0.47	0.45	0.84
	$[72, 360]$	2.11	1.76	1.41	1.11	0.87	0.65	0.49	0.38	0.34	0.33	0.35	0.35	0.71
	$[72, 720]$	3.17	2.66	2.13	1.63	1.20	0.85	0.61	0.47	0.38	0.32	0.32	0.34	0.78
20	$[36, 360]$	3.72	2.94	2.26	1.77	1.26	0.83	0.54	0.37	0.31	0.36	0.42	0.30	0.75
	$[36, 720]$	4.85	3.81	2.93	2.20	1.55	1.08	0.67	0.43	0.29	0.25	0.30	0.24	0.82
	$[72, 360]$	1.78	1.53	1.17	0.89	0.64	0.44	0.29	0.22	0.21	0.23	0.26	0.21	0.69
	$[72, 720]$	2.68	2.20	1.66	1.26	0.89	0.61	0.43	0.29	0.24	0.26	0.32	0.22	0.78

Table 1.A.1: Percentage approximation of the quantilisation procedure as measured by the MAE of the between component share, for different parameter specifications. The last column represents the average fraction of elements in the vector **n** which are lower than the suggested n

required.

All these results are associated to the quantile definition that we suggest. Our choice comes from the comparison of the approximation achieved using the nine different quantile definitions presented in Hyndman and Fan [1996] in the procedure which generates Figure 1.A.1a.As we show in Figure 1.A.2,

the Definition 7 essentially presents the lowest MSE (and MAE) for each choice of n and it ensures computational advantages because the MSE approaches 0 for smaller n.

The outstanding performance of definitions 1 and 2 when the probability is close to 1 are exceptions. Both the definitions rely on a stepwise cumulative probability function which selects the quantiles from the set of values in the starting vector. Thus, if $p = 1$ and $max(\mathbf{n}) = lcm(\mathbf{n})$, the vector of quantiles corresponds to the $\mathbf{z}^k$ of the exact approach, and no approximation is encountered, with evident advantages starting from $p = 0.8$. Nonetheless, in the vast majority of real applications, the vector **n** is much more variable than the bounded vectors used in these simulations. Hence $lcm(\mathbf{n})$ is generally far from $max(\mathbf{n})$ and the Definition 7 from Hyndman and Fan [1996] is definitely recommended.

Actually, the optimal performance associated to the suggested quantiles selection strategy should not come as a surprise. Its outstanding results have a twofold explanation. First, the performance of the proposed choice of n directly derives from its consistency with the exact-approach weighting system. This choice assigns greater weights p_k to the sizes of the most sized groups, which is desirable because these are the groups with the biggest p_k. It is reasonable to preserve their information by choosing a large n and by

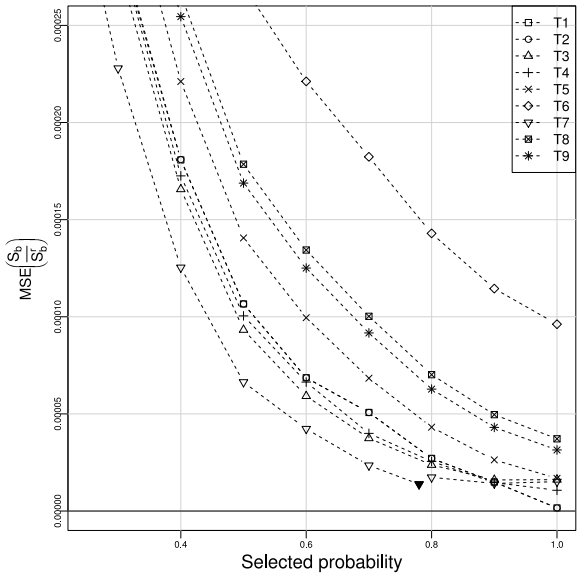


Figure 1.A.2: Between component share approximation of the 9 quantile definitions presented in Hyndman and Fan [1996]. Approximation is measured by the MSE with the procedure of Figure 1.A.1. Using the software R, each definition can be selected by the option *type* of the function *quantile()*. Here, T_j stands for selecting the option type=j

resampling the smallest groups. But if many small groups are present, n is attracted towards their small size. Here, the quantilisation of the biggest groups is preferred to resampling the many small groups. The second explanation for the optimal performance of the suggested strategy is the following. Eq. (1.18) selects the values $prob_j$ so as to partition the interval $[0,1]$ in $n-1$ equal parts, with 0 and 1 two of the n vertices of the partition. It is straightforward to verify that, with our suggestions, $min(\mathbf{x}^k)$ and $max(\mathbf{x}^k)$ are preserved for each n and k. Moreover, if $n_k = n \ \ \forall \ k$, then the vectors $\mathbf{x}^k$ are entirely preserved, too. Both these properties, which ensure robustness w.r.t. outliers, hold at the same time only employing Definition 7 from Hyndman and Fan [1996] and the suggested choice of the values $prob_j$.

1.B The income simulation algorithm

A Monte Carlo algorithm is employed to evaluate the approximation of the quantilisation procedure and to estimate correlations. This section of the appendix provides with the theoretical foundations of the income simulation procedure feeding both these algorithms.

The distribution of $\mathbf{n}$ is a K-variate uniform, where the number of groups K and the extremes of the distribution are determined *ex-ante*. A uniform distribution is also exploited to draw the expected average income of each group: $\mathbb{E}[\mu_k] \sim Unif(m,M)$. The minimum m of this distribution is set to 10^4. As for the

maximum M, it is fixed to $5 \cdot 10^4$ in the simulations described in the first part of this appendix. Differently, in the simulations presented in Section 1.6, M is varied to highlight how correlations depend on the variability of the means of the groups. This is possible because a modification of M directly affects $CV[\mathbb{E}[\mu_k]]$. For the uniform distribution $\mathbb{E}_u[\mathbb{E}[\mu_k]] = (M+m)/2$ and $\mathrm{Var}_u[\mathbb{E}[\mu_k]] = (M-m)^2/12$, therefore the coefficient of variation of $\mathbb{E}[\mu_k]$ is

$$CV[\mathbb{E}[\mu_k]] = \frac{\sqrt{\mathbb{V}_u[\mathbb{E}[\mu_k]]}}{\mathbb{E}_u[\mathbb{E}[\mu_k]]} = \frac{1}{\sqrt{3}}\frac{(M-m)}{(M+m)} \in \left[0, \frac{1}{\sqrt{3}}\right]$$

and, with m fixed, it only depends on the value of M.
The values of M are selected so that the coefficient of variation divides the interval into S equal parts. Denote by $M^{(s)}$, $s = 1 \dots S$ the different values required for this scope. The values $M^{(s)}$ satisfy:

$$\frac{M^{(s)}-m}{M^{(s)}+m} - \frac{M^{(s-1)}-m}{M^{(s-1)}+m} = c$$

with $M^{(0)} = m$ and $c = 1/\left(\sqrt{3}S\right)$. With easy calculations, the following holds:

$$M^{(s)} = \frac{m(cM^{(s-1)} + cm + 2M^{(s-1)})}{(2m - cM^{(s-1)} - cm)}$$

and the $M^{(s)}$ can be calculated iteratively.

Once all the parameters are fixed, we draw the incomes of each group from a lognormal distribution with expected value $\mathbb{E}[\mu_k] \sim Unif(m,M)$. The last requirement is to define a meaningful way to determine the two parameters η and σ^2 of the distribution. As it is well known, for a lognormal distribution the following holds:

$$\mathbb{E}[\mu_k] = e^{\eta_k + \frac{\sigma_k^2}{2}} \tag{1.33}$$

This equation allows an effective way to determine the two elements η_k e σ_k:

$$\eta_k = \alpha_k \ln \mathbb{E}[\mu_k] \tag{1.34}$$

$$\sigma_k^2 = 2(1-\alpha_k)\ln \mathbb{E}[\mu_k] \tag{1.35}$$

Their ratio is

$$c_k = \frac{\sigma_k^2}{\eta_k} = \frac{2(1-\alpha_k)\ln \mathbb{E}[\mu_k]}{\alpha_k \ln \mathbb{E}[\mu_k]} = \frac{2(1-\alpha_k)}{\alpha_k}$$

At this point, we consider the 82 couples of lognormal parameters estimated in Bandourian et al. [2002] using 82 real income distributions from 23 countries over several years (from the end of sixties to the end of nineties). We evaluate all the $c_i = \sigma_i^2/\eta_i$, $i = 1, \dots, 82$.

Verisimilar values for α_k can be obtained sampling a value of i for each group and posing $c_k = c_i$, solving the following equation:

$$c_i = c_k = \frac{2(1-\alpha_k)}{\alpha_k} \implies \alpha_k = \frac{2}{c_i + 2} \tag{1.36}$$

Therefore η_k and σ_k^2 are determined by Eq. (1.34)-(1.36).
The appropriateness of the last step - i.e. sampling a value of i for each group and using the correspondent c_i - is justified by the fact that the 82 values of α in Bandourian et al. [2002] are not influenced by the associated $\mathbb{E}[\mu_k]$: a simple linear regression reports an approximately null coefficient ($5.6 \cdot 10^{-4}$) and a large p-value (0.65) for the regressor $\mathbb{E}[\mu_k]$. Consequently, there are 82 credible proportions to split $\mathbb{E}[\mu_k]$ in η_k and $\sigma_k^2/2$. We exploit them to simulate income.

1.C R Code for implementing the decomposition

The R package implementing the described decomposition technique is available on GitHub. To ensure the `GiniDecA` package can be installed from GitHub, first check if the `devtools` package is installed. If not, install `devtools` using the following command in R:

```
if (!requireNamespace("devtools", quietly = TRUE))
  install.packages("devtools")
```

Afterward, install `GiniDecA` in the R environment

```
devtools::install_github("FedericoAttili/GiniDecA")
```

and use the `GiniDec` function as in the example below:

```
library(GiniDecA)
giniDec(x=c(8,3,6,2),z=c('k','k','h','h'), w = F)
#where x is the population income vector, z is the group label vector and w is
    the vector of weights
```

More details about the package functions are listed below, where the most important code of the package is organised into various subsections to ease its presentation. This code, which already meets most standards for publication on CRAN, will be refined to build an official R package.

1.C.1 Initial Setup and utilities

This part of the code is crucial for setting up the environment. It involves setting the working directory and loading the 'Rcpp' library. This library is particularly important because allows integration of R with C++, which is used to speed up the computation of the decomposition.

```

setwd(getwd())

library(Rcpp)
sourceCpp('ginicpp.cpp') # Replace with the actual path to your C++ file
```

The following snippet consists of various utility functions that facilitate the data processing and analysis:

1. 'mapping" function: This function converts group labels into numeric values, which is essential for analyses that require numeric group identifiers. It takes a vector 'groups' as input, converts it to a factor, and then to a numeric vector. It returns a named vector where the names are the original group labels and the values are the corresponding numeric identifiers.

2. 'gcd" function: A recursive function to calculate the greatest common divisor of two numbers, 'a" and 'b".

3. 'lcm_pair" function: Calculates the least common multiple of two numbers.

4. 'lcm" function: Extends the 'lcm_pair" function to operate on a vector of numbers, calculating the least common multiple of all elements in the vector.

```
########################
#useful functions

mapping=function(groups){#mapping of factor labels for conversion to numeric

  # Convert groups label to factor
  group_factor <- factor(groups)

  # Convert factor to numeric
  groups <- as.numeric(group_factor)

  # Convert numeric back to original categories
  key <- data.frame(label=levels(group_factor)[unique(groups)],id_group=as.
      numeric(unique(groups))-1)
  mapping <- setNames(key$label, key$id_group)
}


# Function to calculate the greatest common divisor (GCD) of two numbers
gcd <- function(a, b) {
  if(b == 0) {
```

```
    return(abs(a))
  } else {
    return(gcd(b, a %% b))
  }
}

# Function to calculate the least common multiple (LCM) of two numbers
lcm_pair <- function(a, b) {
  return(abs(a * b) / gcd(a, b))
}

# Function to calculate the LCM of a vector of numbers
lcm <- function(x) {
  # Check if the vector has at least one element
  if(length(x) < 1) {
    stop("Input vector must have at least one element.")
  }

  # Initialize LCM with the first element
  result_lcm <- x[1]

  # Iteratively calculate LCM for the rest of the elements
  for(i in x[-1]) {
    result_lcm <- lcm_pair(result_lcm, i)
  }

  return(result_lcm)
}

# Example usage
numbers <- c(3, 7, 8)
lcm(numbers) # Should return 24, which is the LCM of 4, 6, and 8
```

1.C.2 Group Size Recalibration Function

The ‘group_size_recalib” function is foundamental for the recalibration of the group sizes detailed in Section 1.5.1. This recalibration is performed to ensure that each group in the dataset has the same number of observations. It takes two arguments: ‘x”, an income vector, and ‘z”, a grouping vector. The function returns

a recalibrated dataset where the group sizes are equal, which is essential to employ the decomposition. The weight recalibration of Eq. (1.16) is performed later, directly in the 'giniDec" function (see further below).

```
1
2  group_size_recalib <- function(x, z) {
3    # The function takes an income vector x and a grouping vector z, then
         recalibrates
4    # the data such that each group has the same number of observations based on
         the
5    # least common multiple of the original group sizes. It scales the data using
6    # inverse proportions to maintain the overall distribution while ensuring
7    # uniform group sizes. The recalibration could be used for certain types of
8    # analysis where uniform group sizes are required for comparison or
         aggregation
9    # purposes.
10
11   # Check if x and z are of the same length
12   if (length(x) != length(z)) {
13     stop("Lengths of 'x' and 'z' must be the same")
14   }
15
16   # Calculate the number of unique groups in z
17   K <- length(table(z))
18
19   # Initialize vector to store the size of each group
20   n_k <- numeric(K)
21
22   # Populate n_k with the size of each group
23   for (i in 1:K) {
24     n_k[i] <- table(z)[[i]]
25   }
26
27   # Calculate the least common multiple (LCM) of group sizes
28   n <- lcm(n_k)
29
30   # Calculate the inverse proportion for each group size relative to LCM
31   p_k_inv <- n / n_k
32
33   # Combine income data with group labels and sort, by group and income
```

```
  x_old <- cbind(x, z)
  x_old <- x_old[order(x_old[, 2], x_old[, 1]), ]

  # Initialize new matrix for recalibrated data
  x_new <- matrix(nrow = n * K, ncol = 3)

  # Recalibrate data for each group
  for (i in 1:K) {
    start_index <- sum(n_k[1:i]) - n_k[i] + 1
    end_index <- sum(n_k[1:i])
    group_data <- x_old[start_index:end_index, 1]

    # Replicate data to fill new group size and assign group label and inverse
        proportion
    x_new[(n * (i - 1) + 1):(n * i), 1] <- rep(group_data, length.out = n)
    x_new[(n * (i - 1) + 1):(n * i), 2] <- rep(i, n)
    x_new[(n * (i - 1) + 1):(n * i), 3] <- rep(p_k_inv[i], n)
  }

  # Sort the recalibrated data by group and income
  x_new <- x_new[order(x_new[, 2], x_new[, 1]), ]

  # Return the recalibrated data
  return(x_new)
}
```

1.C.3 Quantile Calculation Function

The ‘quant” function is fundamental for the *quantilisation* procedure detailed in Section 1.5. It computes quantiles for specified groups in the dataset. The function takes several arguments: ‘gruppi” (the data vector), ‘indice_gruppo” (group index vector), ‘m‘ (number of quantiles to compute), ‘group_sizes‘ (vector of group sizes), and ‘type” (type of quantile calculation, to be chosen between the nine definitions defined in [Hyndman and Fan, 1996]). It returns a list containing the quantile values (‘valori_quantili”) and the corresponding group identifiers. Even in this case, the weight recalibration is performed in the ‘giniDec” function (see further below)

```
library(Hmisc)
quant <- function(gruppi, indice_gruppo, m, group_sizes, type=7, weights=F) {
```

```
  quantili <- seq(0, 1, 1 / (m - 1))  # Sequence of quantiles to calculate
  valori_quantili <- c()  # Initialize vector to store quantile values
  g <- c()  # Initialize vector to store group identifiers

  if (!is.numeric(weights[1])) {
    # If weights are not provided
    for (i in group_sizes) {
      # Calculate quantiles for each group and append to valori_quantili
      valori_quantili <- append(valori_quantili, as.vector(quantile(gruppi[
          indice_gruppo == i], quantili, type = type)))
      # Append group identifier repeated m times
      g <- append(g, rep.int(i, m))
    }
  } else {
    # If weights are provided
    for (i in group_sizes) {
      # Calculate weighted quantiles for each group and append to valori_
          quantili
      valori_quantili <- append(valori_quantili, as.vector(wtd.quantile(gruppi[
          indice_gruppo == i], weights[indice_gruppo == i], quantili)))
      # Append group identifier repeated m times
      g <- append(g, rep.int(i, m))
    }
  }

  return(list(gruppi = valori_quantili, indice_gruppo = g))
}


########################
```

1.C.4 The giniDec function

The ‘giniDec” function is the core function to calculate the decomposition proposed in this chapter. It performs Gini decomposition on the given data. The function takes several arguments, including ‘x’ (input variable vector), ‘z’ (group belonging vector), ‘w’ (individual weights), ‘m’ (quantile level or value), ‘n_equalizer’ (method for normalization, equal to ‘quant’ - the default option - or ‘ricalibrazione’; the default option is advised and it is the only supporting $w \neq F$, while $n\_equalizer =$‘ricalibrazione’ is specifically

designed just to assess, in Section 1.A, the approximation incurred by employing the quantilisation procedure), 'type' (quantile definition). It returns a list of results, including Gini coefficient, within and between group Gini components, and their respective shares. If 'contrib' is set to TRUE (default is FALSE), the function also returns the group and pairwise group contributions to within and between components, as defined in Eq. 1.21 and Eq. 1.23.

```
#CORE FUNCTION

library(dineq)

giniDec=function(x,z,w=F,m=NULL,n_equalizer='quant',type=7,contrib=F){
  #x is the vector with the input variable in the population
  #z is the vector of group belonging; it defines the partition in groups
  #w represents the vector of individual weights
  # m tra 0 e 1 indica il livello del quantile, or if >1 directly the value to
      employ,  if NULL the value suggested in Attili (2020) is employed
  #type indica la def di quantile utilizzata
  if(!is.numeric(w[1])){data=data.frame(x,z,1)}else{data=data.frame(x,z,w)}
  colnames(data)=c('x','groups','w')
  data_store=data
  #force the three columns to be numeric,
  data$x=as.numeric(data$x)
  data$w=as.numeric(data$w)
  data$groups=as.numeric(as.factor(data$groups))
  Gini=gini.wtd(data$x,data$w)

  data=data[order(data$groups),]
  x=data$x
  w=data$w
  z=data$groups


  #recalibration
  #nb if n_equalizer=="ricalibrazione", data should not contain weights

  n_k=as.vector(table(z))
  K=length(n_k)
  kapp=unique(z)

```

```
N_orig=sum(w)
mu=sum(x*w)/N_orig
w_k=as.vector(tapply(w, z, sum))
if(is.null(m)){
  mm=0
  for (i in 1:K){mm=mm+n_k[i]*(w_k[i]/sum(w_k))}
  q=seq(0,1,0.005)
  qq=quantile(n_k,q)
  dif=abs(qq-mm)
  m=q[which.min(dif)]
}

if(n_equalizer=='quant'){
  if(m<=1){m=as.integer(quantile(n_k,m))}
  quantili=quant(x, z, m, kapp,type, w)
  z=quantili$indice_gruppo
  x=quantili$gruppi
  numerosita=c(matrix(rep(w_k,m),m,K,byrow = T))
  p_k_inv=m/numerosita
  x=cbind(x,z,p_k_inv)
  rm(z,p_k_inv)

}

if(n_equalizer=="ricalibrazione") {
  x=ricalibrazione_gruppi(x,z)
  m=mcm(n_k)    #restituisce x=cbind(x,z,p_k_inv) con numerosit? apparate e
      ordinato prima su z e poi su x
}

x[,3]=1/x[,3]

data=as.matrix(x)

output_giniDecomposition <- giniDecomposition(data,contrib=contrib)#contrib=F
    by default
#this function is written in c++. It takes data and return
#the value of the two components (g_w and g_b). If contrib=T,
#also two matrices are provided with the contribution of each
```

```
#pair of units to within and between component
#organise results
G_dot=output_giniDecomposition$g_w+output_giniDecomposition$g_b
#evaluate the components
G_w=output_giniDecomposition$g_w*Gini/G_dot
G_b=output_giniDecomposition$g_b*Gini/G_dot
#evaluate the component shares
within_share=G_w/Gini
between_share=G_b/Gini

if(contrib==T){
  map=mapping(data_store$groups)
  #evaluate (absolute) contribution of groups to within inequality
  group_w_contrib=(tapply(output_giniDecomposition$w_contributions$value,
      output_giniDecomposition$w_contributions$k,sum)+tapply(output_
      giniDecomposition$w_contributions$value,output_giniDecomposition$w_
      contributions$h,sum))/2*Gini/G_dot
  names(group_w_contrib) <- map[names(group_w_contrib)]


  #evaluate (absolute) contribution of groups to between inequality
  group_b_contrib=(tapply(output_giniDecomposition$b_contributions$value,
      output_giniDecomposition$b_contributions$k,sum)+tapply(output_
      giniDecomposition$b_contributions$value,output_giniDecomposition$b_
      contributions$h,sum))/2*Gini/G_dot
  # Replace the names in group_b_contrib with corresponding labels
  names(group_b_contrib) <- map[names(group_b_contrib)]


  #evaluate (absolute) pairwise contribution of groups to between inequality
  group_pairwise_contr=tapply(output_giniDecomposition$b_contributions$value,
      list(output_giniDecomposition$b_contributions$k,output_
      giniDecomposition$b_contributions$h),sum)*Gini/G_dot
  group_pairwise_contr=group_pairwise_contr[upper.tri(group_pairwise_contr)]+
      t(group_pairwise_contr)[upper.tri(t(group_pairwise_contr))]
  #set names
  names(group_pairwise_contr)=apply(combn(names(group_b_contrib),2),2,paste,
      collapse='-')

```

```
99
100     #ADD
101     #evaluate (absolute) contribution of ranks to within and between
102
103   }
104
105   results <- list(
106     Gini = Gini,  # Assuming Gini is calculated somewhere in your function
107     G_w = G_w,
108     G_b = G_b,
109     within_share = within_share,
110     between_share = between_share
111   )
112
113   if (contrib) {
114     # Additional results when contrib is TRUE
115     results$group_w_contrib = group_w_contrib
116     results$group_b_contrib = group_b_contrib
117     results$group_pairwise_contr = group_pairwise_contr
118     # ... [any other additional results you want to include]
119   }
120
121   return(results)
122 }
```

Example usage:

In the following we provide examples to illustrate the application of the above functions. It includes basic scenarios and more complex cases involving different-sized groups and additional parameters.

```
1 ## Example usage
2
3 # Generating 4 random numbers between 1 and 10
4 x = runif(4, 1, 10)
5
6 # Creating a vector of group identifiers (1, 2, 1, 2)
7 groups = c(rep(c(1:2), 2))
8
9
```

```
# Computing the Gini decomposition
res1 = giniDec(x, groups, contrib = T)


##   Example with different-sized groups and weights

# Generating 5 random numbers between 1 and 10
x = runif(5, 1, 10)

# Creating a vector of group identifiers (1, 2, 1, 2, 2)
groups = c(rep(c(1:2), 2), 2)

# Creating a weight vector (1, 2, 1, 2, 2)
w = c(rep.int(c(1, 2), 2), 2)

# Computing the Gini decomposition with additional parameters
res2 = giniDec(x, groups, w = w, n_equalizer = 'quant', type = 7, contrib = T)
```

Chapter 2

Decomposing inequality after asymmetric shocks: an analysis of Italian household consumption

Keywords: Inequality decomposition, Asymmetric shocks, Gini index, COVID-19, Monte Carlo experiment

2.1 Preamble

In the opening chapter, we delved into methods of inequality decomposition, focusing particularly on the Gini index of a population divided into groups. Our analysis scrutinised the 'between-component' of various decompositions, commonly used to evaluate inequality between groups. Recently, the between-component has been increasingly interpreted as a measure of horizontal inequality. We contend that traditional decomposition methods are inadequate for accurately assessing and tracking horizontal inequality. They underestimate it in scenarios lacking stochastic dominance among the distributions of different groups. For this reason, we introduce a novel decomposition approach with a between-component that is more suitable for evaluating horizontal inequality.

In this chapter, leveraging open-source data from the Household Budget Survey (HBS) issued by the Italian National Statistics Institute (ISTAT), we examine the impact of Covid-19 on horizontal inequality of the consumption expenditure of Italian households, according to three inequality factors such as gender, age and geography. Our findings reveal that traditional decomposition methods fall short in accurately capturing the evolution of this inequality. Covid-19's asymmetric impact across and within various population

groups disrupts or establishes stochastic dominance, significantly influencing the assessment of inequality's progression. Through a Monte Carlo experiment, we underscore the necessity of selecting an appropriate measure for assessing horizontal inequality, especially in the face of asymmetric shocks like the Covid-19 pandemic or inflation.

This chapter is a collaborative effort with Michele Costa from the University of Bologna. The preliminary results of this research was presented at the 51st Scientific Meeting of the Italian Statistical Society (SIS) in Caserta, 2022, providing a platform for engaging discussions and further refinement of the study. The manuscript has been recently accepted for publication in the SIS2022 Proceedings volume "Advanced Methods in Statistics, Data Science and Related Applications". I am indebted to editors Antonio Balzanella, Matilde Bini, Lucio Masserini, and Rosanna Verde, along with two anonymous referees, for their invaluable comments and suggestions

2.2 Introduction

In little more than a decade, most of the world's economies have suffered three of the largest shocks since the second world war. The economic repercussions of the financial crisis and the COVID-19 restrictions have been studied extensively. Many efforts are now devoted to the study of high inflation, mainly driven by rising energy prices. All three shocks affected asymmetrically economic agents, business and personal income (Milanovic [2022], Bell et al. [2022], Jaravel [2021], Fana et al. [2020]). As a result, the debate on the implications of asymmetric shocks for monetary policy, prices and risk sharing flourished, receiving adequate attention and support. The relevance of asymmetric shocks and their distributional consequences are also well recognised by Bourguignon [2011] and Berman and Bourguignon [2024], which introduce and employ non-anonymous growth incidence curves to consider initial and final income positions. Similarly, Jenkins and Van Kerm [2016] advocate for a longitudinal approach to assessing income growth, emphasizing the importance of individual income growth profile. These methodologies exist to study the asymmetric effect of growth or shocks in the income distribution dynamics.

In our opinion, the asymmetry of shocks should be considered when studying the dynamics of inequality between groups, particularly in asymmetric contexts like the post-COVID Italian economic landscape. On the contrary, methods to study inequality have barely evolved, while we argue that the usual assessment of inequality is not robust to asymmetric shocks when the focus is on comparisons between population subgroups. Indeed, traditional methods generally compare distributions of subgroups by looking at the difference between their means. However, the comparison of means is not representative of the distance between distributions when the distributions are heterogeneously and asymmetrically affected by shocks.

In this paper, we focus on Italian household consumption after COVID-19, which represented a very strong shock with serious health, psychological, social and economic consequences. We show that usual

methods to assess inequality between subgroups do not fully capture the heterogeneous and asymmetric modifications of the consumption distribution of subgroups. At the same time, we propose a complementary methodology helping to deliver a more complete picture of inequality between groups.

The need to evaluate the effects of COVID on inequality has immediately been recognized. It led to a thriving literature (see e.g. Stantcheva [2022] for a review), to which we contribute by including the methods related to inequality decomposition in the methodological toolbox of researchers and policymakers. Indeed, the between-inequality component represents a strategic tool to explore the effects of asymmetric shocks on divides.

COVID-19 impacted Italian households consumption in 2020 through different channels. The principal ones are income reduction Gallo and Raitano [2020], lockdown, restrictions, and behavioral changes (Hasan et al. [2021], Marques Santos et al. [2020]). Except for the lockdown, which flattened household consumption to unprecedented levels in all the social strata, the other channels shaped the distribution of consumption (and income) with heterogeneous intensity along its deciles and subgroups. The presence of substantial asymmetries which characterize the effects of COVID has been thoroughly investigated and highlighted by Bell et al. [2022]. To fully understand the heterogeneous changes that have occurred, it is therefore important to measure the inequality between subgroups at all distribution points.

We exploit the Italian household budget survey (HBS) to investigate consumption inequality of household residing in Italy. With respect to three different partitions, which define subgroups based on age, gender and geographical factor, we compare how different can be the picture of household consumption inequality when considering decompositions that follow different paradigms. In particular, we compare the summer of 2019 and 2020, which are the two closest comparable periods before and after the lockdown and the main restrictions. We show that the increase in age and gender divide is much smaller if measured by comparing the whole subgroup distributions. The impact of COVID-19 on household consumption is instead homogeneous among the Italian regions and the different points of their distributions, therefore decompositions of inequality agree on a modest reduction of the geographical divide in consumption.

This paper is the extension of a previous version presented at the 51st Scientific Meeting of the Italian Statistical Society (Attili and Costa [2022]) and proceeds as follows. Section 2.3 builds a common framework on the methodologies for studying inequality factors. Section 2.4 theoretically discusses the impact of shocks on between-inequality indicators. Section 2.5 presents the data employed and illustrates their main features. Section 2.6 discusses the response of different measures after simulated shocks and after COVID-19, and exploits a bootstrap algorithm to infer their confidence intervals. Section 2.7 concludes.

2.3 Inequality factors and inequality decomposition

2.3.1 Assessment of inequality factors

Inequality factors such as gender, age, and geographical area significantly impact various societal aspects. By segmenting the overall population into subgroups based on two or more reference values of an inequality factor, we explore the influence of these factors on overall inequality.

Consider a variable of interest x, such as income or consumption, with an average μ in the population of interest. For two subgroups of the population k and h with means μ_k and μ_h, we begin by examining the gap between these subgroups. This gap is quantified by the relative difference between their means:

$$\Delta\mu_{kh} = |\mu_k - \mu_h|/\mu \tag{2.1}$$

which provides a classic aggregate indicator. There are also other aggregated indicators, but averages represent by far the most commonly used tool, so we can consider relation (1) as a benchmark.
A more detailed framework than relation (1) can be developed by extending the comparison to a set of points of the distributions of subgroups k and h. Beyond averages, a detailed comparison involves the distributions of subgroups k and h at various quantiles. As in Chapter 1, let x_q^k and x_q^h, $q = 1, \ldots, Q$, be the Q quantiles of subgroups k and h, respectively. The difference at each quantile level, Δx_q^{kh}, is:

$$\Delta x_q^{kh} = |x_q^k - x_q^h|/\mu$$

The aggregated measure of these differences, d_Q^{kh}, is given by:

$$d_Q^{kh} = \sum_{q=1}^{Q} \Delta x_q^{kh}/Q \tag{2.2}$$

Compared to $\Delta\mu_{kh}$, the quantity d_Q^{kh} enables a comparison based on a broader set of information, which increases with Q. This is particularly useful when the distributions of the two subgroups exhibit different shapes. In the case of a population divided into two groups, this measure coincides with the measure of diversity between income distribution for mean-scaled data (HI) defined in Chapter 1. The extension to the case with more than two groups involves a series of binary comparisons and their synthesis, leading to the more general definition of HI. In this chapter, we explicitly consider in the notation the dependence of HI on the selected number of quantiles Q. In particular, we adopt $Q = 10$ and evaluate d_{10}^{kh}, thus aligning with the traditional inequality framework based on deciles.

We aim to assess whether the relevance of inequality factors and their dynamics are sensitive to the employed methodology, especially when comparing Eq.2.1 and Eq.2.2. We argue that in cases of asymmetric shocks, the dynamics of the difference between means may not fully represent the corresponding evolution of the distributions. This argument is extended to more refined methods of inequality decomposition, with

a particular focus on the between-component, which is crucial in evaluating the importance of underlying inequality factors and the gap between subgroups.

Our discussion applies to various inequality measures, including generalized entropy indexes like the Theil index, the Zenga index (Porro and Zenga [2020]), and the Gini index G. We continue to explore the context of the Gini index and of the three decompositions considered in Chapter 1, where the decomposition methodologies are presented and discussed. We focus on the between-component of inequality, which evaluates the impact of differences between subgroups on overall inequality (Elbers et al. [2008], Giorgi [2011]). Table 2.3.1 summarises the between-group indicators, including these decompositions and the indicators $\Delta\mu_{kh}$ and d_Q^{kh}.

Table 2.3.1: Summary of Between-Group Indicators

Category	**Type**	**Indicator**
Benchmark Indicators	Gap Between Means	$\Delta\mu_{kh} = \lvert\mu_k - \mu_h\rvert/\mu$
	Quantile Comparison	$d_Q^{kh} = \sum_{q=1}^{Q} \Delta x_q^{kh}/Q$
Decomposition Derived Indicators	Bhatt. and Mahalanobis	$G_b^{BM} = \sum_{k=1}^{K}\sum_{h=1}^{K} n_k n_h \lvert\mu_k - \mu_h\rvert/2\mu$ (2.3)
	Yitzhaky and Lerman	$G_b^{YL} = 2cov(\mu_k, \bar{F}_k)/\mu$ (2.4)
	Attili	$G_b^{A} = \sum_{k=1}^{K}\sum_{h=1}^{K}\sum_{j=1}^{n}\left\lvert x_j^k - x_j^h\right\rvert w_j^{kh}/(2\mu n^2 K^2)$ (2.5)

The table above provides a detailed summary of key between-group indicators, each serving a specific role in analyzing inequality.

- **Gap Between Means ($\Delta\mu_{kh}$)**: This fundamental benchmark indicator measures the disparity based on average values within groups, laying the groundwork for more complex inequality assessments.
- **Quantile Comparison (d_Q^{kh})**: This indicator extends the analysis beyond averages to differences at specific quantiles, uncovering disparities across different population segments.

- **Bhattacharya and Mahalanobis Decomposition (G_b^{BM})**: As part of the Gini index decomposition, this method focuses on differences in subgroup means, similarly to the Gap Between Means. It offers a traditional approach to understanding the contribution of mean differences to overall inequality.
- **Yitzhaky and Lerman Decomposition (G_b^{YL})**: This decomposition also stems from the Gini index and is akin to the Gap Between Means in its focus on subgroup means and distributional overlaps. It provides valuable insights, especially where distributional overlaps are significant.
- **Attili's Decomposition (G_b^{A})**: Offering a perspective similar to the Quantile Comparison, Attili's method compares observations ranked similarly in different subgroups. This approach is particularly insightful for detailed, rank-based subgroup comparisons, revealing aspects of inequality that might be missed by average-based methods.

These indicators collectively enable a comprehensive analysis of inequality. While G_b^{BM} and G_b^{YL} align closely with $\Delta\mu_{kh}$ by focusing on subgroup mean differences, G_b^{A}, like d_Q^{kh}, delves into a more granular, distribution-wide examination. In contexts of asymmetric shocks, these tools are invaluable for uncovering the multifaceted impacts on inequality dynamics, guiding nuanced policy and research responses.

This chapter thus seeks to extend the theoretical framework established in Chapter 1, applying it to a nuanced examination of inequality dynamics in various contemporary contexts. The analysis will contribute to a deeper understanding of how different methodologies influence our perception and analysis of inequality factors, particularly when examining changes over time or under specific economic conditions.

2.4 Shocks and assessment of inequality factors

In this section, we study the effect of symmetric and asymmetric shocks on the indicators discussed in the previous section.

We consider a population of size N and we define the income of individual i before the shock as y_i, with $i = 1,\dots,N$. The population is divided into two groups, k and h, of size n_k and n_h ($n_k + n_h = N$). Let the income of the individual i of the group k be y_i^k before the shock, with $i = 1,\dots,n_k$.

We start by discussing symmetric shocks. We can model symmetric shocks as multiplicative or additive shocks. As for multiplicative shocks, we define:

$$x_i = y_i \cdot (1 + c) \tag{2.6}$$

where c is the variation induced by the shock.

It is simple to show that the values of both Eq. (2.1) and Eq. (2.2) stay unchanged after substituting Eq. (2.6). Indeed, such a shock does not even affect inequality measures, according to the principle of scale invariance

(Allison [1978]). We model additive symmetric shocks as:

$$x_i = y_i + c \tag{2.7}$$

In this case, we can show that:

$$\Delta\mu_{kh} = |\mu_k - \mu_h|/(\mu + c) \tag{2.1'}$$

and

$$\Delta x_q^{kh} = |x_q^k - x_q^h|/(\mu + c)$$

so that

$$d_Q^{kh} = \frac{1}{Q \cdot (\mu + c)} \sum_{q=1}^{Q} |x_q^k - x_q^h| \tag{2.2'}$$

Compared to Eq. (2.1) and Eq. (2.2), both Eq. (2.1') and Eq. (2.2') vary in the same direction and proportion. If c is positive, they decrease by a factor $\bar{y}/(\bar{y}+c)$, while they increase for negative shocks. This concordance could weaken when introducing stochastic shocks. We model them both as additive and multiplicative:

Multiplicative		Additive	
$x_i = \begin{cases} y_i \cdot (1+c) & \text{with prob. } \rho \\ y_i & \text{with prob. } 1-\rho \end{cases}$		$x_i = \begin{cases} y_i + c & \text{with prob. } \rho \\ y_i & \text{with prob. } 1-\rho \end{cases}$	(2.8)

It is not simple to treat Eq. (2.2) analytically in the stochastic case, due to the potential change in decile membership of some units after the shock. The study of shock sensitivity is also challenging for the between-inequality indicators.

In general, the analytical study of indicators is even harder when moving toward asymmetric shocks, which, as we argue from the beginning, can significantly reduce the grounds for the concordance of the indicators.

We can generalize the analysis of the effects of shocks by resorting to a mixture distribution

$$X^m = (1-t)Y + tX$$

where Y and X are the income distribution before and after the shock, respectively, while t ranges between 0 and 1, which is the interval in which the shock produces its effects. When $t = 0$ the shock has not yet produced any effect and $X^m = Y$. When $t = 1$ it has generated its maximum consequences and $X^m = X$.

The mean of X^m is

$$\bar{X}^m = \bar{Y} + t(\bar{X} - \bar{Y})$$

and we can evaluate $\delta\bar{X}^m/\delta t$ in order to assess the effects of the shock on the mean.

The linear structure of the mean greatly simplifies the evaluation also in case of asymmetric shocks. For example, if the shock adds a quantity c only to the incomes below y^*, we obtain

$$X = Y + c\mathbb{1}_{Y<y*}$$

where $\mathbb{1}$ is the indicator function. The mean of the after shock distribution is

$$\bar{X} = \bar{Y} + cF(y^*)$$

At this point, we can rewrite the mean of the mixture as

$$\bar{X}^m = \bar{Y} + t(cF(y^*))$$

Looking at its derivative

$$\delta\bar{X}^m/\delta t = cF(y^*)$$

we can study the behaviour of the mean while the shock produces its effects. In this case $\delta\bar{X}^m$ increases linearly with t and its slope $cF(y^*)$ increases with y^*. This approach can be easily generalized to analyse the relative evolution between the averages of multiple groups.

Unfortunately, the extension to inequality measures, from the Gini index to the components of its decompositions, is far more complex (Lambert and Decoster [2005]; Gastwirth [2017]) especially in the case of an empirical (discrete) distribution, where the Gini index is based on the absolute value.

In the following, in order to overcome the difficulties related to the analytical evaluation of the impact of asymmetric shocks on the inequality measures, we exploit a Monte Carlo algorithm. It starts from the data on Italian households' consumption and aims at studying how indicators of between-inequality react to different kinds of shocks. We maintain the distinction between additive and multiplicative shocks and we model the shocks as described in the scheme of Tab. 2.4.1.

Table 2.4.1: Additive and multiplicative shock to individual i of group h as modelled and studied by the Monte Carlo algorithm

Asymmetry	Additive process	Multiplicative process	Probability
Shock asymmetrically affect both groups and deciles	$x_i^k = \begin{cases} y_i^k + c \\ y_i^k \end{cases}$	$x_i^k = \begin{cases} y_i^k \cdot (1+c) \\ y_i^k \end{cases}$	$\begin{cases} \rho_{kq} \\ 1-\rho_{kq} \end{cases}$

The table encapsulates the conceptual framework used in the Monte Carlo algorithm to study the impacts of additive and multiplicative asymmetric shocks on individuals within various groups and deciles. Mathematical representations are provided for the two distinct types of shocks. The additive process adds a constant c to y_i^k, the value of individual i in group k before the shock, with probability ρ_{kq}, or does not affect y_i^k with probability $1-\rho_{kq}$. The multiplicative process, with the same probabilities, involves an increase by a factor of $(1+c)$, indicating a proportional change rather than a fixed increment, or no change at all. The probability ρ_{kq} is a crucial element of the model, and denotes the heterogeneous probability of a shock occurring. This probability is not only specific to each group (k), but also sensitive to the rank of the unit i within its group (q). The different scenarios implemented, expressed in terms of the choice of ρ_{kq}, and the details of the Monte Carlo study, as well as its results, are illustrated in Section 5.

2.5 Data

Our analysis makes use of the household budget survey (HBS), run by Istat, which allows to study consumption expenditure behaviours of households residing in Italy according to their main socio-economic and territorial characteristics. We focus on surveys carried out in July, August and September of 2019 and 2020 (4650 and 6964 households in 2019 and in 2020, respectively), because from July to September 2020 the government removed the shutdown of most activities and many of the restrictions on social distancing and travels. By focusing on summer consumption, we improve the identification of the (potentially structural) effects of income losses and behavioral changes from 2019 and 2020.
Household consumption is equivalised for the number of household members by using the square root equivalence scale. We investigate three different bi-partitions based on age, gender, and geographical factor. They are well-recognised inequality factors, giving rise to age, gender, and geographical divides. Consistently with the theoretical development in Section 2, it is possible to extend the analysis to the case of k subgroups. The rationale behind the choice of two subgroups is to achieve a simpler and clearer assessment of the difference between the subgroups and, consequently, of the effect of the underlying inequality factor.

In Tab. 2.5.1 we report deciles and means of the Italian households' monthly consumption in 2019 and 2020, detailed by subgroups defined by age, gender and geographical factors. In Fig. 2.5.1 we highlight the difference between group deciles. The first panel compare Young-(<65) and Old-headed(≥ 65) households. In 2019 the former have higher consumption than the latter but in the extreme deciles. Consumption of all groups and deciles decreases in 2020, but COVID-19 more evidently reduces rich Old-headed households' consumption, so that Young-headed households show higher consumption at all deciles. Higher values notoriously attract the means, so the difference between subgroup means ($\Delta\mu_{kh}$) increases by 67.7% from 2019 to 2020. However, COVID-19 affected the consumption distributions of the two subgroups asymmetrically, such that, in contrast with the means, the deciles at the bottom of the distribution get closer in 2020

Table 2.5.1: Deciles and means of the Italian households' monthly consumption in 2019 and 2020, detailed by subgroups defined by age, gender and geographical factors. Age subgroups are based on the partition "Young (aged less than 65) vs. Old (aged more than 64)"; the "Gender Divide" panel compares "Men vs. Women"; the "Geographical divide" panel compares "North and center vs. South and islands"

	2019		2020	
	Young	Old	Young	Old
d_1	591	**592**	**584**	572
d_2	**879**	844	**853**	823
d_5	**1470**	1322	**1404**	1293
d_9	**2700**	2548	**2639**	2455
d_{10}	4035	**4273**	**3953**	3723
mean	**1784**	1705	**1731**	1604

	2019		2020	
	Men	Women	Men	Women
d_1	590	**593**	**608**	533
d_2	**873**	852	**867**	789
d_5	**1438**	1368	**1386**	1310
d_9	**2688**	2581	**2635**	2469
d_{10}	4088	**4171**	**3948**	3721
mean	**1772**	1727	**1722**	1614

	2019		2020	
	North and Center	South and Islands	North and Center	South and Islands
d_1	**676**	501	**663**	489
d_2	**981**	716	**947**	695
d_5	**1585**	1122	**1499**	1083
d_9	**2868**	2083	**2779**	2061
d_{10}	**4397**	3253	**4117**	3101
mean	**1926**	1392	**1838**	1356

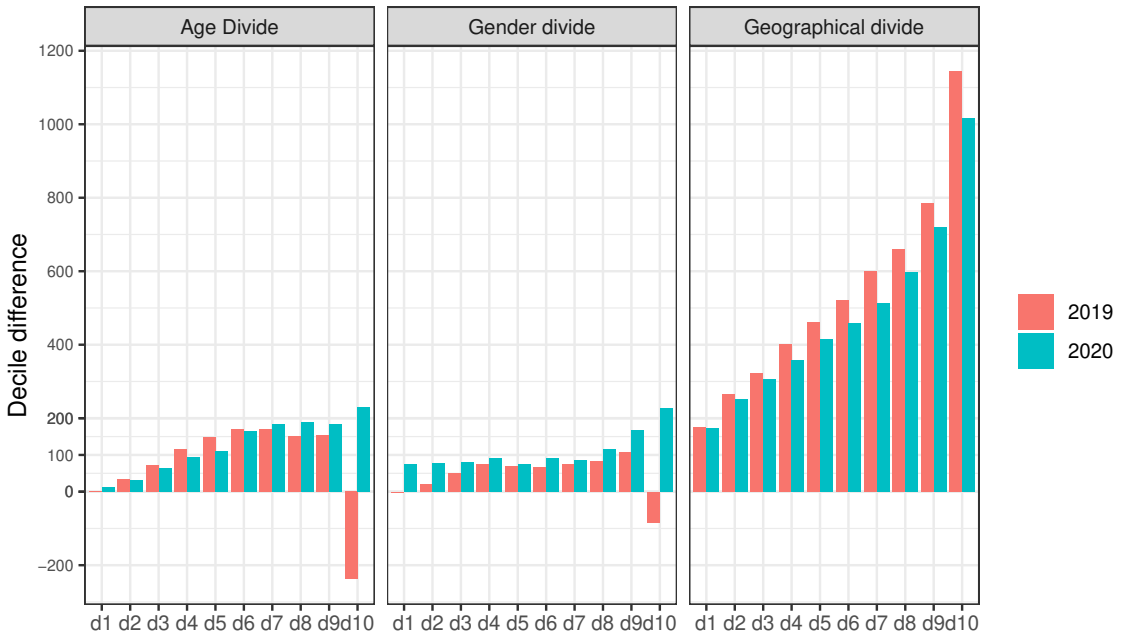


Figure 2.5.1: The difference between deciles of different subgroups. The three panels report the comparison of subgroup deciles (in 2019 and 2020) for three partitions, based on household head characteristics. The "Age Divide" panel compares "Young (aged less than 65) vs. Old (aged more than 64)"; the "Gender Divide" panel compares "Men vs. Women"; the "Geographical divide" panel compares "North and center vs. South and islands"

(Fig. 2.5.1). This is captured by a much smaller increase of d_{10}^{kh} (+5.3%). Are the means representative of the relative dynamic of the distributions?

The second panel explores the gender divide in households' consumption. In 2019 Male-headed households have higher consumption than Female-headed households; again, extreme deciles are an exception. For the same deciles, the gender divide increases a lot in 2020, while it slightly increases in the middle of the distribution. As before, $\Delta\mu_{kh}$ is more responsive (+149.2) than d_{10}^{kh} (+77.7) to fluctuations of richer households.

In the third panel, we investigate the territorial divide. As it is well known, Italian households in the Center and North of Italy are richer than those living elsewhere. The mean and the deciles of consumption of the two subgroups corroborate this fact and suggest that, at least for consumption, the geographical divide is the strongest among the three analysed. The impact of COVID-19 on household consumption is stronger for households in the Centre and the North and, differently from the other partition, it is quite symmetric across deciles. Therefore, the reduction of the distance between subgroups is measured at around 6% by both $\Delta\mu_{kh}$ and d_{10}^{kh}.

2.6 Results

2.6.1 Simulated shocks: a Monte Carlo experiment

As discussed in Section 2.4, alternative indicators of inequality between groups are likely to react in different ways to asymmetric (and stochastic) shocks. Starting from the 2019 data, we use a Monte Carlo experiment to study the behavior of Eq. (2.1)-(2.5) after different shocks. According to the probabilities ρ_{kq} as defined in Tab. 2.4.1, asymmetries produce shocks of heterogeneous intensity both between groups and between quantiles. For the sake of interpretability, we control the settings by modeling the probabilities as follows.

Regarding the source of asymmetry resulting from belonging to different groups, we define three scenarios (we refer to the partition between Group k and Group h for explanatory purposes). In the first one, Group k have a higher probability ($\rho^k = 75\%$) of being hit by the shock than Group h ($\rho^h = 25\%$). In the second one, the two groups have the same probability of 50%. In the last one, Group k has a lower probability ($\rho^k = 25\%$) of being hit than Group h ($\rho^h = 75\%$). Furthermore we introduce the asymmetry of the shock within the groups by defining $\rho_q = 0$ or 1 as the probability of being hit by a shock being in the decile q of the group.

Having ρ^k and ρ_q we are able to define $\rho_{kq} = \rho^k * \rho_q$.

The original consumptions of 2019 correspond to $\rho_q = 0$ for all $q = 1, \ldots, 10$. We start experimenting with $\rho_q = 1$ only for the first decile; then for $q = 1, 2$; until $\rho_q = 1$ for all $q = 1, \ldots, 10$. Then we start excluding the possibility of being hit ($\rho_q = 0$) for $q = 1$; then for $q = 1, 2$; until $\rho_q = 1$ only for $q = 10$.

We simulated all the settings both for additive and multiplicative shocks, setting $c = 200$ for the additive and $c = 20\%$ for the multiplicative case. We run each experiment 100 times and take the average of the 100 replicates of the five indicators.

We summarise the results for the additive positive shock in Fig. 2.6.1. For detailed results regarding multiplicative (encompassing both positive and negative aspects) and additive negative simulations, please refer to the appendix section of this chapter. For each setting and indicator, we express the average of the simulated results as relative to the indicator before the shock. If an indicator takes a value higher (lower) than 1, simulated inequality between groups is higher (lower) than the observed value. First, we notice that, as expected, the behaviour of $\Delta\mu_{kh}$ is the same of G_b^{BM}. It is also similar to that of G_b^{YL}, but less sensitive to changes in the scenario. On the other hand, d_{10}^{kh} and G_b^A accordingly plot a different picture. Where do these differences come from? What information do they provide?

To deal with these questions, we link the trajectories of Fig. 2.6.1 to the consumption gaps the groups (and their deciles) had in 2019 (Tab. 2.5.1 and Fig. 2.5.1). Let's focus for a while only on the gender and age divides, which show interesting similarities. In the first scenario (Probabilities=(0.75,0.25)), the probabilities of a positive shock are higher for Old and Women. In 2019, Old and Women are, respectively, poorer than Young and Men in all the deciles, but in the first and the tenth. All five indicators suggest a decreasing

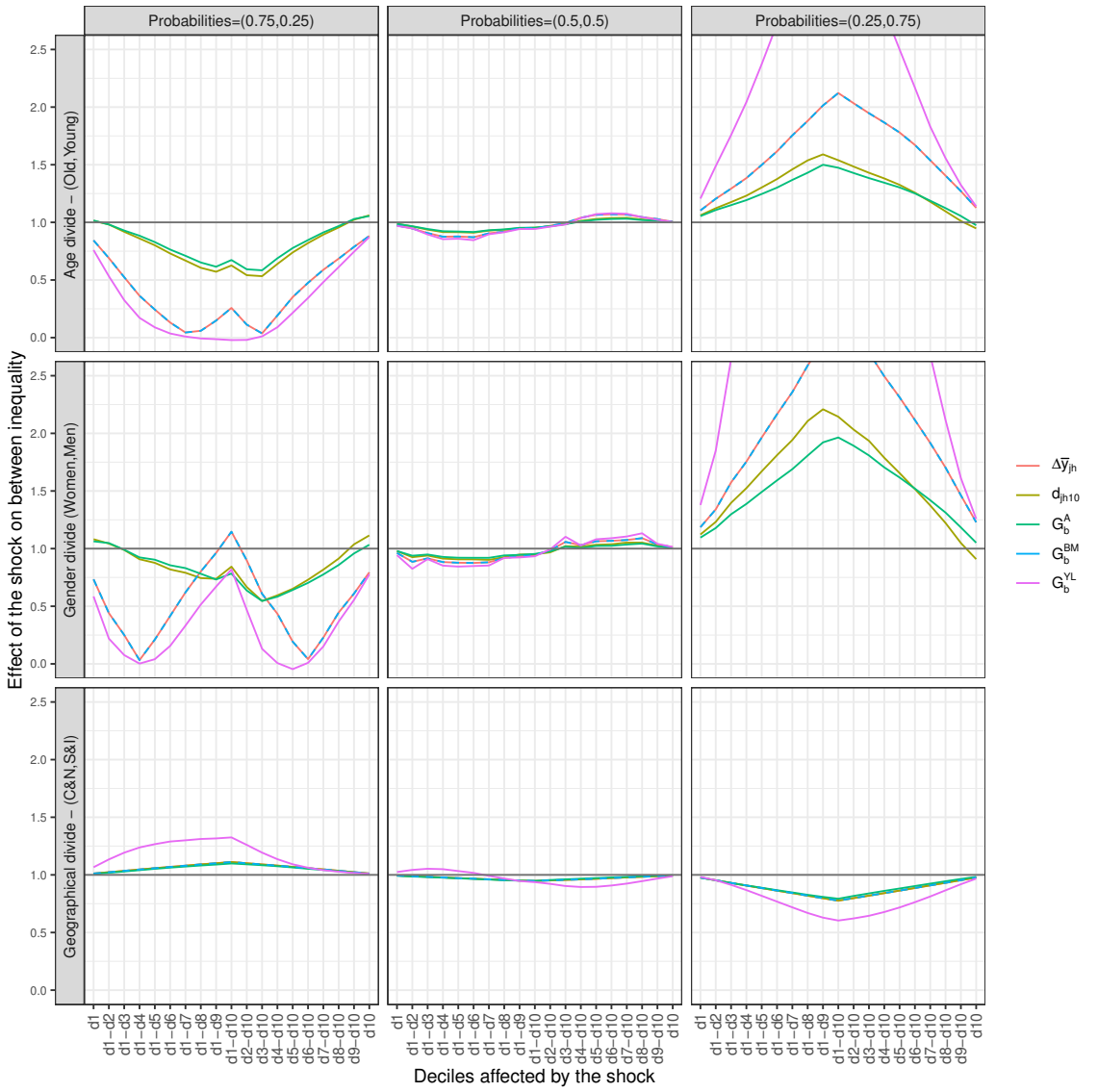


Figure 2.6.1: Summary of simulation results, additive shock. The effect of the shock on the inequality between groups, as expressed by the ratio between the mean of the simulated indicators and their values before the shock, by inequality factor and shock probability

between-group inequality while the shock propagates along the first deciles. However, when the shock only hits the first decile, d_{10}^{kh} and G_b^A increase with respect to their values before the shock. This is because the two first deciles of the two groups are very close before the shock, while they stray after it. More precisely,

before the shock, the average distance at the first decile is less than 20 for both partitions. After the shock, with the probabilities of the first scenario, c=200 produces an expected increase of 150 for Old and Women, and of 50 for Young and Men. The richest groups in the first decile are now Old and Women, and the absolute difference between the two first deciles strays from less than 20 to more than 80.

Measuring the gender divide, a similar argument explains why $\Delta\mu_{kh}$, G_b^{BM} and G_b^{YL} start to increase from $d1 - d5$, while d_{10}^{kh} and G_b^A continue to decrease. Let us focus on the grey bars in Fig. 2.5.1. After the simulated shock, most of the bars become shorter, and the mean of decile differences decreases. On the contrary, the three indicators based on the mean stop decreasing at d1-d4, because the grey bars become negative after the shock and compensate for the positive ones. At d1-d4 the means are approximately equal. However, involving more deciles in the shock shortens more bars and should imply a further decrease in between-group inequality, which is not captured by the three mean-based indicators.

As for the geographical divide, it is always quite stable, because the shock is low relative to the divide. However, the direction of the paths is the one expected, and the indicators are well correlated because, given the distributions before the shock, the shortening of the bars corresponds to the approach of the two averages.

All the trajectories of the other scenarios can be interpreted with the same approach, with the case of symmetric shock between groups leading to no drastic changes in the indicators.
As for the multiplicative case, we inform that the findings follow similar arguments. As natural, the divides are more reactive when extending the shocks to the top deciles.

Table 2.6.1: Observed value of $\Delta\mu_{kh}$, d_{10}^{kh}, G_b^A, G_b^{BM} and G_b^{YL} as measured on partitions - defined by age, gender and geographical factors - of the Italian households' monthly consumption in 2019 and 2020

	2019					**2020**				
	$\Delta\mu_{kh}$	d_{10}^{kh}	G_b^A	G_b^{BM}	G_b^{YL}	$\Delta\mu_{kh}$	d_{10}^{kh}	G_b^A	G_b^{BM}	G_b^{YL}
Age divide	.045	.071	.044	.033	.003	.075	.075	.047	.058	.005
Gender divide	.026	.036	.027	.019	.001	.064	.064	.041	.048	.003
Geogr. divide	.304	.304	.148	.218	.077	.286	.285	.142	.208	.069

2.6.2 A real shock: the COVID-19 impact

To compare the five indicators of between-inequality under analysis ($\Delta\mu_{kh}$, d_{10}^{kh}, G_b^A, G_b^{BM} and G_b^{YL}), Tab. 2.6.1 reports their values for the three bi-partitions under study, in 2019 and 2020 (G_b^A, G_b^{BM} and G_b^{YL}

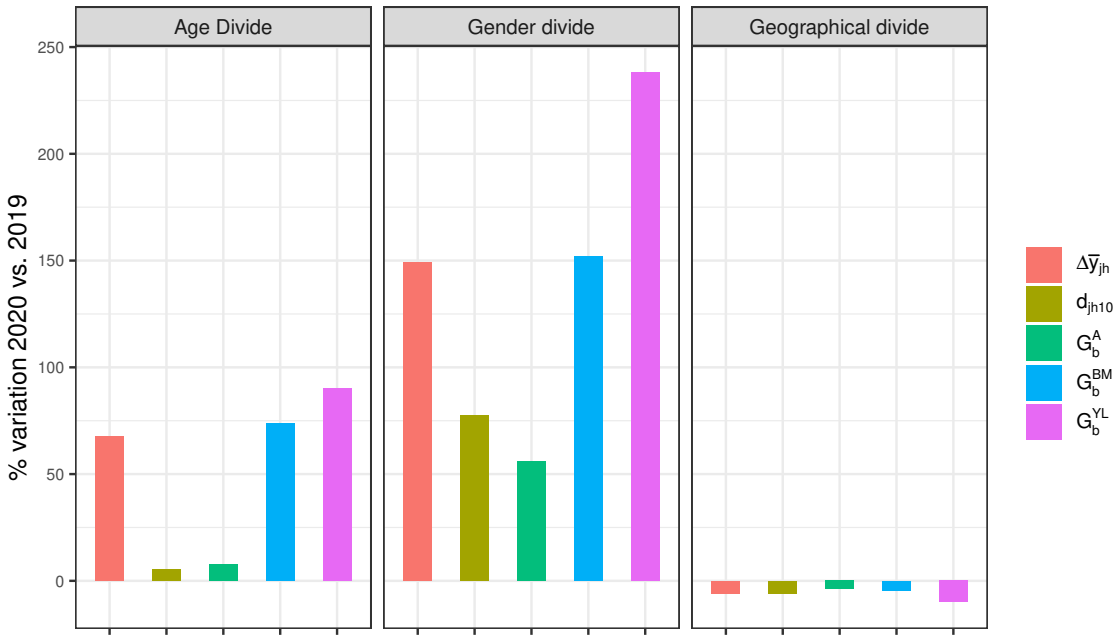


Figure 2.6.2: The three panels show the percentage variation (from 2019 to 2020) of the measures of between-group inequality discussed in Section 2.3 for three partitions, based on household head characteristics. The "Age Divide" panel compares "Young (aged less than 65) vs. Old (aged more than 64)"; the "Gender Divide" panel compares "Men vs. Women"; the "Geographical divide" panel compares "North and center vs. South and islands"

are reported as share of G to take into account inequality changes between the two years). We can observe really different patterns for the three divides analyzed. Geography stands out as the main inequality factors in both years and all indicators agree in providing this suggestion. The same agreement is observed with reference to the gender divide, which is the smallest of the three in 2019, but which increases strongly in 2020, reaching the importance of the age divide.

Comparison on a time basis is illustrated in Fig. 2.6.2, which shows the percentage variation, from 2019 to 2020, of the five indicators. They agree on the direction of the dynamic of the three divides. However, they differ on the intensity of the divides evolution but for the geographical divide. It is interesting to compare the pictures of the two baseline indicators, $\Delta\mu_{kh}$ and d_{10}^{kh}, with the evolution described by the between components from the alternative Gini index decompositions. As for the age divide, from 2019 to 2020, it increases by 67.7% according to the means difference, while just by 5.3% according to d_{10}^{kh}. While G_b^{BM} and G_b^{YL} are in agreement with $\Delta\mu_{kh}$ for a strong increase in the age divide, G_b^A increases slightly according to d_{10}^{kh}. To understand this difference we propose to look at Fig. 2.6.2 while taking Fig. 2.5.1 into account. The negative bar for the tenth decile in 2019 lowers the difference between the means of Young

Table 2.6.2: Bootstrap mean, 90% confidence interval and coefficient of variation of $\Delta\mu_{kh}$, d_{10}^{kh}, G_b^A, G_b^{BM} and G_b^{YL} as measured on partitions - defined by age, gender and geographical factors - of the Italian households' monthly consumption in 2019 and 2020. The bootstrap algorithm consists of 999 replicates. For each replicate, observations are sampled with replacement according to an i.i.d. bootstrap scheme accounting for survey weights

		2019					**2020**				
		$\Delta\mu_{kh}$	d_{10}^{kh}	G_b^A	G_b^{BM}	G_b^{YL}	$\Delta\mu_{kh}$	d_{10}^{kh}	G_b^A	G_b^{BM}	G_b^{YL}
	5%	.015	.054	.037	.012	.001	.050	.051	.035	.038	.002
Age	Mean	.046	.073	.046	.034	.003	.075	.075	.048	.058	.005
divide	95%	.074	.092	.057	.056	.006	.099	.099	.061	.076	.008
	C.V.	.386	.156	.132	.387	.544	.193	.189	.158	.193	.341
	5%	.003	.022	.017	.002	.000	.041	.042	.030	.031	.002
Gender	Mean	.027	.040	.029	.020	.001	.064	.064	.042	.048	.004
divide	95%	.056	.060	.041	.041	.003	.089	.088	.055	.067	.006
	C.V.	.616	.302	.254	.617	.951	.228	.224	.186	.227	.418
	5%	.278	.278	.138	.201	.065	.264	.264	.133	.192	.059
Geogr.	Mean	.305	.304	.149	.219	.077	.285	.285	.142	.208	.069
divide	95%	.331	.331	.160	.238	.090	.305	.305	.151	.222	.078
	C.V.	.053	.053	.046	.052	.097	.045	.046	.038	.044	.084

and Old, thus leading to an underestimation of the age divide by using the three mean-based indicators. The bar for the tenth decile becomes positive in 2020, overcoming the previous underestimation. This is why in 2020 G_b^{BM}, G_b^{YL} and $\Delta\mu_{kh}$ show (Fig.4) a substantial increase. Only G_b^A and d_{10}^{kh} properly represent the information in Fig. 2.5.1.

The picture does not depart when considering the gender partition. All the indicators signal a stronger increase in the divide, ranging from +56% to +238%. Again, the two measures not relying on the mean capture the asymmetric change of the distributions and report the weakest increase in the gender divide. As for the last partition, the geographical one, the agreement between $\Delta\mu_{kh}$ and d_{10}^{kh} on modest reduction of the geographical divide is confirmed by the between components of the Gini index decompositions. In this case, all the indicators do agree due to a symmetric change in the consumption distribution, leading to a quite homogeneous decrease of the divide across all the deciles.

To sum up, when deciles vary heterogeneously, the change of the difference between subgroup means

does not fully account for the evolution of the distance between the subgroup distributions. The latter is better captured by the average change of the differences between quantiles (d_{10}^{kh}), which is well represented by the evolution of Attili's between component.

With the aim of assessing the relevance of our results, and referring to statistical inference for inequality measures (see, e.g., Mills and Zandvakili [1997]), we develop a bootstrap procedure in order to obtain confidence intervals and variability measures for all relevant indicators. Each of the 999 replicates of the bootstrap is obtained by resampling observations from the original datasets according to an i.i.d. sampling with replacement, accounting for survey weights.

Tab. 2.6.2 reports the observed value, the 90% bootstrap confidence intervals and the coefficient of variation for $\Delta\mu_{kh}$ and d_{10}^{kh}, as well as for the between-inequality component of the three Gini index decompositions analyzed throughout the paper. The averages of the 999 replicates of all the indicators are close to the observed values both in 2019 and 2020, meaning that the bootstrap procedure is unbiased. The confidence intervals are quite symmetric around the observed value for all five indicators, so we can directly rely on variability measures to analyse the dispersion of their distributions. Our choice is to use the coefficient of variation to ensure comparability between five indicators ranging on different scales.

Comparing the partitions, both in 2019 and 2020 the dispersion of the indicators varies across partitions in inverse proportion to the intensity of the divide: the geographical divide is measured with the lowest uncertainty by all the indicators, while the dispersion of the indicators is the highest measuring the gender divide. This is because, when the divide is lower, the resampling is more likely to produce larger variations.

Comparing the two years, with few exceptions concerning d_{10}^{kh} and G_b^A, the indicators have higher variability in 2019. As for the comparison of the indicators, for all partitions and both years, G_b^{YL} is the between-inequality indicator with the highest variability over the 999 replicates, while G_b^A is the stablest. In the middle, G_b^{BM}, $\Delta\mu_{kh}$ and d_{10}^{kh} have intermediate values of variability, with G_b^{BM} and $\Delta\mu_{kh}$ reasonably showing similar dispersion.

Both the comparison of the two years and that of the indicators suggest that d_{10}^{kh} and G_b^A are more robust to extreme values, which are more frequent in 2019 and affect in a more decisive manner the indicators based on the means.

2.7 Conclusions

We analyze the effects of asymmetric shocks on the measurement of inequality, with a specific focus on the Gini index and on the components of its decompositions.

As a case study, we evaluate the COVID-19 impact on inequality with respect to three factors: age, gender and geography. Age and gender gaps increase, while territorial divide (which is confirmed as the most important inequality factor in Italy) slightly decreases.

The assessment of these dynamics strongly depends on the employed measure. If it is based on averages, the evaluation reflects an aggregate effect, which may be attributable to some aspects of the distribution only. This is the case of age and gender gaps in Italian household consumption: the average change is driven by the group of households consuming more.
When, on the other hand, the measure of divides refers to the whole distribution, we have a more complete representation of the intervening dynamics.

Our findings are supported by a Monte Carlo study, which allows us to assess the effects of a wide range of asymmetric shocks on different inequality measures, and by a bootstrap procedure, which allows us to infer the confidence intervals and evaluate the responsiveness of indicators to extreme values.

Among the future lines of research, we intend to assess the role of other aggregated comparison measures (useful references being Wasserstein distance and Kullback-Leibler divergence) and analyze other crises as well as the effects of the post-COVID recovery in 2021.

To conclude, inequality decomposition methods, which include both mean-based measures, such as Bhattacharya and Mahalanobis [1967b] and Yitzhaki and Lerman [1991b], and measures referring to the whole distribution, such as Attili [2021], represent an excellent choice for measuring the effects on inequality of symmetric and asymmetric shocks.

Appendix

2.A Additional Simulation Results

This appendix contains the results of the additional simulations conducted for the study. These include the outcomes for both multiplicative (positive and negative) and additive negative shocks.

Additive Negative Shock

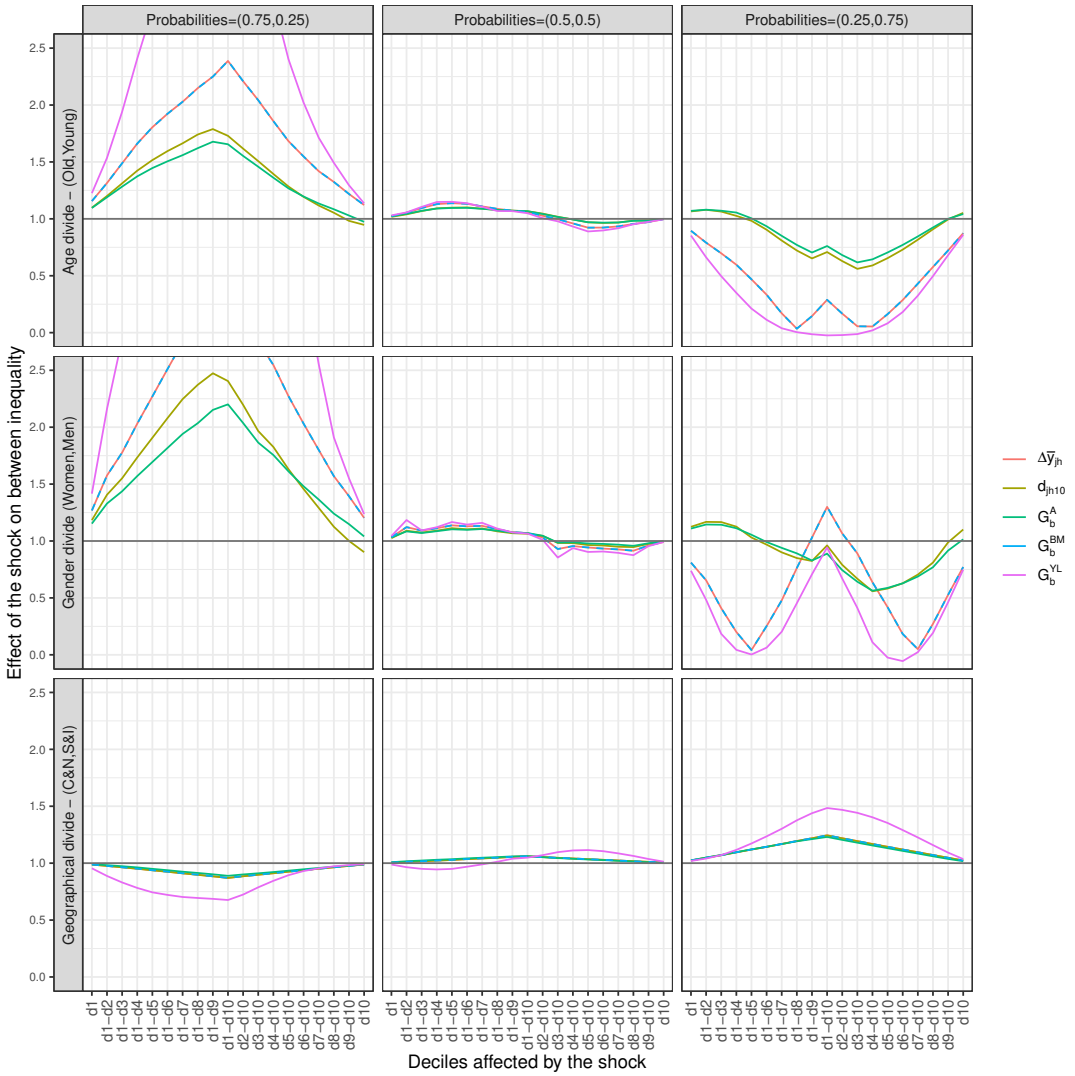


Figure 2.A.1: Results of the simulation with an additive negative shock.

Multiplicative Positive Shock

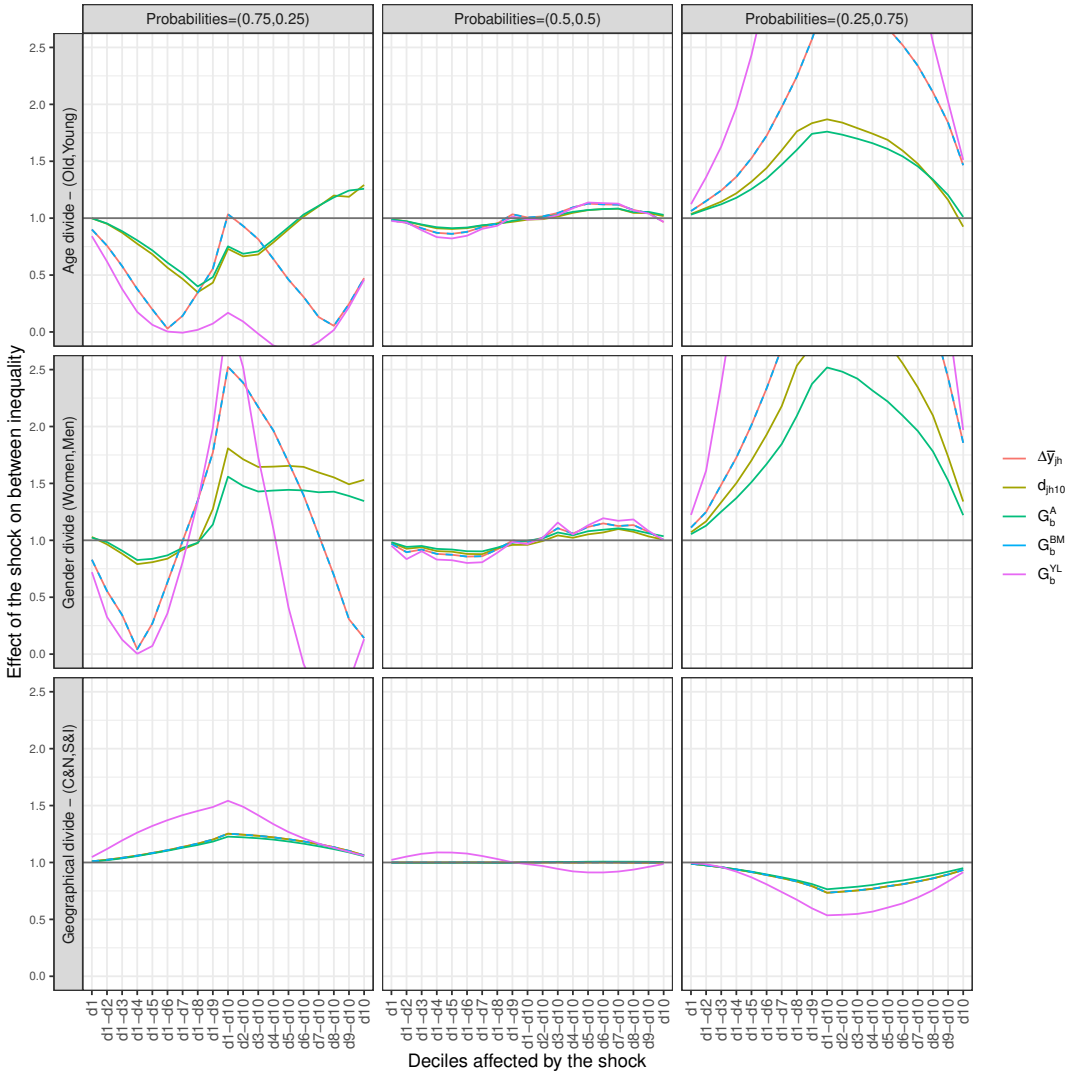


Figure 2.A.2: Results of the simulation with a multiplicative positive shock.

Multiplicative Negative Shock

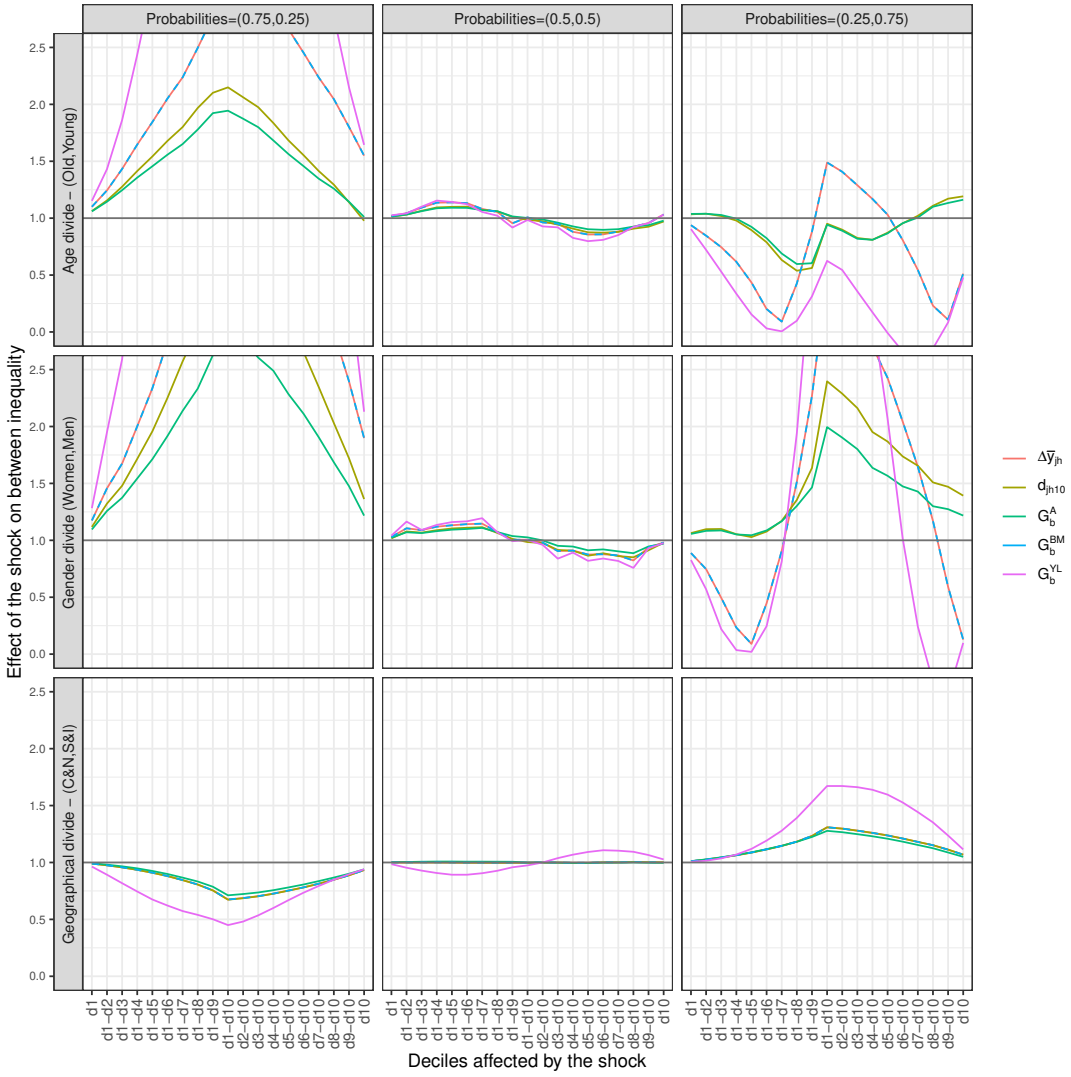


Figure 2.A.3: Results of the simulation with a multiplicative negative shock.

Chapter 3

Reassessing income instability with monthly data

Keywords: Income instability, Earnings dynamics, Squared Coefficient of Variation, Monte Carlo experiment

3.1 Preamble

In the first two chapters, we argued that simply comparing the average income of groups does not fully represent the level of inequality between them. Assessments based solely on averages risk underestimating this inequality or misrepresenting its evolution, especially in the face of significant economic changes, such as the Covid-19 pandemic or inflation. We suggested that a deeper dive into the distribution details of each group is crucial to accurately assess and understand the phenomena under study, particularly in the context of such impactful shocks.

Continuing with this line of reasoning, this chapter focuses on the assessment of income instability, which represents the intensity of short-term - unpredictable - income fluctuations, and posits that analysing individuals' annual income time series is insufficient for a comprehensive understanding of their instability. Particularly in periods of significant labor market volatility, or in the presence of labor markets with minimal regulation, annual earnings may conceal the extent of short-term income fluctuations, leading to an inaccurate comparison of the instability experienced by individuals or groups. We demonstrate how a shift to monthly income data can alter the assessment of instability and recommend a measure that is well-suited to handle the challenges introduced by using monthly data. This work aims to enhance awareness about the importance of data collection and the availability of detailed individual monthly data, and offers a tool for

its effective analysis.

This chapter requires a specific notation since it strongly relies on the time dimension and utilises data from the Survey of Income and Program Participation (SIPP).

The foundational steps of this research, which is coauthored with Emanuele Ciani, commenced during my research internship at the WISE Center of the OECD while developing the OECD [2023] report titled "On Shaky Ground? Income Instability and Economic Insecurity in Europe". I am especially grateful to Francesca Subioli for her significant and insightful contributions to this project. This work was presented in its advanced form at the XXXV conference of the Società Italiana di Economia Pubblica (SIEP) in Verona, 2023, and at the III PhD Workshop in Manciano, 2022. These presentations, along with the initial discussions at the Department of Economics in Bologna, have been instrumental in refining the research.

3.2 Introduction

Income instability consists of short-term income fluctuations arising for instance from job loss, changes in work intensity, or family disruption. Income instability matters for households' well-being as it is a trigger of consumption volatility and economic insecurity [Hardy and Ziliak, 2014, Hannagan and Morduch, 2015]. As such, it contributes to the disruption of family routines, stress, and hardship [Hill et al., 2017]. Income instability has also been found to be associated with mortgage delinquency and poor health outcomes, e.g., clinical depression, food insecurity, and worse educational, health, and behavioral outcomes for children [Leete and Bania, 2010, Avram et al., 2022]. High income instability might become particularly problematic when households are not fully and timely protected by social insurance, public, or private means.

Income instability is often analyzed through the lens of temporary and permanent variance in earnings. Temporary variance, in particular, signifies income instability, reflecting short-term fluctuations mainly due to unpredictable events. While much of the scholarly debate has centered on the methods to estimate this variance, specifically the choice between complex econometric methods and simpler descriptive ones, and on applying them to different contexts (see Jenkins, 2011 for a complete review), it is crucial to recognize the role of data frequency in accurately assessing the transitory variance of income, i.e., income instability. Restricting the attention to descriptive methods to avoid the sensitivity of the temporary variance estimates to arbitrary choices in the model specifications [Shin and Solon, 2011], this paper contributes to the debate by highlighting the advantages of monthly data in unveiling short-term income fluctuations that data with lower frequency may overlook.

Income instability is usually assessed by looking at annual income variations, whether as captured by the standard deviation of the arc-percentage changes over 2 years [Avram et al., 2022] or by the average variance of the logarithm of annual income over several years [Gottschalk and Moffitt, 2009]. However, annual income data may conceal short-term shocks, leading to underestimation of economic instability. For

instance, an individual suffering an income loss in the middle of the first year has zero annual income instability if his income recovers to the initial level in the middle of the second year. The analysis of annual income can, therefore, smooth out variations in income, thus failing to capture short periods of income decline.

Only a few papers analyse income instability looking at monthly income [Hills and Smithies, 2006, Bania et al., 2009, Edwards, 1999, Hannagan and Morduch, 2015]. Although these papers convincingly show that infra-annual instability matters, employing a range of different measures (e.g., average variance of income; measures of episodic poverty; number of income spikes or dips during the year), they lack a methodological discussion on how to best measure instability at the monthly level. In fact, the most popular measures of income instability are not appropriate to deal with monthly data. The standard deviation of the arc-percentage change, a very popular aggregate measure of income instability, is well-defined only for income changes across two periods. The average variance of the logarithm of income can be in principle used but, once the analysis is performed at the monthly level, it becomes problematic to account for zero incomes which are common with monthly data.

Our contribution to this literature is twofold. First, we provide a methodological discussion on how to best measure income instability with high-frequency data, suggesting the Square Coefficient of Variation as a measure of income instability. Second, by developing a Monte Carlo experiment, we study which conditions of the labour market or of income dynamics in general primarily require the use of infra-annual data. It comes out that the underestimation of income instability resulting from the utilization of annual data becomes more pronounced in cases where income dynamics are characterized by frequent negative shocks, such as during economic recessions. It is precisely in these periods that the accurate assessment of instability takes on greater significance.

There are three important requirements when shifting the attention to monthly data. The first is the need to distinguish between inter- and infra-annual instability - i.e., variability across years of annual average income and variability of monthly income within each calendar year. This is key both to assess the relevance of looking at a frequency higher than annual and because the two components help understanding if instability is driven by temporary or permanent shocks. The second is the need to isolate instability coming from seasonality in income patterns. If household members are employed as seasonal workers, e.g., in tourism, they may systematically report higher incomes in certain months, producing large variations in the year that are actually largely predictable and should be measured as a separate source of instability. The last requirement is the possibility to isolate income instability coming from individual trends of upward mobility.

We propose to assess monthly income instability by the squared coefficient of variation of monthly incomes (SCV). It possesses key connections with commonly employed indicators of annual income instability, overcoming their limitations when analysing monthly data. Thanks to the full decomposability of the squared coefficient of variation, our proposal allows to distinguish inter- and infra-annual instability and isolate plausible patterns of seasonality. Following Nichols [2008] and Subioli and Raitano [2021], it also

allows to understand if instability is driven by individual trends of upward mobility or by unstable conditions of the labour market. We also discuss an alternative but related method based on the decomposition of the Gini index developed in Chapter 1.

We finally complement our discussion with an empirical application on monthly income data from the SIPP. This confirms the relevance of measuring the infra-annual component. Thanks to the panel component in the survey, it is possible to estimate monthly income instability in the USA over a period of 48 months and to show how looking at monthly income instability can provide new and novel insights on this harmful economic phenomenon. By decomposing the squared coefficient of variation of monthly household incomes, we show that infra-annual instability accounts for 37% of total instability in the USA between 2014 and 2017. We also warn that the share of the infra-annual component is not constant across different populations, which may produce misleading ranking when using annual instability measures for cross-country or cross-population comparison. Comparing the evidence from the *SCV* methodology with that from the Gini index indicates that especially the most severe income shocks tend more than the others to be leveled out over the course of years. Monthly data are crucial for reporting them.

Next sections flow as follows. Section 3.3 stresses the importance of using infra-annual data to assess income instability and motivates our suggestion to employ the *SCV* to measure it. By a Monte Carlo experiment, Section 3.4 explains the economic conditions under which the utilization of monthly data becomes crucial to assess instability. Section 3.5 describes empirical evidence of the monthly income data from the SIPP. Section 3.6 concludes.

3.3 Measuring Income Instability

3.3.1 The Importance of Monthly Data

Monthly data are important to measure income instability because annual data may conceal infra-annual fluctuations. Table 3.3.1 illustrates a simple scheme with monthly income of two individuals (i and j) over two years. Both have constant annual total income (9000) so that instability measures based on annual income capture zero instability. However, due to infra-annual variations, the two individuals are far from being stable. Also, their levels of total instability should differ, with j being more stable than i in $t+1$.

How frequent are these kinds of patterns? Simple statistics from monthly income data from the 2014-2017 panel of the SIPP help in answering this question, suggesting that these kinds of patterns are not rare. Between 2014 and 2017, we estimate that 49% (27%) of households experienced a monthly income drop of at least 40% (60%) and lasting at least three months. At the same time, almost 7% (4%) of households experienced such a drop without reporting any annual income drop bigger than 10%. This confirms that for a significant share of the population, monthly income fluctuations are cancelled out at least partially when looking at annual incomes.

Table 3.3.1: Annual Earnings May Conceal Short-Term Fluctuations

Individual	Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
Yltirow2*i	t	1500	1500	1500	1500	1500	1500	0	0	0	0	0	0	9000
	$t+1$	0	0	0	0	0	0	1500	1500	1500	1500	1500	1500	9000
Yltirow2*j	t	1500	1500	1500	1500	1500	1500	0	0	0	0	0	0	9000
	$t+1$	0	0	0	1000	1000	1000	1000	1000	1000	1000	1000	1000	9000

Note: Monthly income of individuals i and j over two years

3.3.2 The Issues with Conventional Instability Measures

The most widespread measures of income instability are the standard deviation of the arc-percentage change and the average variance of the logarithm of income. Unfortunately, they are not appropriate to deal with monthly data.

The arc-percentage change of the individual i between two periods t and $t+1$ is defined as:

$$\Delta_{i;\,(t,\,t+1)} = \frac{Y_{i,t+1} - Y_{i,t}}{\frac{1}{2}(Y_{i,t+1} + Y_{i,t})} \tag{3.1}$$

where $Y_{i,t}$ is the individual income of period t. The arc-percentage change $\Delta_{i;\,(t,\,t+1)}$ measures income variation between t and $t+1$ as the change in value divided by the average of the initial and final values. This ensures that positive and negative changes are treated symmetrically, unlike the standard percentage change, which can yield different results when considering the change from t to $t+1$ versus $t+1$ to t. The arc-percentage change provides a consistent measure regardless of the direction of change and ranges from -2 to 2. The standard deviation (between N individuals) of the arc-percentage change between two periods t and $t+1$ [Avram et al., 2022] is defined as:

$$\sigma_{t,t+1}(\Delta) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \Delta^2_{i;\,(t,\,t+1)}} \tag{3.2}$$

$\sigma(\Delta)$ is only defined to measure instability over two periods, usually two consecutive years. This impedes studying instability with monthly data. Another limitation of $\sigma(\Delta)$ is that it is not an individual instability measure as it only provides population estimate of instability.

The second conventional measure we take into consideration is the variance of the logarithm of income, which is defined over multiple periods (we generally refer to the number of months T), so it can be used in principle with monthly data. Defining $\overline{\ln Y_i}$ as the average of the individual i logarithm of income ($\overline{\ln Y_i} =$

$\sum_{i=1}^{T} \ln Y_{i,t}$), the variance of the logarithm of income at the individual level reads:

$$\sigma_i^2 = \frac{1}{T} \sum_{t=1}^{T} \left(\ln Y_{i,t} - \overline{\ln Y_i} \right)^2 \tag{3.3}$$

The population estimate of instability is obtained by the average variance of the logarithm of income, namely by averaging eq. 3.3 over all individuals [Gottschalk and Moffitt, 2009]:

$$\sigma_{pop}^2 = \mathbb{E}(\sigma_i^2) = \frac{1}{N} \sum_{i=1}^{N} \sigma_i^2 \tag{3.4}$$

Unfortunately, once the analysis is performed at the monthly level, it becomes problematic to account for zero incomes - which are common with monthly data - due to the use of logarithm in Eq. 3.3. The presence of zero incomes leads to undefined values when taking logarithms, complicating the use of this formula for monthly data analysis. Moreover, another issue arises with the logarithmic transformation: it tends to diminish the significance of large income fluctuations. This dampening effect on significant income jumps is not desirable in our context, as capturing these fluctuations is crucial for accurately measuring income instability. Thus, the logarithmic approach, while useful in certain analyses, presents substantial limitations in our study of monthly income data. Using the average variance of income would instead introduce an important issue due to the lack of scale invariance, which is fundamental to measuring income instability since its assessment should be independent of the currency considered or the inflation trends.

3.3.3 The Squared Coefficient of Variation

In the presence of high-frequency data, we suggest measuring individual instability by the squared coefficient of variation (SCV) of individual incomes. It is defined as:

$$SCV_i = \frac{1}{T} \sum_{t=1}^{T} \left(\frac{Y_{i,t} - \overline{Y_i}}{\overline{Y_i}} \right)^2 \tag{3.5}$$

where $\overline{Y_i}$ is the mean of individual monthly income. If the population has size N, the population estimate of instability is obtained by averaging the individual SCV_i across the population:

$$SCV_{pop} = \mathbb{E}(SCV_i) = \frac{1}{N} \sum_{i=1}^{N} SCV_i = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{T} \sum_{t=1}^{T} \left(\frac{Y_{i,t} - \overline{Y_i}}{\overline{Y_i}} \right)^2 \tag{3.6}$$

The *SCV* as a measure of income instability is suggested because, in addition to the usual properties inherited from the coefficient of variation (including scale invariance), it possesses several important properties to describe monthly income dynamics.

3.3.4 Properties

Property 1: Individual Based Instability Measure By Eq. 3.5, the squared coefficient of variation provides individual instability estimates. Differently from measures directly defined at the population level (see, e.g., Eq. 3.2), this property allows for investigating instability at the micro-level. For example, one can look for population subgroups suffering more instability or regress instability over individual characteristics.

Property 2: Multiperiod Instability Measure The squared coefficient of variation allows for estimating income instability over a reference period with no limits on income observations. Unlike measures defined over two periods only (see, e.g., Eq. 3.2), this property enables the measurement of instability with monthly data over multiple years.

Property 3: Compatibility with Zeros in Data One of the significant advantages of the *SCV* is its compatibility with zero income values in the data. This is particularly relevant for monthly data, where zero incomes are not uncommon. The *SCV* can effectively handle these instances without the complications associated with logarithmic transformations.

Property 4: Consistency with Widespread Measures of Income Instability The *SCV* shows interesting similarity with other measures commonly employed in the literature to measure income instability. For instance, it relates to the variance of the transitory component σ^2_{pop}. Comparing SCV_{pop} to σ^2_{pop}, Eq. 3.4 averages the individual squared logarithmic differences from the mean, which is approximately equivalent to averaging the squared relative differences of Eq. 3.5. It then takes the average across all individuals, as in Eq. 3.6.

As for the standard deviation of the arc-percentage change $\sigma(\Delta)$, despite its limitations with monthly data, it is widely used in the literature to assess annual instability. Here we show a key relation between Eq. 3.2 and Eq. 3.6, following from a strong connection between the *SCV* and the arc-percentage change Δ. At the individual level, the squared coefficient of variation calculated over two consecutive annual incomes is $1/4$ of the squared arc percentage change Δ^2. At the individual level, in the case of only two periods, the relationship between the squared coefficient of variation and the squared arc percentage change can be mathematically expressed as:

$$\begin{aligned} SCV_i &= \frac{1}{T}\sum_{t=1}^{T=2}\left(\frac{Y_{i,t}-\overline{Y_i}}{\overline{Y_i}}\right)^2 = \\ &= \frac{1}{2}\left[\left(\frac{Y_{i,t}-Y_{i,t+1}}{Y_{i,t}+Y_{i,t+1}}\right)^2+\left(\frac{Y_{i,t+1}-Y_{i,t}}{Y_{i,t}+Y_{i,t+1}}\right)^2\right] = \\ &= \frac{Y^2_{i,t+1}+Y^2_{i,t}-2\cdot Y_{i,t}\cdot Y_{i,t+1}}{4\cdot\overline{Y_i}^2} = \frac{1}{4}\Delta^2_{i;\,(t,\,t+1)} \end{aligned} \tag{3.7}$$

It follows from the derivations that:

$$\mathbb{E}(\Delta^2) = 4 \times \mathbb{E}(SCV_i) = 4 \times SCV_{pop} \tag{3.8}$$

Furthermore, it is established that:

$$\sigma^2(\Delta) = \mathbb{E}(\Delta^2) - \mathbb{E}^2(\Delta) \approx 4 \times SCV_{pop} \tag{3.9}$$

where the approximation holds because $\mathbb{E}^2(\Delta)$ is generally small, even if $\mathbb{E}(\Delta) \neq 0$. Thanks to this relation, we argue that SCV_{pop} appears as the natural extension of $\sigma(\Delta)$ to the case of more than two periods.

Property 5: Decomposability in Infra- and Inter-Annual Components One of the main advantages of measuring instability by SCV_{pop} is its total decomposability into infra-annual and inter-annual components:

$$SCV_{pop} = SCV_{pop,\ infra} + SCV_{pop,\ inter} \tag{3.10}$$

This possibility arises at the individual level, writing the numerator in Eq. 3.5 as:

$$\sum_{t=1}^{T} (Y_{i,t} - Y_{i,\cdot,y})^2 = \sum_{y=1}^{Y} \sum_{m=1}^{M} (Y_{i,m,y} - Y_{i,\cdot,y})^2 + M \cdot \sum_{y=1}^{Y} \left(Y_{i,\cdot,y} - \overline{Y_i}\right)^2 \tag{3.11}$$

where M is the number of months in a year, $Y_{i,m,y}$ is the income of individual i in month m of year y, and $Y_{i,\cdot,y}$ is the average income of individual i in year y. To achieve this result, we first notice that the numerator of the variance of income for individual i can be calculated as $\frac{1}{T}\sum_{y=1}^{Y}\sum_{m=1}^{M}(Y_{i,m,y} - \bar{Y}_i)^2$ where Y is the total number of years and M is the number of months per year. We can add and subtract $Y_{i,\cdot,y}$ within the variance formula without changing the result:

$$(Y_{i,m,y} - \bar{Y}_i)^2 = ((Y_{i,m,y} - Y_{i,\cdot,y} + Y_{i,\cdot,y} - \bar{Y}_i))^2$$

Expanding the square and grouping terms:

$$(Y_{i,m,y} - \bar{Y}_i)^2 = (Y_{i,m,y} - Y_{i,\cdot,y})^2 + (Y_{i,\cdot,y} - \bar{Y}_i)^2 + \text{cross terms}$$

The remaining terms in the variance decomposition are:

1. **Infra-annual fluctuations:** $\sum_{y=1}^{Y}\sum_{m=1}^{M}(Y_{i,m,y} - Y_{i,\cdot,y})^2$
2. **Inter-annual fluctuations:** $\sum_{y=1}^{Y}\sum_{m=1}^{M}(Y_{i,\cdot,y} - \bar{Y}_i)^2$

Individual infra-annual instability arises from the first group of addenda, which compares the income of each month with the average of its year:

$$SCV_{i,\ infra} = \frac{1}{M \cdot Y} \sum_{y=1}^{Y} \sum_{m=1}^{M} \left(\frac{Y_{i,m,y} - Y_{i,\cdot,y}}{\overline{Y_i}} \right)^2 \tag{3.12}$$

Defining the instability within each calendar year y as:

$$SCV_{i,\ infra,\ y} = \frac{1}{M}\sum_{m=1}^{M}\left(\frac{Y_{i,m,y} - Y_{i,\cdot,y}}{Y_{i,\cdot,y}}\right)^2 \tag{3.13}$$

Eq. 3.12 comes as weighted sum of each $SCV_{i,\ infra,\ y}$, so that individual infra-annual instability arises summing up the contributions of individual instability within each calendar year.

Averaging Eq. 3.12 (Eq. 3.13) across the population provides the infra-annual component (within each calendar year) of instability in the population. We have:

$$SCV_{pop,\ infra} = \mathbb{E}(SCV_{i,\ infra}) = \frac{1}{N}\sum_{i=1}^{N} SCV_{i,\ infra} \tag{3.14}$$

and

$$SCV_{pop,\ infra,\ y} = \mathbb{E}(SCV_{i,\ infra,\ y}) = \frac{1}{N}\sum_{i=1}^{N} SCV_{i,\ infra,\ y} \tag{3.15}$$

With the same approach, income instability between years comes from averaging the individual inter-annual instability. The individual inter-annual instability, which compares individual yearly averages with the overall mean, is given by:

$$SCV_{i,\ inter} = \frac{1}{Y}\sum_{y=1}^{Y}\left(\frac{Y_{i,\cdot,y} - \overline{Y_i}}{\overline{Y_i}}\right)^2 \tag{3.16}$$

Therefore, the population estimate of inter-annual instability reads:

$$SCV_{pop,\ inter} = \mathbb{E}(SCV_{i,\ inter}) = \frac{1}{N}\sum_{i=1}^{N} SCV_{i,\ inter} \tag{3.17}$$

Property 6: Disentangle Seasonality Contribution in a Three-Component Decomposition By the same approach, adding and subtracting the overall mean $\overline{Y_i}$ and the mean of each month $Y_{i,m,\cdot}$ to the infra-annual component of instability $SCV_{pop,infra}$, it can be further decomposed to disentangle the contribution of seasonality. In this way, we isolate the impact of seasonal income fluctuations, obtaining:

$$\sum_{y=1}^{Y}\sum_{m=1}^{M}\left(Y_{i,m,y} - Y_{i,\cdot,y}\right)^2 = \sum_{y=1}^{Y}\sum_{m=1}^{M}\left(Y_{i,m,y} - Y_{i,\cdot,y} + \overline{Y_i} - Y_{i,m,\cdot}\right)^2 + M\cdot\sum_{y=1}^{Y}\left(Y_{i,m,\cdot} - \overline{Y_i}\right)^2$$

The terms in the first summation take each month-year income deviation from the year average and correct for the peculiarity of its month (i.e., the difference between the overall mean and the month average across years $Y_{i,m,\cdot}$); terms in the second summation compare each month average across years with the overall mean. Consequently, we define the infra-annual component of instability net of seasonality and the seasonal component. At the individual level, the former reads:

$$SCV_{i,\ net-infra} = \frac{1}{M\cdot Y}\sum_{y=1}^{Y}\sum_{m=1}^{M}\left(\frac{Y_{i,m,y} - Y_{i,\cdot,y} + \overline{Y_i} - Y_{i,m,\cdot}}{\overline{Y_i}}\right)^2 \tag{3.18}$$

The population estimate of infra-annual instability net of seasonality is calculated by averaging Eq. 3.18 across the population:

$$SCV_{pop,\ net-infra} = \frac{1}{N}\sum_{i=1}^{N} SCV_{i,\ net-infra} \tag{3.19}$$

Similarly, we define the contribution of seasonal variations to individual and population instability as:

$$SCV_{pop,\ seasonal} = \frac{1}{N}\sum_{i=1}^{N} SCV_{i,\ seasonal} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M}\sum_{y=1}^{Y}\left(\frac{Y_{i,m,\cdot} - \overline{Y_i}}{\overline{Y_i}}\right)^2 \tag{3.20}$$

To estimate seasonality, we assume that individual seasonal differences are persistent during the years of the reference period. Deviations from this hypothesis may lead to overestimation of the seasonal component. Also, the use of this methodology in short panels increases the seasonal component. However, we observed that four-year panels are long enough to provide a sufficiently reliable estimate of seasonality.

Summing up, the average squared coefficient of variation is decomposable as follows:

$$\begin{aligned} SCV_{pop} &= SCV_{pop,\ infra} + SCV_{pop,\ inter} = \\ &= SCV_{pop,\ net-infra} + SCV_{pop,\ seasonal} + SCV_{pop,\ inter} \end{aligned} \tag{3.21}$$

Property 7: Disentangle Upward Mobility from Instability Upward income mobility gives positive contribution to instability measures. Keeping fixed the individual average income, an increasing income path is not optimal if compared to a smoother path. Despite this, a decreasing or unpredictable income path is far worse. For this reason it is important to disentangle the contribution of upward mobility from total instability. For example, this is important to understand if income instability is due to instable conditions in the job market or to recovery after a crisis; or to investigate to what extent the usually higher instability of young people is driven by upward mobility. The squared coefficient of variation is compatible with the method proposed by Nichols [2008] and recently applied to Italian earnings by Subioli and Raitano [2021] to isolate the contribution of upward mobility from other components of income instability. They model individual monthly income as a linear time trend with stochastic variations around that trend:

$$Y_{i,t} = \alpha_i + \beta_i \widetilde{t} + e_{i,t} \tag{3.22}$$

where $\widetilde{t} = t - \frac{(T+1)}{2}$ is centered to ensure $\alpha_i = \overline{Y_i}$, and $\mathbb{E}(e_{i,t}|\widetilde{t}) = 0$, so that $e_{i,t}$ are deviations around the trend. Thanks to the decomposability of SCV_i, it can be shown that:

$$SCV_i = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\hat{\beta}_i\widetilde{t}}{\overline{Y_i}}\right)^2 + \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\hat{e}_i}{\overline{Y_i}}\right)^2 \tag{3.23}$$

The first component of Eq. 3.23 captures the instability due to the overall trend, and the residual $\hat{e}_i$ captures fluctuations around the trend. When the slope of the trend is positive ($\hat{\beta}_i > 0$), the contribution to

instability of the trend component can be interpreted as upward mobility. Conversely, a downward trend ($\hat{\beta}_i < 0$) indicates downward mobility. Following Subioli and Raitano [2021], we suggest considering downward mobility and instability around the trend as the "bad" component of instability, in contrast to the "good" source provided by upward mobility.

3.3.5 An Alternative Measure Based on the Gini Index

As every variance-based indicator, the squared coefficient of variation amplifies the contribution of high deviations from the mean. The standard deviation of the arc-percentage change Δ may produce a similar issue by giving higher importance to significant transitions. In contrast, the variance of the transitory component σ_i^2 does the opposite due to the use of logarithms. We propose considering the Gini index, a linear measure of income inequality [Mehran, 1976], to assess individual instability:

$$G_i = \sum_{t=1}^{T} \sum_{t'=1}^{T} \frac{|Y_{i,t} - Y_{i,t'}|}{2\overline{Y}_i T^2} \tag{3.24}$$

The Gini index is appropriate for measuring income instability, i.e., to gauge how income is equally distributed (stability) or concentrated (instability) across periods. The population estimate of the Gini index is defined as:

$$G_{pop} = \mathbb{E}(G_i) = \frac{1}{N} \sum_{i=1}^{N} G_i \tag{3.25}$$

Gini subgroup decompositions (Bhattacharya and Mahalanobis, 1967a;Yitzhaki and Lerman, 1991a) can be used to distinguish between infra- and inter-annual instability. This involves dividing the monthly incomes of each individual into M groups – each corresponding to a calendar year - of M observations. In this way the within and between groups components of the decompositions can be interpreted as infra- and inter-annual instability. However, these decompositions include a third component that has no interpretation in this context. Therefore, we prefer to employ the decomposition proposed in Chapter 1, which only has within and between components. At the individual level, employing the notation of this chapter, the within component is given by:

$$G_{i,\,w} = \frac{1}{2\overline{Y}_i T^2} \sum_{y=1}^{Y} \sum_{m=1}^{M} \sum_{m'=1}^{M} w^y_{mm'} \left| Y_{i,m,y} - Y_{i,m',y} \right| \tag{3.26}$$

Infra-annual instability is measured by comparing monthly incomes of the same year, with each absolute difference weighted by the factor $w^y_{mm'}$, which is extensively discussed in Chapter 1. The population infra-annual inequality is measured averaging Eq. 3.26 across the population:

$$G_{pop,\,w} = \mathbb{E}(G_{i,\,w}) = \frac{1}{N} \sum_{i=1}^{N} G_{i,\,w} \tag{3.27}$$

The individual and population between component for measuring inter-annual instability are defined as:

$$G_{i,\,b} = \frac{1}{2\overline{Y}_i T^2} \sum_{y=1}^{Y} \sum_{y'=1}^{Y} \sum_{m=1}^{M} w_m^{yy'} \left| Y_{i,m,y} - Y_{i,m,y'} \right| \tag{3.28}$$

and

$$G_{pop,\,b} = \mathbb{E}(G_{i,\,b}) = \frac{1}{N} \sum_{i=1}^{N} G_{i,\,b} \tag{3.29}$$

Inter-annual instability is measured by comparing same month income from different years, with each absolute difference weighted by the factor $w_m^{yy'}$. Again, for proper interpretation of these weights, we refer to Chapter 1. It is important to note that these weights preserve the information of the pairwise differences that they multiply and ensure the total Gini index to be the sum of the within and between components:

$$\begin{aligned} G_i &= G_{i,\,w} + G_{i,\,b} \\ G_{pop} &= G_{pop,\,w} + G_{pop,\,b} \end{aligned} \tag{3.30}$$

Unfortunately, in the case of the Gini index, there is no possibility to isolate a seasonality component. To elucidate the main differences between *SCV* and *G*, we revisit the example of Tab. 3.3.1. Tab. 3.3.2 compares the values of the two instability measures (and their components) for individuals i and j. Both indicators consistently agree that individual i exhibits higher total and infra-annual instability than individual j. Notably, the difference in instability between individuals i and j is more pronounced when using the *SCV*, which assigns more significance to the larger deviation from the mean experienced by individual i.
Another crucial difference between the two indicators is the interpretation of the inter-annual component. Unlike *SCV*, *G* does not compare the means of the groups but instead compares the distribution of monthly income between years. In the case of individual j, the two distributions differ even though they have the same mean, resulting in a positive inter-annual component.

Table 3.3.2: The *SCV* is More Sensitive to Big Deviations from the Individual Average Income
Note: Instability for individuals A and B as measured by the *SCV* and the Gini index

	Overall	Inter-annual	Infra-annual
SCV_i	1.00	0.00	1.00
SCV_j	0.67	0.00	0.67
G_i	0.50	0.00	0.50
G_j	0.44	0.13	0.31

Note: Instability for individuals i and j, as measured by the *SCV* and the Gini index

3.4 Annual Data and Instability Underestimation under Plausible Income Patterns: A Monte Carlo Experiment

How relevant is the underestimation of instability due to the lack of monthly data? Which labor market conditions increase the importance of using them? This section presents a Monte Carlo experiment to explore these questions. It examines the difference between annual and monthly income instability under various labor market conditions, aiming to identify those associated with higher underestimation of instability when using annual income data.

This experiment compares annual and monthly income instability as measured by Eq. 3.6 under different monthly income dynamics. We tested various specifications of income dynamics. The main results are robust to different specifications, so we present the simplest model to privilege interpretability. We consider a population of N individuals each having an initial income[1] $Y_{i,0} = 1000$. The earnings dynamic of each individual is generated by the following model:

$$Y_{i,t+1} = \max\left(Y_{i,t} + \varepsilon_{i,t}, 0\right) \tag{3.31}$$

where $\varepsilon_{i,\,t}$ represents the income shock of individual i at time t and is defined as:

$$\varepsilon_{i,t} = \delta_{i,t} \cdot s_{i,t} \cdot i_{i,t} \cdot \max\left(Y_{i,0} = 1000,\ Y_{i,t}\right) \quad \text{with} \begin{cases} \delta \sim \text{Bern}(p_\delta) \\ s \sim 1 - 2 \cdot \text{Bern}(p_s)\ t = 1, \ldots, T \\ i \sim \text{Unif}(0, I) \end{cases}$$

The dynamic of the model in Eq 3.31 is constrained by the $\max(\cdot)$ operator, ensuring that simulated incomes remain non-negative. The values $\delta_{i,t}$, $s_{i,t}$, and $i_{i,t}$ are realisations from the random variables δ, s and i. They establish the conditions in the labor market through their parameters. The parameter p_δ in the Bernoulli distribution of δ represents the probability of income shocks, which happen when $\delta = 1$. The realisation $s_{i,t}$ determines the direction of the shock and takes values in $\{$-1, 1$\}$ based on its parameter p_s controlling the probability of a negative shock. Finally, $i_{i,t}$ represents the intensity of the shock, uniformly distributed between 0 and I, which is the intensity of the largest possible shock. Each realization $i_{i,t}$ expresses the share of the shocks over the previous level of earnings $Y_{i,t-1}$ or over 0 if $Y_{i,t-1}$ is negative. Therefore, it determines the absolute variation in earnings after each shock. The $\max(\cdot)$ operator in the equation determining $\varepsilon_{i,t}$ ensures the possibility to exit from zero-earnings periods and preserves the proportionality of the shocks with respect to current income when it is higher than initial income.

To compare the *SCV* for annual and monthly data, we fix $I = 1$, the number of periods $T = 24$, and the population size $N = 10^4$. These parameters mainly affect the level and variability of instability but do not significantly affect the comparison we are interested in. The two parameters playing a crucial role in

[1] The distribution of $Y_{i,0}$ does not impact the results. Indeed, the *SCV* is defined at the individual level and is scale invariant.

comparing *SCV* for annual and monthly data are p_s and p_δ, which determines how frequent shocks are and what is the prevalence of negative over positive shocks. In this way, we control for the most relevant conditions of the labour market. Focusing on these two parameters, the MC experiment operates as shown in Fig. 3.4.1. First, the algorithm generates the earning dynamic of each individual. Second, it evaluates monthly and annual instability by Eq. 3.6. Then, it replicates Step I-II for 499 times. Finally, it varies the values of the parameters p_s and p_δ.

Fig. 3.4.2 presents three panels, each for each $p_s \in \{0.2; 0.5; 0.8\}$, showing, for $p_\delta \in [0; 0.1]$, annual and monthly instability as measured by Eq. 3.6. If $p_\delta = 0$, there are no shocks and consequently no instability. Then instability monotonically grows with p_δ. The monthly *SCV* grows faster, capturing all the variability of earnings dynamics. When the probability of negative shocks is low (left panel), the annual and monthly *SCV* are low and close. They increase and, more importantly, diverge faster when negative shocks are more frequent (higher p_s). The ratio between the monthly and annual lines is not constant, also because the uncertainty of monthly instability increases with p_s. This suggests that, assessing instability, the importance of using monthly data tu void underestimation of income instability increases precisely when the labor market instability is high (p_δ is high), and, even more, during recessions (high p_s).

3.5 Monthly Income Instability in the USA

In this section, we exploit monthly income data from the SIPP. We use a panel covering the period 2014-2017, composed of 4 annual waves. They provide non-missing monthly personal income over 48 months for 21,253 individuals covering 11,092 households. Our analysis is at the individual level, and for each individual we consider the equivalised household income. It is obtained by cumulating individual income of members in the same household and then equivalising by dividing by the square root of household size. Negative equivalised household incomes are replaced with zero, and we remove 12 households with constant

Figure 3.4.1: Scheme describing the algorithm behind the Monte Carlo experiment

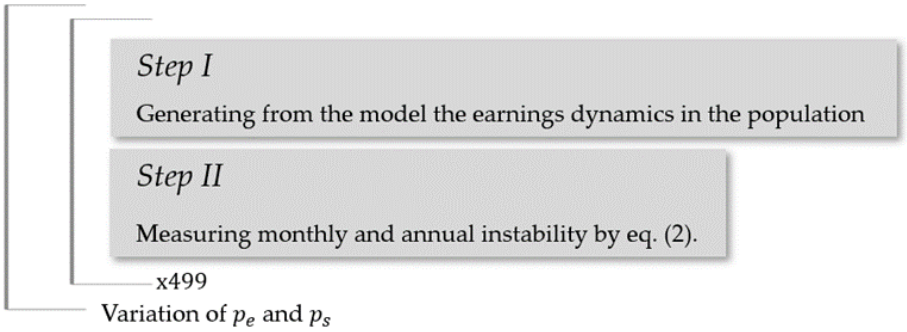
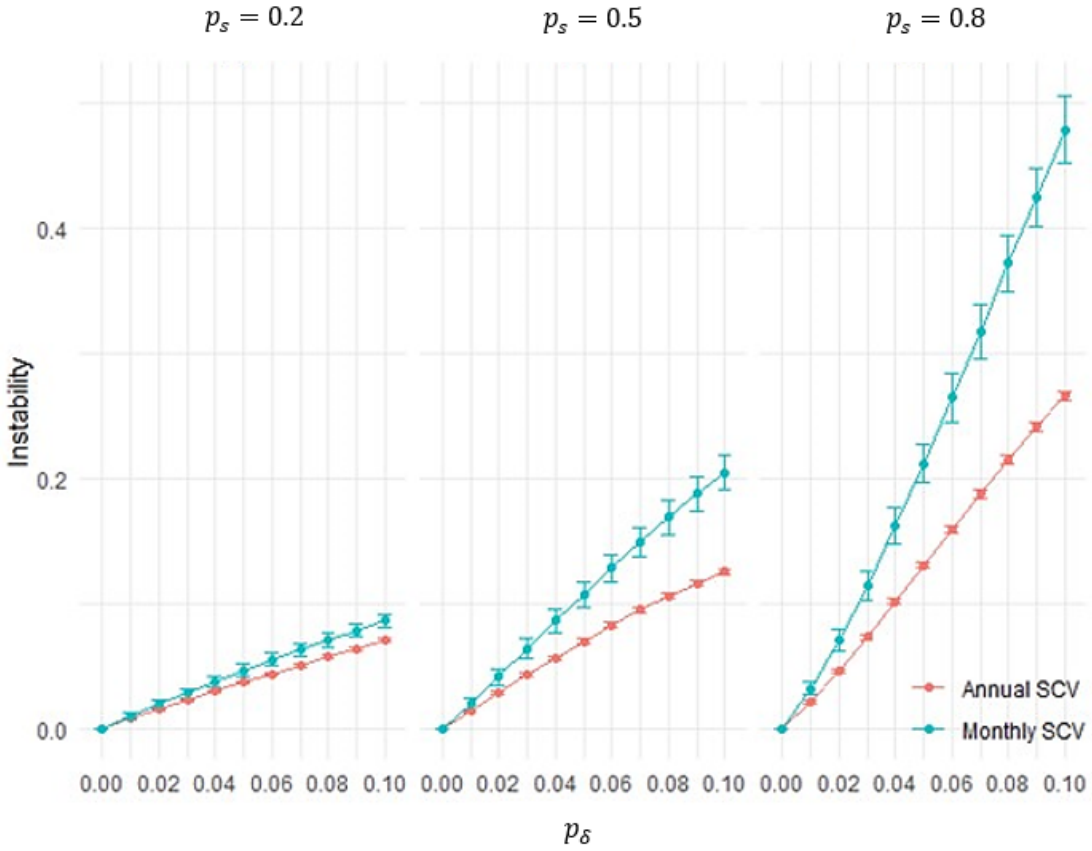


Figure 3.4.2: Using monthly data is crucial in presence of earnings dynamics with frequent negative shocks



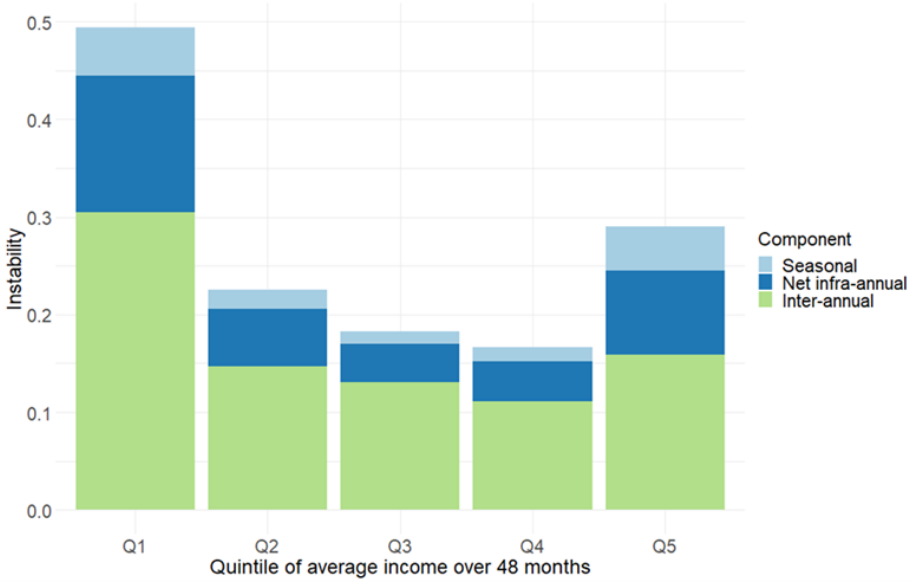
Note: Instability as measured by monthly and annual *SCV* for $p_\delta \in [0;0.1]$ and $p_s \in \{0.2;0.5;0.8\}$

zero-equivalised household income over 48 months.

Fig. 3.5.1 clusters the household based on their equivalised household income quintile and shows that infra-annual income instability matters particularly for poor individuals, i.e., those in the lowest quintile of average household equivalised income over 48 months. High instability also affects top quintile income individuals, while people in the middle of the distribution are more stable. Net infra-annual and seasonal components play a crucial role in comparing different groups. For example, inter-annual instability in the top quintile is 7.5% higher than in the second quintile, but almost 30% higher when considering overall monthly income instability.

The relevance of monthly data becomes even more evident when comparing individuals belonging to different racial groups. Fig. 3.5.2 clusters the household based on the race of the household head and shows

Figure 3.5.1: The Instability Gap for Low-Income Households Widens Once We Include Infra-Annual Instability



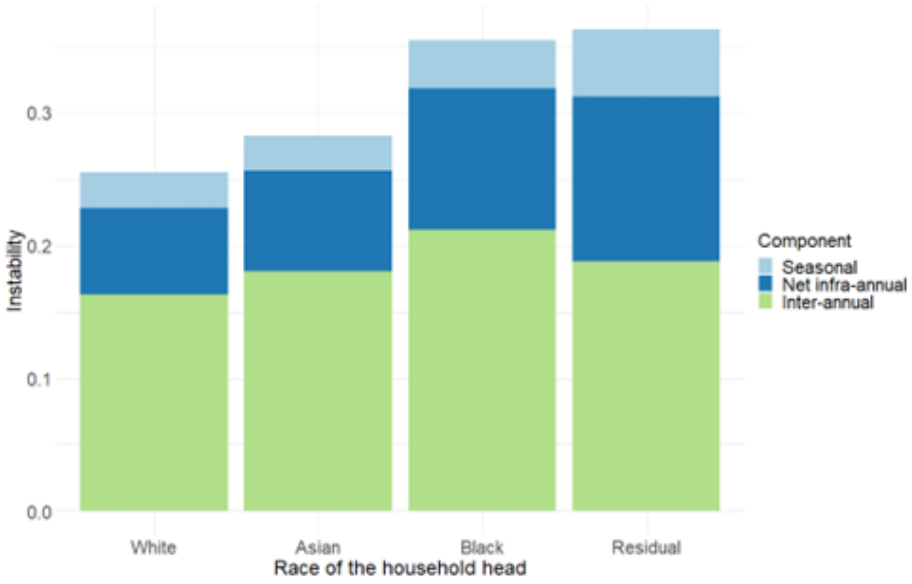
Note: Households are clustered by quintile of average income over 48 months. For each group, overall instability is decomposed into three components: inter-annual instability, infra-annual instability net of seasonality, and seasonality.

that Whites are unsurprisingly the most stable. The group with "Residual" races has the highest value of overall instability. However, excluding seasonality, it is second in the ranking and close to third if only considering inter-annual instability. Individuals with black household heads report the highest inter-annual instability, while Asians, standing out for second highest inter-annual instability, suffer relatively less of infra-annual shocks.

The two panels of Fig. 3.5.3 compare instability as measured by the *SCV* and the Gini index, focusing on the infra- and inter-annual components as measured by their decompositions. The grouping rule follows that of Fig. 3.5.1, with the right panel depicting the same plot as Fig. 3.5.1. The inter-annual component of the five groups shows similar values regardless of the selected measure. Only in $Q1$, it is higher for the *SCV*, informing that inter-annual shocks are relatively bigger in Q1. However, the infra-annual component is consistently higher for the *SCV* across all quintiles, particularly in $Q1$ and $Q5$. This indicates that the strongest shocks typically occur at the infra-annual level. This is an additional reason to avoid smoothing them out by looking at annual income variations.

Finally, considering the most recent panel (2018-2021), Fig. 3.5.4 examines the evolution of instability during the period 2014-2021. We compare the *SCV* of annual income between consecutive years with the annual contribution to infra-annual instability as defined in Eq. 3.13. The plot shows a discontinuity between 2017 and 2018 due to COVID-19-related difficulties in reaching out the survey partecipants of the 2018-2021

Figure 3.5.2: The Instability Gaps and Ranking Between Racial Groups Changes When Considering Infra-Annual Instability

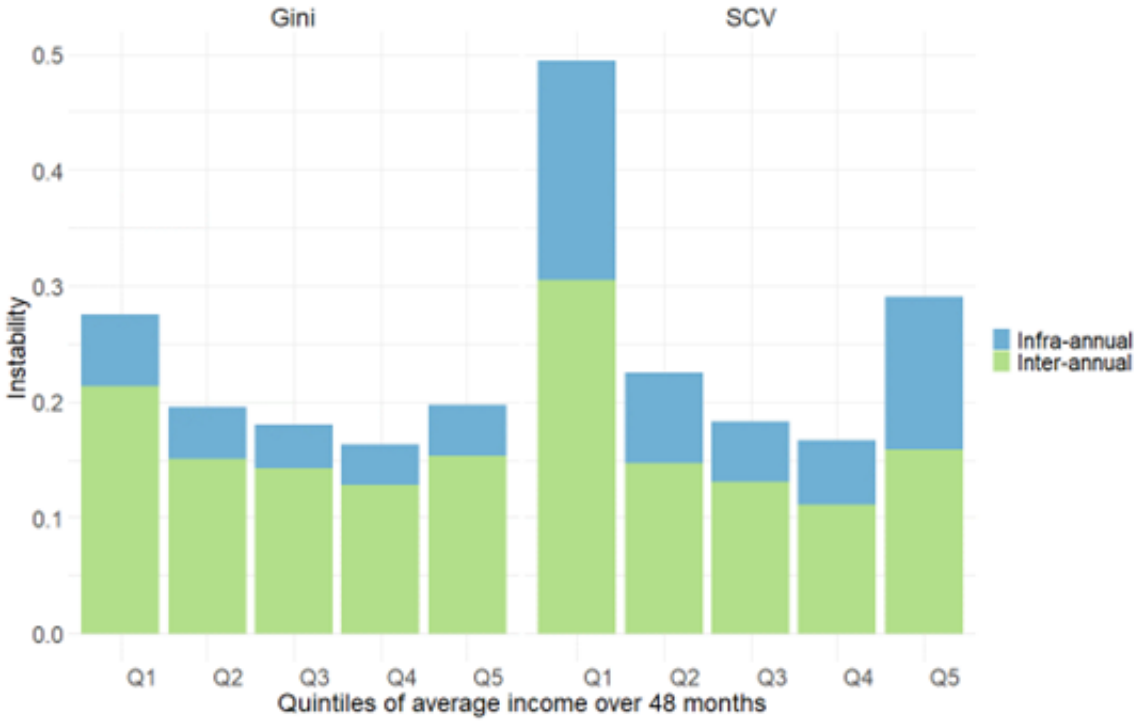


Note: Households are clustered by groups based on the race of the household head.

panel. For both the panels, the dynamics of the two measures are quite different, confirming the importance of infra-annual instability to study the evolution of instability.

This section concludes with findings on the contribution of upward mobility to income instability. For both the 2014-2017 and 2018-2021 panels, we evaluate the share of instability explained by upward mobility. We distinguish between young and old-headed households, using a 40 year-old cut-off on the age of the household head. Tab. 3.5.1 confirms that young-headed households generally suffer higher instability, but experience positive trends more frequently and exhibit higher shares of instability due to upward mobility. There is a slight increase in instability and the relevance of upward mobility for both age groups between the two panels. Caution is advised in considering this trend as reliable, and it is suggested to use these findings only for comparing young- and old-headed households.

Figure 3.5.3: The Largest Income Shocks Happen at Infra-Annual Level



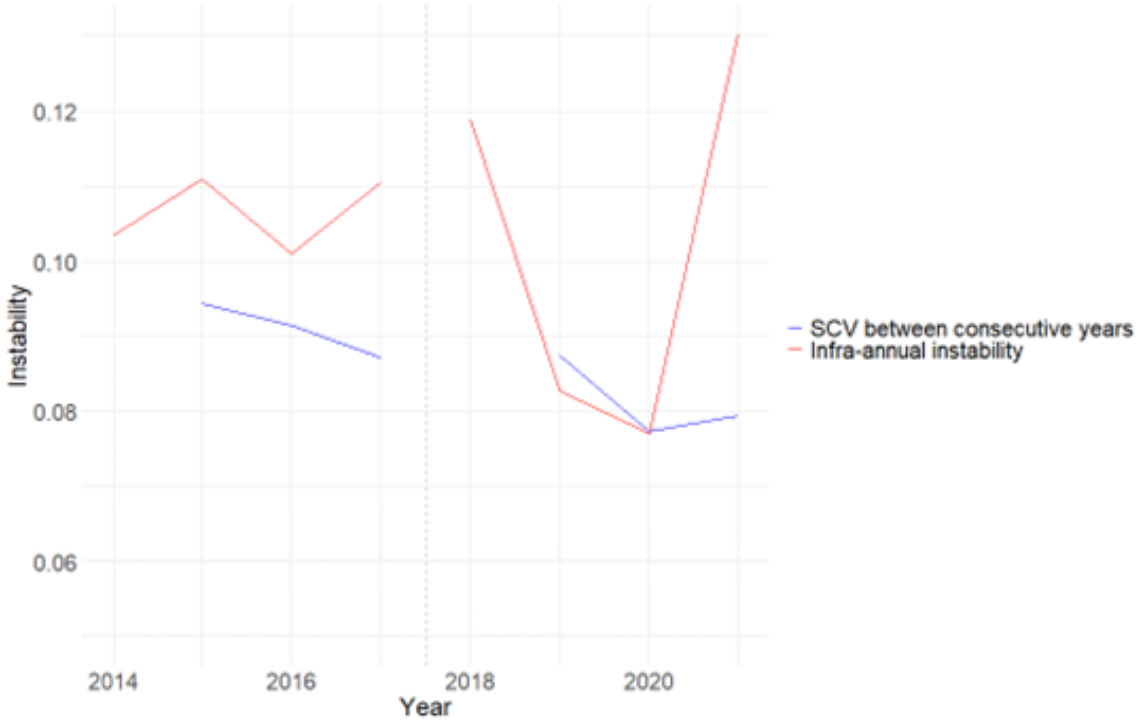
Note: Households are clustered by quintile of average income over 48 months. For each group, overall instability as measured by the Gini index (left) and *SCV* (right) is decomposed into two components: inter-annual and infra-annual instability.

Table 3.5.1: Instability and Upward Mobility in Young and Old-Headed Households

Panel	Age Group	E(SCV)	Upward Mobility Share (%)
2014-2017	Young	0.28	18.6
2014-2017	Old	0.24	14.5
2018-2021	Young	0.30	21.8
2018-2021	Old	0.25	19.4

Note: Average squared coefficient of variation and upward mobility contribution to it for young- and old-headed households in the two most recent panels, identified by a 40-year-old cut-off on the age of the household head.

Figure 3.5.4: Annual Income Instability Dynamic Differs from That of Infra-Annual Instability



Note: *SCV* between annual incomes of consecutive years (blue) and annual contribution to infra-annual instability (red) for the two panels 2014-2017 and 2018-2021.

3.6 Conclusions

This chapter focuses on the measurement of income instability, emphasizing the need to consider short-term income fluctuations that may be concealed when using annual income data. After motivating the importance of income data with a frequency higher than annual, the paper briefly discusses the issues of conventional measures of instability when data have a frequency higher than annual. Then, it develops a methodological discussion on how to measure income instability using monthly data. It suggests using the average squared coefficient of variation (*SCV*) as a measure of monthly income instability, highlighting several properties of the *SCV* that make it suitable for analyzing monthly income dynamics.

After discussing the importance of distinguishing between inter-annual and infra-annual instability, as well as the need to isolate instability caused by seasonal income patterns and individual trends of upward mobility, the paper demonstrates how the *SCV* allows for the decomposition of instability into these components, providing a more detailed understanding of income dynamics.

By conducting a Monte Carlo experiment, we confirm that instability measures underestimate instability when relying solely on annual income changes. We also show that underestimation is particularly significant during recessions. Especially in such periods, it is strongly suggested to seek for higher frequency data, such as monthly data, to effectively assess instability.

The work includes an empirical analysis of monthly income data from the U.S. Survey of Income and Program Participation, confirming the relevance of measuring infra-annual income instability. This analysis shows that infra-annual instability accounts for a significant share of total income instability in the USA between 2014 and 2017. Furthermore, it warns against using annual instability measures for cross-country or cross-population comparisons, as the share of the infra-annual component varies across different populations. The comparison of the proposed methodology with an alternative procedure based on the decomposition of the Gini index emphasizes the need to consider monthly data to accurately capture the most severe income shocks.

In conclusion, this paper highlights the limitations of relying solely on annual income data to measure income instability and suggests the use of the *SCV* to assess instability with monthly data. The empirical analysis supports the importance of considering infra-annual income fluctuations and demonstrates the value of the proposed measures in providing a more detailed picture of income instability. Overall, the paper contributes to the understanding of income dynamics and provides methodological guidance for policymakers and researchers studying income instability.

Bibliography

Paul D Allison. Measures of inequality. *American sociological review*, pages 865–880, 1978.

Federico Attili. Within-between decomposition of the Gini index: a novel proposal. *Quaderni-Working Paper DSE*, 2021.

Federico Attili. Uncovering complexities in horizontal inequality: A novel decomposition of the Gini index. *Social Indicators Research*, 2024. doi: https://doi.org/10.1007/s11205-024-03343-6.

Federico Attili and Michele Costa. Covid-19 impact assessment and inequality decomposition. In *Book of short papers SIS 2022*, volume 1, pages 1189–1094. Pearson, 2022.

Silvia Avram, Mike Brewer, Paul Fisher, and Laura Fumagalli. Household earnings and income volatility in the uk, 2009–2017. *The Journal of Economic Inequality*, 20(2):345–369, 2022.

Ripsy Bandourian, James McDonald, and Robert S Turley. A comparison of parametric models of income distribution across countries and over time. *Luxembourg income study working paper*, pages 1–47, 2002.

Neil Bania, Laura Leete, et al. Monthly household income volatility in the us, 1991/92 vs. 2002/03. *Economics Bulletin*, 29(3):2100–2112, 2009.

Brian Bell, Nicholas Bloom, and Jack Blundell. Income dynamics in the United Kingdom and the impact of the covid-19 recession. *Quantitative Economics*, 13(4):1849–1878, 2022.

Yonatan Berman and François Bourguignon. Evaluating the distributive incidence of growth using cross-sections and panels. *Review of Income and Wealth*, 2024.

Nath Bhattacharya and B Mahalanobis. Regional disparities in household consumption in India. *Journal of the American Statistical Association*, 62(317):143–161, 1967a.

Nath Bhattacharya and B Mahalanobis. Regional disparities in household consumption in India. *Journal of the American Statistical Association*, 62(317):143–161, 1967b.

Francois Bourguignon. Decomposable income inequality measures. *Econometrica: Journal of the Econometric Society*, pages 901–920, 1979.

François Bourguignon. Non-anonymous growth incidence curves, income mobility and social welfare dominance. *The Journal of Economic Inequality*, 9:605–627, 2011.

Carla Canelas and Rachel M Gisselquist. Horizontal inequality and data challenges. *Social Indicators Research*, 143(1):157–172, 2019.

Lars-Erik Cederman, Nils B Weidmann, and Kristian Skrede Gleditsch. Horizontal inequalities and ethnonationalist civil war: A global comparison. *American Political Science Review*, 105(3):478–495, 2011.

Lidia Ceriani and Paolo Verme. The origins of the Gini index: extracts from variabilità e mutabilità (1912) by corrado gini. *The Journal of Economic Inequality*, 10(3):421–443, 2012.

Lidia Ceriani and Paolo Verme. Individual diversity and the Gini decomposition. *Social Indicators Research*, 121(3):637–646, 2015.

Michele Costa. The gini index decomposition and the overlapping between population subgroups. *Gini Inequality Index: Methods and Applications*, pages 63–91, 2021.

Camilo Dagum. Inequality measures between income distributions with applications. *Econometrica (pre-1986)*, 48(7):1791, 1980.

Camilo Dagum. A new approach to the decomposition of the gini income inequality ratio. *Empirical Economics*, pages 515–531, 1997.

Joseph Deutsch and Jacques Silber. Inequality decomposition by population subgroups and the analysis of interdistributional inequality. *Handbook of income inequality measurement*, pages 363–403, 1999.

Udo Ebert. Measures of distance between income distributions. *Journal of Economic Theory*, 32(2):266–274, 1984.

Udo Ebert. The decomposition of inequality reconsidered: Weakly decomposable measures. *Mathematical Social Sciences*, 60(2):94–103, 2010.

Ashley N Edwards. Dynamics of economic well-being: Poverty, 2009–2011. *Journal of Human Resources*, 34:557–88, 1999.

Chris Elbers, Peter Lanjouw, Johan A Mistiaen, and Berk Özler. Reinterpreting between-group inequality. *The Journal of Economic Inequality*, 6(3):231–245, 2008.

Joan Esteban and Debraj Ray. Linking conflict to inequality and polarization. *American Economic Review*, 101(4):1345–1374, 2011.

Marta Fana, Sergio Torrejón Pérez, and Enrique Fernández-Macías. Employment impact of covid-19 crisis: from short term effects to long terms prospects. *Journal of Industrial and Business Economics*, 47(3): 391–410, 2020.

Ivan Gachet, Diego F Grijalva, Paúl A Ponce, and Damián Rodríguez. Vertical and horizontal inequality in ecuador: The lack of sustainability. *Social Indicators Research*, 145(3):861–900, 2019.

Giovanni Gallo and Michele Raitano. Sos incomes: Simulated effects of covid-19 and emergency benefits on individual and household income distribution in italy. Technical report, ECINEQ, Society for the Study of Economic Inequality, 2020.

Joseph L. Gastwirth. Is the gini index of inequality overly sensitive to changes in the middle of the income distribution? *Statistics and Public Policy*, 4(1):1–11, 2017.

Giovanni M Giorgi. The gini inequality index decomposition. an evolutionary study. *The measurement of individual well-being and group inequalities: Essays in memory of ZM Berrebi*, pages 185–218, 2011.

Giovanni M Giorgi et al. Gini's scientific work: an evergreen. *Metron*, 63(3):299–315, 2005.

Peter Gottschalk and Robert Moffitt. The rising instability of us earnings. *Journal of Economic Perspectives*, 23(4):3–24, 2009.

Anthony Hannagan and Jonathan Morduch. Income gains and month-to-month income volatility: Household evidence from the us financial diaries. *NYU Wagner research paper*, (2659883), 2015.

Bradley Hardy and James P Ziliak. Decomposing trends in income volatility: The "wild ride" at the top and bottom. *Economic Inquiry*, 52(1):459–476, 2014.

Shahedul Hasan, Md Amanul Islam, and Md Bodrud-Doza. Crisis perception and consumption pattern during covid-19: do demographic factors make differences? *Heliyon*, 7(5):e07141, 2021.

Vesa-Matti Heikkuri and Matthias Schief. Subgroup decomposition of the Gini coefficient: A new solution to an old problem. Technical report, working paper, 2022.

Heather D Hill, Jennifer Romich, Marybeth J Mattingly, Shomon Shamsuddin, and Hilary Wething. An introduction to household economic instability and social policy, 2017.

John Hills and Rachel Smithies. Tracking income: How working families incomes vary through the year. *LSE STICERD Research Paper No. CASEREPORT32*, 2006.

John D Huber and Laura Mayoral. Group inequality and the severity of civil conflict. *Journal of Economic Growth*, 24(1):1–41, 2019.

Rob J Hyndman and Yanan Fan. Sample quantiles in statistica packages. *The American Statistician*, 50(4): 361–365, 1996.

Xavier Jaravel. Inflation inequality: Measurement, causes, and policy implications. *Annual Review of Economics*, 13:599–629, 2021.

Stephen P Jenkins. *Changing fortunes: Income mobility and poverty dynamics in Britain*. OUP Oxford, 2011.

Stephen P Jenkins and Philippe Van Kerm. Assessing individual income growth. *Economica*, 83(332): 679–703, 2016.

I Josa and A Aguado. Measuring unidimensional inequality: Practical framework for the choice of an appropriate measure. *Social Indicators Research*, 149(2):541–570, 2020.

Peter J Lambert and J Richard Aronson. Inequality decomposition analysis and the Gini coefficient revisited. *The Economic Journal*, 103(420):1221–1227, 1993.

Peter J. Lambert and André Decoster. The gini coefficient reveals more. Technical report, Social Science Research Network, 2005.

Laura Leete and Neil Bania. The effect of income shocks on food insufficiency. *Review of Economics of the Household*, 8:505–526, 2010.

A Marques Santos, C Madrid, K Haegeman, and A Rainoldi. Behavioural changes in tourism in times of covid-19. *JRC121262*, 22(3):121–147, 2020.

Omar Shahabudin McDoom, Celia Reyes, Christian Mina, and Ronina Asis. Inequality between whom? patterns, trends, and implications of horizontal inequality in the philippines. *Social indicators research*, 145(3):923–942, 2019.

Farhad Mehran. Linear measures of income inequality. *Econometrica*, 44(4):805–809, 1976. ISSN 00129682, 14680262. URL `http://www.jstor.org/stable/1913446`.

Branko Milanovic. After the financial crisis: the evolution of the global income distribution between 2008 and 2013. *Review of Income and Wealth*, 68(1):43–73, 2022.

Jeffrey A Mills and Sourushe Zandvakili. Statistical inference via bootstrapping for measures of inequality. *Journal of Applied econometrics*, 12(2):133–150, 1997.

Jørgen Modalsli. Decomposing global inequality. *Review of Income and Wealth*, 63(3):445–463, 2017.

Domenico Moramarco. Fairness and Gini decomposition. *Economics Letters*, 233:111409, 2023.

Austin Nichols. Trends in income inequality, volatility and mobility risk. 2008.

OECD. *On Shaky Ground? Income Instability and Economic Insecurity in Europe*. 2023. doi: https://doi.org/https://doi.org/10.1787/9bffeba6-en. URL `https://www.oecd-ilibrary.org/content/publication/9bffeba6-en`.

Francesco Porro and Michele Zenga. Decomposition by subpopulations of the zenga-84 inequality curve and the related index ζ: an application to 2014 bank of italy survey. *Statistical Methods & Applications*, 29(1):187–207, 2020.

Sergio J Rey and Richard J Smith. A spatial decomposition of the gini coefficient. *Letters in Spatial and Resource Sciences*, 6(2):55–70, 2013.

Sandip Sarkar. Gini decomposition: An inequality of opportunity perspective. *Economics Letters*, 223: 110975, 2023.

Donggyun Shin and Gary Solon. Trends in men's earnings volatility: What does the panel study of income dynamics show? *Journal of public Economics*, 95(7-8):973–982, 2011.

Anthony F Shorrocks. The class of additively decomposable inequality measures. *Econometrica: Journal of the Econometric Society*, pages 613–625, 1980.

Anthony F Shorrocks. On the distance between income distributions. *Econometrica: Journal of the Econometric Society*, pages 1337–1339, 1982.

Stefanie Stantcheva. Inequalities in the times of a pandemic. Technical report, National Bureau of Economic Research, 2022.

Francesca Subioli and Michele Raitano. Persistent, mobile, or volatile? long-run trends of earnings dynamics in italy. 2021.

Shlomo Yitzhaki. Economic distance and overlapping of distributions. *Journal of Econometrics*, 61(1): 147–159, 1994.

Shlomo Yitzhaki and Robert I Lerman. Income stratification and income inequality. *Review of income and wealth*, 37(3):313–329, 1991a.

Shlomo Yitzhaki and Robert I Lerman. Income stratification and income inequality. *Review of income and wealth*, 37(3):313–329, 1991b.