Contemporary Approaches to Classical Econometrics: Measure Transport Applications in Multivariate Inequality Analysis and Fixed Effects

Jorge Andres Rivero

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2024

Reading Committee:

Yanqin Fan, Chair

Jing Tao

Rachel M. Heath

Program Authorized to Offer Degree:

Economics

University of Washington

Abstract

Contemporary Approaches to Classical Econometrics:
Measure Transport Applications in Multivariate Inequality Analysis and Fixed Effects

Jorge Andres Rivero

Chair of the Supervisory Committee:
Professor Yanqin Fan
Economics

This dissertation contributes to two major research areas: multivariate analysis of economic inequality and panel data methodology. The focus is on extending popular established approaches while retaining the features responsible for their enduring appeal. I achieve this by applying optimal transport theory directly to develop inequality orderings based on Lorenz curves, and indirectly to relax structure imposed in traditional fixed effects models. The first chapter offers a brief introduction on optimal transport.

In the second chapter, we propose a multivariate extension of the Lorenz curve based on multivariate rearrangements of optimal transport theory. We define a vector Lorenz map as the integral of the vector quantile map associated with a multivariate resource allocation. Each component of the Lorenz map is the cumulative share of each resource, as in the traditional univariate case. The pointwise ordering of such Lorenz maps defines a new multivariate majorization order, which is equivalent to preference by any social planner with inequality averse multivariate rank dependent social evaluation functional. We define a family of multi-attribute Gini index and complete ordering based on the Lorenz map. We propose the level sets of an Inverse Lorenz Function as a practical tool to visualize and compare inequality in two dimensions, and apply it to income-wealth inequality in the United States between 1989 and 2022.

In the third chapter, I extend the linear grouped fixed effects (GFE) panel model to allow for heteroskedasticity from a discrete latent group variable. Key features of GFE are preserved, such as individuals belonging to one of a finite number of groups and group membership is unrestricted and estimated. Ignoring group heteroskedasticity is shown to lead to poor classification, which causes significant finite-sample bias. I introduce the "weighted grouped fixed effects" (WGFE) estimator that minimizes a weighted average of group sum of squared residuals. I establish $\sqrt{NT}$-consistency and normality under a concept of group separation based on second moments. A test of group heteroskedasticity is proposed. A fast computation procedure is provided. Simulations show that WGFE outperforms alternatives that exclude second moment information. I demonstrate this approach by revisiting studies on the effect of unionization on earnings and the link between income and democratization.

In the fourth chapter, I reexamine the Rational Addiction model by introducing the type fixed effects (TFE) panel model. The TFE model incorporates heterogeneous coefficients and time-varying patterns of heterogeneity, which reflect differences in preferences and the addiction process. The model assumes the existence of a latent, time-invariant continuous variable referred to as a "type", which drives the heterogeneity in the parameters. Smoothness of the parameters as functions of the type is key to identification, allowing individuals of similar types to have similar parameter values. Correlation between the parameters, covariates, and instruments stem from type heterogeneity. I propose the type fixed effects generalized method of moments (TFE-GMM) estimator and establish consistency. I provide fast computation procedures based on a stochastic gradient descent algorithm. Simulations demonstrate good performance of this estimator. Using yearly household cigarette purchase data to estimate the model shows that most households follow cyclical consumption patterns and insensitivity to prices changes, giving support to educational interventions to curb smoking.

TABLE OF CONTENTS

Page

List of Figures v

List of Tables vii

Chapter 1: Introduction 1
1.1 Literature 6

Chapter 2: Lorenz map, inequality ordering and curves based on multidimensional rearrangements 7
2.1 Introduction 7
2.2 Vector Lorenz Map 11
2.2.1 Definition of the Lorenz map 11
2.2.2 Computation and examples 16
2.2.3 Relative scale and normalization 22
2.2.4 Properties and comparisons with other multivariate Lorenz concepts 23
2.3 Multi-attribute inequality comparisons 28
2.3.1 Lorenz dominance 28
2.3.2 Visualization of Lorenz dominance 35
2.3.3 Multivariate Gini inequality index 38
2.4 Empirical Illustration 43
2.4.1 Income-wealth α-Lorenz curves 43
2.4.2 Resource shares 45
2.4.3 Gini indices 45
2.5 Concluding Remarks 48

Chapter 3: Unobserved Grouped Heteroskedasticity and Fixed Effects 49
3.1 Introduction 49

3.1.1 Related Literature . . . 52
3.2 The GFE Model with Group Heteroskedasticity . . . 55
3.2.1 Identification of groups . . . 56
3.2.2 Estimation . . . 59
3.2.3 Computation . . . 60
3.2.4 Testing for group heteroskedasticity . . . 62
3.3 Asymptotic Theory . . . 63
3.3.1 Infeasible WGFE estimation . . . 64
3.3.2 Consistency of the WGFE estimator . . . 67
3.3.3 Consistency of group assignments . . . 68
3.3.4 Asymptotic Normality of the WGFE Estimator . . . 72
3.4 Simulations . . . 73
3.5 Empirical Applications . . . 75
3.5.1 The Effect of Unionization on Wages . . . 75
3.5.2 Income and Democracy . . . 79
3.6 Concluding Remarks . . . 84

Chapter 4: Type Fixed Effects and Rational Addiction: A GMM Framework for Latent Type Heterogeneity . . . 86
4.1 Introduction . . . 86
4.1.1 Related Literature . . . 90
4.2 GMM Framework with Type Heterogeneity . . . 93
4.2.1 Identification of type-specific parameters . . . 94
4.2.2 Identification of types . . . 97
4.2.3 The TFE-GMM Estimator . . . 101
4.2.4 Computation . . . 103
4.3 Asymptotic Theory: Consistency . . . 106
4.3.1 Sketch for consistency . . . 107
4.3.2 Consistency of type-specific parameters . . . 107
4.4 Simulation Evidence . . . 110
4.5 Rational Addiction with Type Fixed Effects . . . 114
4.5.1 Background . . . 115
4.5.2 Data . . . 117

4.5.3 Identification . . . 119
4.5.4 Results . . . 122
4.6 Concluding Remarks . . . 124

Bibliography . . . 126

Appendix A: Lorenz map, inequality ordering and curves based on multidimensional rearrangements . . . 146
A.1 User's implementation guide . . . 146
A.2 Additional Results: 3-D Inequality Analysis . . . 147
A.3 Vector quantiles as solutions to the semi-discrete transport problem . . . 148
A.3.1 Multiscale approach for minimizing L . . . 152
A.4 Specific features and issues with the data source . . . 153
A.5 Additional details and results . . . 157
A.5.1 Vector ranks and quantiles . . . 157
A.5.2 Egalitarian multi-attribute allocations . . . 158
A.5.3 Uniform Majorization . . . 167
A.6 Inequality Dominance based on the Inverse Lorenz Function . . . 169
A.7 Proofs of the main results . . . 170

Appendix B: Unobserved Grouped Heteroskedasticity and Fixed Effects . . . 177
B.1 Additional Results . . . 177
B.2 Variance Estimation . . . 185
B.3 Generalized Grouped Fixed Effects Estimation . . . 186
B.3.1 Computation . . . 186
B.3.2 Gradient of the criterion function with respect to assignments . . . 187
B.4 Computation . . . 188
B.4.1 A Variable Neighborhood Search (VNS) algorithm . . . 188
B.4.2 Initialization . . . 189
B.5 Proofs of the main results . . . 189
B.5.1 Proof of Theorem 1 . . . 189
B.5.2 Proof of Asymptotic Normality of the infeasible WGFE estimator . 194
B.5.3 Proof of Theorem 3 (Consistency of the WGFE Estimator) . . . 197
B.5.4 Proof of Theorem 4 (Consistency of group assignments) . . . 207

Appendix C: Type Fixed Effects and Rational Addiction: A GMM Framework for Latent Type Heterogeneity 228
C.1 Plots & Additional Results 228
C.2 Proofs for the main results 231
C.2.1 Proof of Theorem 5 231
C.2.2 Proof of consistency of $\left(\widehat{\theta}, \widehat{\alpha}\right)$ in endogenous linear model (4.6) . . . 232

LIST OF FIGURES

Figure Number — Page

1.1 Visualization of the Monge map between two densities 2
1.2 Total Variation and Kantorovich distance comparison 3
1.3 Comparison of metrics between uniform densities of disjoint support . . . 3
1.4 Arithmetic average versus the barycenter 4

2.1 Visualization of vector quantile and Lorenz map 20
2.2 Effects of correlation on α-Lorenz curves 37
2.3 Effects of correlation and marginal dispersion on α-Lorenz curves 38
2.4 Multivariate Lorenz ordering does not imply marginal Lorenz dominance . 39
2.5 Visualization of US income-wealth inequality across decades 44
2.6 Resource shares measured by the Lorenz map for US between 1989-2022 . 46
2.7 Gini indices for income-wealth inequality in US between 1989-2022 47
2.8 Multivariate Gini indices for inequality of racial and age groups 47

3.1 Example of groups that satisfy strong separability 58
3.2 Simple case of WGFE misclassification as T grows large 70
3.3 Median incomes across estimated groups 77
3.4 Proportion of workers under union contract 2001-2019 78
3.5 Group means of proportion/log earnings pairs 79
3.6 GFE time effects, within-group average incomes and democracy ($G = 4$) . . 82
3.7 Diagram of countries in groups determined by WGFE and GFE 85

4.1 Sketch of consistency of TFE-GMM estimator of types 108
4.2 Simulation results using common specifications: type fixed effects 112
4.3 Simulation results using unusual specification 115
4.4 Example of saddle-point equilibrium 117
4.5 Population density of consumer panel respondents 119
4.6 Aggregate median packs purchased and average unit price paid 120

4.7 By age cohort: median packs purchased and average unit price paid 120
4.8 By individual: packs purchased and average unit price paid 121
4.9 Diagram illustrating identification of heterogeneous effects 123
4.10 Main empirical result: majority of sample following cyclical behavior . . . 124
4.11 Main empirical result showing low price sensitivity. 125

A.1 Three dimensional Gini of income, wealth, and consumption 150

B.1 WGFE and GFE estimates of group sizes and variances for $G = 2, 3, 5$. . . . 178
B.2 GFE time effects, within-group average incomes and democracy for $G = 2$ 179
B.3 GFE time effects, within-group average incomes and democracy for $G = 3$ 180
B.4 GFE time effects, within-group average incomes and democracy for $G = 5$ 181
B.5 $G = 5$ map with Top: WGFE groups, and Bottom: GFE groups 184

C.1 Histograms of TFE-GMM estimates 229
C.2 Showing consumption paths are stationary AR(2). 230

LIST OF TABLES

Table Number Page

3.1 Simulation specification of latent variable's variance and mass function . . 74
3.2 RMSE/Misclassification rate for WGFE and GFE estimators 74
3.3 Descriptive statistics for PSID data set 77
3.4 WGFE estimates for $G = 2, \ldots, 7$ of democracy-income panel model 81
3.5 WGFE and GFE estimates of group sizes and variances 83

4.1 Bias of mean and variance of the heterogeneous effects 113
4.2 Features of the RMSE distribution of the heterogeneous effects 114
4.3 Descriptive statistics of the consumer panel 118

A.1 3-D, 2-D, 1-D Gini indices of income-wealth-consumption (1989-2016) . . . 151

B.1 Misclassification rates for WGFE and GFE estimators 177
B.2 List of countries in each group determined by WGFE assignment ($G = 4$). . 182
B.3 List of countries in each group determined by GFE assignment ($G = 4$). . . 183

C.1 Simulation results of TFE-GMM estimation with $h = 0.15$ 228
C.2 Simulation results of TFE-GMM estimation with $h = 0.01$ 230

ACKNOWLEDGEMENTS

I wish to express sincere gratitude to Yanqin Fan for the time, patience, and guidance she readily provided throughout my studies. She has had a profound influence on the researcher I have become and the one I look forward to being in the future. I'd also like to thank Jing Tao, Rachel Heath and Melissa Knox for providing invaluable advice for my research and listening to all of my ideas. From the other side of campus, many thanks to Soumik Pal for spurring my curiosity in mathematics and for being an incredible member of the PNW optimal transport community. To my coauthors Marc Henry and Brendan Pass, thank you for being excellent collaborators and mentors.

I would also like to thank Alfonso Rodriguez at Florida International University for convincing me to pursue a PhD and for being a constant voice of support and encouragement. From the math department, Laura De Carli encouraged me to pursue my MS in mathematics and I thank her for her generosity, kindness, and support for my career. I must also thank Pierluigi Vellucci for being an amazing coauthor and for his support throughout the PhD. I'd also like to thank: Julian Edward for kick-starting my teaching career in the Learning Assistant program; and Yu-chin Chen at the UW for her help as the graduate program director, for providing a great example of a well-taught course, and always being open to having a conversation about my research.

The staff at the UW have made life easier as a graduate student, starting with Simon Reeve-Parker in my first few years and ending with Heidi Hannah who meticulously approached my funding concerns, progress in the program, and job applications. I appreciate all that they do. I am also grateful for the people at the Office of Graduate Student Equity & Excellence (GSEE), who provided the financial support to write this dissertation.

To my friends I have made at the UW, thank you. The PhD was fun because of: Sean Ewen, Yigit Okar, Resem Makan, Raj Datta, John Kim, Seungryul Jeong, Amre

Abkem, Ryan Cummings, Dadmehr Didgar, Wendao Xue, Aochun Di, Abby Schamp, Reina Kawai, Kovid Puria, Yvonne Ng, Yoon Choi, Alejandro Gonzalez, Danial Salman, and Hyeonseok Park. Along with my friends back home: Rafael Badui, Lazaro Diaz, and Kola Oluwo– thank you for your support and encouragement throughout all these years. Special thanks to Lenox Li for all the great suggestions to improve my writing.

Thank you to Lily Engel and Sean Ewen for being the best of friends throughout the challenging and enjoyable times in Seattle. To my cousins in the Leon family, thank you for all your love and support. Moving to and living in Seattle was made better because of you. Thank you to the Henshaw family for your support and encouragement. Thank you to my dearest friend Charles Brown whose friendship has made all the difference in my life and my career.

I wish to express my deepest gratitude for the unconditional love and support from my grandparents and my parents. Their faith in me has kept me motivated to improve. To my brothers and sisters, I appreciate you. You are all so inspiring.

To my partner Katie, your patience and devotion got me through the PhD program. As one of many examples, you introduced me to our dear cats Julep and Mocha, who provided their company whether I asked for it or not. Your love and care made it possible for me to produce this dissertation and all the ideas herein. Thank you.

In Loving Memory of

Marcelo Rodriguez

Rosendo S. Rivero

Octavio Rivero

Chapter 1
INTRODUCTION

The analysis of economic inequality and the challenges of unobserved confounders are classical topics with many remaining open problems. For example, Lorenz curves are a popular tool to easily quantify and visualize inequality of a single resource, but unsuited for multivariate analysis due to the absence of a canonical ordering between bundles of resources. This is a major drawback since overall well-being of individuals should be measured jointly across several dimensions as argued by Stiglitz et al. [2009][1]. Another example, the wide availability of panel data and the use of fixed effects models are a powerful combination to control for unobserved heterogeneity, but rely on the assumption that these confounders can be captured by lower dimensional objects, such as a time-invariant, individual fixed effect[2]. In this dissertation, I'll revisit these classical approaches under the scope of measure (optimal) transport theory.

Optimal transport theory revolves around the problem of costly transport of mass from one distribution to another. The solution to this problem provides two powerful objects that have found to be useful in many applied fields. One is the transport map (or plan), which optimally associates points (and their mass) from the supports of each distributions; see Figure 1.1. This object conveys information on how one distribution relates to another based on a cost function that is chosen to reflect key properties. For example, given a user specified reference distribution and under standard Euclidean cost, the transport map between this reference and the data can be viewed as a multivariate

[1]The Lorenz curve was first proposed in Lorenz [1905].

[2]The connection between fixed effects and ANOVA predates its use in econometrics (Fisher [1925]).

quantile function or *vector quantile* where the rankings depend crucially on the chosen reference (e.g., uniform on the unit square/sphere). This is justified by the fact that the vector quantile is the *unique* gradient of some convex function which, like the univariate quantile, must satisfy some notion of monotonicity in ranks[3].

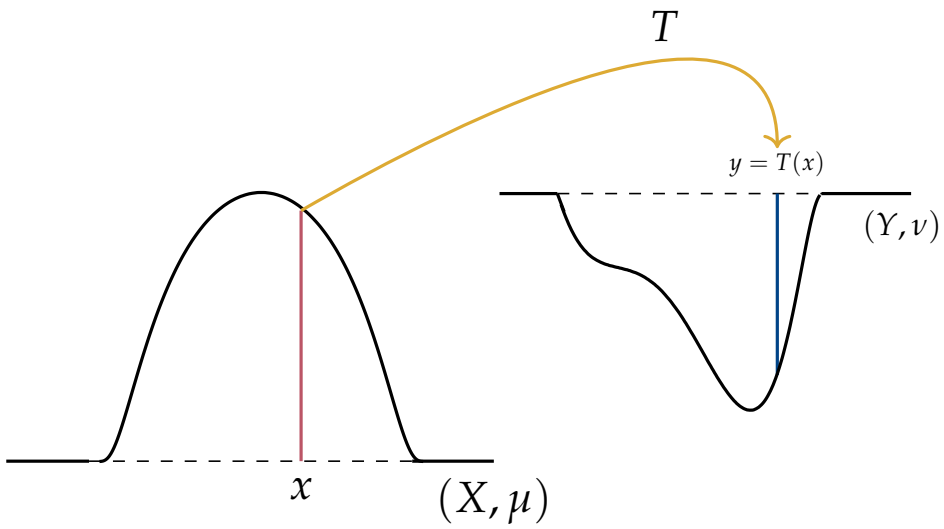


Figure 1.1: Visualization of the optimal transport (Monge) problem between two probability densities μ and ν. The density ν is flipped to facilitate the physical interpretation of moving soil from a mound into a ditch, which was the original consideration of this problem found in Monge [1781]. The deterministic mapping T associates mass at x to mass at y, which minimizes some cost.

The second object is the Kantorovich distance[4], which is the value function of the optimization problem of transporting mass from one distribution to another. Unlike many popular discrepancy measures, such as the Kullback-Leibler (KL) divergence, it is

[3]Univariate convex functions have the property that their first derivative is an increasing function. Multivariate convex functions have gradients that are *cyclically monotone*; see Rockafellar [1970] or Chapter 2 equation 2.3.

[4]The Kantorovich distance is also known as the Wasserstein distance, Kantorovich-Rubinstein metric, or earth mover's distance.

a proper metric on spaces of distributions satisfying symmetry, positive definiteness, and the triangle inequality. It also carries qualitative advantages over other metrics. For one, it provides a well-defined and meaningful distance between continuous and discrete distributions; see Figure 1.2 for an example comparison between total variation and Kantorovich distance. More generally, it captures key geometric features of distributions, e.g., support and shape information between densities; see Figure 1.3. A consequence of this is that the Kantorovich distance is well-defined between distributions with disjoint support unlike the KL divergence.

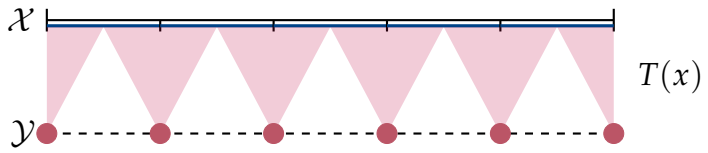


Figure 1.2: Consider the uniform distribution X over $\mathcal{X} = [0, 1]$ and the uniform distribution Y over $\mathcal{Y} = \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. The mass and support of these two distributions are close, intuitively. However, the total variation distance between these distributions is 1 ($\mathcal{Y}$ is X-a.s. zero), while the Kantorovich distance using Euclidean cost is about 0.06.

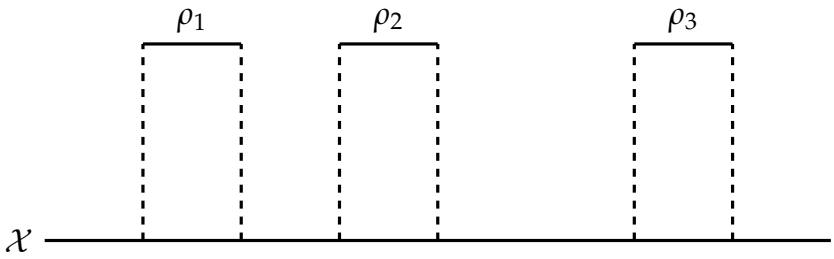


Figure 1.3: Three disjoint, but of same height uniform densities ρ_1, ρ_2 and ρ_3 on a space $\mathcal{X}$. The Hellinger distances and total variation distances are pairwise equal, while the KL divergence is undefined. However, ρ_1 and ρ_2 are "closer" than ρ_3 is to any of the others. The Kantorovich distance yields more intuitive results.

Given the advantages of the Kantorovich distance, many have proposed its use as a loss function to estimate parametric models in place of maximum likelihood (MLE). As an example in econometrics, Fan and Park [2022] proposed a "sliced" variation of the Kantorovich distance to estimate parameters that may lie on the boundary of a discontinuity– a commonly encountered situation where MLE performs poorly. On the other hand, the Kantorovich distance being a proper distance function admits an appropriate notion of "average" of probability distributions or *Fréchet mean*; see Agueh and Carlier [2011]. The Kantorovich-based Fréchet mean is known as the barycenter and it is a probability distribution that minimizes the sum of squared Kantorovich distances from the other distributions to itself, analogous to the average of a scalar finite sequence minimizing the sum of squared deviations. Figure 1.4 shows an example comparing the arithmetic average and barycenter of normal densities with equal variance. The barycenter in this example is also a normal density with a mean that is the average of the means of the other densities.

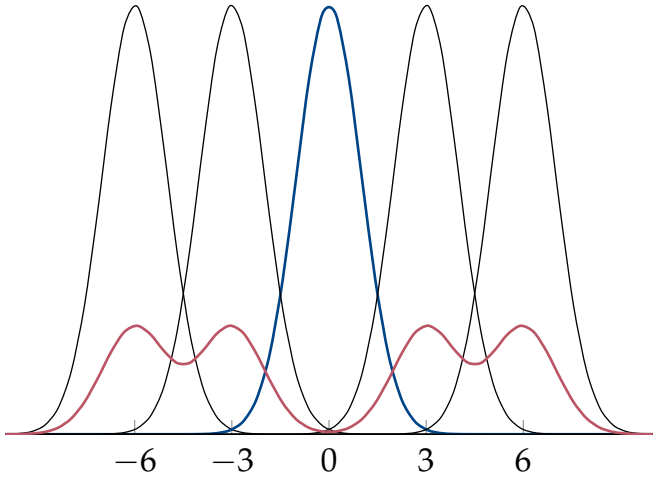


Figure 1.4: Black: 4 normal densities varying by mean; Red: Their arithmetic average; Blue: Their barycenter, a normal density with mean as the average of their means.

The barycenter has been shown to be a useful perspective for latent factor discovery. Yang and Tabak [2022] formulate a criteria to learn the latent variable by minimizing the variance of the barycenter of the conditional distributions given a candidate latent variable density. That is, we seek the distribution of the latent variable that maximizes the information gained from filtering it from the data. This is related to the problem of k-means and may be regarded as a generalization. From here, we can consider clustering that uses many features of the latent conditional distributions, which leads to so-called unobserved group heteroskedasticity where an unobserved discrete variable contributes to the heteroskedasticity (varying second moments) of a model. This is the topic of Chapter 3 where I include covariates in a linear model with discrete heterogeneity.

Another appealing quality of optimal transport is its relation to linear programming and convex optimization, therefore amenable to a vast library of efficient computational techniques. For example, when both distributions being transported are discrete, the problem is explicitly a linear program with equality constraints and can be solved efficiently, e.g., using variations of the simplex or interior-point algorithm. When one distribution is discrete, also known as the *semidiscrete* problem, the dual formulation can be exploited to solve for the transport map by using popular unconstrained convex minimization routines such as Newton's method and its variations.

In summary, the methodological core of this dissertation is optimal transport theory. The vector quantile satisfies key properties to be a suitable notion of a multivariate quantile. It is well-defined between continuous and discrete distributions, enabling calculations using data. It is also the unique gradient of a convex function implying monotonicity and an ordering of vectors. Using this, Chapter 2 extends the integrated quantile formulation of the Lorenz curve due to Gastwirth [1971][5]. On the other hand, Chapters 3 and 4 borrow inspiration from barycentric clustering. Using panel data, repeated measurements of units are used to identify the lower dimensional fixed effect in

[5]This chapter is based on joint work from Fan et al. [2024].

panel models. Chapter 3 considers the case where the fixed effect is a discrete random variable, generates time varying heterogeneity and heteroskedasticity. Chapter 4 extends to a continuous fixed effect called a *type* that describes heterogeneous effects and time-varying heterogeneity in a generalized method of moments (GMM) framework.

1.1 Literature

First studied in Monge [1781], the original problem was to solve for a deterministic mapping that minimized the absolute value transportation cost of displacing soil from one site to another. Such a deterministic mapping may not always exist, and the seminal work by economist Leonid Kantorovich (Kantorovich [2006a,b]) showed that a relaxation of the problem from pure assignments to *mixed assignments* will always yield a solution under weak assumptions. For example, Kantorovich [1960] present an application of transporting raw minerals from a distribution of mines to a distribution of refineries, where raw minerals from one mine may be sent to multiple refineries. This solution is known as an optimal *coupling* and is the solution to an infinite dimensional linear program. This implies that the optimal transport problem can be reformulated as a dual convex optimization problem.

Optimal transport theory has been involved in "booms" across many disciplines. These include physics, biology, statistics, machine learning, operations research, and economics. The standard mathematical references are Villani [2003, 2009]. Santambrogio [2015] is geared more towards applied fields with computational considerations. Galichon [2018] provides an excellent introduction to economists with an online resource for computations. Another economics reference by Ekeland [2010] introduces the topic under the lense of the theory of incentives. For a reference to all-things computational optimal transport, see Peyré and Cuturi [2018]. These serve as the most standard references for optimal transport. Each chapter herein contains additional references specialized to the topic discussed and I defer them until further on.

Chapter 2

LORENZ MAP, INEQUALITY ORDERING AND CURVES BASED ON MULTIDIMENSIONAL REARRANGEMENTS

2.1 Introduction

The Lorenz curve, first proposed in Lorenz [1905], is a compelling visual and simple quantification tool for the analysis of dispersion in univariate distributions. It allows easy visualization of dispersion from the curvature of a convex curve and its distance from the diagonal. The diagonal itself is the Lorenz curve of a degenerate distribution– an egalitarian allocation where all individuals have the same amount of resource. It also enables quick computations, reading off the curve, as it were, of the share of a resource held by the top or bottom of the allocation distribution for that resource. These features of the Lorenz curve account for much of its enduring appeal among practitioners, policy analysts and policy makers. This appeal is further enhanced by the relation between majorization and the pointwise ordering of Lorenz curves, which provides a way to visualize inequality comparisons between populations and within a given population across time periods. Comprehensive accounts are given in Marshall et al. [2011] and Arnold and Sarabia [2018].

The appealing properties of the Lorenz curve are well captured by the formulation given in Gastwirth [1971]. In that formulation, the Lorenz curve is the graph of the Lorenz map, and the latter is the cumulative share of individuals below a given rank in the distribution, i.e., the normalized integral of the quantile function. The relation

---

This chapter is based on joint work with Yanqin Fan, Marc Henry and Brendan Pass.

to majorization and the convex order follows immediately, as shown in Section C of Marshall et al. [2011]. As pointed out by Arnold [2008], this makes the Lorenz ordering an uncontroversial partial inequality ordering of univariate distributions, and most open questions concern the higher dimensional case.

Dispersion in multivariate distributions is not adequately described by the Lorenz curve of each marginal, and a genuinely multidimensional approach is needed. Even for utilitarian welfare inequality, Atkinson and Bourguignon [1982] motivate the need for the multidimensional approach initiated by Fisher [1956]. More generally, the literature on multidimensional inequality of outcomes and its measurement is vast, as evidenced by many recent surveys, see for instance Decancq and Lugo [2012], Aaberge and Brandolini [2014], Andreoli and Zoli [2020]. We only discuss it insofar as it relates to the Lorenz curve.

Multivariate extensions have been proposed for the Lorenz curve: Taguchi [1972a], Arnold [1983], and Koshevoy and Mosler (1996, 1999)[1]. They are reviewed in Marshall et al. [2011] and Sarabia and Jorda [2014] and discussed in more details in Section 2.2.4, where we compare them to our proposal. We contribute to this literature with a vector version of the Gastwirth [1971] formulation of the Lorenz curve. We provide an implementable criterion to measure and compare inequality in multivariate distributions, which emulates the features of the Lorenz curve that most contributed to its success.

The traditional Gastwirth [1971] formulation of the Lorenz curve is an integrated quantile over the lowest ranked individuals. To simplify the argument in the univariate case, model the population as a continuum on $[0,1]$ and suppose the distribution of incomes in the population is continuous. Then the Gastwirth [1971] formulation can be thought of involving two stages. Take an income allocation Y, which is a random variable on $\mathbb{R}_+$ with cumulative distribution function F_Y. First, reorder individuals in the

[1]More recently, subsequent to our work, Hallin and Mordant [2022] also adopt a multivariate rearrangement approach to the definition of multi-attribute Lorenz curves. They adopt a center-outward approach, which is better suited to define notions of middle class.

population so that they are ranked in increasing incomes. Then compute the cumulative share of lowest ranked individuals by integrating F_Y^{-1} from 0 to r and dividing by the mean. The first step involves the probability integral transform $F_Y(Y)$, which should be thought of in this context as a cardinal to ordinal transformation, since $F_Y(Y)$ is uniform on $[0,1]$, so cardinal information is purged, but F_Y is increasing, so ordinal information is preserved. Our proposal is based on a multivariate version of the cardinal to ordinal transformation involved in the first stage. The latter is the unique map that transforms a d dimensional allocation into a uniform one on $[0,1]^d$, and is cyclically monotone[2] and hence preserves ordinal information. This motivates our definition of the vector Lorenz map as the cumulative integral of the multivariate quantile of Chernozhukov et al. [2017].

The vector Lorenz map we propose, therefore, is the vector of shares of each resource held by individuals below a given rank. The associated Lorenz inequality dominance criterion deems a multivariate allocation more equal if this share of resources is larger for each rank. Hence, our proposal shares the interpretation of the traditional Lorenz curve and Lorenz dominance. It also shares the desirable properties of the Lorenz curve and dominance ordering. Like the Lorenz zonoid of Koshevoy and Mosler (1996, 1999), it characterizes the distribution of an allocation (see Section 2.2.4 for a definition and discussion). Unlike the Lorenz zonoid, the vector Lorenz map we propose can be efficiently computed as an unconstrained convex optimization problem and connected to recent developments in computational optimal transport theory [3]. Hence, the Lorenz dominance order we propose is an implementable inequality dominance criterion. Using recent advances on the asymptotic properties of multivariate quantiles, surveyed in Hallin [2022], our Lorenz dominance criterion can be the basis for inequality dominance testing that accounts for sampling uncertainty. This contrasts our proposal with the growing literature on multivariate inequality dominance criteria for finite populations. See Gravel and Moyes [2012], Banerjee [2016], Faure and Gravel [2021] and references within.

[2]Existence and uniqueness are shown in McCann [1995]. See Section 2.2.1 for details and definitions.

[3]See Peyré and Cuturi [2018] for account of recent advances

Other implementable inequality dominance criteria are proposed in the literature, in Koshevoy [1995], Koshevoy and Mosler [1996], Koshevoy and Mosler [2007] and Banerjee [2016] and other references surveyed in Arnold and Sarabia [2018]. However, they do not provide an equivalence between the Lorenz dominance criterion and a class of compatible social evaluation functionals. An exception is Gravel and Moyes [2012] and Faure and Gravel [2021] who give a comprehensive treatment of the special case of a finite population with a single cardinal transferable attribute combined with an ordinal non transferable one. We characterize the class of social evaluation functionals that are inequality averse in the sense that they are increasing in the Lorenz dominance order. We build on the multivariate extension of the Quiggin [1992] and Yaari [1987] rank dependent decision theory in Galichon and Henry [2012] to show that, as in Weymark [1981] for the univariate case, social evaluation functionals are inequality averse if and only if they are rank dependent social evaluation functionals with attribute specific weights decreasing in ranks. We also characterize the class of transfers that increase inequality according to the Lorenz dominance criterion as rank preserving transfers of any attribute from a lower to a higher ranked individual. A special case of such transfers, which we call *monotone regressive transfers* weakly increase marginal inequality and dependence between attributes.

To visualize Lorenz dominance, we define an *Inverse Lorenz Function* at a given vector of resource shares as the share of the population that cumulatively holds those shares. It is characterized by the cumulative distribution function of the image of a uniform random vector by the Lorenz map. Hence, it is a cumulative distribution function by construction, like the univariate inverse Lorenz function. In two dimensions, the α-level sets of this cumulative distribution function, which we call α-Lorenz curves, are non crossing downward sloping curves that shift to the south-west when inequality increases, as defined by the Lorenz ordering. Finally, we propose an illustration to the analysis of income-wealth inequality in the United States between 1989 and 2022.

In the first section, we define the Lorenz map, explain its computation, detail its

properties and how it compares with alternative proposals. In Section 2.3, we introduce the Lorenz dominance ordering, its characterization in terms of classes of social evaluation functions and in terms of transfers compatible with it. Section 2.4 illustrates the visualization of Lorenz dominance and the final Section concludes.

2.2 *Vector Lorenz Map*

2.2.1 Definition of the Lorenz map

The Lorenz curve was originally proposed in Lorenz [1905] to provide a graphical representation of inequality within a single resource. Let Y be a random variable on $\mathbb{R}_+$ with cumulative distribution function F_Y, which represents the allocation of a resource in a population. The population is modeled as the continuum $[0,1]$.

The Lorenz curve is traditionally defined as the set of points in $[0,1]^2$, parameterized by y, with coordinates

$$\left(F_Y(y), \frac{1}{\mu_Y}\int_0^y v dF_Y(v)\right), \tag{2.1}$$

where μ_Y is the expectation of Y. See for instance page 149 of Arnold and Sarabia [2018]. Gastwirth [1971] points out that the Lorenz curve is given by the graph of the map on $[0,1]$:

$$q \mapsto L_Y(q) = \frac{1}{\mu_Y}\int_0^q F_Y^{-1}(v)\, dv, \tag{2.2}$$

where $F^{-1}(v) := \inf\{y : v \leq F_Y(y)\}$ is the traditional quantile function. Formulation (2.2) provides a closed form expression and simple interpretation: For each proportion $q \in [0,1]$, the Lorenz map gives the cumulative share of the resource held by the poorest proportion q of the population. This relies on the well known fact that the quantile function F_Y^{-1} is the only increasing function such that for any uniformly distributed random variable V on $[0,1]$, $F_Y^{-1}(V)$ is distributed identically to Y. Hence, integrating F_Y^{-1} from 0 to q and normalizing produces the cumulative share held by the individuals ranked below q.

Conversely, when F_Y admits a density, the probability integral transform $V := F_Y(Y)$ produces a uniformly distributed random variable V on $[0,1]$, which preserves the ranks of individuals in the population. This holds because F_Y is an increasing map. Hence, the probability integral transform removes cardinal information (by producing a uniformly distributed outcome), while preserving ordinal information (by keeping the rank order of individuals in the population). The probability integral transform $V := F_Y(Y)$ is the rank associated with allocation Y.

If F_Y is not continuous, then $F_Y(Y)$ is no longer uniformly distributed (positive masses of individuals have identical ranks). However, defining F_Y^{-1} as the generalized inverse of F_Y, i.e.,

$$q \mapsto F_Y^{-1}(q) := \inf_y \{y : F_Y(y) \geq q\},$$

it is still the case that for any uniformly distributed random variable V on $[0,1]$, $F_Y^{-1}(V)$ is distributed identically to Y. Hence, the closed form solution for the Lorenz map (2.2) still holds with the same interpretation: Integrating F_Y^{-1} from 0 to q and normalizing still produces the cumulative share held by the individuals ranked below q.

Consider now an allocation $X := (X_1, \ldots, X_d)$ of d resources in the population. To analyze inequality in allocation X, we can first look at inequality in each marginal allocation $X_1, \ldots, X_d$, using the univariate Lorenz curves $L_1 := L_{X_1}, \ldots, L_d := L_{X_d}$. However, this strategy disregards the effect of dependence. The latter is relevant to inequality, as can be trivially illustrated by the fact that for given wealth and income marginal allocations, the comonotonic allocation (the wealthier individuals have higher income) is more unequal than the admittedly unrealistic counter-monotonic allocation (the wealthier individuals have lower income).

To take dependence into account, we propose to emulate the Gastwirth [1971] formulation by measuring cumulative shares of each resource for all individuals, *below a certain rank*. Conceptually, this is achieved in two steps. First, we find a transformation that removes cardinal information while preserving individual's ranking in the population, i.e.,

a cardinal to ordinal transformation. Then we integrate the shares of individuals with lowest rank. The difficulty here, of course, is the absence of a canonical order in $\mathbb{R}^d$ to define the rank.

As noted in Faugeras and Rüschendorf [2017], by Borel's isomorphism Theorem[4], there exist measurable bijective maps $T : [0,1] \to \mathbb{R}^d$ such that for any uniformly distributed random variable V on $[0,1]$, $T(V)$ is distributed identically to X. However, such maps are unsuitable cardinal to ordinal transformations for two main reasons. First, there is no known explicit construction, hence no way to compute them. Second, even if we could compute such a map, its choice would imply an implicit ad hoc aggregation of the different resources in allocation X in order to arrive at a scalar ranking of individuals in the population.

In order to avoid an implicit ad hoc aggregation of the different resources in X, the cardinal to ordinal transformation must be between $\mathbb{R}^d$ and $[0,1]^d$. Hence, we model the population as a continuum on $[0,1]^d$ and individual ranks are points in $[0,1]^d$. The multivariate quantile transform, and its inverse (the cardinal to ordinal transform, or rank transform), must satisfy the same requirements as in the univariate case: It must map the uniform distribution (no cardinal information) to the distribution of the allocation, and it must be monotonic (so as to preserve ordinal information). The monotonicity of the quantile in the univariate case ensures that the cardinal to ordinal transformation does indeed preserve the rankings of individuals in the population.

To construct an analogue of the Gastwirth [1971] Lorenz curve formulation, we therefore need the cardinal to ordinal transformation to satisfy a form of multivariate monotonicity. The classical notion of monotonicity in $\mathbb{R}^d$, also known as 2-monotonicity, of a map $T : \mathbb{R}^d \to \mathbb{R}^d$, requires

$$\left(T(x') - T(x)\right)' \left(x' - x\right) \quad \geq \quad 0$$

for any pair of vectors $x, x' \in \mathbb{R}^d$. It can be interpreted as monotonicity on average.

[4]See for instance section 13.1 page 487 of Dudley [2002].

For uniqueness of the cardinal to ordinal transformation, we need the stronger version of monotonicity, called *cyclical monotonicity*, which characterizes the gradients of convex functions and was introduced by Rockafellar [1966]. Cyclical monotonicity requires

$$\sum_{i=1}^{K}(T(x_i) - T(x_{i+1}))'(x_i - x_{i+1}) \quad \geq \quad 0 \tag{2.3}$$

for any K, and any collection of vectors $(x_1, \ldots, x_K)$, setting $x_{K+1} = x_1$. Cyclical monotonicity also characterizes maps T that minimize distortion in the sense that in case the allocation X has continuous distribution with finite variance, T minimizes $\mathbb{E}\|X - T(X)\|^2$ among all the maps such that $T(X)$ is uniformly distributed on $[0,1]^d$.

The following definition summarizes the properties needed for a cardinal to ordinal transformation as a first step in the Lorenz map construction.

Definition 1 (Vector quantile)**.** A vector quantile Q_X associated with random vector X on $\mathbb{R}^d$ is a map $Q_X : [0,1]^d \to \mathbb{R}^d$ with the following properties:

i. If U is uniform on $[0,1]^d$, then $Q_X(U)$ is distributed identically to X.

ii. If X is absolutely continuous, then $Q_X^{-1}(X)$ exists and is uniform on $[0,1]^d$.

iii. The map Q_X is cyclically monotone.

Properties (i) and (iii) imply Q_X is the traditional quantile function when $d = 1$. As shown in McCann [1995], there exists a transformation that conforms with Definition 1 and it is unique in the sense that two such transformations are equal almost everywhere. It is proposed as a vector quantile notion in Chernozhukov et al. [2017], and we will refer to it as the vector quantile associated with X.

Once we model the population as a continuum on $[0,1]^d$, interpret each point on $[0,1]^d$ as a rank, and define the vector quantile Q_X as a multidimensional rearrangement of the allocation X in rank order, we simply integrate the quantile over the lowest ranks to define a multivariate version of the Gastwirth [1971] formulation of the Lorenz curve.

Definition 2 (Lorenz map)**.** Let U be a uniformly distributed random vector on $[0,1]^d$, and let $X := (X_1, \ldots, X_d)$ be an allocation, i.e., a random vector on $\mathbb{R}^d_+$ with finite mean $\mu = (\mu_1, \ldots, \mu_d)$. Call $\widetilde{X}$ the normalized version of X, i.e.,

$$\widetilde{X} := \left(\frac{X_1}{\mu_1}, \ldots, \frac{X_d}{\mu_d} \right),$$

and let $Q_{\widetilde{X}}$ be the vector quantile of $\widetilde{X}$. The *Lorenz map* of allocation X is the vector-valued function $\mathcal{L}_X : [0,1]^d \to [0,1]^d$ defined for each $r := (r_1, \ldots, r_d) \in [0,1]^d$ by

$$\mathcal{L}_X(r_1, \ldots, r_d) \;=\; \int_0^{r_1} \cdots \int_0^{r_d} Q_{\widetilde{X}}(u_1, \ldots, u_d)\, du_1 \ldots du_d. \tag{2.4}$$

The transformation of X into its normalized version $\widetilde{X}$ prior to integrating the vector quantile is required to remove dependence of the Lorenz map of Definition 2 on units of measurements. Different resources, such as earnings and health, may not be measured with the same units of measurement. The transformation into $\widetilde{X}$ makes the allocation unit free. Hence the Lorenz map satisfies ratio-scale invariance (i.e., invariance to rescaling of the different attributes, or change of units of measurement). Section 2.2.3 discusses an alternative unnormalized version of the definition in the spirit of Shorrocks [1983].

When X has absolutely continuous distribution P_X, its quantile function is P_X-almost everywhere invertible (see for instance Theorem 2.1 in Chernozhukov et al. [2017]). In that case, the transformation $U = Q_X^{-1}(X)$ is the vector analogue of the probability integral transform $V = F_Y(Y)$ discussed above. The random vector $U = Q_X^{-1}(X)$ is uniformly distributed on $[0,1]^d$, and is the vector rank of the individual with endowment X, in the terminology of Chernozhukov et al. [2017]. The Lorenz map of Definition 2 can then be rewritten as:

$$\mathcal{L}_X(r) \;=\; \mathbb{E}\left[\widetilde{X}\, \mathbb{1}\left\{ Q_{\widetilde{X}}^{-1}(\widetilde{X}) \leq r \right\} \right]. \tag{2.5}$$

This clarifies the interpretation of $\mathcal{L}_X(r)$ as the cumulative share of all individuals with vector rank below r in the partial order of $\mathbb{R}^d$.

In the scalar case discussed above, inverting the Lorenz curve L_Y defined in (2.2) yields the inverse Lorenz curve

$$L_Y^{-1}(y) = \int_0^{L_Y^{-1}(y)} dv = \int_0^1 \mathbb{1}\{L_Y(v) \leq y\}\, dv = \mathbb{P}(L_Y(V) \leq y), \tag{2.6}$$

where the probability is taken with respect to a uniformly distributed random variable V on $[0,1]$. The scalar inverse Lorenz curve at y is therefore shown in (2.6) to be equal to the maximum proportion of the population with cumulative share of the resource equal to y. In the vector case, the analogue of the right-hand side of (2.6) can still be used to define an inverse Lorenz function.

Definition 3 (Inverse Lorenz Function)**.** The Inverse Lorenz Function (ILF) of a random vector X is the function $l_X : [0,1]^d \to [0,1]$ defined for each $z = (z_1, \ldots, z_d) \in [0,1]^d$ by $l_X(z) := \mathbb{P}(\mathcal{L}_X(U) \leq z)$, where $z = (z_1, \ldots, z_d) \in [0,1]^d$, inequality $\leq$ is understood component-wise, and the probability is taken with respect to the uniform random vector U on $[0,1]^d$.

The expression above is no longer the mathematical inverse of the Lorenz map $\mathcal{L}_X$, but it can still be interpreted as the share of the population with cumulative shares of all resources equal to a predetermined proportion $z = (z_1, \ldots, z_d)$.

2.2.2 Computation and examples

Computation

We now give a step-by-step method to compute the vector quantile, Lorenz map, and ILF of a discrete distribution, which may be the allocation in a finite population, or the empirical distribution of a (possibly weighted) sample from an underlying (possibly mixed discrete-continuous) distribution. The full algorithm and a step-by-step guide to implementation in `R` are given in Appendix A.1.

Let X be a random vector in $\mathbb{R}^d_+$ with discrete distribution. The probability mass function of the distribution of X is given by $\{(x^1, w_1), \ldots, (x^n, w_n)\}$, where $x^1, \ldots, x^n$ are vectors in $\mathbb{R}^d_+$ and $w_1, \ldots, w_n$ are positive scalar weights summing to 1.

Vector Quantiles. First, normalize the allocation vector X and form $\widetilde{X}$[5]. Then compute the vector quantile $Q_{\widetilde{X}}$ of $\widetilde{X}$. According to Definition 1 (requirements (*i*) and (*iii*)), the vector quantile $Q_{\widetilde{X}}$ must satisfy the following requirements:

1. For any uniformly distributed random variable U on $[0,1]^d$, $Q_{\widetilde{X}}(U)$ is distributed identically to $\widetilde{X}$. Hence:

   (a) For all $u \in [0,1]^d$, $Q_{\widetilde{X}}(u) \in \{x^1, \ldots, x^n\}$;

   (b) For all $i = 1, \ldots, n$,

   $$W_i \;:=\; Q_{\widetilde{X}}^{-1}(x^i) = \{u \in [0,1]^d : Q_{\widetilde{X}}(u) = x^i\} \tag{2.7}$$

   has measure w_i.

2. The map $Q_{\widetilde{X}}$ is cyclically monotone. Hence, by Rockafellar [1966], there is a convex function $\psi_{\widetilde{X}} : [0,1]^d \to \mathbb{R}$ such that $Q_{\widetilde{X}}$ is almost everywhere equal to the gradient of $\psi_{\widetilde{X}}$.

Since $Q_{\widetilde{X}}$ takes a finite number of values and is constant and equal to x^i on each W_i, the computation of $Q_{\widetilde{X}}$ is equivalent to the computation of the partition of $[0,1]^d$ in regions $W_1, \ldots, W_n$. As $Q_{\widetilde{X}}$ is the gradient of the convex function $\psi_{\widetilde{X}}$ and is constant on each of the W_i, $\psi_{\widetilde{X}}$ is affine on each of the W_i, and each W_i is a convex polytope in $[0,1]^d$. Aurenhammer et al. [1998] show

$$\psi_{\widetilde{X}}(u) \;=\; \max_{i=1,\ldots,n} \{u' x^i - h^i\},$$

[5]The issue of normalization is discussed in Section 2.2.3.

where $h := (h^1, \ldots, h^n)$ is the solution to the convex optimization program

$$\min \left\{ \sum_{i=1}^{n} w_i h^i + \int_{[0,1]^d} \max_{k=1,\ldots,n} \{u' x^k - h^k\} du \right\} \tag{2.8}$$

with first order condition that satisfies $1(b)$, i.e., for all $i = 1, \ldots, n$:

$$w_i = \lambda(W_i^h) \tag{2.9}$$

where the Lebesgue measure λ is simply the ordinary area of the convex polytope W_i^h. That is, the optimal h will weigh the partitions of the rank space $[0,1]^d$ in the same way to the corresponding sample points.

The algorithm minimizes (2.8) to find h, from which the regions W_i^h are obtained as

$$W_i^h \;=\; \{u \in [0,1]^d : u' x^i - h^i \geq u' x^j - h^j, 1 \leq j \leq n\}, \tag{2.10}$$

and $Q_{\widetilde{X}}$ is the map that takes value x^i on W_i^h, for each $i = 1, \ldots, n$.

Lorenz maps. Once we have computed the vector quantile map $Q_{\widetilde{X}}$, the Lorenz map at $r := (r_1, \ldots, r_n) \in [0,1]^d$ is obtained straightforwardly as the integral of the piece-wise constant map $Q_{\widetilde{X}}$ over $[0, r] := [0, r_1] \times \ldots \times [0, r_d]$:

$$\mathcal{L}_X(r) \;=\; \sum_{i=1}^{n} \lambda \left(W_i^h \cap [0, r] \right) \; x^i, \tag{2.11}$$

where the term under λ is the ordinary area of the convex polytope formed by the intersection of the cell W_i^h and the rectangle $[0, r]$.

Inverse Lorenz Function. Finally, $\mathcal{L}_X$ can be used to generate a pseudo sample $\{\mathcal{L}_X(U_k)\}$, where $\{U_k\}_{k=1}^m$ is a uniformly distributed random sample or any pseudo-random (aka minimum discrepancy) sequence that approximates the uniform distribution on $[0,1]^d$. The Inverse Lorenz function l_X can then be approximated with the empirical distribution of this pseudo-sample:

$$l(z) \;:=\; \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}\{\mathcal{L}_X(U_j) \leq z\} \;\; z \in [0,1]^d. \tag{2.12}$$

Examples

To illustrate the definition and the computation of the Lorenz map, we now explore examples of specific allocations and compute the corresponding Lorenz maps. First, we illustrate the computation of the Lorenz map for a discrete allocation.

Example 1 (Discrete allocations)**.** Let X be the allocation with probability mass function $\{(x^1, 1/n), \ldots, (x^n, 1/n)\}$. We select the support points $(x^1, \ldots, x^n)$ as the realizations of $n = 100$ i.i.d. draws from a bivariate standard normal distribution. The vector quantile $Q_{\widetilde{X}}$ of the normalized allocation $\widetilde{X}$ is characterized by its value x^i on the convex polygon W_i, such that $(W_1, \ldots, W_n)$ form the partition of $[0,1]^2$ shown on the left panel of Figure 2.1. As shown on the right panel of Figure 2.1, the Lorenz map at $r = (r_1, r_2) \in [0,1]^2$ is equal to the sum of the x^i's times the area of W_i intersected with $[0, r_1] \times [0, r_2]$.

Next, we consider the special case, where all individuals are endowed with the same quantity of resources.

Example 2 (Identical allocations)**.** Let X be the constant allocation $X = (1, 1, ..., 1) \in \mathbb{R}^d$. Then $Q_X(u) = (1, 1, ..., 1)$ for all $u \in [0,1]^d$, so that each entry of $\mathcal{L}_X(r)$ is $r_1 r_2 \cdots r_d$ so that the image of $\mathcal{L}_X$ is the diagonal in $[0,1]^d$. The Inverse Lorenz function $l_X(z)$ of X is 0 when $z_1 z_2 \cdots z_d = 0$. For $d \geq 1$ and $z = (z_1, z_2, ...z_d) \in (0,1]^d$, and letting $\underline{z} := \min\{z_1, z_2, ..., z_d\}$, the Inverse Lorenz function $l_X(z)$ of X is

$$\begin{aligned} l_X(z) &= \mathbb{P}(U_1 U_2 \cdots U_d \leq z_1, U_1 U_2 \cdots U_d \leq z_2, ..., U_1 U_2 \cdots U_d \leq z_d) \\ &= \mathbb{P}(U_1 U_2 \cdots U_d \leq \min\{z\}) \\ &= \min\{z\} \sum_{k=1}^{d} \frac{(-1)^{k-1}}{(k-1)!} \left[\log(\min\{z\})\right]^{k-1}. \end{aligned}$$

We also check that our definition is compatible with scalar definitions when all resources are mutually independent.

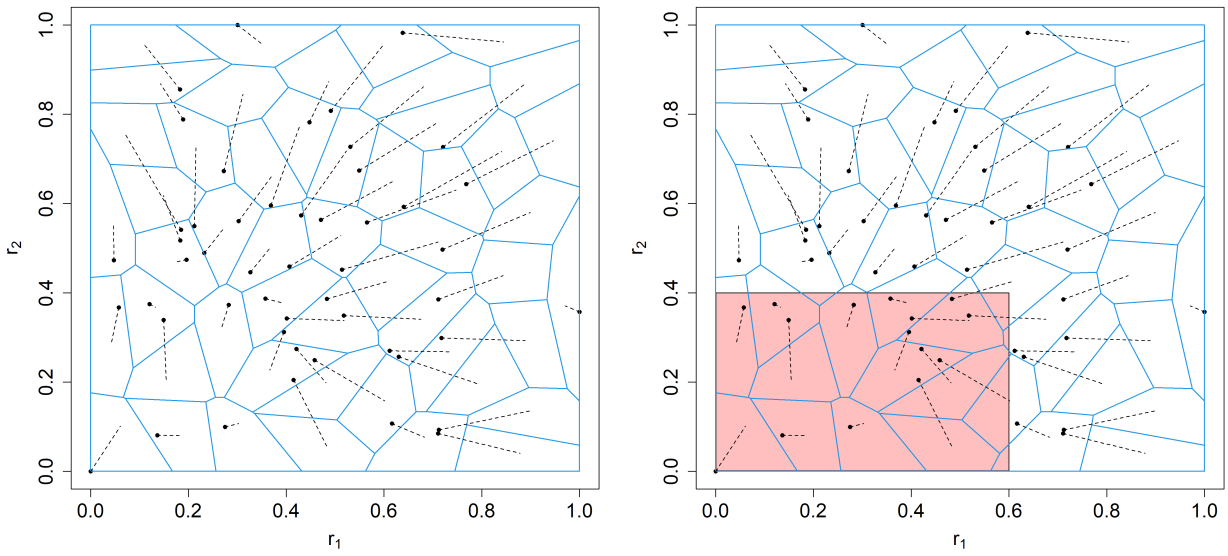


Figure 2.1: The left panel is a visualization of the quantile map. On each cell of the blue partition of $[0,1]^2$, the quantile map takes the value of the sample point indicated by a black dot and connected to it by a dashed line. Note that sample points are drawn from a standard bivariate normal, then shifted and rescaled to fit into $[0,1]^2$. The right panel shows a visualization of the computation of the Lorenz map at $r := (0.6, 0.4)$. The latter is equal to the sum of areas of the intersection between the red shaded area and the elements of the blue partition weighted by the corresponding sample realization.

Example 3 (Independent Resources)**.** Let the components $X_1, ..., X_d$ of X be mutually independent with marginal Lorenz curves $L_1, ..., L_d$, respectively. Then, the ith component of the Lorenz map is $L_i(r_i)\Pi_{j=1,j\neq i}^{d} r_j$. This expression of the Lorenz map has the following interpretation. Consider the first component $r_2 \cdots r_d L_1(r_1)$. The share of resource 1 held by people with multivariate rank in $[0, r_1] \times [0,1]^{d-1}$ is the marginal share, equal to the marginal Lorenz function. Since the resources are independent, this share is uniformly distributed along the other dimensions, so that people with ranks in $[0, r_1] \times [0, r_2] \times \cdots \times [0, r_d]$ command a share $r_2 \cdots r_d L_1(r_1)$. The other components

are interpreted analogously. When $d = 2$ and $r_1 = 1$, the Lorenz map takes values $\mathcal{L}_X(1, r_2) = (r_2, L_2(r_2))$. That is, the image of $\{(1, r_2) : 0 \leq r_2 \leq 1\}$ under $\mathcal{L}_X$ is the Lorenz curve L_2 of the second resource X_2 (and symmetrically when $r_2 = 1$).

The Inverse Lorenz Function $l_X(z)$ of allocation X with independent components is

$$\begin{aligned} l_X(z) &= \mathbb{P}(\, U_2 \cdots U_d L_1(U_1) \leq z_1, U_1 U_3 \cdots U_d L_2(U_2) \leq z_2, ..., U_1 \cdots U_{d-1} L_d(U_d) \leq z_d) \\ &= \int_{[0,1]^{d-1}} \min \left\{ \frac{z_1}{L_1(u_1)\Pi_{k \neq 1,d} u_k}, \frac{z_2}{L_2(u_2)\Pi_{k \neq 2,d} u_k}, ..., l_d \left(\frac{z_d}{\Pi_{k \neq d} u_k} \right) \right\} du_1 ... du_{d-1}, \end{aligned}$$

where l_d is the univariate inverse Lorenz curve of X_d.

Next, we derive the Lorenz map for $d = 2$ in the case of allocations $X = (X_1, X_2)$ with the same components, i.e., $X_1 = X_2$ almost surely.

Example 4 (Comonotonic Resources)**.** Let the components X_1 and X_2 of the allocation X be almost surely equal. Then, X_1 and X_2 have identical distributions. Since the distribution of $X = (X_1, X_2)$ concentrates on the line $x_1 = x_2$, the cost function $c(x, u) = -x \cdot u = -\frac{1}{2}[(x_1 + x_2)(u_1 + u_2) + (x_1 - x_2)(u_1 - u_2)] = -\frac{1}{2}(x_1 + x_2)(u_1 + u_2)$ in the definition of the multi-variate quantile depends on u only through $u_1 + u_2$ (it is an index cost in the terminology of Chiappori et al. [2017]). Therefore, Q_X depends only on $u_1 + u_2$. More precisely, it is $(u_1, u_2) \mapsto (\psi'(u_1 + u_2), \psi'(u_1 + u_2))$, where $z \mapsto \psi'(z) = Q_X(z)$ is the map from σ to the distribution of X_1, where σ has density on $[0, 2]$ given by $1 - |1 - z|$. Each component of the Lorenz curve is then given by

$$\mathcal{L}_1(r_1, r_2) = \mathcal{L}_2(r_1, r_2) \;\; = \;\; \int_0^{r_2} \int_0^{r_1} \psi'(u_1 + u_2) du_1 du_2 = \int_0^{r_2} [\psi(u_2 + r_1) - \psi(u_2)] du_2.$$

In case X_1 and X_2 are uniformly distributed on $[0, 2]$, the optimal transport map ψ' for $z < 1$ is given by $\psi'(z) = z^2$, so that $\psi(z) = z^3/3$. We then have,

$$\mathcal{L}_1(r_1, r_2) = \frac{r_1^3 r_2}{3} + \frac{r_1 r_2^3}{3} + \frac{r_1^2 r_2^2}{2},$$

when $r_1 + r_2 \leq 1$, and

$$\mathcal{L}_1(r_1, r_2) = \frac{2}{3}(r_1 + r_2)^3 - \frac{1}{12}(r_1 + r_2)^4 - (r_1 + r_2)^2 - \frac{r_1^4}{12} - \frac{r_2^4}{12} + \frac{2}{3}(r_1 + r_2) - \frac{1}{6},$$

when $r_1 + r_2 > 1$. The image of this Lorenz map is once again the diagonal in $[0,1]^2$.

The Inverse Lorenz function $l_X(z)$ of allocation $X = (X_1, X_2)$ with $X_2 = X_1$ almost surely, is

$$\begin{aligned} l_X(z) &= \mathbb{P}\left(\mathcal{L}_1(R) \leq z_1, \mathcal{L}_2(R) \leq z_2\right) \\ &= \mathbb{P}\left(\mathcal{L}_1(R) \leq \min\{z_1, z_2\}\right) \\ &= h\left(\min\{z_1, z_2\}\right), \end{aligned}$$

where $\mathcal{L}_j(z)$ is the j-th component, $j = 1, 2$, of $\mathcal{L}_X(r)$, and h is the distribution function of $\mathcal{L}_1(R)$ for R uniform on $[0,1]^2$.

2.2.3 *Relative scale and normalization*

Let X be the original allocation. Normalizing X into $\widetilde{X}$ as in Definition 2 by dividing each component by its mean, removes any sensitivity to (changes in) units of measurements. It is a standard approach to achieve ratio-scale invariance. See for instance Banerjee (2010, 2016). However, by construction, it comes with the disadvantage of removing scale effects. In the univariate case, Shorrocks [1983] proposes to eschew normalization in order to take scale effects into account in the measurement of inequality.

When units of measurement are not a concern, an alternative definition without the feature described above is defined as

$$\int_0^{r_1} \cdots \int_0^{r_d} Q_X(u_1, \ldots, u_d)\, du_1 \ldots du_d,$$

where Q_X is the vector quantile of the original allocation X (not the normalized one). This alternative version also allows the weighting of different resources according to a priori importance to overall inequality. Call $X^\epsilon = (\lambda_1(\epsilon)X_1, \ldots, \lambda_d(\epsilon)X_d)$ the suitably rescaled version of the initial allocation X. Define the sequence of weights $(\lambda_1(\epsilon), \ldots, \lambda_d(\epsilon))$ in such a way that $\lambda_{k+1}(\epsilon)/\lambda_k(\epsilon) \to 0$ as $\epsilon \to 0$. In this way, resources are ordered in decreasing importance to inequality, and we can entertain the extreme lexicographic

case, where $\epsilon \to 0$. From there, we can construct the alternative Lorenz map $\widetilde{\mathcal{L}}_{X^\epsilon}$ of the rescaled allocation vector.

It follows from Carlier et al. [2010] that, when X has an absolutely continuous distribution, as ϵ tends to 0, the alternative Lorenz map tends to the component-wise demeaned map

$$\int_0^{r_1} \cdots \int_0^{r_d} Q_X^{KR}(u_1, \ldots, u_d)\, du_1 \ldots du_d,$$

where Q_X^{KR} is the inverse of the Knothe-Rosenblatt transform $T^{KR} : \mathbb{R}^d \mapsto [0,1]^d$ of the original allocation X proposed by Rosenblatt [1952] and Knothe [1957], and defined as follows

$$T^{KR}(x_1, x_2, \ldots, x_d) \quad := \quad \begin{bmatrix} F_{X_1}(x_1) \\ F_{X_2|X_1}(x_2|x_1) \\ \vdots \\ F_{X_d|X_1,\ldots,X_{d-1}}(x_d|x_1, \ldots, x_{d-1}) \end{bmatrix}.$$

The Knothe-Rosenblatt quantile map is the only multivariate quantile map from the uniform on $[0,1]^d$ to $\mathbb{R}^d$ proposed in the literature other than optimal transport based vector quantiles as in Definition 1. The result above shows that the Knothe-Rosenblatt quantile is not a good alternative to vector quantiles of Definition 1 to base an integrated quantile definition for the Lorenz map, since it relies on an a priori lexicographic ordering of the different resources in the allocation.

2.2.4 Properties and comparisons with other multivariate Lorenz concepts

In this section, we detail previous proposals for multivariate extensions of the Lorenz curve and list the properties that distinguish our proposal from the former.

Alternative multivariate Lorenz proposals

Until now, the development of multivariate extensions of the Lorenz curve was hampered by the lack of simple multivariate analogues of ranks and quantiles. Early propos-

als for bivariate extensions of the Lorenz curve in Taguchi (1972a,1972b) and Arnold (1983,2012) are based on a direct ad-hoc extension of the traditional formula given in (2.1). Let $(x_1, x_2) \mapsto F(x_1, x_2)$ be the CDF of a bivariate allocation with density f and mean (μ_1, μ_2). Taguchi (1972a,1972b) proposes the bivariate Lorenz surface $L : [0,1]^2 \to [0,1]$ defined implicitly by $(s, t, L(s,t)) :=$

$$\left(F(x_1, x_2)\,,\, \frac{1}{\mu_1} \int_0^{x_1}\!\!\int_0^{x_2} u_1 f(u)\, du\,,\, \frac{1}{\mu_2} \int_0^{x_1}\!\!\int_0^{x_2} u_2 f(u)\, du \right). \tag{2.13}$$

In order to treat both dimensions of the allocation symmetrically, Arnold (1983,2012) proposes the alternative Lorenz surface parameterized by (x_1, x_2) as the set of points

$$\left(F_1(x_1)\,,\, F_2(x_2)\,,\, \frac{1}{\mu_{12}} \int_0^{x_1}\!\!\int_0^{x_2} u_1 u_2 f(u)\, du \right), \tag{2.14}$$

where F_1 and F_2 are the marginal CDFs associated with F, and μ_{12} is the expectation of the product $X_1 X_2$. A closed form solution, given in Sarabia and Jorda [2014], makes the Lorenz surface (2.14) amenable to parametrization and statistical analysis. However, it does not share the interpretation or any of the properties of the univariate Lorenz curve.

A more successful proposal in that respect, is the Lorenz zonoid of Koshevoy and Mosler [1996]. Again, take (2.1) in the univariate case as the point of departure. It associates a fraction p of the population to the share of the resource collectively held by the poorest fraction p of the population. Koshevoy and Mosler [1996] eschew the need to order the population by associating with a fraction p of the population the share of resources held by any group of individuals making up a fraction p of the population, poor, rich, or mixed. The lower bound is the share held by the poorest individuals (the traditional Lorenz curve), and the upper bound is the share held by the richest individuals (a reverse Lorenz curve). The Lorenz zonoid is defined in Koshevoy and Mosler [1996] as the collection of all such shares for each fraction of the population. It is a convex region in $[0,1]^2$ bounded below by the Lorenz curve and above by the reverse Lorenz curve. More precisely, the Lorenz zonoid is defined as the set of points

$$L(Y) \;:=\; \left\{ \left(\int_0^{\infty} \phi(v)\, dF_Y(v)\,,\, \frac{1}{\mu_Y} \int_0^{\infty} v\phi(v)\, dF_Y(v) \right) \;:\; \phi \in \Phi \right\},$$

where the function ϕ ranges over the set Φ of measurable functions from $\mathbb{R}_+$ to $[0,1]$. The lower (resp. upper) bound is obtained with the collection of functions $\phi(v) := \mathbb{1}\{v \leq y\}$ (resp. $\mathbb{1}\{v > y\}$), $y \in \mathbb{R}_+$.

Since the definition of the Lorenz zonoid does not rely on ranks or quantiles, the extension to higher dimensions is straightforward. Let Φ now be the set of measurable functions from $\mathbb{R}^d_+$ to $[0,1]$ and $X = (X_1, \ldots, X_d)$ be a multivariate allocation with CDF F and mean $(\mu_1, \ldots, \mu_d)$. The Lorenz zonoid of Koshevoy and Mosler [1996] is defined as the set of points $L(X) :=$

$$\left\{ \left(\int_0^\infty \phi(u)\, dF(u)\,,\, \frac{1}{\mu_1} \int_0^\infty u_1 \phi(u)\, dF(u)\,,\, \cdots\,,\, \frac{1}{\mu_d} \int_0^\infty u_d \phi(u)\, dF(u) \right) \,:\, \phi \in \Phi \right\}.$$

The Lorenz zonoid is an American football-shaped region in $[0,1]^{d+1}$ with poles at points $(0, \ldots, 0)$ and $(1, \ldots, 1)$. The Lorenz surface of Taguchi (1972a,1972b) is a subset of the Lorenz zonoid, obtained when Φ is restricted to the set of functions $\phi_x(\cdot) := \mathbb{1}\{\cdot \leq x\}$, all $x \in \mathbb{R}^2$. A function $\phi \in \Phi$ defines a point in the zonoid. The interpretation is simple in the case of indicator functions. The latter pick out specific groups of individuals in the population and the corresponding point in the zonoid has first coordinate equal to the fraction of the population involved. The other coordinates are the shares of each of the resources held by this group of individuals.

Definition 4.1 of Banerjee [2016] proposes a multivariate inequality ordering in the finite population case. The analogue multivariate Lorenz map in the general case of a (possibly mixed discrete continuous) vector allocation X with normalized version $\widetilde{X}$ can be defined for all $r = (r_1, \ldots, r_d) \in [0,1]^d$ by

$$L^B(r) \;:=\; \left(\int_0^{r_j} Q^{B,j}(u)\, du \right)_{j=1}^{d},$$

where, for each $j = 1, \ldots, d$, and $Q^{B,j}$ is the quantile function associated with the random variable $\Sigma_{k=1}^d (\widetilde{X}_j + \widetilde{X}_k)/2d$. If the latter were replaced by $\widetilde{X}_j$, the Lorenz map would be the vector of marginal Lorenz curves. Mixing with the average across allocation introduces sensitivity to dependence between the marginal allocations. Note however,

that $Q^B := (Q^{B,j})_{j=1}^d$ is not a valid vector quantile for $\widetilde{X}$, since $Q^B(U)$ is not distributed like $\widetilde{X}$ when U is uniform on $[0,1]^d$. As a result, $L^B(r)$ is not a vector of resource shares, as is the case for the marginal Lorenz curve.

Properties of the Lorenz map and Inverse Lorenz function

The proposed multivariate extensions of the Lorenz curve in both Taguchi (1972a,1972b) and Koshevoy and Mosler [1996] relate population proportions to a vector of resource shares. Our proposal differs substantially from these in that it directly relates a specific subset of the population, namely individuals with multivariate rank below r to their share of both resources. Beyond this major conceptual difference, we now investigate properties of our multivariate extension of the Lorenz curve that make it a valuable contribution.

Interpretation. Unlike other multivariate proposals, the Lorenz map shares the interpretation of the traditional Lorenz curve as the cumulative share of resources held by the lowest ranked individuals.

Computation. As shown in Section 2.2.2, the Lorenz map can be efficiently computed as a convex program without imposing parametric form. As an integrated vector quantile, it also relies on the growing literature on computational geometry and computational optimal transport, where algorithms and implementations abound and are tested in a variety of applied fields. This is in sharp contrast with the other proposals, specifically the Lorenz zonoid proposed by Koshevoy and Mosler [1996] which is notoriously difficult to compute.

Statistical inference. As an integrated quantile, the Lorenz map is amenable to statistical inference. The convergence of sample analogues of vector quantiles to their theoretical counterpart was shown in Chernozhukov et al. [2017] and Figalli [2018]. Vector ranks are distribution free and can be used in rank based statistical procedures that emulate scalar rank-based inference, as shown in Deb and Sen [2023], Ghosal and Sen [2022] and

Shi et al. [2022]. See the survey in Hallin [2022] and references within.

Uniqueness. The Lorenz map characterizes the distribution of the allocation it is associated with. This property is shared with the Lorenz zonoid of Koshevoy and Mosler [1996], but not the other alternative proposals in the literature.

Proposition 1. The Lorenz map $\mathcal{L}_X$ characterizes the distribution of X in the sense that X and $\widetilde{X}$ are identically distributed if and only if $\mathcal{L}_X = \mathcal{L}_{\widetilde{X}}$.

Lorenz curve as a CDF. The Lorenz map is a map from $[0,1]^d$ to $[0,1]^d$. Hence, unlike the traditional scalar Lorenz curve, it cannot be a CDF. However, the Inverse Lorenz Function is the cumulative distribution function of a random vector on $[0,1]^d$ by construction. This property is not shared by the alternative proposals in the literature.

Decomposition under independent attributes. As shown in Example 3, the Lorenz map reduces to a simple function of the marginal Lorenz curves in case marginal attribute allocations are mutually independent. This feature is shared with the multivariate Lorenz proposal in Arnold (1983,2012) but not the alternative proposals.

Dominance of egalitarian allocations. In the univariate case, the Lorenz curve of the identical allocation $Y = 1$ almost surely, is $L_Y(q) = q$, which is sometimes called the egalitarian line. The Lorenz curve of any other allocation $Y \geq 0$ is below the egalitarian line, i.e., $L_Y(q) \leq q$, for all $q \in [0,1]$. For $d > 1$, the identical allocation of Example 2 is a direct extension of the univariate notion of egalitarian. We show here, that the Lorenz map and Inverse Lorenz Function of the identical allocation provide similar bounds in the multi-attribute case. For this, we require allocations with components that display a form of positive association defined in Assumption 1.

Assumption 1. *The vector quantile $Q_{\widetilde{X}} := (Q_1, \ldots, Q_d)$ of $\widetilde{X}$ is such that, for each j,*

$$\mathbb{E}\left[Q_j(U_1, \ldots, U_d) \,|\, U_k = u_k, \text{all } k \neq j\right]$$

is monotonically increasing in each of the u_k, $k \neq j$, where the vector $(U_1, \ldots, U_d)$ is uniform on $[0,1]^d$.

This assumption imposes a type of positive dependence between the components of X through their ranks. More precisely, Assumption 1 imposes a form of *positive regression dependence*, as in Lehmann [1966], between one resource and the others' ranks. For allocations satisfying Assumption 1, we show that Lorenz map and Inverse Lorenz Function of the identical allocation serve as upper and lower bounds, respectively.

Proposition 2. The Lorenz map of any allocation X satisfying assumption 1 is component-wise dominated by the Lorenz map of the identical allocation in example 2. Moreover, the Inverse Lorenz Function of allocation X is bounded below by the Inverse Lorenz Function of the identical allocation.

We argue in Appendix A.5.2 that defining egalitarianism solely by identical allocations is too restrictive in the case of multiple resources. In case $d = 2$, we show that a much larger class of allocations have Lorenz maps dominated by an egalitarian allocation from Definition 9, which includes the identical allocation.

2.3 Multi-attribute inequality comparisons

We can use the vector Lorenz map $\mathcal{L}_X$ of an allocation X introduced in Section 2.2.1 as a tool to compare inequality of different allocations. We base an inequality dominance criterion to compare different allocations on the dominance of Lorenz maps. We develop a visualization tool for inequality dominance, and an inequality index for the cases where Lorenz maps cross.

2.3.1 Lorenz dominance

Consider two allocations X and X', with respective Lorenz maps $\mathcal{L}_X$ and $\mathcal{L}_{X'}$. If $\mathcal{L}_X(r) \geq \mathcal{L}_{X'}(r)$ for some vector rank r, the same proportion of the population with vector ranks below r commands a larger share of all resources in allocation X than in allocation X'. If this is true for any vector rank r in $[0,1]^d$, then, we say that allocation X' is more unequal than allocation X.

Definition 4. An allocation X' is said to be more unequal in the Lorenz order than an allocation X if $\mathcal{L}_X(r) \geq \mathcal{L}_{X'}(r)$ for all $r \in [0,1]^d$. We denote this $X \succcurlyeq_{\mathcal{L}} X'$.[6]

The Lorenz partial order of Definition 4 is an implementable dominance criterion: The Lorenz maps can be computed and compared. The relation $X \succcurlyeq_{\mathcal{L}} X'$ is equivalent to stochastic dominance of the random vector $\mathcal{L}_X(U)$, with $U \sim U[0,1]^d$, over $\mathcal{L}_{X'}(U)$ (see section 3.8 of Müller and Stoyan [2002]). Hence, dominance tests can be derived on the basis of sample analogues of the Lorenz maps to emulate the large literature on inference techniques to compare inequality of distributions of a single attribute. See Davidson and Duclos [2000] and references within.

Following the literature on the measurement of inequality, we assess the value of this implementable dominance criterion for inequality comparisons in two ways. First, we analyze the class of social evaluation functionals that are compatible with the Lorenz order, and show that they are rank-dependent social evaluation functions, with weights decreasing in rank. Second, we identify the class of transfers that increase inequality as defined by this Lorenz criterion.

Rank-dependent social evaluation functionals

The first way to gain insight into the relevance of our multivariate Lorenz dominance criterion is to characterize the set of social evaluation functionals that are compatible with it. A social evaluation functional is a map S from an allocation X, i.e., a random vector in $\mathbb{R}^d_+$, to $\mathbb{R}$, which orders allocations in their social desirability. A social evaluation functional S is compatible with the dominance criterion if $X \succcurlyeq_{\mathcal{L}} X' \Rightarrow S(X) \geq S(X')$. Compatibility with Lorenz dominance is a form of inequality aversion of the social evaluation functional, since more equal allocations are deemed socially more desirable.

[6] As a partial ordering based on cumulative sums of vector quantiles, the relation $X \succcurlyeq_{\mathcal{L}} X'$ is a multivariate extension of the concept of majorization of Hardy et al. [1934]. It is different from existing multivariate notions of majorization reviewed in Marshall et al. [2011] and Arnold and Sarabia [2018], in that it relies on a multivariate reordering of the random vector allocation.

By construction, a social evaluation functional that is compatible with the Lorenz dominance order must satisfy anonymity and ratio-scale invariance. Anonymity, also called law-invariance or symmetry in the literature, refers to the fact that $S(X) = S(X')$ whenever X and X' are identically distributed. The identity of individuals does not matter in the social evaluation, so that a permutation of individuals in the population leaves S unchanged. Ratio-scale invariance refers to the fact that $S(\lambda' X) = S(X)$ for any positive vector λ. Hence, the social evaluation is not affected by a change in units of measurement.

Next, and more substantively, all social evaluation functionals that are compatible with the Lorenz dominance criterion are rank-dependent social evaluation functions. Individuals are weighted in the social evaluation according to their rank in the distribution. To define and formalize this statement, start with the case of a single attribute. Weymark [1981] shows that social evaluations that satisfy the comonotonic independence property defined below take the form of weighted sums of quantiles.

Property CI (Comonotonic Independence). A social evaluation functional S is said to satisfy comonotonic independence if, whenever X, X' and Z are comonotonic allocations, and $S(X) \geq S(X')$, then, for all $\mu \in (0,1)$, $S(\mu X + (1-\mu)Z) \geq S(\mu X' + (1-\mu)Z)$.

In the univariate case, two allocations are called comonotonic if one is a positive increasing function of the other. In other words, individuals are ranked identically in both allocations. Comonotonic independence means that the comparison of two allocations with a common component only depends on the comparison between the two variable components, as long as rankings stay unchanged. As an illustration, when assessing the effect on household income distributions of a policy that only affects women, under perfect assortative matching, one need only look at the change in the distribution of women's income.

The same property of comonotonic independence can be entertained in the multi-attribute case, with the same interpretation. Two allocations are comonotonic if individuals are ranked identically in both allocations. Now, in case of random vectors X and X',

comonotonicity is defined in the same way by the fact that X and X' have the same vector ranks. The following definition follows Galichon and Henry [2012] and Ekeland et al. [2012], where it is called μ-comonotonicity[7] of $\widetilde{X}_1, \ldots, \widetilde{X}^J$.

Definition 5 (Vector comonotonicity)**.** Random vectors $X^1, \ldots, X^J$ on $\mathbb{R}^d$ are said to be *comonotonic* if there exists a uniform random vector U on $[0,1]^d$ such that $\widetilde{X}^j = Q_{\widetilde{X}^j}(U)$, $j = 1, \ldots, J$, almost surely, where Q_X is the vector quantile of Definition 1 associated with the distribution of X, and $\widetilde{X}$ is the component-wise demeaned version of X.

With this definition of comonotonicity (which coincides with the usual definition in the single attribute case), comonotonic independence of a social evaluation function is still defined as property CI. If individuals are ranked identically in allocations X and X', and X' is socially less desirable than X, then, adding to both X and X' a third common allocation Z cannot reverse the ordering, if Z ranks individuals as X and X' do.

As in Weymark [1981] for the single attribute case, Galichon and Henry [2012] show that comonotonic additive social evaluation functions are rank dependent, i.e., of the form

$$S_\phi(X) \;:=\; \int_{[0,1]^d} \phi(u)' Q_{\widetilde{X}}(u)\, du, \tag{2.15}$$

for some function $\phi : [0,1]^d \to \mathbb{R}^d_+$. To each vector rank u, ϕ associates the attribute-specific weights of ranked u individual in the social evaluation. We show that social evaluation functionals are only compatible with the Lorenz dominance order if they satisfy comonotonic additivity, hence if they are rank-dependent social evaluation functions.

Inequality aversion of a rank dependent social evaluation functional implies a weighting scheme that gives more weight to lower ranked individuals. In the scalar case analyzed in Weymark [1981], an inequality averse rank dependent social evaluation func-

[7] See Puccetti and Scarsini [2010] for a discussion of this and other multivariate comonotonicity concepts.

tional is characterized by decreasing weights as ranks increase. We show a similar result in the multivariate case. Social evaluation functionals that are compatible with the Lorenz order of Definition 4 are rank dependent social evaluation functionals with rank-specific weights of the form

$$\phi_m(u) := \left(\int_{[0,1]^d} \mathbb{1}\{u \leq r\}\, dm_1(r), \ldots, \int_{[0,1]^d} \mathbb{1}\{u \leq r\}\, dm_d(r) \right)', \tag{2.16}$$

where m_j is a non negative measure on $[0,1]^d$, all $j \leq d$.

Proposition 3. A social evaluation functional is compatible with the Lorenz dominance order of Definition 4 if and only if it is of the form

$$S(X) := \int_{[0,1]^d} \phi_m(u)' Q_{\widetilde{X}}(u)\, du \tag{2.17}$$

A special case of weighting scheme satisfying Proposition 3 is the case $m_j = \delta_r$ all j, where all individuals below rank r receive weight 1 and all individuals above rank r receive weight 0. More generally, individuals can be given different weights for different resource dimensions, but as non negative mixtures of the indicators $\mathbb{1}\{u \leq r\}$, the weights are always decreasing in ranks.

Increasing marginal inequality and increasing correlation

The second way we evaluate our Lorenz dominance criterion is by identifying transfers of resources between individuals that increase inequality according to this criterion. Since inequality is a cardinal aspect of the distribution, we consider a class of transfers that preserves the multivariate ranks. The transfers we consider are functions $T : [0,1]^d \to \mathbb{R}^d$. If the j-th component of transfer T is positive (resp. negative), it is added to (subtracted from) the endowment in resource j of individual with rank u.

Definition 6. A rank preserving transfer from allocation X to allocation X' is a transfer such that pre-transfer and post-transfer allocations are comonotonic (individuals preserve the same rank). Equivalently, it is a function $T : [0,1]^d \to \mathbb{R}^d$ such that $Q_{\widetilde{X}'}(u) =$

$Q_{\widetilde{X}}(u) + T(u)$ for all $u \in [0,1]^d$, where Q_X is the vector quantile of Definition 1 associated with the distribution of X, and $\widetilde{X}$ is the component-wise demeaned version of X.

First we show that the transfers that increase inequality according to the Lorenz criterion are the arbitrary combinations of rank preserving transfers of a non negative quantity of one of the resources from an individual with rank u_1 to an individual with rank $u_2 \geq u_1$.

Proposition 4. An allocation X' is more unequal than X, i.e., $X' \preccurlyeq_{\mathcal{L}} X$, if and only if an allocation with the same distribution as X' can be obtained from X via an arbitrary sequence of rank preserving transfers T such that for all $r \in [0,1]^d$,

$$\int_{[0,1]^d} T(u) \mathbb{1}\{u \leq r\} \, du \leq 0. \tag{2.18}$$

The inequality in (2.18) expresses the fact that mass is transferred from lower ranked to higher ranked individuals.

A desirable feature of the Lorenz inequality ordering of Definition 4 is its ability to rank two allocations X and X', when the latter is obtained from the former through a transfer that increases inequality of the marginals or that increases the degree of positive dependence between the marginals. We formalize this feature with a specific type of multivariate transfer we call *Monotone Regressive Transfers*. We specialize the discussion to bivariate allocations to avoid wading into concepts of increasing multivariate dependence when $d > 2$.

Definition 7 (Monotone Regressive Transfer, MRT)**.** A transfer $T : [0,1]^2 \to \mathbb{R}^2$ is a *monotone regressive transfer* if T is rank preserving and has non-negative Jacobian (i.e., the Jacobian's entries are all non negative).

In the univariate case, a monotone regressive transfer reduces to a monotone mean preserving spread (Quiggin [1992]), also called Bickel-Lehmann increase in dispersion (Bickel and Lehmann [1976]). In the multivariate case[8], monotone regressive transfers

[8] A related extension in the theory of multivariate risks was proposed in Charpentier et al. [2016].

weakly increase both marginal inequality and positive dependence. The former happens because each component of the transfer has non negative own derivative, hence is increasing in each component of the rank. The latter happens because the transfer has non negative cross derivative, hence increases the degree of positive dependence between the two resources.

Proposition 5 (Monotonicity in MRT)**.** If an allocation X' is obtained from an allocation X through a monotone regressive transfer, then $X \succcurlyeq_{\mathcal{L}} X'$, i.e., X' is more unequal than X as defined by the Lorenz dominance partial order of Definition 4.

Proposition 5 shows that the multivariate Lorenz dominance order of Definition 4 therefore ranks an allocation as more unequal if the marginal resource allocations are weakly more unequal and if the marginal resource allocations are weakly more positively dependent. This is in contrast with the Lorenz dominance order based on inclusion of Lorenz zonoids proposed in Koshevoy and Mosler [1996]. Indeed, by Proposition 8 in Koshevoy and Mosler [2007], if two allocations have identical marginals, and X dominates X' in the Lorenz dominance order based on Lorenz zonoid inclusion, then X and X' are identically distributed.

The Lorenz dominance ordering of Definition 4 does not satisfy the *uniform majorization principle* proposed by Kolm [1977]. In case of discrete populations, the uniform majorization principle stipulates that inequality should be reduced through multiplication by a doubly stochastic matrix different from a permutation. However, as Dardanoni [1993] points out, such transformations can increase correlation and therefore increase inequality in an egregious way. See the discussion in Savaglio [2006]. We show that a similar issue arises with the continuous version of the uniform majorization principle. The latter requires an inequality dominance order to be monotonic with respect to the concave order. From Theorem 4 of Galichon and Henry [2012], we deduce that a social evaluation functional that satisfies uniform majorization and comonotonic independence

must be equal to

$$S_{UM}(X) := 1 - \int_{[0,1]^d} u' Q_{\widetilde{X}}(u)\, du \tag{2.19}$$

up to an affine transformation. We show in Appendix A.5.3 that in the case of bivariate allocation $X = (X_1, X_2)$, S_{UM} is minimized when the two components X_1 and X_2 of allocation X are independent, which runs against the intuition that increased dependence can increase inequality[9].

2.3.2 *Visualization of Lorenz dominance*

Failures of Lorenz dominance can be visualized with the Inverse Lorenz function. Consider two allocations X and X', with respective Inverse Lorenz Functions l_X and $l_{X'}$. If $l_X(z) \leq l_{X'}(z)$ for some vector of shares z, a larger proportion of the population commands the same share of resources in allocation X' than in allocation X. Lorenz dominance of allocation X over allocation X' (in the sense of Definition 4) implies that the relation $l_X(z) \leq l_{X'}(z)$ holds for each resource share vector z (see Proposition 11 in the appendix).

In case of bivariate allocations, the latter can be easily visualized on $[0,1]^2$ through the relative positions of the level sets of the Inverse Lorenz function, which we call α-Lorenz curves, denoted

$$\mathcal{C}_X^\alpha := \{z \in [0,1]^2 : l_X(z) = \alpha\}, \text{ for each } \alpha \in (0,1).$$

The α-Lorenz curves provide a visualization of Lorenz dominance. We can compare the inequality of different allocations based on the shape and relative positions of their respective α-Lorenz curves. Suppose X is less unequal in the Lorenz dominance order than X'. Then, by Proposition 11, for any $z \in \mathcal{C}_{X'}^\alpha$, $l_X(z) \leq l_{X'}(z) = \alpha$. So $z \in \mathcal{C}_X^{\widetilde{\alpha}}$

[9]Note that we are considering inequality over outcomes, not welfare inequality. Hence, the point made in Atkinson and Bourguignon [1982], that increased correlation may decrease utilitarian welfare inequality when resources are complements, doesn't apply here.

with $\widetilde{\alpha} \leq \alpha$. This can be visualized as a shift to the north-east of the α-Lorenz curves of the more unequal allocation X' to the α-Lorenz curves of the less unequal allocation X.

Proposition 6. (i) The α-Lorenz curves $\mathcal{C}_X^\alpha$ are the level curves of a bivariate cdf, hence they are downward sloping, non decreasing in α and they do not cross. In addition, (ii) The α-Lorenz curves $\mathcal{C}_X^\alpha$ are convex if

$$\frac{\partial l_X}{\partial z_2}\frac{\partial l_X^2}{\partial z_1 \partial z_2} - \frac{\partial l_X}{\partial z_1}\frac{\partial l_X^2}{\partial z_2^2} \geq 0.$$

Visually, inequality can be assessed by the departure of α-Lorenz curves from those of the identical allocation. This visual comparison is facilitated by the fact that they are shaped like indifference curves. In addition, correlation information is preserved through the curvature of the α-Lorenz curves, which decreases with dependence.

Figure 2.2 and 2.3 displays α-Lorenz curves of the multivariate lognormal allocation X such that

$$\ln X \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\sigma_1^2 & \rho\\ \rho & \sigma_2^2\end{bmatrix}\right) \tag{2.20}$$

where $\sigma_1, \sigma_2 > 0$ control the dispersion of the respective marginals and ρ tunes the degree of dependence.

In Figure 2.2, we revisit the three special cases of identical allocations, independent attributes and comonotonic attributes. We also add the countermonotonic case, where individuals are ranked in opposite order for the two resources, as well as intermediate dependence cases. Figure 2.2 shows the α-Lorenz curve (with $\alpha = 0.9$) of the identical allocation compared to multivariate lognormal allocations with variances $\sigma_1 = \sigma_2 = 1$ and correlation coefficients $\rho = -0.99$ (countermonotonic), $\rho = -0.6, -0.3$, $\rho = 0$ (independent), $\rho = 0.3, 0.6$, and $\rho = 0.99$ (comonotonic). Figure 2.3 compares α-Lorenz curves of 6 different allocations that are multivariate lognormally distributed as in (2.20), for $\alpha = 0.9$. The parameters σ_1^2, σ_2^2 take values 1 or 2, whereas ρ takes values 0.2 or 0.8. In case of marginals with different σ, the asymmetry is reflected in the α-Lorenz curves.

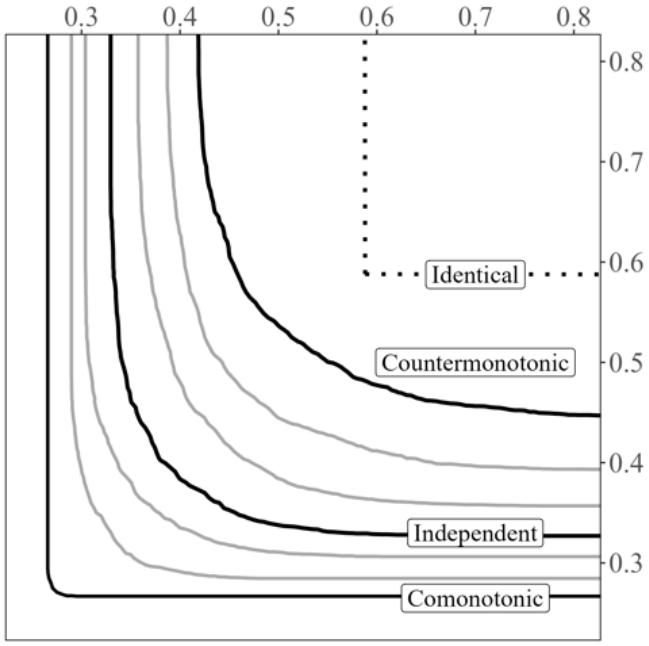


Figure 2.2: 0.9-Lorenz curves of multivariate lognormal random vectors defined in (2.20). The scale parameters are $\sigma_1 = \sigma_2 = 1$. The curves correspond to different correlation coefficient ρ (from bottom left to top right: $0.99, 0.6, 0.3, 0, -0.3, -0.6, -0.99$). The nested curves show increasing dependence between fixed marginals increases inequality.

Moreover, other things equal, inequality increases with σ, which measures inequality in the marginals, and with ρ, which measures correlation. Finally, Figure 2.4 shows an example of two multivariate lognormally distributed allocations X and X' such that the marginals of X are more unequal than those of X', but X' is more unequal overall due to positive dependence of its marginals. Specifically, the marginals of X are independent and have the same scale parameter $\sigma^2 = 1.2$, while X' has marginals with correlation parameter $\rho = 0.9$ and scale $\sigma^2 = 1$. This shows that the Lorenz inequality dominance ordering does not imply dominance of the marginals, and how it can instead incorporate trade-offs between marginal inequality and dependence.

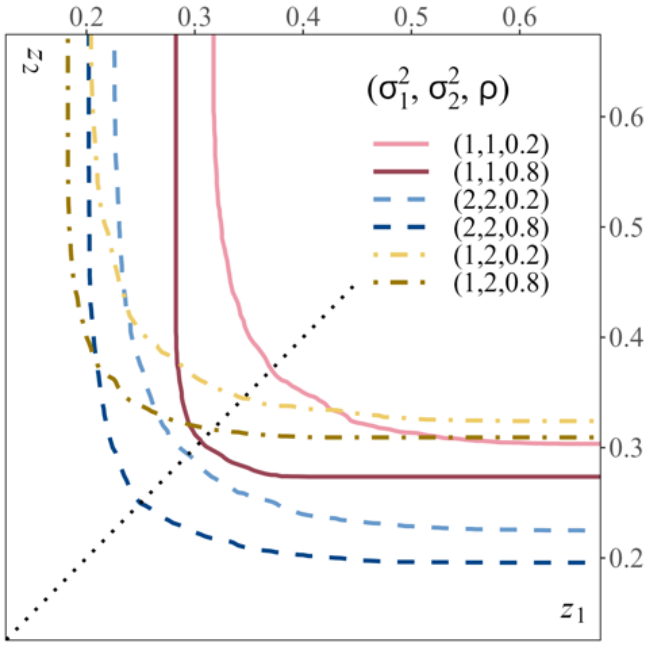


Figure 2.3: 0.9-Lorenz curves of multivariate lognormal random vectors that vary by correlation and scale parameters defined in (2.20). Notably the dotted-dashed lines display when $\sigma_1 \neq \sigma_2$ showing a skewness towards the axis z_2 of the marginal with larger value. The diagonal black dotted line mark where the identical allocation lies and helps to show skewness towards an axis.

2.3.3 Multivariate Gini inequality index

The Lorenz dominance ordering is a partial ordering of multivariate distributions. In many cases, α-Lorenz curves may cross. For a complete inequality ordering, we also propose an extension of the classical Gini index to compare inequality in multi-attribute allocations. Gajdos and Weymark [2005] propose a multivariate Gini coefficient based on aggregation across individuals first, then across dimensions, which removes the effect of dependence across attributes. Decancq and Lugo [2012] propose to aggregate across dimensions first, then across individuals, in order to keep track of correlation. From the

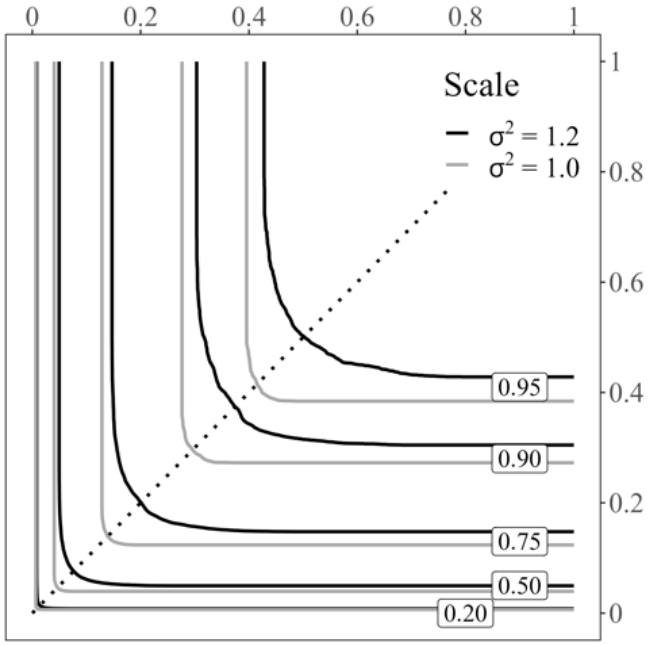


Figure 2.4: α-Lorenz curves of X with independent lognormal marginals with $\sigma^2 = 1.2$ in black and X' with dependent lognormal marginals with $\sigma^2 = 1$ and correlation $\rho = 0.9$ in gray.

volume of the Lorenz zonoid, a multidimensional Gini coefficient can be derived naturally as in Koshevoy and Mosler [1997]. An alternative strategy is followed by Arnold [1983], Koshevoy and Mosler [1997], who extend the definition based on the sum of all distances between pairs of individuals.[10]

The univariate Gini index can be interpreted as the average deviation from the egalitarian allocation, the univariate version of our identical allocation. We emulate this interpretation and define a multivariate Gini based on an average deviation from the

[10]Other multivariate Gini indices based on multivariate Lorenz curve proposals include Banerjee [2010], Grothe et al. [2022], and Sarabia and Jorda [2020].

Lorenz map $r \mapsto (r_1 r_2 \cdots r_d, \ldots, r_1 r_2 \cdots r_d)$. The deviation measure we propose is

$$\sum_{j=1}^{d} \int_{[0,1]^d} \left[\prod_{k=1}^{d} r_k - \mathcal{L}_j(r) \right] dr, \tag{2.21}$$

where $\mathcal{L}_j$ is the j-th component of the Lorenz map $\mathcal{L}_X$, with $j = 1, \ldots, d$. After normalization, (2.21) becomes

$$G(X) \;=\; 1 - \frac{2^d}{d} \left(\int_{[0,1]^d} \sum_{j=1}^{d} \mathcal{L}_j(r)\, dr \right), \tag{2.22}$$

which yields the following definition.

Definition 8 (Gini Index)**.** (2.22) defines the multivariate Gini index of allocation X.

The traditional Gini index of a univariate allocation can also be characterized as a weighted sum of outcomes, where the weights are increasing linearly in the rank of the individual in the population. Hence, the negative of the Gini, seen as a social evaluation functional displays inequality aversion by giving more weight to the outcomes of lower ranked individuals than to those of higher ranked ones.

The same interpretation is valid for our multivariate Gini. Specifically,

$$\begin{aligned} S(X) \;&:=\; \frac{d}{2^d}\,(1 - G(X)) \\ &=\; \int_{[0,1]^d} \left(\int_{[0,1]^d} (\mathbb{1}\{u \le r\}, \ldots, \mathbb{1}\{u \le r\})'\, Q_X(u)\, du \right) dr \\ &=\; \int_{[0,1]^d} \left(\prod_{j=1}^{d} (1 - u_j), \ldots, \prod_{j=1}^{d} (1 - u_j) \right)' Q_X(u)\, du \end{aligned} \tag{2.23}$$

is an inequality averse social evaluation functional of the form (2.17) with uniform measures on $[0,1]^d$ in (2.16). So $S(X)$ evaluates the social desirability of an allocation with a weighted sum of individual endowments, where the weights $\Pi_{j=1}^{d}(1 - u_j)$ are component-wise decreasing in the individual's rank $u = (u_1, \ldots, u_d)$.

The Gini index correspondingly takes the form $G(X) = 1 - (2^d/d)S(X)$, with S as in (2.23). For instance, in the case of bivariate allocations, (2.22) takes the form

$$G(X) \;=\; 2 \int_0^1 \int_0^1 (u_1 + u_2 - u_1 u_2)\, (Q_1(u) + Q_2(u))\, du_1 du_2 \;-\; 3. \tag{2.24}$$

In expression (2.24), $Q_1(u) + Q_2(u)$ is the sum of the two resource allocations of the individual in the population with vector rank (u_1, u_2). Hence, the Gini index is indeed a weighted sum of outcomes, with weights $(u_1 + u_2 - u_1u_2)$ increasing with the vector ranks (u_1, u_2). It is a genuinely multivariate extension in that the weighting scheme, hence the social evaluation of inequality, depends on multivariate ranks of individuals.

Examples 1 continued The Gini coefficient for discrete distributions can be computed by plugging in the Lorenz map from Section 2.2.2 into (2.22).

Examples 3 and 4 continued We compare Gini indices in the independent case with the perfect comonotonicity case, where X_1 and X_2 have the same (uniform on $[0, 2]$) marginal distributions. We verify (analytically for $r_1 + r_2 < 1$ and numerically using R for $r_1 + r_2 \geq 1$) that $Q_1(u) + Q_2(u)$ is smaller in the comonotonic case, than in the independent case. Hence the Gini index (and the measure of inequality) is larger in the comonotonic case.

Example 5 (Countermonotone Resources)**.** If we have $X_1 + X_2 = 2$ a.s., then $Q_1(u) + Q_2(u) = 2$ for almost all u, and we obtain $\mathcal{L}_1(r_1, r_2) + \mathcal{L}_2(r_1, r_2) = 2r_1r_2 \geq r_2L_1(r_1) + r_1L_2(r_2)$, so that, in particular, the Gini index in the countermonotone case is the same as in the case of the identical allocation, i.e., equal to 0, and both are smaller than the Gini of the allocation with independent resources. This is consistent with the fact that these allocations X are considered egalitarian according to Definition 9 in Appendix A.5.2.

The Gini index of definition 8 is in $[0, 1]$ under Assumption 1. It equals 0 for the identical allocation. It tends to 1, when the Lorenz map tends to 0 (extreme inequality). The Gini index of an allocation with independent components reduces to the average of classical scalar Gini's of both components. Like the classical scalar Gini index, it preserves the Lorenz inequality ordering, in the sense that higher inequality according to $\succcurlyeq_{\mathcal{L}}$ implies a larger value of the Gini index. In other words, $X \succcurlyeq_{\mathcal{L}} X'$ implies $G(X) \leq G(X')$, so that the negative of the Gini is a compatible social evaluation functional. Hence it inherits the properties of anonymity, scale invariance and comonotonic independence.

Multivariate S-Gini

The multivariate Gini in expression (2.22) is the suitably normalized negative of an inequality averse social evaluation functional of the form (2.17) with uniform measures on $[0,1]^d$ in (2.16). It can be extended to reflect varying concern for inequality in different attributes. To achieve this, a multivariate Gini coefficient can be defined as $1 - cS(X)$, where c is a normalizing constant and S is a social evaluation functional that reflects different degrees of inequality aversion in different attributes.

In the univariate case, to reflect varying degrees of inequality aversion, Donaldson and Weymark [1980] propose a single parameter family of Gini coefficients, called S-Gini, defined by

$$G_\delta(X) \;:=\; 1 - \delta(\delta-1)\int_{[0,1]}(1-r)^{\delta-2}\mathcal{L}(r)\,dr,$$

where $\mathcal{L}$ is the traditional Lorenz curve, and δ ranges from 1, corresponding to indifference to inequality, to the Rawlesian extreme at the limit $\delta \to \infty$, where only the poorest individual matters[11].

The S-Gini family of Donaldson and Weymark [1980] can be extended to the assessment of multivariate inequality within our framework. Let $\delta = (\delta_1, \ldots, \delta_d)$ be a d-dimensional parameter, where $\delta_j \in [1, \infty)$, $j = 1, \ldots, d$, reflects the concern for inequality in attribute j. We define the family of multivariate S-Gini coefficients of inequality of an allocation X with Lorenz map $\mathcal{L}_X = (\mathcal{L}_1, \ldots, \mathcal{L}_d)$ as

$$G_\delta(X) \;:=\; 1 - c_\delta S_\delta(X),$$

where $c_\delta := 2^{d-1} / \sum_{j=1}^d \delta_j^{-1}$ is a normalizing constant, and S_δ is the social evaluation functional

$$S_\delta(X) \;:=\; \sum_{j=1}^d (\delta_j - 1)\int_{[0,1]^d}(1-r_j)^{\delta_j-2}\mathcal{L}_j(r)\,dr.$$

[11] We were unable to locate a precise statement of this in the literature, so we include it in proposition 12 in the appendix with a proof for completeness.

The normalizing constant c_δ is chosen such that the multivariate S-Gini G_δ lies in $[0,1]$ and is zero in case of the identical allocation. There remains to verify that the social evaluation functional S_δ is indeed of the form (2.17), and hence compatible with the Lorenz order. Indeed, we have

$$S_\delta(X) = \sum_{j=1}^{d} \int_{[0,1]^d} \mathcal{L}_j(r)\, dm_j^{(\delta)}(r),$$

with $m_j^{(\delta)}(r) = \left(\prod_{l=1, l\neq j}^{d} r_l\right) [1-(1-r_j)^{\delta_j - 1}]$, for each $j = 1, \ldots, d$. The multivariate S-Gini thereby incorporates varying degrees of inequality aversion for different attributes. We recover the S-Gini of Donaldson and Weymark [1980] when $d = 1$, and the multivariate Gini of Section 2.3.3 when $\delta_j = 2$, for all $j = 1, \ldots, d$, as desired.

2.4 Empirical Illustration

In this section, we apply our methodology to the analysis of the evolution of income-wealth inequality in the United States between 1989 and 2022, based on the public version of the triennial Survey of Consumer Finances (SCF). Wealth refers to all assets, financial and otherwise. Details of the sampling technique and a discussion of specific features and issues with the data set are given in Appendix A.4. A guide for practical implementation of the computational procedure outlined in Section 2.2.2 is given in Appendix A.1. We refer to inequality displayed by any of our measures as overall inequality, while specific marginal inequality is described as wealth or income inequality.

2.4.1 Income-wealth α-Lorenz curves

Figure 2.5 shows the α-Lorenz curves for $\alpha = 0.6, 0.8, 0.95$ for the years 1989, 2007, 2010, and 2022. There is a general worsening of overall inequality over 3 decades since the curves shift away from the north-east corner. The tight curvature also reflects the positive correlation of income and wealth as in Figure 2.2. Using Figure 2.3 as reference, the

skew towards the wealth axis indicates inequality from the wealth marginal is dominant at these α-levels, as expected.

Moving from 1992 to 2007, the curve pulls disproportionately towards the income axis. This signals a worsening of income inequality that dominates the effect on inequality from the wealth marginal. From 2007 to 2010, inequality recedes while pulling away from the income axis. Onward to 2022, we see a worsening shift that favors the income dimension, but still indicating disproportionately worse wealth inequality. This signals a worsening of inequality in income, improvement in wealth inequality, increase in positive association of the marginals, or all three.

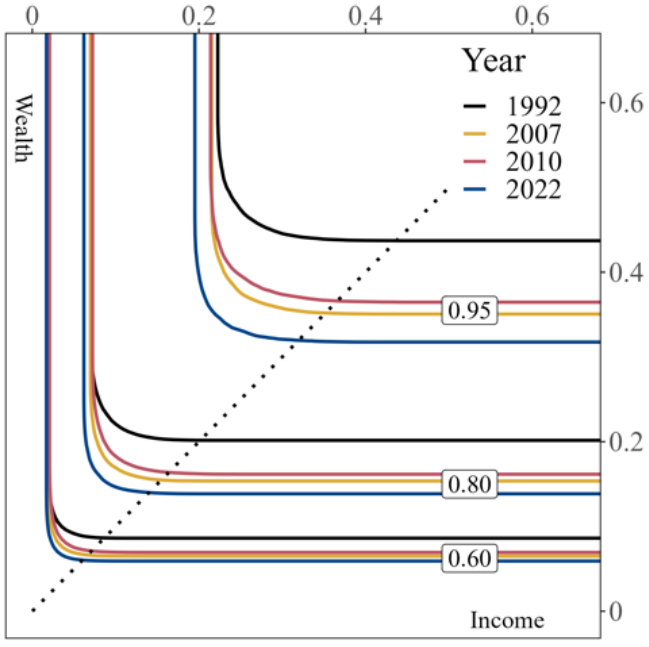


Figure 2.5: α-Lorenz curves ($\alpha = 0.6, 0.8, 0.95$) for US Income-Wealth across select years (1989, 2007, 2010, and 2022).

2.4.2 *Resource shares*

The Lorenz map $\mathcal{L}(r) = (\mathcal{L}_1(r), \mathcal{L}_2(r))$ is the vector of shares of both resources held by the fraction of the population below rank $r = (r_1, r_2)$. It provides insight on the inequality condition of specific strata commonly considered in the literature. Moreover, measuring the share held by the bottom proportions by considering attributes jointly avoids the need to manually, and in an ad hoc manner, match individuals in different strata for each resource to assess inequality. For example, gathering the individuals in the bottom 95% of income that are also in the bottom 50% of wealth.

Figure 2.6 shows resource shares of the 25% fraction of the the population below rank $r = (0.5, 0.5)$ as well as the 90% fraction[12] of the population below rank $r = (0.95, 0.95)$. Consistent with the visual assessment in Figure 2.5, the shares in both resources of the bottom 90% have been steadily declining and the shares of wealth are far less than of income. Between 2007 and 2010, we see income shares increasing relatively more than the decrease in wealth shares. Wealth and income shares both fell from 2010 to 2022, explaining the shift in the curves from Figure 2.5. As for the bottom 25%, changes over time are minor compared to those of the bottom 95%.

2.4.3 *Gini indices*

Figure 2.7 displays the marginal Gini indices for income and for wealth, the multivariate Gini index based on (2.22), as well as Kendall's τ for the dependence between income and wealth over time. Some of the visual indicators are retained in the overall Gini, such as displaying a steady increase in overall inequality. Recall that if the resources were independent, the multivariate Gini would simply be the average of the marginal versions. In this case, the multivariate Gini is heavily skewed towards the marginal Gini of wealth indicating that wealth is a driver of overall inequality along while displaying a positive association between resources since it is far from the average of marginals.

[12]The fraction of the population is exactly $0.95^2 = 0.9025$.

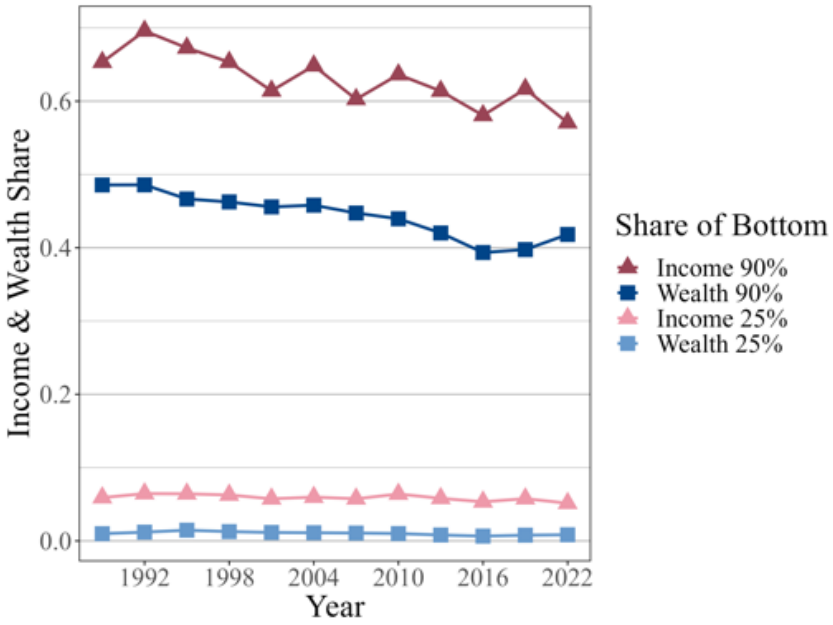


Figure 2.6: The income and wealth shares of the 90.25% and 25% fraction of the population in terms of ranks below $r = (0.95, 0.95), (0.5, 0.5)$, respectively.

The multivariate Gini shows reduced overall inequality between 2007 and 2010. The decreased correlation and income inequality may have been sufficient to offset the rise in wealth inequality.

Inequality analysis across groups can reveal further insights. Figure 2.8 shows Gini indices among White and Black respondents as well as among the working age (younger than 65 years) and retiring age populations. Across racial groups, there is an increasingly large gap in overall inequality that is familiar to the income inequality gaps that have been observed in the literature. While overall inequality has worsened among White respondents, the inequality among Black respondents has remained steady. When comparing inequality across age groups, they both exhibit a steady increase in overall inequality, however the multivariate Gini among the retiring age group inherits the variability in the income marginal.

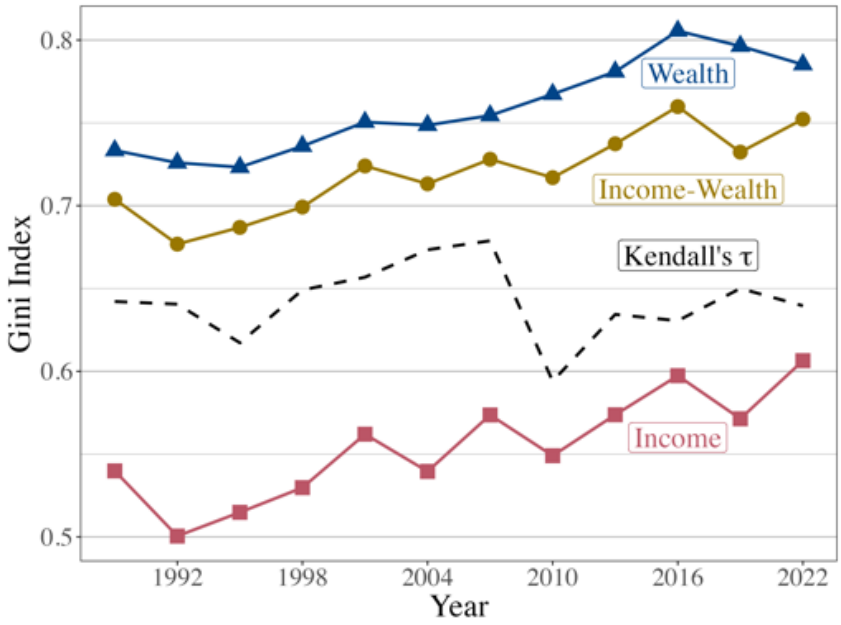


Figure 2.7: Top: Gini indices for income and for wealth, multivariate Gini index, and Kendall's τ (dashed) for US Income-Wealth across 1989-2022.

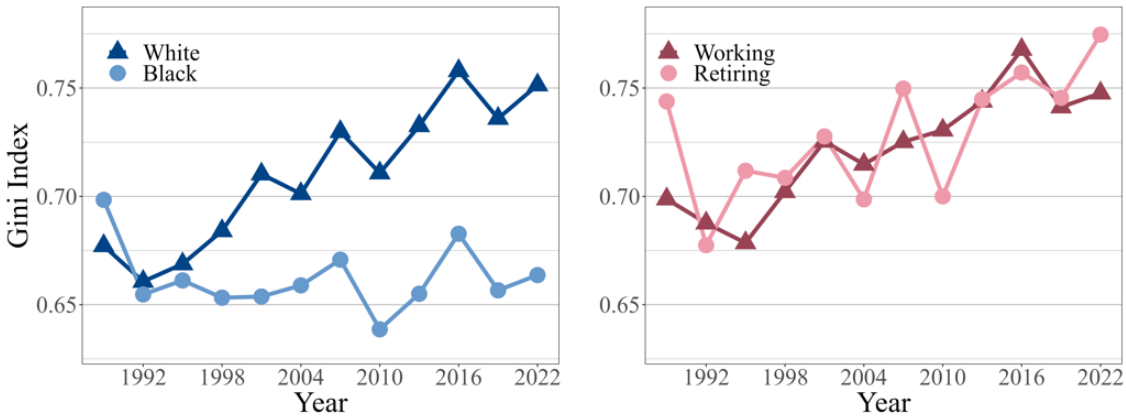


Figure 2.8: Left: Multivariate Gini indices among White and Black respondents. A significant overall inequality gap has formed over time between the two segments. Right: Multivariate Gini indices among working age ($<$ 65 years) and retiring age respondents.

2.5 *Concluding Remarks*

We proposed a new multivariate extension of the Lorenz curve. We propose to emulate the Gastwirth [1971] formulation of the Lorenz curve and define a Lorenz map by integrating vector quantiles of Chernozhukov et al. [2017]. The value of the Lorenz map is a vector of shares of each resource held by the poorer section of the population, as in the scalar case. Dominance of Lorenz maps defines a multi-attribute inequality dominance partial ordering. This Lorenz ordering is, like its scalar counterpart, an implementable criterion to compare inequality in allocations. It is, also like its scalar counterpart, equivalent to preference by any inequality averse rank dependent social evaluation functional. We propose an Inverse Lorenz Function and its level sets as a visualization tool, and apply it to income and wealth in the United States between 1989 and 2022.

Multi-attribute inequality can vary substantially across population groups, as shown in Maasoumi and Racine [2016] within the information theoretic framework of Maasoumi [1986]. There is tension between heterogeneity across covariates and anonymity, according to which inequality measurement should not depend on individual's identities, but only on the distribution of resource allocations. As Kolm [1977] pointed out, this tension is alleviated in part by including more variables in the allocation. This reinforces the motivation for a multidimensional approach to inequality measurement. As for the other potential sources of individual heterogeneity that matters to the social planner, anonymity can be restored by measuring inequality in each subgroup. Beyond this, the conditional approach of Maasoumi and Racine [2016] could be extended with the use of conditional vector quantiles (Carlier et al. [2016]).

Finally, we argue that a formal test of multi-attribute inequality dominance can be based on our Lorenz map, in analogy to dominance testing based on the traditional Lorenz curve in Davidson and Duclos [2000] and references within. The statistical theory for such a test relies on multivariate stochastic dominance testing and the regularity of optimal transport maps, and is left for future research.

Chapter 3

UNOBSERVED GROUPED HETEROSKEDASTICITY AND FIXED EFFECTS

3.1 Introduction

Agents that exhibit unobserved differences often confound the true relationships between variables in economic models. Panel data provides a solution where it is possible to make intuitive, though restrictive, assumptions on the form of heterogeneity such as with two way fixed effects or interactive fixed effects (see Wooldridge [2010] and Bai [2009], respectively). A middle-ground approach is group fixed effects where the unobservable is assumed to be discrete and generates within-group parameters that vary across time. This method requires estimating the unobservable, where ignoring important features of its latent distribution could be harmful.

A linear panel model with group fixed effects is written as

$$y_{it} = x_{it}'\theta + \alpha_{g_i t} + u_{it}. \tag{3.1}$$

The exogenous group membership variable g_i is unobserved, discrete, and takes a known or estimable G number of values. It is assumed that G is relatively small and group membership of individuals does not change over time. The associated group-specific time effects $\alpha_{g_i t}$ are assumed to be exogenous, estimated from the data, and imply that individuals in the same group g follow the same time path of effects α_{gt}. There are no restrictions placed on the relationship between covariates x_{it} and $\alpha_{g_i t}$. A least squares estimator was introduced in Bonhomme and Manresa [2015] and several alternative estimators have since been proposed (Chetverikov and Manresa [2022], Mugnier [2022b]).

The key identification assumption is that groups can be distinguished by the group

fixed effects $\alpha_g = (\alpha_{gt})_t$ for all $g = 1, \ldots, G$. From an error component perspective, $v_{it} = y_{it} - x_{it}'\theta = \alpha_{g_i t} + u_{it}$ has the group fixed effect α_{gt} as the conditional mean given $g_i = g$ so that identification requires the conditional means of the error components to be different, in some sense. In addition to varying by this first moment, the error component may also exhibit heteroskedasticity with respect to the unobserved groups.

Unobserved group heteroskedasticity can be expressed as

$$\mathbb{E}[u_{it}^2 | g_i = g] = \sigma_g^2, \tag{3.2}$$

where $\sigma_g > 0$ for any $g = 1, \ldots, G$. This could arise, using worker's earnings as an example, when there is different variability in wages across groups. If groups are skill tiers, workers that are part of a highly skilled group may display consistently high earnings, with little variation. On the other hand, workers in a lower skilled, but improving skill group may show a larger variation in earnings. This example highlights the importance of the conditional second moment in associating workers to their correct group, and discourages solely relying on α_g to distinguish groups.

In view of this, I extend the group fixed effects model to account for group heteroskedasticity. I introduce the "weighted grouped fixed-effects" (WGFE) estimator, where the name serves as an analogy to the estimators connection to the GFE estimator and resemblance to the weighted least squares estimator. The WGFE estimator simultaneously estimates parameters and the optimal grouping. This is done by alternating between organizing units with outcomes net the effect of covariates into groups that they are most similar to and then updating parameters using that grouping. This describes a version of Lloyd's algorithm (k-means) that I adapted with a modified Mahalanobis distance to measure nearness to the mean and variance of each group.

I provide an asymptotic theory for the WGFE estimator where I allow the number of individuals N and time periods T to approach infinity simultaneously. This estimator shares the property with the GFE estimator that N can grow much faster than T as opposed to the fixed-effects estimator of dynamic linear models, which suffers from bias of

order $1/T$ as N/T converges to a constant (Nickell [1981], Arellano and Hahn [2007a]). As with GFE estimation, to my knowledge this is the first paper to provide asymptotic theory for clustering problems with covariates when clusters differ in variance; see Pollard [1981] and Pollard [1982] for the theory for k-means in the absence of covariates. The WGFE estimator is consistent and asymptotically normal as N/T^{ν} tends to zero for some $\nu > 0$ under conditions presented in Section 3.3, which among them require that groups are *strongly* separated.

Strong separation is a strengthening of the usual identification condition. It requires the distance between each pair of group fixed effects α_g, α_h to not only be non zero, but be larger than a function of the differences between their associated variances σ_g, σ_h. This assumption excludes population groupings where differences in group variances are relatively larger than differences in group means in a way that there is significant overlap. In other words, groups must be mean-separated as a function of relative differences in the group variances. If this is achieved, the quality of group assignments may improve faster than GFE as T grows large, so the WGFE estimator is expected to outperform across most short panel data sets in the finite-sample setting. I present simulation evidence that suggests this: ignoring group heteroskedasticity results in poor estimation of group memberships, leading to severe finite-sample bias of parameter estimators.

The WGFE estimator is equivalent to the concentrated version of the GSR estimator that was concurrently and independently developed by Aguilar Loyo and Boot [2024]. The WGFE estimator is concentrated upon the optimal group variance parameter from the GSR first order conditions. This connection can be used to show that the GFE estimator is equal to the restricted WGFE estimator under the linear restriction where all group variances are the same. Therefore, I propose a quasi-likelihood ratio test of group homoskedasticity provided G is known.

I showcase the WGFE estimator in two illustrations. The first is an application to study the effect of national income on democracy, adding to the insights found by Bonhomme and Manresa [2015] and originally by Acemoglu et al. [2008]. Following Gled-

itsch and Ward [2006] and Ahlquist and Wibbels [2012], it is empirically observed that there are grouped patterns of democratization that transition in both time and space. WGFE bolsters the evidence by finding groupings with more contiguous nations, clearly defined correlation patterns with covariates, and parameter estimates the reflect even less of an effect of income.

The second empirical example revisits the question of the effects of unionization on wages. As with the skill group example from earlier, employers may select more able workers while facing higher labor costs stemming from a union contract. Using WGFE, workers across various unobserved skill groups experience different rates of expansion and deterioration of worker bargaining power and real earnings across the sample period. The union effect is estimated to be about 6%, which lies outside of estimates from pooled OLS (9%) and two way fixed effects (22%).

Section 3.2 presents the issues of identification with different notions of separability, the estimator and computational approach. Section 3.3 develops the asymptotic theory of the infeasible WGFE estimator and the WGFE estimator including consistency of parameter estimators and sample group assignments, and asymptotic normality. In Sections 3.4 and 3.5, I compare the WGFE and GFE estimators in a simulation study and WGFE to study the two empirical applications on unionization-wages and income-democracy.

3.1.1 Related Literature

Modeling with latent group heterogeneity in panel models has received attention as a useful alternative to standard fixed effects approaches. Sun [2005] estimate a multinomial logistic regression with unobserved groupings, but known number of groups G via maximum likelihood. Chang-Ching and Serena [2012] in one of their proposed methods develop a k-means algorithm to group individuals based on the regression estimate of one of G groups. When groupings are known, Bester and Hansen [2016] propose an estimator for a random effects model that assumes individuals share a fixed-effect at

some (undetermined) level of grouping. Hahn and Moon [2010] consider a class of game theoretic models and show that the incidental parameter problem vanishes quickly as N approaches infinity since the number of equilibria are predicted to be finite, hence the support of the fixed-effects is finite. In the case of multidimensional heterogeneity, Cheng et al. [2019] consider individuals who may belong to different groups modeled by several latent variables. None of these papers consider time-varying heterogeneity. Mugnier [2022a] introduce the single index nonlinear GFE model and estimation that does allow for time-varying heterogeneity, but not the possibility of group heteroskedasticity.

When the number of groups are unknown it must be inferred from the data. Bonhomme and Manresa [2015] follow Bai and Ng [2002] and Bai [2009] by using a Bayesian information criteria (BIC) and setting a maximum number of groups. Recently there have been a number of papers that use penalization to classify and estimate parameters. Su et al. [2016] propose the C-Lasso related to group Lasso that classifies and shrinks individual coefficients to the unobserved group coefficients. Liangjun Su [2017] continue with a testing procedure to determine the number of groups based on C-Lasso. Su and Ju [2018] extend the C-Lasso to dynamic panel data models with interactive fixed-effects and cross sectional dependence and propose a BIC to estimate the number of factors and groups. Ando and Bai [2016] consider grouped factor structure and penalization for coefficient estimates and their C_p-type criteria to estimate the number of groups and number of factors in each group. Mehrabani [2022] take these penalization approaches further by allowing the number of groups to diverge along with the sample dimensions. Mugnier [2022b] propose a simple nuclear-norm regularized estimator that estimates the number of groups along with parameters.

Zhang [2020] propose a Bayesian approach that treats the number of groups as a parameter to be estimated and use a Dirichlet process prior that allows for infinitely many groups and allows for group heteroskedasticty, but require strictly normal errors while WGFE imposes no such structure. Along these lines, Kim and Wang [2019] give evidence that in the normal errors case that accounting for group heteroskedasticity in the income

and democracy application shows the correlation between the estimate group variable and income is larger than predicted by GFE estimation and that the total number of groups estimated might depend on the specification of group heteroskedasticity, which coincides with the prediction using WGFE.

Finite mixture models are closely related to our approach, see the textbook by McLachlan and Peel [2004] for a comprehensive review. In these models the data is not required to label individuals into groups and group membership is instead estimated. It is typical to assume that the group distributions belong to the same family e.g. Gaussian mixture models so group heteroskedasticity is modeled. For identifiability of models with covariates there needs to be restrictions the values covariates can take on and on the interaction between covariates and the latent variable; see Kasahara and Shimotsu [2009] for a result on identification of discrete choice models with panel data where both the group distribution and choice probabilities are nonparametrically specified. I allow covariates to be arbitrarily correlated with the latent variable as in typical fixed-effects and leave group membership probabilities unrestricted. Mixture models can incorporate fixed-effects, see Deb and Trivedi [2013] who provide a solution to the incidental parameter problem for Gaussian and Poisson families. Heckman and Singer [1984] apply them to duration models where they take the distribution of the unobservable nonparametrically, while imposing a parametric assumption on the group distributions. A related setting is Markov-Switching models for time series, which can be seen as a mixture model, see Frühwirth-Schnatter [2006]. Related to switching models, k-means clustering is similar to greedy approximations of step function signals, which can be thought of as a estimating a switching model of intercepts with no covariates or parametric assumptions on the error, see Rivero [2023]. In this setting, the user does not need to specify the number of groups and instead could be determined via penalty methods.

The Expectation Maximization (EM) algorithm commonly used in the maximum likelihood estimation of mixture models can be seen as a clustering algorithm, see Redner and Walker [1984]. In fact, in the case of a Gaussian mixture, the Mahalanobis distance

is a key part of the algorithm, which suggests a connection to our clustering algorithm which also incorporates second moment information. Indeed GFE is also the maximizer of the pseudo-likelihood of a Gaussian mixture model, where the mixing probabilities are individual-specific and unrestricted e.g. independent of covariates.

An inspiration for considering unobserved group heteroskedasticity is the observation that the k-means problem of assigning a number of individuals into a finite number of groups; see (MacQueen [1967], Lloyd [1982], Forgy [1965]). The k-means criterion is a least squares criterion, which implicitly assumes that the groups are separated enough in mean and are of identical variance. When the data violates any of these assumptions, then group assignment at the sample level may be compromised. See the recent survey Ahmed et al. [2020] on the performance of the standard k-means algorithm in these contexts.

The estimator is tied to optimal measure transport theory (Villani [2009], Santambrogio [2015]), specifically the Kantorovich distance between probability distributions; see Kolouri et al. [2017] for more on Kantorovich distance and its applications. The Kantorovich distance function defines a distance function between probability measures and so permits a Fréchet mean of distributions that is also itself a distribution known as the Kantorovich barycenter, see Agueh and Carlier [2011]. The WGFE estimator for a model without covariates can be seen as the minimization problem of optimally forming two parameter (location-scale) groups such that these groups have a barycenter with minimum variance over other possible groupings and barycenters.

3.2 The GFE Model with Group Heteroskedasticity

The slope parameters are contained in the vector $\theta \in \Theta \subset \mathbb{R}^p$ and group-specific time-effects $\alpha_{gt} \in \mathcal{A} \subset \mathbb{R}$ for any $g \in \{1, \ldots, G\} := \Gamma_G$ and $t \in \{1, \ldots, T\} := \mathcal{T}$. Denote the vector of time-effects as $\alpha \in \mathcal{A}^{GT}$. The group assignment parameters are categorical: $g_i \in \Gamma_G$ for every $i \in \{1, \ldots, N\} := \mathcal{N}$ and an arbitrary partition using g_i is denoted

as $\gamma = (g_1, \ldots, g_N) \in \Gamma_G^N$. The covariates x_{it} can contain lagged outcomes along with strictly exogenous regressors. The covariates are also allowed to be arbitrarily correlated to the time-effects α_{gt}. Throughout, the absence of a time subscript implies a vector representing a time series. Lastly, $x \mapsto \|x\|$ is the Standard Euclidean norm of a finite-dimensional vector x. True parameter values are denoted with a "0" superscript.

3.2.1 Identification of groups

Identification of group memberships in the GFE model requires that groups are "separated" in the mean squared sense. In other words, groups can be distinguished by the comparison of their group fixed effects α_g. Formally, for any $(g, \widetilde{g}) \in \Gamma_G^2$ such that $g \neq \widetilde{g}$:

$$\operatorname*{plim}_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \left(\alpha_{gt}^0 - \alpha_{\widetilde{g}t}^0\right)^2 = \operatorname*{plim}_{T\to\infty} \frac{1}{T} \left\|\alpha_g^0 - \alpha_{\widetilde{g}}^0\right\|^2 > 0. \tag{3.3}$$

I call this weak separability and is sufficient for designing a rule to assign individuals into groups using their time series information. For any $i \in \mathcal{N}$, suppose that $g_i^0 = g$. Then, the mean squared distance between $y_i - x_i\theta^0$ and α_h^0 for any $h \in \Gamma_G$ is minimized when $h = g$. Recall the true model (3.1) and consider the following:

$$\begin{aligned}
&\operatorname*{plim}_{T\to\infty} \frac{1}{T} \left\|y_i - x_i\theta^0 - \alpha_h^0\right\|^2 - \frac{1}{T} \left\|y_i - x_i\theta^0 - \alpha_g^0\right\|^2 && (3.4)\\
=&\operatorname*{plim}_{T\to\infty} \frac{1}{T} \left\|\left(\alpha_g^0 - \alpha_h^0\right) + u_i\right\|^2 - \frac{1}{T} \|u_i\|^2 && (3.5)\\
=&\operatorname*{plim}_{T\to\infty} \frac{1}{T} \left\|\alpha_g^0 - \alpha_h^0\right\|^2 + o_p(1) && (3.6)\\
>&0
\end{aligned}$$

where the $o_p(1)$ results from an exogeneity condition applied on $\alpha_g^0 - \alpha_h^0$ that will be made in more detail in Section 3.3. The converse is also true– for any $i \in \mathcal{N}$, if we assume the strict inequality holds for some g, it uniquely minimizes the distance between $y_i - x_i\theta^0$ and the group fixed effect α_g of group g. Therefore, since groups are weakly separated, an individual must belong to a group g if their observed $y_i - x_i\theta$ is closest

to α_g compared to the other group fixed effects, thus identifying group assignments for each individual provided parameters are known.

Weak separability, among other conditions, delivers identification of group memberships even when clusters have many distributional features that are heterogeneous. However, assignment based solely on first moment discrepancies may fail to account for clusters of significant overlap in the finite-sample setting. Accounting for group heteroskedasticity in assignment will properly weigh these discrepancies, but will necessitate steeper identification requirements.

A natural criteria to serve this purpose can be based on the Mahalanobis distance, which normalizes distances to the cluster mean based on the "dispersion" or noise of the cluster. Letting $\sigma_g^0 > 0$ denote the standard deviation of any group g, a candidate rule may be written for all $i \in \mathcal{N}$: $g_i^0 = g$ if and only if, for all $h \neq g$,

$$\operatorname*{plim}_{T\to\infty} \frac{1}{T} \frac{\left\| y_i - x_i'\theta^0 - \alpha_g^0 \right\|^2}{\sigma_g^0} < \operatorname*{plim}_{T\to\infty} \frac{1}{T} \frac{\left\| y_i - x_i'\theta^0 - \alpha_h^0 \right\|^2}{\sigma_h^0}. \tag{3.7}$$

For this rule to be valid, we require "strong" separability of groups:

$$\operatorname*{plim}_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \left(\alpha_{gt}^0 - \alpha_{ht}^0 \right)^2 > \frac{|\sigma_h^0 - \sigma_g^0|}{\min_{f\in\Gamma_G} \sigma_f^0} \operatorname*{plim}_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} u_{it}^2 \tag{3.8}$$

which explicitly requires that the model noise is finite, i.e., $\operatorname*{plim}_{T\to\infty} \sum_{t=1}^{T} u_{it}^2 < \infty$ for all $i \in \mathcal{N}$ and that there are no degenerate groups: $\sigma_f^0 > 0$ for all $f = 1, \ldots, G$. This rule then says that identification of groups still relies on separation in means, but moreover as a function of separation in group variances. The separation becomes more demanding if one or both of the following occur: the model variance is large or the differences in group variances is large relative to the smallest group. We recover weak separability (3.3) and GFE assignment when there is group homoskedasticity.

As an example with no covariates, with $T = 2$, and errors that are conditionally normal given groupings, fixing the mean squared error of the two group fixed effects to 4 and one of the groups variances to 1, (3.8) implies a valid range of variances for the

other group in the interval $(0.2, 2)$ where both groups would remain identifiable. Figure 3.1 displays a visual on how these groups may appear when the variable standard error is relatively close to these bounds. The visuals suggests that strong separation is a reasonable assumption to make, still allowing for considerable overlap, but ruling out identification of groups that solely relies on variance comparisons.

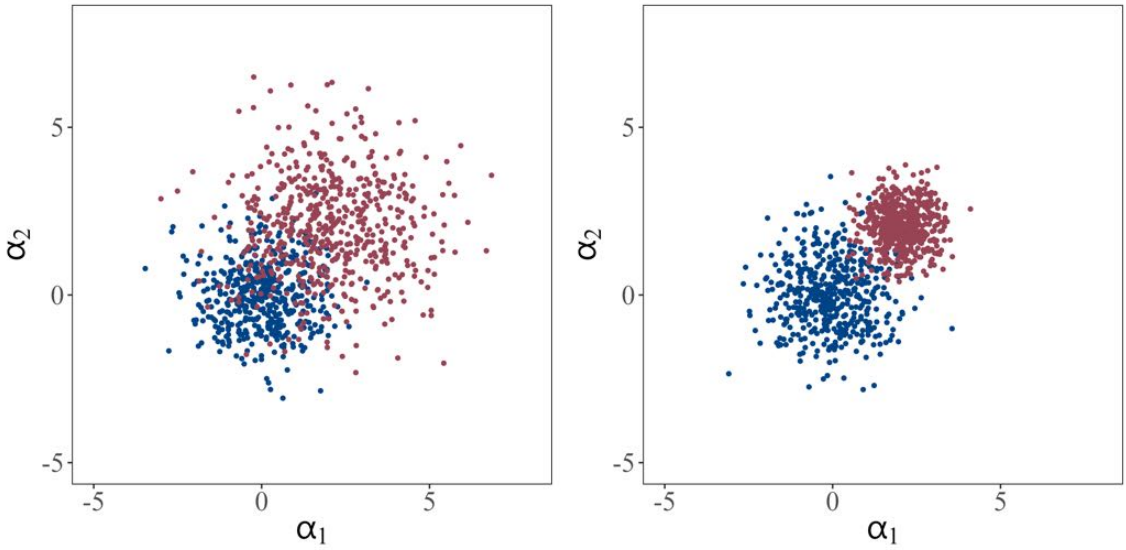


Figure 3.1: Groups that satisfy strong separability. Blue group with standard error 1, Red group with variable σ. Left: $\sigma = 1.55$, Right: $\sigma = 0.65$

Constructing a sample criterion based on the rule (3.7) and taking it to data may perform poorly by favoring the creation of groupings with relatively large variances. To counteract this, a simple additive penalty may be imposed that vanishes asymptotically with larger T so that strong separability is preserved. I propose the following correction:

$$\operatorname*{plim}_{T\to\infty} \frac{1}{T} \frac{\left\| y_i - x_i'\theta^0 - \alpha_g^0 \right\|^2}{\sigma_g} + \frac{1}{T}\sigma_g < \operatorname*{plim}_{T\to\infty} \frac{1}{T} \frac{\left\| y_i - x_i'\theta^0 - \alpha_h^0 \right\|^2}{\sigma_h} + \frac{1}{T}\sigma_h \tag{3.9}$$

so that noisy groups are penalized and less noisy groups are assigned primarily by closeness to group mean. The rule (3.9) is remarkably the same as in barycentric clustering without covariates and assuming groups are part of the same location-scale family

restricted to diagonal covariances matrices; see Yang and Tabak [2022]. This suggests that this assignment problem can be extended to higher heterogeneous moments. In Chapter 4, I replace the standard Euclidean norm with a GMM criteria that has weight matrix independent of the unobservable. This suggests that if there is so-called latent heteroskedasticity then additional identification conditions must be made.

3.2.2 Estimation

The *weighted grouped fixed-effects* (WGFE) estimator for model (3.1) is defined as

$$\left(\widehat{\theta}, \widehat{\alpha}, \widehat{\gamma}\right) = \underset{(\theta,\alpha,\gamma)\in\Theta\times\mathcal{A}^{GT}\times\Gamma_G^N}{\operatorname{argmin}} \sum_{g=1}^{G} P_g^{\gamma} \sqrt{\widehat{Q}_g(\theta,\alpha,\gamma)} \tag{3.10}$$

where

$$\widehat{Q}_g(\theta,\alpha,\gamma) = \frac{1}{T\sum_{j=1}^{N} P_g^j(\gamma)} \sum_{i=1}^{N}\sum_{t=1}^{T} P_g^i(\gamma)(y_{it} - x_{it}'\theta - \alpha_{gt}^0)^2 \tag{3.11}$$

and $P_g^i(\gamma) = \mathbb{1}\{g_i = g\}$ and $P_g^{\gamma} = N^{-1}\sum_{i=1}^{N} P_g^i$.

The objective function is a departure from the pooled least squares criteria of Bonhomme and Manresa [2015] and to k-means when covariates are absent. It is instead a minimization of a weighted transformation of within-group least squares that is shown to coincide to choosing $\widehat{g}_i = \widehat{g}_i(\theta,\alpha)$ that satisfies (3.9) for every $i \in \mathcal{N}$. This estimator is equivalent to the GSR estimator of Aguilar Loyo and Boot [2024] by concentrating out their variance parameters and reducing to homogeneous slope coefficients.

Let $\sigma_{g_i}(\theta,\alpha,\gamma) = \sqrt{\widehat{Q}_{g_i}(\theta,\alpha,\gamma)}$. With a fixed grouping scheme γ^* the first-order conditions with respect to θ^* and α^* are

$$0 = \sum_{i=1}^{N}\sum_{t=1}^{T} \sigma_{g_i^*}^{-1}(\theta^*,\alpha^*,\gamma^*) x_{it}\left(y_{it} - x_{it}'\theta^* - \alpha_{g_i^* t}^*\right) \tag{3.12}$$

$$0 = \frac{\sum_{i=1}^{N}\mathbb{1}\{g_i^* = g\}\left(y_{it} - x_{it}'\theta^* - \alpha_{gt}^*\right)}{\sum_{j=1}^{N}\mathbb{1}\{g_i^* = g\}} \tag{3.13}$$

for all $t \in \mathcal{T}$, $g \in \Gamma_G$. With these groupings, θ^* is the solution to the nonlinear equation

$$\theta^* = \left[\sum_{i=1}^{N} \sum_{t=1}^{T} \sigma_{g_i}^{-1}(\theta^*, \alpha^*, \gamma^*)(x_{it} - \overline{x}_{g_i t})(x_{it} - \overline{x}_{g_i t})' \right]^{-1} \times \sum_{i=1}^{N} \sum_{t=1}^{T} \sigma_{g_i}^{-1}(\theta^*, \alpha^*, \gamma^*)(x_{it} - \overline{x}_{g_i t})(y_{it} - \overline{y}_{g_i t}) \tag{3.14}$$

in the system of equations with

$$\alpha_{gt}^* = \overline{y}_{gt} - \overline{x}_{gt}' \theta^* \tag{3.15}$$

where $\overline{y}_{gt}$ and $\overline{x}_{gt}$ are the within-group averages of the respective variables according to grouping γ^*. The form of (3.14) displays similarities to weighted least squares and solvable by the Newton-Raphson method or by fixed-point iteration.

Remark 1. If slope coefficients were group heterogenous (θ_g) then the weighted form of (3.14) reduces to the usual grouped estimator without the weighting by the group standard errors.

3.2.3 Computation

The partitioning problem of assigning N units into $G < N$ groups is NP-hard (Aloise et al. [2008], Dasgupta [2008]) and, with additional parameters such as θ, rules out exhaustive search [1]. I follow convention and propose a heuristic algorithm based on Lloyd's (k-means) algorithm (Lloyd [1982]) that after initializing will alternate between assignment into groups and then updating parameters based on those assignments.

[1] Clausen [2003], Brusco [2006] and Aloise et al. [2012] for some global and exact methods that are prohibitive in my application.

Algorithm 1. *(Lloyd's Algorithm for WGFE).*

1. *Initialize $\theta^{(0)}$; set the $\alpha^{(0)}$ as G randomly chosen $v_i = y_i - x_i\theta^{(0)}$; create initial assignments $g_i^{(0)}$ by assigning v_i to the closest $\{\alpha_{g_i^0}^{(0)}\}$; and calculate weights $\{P_g^{(0)}\}$ and $\{\sigma_g^{(0)}\}$ for all $g \in \Gamma_G$. Set $s = 1$.*

2. *(Assignment Step). For all i, assign according to*

$$g_i^{(s+1)} \leftarrow \underset{g}{\operatorname{argmin}} \left\{ \frac{\left\| y_i - x_i\theta^{(s)} - \alpha_g^{(s)} \right\|^2}{\sigma_g^{(s)}} + \sigma_g^{(s)} \right\}$$

   and collect them in $\gamma^{(s+1)} = (g_1^{(s+1)}, \ldots, g_N^{(s+1)})$.

3. *(Update Step). Update $\alpha^{(s+1)}$, $\theta^{(s+1)}$ according to* (3.14) *and* (3.15)*, respectively, then update $\{P_g^{(s)}\}$ and $\{\sigma_g^{(s+1)}\}$ for all $g \in \Gamma_G$.*

4. *If $g_i^{(s)} = g_i^{(s+1)}$ for all i, stop. Otherwise, set $s \leftarrow s + 1$ and go back to step 2.*

This is a variance augmented *k*-means algorithm similar to gradient descent where part of the set of parameters take on discrete values and each iteration improves on the function value until a local minimum is reached (see Forgy [1965], MacQueen [1967] and Lloyd [1982]). This procedure converges quickly, however it is a heuristic algorithm that should be repeated many times with a randomly drawn starting value and choosing the result with the smallest local minima; see Brusco and Steinley [2007] and Hansen et al. [2010] for efficient versions of this algorithm. I adapt a variable neighborhood search (VNS) algorithm described in the Supplementary Appendix of Bonhomme and Manresa [2015] and originally from Pacheco and Valencia [2003] to incorporate covariates and utilize it in simulations and the empirical applications. These algorithms combine both stochastic initialization and deterministic search to comb the domain for the smallest local minimum and can be more efficient than repeated applications of Algorithm 1. See Appendix B.4.1 for a description of the VNS Algorithm used in this paper.

The rule (3.9) appears in the assignment step and serves as a balanced criteria to assign an individual into a group it is closest in group fixed effect relative to the noise in the group. While very noisy groups make the first term smaller, they are penalized by the second term preventing assignment solely based on the standardized distance. This is not the first clustering criteria incorporating second moment information, see Zhao et al. [2015]. The Mahalanobis distance can be used, however too much normalization without the penalty term in (3.9) may result in an assignment rule that favors the noisiest group and the clustering objective function becomes trivial (Krishnapuram and Kim [1999]).

3.2.4 Testing for group heteroskedasticity

The GFE estimator is defined as

$$\left(\widetilde{\theta}, \widetilde{\alpha}, \widetilde{\gamma}\right) = \underset{(\theta,\alpha,\gamma)\in\Theta\times\mathcal{A}^{GT}\times\Gamma_G^N}{\operatorname{argmin}} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it} - x_{it}'\theta - \alpha_{g_i t}\right)^2 \tag{3.16}$$

and the assignment step is given by

$$g_i \leftarrow \underset{g}{\operatorname{argmin}} \left\|y_i - x_i\theta - \alpha_g\right\|^2. \tag{3.17}$$

The estimators have a closed-form for given grouping $\widetilde{g}$:

$$\widetilde{\theta} = \left[\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \overline{x}_{\widetilde{g}_i t})(x_{it} - \overline{x}_{\widetilde{g}_i t})'\right]^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \overline{x}_{\widetilde{g}_i t})(y_{it} - \overline{y}_{\widetilde{g}_i t}) \tag{3.18}$$

and

$$\widetilde{\alpha}_{\widetilde{g}t} = \overline{y}_{\widetilde{g}t} - \overline{x}_{\widetilde{g}t}'\widetilde{\theta}. \tag{3.19}$$

The GFE criterion function weighs the residual variance the same across all groups while the WGFE criterion separates them by weighing them according to the size of groups. Furthermore, they are connected in a way economists might describe risky behavior. The GFE criterion is to the risk neutral agent while the WGFE is to the risk averse agent. The following is a consequence of Jensen's inequality.

Proposition 7. Let $\widetilde{Q}$ and $\widehat{Q}$ denote the sample criterion functions for the GFE and WGFE estimators, respectively. Then,

$$\left[\widehat{Q}(\theta, \alpha, \gamma)\right]^2 \leq \widetilde{Q}(\theta, \alpha, \gamma) \tag{3.20}$$

for any $(\theta, \alpha, \gamma) \in \Theta \times \mathcal{A}^{GT} \times \Gamma_G^N$.

Furthermore, the GFE estimator is equivalent to the restricted GSR estimator under the $G-1$ linear restrictions that $\sigma_g = \sigma_h$ for all $g, h \in \Gamma_G$. Therefore, this fact and Proposition 7 suggests that the value functions of the GFE and WGFE estimators can be used to define a quasi-likelihood ratio test statistic for the null hypothesis that there is no group heteroskedasticity:

$$\tau_{NT} = 2NT \left(\widetilde{Q}\left(\widetilde{\theta}, \widetilde{\alpha}, \widetilde{\gamma}\right) - \left[\widehat{Q}\left(\widehat{\theta}, \widehat{\alpha}, \widehat{\gamma}\right)\right]^2 \right) / \left[\widehat{Q}\left(\widehat{\theta}, \widehat{\alpha}, \widehat{\gamma}\right)\right]^2 \geq 0.$$

Under the null of group-homoskedasticity and consistency of criterion functions and estimators, we should have $\tau_{NT} \to_p 0$ as N, T approaches infinity. Furthermore, τ_{NT} is expected to be distributed as chi-squared with $(G-1)$ degrees of freedom, pending the asymptotic normality results in the next section.

3.3 Asymptotic Theory

Consider the data generating process

$$y_{it} = x_{it}'\theta^0 + \alpha^0_{g_i^0 t} + u_{it},$$

where $g_i^0 \in \Gamma_G$ denotes the true group membership for each individual $i \in \mathcal{N}$ and zero superscripts denote true values. The triple $\{(y_i, x_i, g_i^0)\}_{i \in \mathcal{N}}$ is regarded as a random sample from some joint distribution. I provide conditions that enable the WGFE estimators of group assignments to converge to their true values and use this to show that the WGFE parameter estimators are asymptotically equivalent to their infeasible version where the true group memberships are known. I establish conditions for which

this infeasible estimator is consistent and asymptotically normal using standard theory of extremal estimation (Newey and McFadden [1986]). In this setting both N and T are allowed to approach infinity, but T may grow slower than N.

3.3.1 Infeasible WGFE estimation

Redefine $\left(\widetilde{\theta}, \widetilde{\alpha}\right)$ to be the infeasible version of the WGFE estimator where the true group memberships are known:

$$\left(\widetilde{\theta}, \widetilde{\alpha}\right) = \underset{(\theta,\alpha)\in\Theta\times\mathcal{A}^{GT}}{\operatorname{argmin}} \sum_{g=1}^{G} P_g \sqrt{\widetilde{Q}_g(\theta,\alpha)} \tag{3.21}$$

where

$$\widetilde{Q}_g(\theta,\alpha) = \frac{1}{T\sum_{j=1}^{N} P_g^j} \sum_{i=1}^{N}\sum_{t=1}^{T} P_g^i (y_{it} - x_{it}'\theta - \alpha_{gt})^2 \tag{3.22}$$

and $P_g^i = \mathbb{1}\{g_i^0 = g\}$ and $P_g = N^{-1}\sum_{i=1}^{N} P_g^i$.

Unlike GFE estimation, (3.21) is not a pooled least squares estimator so it must be shown that the infeasible estimator has the property of consistency and asymptotic normality under some conditions. The population criterion function is

$$Q(\theta,\alpha) = \sum_{g=1}^{G} P_g^0 \sqrt{Q_g(\theta,\alpha)} \tag{3.23}$$

where $P_g^0 = \mathbb{P}(g_i^0 = g)$ and $Q_g(\theta,\alpha) = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\{g_i^0 = g\}(y_{it} - x_{it}'\theta - \alpha_{gt})^2\right]$.

Assumption 2. (Infeasible WGFE). *There exists $M > 0$ such that*

a. The parameter space $\Theta \times \mathcal{A}^{GT} \subset \mathbb{R}^p \times \mathbb{R}^{GT}$ is compact.

b. $\mathbb{E}\left[u_{it}^4\right] < M$, $\mathbb{E}\left[u_{it}\right] = 0$ and $\mathbb{E}\left[\|x_{it}\|^4\right] < M$

c. $\sum_{i=1}^{N} \mathbb{1}\{g_i^0 = g\} \to \infty$ as $N \to \infty$.

d. For every $g = 1, \ldots, G$, $\mathbb{P}(g_i^0 = g) > 0$.

e. The matrix

$$\operatorname*{plim}_{T\to\infty} \sum_{g=1}^{G} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[\left(x_{it} - \mathbb{E}\left[x_{it}|g_i^0 = g\right]\right)\left(x_{it} - \mathbb{E}\left[x_{it}|g_i^0 = g\right]\right)' |g_i^0 = g\right]$$

has a positive minimum eigenvalue.

The first two assumptions are standard, while (c) and (d) are non degeneracy conditions on each group. I require all groups to have many observations in them and thus group variances are bounded away from zero. Assumption 2(e) is a full rank condition analogous to those on design matrices for the basic linear models. A sufficient condition for (e) is that for each $g = 1, \ldots, G$ and setting $\widetilde{X}_i = X_i - \mathbb{E}\left[X_i|g_i^0\right]$ as the stacked $x_{it} - \mathbb{E}\left[x_{it}|g_i\right]$, the matrix $\mathbb{E}\left[\widetilde{X}_i'\widetilde{X}_i|g_i^0 = g\right]$ has full rank. Since these matrices are symmetric, they are positive definite and so is their sum. Hence the minimum eigenvalue of the sum is positive. This is similar to fixed effects but, instead of a within-individual transformation, it is at the group level which requires removing the time-varying group mean for which individual i belongs.

The following consistency result does not require T to be large and follows from the standard result on consistency of M-estimators; see Appendix B.5 for a proof.

Theorem 1. *Under Assumption 1, $\widetilde{\theta} \to_p \theta^0$ and $\widetilde{\alpha} \to_p \alpha^0$, as $N \to \infty$.*

With $\sqrt{N}$-consistency of the infeasible WGFE estimator, I turn to the asymptotic distribution of the estimator. Under a few more standard assumptions, the infeasible WGFE

estimator is asymptotically normally distributed as $N, T \to \infty$. This follows from the standard asymptotic normality theorem for M-estimators. First, define the estimator for the group variance when groupings are known as $\widetilde{\sigma}^2_{g_i^0}(\widetilde{\theta}, \widetilde{\alpha})$. Since the infeasible estimator is consistent then $\widetilde{\sigma}^2_{g_i^0}(\widetilde{\theta}, \widetilde{\alpha}) \to_p \sigma^2_{g_i^0}$ as N approaches infinity using a weak law of large numbers. The following assumptions are sufficient to characterize the asymptotic distribution of $\left(\widetilde{\theta}, \widetilde{\alpha}\right)$.

Assumption 3. (Normality of infeasible WGFE).

a. For all $i, j = 1, \ldots, N$ and $t = 1, \ldots, T$: $\mathbb{E}\left[x_{jt} u_{it}\right] = 0$.

b. There exists positive definite matrices B_θ and V_θ such that, as $N, T \to \infty$

$$\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \widetilde{\sigma}^{-1}_{g_i^0}(\widetilde{\theta}, \widetilde{\alpha})(x_{it} - \overline{x}_{g_i^0 t})(x_{it} - \overline{x}_{g_i^0 t})' \longrightarrow_p B_\theta$$

$$\mathbb{E}\left[\frac{1}{NT} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=1}^{T} \sum_{s=1}^{T} \left[\widetilde{\sigma}_{g_i^0}(\widetilde{\theta}, \widetilde{\alpha}) \widetilde{\sigma}_{g_j^0}(\widetilde{\theta}, \widetilde{\alpha})\right]^{-1} (x_{it} - \overline{x}_{g_i^0 t})(x_{js} - \overline{x}_{g_j^0 s})' u_{it} u_{js}\right] \longrightarrow V_\theta.$$

c. As $N, T \to \infty$, $\frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} \sigma^{-1}_{g_i^0}(x_{it} - \overline{x}_{g_i^0 t}) u_{it} \to_d N(0, V_\theta)$.

d. For all g and t and as $N \to \infty$, $\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbb{E}\left[\mathbb{1}\{g_i^0 = g\} \mathbb{1}\{g_j^0 = g\} u_{it} u_{jt}\right] \to v_{gt} > 0$.

e. For all g and t and as $N \to \infty$, $\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbb{1}\{g_i^0 = g\} u_{it} \longrightarrow_d N(0, v_{gt})$.

Assumption 3(a) allows covariates that are strictly exogenous and also time lagged outcomes. The following establishes asymptotic normality of the infeasible estimator.

Theorem 2. *Suppose Assumptions 2 and 3 hold. As N and T approach infinity,*

$$\sqrt{NT}\left(\widetilde{\theta} - \theta^0\right) \longrightarrow_d N(0, B_\theta^{-1} V_\theta B_\theta^{-1})$$

and, for all $g \in \Gamma_G$ and $t \in \mathcal{T}$,

$$\sqrt{N}\left(\widetilde{\alpha}_{gt} - \alpha^0_{gt}\right) \longrightarrow_d N\left(0, v_{gt} / \left[\mathbb{P}(g_i^0 = g)\right]^2\right).$$

The form of the asymptotic variance of the infeasible estimator is different than the pooled least squares case in that observations are weighted according to the noise in their group. It appears similar to weighted least squares however the weight is the standard deviation instead of the variance.

3.3.2 Consistency of the WGFE estimator

This section establishes conditions for the consistency of the WGFE estimator $\left(\widehat{\theta}, \widehat{\alpha}\right)$ of $\left(\theta^0, \alpha^0\right)$. Remarkably these assumptions are identical to those made in Bonhomme and Manresa [2015] for consistency of the GFE estimator.

Assumption 4. (Consistency of parameters). *There exists $M > 0$ such that*

a. $\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}\left[u_{it}u_{is}x_{it}'x_{is}\right]\right| \leq M.$

b. $\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[v_{it}v_{jt}\right]\right| \leq M.$

c. $\left|\frac{1}{N^2T}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathrm{Cov}(v_{it}v_{jt}, v_{is}v_{js})\right| \leq M.$

d. Let $\overline{x}_{g\wedge\widetilde{g}t}$ denote the average of x_{it} in the intersection of groups $g_i^0 = g$ and $g_i = \widetilde{g}$. For all $g \in \Gamma_G$, and for any $\gamma = (g_1, \ldots, g_N) \in \Gamma_G^N$ define $\widehat{\rho}(\gamma)$ as the minimum eigenvalue of the matrix

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x_{it} - \overline{x}_{g_i\wedge g_i^0 t}\right)\left(x_{it} - \overline{x}_{g_i\wedge g_i^0 t}\right)'.$$

Then, $\min_{\gamma\in\Gamma_G}\{\widehat{\rho}(\gamma)\} \to_p \rho > 0$ as $N, T \to \infty$.

To summarize from previous work, weak dependence conditions in the form of Assumptions 4(b, c, d) are imposed. These assumptions are related to those made in Stock and Watson [2002] and Bai and Ng [2002] on large factor models. Assumption 4 (b) allow for lagged outcomes or any predetermined regressor as covariates e.g. $\mathbb{E}\left[u_{it}|x_{it}, x_{it-1}, u_{it-1}\right] = 0$. Assumptions 4 (b) and (d) are conditions on time series dependence of the errors and

covariates while Assumption 4(c) limits cross-sectional dependence. Assumption 4 (e) is akin to Assumption 2 (e), so it is similar to a full rank condition found in other linear regression models. This sample version is stronger since it requires sufficient variation in the covariates across time and individuals within groups generated by *any* grouping scheme $\widetilde{\gamma} \in \Gamma_G^N$.

Consistency largely follows that of Bonhomme and Manresa [2015] with additional algebraic steps taken given the additional complexity of the WGFE criterion.

Theorem 3. *Suppose Assumption 2 and 4 hold. Then, as* $N, T \to \infty$,

$$\widehat{\theta} \to_p \theta^0 \quad \textit{and} \quad \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\widehat{\alpha}_{\widehat{g}_i t} - \alpha^0_{g_i^0 t}\right)^2 \longrightarrow_p 0.$$

3.3.3 Consistency of group assignments

In this section I discuss properties of the WGFE estimator for group assignment in a large N and T setting. The assignment of individual i into any of the groups depends on their time series information. Therefore it is expected that group assignment should be more accurate as T grows large. I refer to the assignment rules (3.9) and (3.17) as the WGFE and GFE assignments, respectively. I will first present a simple case of two groups for intuition and then will move to the general case where I provide conditions for any number of groups G.

Simple Case

Consider a simplification of the main model where there are no covariates, $G = 2$ and errors are independent, but drawn from different normal distributions depending on g_i^0 with variance $\sigma^2_{g_i^0}$. Formally,

$$y_{it} = \alpha^0_{g_i^0} + u_{it}, \;\; u_{it} \sim N(0, \sigma^2_{g_i^0}), \;\; u_{it} \perp\!\!\!\perp u_{jt} \text{ for } i \neq j.$$

Assume that $\sigma_1 \geq \sigma_2$. Then, the probability of incorrectly assigning an individual

into group 2 while they belong to group 1 using WGFE assignment is

$$\mathbb{P}(\widehat{g}_i(\alpha^0)) = 2|g_i^0 = 1) \tag{3.24}$$

$$= \mathbb{P}\left(\frac{1}{\sigma_2}\sum_{t=1}^{T}\left(\alpha_1^0 + u_{it} - \alpha_2^0\right)^2 + \sigma_2 < \frac{1}{\sigma_1}\sum_{t=1}^{T}\left(\alpha_1^0 + u_{it} - \alpha_1^0\right)^2 + \sigma_1\right) \tag{3.25}$$

$$= \mathbb{P}\left(\left(1 - \frac{\sigma_2}{\sigma_1}\right)\sum_{t=1}^{T} u_{it}^2 + 2\left(\alpha_1^0 - \alpha_2^0\right)\sum_{t=1}^{T} u_{it} < (\sigma_1 - \sigma_2)\,\sigma_2 - T\left(\alpha_1^0 - \alpha_2^0\right)^2\right) \tag{3.26}$$

This expression on the left hand side is the distribution function of some generalized chi-squared distribution and has no closed-form; see Davies [1980]. If $\sigma_1 = \sigma_2$, then this is simply GFE assignment.

In the case that $\alpha_1^0 = \alpha_2^0$ are equal:

$$\mathbb{P}\left(\widehat{g}_i(\alpha^0) = 2|g_i^0 = 1\right) = \mathbb{P}\left(\frac{\sum_{t=1}^{T} u_{it}^2 - T}{\sqrt{2T}} < \frac{\sigma_1\sigma_2 - T}{\sqrt{2T}}\right) \approx \Phi\left(\frac{\sigma_1\sigma_2 - T}{\sqrt{2T}}\right) \longrightarrow 0$$

as $T \to \infty$ where Φ is the standard normal cdf. However, if $\sigma_2 > \sigma_1$, then this is no longer the case as the leading term in (3.26) is now negative and the inequality will reverse:

$$\mathbb{P}\left(\widehat{g}_i(\alpha^0)\right) = 2|g_i^0 = 1) = \mathbb{P}\left(\frac{\sum_{t=1}^{T} u_{it}^2 - T}{\sqrt{2T}} > \frac{\sigma_1\sigma_2 - T}{\sqrt{2T}}\right) \approx 1 - \Phi\left(\frac{\sigma_1\sigma_2 - T}{\sqrt{2T}}\right) \longrightarrow 1. \tag{3.27}$$

Therefore, one group will be left empty since this assignment rule can't distinguish between the smaller density with σ_1 and the larger density with σ_2. This shows that separation of group fixed effects is critical even if there is other information to use in classification.

Figure 3.2 shows the misclassificaton probability (3.25) as T grows larger by calculating the probability numerically. In this example, $\alpha_1 = 1$, $\alpha_2 = 0$, $\sigma_1 = 1$, and $\sigma_2 = 1.1$. We see that the probability of incorrectly assigning an individual falls exponentially as T becomes large, mirroring the same property from GFE estimation.

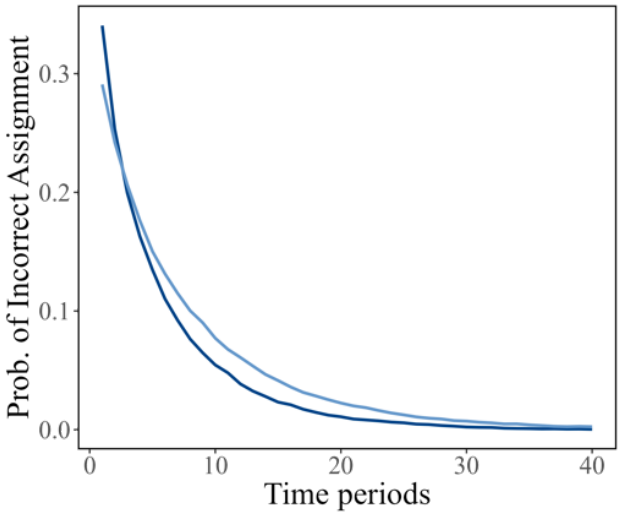


Figure 3.2: Dark Blue: Misclassification probability (3.25) (and the opposite case of assigning into group 1 when group 2 is correct) with $\alpha_1 = 1$, $\alpha_2 = 0$, $\sigma_1 = 1$, and $\sigma_2 = 1.1$.

The General Case

Let $\eta > 0$ and define a neighborhood $\mathcal{N}_\eta$ around the true parameter values θ^0 and α^0 as the subset of $(\theta, \alpha) \subset \Theta \times \mathcal{A}^{GT}$ that satisfy $\left\|\theta - \theta^0\right\| < \eta$ and $T^{-1}\sum_{t=1}^{T}(\alpha_{gt}^0 - \alpha_{gt})^2 < \eta$ for all $g \in \Gamma_G$. Define $\sigma_g^2(\theta, \alpha, \gamma) = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\mathbb{1}\{g_i = g\}\left(u_{it} + x_{it}'(\theta^0 - \theta) + \alpha_{gt}^0 - \alpha_{gt}\right)^2\right]$.

Assumption 5. (Consistency of group assignments).

a. For all $g \in \Gamma_G$, $\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\{g_i^0 = g\} \to_p P_g > 0$.

b. There exists constants $a > 0$ and $d_1 > 0$ and a sequence $\alpha[t] \leq e^{-at^{d_1}}$ such that, for all $i \in \mathcal{N}$ and $g \in \Gamma_g$, $\{u_{it}\}_t$ and $\{\alpha_{gt}^0\}_t$ are strongly mixing processes with mixing coefficients $\alpha[t]$. Moreover, $\mathbb{E}\left[\alpha_{gt}^0 u_{it}\right] = 0$ for all $g \in \Gamma_G$.

c. There exists constants $b > 0$ and $d_2 > 0$ such that $\mathbb{P}(|u_{it}| > m) \leq e^{1-\left(\frac{m}{b}\right)^{d_2}}$ for all i, t, and $m > 0$.

d. There exists a constant $M^* > 0$ such that, as N, T tend to infinity I have

$$\sup_{i \in \mathcal{N}} \mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T} \|x_{it}\| \geq M^*\right) = o\left(T^{-\delta}\right) \quad \textit{for all } \delta > 0.$$

e. For all $g, \widetilde{g} \in \Gamma_g$ where $g \neq \widetilde{g}$,

$$\text{plim}_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^0 - \alpha_{\widetilde{g}t}^0\right)^2 > \frac{|\sigma_h^0 - \sigma_g^0|}{\min_{f\in\Gamma_G}\sigma_f^0} \underset{T\to\infty}{\text{plim}} \frac{1}{T}\sum_{t=1}^{T} u_{it}^2.$$

The assumptions made here do not deviate much from those in Bonhomme and Manresa [2015] besides the strong separability assumption in part (*e*) that requires a specific bound depending on other features of the model. Assumptions 5(*b*) and (*c*) strengthen restrictions on the dependence and tail properties of the error, respectively. I assume the error is strongly mixing with faster-than-polynomial decay rate and with tails that also decay faster than any polynomial. The mixing property is stronger than ergodicity and says that, eventually, the process will forget its history. Mathematically, for any stochastic process X_t on some probability space $(\Sigma, \mathcal{F}, \mathbb{P})$ define the function $\alpha[s]$ as a strongly mixing coefficient:

$$\alpha[s] = \sup\{|\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)| : t \geq 0, A \in X_0^t, B \in X_{t+s}^\infty\}$$

where X_a^b denote the sub σ-algebra of $\mathcal{F}$ specified between times a and b. If $\alpha[s] \to 0$ as $s \to \infty$, then X_t is said to be a strongly mixing process. I also assume that the group time-effects are strongly mixing and contemporaneously uncorrelated with the error u_{it}. These assumptions enable us to use exponential inequalities for weakly dependent processes and allows us to bound misclassification probabilities.[2] Assumption 5 (*d*) is a condition on the distribution of the covariates. A sufficient condition would be if the covariates

[2]For more on asymptotic theory of weakly dependent process see Rio [2017] and for useful exponential inequalities see Merlevède et al. [2011]. This condition for strongly mixing was established in Rosenblatt [1956] to prove a central limit theorem.

have bounded support or, alternatively, satisfy similar dependence and tail conditions as the error term. In their supplementary appendix, Bonhomme and Manresa [2015] discuss in detail the latter condition since the strong mixing property may not hold with lagged outcomes.

The following proposition establishes an asymptotic equivalence of WGFE estimation to infeasible WGFE estimation and consistency of group membership as N, T approach infinity, but T may grow slower than N.

Theorem 4. *Suppose Assumptions 2, 4 and 5 hold. Then, for all $\delta > 0$,*

$$\mathbb{P}\left(\sup_{i\in\mathcal{N}} |\widehat{g}_i - g_i^0| > 0\right) = o(1) + o(NT^{-\delta}) \tag{3.28}$$

and for all g and t

$$\widehat{\theta} = \widetilde{\theta} + o_p(T^{-\delta}) \tag{3.29}$$

$$\widehat{\alpha}_{gt} = \widetilde{\alpha}_{gt} + o_p(T^{-\delta}). \tag{3.30}$$

3.3.4 Asymptotic Normality of the WGFE Estimator

With conditions guaranteeing the asymptotic equivalence of WGFE estimator and the infeasible WGFE estimator, the asymptotic distribution of the WGFE estimator for θ^0 and α^0 is that of the infeasible version where groups are known.

Proposition 8. Let Assumptions 2, 3, 4 and 5 hold. As N and T approach infinity such that for some $\nu > 0$, $N/T^{\nu} \to 0$ I have

$$\sqrt{NT}\left(\widehat{\theta} - \theta^0\right) \longrightarrow_d N\left(0, B_\theta^{-1} V_\theta, B_\theta^{-1}\right)$$

and, for all $g = 1, \ldots, G$ and $t = 1, \ldots, T$,

$$\sqrt{N}\left(\widehat{\alpha}_{gt} - \alpha_{gt}^0\right) \longrightarrow_d N\left(0, v_{gt}/\left[P_g\right]^2\right).$$

3.4 Simulations

In this section I show evidence of the finite sample performance of the WGFE estimator. The expectation is that the WGFE estimator will classify more accurately than GFE when there is group heteroskedasticity and do no better than GFE when errors are homoskedastic.

The simulation design is based on that of Bonhomme and Manresa [2015] where they revisit the Acemoglu, Johnson, Robinson, & Yared (2008) study on the association between income and democracy. The data generating process is

$$y_{it} = \widehat{\theta}_1 y_{it-1} + \widehat{\theta}_2 x_{it} + \widehat{\alpha}_{\widehat{g}_i t} + u_{it}$$

where y_{it} is an index of democracy and x_{it} is log-income per capita. I split this DGP into two depending on the estimates, groupings and variance of normally distributed errors with zero mean:

1. GFE estimates, groupings and homoskedastic $\widehat{\sigma}^2$ as the mean of squared residuals.

2. WGFE estimates, groupings and group heteroskedastic $\widehat{\sigma}^2_{\widehat{g}_i^{WGFE}}$

I generate data for different specifications of G while holding the income vector x_{it} fixed. I collect a sample of 1,000 GFE and WGFE estimates for each specification.

In Figure 3.1, the values of the heterogeneous variances and sizes are provided for DGP 2. In Figure 3.2, I compare the root mean-squared error of the parameter estimates using WGFE, GFE, and two-way fixed effects as well as the average rates of misclassification across different G for each DGP. We see that for DGP 1 there is not much difference between WGFE and GFE estimation since errors are homoskedastic. However, for DGP 2 it appears that WGFE is a better performer across all metrics. Unlike group-wise heteroskedasticity with observed groups, ignoring it in this context not only affects standard errors, but also causes bias since the estimators are functions of group assignments.

G	σ_1	σ_2	σ_3	σ_4	σ_5
2	0.219	0.086	–	–	–
3	0.230	0.076	0.114	–	–
4	0.235	0.105	0.152	0.059	–
5	0.141	0.206	0.064	0.096	0.209

G	P_1	P_2	P_3	P_4	P_5
2	0.64	0.36	–	–	–
3	0.43	0.33	0.24	–	–
4	0.33	0.23	0.14	0.29	–
5	0.14	0.21	0.30	0.20	0.14

Table 3.1: DGP# 2: latent variable features across groups. Left: group variances, Right: group mass function.

DGP# 1	Lagged Term			Income Term			Missclassified	
G	WGFE	GFE	FE	WGFE	GFE	FE	WGFE	GFE
2	0.0484	0.0479	0.3146	0.0173	0.0173	0.1095	10.38%	10.28%
3	0.0415	0.0416	0.1211	0.0133	0.0132	0.1029	6.55%	6.56%
4	0.0423	0.0422	0.0623	0.0109	0.0109	0.0708	4.74%	4.69%
5	0.0458	0.0449	0.0527	0.0105	0.0105	0.0855	4.13%	4.19%
DGP# 2								
2	0.042	0.069	0.2538	0.012	0.026	0.1134	3.71%	18.51%
3	0.049	0.067	0.1211	0.010	0.022	0.1029	4.85%	16.93%
4	0.053	0.066	0.1607	0.010	0.020	0.0983	8.86%	20.41%
5	0.047	0.061	0.2072	0.008	0.013	0.0972	7.78%	13%

Table 3.2: Root Mean Squared Errors and average misclassification rate for the WGFE, GFE and fixed effects (FE) estimators of θ_1 and θ_2 across 1,000 simulations.

3.5 Empirical Applications

3.5.1 The Effect of Unionization on Wages

In studies on the effect of unionization on earnings it is often argued that controlling for unobserved ability or skill is essential since employers who face a union contract may select more able candidates than those employers who do not (Abowd and Farber [1982]). A solution for this selection problem would be a fixed-effects approach, but measurement error resulting from misclassification of unionized jobs is more pronounced with common short panels (Freeman [1984]).

Estimates of the effect using the Panel Study of Income Dynamics (PSID) range from 8%–23% and likely affected by union misclassification. Card [1996] estimate a discrete proxy for ability that is used to estimate five separate models for each ability level, controlling for misclassification in the process. He finds at "high" levels of ability there is negative bias when ignoring high skill heterogeneity versus the positive bias found with low ability workers.

I revisit the effect of unionization using data from the PSID on $N = 1,158$ workers who are the heads of their household between 2001 and 2019 ($T = 10$, odd numbered years). I add onto and draw comparisons to both of these studies by showing the WGFE estimate lies outside the range between pooled and fixed effects estimates as postulated by Freeman [1984] and the estimated groups follow a similar correlation pattern to the skill groups found in Card [1996].

I justify unobserved groups by arguing that employers form discrete-level impressions of ability of workers based on comparisons to other job candidates. For example, they may classify candidates as "avoid hire", "maybe hire" or "strong hire". There may also exist a complicated correlation structure between working at a union job and skill. On one hand, employers are more selective in jobs covered by a union. On the other, jobs that are not covered by unions, but compensate well may also be selective. It is also reasonable that this discrete classification is correlated with wages across time, which

may be due to the bargaining power of unions and overall economic conditions. For example, jobs with a strong union may be more robust to poor economic conditions such as a recession or worsening wage inequality. In addition the response from each skill group to earnings shocks might be more variable suggesting unobserved group heteroskedasticity.

I control for this discrete unobserved variable by estimating a linear model of unionization and wages with a group fixed-effects term:

$$logwage_{it} = \delta union_{it} + x'_{it}\theta + \alpha_{g_i t} + u_{it} \tag{3.31}$$

where $logwage_{it}$ is the log real labor earnings of worker i in year t in 2001 USDs, $union_{it}$ is a dummy variable indicating if worker i had a job at time t that was under a union contract and x_{it} contains time-varying and time-invariant characteristics such as years of schooling, if they are non White, if they are female, number of years of experience at current job, and if there job is classified as "blue collar". I also include a time-varying dummy variable indicating if the worker resides in the south, along with their marital status and age.

Figure 3.3 shows summary statistics for the data where the majority of respondents are white, male, and completed grade school. Figure 3.3 displays median incomes across different groups. Clearly, unions have a positive impact on real earnings despite the erosion of numbers of unionized jobs post 2007 as shown in Figure 3.4. This effect persists across race and sex.

Estimating model (3.31) with pooled least squares and two way fixed effects estimation indicates union membership corresponds to 9.32% and 22% increases in real earnings, respectively. This observation is surprising since Freeman [1984] pointed out that pooled OLS would produce results larger than two-way fixed effects, and worker bargaining power seems to be waning, especially compared to the proportion of unionized jobs in the older studies. One empirical explanation is that of Card [1996] where unobserved ability has a complicated correlation structure with unionized job selection.

	Mean	Median	Min	Max
Wages	$53,272	$40,496	$694	$3,365,385
Age	46.41	46	19	87
Experience	9.88	8	0	57

Table 3.3: Wages in 2001 USD, experience in years.

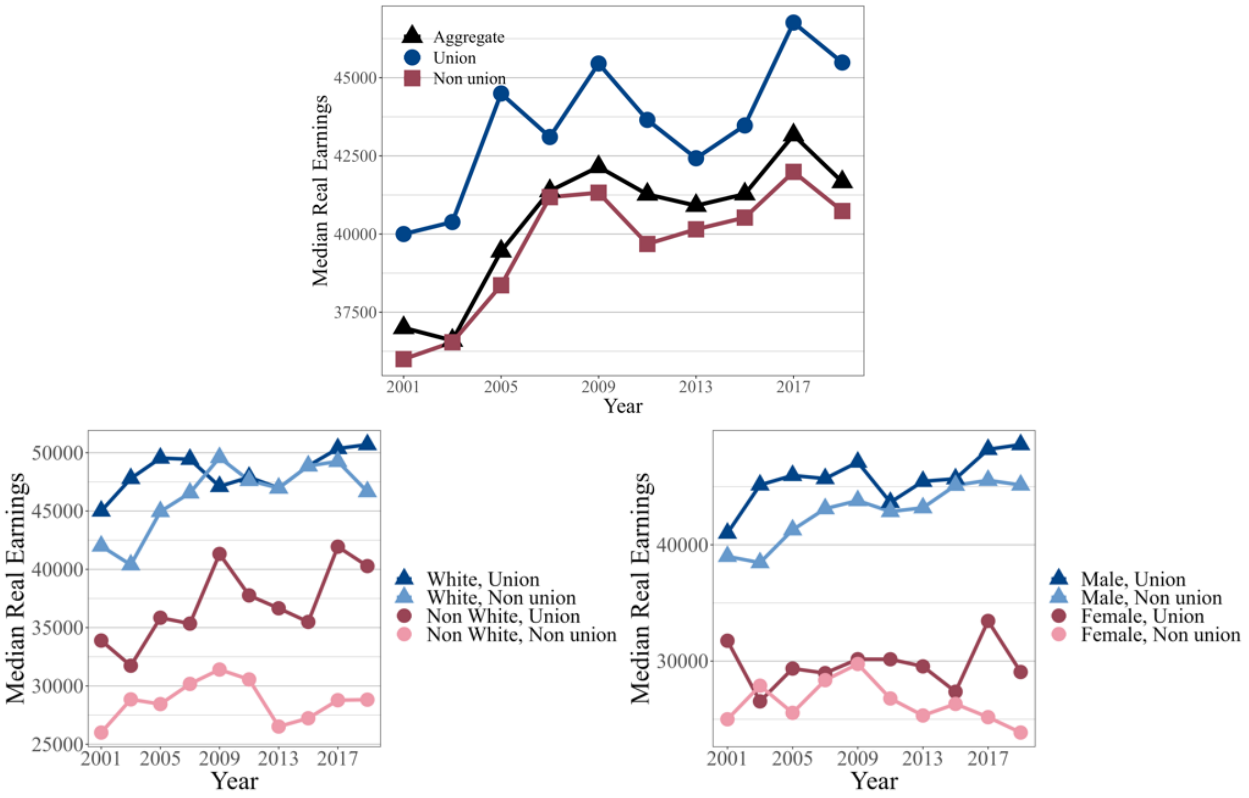


Figure 3.3: Differences in median incomes between observable groups: Aggregate, race, and sex.

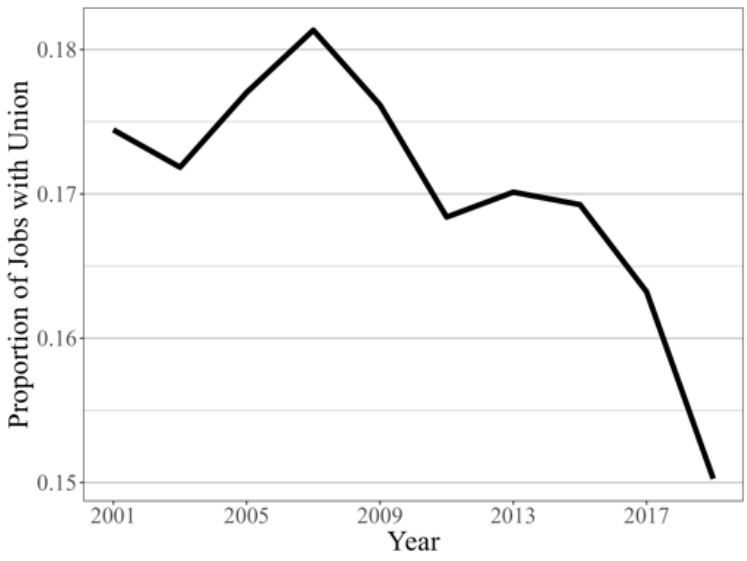


Figure 3.4: Proportion of workers with jobs with in a union contract

This may be due to a modern labor market with more variable job outcomes due to technological advancements and the need for the associated skills. The worker has options beyond the traditional union job and so I observe in the sample consistent high earners who are compensated well outside of a union contract. Therefore, the more able may not value a union as much since employers make more competitive offers in their ability category. Wage inequality may also play as a secondary factor as the dominating wages this group might earn may induce negative correlation with union job employment, explaining why these estimates are much larger than pooled OLS estimates.

A Bayesian Information Criteria[3] is used to estimate $\widehat{G} = 16$ and the WGFE estimate shows a significant 6.4% (standard error of 0.008) union-earnings effect, which is less than pooled OLS and two-way fixed effects. Figure 3.5 shows nonlinear correlation where the proportion of union jobs and mean earnings averaged over time are positively correlated with unionization until earnings are much larger when it becomes negatively

[3]See the appendix of Aguilar Loyo and Boot [2024] and Bai and Ng [2002] for the BIC method of estimating the number of groups.

correlated. In other words, earnings are lowest among the lowest skill groups not covered by unions. As incomes rise, the prevalence of unions also rise, until reaching outlier groups of high skill with low union coverage. Those in the outlier group may contain doctors, attorneys, business leaders, etc.

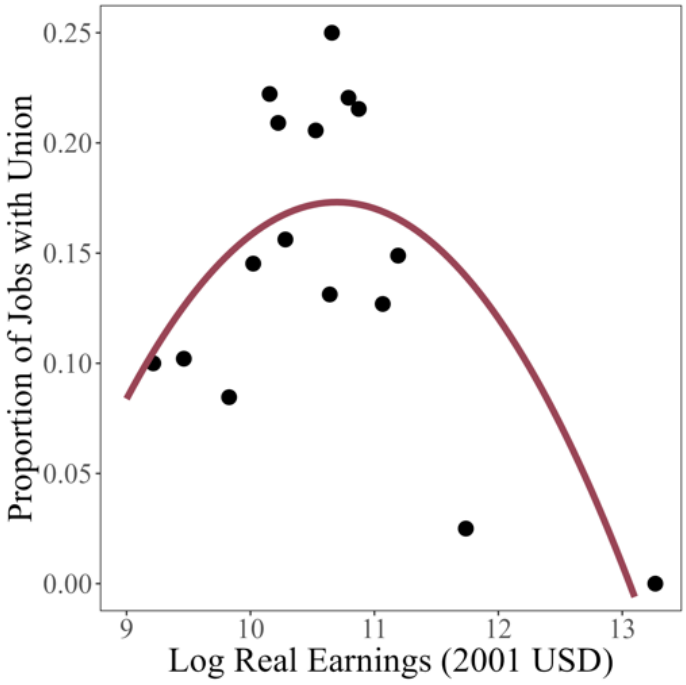


Figure 3.5: Group means of proportion/log earnings pairs showing concave relationship.

3.5.2 Income and Democracy

In this section I revisit Acemoglu et al. [2008] where they incorporate country fixed-effects to study the link between income and democracy in nations. They find that the positive effects of income on democracy vanish with the inclusion of country fixed effects arguing that this is consistent with unobserved historical factors that initiated divergent economic and political paths across nations. The purpose of this section is to make comparisons and find insights on how to interpret results using GFE or WGFE.

As in Bonhomme and Manresa [2015], I consider a regression of the Freedom House index of democracy on time-lagged democracy, lagged log-GDP per capita and a grouped fixed-effect term $\alpha_{g_i t}$, which captures group-specific time varying heterogeneity within group $g_i = g$. The model is written as

$$democracy_{it} = \theta_1 \, democracy_{it-1} + \theta_2 \, logGDPpc_{it-1} + \alpha_{g_i t} + u_{it}.$$

Table 3.4 shows the GFE and WGFE estimates and standard errors of the lagged democracy and income coefficients over different total number of groups G using the 1970-2000 balanced subsample of Acemoglu et al. [2008]. First note the decrease in both estimates from the pooled estimate, which is consistent with positive correlation between groups and lagged democracy due to historical/political events. WGFE standard errors are smaller than the GFE standard errors of the estimator of the lag term coefficient with the exception of $G = 2$. However, for the income term in larger group numbers the standard errors for GFE are smaller. This can likely be attributed to multicollinearity between the unobservable and covariates. In other words, the latent factor that is uncovered under WGFE is more correlated with income, which was also found by Kim and Wang [2019] and Aguilar Loyo and Boot [2024]. As for the parameter estimates, the WGFE estimates of the lagged democracy term appear to be more robust to the number of groups G.

In Figure 3.6 I plot the WGFE and GFE time effects in the case of $G = 4$ groups. We see heterogeneous time patterns that are clearly distinct from each other. We also see two, well separated time paths, which are known as the low and high democracy groups following the convention of previous work. The high-democracy groups include most European Union countries, the USA, and also Colombia, Japan and India. The low-democracy groups are comprised of China, Iran, and many African countries. These two groups run parallel to each other across time.

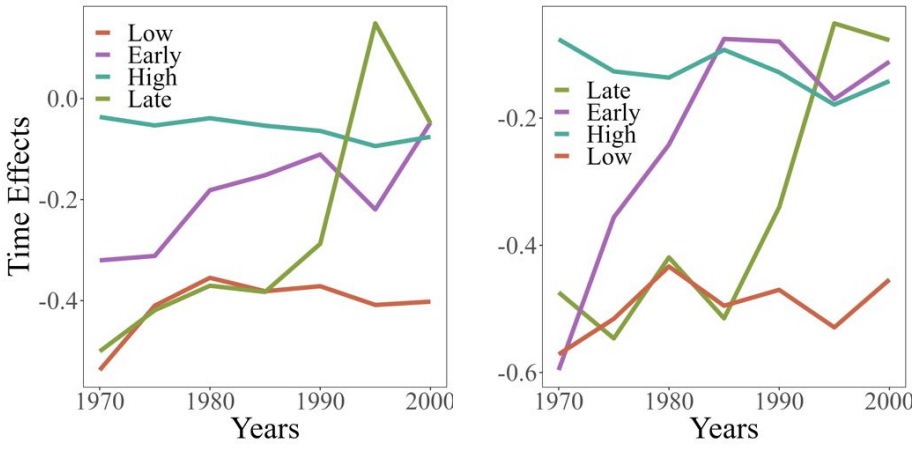
In contrast, there are two additional groups detected that exhibit transitions from low-democracy to high-democracy in the sample time periods. Keeping with convention, I identify the group that makes the transition sooner as the early-transition group and the

other as the late-transition group. As for the income and democracy plots, the WGFE assignments of groups display more positive correlation between the latent variable and income than GFE. This might explain the multicollinearity issue raised for the coefficient estimate of income using the WGFE estimator.

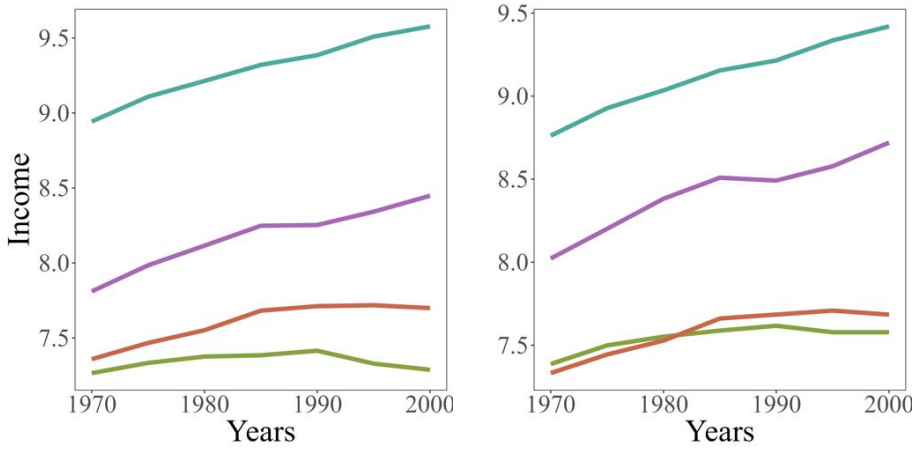
	Objective		Lagged Dem. Term (θ_1)		Income Term (θ_2)	
G	WGFE	GFE	WGFE	GFE	WGFE	GFE
1	–	–	–	0.665 (0.049)	–	0.083 (0.014)
2	0.1719	19.847	0.554 (0.045)	0.600 (0.044)	0.062 (0.0095)	0.061 (0.0118)
3	0.1522	16.599	0.403 (0.048)	0.407 (0.055)	0.070 (0.0089)	0.089 (0.0113)
4	0.1415	14.319	0.425 (0.051)	0.302 (0.058)	0.065 (0.0091)	0.082 (0.0098)
5	0.1325	12.593	0.455 (0.047)	0.255 (0.048)	0.062 (0.0089)	0.079 (0.0086)
6	0.1252	11.132	0.539 (0.034)	0.465 (0.046)	0.042 (0.0083)	0.064 (0.0075)
7	0.1182	10.059	0.483 (0.038)	0.403 (0.0423)	0.040 (0.0080)	0.065 (0.0079)
FE	–	–	–	0.284 (0.058)	–	−0.044 (0.069)

Table 3.4: WGFE and GFE estimates of θ_1 and θ_2 over number of groups $G = 1, \ldots, 7$, and fixed-effects (FE). Standard errors in parenthesis calculated via formulas in Appendix B.2. For reference, the results of the pooled regression ($G = 1$) is reported under the GFE columns.

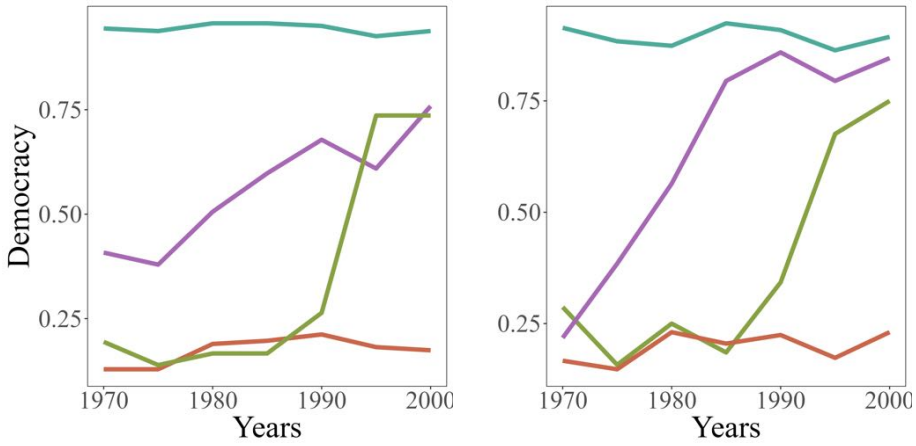
The group assignments are significantly different between the two methods, however both uncontroversially detect Argentina, South Korea, Spain and Greece as early transitioners and Panama, Romania and Taiwan as late transitioners. Differences of the WGFE



(a) Time effects α estimated via Left: WGFE, Right: GFE



(b) Average income within groups estimated via Left: WGFE, Right: GFE



(c) Average democracy index within groups estimated via Left: WGFE, Right: GFE

Figure 3.6

assignments come from the early and late transitioners poaching members from either the high-democracy or low-democracy groups. WGFE may provide arguably more consistent groupings based on historical account. For instance, the Dominican Republic had unstable democratic institutions throughout the 1960's and their first elected president was the proxy president for their last dictator, who remained in power until 1978. His regime was marked by human rights and civil abuses with restrictions were placed on opposition parties. Despite these factors and how the Freedom index measures democracy, GFE assigns the Dominican Republic to the high democracy group while WGFE places the country in the early transitioners. This and some other examples (Greece, Cyprus and Turkey) display a potential advantage of WGFE assignment detecting groupings more robust to the instability that might be found in transitioning groups. For a complete list of assignments and differences between WGFE and GFE see Appendix B.1, Table B.2. I also report results from alternative specifications of $G = 2, 3, 5$ in Figure B.4.

Group	High	Early	Late	Low
P_g	26	13	18	33
σ_g	0.133	0.213	0.212	0.181

Group	High	Early	Late	Low
P_g	27	29	12	22
σ_g	0.186	0.293	0.217	0.208

Table 3.5: GFE (Left) and WGFE (Right) estimates of group sizes and variances.

Figure 3.5 displays estimates of group sizes and variances. Concerning variances, the late and early transitioners are found to be significantly different with WGFE estimates where the early transitioners have the most variable democracy outcomes. As for group sizes, Huntington's theory suggests democracy was on the rise in this period, so we should observe that most countries are either democracies or early transitioners. From this figure and Table B.2 we see that democracy under WGFE is more dominant in

the sample period with more early transitioners and high democracies, as opposed to predictions from GFE which displays much less early transitioners and a large group of late transitioners. WGFE agrees more with Huntington's theory in this sense.

In Figure 3.7 we see more local clustering of democracies and early transitioners under WGFE assignments. South America seems to be immersed in transition while Africa being unstable in the sample period having a mix of early, late transitioners, and low democracy countries. The Eastern Mediterranean contains Turkey, Greece and Cyprus all transitioning as opposed to having Turkey as a high democracy country. In Asia, transitioners sandwiched by low democracy China and high democracies Australia, India and Japan. I emphasize the absence of any structure imposed on groupings so the apparent contiguity of democracy is all the result of estimation.

3.6 Concluding Remarks

This chapter extended the linear grouped fixed effects model to incorporate heteroskedasticity stemming from the unobserved groups. Simulations showed that ignoring this feature of the data leads to erroneous classification of individuals, harming finite-sample performance of the parameter estimators. The empirical applications showcased that the WGFE estimator can detect more compelling groupings in some space, e.g., skill-wise or geographically. Identification of groups required separability of group fixed effects as a function of the difference in group variances. A quasi-likelihood ratio test of group heteroskedasticity was proposed.

Extensions of the WGFE estimator to a linear model of group heterogeneous parameters is straightforward. For nonlinear models, the separability assumption must be modified as in Mugnier [2022a] for the case of single index models. Heteroskedasticity from groups or other forms of estimated latent variables may pose issues in a large sample, specifically I leave the topic of optimal generalized method of moments under group heteroskedasticity for future work.

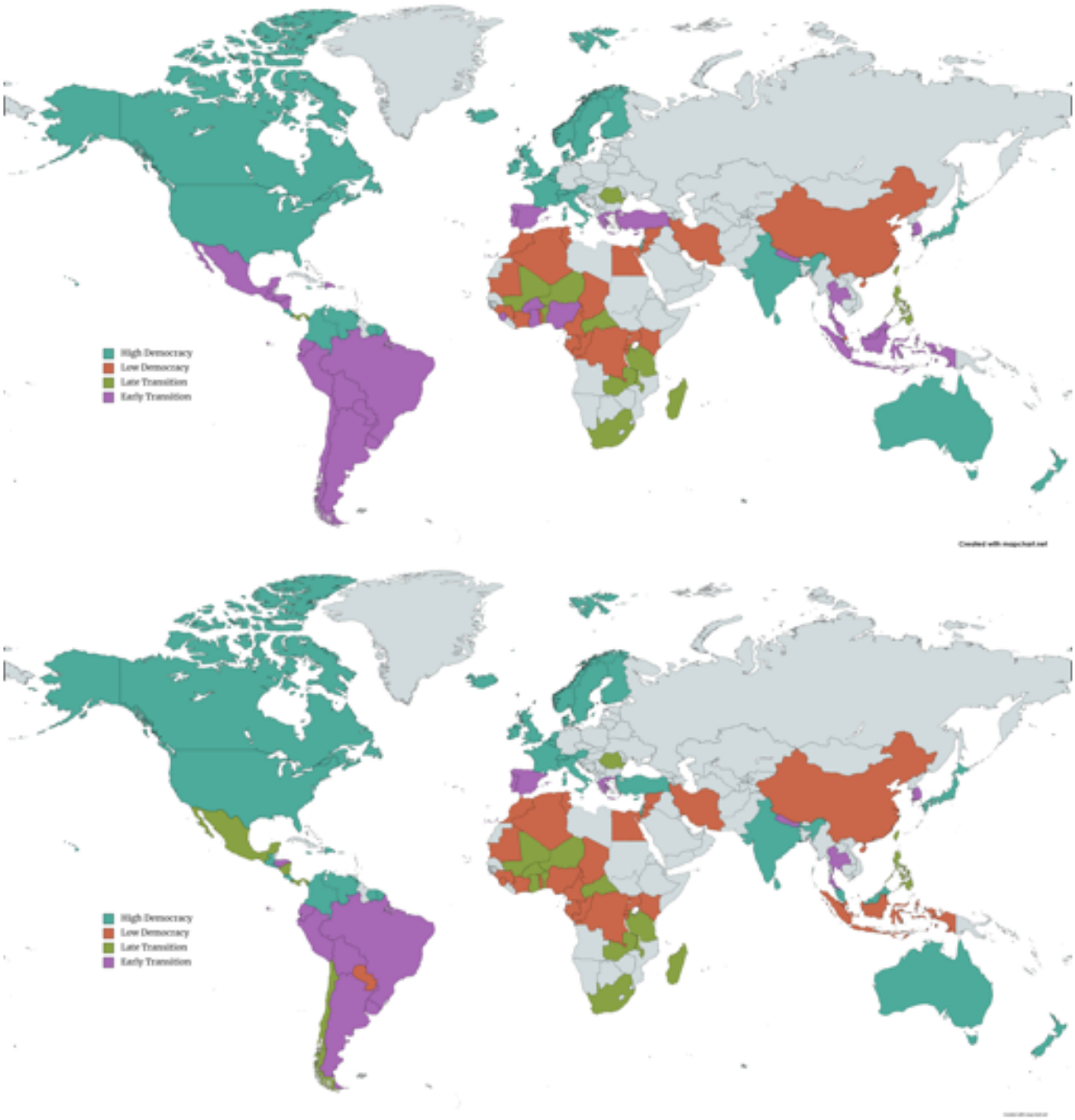


Figure 3.7: $G = 4$ map with Top: WGFE groups, and Bottom: GFE groups

Chapter 4

TYPE FIXED EFFECTS AND RATIONAL ADDICTION: A GMM FRAMEWORK FOR LATENT TYPE HETEROGENEITY

4.1 Introduction

The consumption of harmful addictive goods is a paradox involving rational and self-destructive behavior, yielding brief benefits to the consumer while causing prolonged damage to their health. The Rational Addiction (RA) framework initiated by Becker and Murphy [1988] reconciles this dichotomy by incorporating awareness of these beneficial and harmful effects. It carries important policy implications, such as endorsing taxes as a public health measure since it predicts consumers are sensitive to price changes in the long run. However, while the majority of health economists view its real world implications favorably, its empirical support is recognized as weak (Melberg [2009]).

Acknowledging the complicated diversity of addiction across individuals, I introduce unobserved type heterogeneity in a RA model that induces type-heterogeneous effects and type-specific time series for each consumer type. The type heterogeneity framework is flexible yet parsimonious by assuming individuals with similar types have similar type parameters, rather than considering complete heterogeneity of parameters. Estimation provides evidence that cigarette users exhibit cyclical consumption patterns across years and are less price sensitive than previously thought. This implies that policy aiming to educate on the dangers of use, may lengthen periods of temperance, similar to successes in addressing binge-eating disorders (McElroy et al. [2015]).

The basic RA model is of a rational finite-lived forward-looking representative consumer of an addictive good that is aware of the potential harm (or benefit) from consumption. Past consumption accumulates a "consumption capital" or addiction "stock"

that incurs disutility and depreciates over time. The agent follows a reinforcement mechanism known as "adjacent complementarity" where the marginal utility of current consumption increases with past consumption and also displays tolerance, which amounts to diminishing marginal returns of past consumption. This is the most popular iteration of the RA model as it incorporates important features of addiction while remaining simple.

The model with additional assumptions yields a linear consumption plan that depends on consumption in the previous and subsequent period. Along with covariates and a fixed effects specification, this is the most commonly estimated RA model, which was first proposed in Becker et al. [1994]. Both state-level and household studies offer similar conclusions: consumers display saddle-point dynamics, consuming high/low amounts early in their lifetimes, approach an equilibrium, and then return or exceed previous levels of consumption for the rest of their lives. My claim is that the empirical view of rational addiction has been narrow and expanding the scope of dynamics and heterogeneity would offer additional insights on demand for addictive goods.

The theory of rational addiction provides conditions for cyclical consumption, but to my knowledge has not been detected empirically. Dockner and Feichtinger [1993] show that if there are two addictions stocks accumulating based on past consumption, then the consumer may display cyclicality. As an example, there may be the traditionally considered habituation stock that displays adjacent complementarity as well as an additional "poor health" stock that displays "adjacent substitutability", meaning that marginal utility will fall when health begins to deteriorate. We should expect the habituation stock to be discounted at a larger rate than health so that eventually the harm surpasses the benefits from satisfying cravings and consumption falls. When health improves, the habituation stock may once again drive consumption to higher levels, implying cyclical consumption patterns. In addition, nothing is preventing both stocks from exhibiting adjacent complementarity, thus simultaneously reinforcing future consumption. Therefore, without controlling for this unobserved time-varying process there is potential for

bias since past consumption is summarizing two stock variables that are at odds, where a common time factor or individual fixed effect will not suffice for individual-specific time-varying effects.

Household data is preferred over aggregate data since the unit of interest is the individual's behavior and addiction. In addition, aggregate data are usually short panels, which may lead to inconsistent estimators for dynamic panel models (Baltagi and Griffin [2001]). For example, Auld and Grootendorst [2004] demonstrate how aggregate studies may be misleading by estimating an RA model to show milk consumers appear more addicted than cigarette users. Nevertheless, despite using household data, the implied saddle-point dynamics do not explain why many cigarette consumers quit and then return to habit, or exhibit binging throughout their lifetime. The inclusion of a second addiction stock may rectify this issue, but unobserved differences in preferences might also be a driving factor.

These dynamics are associated to the pharmacology of nicotine dependence (Benowitz [2010]) where after periods of abstinence the sensitivity to rewards from nicotine are refreshed, facilitating relapse. The effects of nicotine dependence manifest very differently across the population, for instance, Saunders et al. [2022] survey 3.4 million people and found almost 4,000 genetic variants associated to the use of tobacco, which in addition to cultural norms pose a high degree of unobserved heterogeneity in the data. They may also impact forward looking decision making, sensitivity to prices, and age-related factors that impact consumption in the life-cycle; see Grant et al. [2010] and Carroll [2021]. This discussion alludes to the potential for there being many heterogeneous consumer types with corresponding addiction processes and responses impacting their consumption of addictive goods.

I address this by augmenting the RA model with "type heterogeneity" written as:

$$C_{it} = \alpha_t(\xi_i) + \theta_l(\xi_i)\, C_{it-1} + \theta_f(\xi_i)\, C_{it+1} + \beta(\xi_i)' X_{it} + U_{it} \tag{4.1}$$

where C_{it} denotes consumption of the addictive good, X_{it} are observable characteristics

such as price, income, race and age, and U_{it} is an error term. Note that C_{it-1} and C_{it+1} are both endogenous and I assume that valid instruments Z_{it} are available.

The variable ξ_i is a continuous, time-invariant and unobserved variable I call a "type" that belongs to a compact subset $\Xi \subset \mathbb{R}$ and randomly sampled along with observables from some joint density. All covariates and instruments may be arbitrarily correlated with this type and thus with the type-specific parameters $\alpha_t(\xi_i)$ $\theta_l(\xi_i)$, $\theta_f(\xi_i)$ and $\beta(\xi_i)$. I call a panel model with unobserved types, type-specific parameters, and arbitrary correlation structure with observables a type fixed effects (TFE) model. To my knowledge this is the first paper to propose such a model.

The TFE RA model is a flexible way to incorporate complex forms of consumer heterogeneity towards addictive goods without assuming complete heterogeneity of parameters over individuals and time. The type variable ξ_i governs the partial effects $\theta(\cdot)$ of covariates on consumption, reflecting heterogeneous preferences. This determines a dynamic profile of effects from the testable implications of the RA model for each individual allowing for some segments of the population to follow saddle-point dynamics while others are in cycles. The time-varying term $\alpha_t(\xi_i)$ is known as the type fixed effect and captures dynamic-idiosyncratic effects such as the variation from an alternative latent addiction stock driving consumption. To my knowledge, this is the first paper to propose estimating a RA model with many heterogeneous parameters and to consider heterogeneous consumption dynamics.

A key identification condition for this model is that those of similar types must have similar parameter values. The parameters are regarded as curves (functions) parametrized by the type and in this sense must be sufficiently smooth to allow such comparisons. This is similar to semiparameteric models with varying coefficients (Hastie and Tibshirani [1993]), however the coefficients vary according to an unobservable. Along these lines, $\alpha_t(\cdot)$ for $t = 1, \ldots, T$ can be thought of as a nonparametric specification of the functional form of the heterogeneity, where important examples are two-way fixed effects and interactive fixed effects (see Wooldridge [2010] and Bai [2009]). This observation along

with the extension of varying coefficients to an unobserved argument and parameter with dimension growing with the sample are two contributions to the semiparametric literature.

In Section 4.2, I introduce the TFE-GMM framework for models with type heterogeneity, discuss identification of types, define an estimator, and provide a computational algorithm based on stochastic gradient descent. The model (4.1) is a special example of this framework . In Section 4.3, I give conditions for consistency of the TFE-GMM estimator for parameters of a TFE linear panel model with endogenous covariates and discuss the consistency for types. In Section 4.4, I present simulation results indicating good performance of the estimator under various data generating processes. In Section 4.5, I provide results from estimating the TFE RA model for cigarette consumption. Most individuals in the sample are following cyclical patterns and are less price sensitive than previous studies suggest.

4.1.1 Related Literature

Cawley and Ruhm [2011] surveys models of addiction and habit that includes the RA model. The model has been extensively applied to many goods considered addictive in both aggregate and micro panel studies such as tobacco, alcohol and coffee (Becker et al. [1994], Grossman et al. [1998] and Olekalns and Bardsley [1996], respectively). The RA model is typically tested against the alternative that the coefficients of the lead and lag consumption covariates are positive, so that if the null is rejected then the consumer is taken to follow addiction (lag) and is forward looking (lead). Household studies are preferred since they capture individual behavior more closely and can be aggregated to describe municipality and state-level dynamics (Chaloupka [1991], Grossman et al. [1998]). They also tend to produce more plausible estimates of the discount factor, but results remain mixed. Laporte et al. [2017] argue with simulation evidence that saddle-point

dynamics may cause identification issues when an unstable root is dominant, meaning stability can't be used to pin down values of parameters. Unstable roots are a feature of dynamic micro panel models and so they claim it may be difficult to estimate RA models in general. Considering type heterogeneity allows consumers to display different dynamics and stability properties, which may give more credibility to estimates. The closest application of RA to this paper is found in Fernández-Val and Lee [2013], but they do not include unobserved time-varying fixed effects and heterogeneity of the lead and lag of consumption thereby excluding heterogeneous dynamics.

The TFE model is related to random coefficients models, except it includes both individual and time-varying heterogeneity, see Hsiao [2022] for a text on these and other panel data methods. It is well known that ignoring parameter heterogeneity by imposing fixed coefficients generally results in inconsistent estimation of the mean of random coefficients (Yitzhaki [1996], Heckman and Vytlacil [1998], Angrist et al. [2000], Angrist [2004]) and moreover denies estimation of the other features of the coefficients. However, identification requires additional care such as limiting the degree of heterogeneity of the random coefficients, which may not be justified empirically and in theory (Arellano and Bonhomme [2012], Graham and Powell [2012], Laage [2020]). Nonparametric and semiparametric techniques are prevalent in economics in part for their robustness to misspecification at the cost of requiring a large sample, see Unit 1 of Li and Racine [2007] for kernel smoothing techniques. This is not the first application of a panel model of varying coefficients, however it is the first to include parameters that vary according to an unobservable and also vary across time, see Hoover et al. [1998] and Fan and Zhang [2000] that focus on the case of coefficients as unspecified functions of time. The TFE-GMM criterion function is similar to the criterion of Fernández-Val and Lee [2013], which is based on the two-step procedure of Hansen [1982] but is aggregated on the cross-sectional level using each individual unit's time series GMM criteria. There is finite-sample bias for GMM and this does not exclude the TFE-GMM estimator, however there does exist bias correction measures that may be applied to this case (Newey and

Smith [2004]). Incidental parameter bias may affect estimates of dynamic models using short panels and model (4.1) is no exception; see Neyman and Scott [1948], Chamberlain [1980] and Nickell [1981] and Arellano and Hahn [2007b] for bias correction approaches for fixed effects models.

There has been a large focus on discrete types, commonly known as groups, with and without time-varying heterogeneity (Sun [2005], Chang-Ching and Serena [2012], Bester and Hansen [2016] among others). Bonhomme and Manresa [2015] introduce the grouped fixed effects (GFE) model and estimation where the heterogeneity forms an unobserved group structure and allow for slope parameters to vary across groups in addition to a group time-varying heterogeneity term similar to TFEs. Identification of groups is similar to identification of types requiring separability in the model, specifically for GFE, that the group fixed effects terms are different in the mean squared sense. Other estimators of this model have been proposed and require this assumption (Chetverikov and Manresa [2022], Mugnier [2022b]) and when there is unobserved group heteroskedasticity there are steeper conditions needed in order to identify groups; see Rivero [2023] (Chapter 3) that requires group fixed effects separation as a function of group variances. Su et al. [2016] and Mehrabani [2022] consider linear and nonlinear models with unknown group structure where the random coefficients are heterogeneous across groups, but homogeneous across individuals within the same group. For linear models with endogeneity they specialize to a penalized GMM (PGMM) estimation framework that contributes to the fused-Lasso literature where some of the individual coefficients share the same value, hence forming groups through penalizing coefficients into clusters. Cheng et al. [2019] estimate a model of time-invariant multi-group heterogeneity and covariates that are endogenous despite the groups. My proposal can be viewed as a continuous extension of the discrete case where, instead of grouped patterns of heterogeneity shared among those in the same group, we have type heterogeneity that is shared within types.

The discreteness assumption may be viewed as too strict and the TFE model may be

more flexible in this regard. Additionally, not much is known of the consequences of violating the group structure assumption or how to control for a continuous latent variable of this kind. One example is Bonhomme et al. [2022] who propose a two-step procedure by first discretizing the latent variable by clustering moments of observables that are informative of types and then estimating parameters and the time-varying heterogeneity terms via maximum likelihood. This approach is particularly useful for nonlinear models where identification can be troublesome. I approach the problem of a continuous variable head-on and avoid the need for additional auxiliary moments by simply relying on the moments that identify parameters, which are smooth curves parametrized by types.

First initiated by Robbins and Monro [1951], the stochastic gradient descent (SGD) algorithm is an incredibly popular technique in machine learning due to the size of data sets and wide applicability, see Bottou [2010]. The version of stochastic gradient descent proposed in this paper is related to the k-means algorithm (Forgy [1965], Lloyd [1982]) where the assignment step in this case is a "soft" assignment that puts a weights on observations depending on relative positions of types of other observations, locally estimating heterogeneous parameters based on proximity in the type space Ξ. SGD can be applied to many extremal estimation problems since all that is required is the form of the gradient or subgradient of an objective function, for example see Lee et al. [2023] which involves quantile regression where there the non differentiability of the criterion is not an obstacle.

4.2 GMM Framework with Type Heterogeneity

Denote $i \in \{1, \ldots, N\} = \mathcal{N}$ and $t \in \{1, \ldots, T\} = \mathcal{T}$ as the index of individuals and the index of the observations of the individuals, respectively. Suppose we have a balanced panel data set $\{w_{it}\}$ with support $\mathcal{W}$ that is independent and identically distributed (iid) over $i \in \mathcal{N}$ from a density function f, with bounded fourth moments, and stationary

and strongly mixing over $t \in \mathcal{T}$ with mixing coefficients that decay exponentially. Let $\theta \in \Theta$ and let $\alpha = (\alpha_1, \ldots, \alpha_T) \in \mathcal{A}^T$ be infinite-dimensional parameters both defined as functions on some common set $\Xi \subset \mathbb{R}$. Both of these parameters are unknown functions of types, which are the realizations of an unobserved random variable ξ_i that has support on the interior of the type space $\mathring{\Xi}$ according to density ν and may be arbitrarily correlated with some or all elements of w_{it}. The set of moment conditions will depend on the true types $\{\xi_i^0\}_{i \in \mathcal{N}}$ that are assumed to be iid draws from a type density ν^0. Let $x \mapsto \|x\|$ denote the standard Euclidean norm for vectors or the L^2 norm for functions, whichever is applicable.

For exposition, we distinguish two kinds of parameters: the type-specific parameters that enter the model directly (θ and α) and the unobserved types ξ_i. The key strategy for identification is to treat the two types of parameters separately by first assuming types are known, then identifying the type parameters, and then establishing the converse. This defines a system with the true parameters as unique solutions. A similar argument can be found in Bai [2009] for interactive fixed effects models. The endogenous linear model with type fixed effects is an important example for studying rational addiction and will be discussed in the context of the following GMM framework, which include the necessary identifying assumptions.

4.2.1 Identification of type-specific parameters

Assume first that the true types ξ_i^0 are known and suppose that the true parameters $\left(\theta^0(\xi_i^0), \alpha^0(\xi_i^0)\right)$ are (over) identified by the conditional moment conditions:

$$\mathbb{E}\left[g(w_{it}; \theta^0(\xi_i^0), \alpha_t^0(\xi_i^0)) \middle| \xi_i^0\right] = 0 \tag{4.2}$$

for all $(i,t) \in \mathcal{N} \times \mathcal{T}$ where $g : \mathcal{W} \times \Theta \times \mathcal{A} \to \mathbb{R}^\ell$ are known functions with $\ell \geq p$. Suppose further that the infinite-dimensional time parameters α_t are just identified by:

$$\mathbb{E}\left[\rho_t(w_{it}; \theta^0(\xi_i^0), \alpha_t^0(\xi_i^0)) \middle| \xi_i^0\right] = 0 \tag{4.3}$$

for all $t \in \mathcal{T}$ and, ν^0-almost surely, for all $\xi \in \Xi$ and where $\rho_t : \mathcal{W} \times \Theta \times \mathcal{A} \to \mathbb{R}$ is known. That is to say the conditions (4.3) are also included in (4.2). I assume that for all $t \in \mathcal{T}$ the function ρ_t is strictly monotonic with respect to α_t so that there exists a function φ_t defined by

$$\alpha_t^0(\xi_i^0) = \varphi_t(\xi_i^0; \theta^0) \tag{4.4}$$

where φ_t depends on the sample information at time t. We can solve for the infinite-dimensional parameters by using (4.3) to get (4.4) and essentially substitute it in the moment conditions (4.2) for θ^0. Note that for fixed $i \in \mathcal{N}$ the type ξ_i^0 is a constant random variable so that (4.2) can be written as

$$\mathbb{E}\left[g(w_{it}; \theta^0(\xi_i^0), \alpha_t^0(\xi_i^0))\right] = 0, \text{ for any } i \in \mathcal{N}, \tag{4.5}$$

which is an expectation over individual i's time series. This property allows inspection of an individual i's time series information to extract their true type ξ_i^0, which will be important in Section 4.2.2 for identification of types.

Example 6 (Endogenous linear panel model with type fixed effects)**.** Suppose that $\{w_{it}\} = \{(y_{it}, x_{it}, z_{it})\}_{(i,t)\in\mathcal{N}\times\mathcal{T}}$ and that iid types $\{\xi_i^0\}_{i\in\mathcal{N}}$ are observed. Consider the structural linear model with type heterogeneity

$$y_{it} = x_{it}'\theta(\xi_i^0) + \alpha_t(\xi_i^0) + u_{it} \tag{4.6}$$

where for all $i \in \mathcal{N}$ the covariates x_{it} are contemporaneously correlated with ξ_i^0, $\alpha_t(\xi_i^0)$, u_{it} and $\theta(\xi_i^0)$. We assume that we have weakly exogenous instruments z_{it} and that types ξ_i^0 are also exogenous. Specifically, $\mathbb{E}\left[u_{it}|\xi_i^0, \alpha_t(\xi_i^0)\right] = 0$, $\mathbb{E}\left[z_{it}u_{it}|\xi_i^0\right] = 0$ and $\mathrm{Cov}(z_{it}, x_{it}|\xi_i^0) \neq 0$ for all $(i,t) \in \mathcal{N} \times \mathcal{T}$[1]. It is useful to write the model (4.6) as

$$y_{it} = x_{it}'\theta(\xi_i^0) + \alpha(\xi_i^0)\delta_t + u_{it} \tag{4.7}$$

where δ_t is the Kronecker delta function (dummy variable) defined by $\delta_t = (\mathbb{1}\{t = s\})_{s=1}^T$, i.e., a vector of zeros in each entry except for a 1 in entry t. Since α_t is an

[1]Weak dependence assumptions can also be made to guarantee consistency, see Section 4.3.

exogenous variable, we use δ_t for all $t \in \mathcal{T}$ as an instrument and allow the non constant elements of z_{it} be correlated to $\alpha_t(\xi_i^0)$. Therefore, $\ell = K + T$ where $K \geq p$. Note that the type ξ_i^0 drives the correlation between covariates and the parameters of the model[2].

The approach of Robinson [1988] for identification of partially linear models is used to identify $\theta(\xi_i^0)$ for any $i \in \mathcal{N}$ and highlight the connection to traditional fixed effects. Taking conditional expectations:

$$\mathbb{E}\left[y_{it}|\xi_i^0\right] = \mathbb{E}\left[x_{it}|\xi_i^0\right]' \theta(\xi_i^0) + \alpha_t(\xi_i^0) \tag{4.8}$$

and then differencing out the type conditional means

$$y_{it} - \mathbb{E}\left[y_{it}|\xi_i^0\right] = \left(x_{it} - \mathbb{E}\left[x_{it}|\xi_i^0\right]\right)' \theta(\xi_i^0) + u_{it} \tag{4.9}$$

eliminates the type fixed effects and can be regarded as a "within-type" transformation.

Let $\widetilde{y}_{it} = y_{it} - \mathbb{E}\left[y_{it}|\xi_i^0\right]$ and $\widetilde{x}_{it} = x_{it} - \mathbb{E}\left[x_{it}|\xi_i^0\right]$. Provided that the usual rank conditions hold, i.e., $\operatorname{plim} \sum_{t=1}^T z_{it} z_{it}'$ and $\operatorname{plim} \sum_{t=1}^T z_{it} \widetilde{x}_{it}'$ are of full rank for any $i \in \mathcal{N}$, then $\theta^0(\xi_i^0)$ is (over) identified.

The curve $\theta^0(\cdot)$ is also over identified by

$$\mathbb{E}\left[z_{it}\left(y_{it} - \mathbb{E}\left[y_{it}|\xi_i^0\right]\right)\middle|\xi_i^0 = \xi\right] = \mathbb{E}\left[z_{it}\left(x_{it} - \mathbb{E}\left[x_{it}|\xi_i^0\right]\right)\middle|\xi_i^0 = \xi\right]' \theta^0(\xi) \tag{4.10}$$

for any $t \in \mathcal{T}$ and ν^0-almost surely $\xi \in \Xi$ provided full rank conditions conditional on types hold. In other words there must be sufficient variation within-types across time. Since the types must be estimated simultaneously with parameters, stronger identification conditions will be required as shown in Section 4.3.

Relating back to the moment conditions (4.2) gives us

$$g(w_{it}; \theta(\xi_i^0), \alpha_t(\xi_i^0)) = z_{it}\left(y_{it} - x_{it}'\theta(\xi_i^0) - \alpha_t(\xi_i^0)\right) \tag{4.11}$$

$$\rho_t(w_{it}; \theta(\xi_i^0), \alpha_t(\xi_i^0)) = y_{it} - x_{it}'\theta(\xi_i^0) - \alpha_t(\xi_i^0), \tag{4.12}$$

[2]A connection can be made from these type coefficients to stationary random coefficients: $\beta(\xi_i^0) = \beta + \xi_i^0$ where $\xi_i^0 \sim (0, \sigma^2)$ iid so that the mean and variance of the coefficients are the constants $\beta \in \mathbb{R}^p$ and $\sigma^2 \geq 0$, respectively.

for all $(i,t) \in \mathcal{N} \times \mathcal{T}$, and

$$\alpha_t(\xi_i^0) = \mathbb{E}\left[y_{it}|\xi_i^0\right] - \mathbb{E}\left[x_{it}|\xi_i^0\right]' \theta(\xi_i^0) \tag{4.13}$$

$$= \varphi_t(\xi_i^0, \theta(\xi_i^0)) \tag{4.14}$$

for all $t \in \mathcal{T}$. This expectation is over the cross-sectional dimension so it is understood as an "average" over individuals with type ξ_i^0 at time t.

4.2.2 Identification of types

The identification of types amounts to distinguishing time series patterns between individuals. Towards this, I introduce notation to accommodate the individual's time series GMM criteria:

$$g_i(\theta^0(\xi), \alpha^0(\xi)) = \frac{1}{T}\sum_{t=1}^{T} g(w_{it}, \theta^0(\xi), \alpha_t^0(\xi)). \tag{4.15}$$

Fix $i \in \mathcal{N}$ and assume we know the true parameters and the type space Ξ. Then, given individual i's time series moment condition (4.5) and exponentially decaying mixing coefficients, we have $\text{plim}_{T\to\infty} g_i(\theta^0(\xi), \alpha^0(\xi)) = 0$. With this we can identify their type ξ_i^0 with a simple rule made valid by an invertibility condition on the moment functions.

Assumption 6. *For any $i \in \mathcal{N}$,*

$$\operatorname*{plim}_{T\to\infty} \left\| g_i(\theta^0(\xi), \alpha^0(\xi)) - g_i(\theta^0(\widetilde{\xi}), \alpha^0(\widetilde{\xi})) \right\| = 0$$

if and only if $\xi = \widetilde{\xi}$.

This assumption allows the individual's GMM time series moments to be informative of types, all else equal. The following uses this and (4.5) to define a rule to assign types to each individual.

Lemma 1. *For all $i \in \mathcal{N}$ and provided* (4.2) *and Assumption 6 hold, the true realized type is $\xi_i^0 = \xi$ if and only if, for any $\widetilde{\xi} \in \Xi$ such that $\widetilde{\xi} \neq \xi$,*

$$0 = \operatorname*{plim}_{T\to\infty} \left\| W^{1/2} g_i(\theta^0(\xi), \alpha^0(\xi)) \right\|^2 < \operatorname*{plim}_{T\to\infty} \left\| W^{1/2} g_i(\theta^0(\widetilde{\xi}), \alpha^0(\widetilde{\xi})) \right\|^2 \tag{4.16}$$

where W is any symmetric positive definite matrix that does not depend on Ξ[3].

This lemma implies the conditional distribution of the true types given the time series $w_i = \{w_{it}\}_{t\in\mathcal{T}}$ is a degenerate distribution around the function:

$$F(w_i;\theta^0,\alpha^0) = \underset{\xi\in\Xi}{\operatorname{argmin}} \operatorname*{plim}_{T\to\infty} \left\| g_i(\theta^0(\xi),\alpha^0(\xi)) \right\|^2 \tag{4.17}$$

where the weight matrix W is dropped since the value of the minimum is zero due to the moment condition being satisfied. In view of this, let $\nu^0(\xi|w_i) = \delta(\xi - F(w_i;\theta^0,\alpha^0))$, where δ is the the Dirac delta function defined by the property $\int h(x)\delta(x)\,dx = h(0)$ for continuous function h with compact support or rapidly shrinking tails[4].

The density ν^0 can then be derived under some additional smoothness conditions on the moment functions g and on the parameters (θ^0,α^0). Such conditions guarantee a differentiable F via the implicit function theorem and allow for a change-of-variables to apply, which defines a push forward mapping from the observables to the type space.

Assumption 7 (Smoothness)**.** *The following hold:*

a. The set $\Xi \subset \mathbb{R}$ is connected and compact and its interior $\mathring{\Xi}$ is the support of ν^0.

b. There exists a constant $M > 0$ such that for all $\xi \in \mathring{\Xi}$ we have $\left\|\frac{\partial^2\theta^0(\xi)}{\partial\xi^2}\right\| < M$, $\operatorname{plim}_{T\to\infty} T^{-1}\left\|\frac{\partial^2\alpha^0(\xi)}{\partial^2\xi}\right\|^2 < M$, $\left\|\frac{\partial\theta^0(\xi)}{\partial\xi}\right\| > 0$, and $\operatorname{plim}_{T\to\infty} T^{-1}\left\|\frac{\partial\alpha^0(\xi)}{\partial\xi}\right\|^2 > 0$.

c. The second partial derivatives of $g(w,\theta^0(\xi),\alpha^0(\xi))$ exist and are bounded.

d. For any $i \in \mathcal{N}$, if $\xi = F(w,\theta^0,\alpha^0)$ then

$$\operatorname*{plim}_{T\to\infty} \left\| \frac{1}{T}\sum_{t=1}^{T} \frac{\partial g(w_{it};\theta^0(\xi),\alpha_t^0(\xi))}{\partial\theta^0}\cdot\frac{\partial\theta^0(\xi)}{\partial\xi} + \frac{\partial g(w_{it};\theta^0(\xi),\alpha_t^0(\xi))}{\partial\alpha_t^0}\cdot\frac{\partial\alpha_t^0(\xi)}{\partial\xi} \right\|^2 > 0$$

[3]The weight matrix may depend on the type if errors are heteroskedastic with respect to the type. In this case, identification restrictions may need to be stronger, which is a topic of Chapter 3.

[4]An example is the Gaussian density, which is itself an example of a Schwartz function that has rapidly decreasing derivatives.

e. There exists constants $a > 0$ *and* $b > 0$ *such that* $\mathbb{P}(\|w_{it}\| > m) \leq e^{1-(\frac{m}{b})^a}$ *for all* $(i,t) \in \mathcal{N} \times \mathcal{T}$ *and* $m > 0$.

Assumption 7(a) requires that the range of F is the interior of Ξ, excluding boundary cases as minima. Assumption 7(b) is a collection of smoothness conditions on the parameters. Bounded second derivatives lean on the interpretation that individuals with similar types will display similar parameter values. First derivatives bounded from zero imply that the parameters are smooth (regular) curves parametrized by the types $\xi \in \Xi$, so geometrically they will not halt or retrace themselves locally. Assumption 7(c) ensures that F is sufficiently smooth and Assumption 7(d) is a convexity condition enabling F to be of a unique minimizer, i.e., injective. Assumption 7(e) imposes a faster-than-any-polynomial tail decay property on observables, which satisfies the property associated to the Dirac delta function so the marginal density ν^0 is well-defined by the change-of-variables. These properties are sufficient to apply a change-of-variables.

Theorem 5. *Suppose that Assumption 6 and 7 holds. Then, the type density is given as*

$$\nu^0(\xi) = \int_{\mathcal{W}} f(w_i)\delta\left(\xi - F(w_i;\theta^0,\alpha^0)\right)\, dw_i. \tag{4.18}$$

Theorem 5 identifies the density of types using the individual's type rule (4.16) and smoothness of parameters. It is important to verify on a case-by-case basis that Assumptions 6 and 7 hold. For the endogenous linear model (4.6) the moment functions are known to be sufficiently smooth, but conditions for Assumption 6 must be found.

Example 6 (Continued)**.** With knowledge of the true parameters θ^0 and α^0, the goal is to write conditions that ensure Assumption 6 holds. Using (4.11) and the model definition (4.6) substituted in for y_{it}, let $\xi, \widetilde{\xi} \in \Xi$ and $\widetilde{\xi} \neq \xi$, and assume without loss of generality

that $\xi_i^0 = \xi$. Consider the following:

$$
\begin{aligned}
& g_i(\theta^0(\widetilde{\xi}), \alpha^0(\widetilde{\xi}))' g_i(\theta^0(\widetilde{\xi}), \alpha^0(\widetilde{\xi})) - g_i(\theta^0(\xi), \alpha^0(\xi))' g_i(\theta^0(\xi), \alpha^0(\xi)) \\
&= \left\| \frac{1}{T} \sum_{t=1}^{T} z_{it} \left(u_{it} + x_{it}' \left(\theta^0(\widetilde{\xi}) - \theta^0(\xi) \right) + \left(\alpha_t^0(\widetilde{\xi}) - \alpha_t^0(\xi) \right) \right) \right\|^2 - \left\| \frac{1}{T} \sum_{t=1}^{T} z_{it} u_{it} \right\|^2 \\
&= \left\| \frac{1}{T} \sum_{t=1}^{T} z_{it} \left(x_{it}' \left(\theta^0(\widetilde{\xi}) - \theta^0(\xi) \right) + \left(\alpha_t^0(\widetilde{\xi}) - \alpha_t^0(\xi) \right) \right) \right\|^2 + o_p(1)
\end{aligned}
$$

where the $o_p(1)$ term arises from the fact that the instruments are weakly exogenous so cross-terms will vanish asymptotically as T tends to infinity. We require that this must be asymptotically bounded away from 0 whenever $\xi \neq \widetilde{\xi}$. Indeed, this is precisely the difference in moment functions of Assumption 6.

Assumption 8 (Separability)**.** *There exists a function $C : \Xi \times \Xi \to [0, \infty)$ such that:*

$$
\operatorname*{plim}_{T \to \infty} \left\| \frac{1}{T} \sum_{t=1}^{T} z_{it} \left(x_{it}' \left(\theta^0(\widetilde{\xi}) - \theta^0(\xi) \right) + \left(\alpha_t^0(\widetilde{\xi}) - \alpha_t^0(\xi) \right) \right) \right\|^2 \geq C(\xi, \widetilde{\xi}), \tag{4.19}
$$

for any $i \in \mathcal{N}$, and $C(\xi, \widetilde{\xi}) = 0$ if and only if $\xi = \widetilde{\xi}$.

Assumption 8 satisfies Assumption 6 for the endogenous linear model and is similar to group separability in the group heterogeneous coefficients case in the supplementary appendix of Bonhomme and Manresa [2015][5]. This separability condition is a departure from the others in that it involves a continuous variable and is determined by the time series dependence between observables and type fixed effects. It also rules out perfect collinearity between covariates, instruments and the type fixed effects.

I have shown for arbitrary $\widetilde{\xi} \neq \xi \in \Xi$ that the difference between GMM criterion is bounded away from zero and zero only when $\widetilde{\xi} = \xi = \xi_i^0$, for some $i \in \mathcal{N}$. Therefore, each individuals type the type density is identifiable based on individual's time series information. Finally, if we know the types we can extract the parameters and, on the

[5]Identification of groups in the panel data with discrete latent variable literature have required this assumption in some way or form, see Bonhomme and Manresa [2015], Cheng et al. [2019], Chetverikov and Manresa [2022], Mugnier [2022b,a].

other hand, if we know the parameters we can find each individuals type and type density. This defines a system that uniquely determines parameters and types.

The identification conditions do not immediately rule out lagged/forwarded outcomes or time-invariant covariates. These full rank conditions require that covariates and instruments must display sufficient variation *within types* across individuals since the expectation is taken with respect to the cross-sectional dimension unlike in the within-individual transformation of ordinary fixed effects, which transforms variables into functions of the entire time series. This is similar to group fixed effects models that require sufficient within-group variation (Assumption 1.g.) and within-factor variation (Assumption A in Bai [2009]), although they are specialized to their unknown factors so stronger conditions for within-type variation will be provided in Section 3. □

4.2.3 The TFE-GMM Estimator

The proposed population GMM criterion function is the following:

$$Q = \operatorname*{plim}_{T\to\infty} \mathbb{E}\left[g_i(\theta(\xi_i), \alpha(\xi_i))' W g_i(\theta(\xi_i), \alpha(\xi_i))\right] \geq 0 \tag{4.20}$$

with equality if and only if $\theta = \theta^0$, $\alpha = \alpha^0$ and ξ_i equal in distribution to ξ^0. Since ξ_i^0 has a density, we can constrain the set of possible densities to those that are absolutely continuous with respect to the Lebesgue measure and supported on the interior of Ξ. Using the law of iterated expectations, we can rewrite this population objective function in terms of the conditional density $\nu(\xi|w_i)$ of candidate type random variables with respect to observables:

$$Q(\theta, \alpha, \nu(\cdot|w_i)) = \operatorname*{plim}_{T\to\infty} \mathbb{E}\left[\int_\Xi g_i(\theta(\xi), \alpha(\xi_i))' W g_i(\theta(\xi), \alpha(\xi_i)) \nu(\xi|w_i)\, d\xi\right]. \tag{4.21}$$

which is zero at the true values of the parameters and $\nu(\xi|w_i) = \nu^0(\xi|w_i)$ as defined in (4.17) since Ξ is compact making this integral well-defined.

To define a sample criterion function from (4.21) I smooth the Dirac mass $\nu^0(\xi|w_i) = \delta(\xi - F(w_i; \theta^0, \alpha^0))$. Let $h > 0$ be a bandwidth that may depend on N and T and let K

be a symmetric, continuously differentiable density function. Let $\{\widehat{W}_i\}_{i\in\mathcal{N}} \subset \mathbb{R}^{\ell\times\ell}$ be a collection of positive definite weight matrices that may depend on the observable sample. I define the one-step type fixed effects GMM (TFE-GMM) estimator as the solution to

$$\left(\widehat{\theta},\widehat{\alpha},\widehat{\mu}\right) = \underset{(\theta,\alpha,\mu)\in\Theta\times\mathcal{A}^T\times\Xi^N}{\operatorname{argmin}} \frac{1}{N}\sum_{i=1}^{N}\int_{\Xi}\left\|\widehat{W}_i^{1/2} g_i(\xi;\theta,\alpha,\mu)\right\|^2 K_h(\xi-\mu_i)\,d\xi \tag{4.22}$$

where $g_i(\xi;\theta,\alpha,\mu) = T^{-1}\sum_{t=1}^{T} g(w_{it};\theta(\xi),\alpha(\xi))$ and $K_h(\xi-\mu) = h^{-1}K((\xi-\mu)/h)$.

This objective resembles the individual-separated GMM criterion of Fernández-Val and Lee [2013] and Cheng et al. [2019] as a cross-sectional conditional average of individual time series GMM criteria. This form of objective function explicitly separates each individual's GMM criteria and places weights on them through the type kernel to locally estimate the type fixed effects and the type-specific coefficients with those individuals in close proximity within the type space. Because of this, μ can be thought of as a location parameter– they cluster individuals with similar types to use their similarities in estimation of the type-dependent parameters.

Example 6 (Continued)**.** The TFE-GMM estimator for the model parameters of (4.6) is defined as the solution to

$$\min_{(\theta,\mu)\in\Theta\times\Xi^N} \frac{1}{N}\sum_{i=1}^{N}\int_{\Xi}\left\|\widehat{W}_i^{1/2}\left(\sum_{t=1}^{T} z_{it}\left(\widetilde{y}_{it}-\widetilde{x}_{it}'\theta(\xi)\right)\right)\right\|^2 K_h(\xi-\mu_i)\,d\xi. \tag{4.23}$$

where I've applied the within-type transformation:

$$\widetilde{d}_{it} = \widetilde{d}_{it}(\xi,\mu) = d_{it} - \frac{1}{N}\sum_{i=1}^{N} d_{it}\frac{K_h(\xi-\mu_i)}{\sum_{j=1}^{N}K_h(\xi-\mu_j)},\quad d\in\{x,y\}. \tag{4.24}$$

The local constant estimator appears as a plug-in for the conditional expectation given types (4.13). In principle, any local polynomial estimator could be used instead, however I consider the simplest example for discussion and computation.

Using any kernel K, the first-order conditions for $\widehat{\theta}$ can be taken by using the fact Ξ

is an interval subset of $\mathbb{R}$ yielding Euler equations that give us

$$\widehat{\theta}(\xi;\mu) = \left[\sum_{i=1}^{N}\left(\sum_{t=1}^{T} z_{it}\widetilde{x}'_{it}\right)' \widehat{W}_i \left(\sum_{t=1}^{T} z_{it}\widetilde{x}'_{it}\right) K_h(\xi-\mu_i)\right]^{-1} \times \sum_{i=1}^{N}\left(\sum_{t=1}^{T} z_{it}\widetilde{x}'_{it}\right)' \widehat{W}_i \left(\sum_{t=1}^{T} z_{it}\widetilde{y}_{it}\right) K_h(\xi-\mu_i) \tag{4.25}$$

revealing that this GMM estimator is a local estimator based on proximity of observations controlled by μ to the type value ξ. Due to (4.13), the type fixed effects have the form, for each $t \in T$,

$$\widehat{\alpha}_t(\xi;\theta,\mu) = \sum_{i=1}^{N}\left(y_{it} - x'_{it}\theta\right)\frac{K_h(\xi-\mu_i)}{\sum_{j=1}^{N} K_h(\xi-\mu_j)}. \tag{4.26}$$

The first-order conditions for μ require interchangeability of differentiation and integration, which is satisfied by the smoothness properties of Assumption 7. With a Gaussian kernel function, the partial derivative with respect to μ_j of the sample GMM criteria $\widehat{Q}$, for any $j \in \mathcal{N}$, is

$$\frac{\partial \widehat{Q}(\theta,\mu)}{\partial \mu_j} = \frac{1}{Nh^2}\int_{\Xi}\left\|\widehat{W}_i^{1/2} g_j(\xi;\theta,\mu)\right\|^2 (\xi-\mu_j)\, K_h(\xi-\mu_j)\, d\xi \tag{4.27}$$

$$-\frac{2}{Nh^2}\sum_{i=1}^{N}\int_{\Xi}\left(\frac{1}{T}\sum_{t=1}^{T} z_{it}\left(\widetilde{y}_{jt} - \widetilde{x}'_{jt}\theta(\xi)\right)\right)' \widehat{W}_i g_i(\xi;\theta,\mu)\,(\xi-\mu_j)\,\frac{K_h(\xi-\mu_j)K_h(\xi-\mu_i)}{\sum_{i=1}^{N} K_h(\xi-\mu_i)}\, d\xi. \tag{4.28}$$

4.2.4 Computation

This section focuses on calculating the TFE-GMM estimator for a general model. For all $i \in \mathcal{N}$, denote

$$\widehat{Q}_i(\xi;\theta,\alpha,\mu) = \left\|\widehat{W}_i^{1/2} g_i(\xi;\theta,\alpha,\mu)\right\|^2 K_h(\xi-\mu_i) \tag{4.29}$$

as the individual's weighted time series GMM criteria. To calculate $\left(\widehat{\theta},\widehat{\alpha},\widehat{\mu}\right)$ we can use a first order method such as gradient descent provided the gradient of the objective function has a closed-form. We can write a smoothed version of the marginal density of

types as follows:

$$\widehat{v}(\xi;\mu) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{\xi-\mu_i}{h}\right). \tag{4.30}$$

Then we can write the gradient in terms of a sample average of expectations with respect to the type:

$$\nabla\widehat{Q}(\theta,\alpha,\mu) = \frac{1}{N}\sum_{i=1}^{N}\int_{\Xi}\frac{\nabla\widehat{Q}_i(\xi;\theta,\alpha,\mu)}{\widehat{v}(\xi;\mu)}\,d\widehat{v}(\xi;\mu). \tag{4.31}$$

We cannot use an ordinary gradient descent algorithm since this gradient is an expectation and the calculation depends on the support and density of an unobserved variable.

I propose a "double-online" gradient descent where we sample a single type from the current iterate density and a batch of observations w_i at each iterate. In the case of the type, I am using the most crude approximation of the expectation and rely on large numbers of iterations to approximate the gradients. For identification and the asymptotic theory, the type space Ξ was assumed to be known, but in practice it is likely to be unknown. I introduce an additive penalty $\sqrt{N^{-1}\sum_{i=1}^{N}\mu_i^2}$ in the gradient descent to penalize types from being excessively spread out, placing a constraint on their sample standard deviation. I also constrain the types to have sample mean 0, making this an example of penalized SGD where the constraints are expectation constraints on the first and second moments; see Xiao [2019].

This is purely a computational device and not studied in the asymptotic theory, where it is assumed that the type space is known. This empirical correction can be justified by noticing that the types only enter the model through the parameters and only their relative locations matter in the calculations. Therefore there is no need to precisely define the support to obtain estimates, but one must ensure that there are bounds so that estimated types do not travel arbitrarily far away causing numerical instability. The following is the basic algorithm.

Algorithm 2 (SGD for TFE-GMM)**.** *Devise a learning rate schedule* η*, penalty parameter* $\lambda > 0$*, convergence threshold* $\kappa > 0$*, and bandwidth* $h > 0$*. Initialize* $\mu^{(0)}$ *randomly such that* $\overline{\mu}^{(s)} = 0$*, and set* $s \leftarrow 0$*.*

1. *Sample: A type value* $\xi^{(s)} \sim \widehat{\nu}(\cdot; \mu^{(s)})$ *and a time series* w_i *randomly from* $\mathcal{N}$*.*

2. *Update parameters: at the sampled value* $\xi^{(s)}$ *and with all of the observations:*

$$\left(\theta^{(s)}, \alpha^{(s)}\right) = \left(\theta^{(s)}(\xi^{(s)}), \alpha^{(s)}(\xi^{(s)})\right) \longleftarrow \underset{(\theta,\alpha)}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \frac{\widehat{Q}_i(\xi^{(s)}; \theta, \alpha, \mu^{(s)})}{\widehat{\nu}(\xi^{(s)}; \mu^{(s)})}$$

3. *Update types:*

$$\mu^{(s+1)} \longleftarrow \mu^{(s)} + \eta^{(s)} \frac{\nabla_\mu \widehat{Q}(\xi^{(s)}; \theta^{(s)}, \alpha^{(s)}, \mu^{(s)})}{\widehat{\nu}(\xi^{(s)}; \mu^{(s)})} + \eta^{(s)} \lambda \frac{\mu^{(s)}}{\sqrt{N^{-1} \sum_{i=1}^{N} (\mu_i^{(s)})^2}}$$

   and update the Ruppert-Polyak averages:

$$\overline{\mu}^{(s+1)} \longleftarrow \left(\frac{s-1}{s}\right) \overline{\mu}^{(s)} + \frac{1}{s} \mu^{(s+1)}.$$

4. *Check: if* $\left\| \mu^{(s)} - \mu^{(s+1)} \right\| < \kappa$*, then stop and report* $\overline{\mu}^{(s+1)}$*. Otherwise, set* $s \leftarrow s + 1$ *and go to step 1.*

Step 2 is simply stated as solving for a weighted GMM estimator. For example, in the case of the endogenous linear model (4.6), it can be simplified by updating via (4.25) and (4.26). The basic algorithm does not guarantee a global minima since the objective function is not convex in μ (Kiwiel [2001]), but several modifications exist to improve performance. Early stopping and patience techniques can be used to prevent overfitting, see Prechelt [1998]. Ruppert-Polyak averages tend to improve the rate of convergence of the algorithm (Ruppert [1988], Polyak and Juditsky [1992]) and statistical properties (Lee et al. [2023]). Instead of using each w_i, a batch (subsample) of them can be used to reduce the noise in the estimates and improve stability by trading off CPU time.

Initialization and the selection of the learning rate schedule are also crucial for good performance. I follow a multi-start technique by running the algorithm many times with different initial values and then choosing the result with the smallest minima (Hu et al. [2009], Martí et al. [2016], Ahuja et al. [2020]). I find that initializing the location parameters near estimates of individual fixed effects of a linear panel model provides a spread of types that seems to lead to good performance of the algorithm. To set the learning rate schedule, I use the Adaptive Moment Estimation (Adam) class of learning rates (Kingma and Ba [2014]) that combines elements of momentum-based rules that remembers previous updates to keep updates tending in the same direction (Rumelhart et al. [1986]) and Adaptive gradient (AdaGrad) (Duchi et al. [2011]) or root mean square propagation (RMSProp) where the learning rate schedule is adapted to each parameter and decreasing over iterates to decrease the influence of older updates. See Spall [2005] for more classical refinements to the standard algorithm.

4.3 Asymptotic Theory: Consistency

Conditions for consistency will be given for the TFE-GMM estimator of the type-specific parameters of the endogenous linear model (4.6) in the case where Ξ is known[6]. I consider the cross-sectional and time dimensions N and T approaching infinity. The bandwidth not only controls the approximation of the type conditional expectations, but also controls the concentration within the integrand directly leveraging the weak convergence property of the kernel function to the Dirac mass. Therefore there must be care in choosing the rate at which h tends to zero.

[6]In the case where Ξ is unknown, consistency with respect to the Fréchet distance might be considered which accounts for all possible reparametrizations of the parameter curves (θ, α). This is essentially accounting for different type spaces as different indexing sets for these curves. Reparametrization is similar to relabeling of discrete groups, except that the reparametrizations must preserve smoothness properties of the curves.

4.3.1 Sketch for consistency

Recall the definition of the estimator

$$\min_{(\theta,\alpha,\mu)\in\Theta\times\mathcal{A}^T\times\Xi^N}\sum_{i=1}^{N}\int_{\Xi}\left\|\widehat{W}_i^{1/2}\left(\sum_{t=1}^{T}z_{it}\left(y_{it}-x_{it}'\theta(\xi)-\alpha_t(\xi)\right)\right)\right\|^2 K_h(\xi-\mu_i)\,d\xi \tag{4.32}$$

and the sufficient statistic for some individual i's type:

$$\xi_i^0 = F(w_i;\theta^0,\alpha^0) = \operatorname*{argmin}_{\xi\in\Xi}\operatorname*{plim}_{T\to\infty}\left\|\frac{1}{T}\sum_{t=1}^{T}g(w_{it};\theta^0(\xi),\alpha_t^0(\xi))\right\|^2 \tag{4.33}$$

where $g(w_{it};\theta^0(\xi),\alpha_t^0(\xi)) = z_{it}\left(y_{it}-x_{it}'\theta^0(\xi)-\alpha_t^0(\xi)\right)$. With the appropriate conditions, as the panel dimensions and bandwidth tend to their limits, the sequence of value functions tend to their population counterpart. In particular, as h gets smaller the kernel density, e.g., Gaussian with variance h^2, become tall and narrow, they begin concentrating on the type estimates μ_i that approximate the sample moment conditions closest to 0. All the while the number of estimated types μ_i becomes large and covers more of the type space so the local constant estimators approach the conditional expectations forming the type fixed effects (4.13) and similarly for θ^0 (4.10).

For sufficiently small h and large N,T, the integrands then concentrate around the sufficient statistic (4.33) in the type space, making $\widehat{\mu}$ defined as the minimizer a good approximation of the true types, thus providing a good sample for local estimation of parameters. See Figure 4.1 for an illustration of consistency of the type estimators.

4.3.2 Consistency of type-specific parameters

Define a norm for a vector-valued function $\psi : I \to \mathbb{R}^K$ for $K \geq 1$ and $I \subset \mathbb{R}$ compact as

$$\|\psi\|_2 = \left(\int_I \|\psi(u)\|^2\,du\right)^{1/2} \tag{4.34}$$

where the inner norm is the Standard Euclidean norm for vectors.

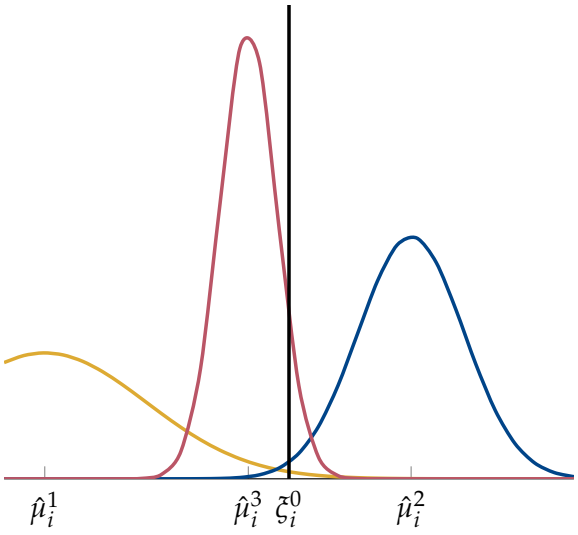


Figure 4.1: The type estimates $\widehat{\mu}_i^1, \widehat{\mu}_i^2, \widehat{\mu}_i^3$ for ξ_i^0 are representative of a growing sample NT and decreasing bandwidth h. As the sample size grows, the kernels concentrate more on their means, i.e., the type estimates, while also approaching the true type ξ_i^0 since the integrals in the objective (4.32) will concentrate on the population moment conditions.

Assumption 9. *There exists $M > 0$ such that*

a. *There exists a collection of non random positive definite matrices $\{W_i\}_{i \in \mathcal{N}} \subset \mathbb{R}^{\ell \times \ell}$ such that $\max_{i \in \mathcal{N}} \left\| \widehat{W}_i - W_i \right\| \to_p 0$ for some suitable matrix norm $\|\cdot\|$ and denoting the minimum eigenvalue among W_i for all $i \in \mathcal{N}$ as $\widehat{\tau}$, then $\widehat{\tau} \to \tau > 0$ as $N, T \to \infty$.*

b. *The parameter spaces are of compactly supported, bounded functions:*

$$\Theta = \{\theta : \Xi \to \mathbb{R} : \|\theta\|_2 < \infty\} \text{ and } \mathcal{A} = \{\alpha : \Xi \to \mathbb{R} : |\alpha(\xi)| < \infty, \text{ for all } \xi \in \Xi\}.$$

c. *For $(i,t) \in \mathcal{N} \times \mathcal{T}$, $\mathbb{E}\left[\left\|z_{it} x_{it}'\right\|\right] \leq M$, $\mathbb{E}\left[\left\|z_{it}\right\|^2\right] \leq M$, $\mathbb{E}\left[u_{it}\right] = 0$ and $\mathbb{E}\left[u_{it}^4\right] \leq M$.*

d. *For all $i \in \mathcal{N}$, $\left| \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{T} \mathbb{E}\left[z_{it}' z_{is} u_{it} u_{is}\right] \right| \leq M$.*

e. The density v^0 is twice continuously differentiable on its support and the kernel function K must satisfy: $\int_{\mathbb{R}}[K(u)]^2\,du \leq M$ and $\int f(u)K_b(u-x)\,du \to f(x)$ as $b \to 0^+$ where f is any function that is continuous on a compact domain or a Schwartz function.

f. Let $Z_{it} \in \mathbb{R}^K$ denote the nonconstant elements of z_{it}. For any vector of type assignments $\mu = (\mu_1, \ldots, \mu_N) \in \Xi^N$ and for any $\xi \in \Xi$, define $\mathcal{M}(\mu, \xi)$ as the following $(p+T) \times (p+T)$ matrix:

$$\sum_{i=1}^{N} \frac{K_h(\xi - \mu_i)K_b(\xi - \xi_i^0)}{\sum_{j=1}^{N} K_h(\xi - \mu_j)K_b(\xi - \xi_j^0)} \begin{bmatrix} \frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T} x_{it}Z_{it}'Z_{is}x_{is}' & \frac{1}{T^{3/2}}\sum_{t=1}^{T}\sum_{s=1}^{T} x_{it}Z_{it}'Z_{is}\delta_s' \\ \frac{1}{T^{3/2}}\sum_{t=1}^{T}\sum_{s=1}^{T} \delta_t Z_{it}'Z_{is}x_{is}' & \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T} \delta_t Z_{it}'Z_{is}\delta_s' \end{bmatrix},$$

where $\widehat{\rho}(\mu, \xi)$ denotes its minimum eigenvalue. Then, $\operatorname{plim}_{N,T\to\infty} \min_{\mu\in\Xi^N} \widehat{\rho}(\mu, \xi) > 0$, for all $\xi \in \Xi$.

Assumption 4(a) requires that the chosen weight matrices converge in probability to positive definite matrices. Assumption 4(b) requires the parameters be bounded as functions of the types. Assumption 4(c) rules out non stationary processes and perfect collinearity between instruments and covariates. Assumption 4(d) is a weak exogeneity condition, bounding the time series dependence between instruments and errors for every individual in the sample. A simple sufficient condition would be instruments are independent from errors. Assumption 4(e) contains standard assumptions from kernel density estimation requiring a sufficiently smooth true density and small tailed kernel function. Additionally, only kernels that satisfy weak convergence to the Dirac delta function are permissible as it is crucial to concentrate the estimated types around the true types to obtain consistency of parameter estimators. The Gaussian kernel would satisfy these requirements along with many other commonly used kernel functions.

Assumption 4(f) is a relevance condition similar to Assumption 1(g) found in Bonhomme and Manresa [2015] and their supplementary appendix for group heterogeneous parameters. It requires that z_{it} display sufficient within-type variation over time and across individuals to serve as relevant instruments for covariates x_{it}. Some time series

dependence is required, but must be limited. In this sense, Assumptions $4(d, f)$ are analogous to the classic conditions for validity of a set of instruments. Notably types do not enter the former since they are assumed exogenous of the error terms, but types can induce correlation between instruments and covariates.

The following establishes consistency with respect to the norm (4.34).

Theorem 6. *Suppose that Assumption 7 and 9 hold. As $h \to 0$, $N, T \to \infty$ and $Nh \to \infty$:*

$$\left\|\widehat{\theta} - \theta^0\right\|_2^2 \to_p 0 \;\; \textit{and} \;\; \frac{1}{T}\sum_{t=1}^{T}\left\|\widehat{\alpha}_t - \alpha_t^0\right\|_2^2 \to_p 0.$$

4.4 Simulation Evidence

In this section, a demonstration of the finite sample properties of the TFE-GMM estimator are presented across relevant data generating processes (DGP). I assume that data is generated randomly by

$$y_{it} = \theta_1(\xi_i)x_{1it} + \theta_2(\xi_i)x_{2it} + \alpha_t(\xi_i) + u_{it} \tag{4.35}$$

$$x_{kit} = \gamma' z_{it} + k^{-1}\alpha_t(\xi_i) + v_{kit}, \;\; k \in \{1, 2\} \tag{4.36}$$

$$\begin{bmatrix} z_{1it} \\ z_{2it} \end{bmatrix} \sim N\left(\begin{bmatrix} \alpha_t(\xi_i) \\ \alpha_t(\xi_i) \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right), \;\; \gamma = (1, -1) \tag{4.37}$$

$$\begin{bmatrix} u_{it} \\ v_{1it} \\ v_{2it} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix}\right). \tag{4.38}$$

where $N(\cdot, \cdot)$ denotes the normal distribution. I emphasize that the form of this DGP implies covariates x_{it} are correlated with types via $\alpha_t(\xi_i)$ and z_{it} and therefore correlated with coefficients $\theta(\xi_i)$.

The additional endogeneity and absence of heteroskedasticity partially justifies the use of the mean group two stage least squares (MG2SLS) estimator of Pesaran and Smith [1995] by estimating each individual time series model separately. However it is expected that this estimator performs poorly due to the presence of type heterogeneity.

I consider various specifications of type-specific coefficients, such as linear and logarithmic and type fixed effects, including traditional two-way and interactive fixed effects specifications and a dynamic AR(1) form as in the simulation of Mugnier [2022a]. Throughout I sample types from a beta distribution, $\text{Beta}(2,2) - 0.5$, reflecting the need for a compact type space. A sample of $(N,T) = (500,15)$ is taken in each simulation and this is repeated 100 times.

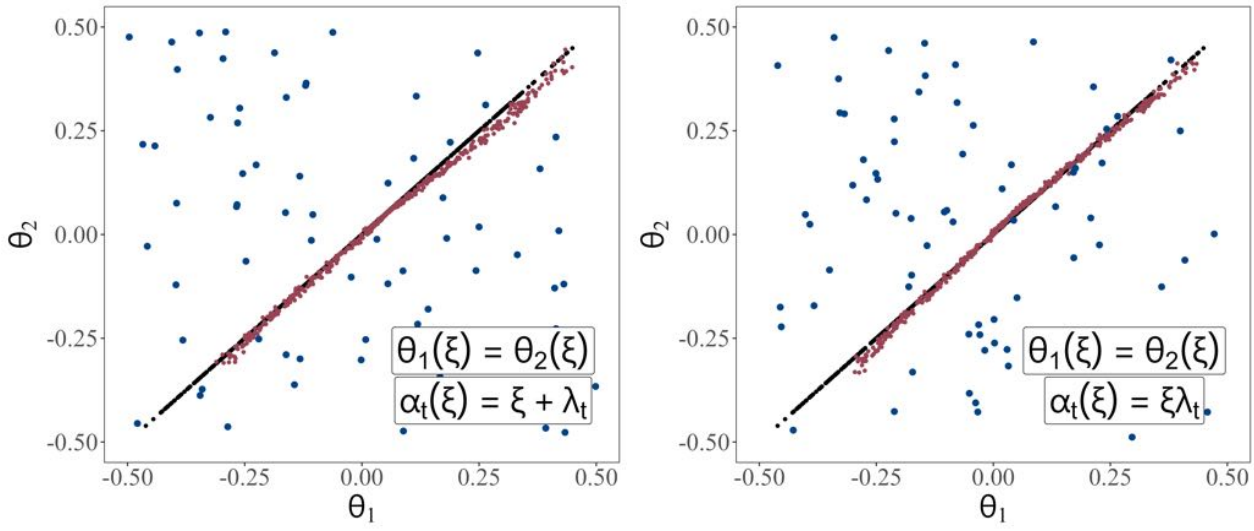
Let $\theta_1(\xi) = \theta_2(\xi) = \xi$ and let λ_t be iid $N(0, 0.25)$. Consider two specifications of TFEs: two-way fixed effects $\alpha_t(\xi_i) = \xi_i + \lambda_t$ and one factor interactive fixed effects $\alpha_t(\xi_i) = \xi_i \times \lambda_t$. Note that estimates result in $N = 500$ pairs of type coefficients and time series of TFEs and so visualization may be challenging. For estimation throughout, I take all of the weight matrices as the identity in the one-step estimation. I also initialize Algorithm 1 at the true values, set penalty to $\lambda = 0.3$ and set the bandwidth to $h = 0.073$, which is obtained from an admittedly ad-hoc use of Silverman's rule-of-thumb (Silverman [1986])[7].

Figure 4.2a shows the geometric properties of the average of the TFE-GMM estimates for the type coefficients. Figure 4.2b shows a time-varying box-and-whisker plot displaying bias properties of the average of TFE estimates. At a glance, the TFE-GMM bias appears to be mild, while the MG2SLS estimates (in blue) struggle to capture the shape of the line. As for the TFEs, since we are using a local constant plug-in we can interpret the wide range of bias as boundary bias, where outlier types have limited information to produce good estimates. This occurs for both coefficients and TFEs, where the coefficients do not adequately reach the boundaries of the line. This suggests that a local linear plug-in for the type conditional expectation functions may be more suitable.

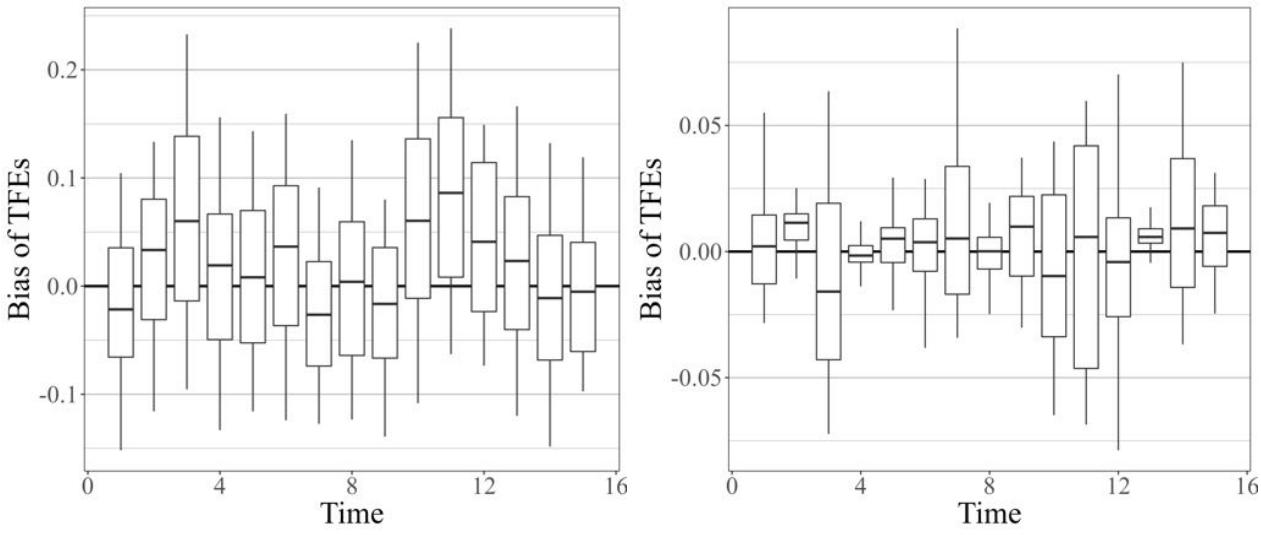
Table 4.1 displays the statistical properties of the TFE-GMM estimator for specific features of the type-heterogeneous coefficients across various specifications. Table 4.2 shows the distribution of the root mean squared error of the average of coefficients.

---

[7] A cross-validation technique may make a more justifiable choice, but this rule-of-thumb seems to work reasonably well likely due to the symmetric and regular type density



(a) Estimates of type coefficients using TFE-GMM (Red) and with MG2SLS (Blue). True values in Black. Left: two-way fixed effects, Right: interactive fixed effects specifications.



(b) Per-time period box-and-whisker for bias of TFE estimates using TFE-GMM. Left: two-way fixed effects, Right: interactive fixed effects specifications.

Figure 4.2: Simulation results specifying linear relationship among type coefficients and traditional fixed effects (two-way and interactive).

Bias	$\mathbb{E}[\theta_k(\xi_i)]$		$\text{Var}(\theta_k(\xi_i))$	
Specification	$\theta_1(\xi_i)$	$\theta_2(\xi_i)$	$\theta_1(\xi_i)$	$\theta_2(\xi_i)$
$\theta_1 = \theta_2$ & FE	0.068	0.053	-0.01	-0.01
$\theta_1 = \theta_2$ & TWFE	0.067	0.056	-0.007	-0.009
$\theta_1 = \theta_2$ & IFE	0.059	0.056	-0.010	-0.009
$\theta_1 = \theta_2$ & AR(1), $\rho = 0.9$	0.066	0.051	-0.009	-0.007
$\log(\theta_1) = \theta_2$ & AR(1), $\rho = 0.75$	0.062	0.065	-0.006	-0.083

Table 4.1: Bias of the mean and variance estimators of the type-heterogeneous coefficients over 100 simulations across different specifications. Bandwidth chosen is $h = 0.073$. Estimator performs just as well with traditional TFE specifications as with AR(1) form. The concavity of the second parameter introduces more downward bias when estimating the variance (bottom row).

For robustness to bandwidth choice I repeated some of the experiments with different bandwidths and reported these additional results in Appendix C.1. Notably, intuition from over and under smoothing a kernel density estimate seems to carry over to TFE-GMM. This indicates that consistency for type estimates relies on choosing a bandwidth that depends on both N, T.

For the next experiment, let $\theta_1(\xi) = \xi$ and $\theta_2(\xi) = \log(\theta_1(\xi))$ so that the curve appears as the graph of the natural logarithm. Take TFEs as an AR(1) process with idiosyncratic shocks: $\alpha_t(\xi_i) = 0.75\,\alpha_{t-1}(\xi_i) + U_{\xi_i t}$ where $U_{\xi_i t} \sim \text{Unif}(-0.1, 0.1)$ is specific to $i = 1, \ldots, 500$. Figure 4.3 shows the geometry of the curve and average of TFE-GMM estimates. The bias is more apparent in segments with more curvature and sparsity of points. This is expected given that conditions for identification rely on curvature, which would make estimation more challenging. There are also moderately small biases even

RMSE	Mean		Median		25th Percentile		75th Percentile	
Specification	$\theta_1(\xi_i)$	$\theta_2(\xi_i)$	$\theta_1(\xi_i)$	$\theta_2(\xi_i)$	$\theta_1(\xi_i)$	$\theta_2(\xi_i)$	$\theta_1(\xi_i)$	$\theta_2(\xi_i)$
$\theta_1 = \theta_2$ & FE	0.110	0.107	0.105	0.010	0.086	0.086	0.138	0.134
$\theta_1 = \theta_2$ & TWFE	0.112	0.107	0.107	0.010	0.091	0.084	0.140	0.136
$\theta_1 = \theta_2$ & IFE	0.105	0.105	0.098	0.096	0.083	0.082	0.134	0.131
$\theta_1 = \theta_2$ & AR(1), $\rho = 0.9$	0.111	0.102	0.103	0.099	0.087	0.085	0.139	0.124
$\log(\theta_1) = \theta_2$ & AR(1), $\rho = 0.75$	0.106	0.171	0.099	0.122	0.086	0.101	0.127	0.225

Table 4.2: Features of the root mean squared error distribution of the type-heterogeneous coefficients TFE-GMM estimates over 100 simulations across different specifications. Bandwidth chosen is $h = 0.073$. Estimator performs just as well with traditional TFE specifications as with AR(1) form. Concavity of the second parameter makes estimation more difficult, resulting in larger ranges of RMSE (bottom row).

at the boundaries and a single example TFE drawn randomly from the sample shows that it follows the true path reasonably well.

4.5 Rational Addiction with Type Fixed Effects

In this section I estimate a RA model with type fixed effects using household cigarette purchase data. I start with a background on the testable implications of the RA model, a description of the data, and providing initial evidence of cyclical consumption. I follow with an explanation of how exogenous variation from type heterogeneity and dynamic prices as instruments address the endogeneity problem. Lastly, I provide the results.

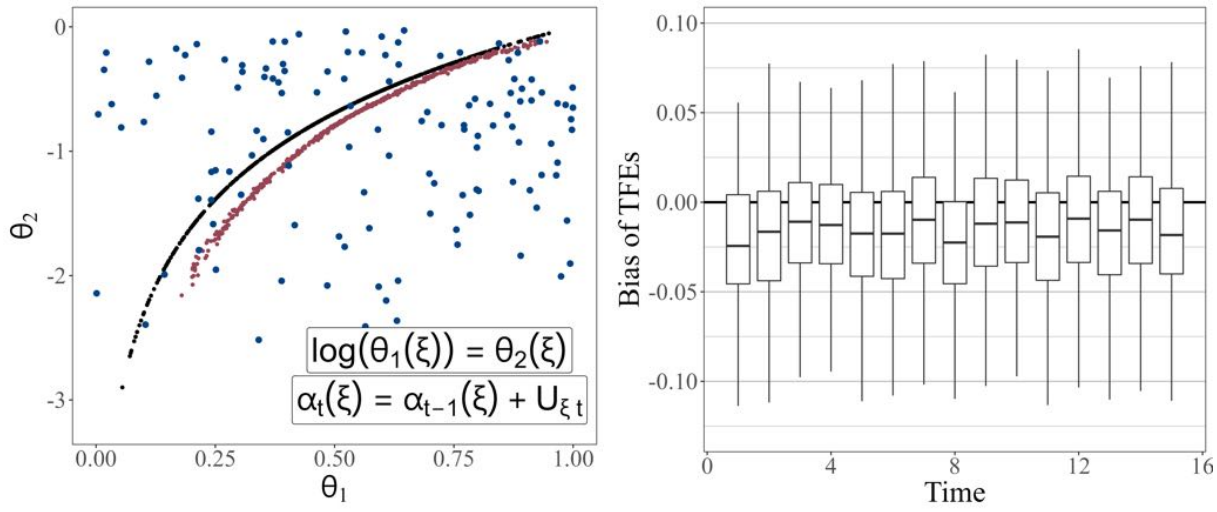


Figure 4.3: Top: Estimates of type coefficients using TFE-GMM (Red). True values in Black. Bottom: Per-time period box-and-whisker for bias of TFE estimates using TFE-GMM.

4.5.1 Background

The RA model is of a representative consumer that maximizes their discounted lifetime utility subject to their budget. Agents devise an optimal consumption plan that takes into account an accumulation of addiction stock that may provide utility or disutility. This stock is assumed to exhibit reinforcement and tolerance, which are modeled respectively as adjacent complementarity and as diminishing marginal utility from consumption. The stock also depreciates according to some positive factor so that periods of abstinence will lead to addiction vanishing. For complete details regarding the derivations of the RA model, see Ferguson [2000].

Assuming a quadratic utility and lifetime budget specification, the consumption plan is linear in past and future consumption C_t along with other explanatory variables and an error term:

$$C_t = \theta_l \, C_{t-1} + \theta_f \, C_{t+1} + \beta' X_t + e_t. \tag{4.39}$$

It is common to test the null that (θ_f, θ_l) are zero against the alternative that they are positive. When they are positive, it is taken that there is addiction (θ_l) and that consumers are forward-looking (θ_f) . In addition, there are a few other testable implications of RA that are not always analyzed, but are important for our purposes. These conditions as given in Becker and Murphy [1988] are

$$\theta_l \, \theta_f < 0.25, \tag{4.40}$$

$$\theta_l + \theta_f < 1. \tag{4.41}$$

The inequality (4.40) guarantees that the roots of the second order difference equation associated to the optimal control problem are real-valued, which rules out cyclical behavior. The inequality (4.41) guarantees that there is a saddle-point equilibrium. Figure 4.4 shows an example saddle-point path devised by the traditional rational addict in view of the harmful (beneficial) addiction they may develop.

In micro panels, there are many individuals at different points in the life cycle and on their idiosyncratic consumption paths. Estimation of (4.39) with ordinary fixed effects regression will therefore yield weighted averages of these many different points and dynamics of consumption. As pointed out by Laporte et al. [2017], if indeed consumers follow a saddle-point equilibrium, the presence of the unstable root will have, for example, different age cohorts of individuals follow stable and unstable branches at different points in time. In segments where both are occurring, the unstable root prevents identification of the parameters of interest due to combinations with a non stationary process. This would also prohibit identification of these parameters if a portion of consumers are following cyclical dynamics.

In this sense, the type augmented RA model (4.1) is better equipped to reconcile individuals on different time paths with varying dynamics. There is no a priori reason why saddle-point dynamics is the only possibility with rational addiction so inclusion

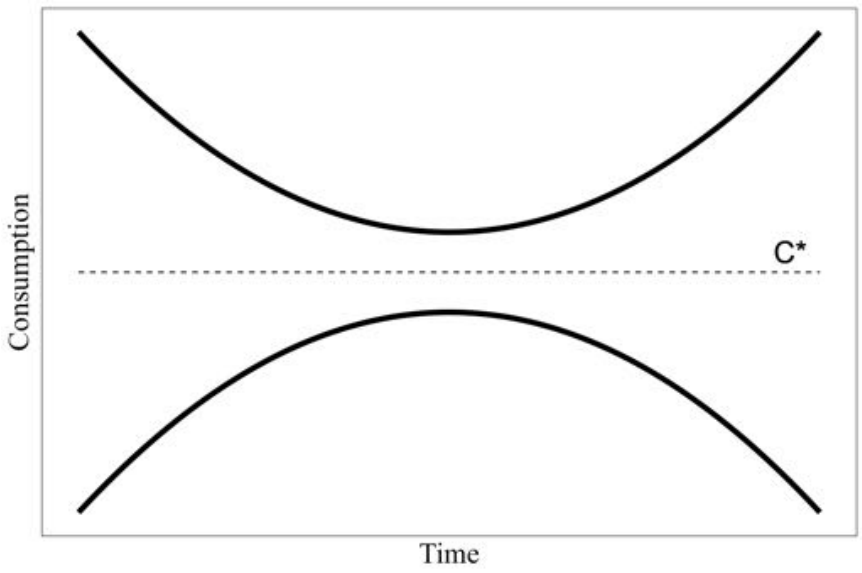


Figure 4.4: Two examples (top and bottom curves) of optimal consumption plan under saddle-point dynamics. At first, the consumer tends to the equilibrium quantity but over time the unstable branch dominates leading the consumer to veer off.

of the type fixed effect may, for example, capture an omitted second addiction stock that may exhibit adjacent complementarity or substitutability as argued by Dockner and Feichtinger [1993]. With type-heterogeneous parameters, we can inspect each consumers parameter estimates and classify as following cyclical (4.40) or "traditional" saddle-point (4.41) behavior.

4.5.2 Data

The data is sourced from the NielsenIQ Consumer Panel[8] that follows household weekly purchases using a survey-provided scanner. I aggregated to the year-level by summing

[8]Analysis calculated (or derived) based in part on data from Nielsen Consumer LLC and marketing databases provided through the NielsenIQ Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the NielsenIQ data are my own and do not reflect the views of NielsenIQ. NielsenIQ is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

the quantity of packs purchased and taking the average price paid per pack in the year. I ignore substitution to other goods such as e-cigarettes. I maintain a balanced panel with $N = 3,296$ households and $T = 16$ years between 2004 and 2019. The data also contains demographic information such as the household size, head of household employment, education, marital status, race, sex and state of occupancy. Income is recorded in interval ranges and values reported are from 2 years prior to the panel year.

Figure 4.3 presents aggregate summary statistics. At the start of the sample period nearly 80% of individuals are over 45 years old and another non inclusive 80% have some college education, either partially or fully completed, or with post graduate study. A majority of the sample (83%) are recorded as White while 8% are Black and 4% Asian. Only 5% of the sample identify as Hispanic. Figure 4.5 shows the location density of respondents in the USA and displays somewhat representative population density according to state size in the contiguous states.

Variable	Mean	Median	St. Error	Min	Max
Packs Purchased	20.27	14	24.78	0.20	603
Price	5.70	5.28	2.28	1.35	18.69
Income	50−59.9	60−69.9	−	< 5	> 200
Age	50−54	55−64	−	< 25	> 65

Table 4.3: Descriptive statistics of the sample. Price and income in 2012 USDs. Income is in thousands and income and age are given in ranges.

Figure 4.6 displays the median amount of packs purchased along with the median average unit price paid over the sample period. It appears that demand is sensitive to price changes over time. This can also be said of each age cohort as seen in Figure 4.7.

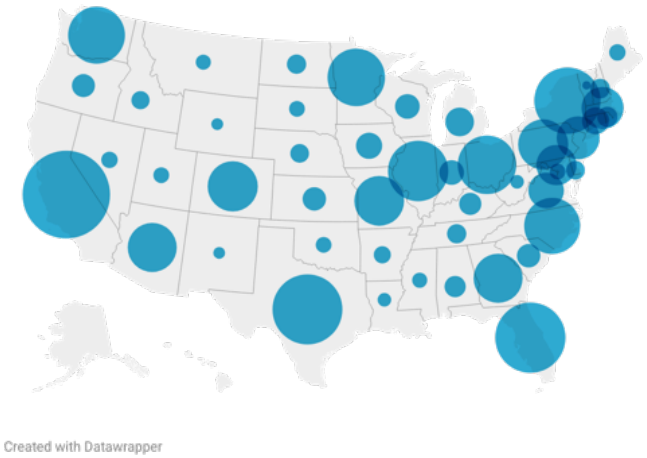


Figure 4.5: Size of circles indicate density of respondents in each state.

However, Figure 4.8 shows some individual's consumption profiles that appear much different than the aggregate. Some display cyclical patterns across years and follow these dynamics at different time periods. The fixed effects two-stage least squares estimates are $(0.247, 0.369)$ for the lag and lead consumption covariates, respectively, and are significant and positive meaning that we would conclude cigarette smokers are rationally addicted and follow saddle-point dynamics since they satisfy (4.40) and (4.41). However, judging by the consumption profiles in Figure 4.8, there remains important unexplained variation that could align with cyclical behavior.

4.5.3 Identification

Identification is similar to fixed effects regression with instruments. The model (4.1) contains lagged and forwarded outcomes and type heterogeneity that does not address all endogeneity concerns. First, I assume that inclusion of the type fixed effect eliminates components of the error term that would be serially correlated. The error term in this case can be interpreted as life-cycle shocks such as loss of job or other personal period-

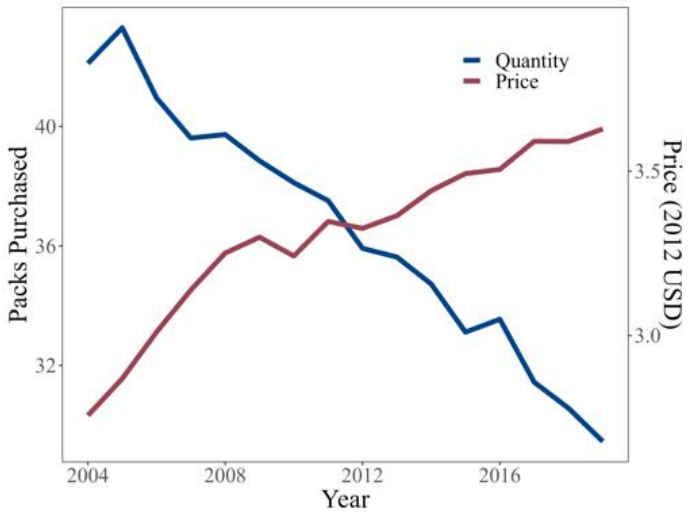


Figure 4.6: Aggregate median packs purchased and average unit price paid.

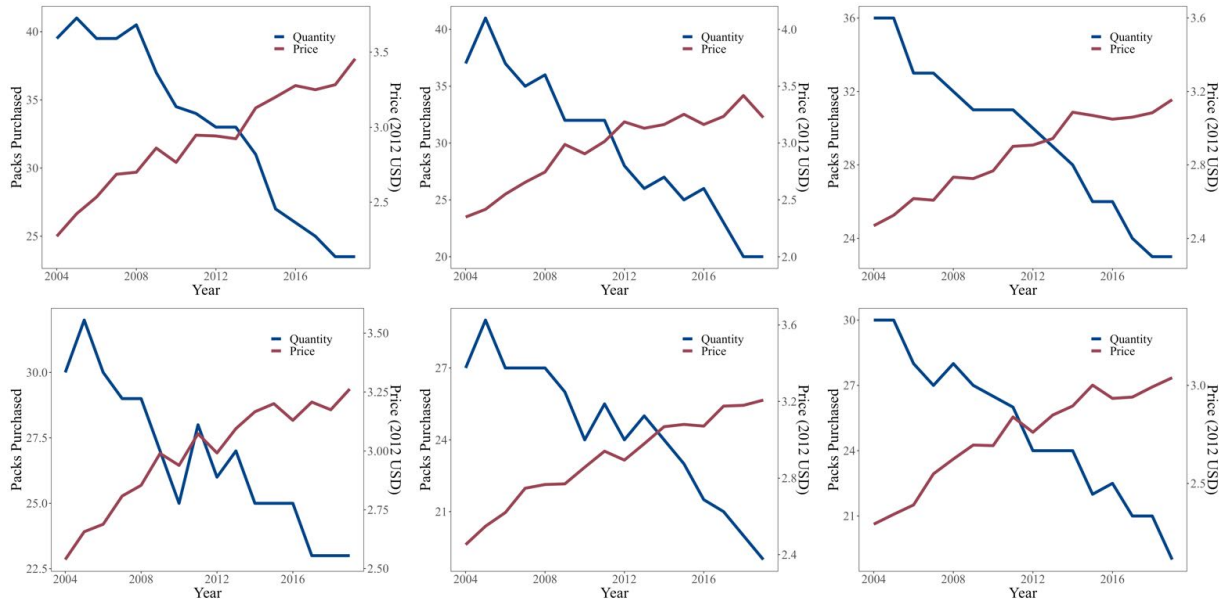


Figure 4.7: By age cohort: median packs purchased and average unit price paid. Left vertical axis follows purchases and right vertical axis follows price in 2012 USDs.

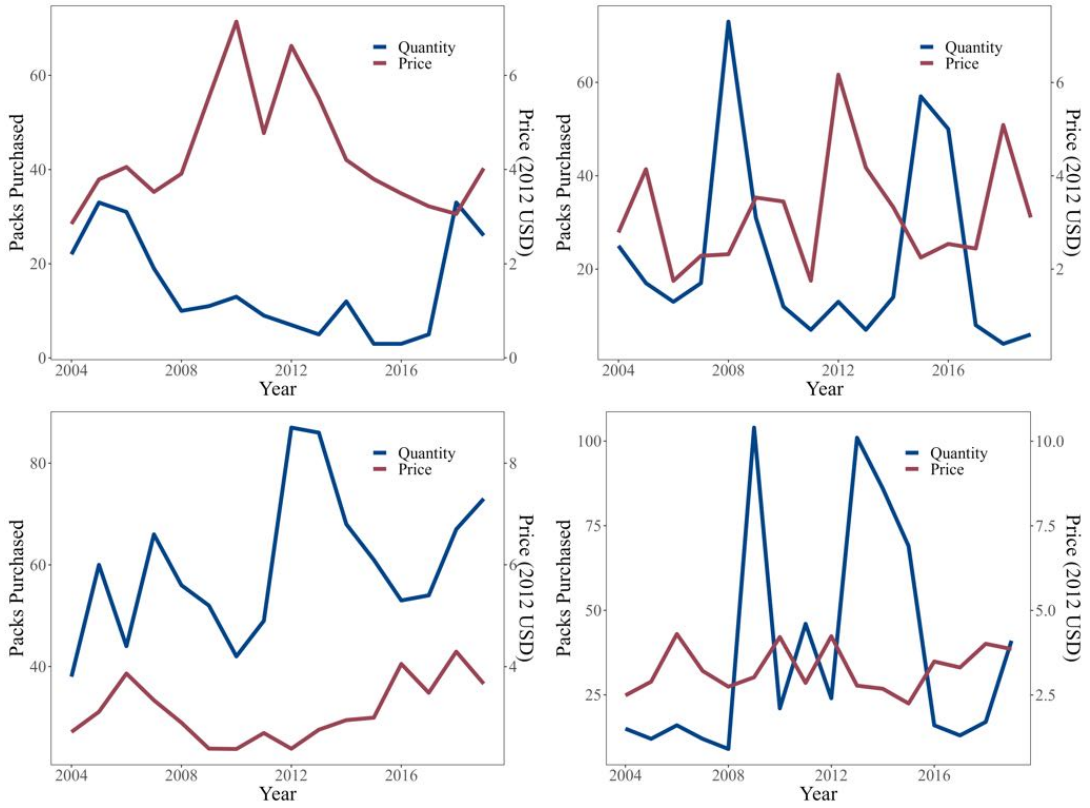


Figure 4.8: For specific individuals: packs purchased and average unit price paid. Left vertical axis follows purchases and right vertical axis follows price in 2012 USDs. Individuals follow complicated consumption paths. Cycles evident in bottom row. Top row appears similar to saddle-point dynamics.

specific shocks. The type fixed effects contains the components of addiction that are not directly caused by habit captured by lag consumption, which may include a stock of health: physical or mental. Regardless of what it may be, the habit stock may be arbitrarily correlated with this stock, meaning the coefficient on past consumption will be biased if unaccounted for.

Type heterogeneity is an important feature of this model to allow for different addiction profiles. For example, both stocks could be positively correlated with consumption

in which case the consumer could be classified as fully addicted, borrowing the naming convention from Dockner and Feichtinger [1993]. Another example, while the habit stock is positively associated to consumption, the other could be a health stock that is negatively correlated with consumption indicating a partial addiction. These properties depend upon the realization of the type of the consumer in the sample period, putting them on these different trajectories and if type heterogeneity is ignored then estimates of the coefficient corresponding to lag consumption could be understated in the case of full addiction and overstated in the case of partial addiction. Cigarettes are highly addictive and harmful goods due to nicotine dependence (Benowitz [2010]) so it is expected that omission of type heterogeneity produces estimates that are generally downward biased.

The inclusion of the lead of consumption results in endogeneity since the consumer realizes their life-cycle shock for the period and uses that information to infer the amount of cigarettes to consume in the next period. The TFE does not solve this since the consumer also realizes at the beginning of the period their health stock level, how much they smoked last period, prices and income level. Then, when considering how much to smoke in the current period, they can infer the consequences of smoking today on their health stock in the next period followed by how their addiction would affect future consumption. To correct for this I follow tradition and use the lead and lag of prices as instruments for lead and lagged consumption. Justification for using future prices as instruments come from the fact that price increases on cigarettes are announced in advance due to government tax changes and that past prices on their own may be a poor instrument (Nelson and Startz [1990]). Figure 4.9 summarizes this discussion in a diagram.

4.5.4 Results

Consumption is measured per household member and all relevant values are in 2012 USDs. Figure 4.10 shows the estimated type-specific parameters (θ_f, θ_l) in blue, the

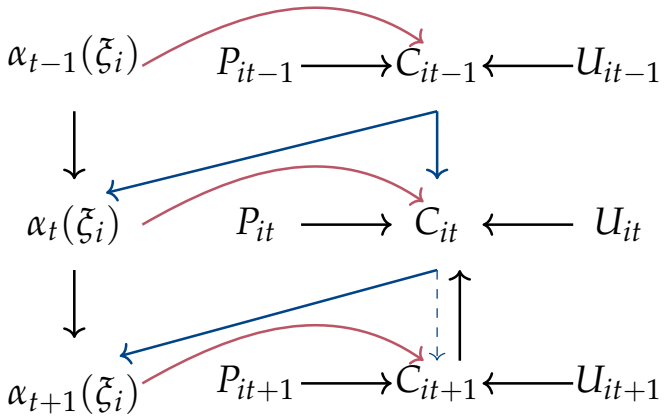


Figure 4.9: Diagram illustrating identification of heterogeneous effects. Blue arrows indicate direct effects from habit and red arrows indicate effects from the health stock. Dashed line indicates a endogeneity concern despite controlling for type heterogeneity.

region corresponding to violation of (4.40) in purple, and region corresponding to traditional evidence for the RA model in green. In the sample, 73% of individuals lie in the purple area, which corresponds to cyclical behavior while only 24% are what typically is taken as rational addicts. All processes in the sample are on stable paths, see C.2.

Figure 4.11 shows the absolute price and income effects on consumption of cigarettes. In the figure, 95% of the sample is shown since there are significant outliers that would prevent visualization of the concentration of effects at the origin. Most of the sample would show less than a quarter of a pack drop in demand in response to a dollar increase in price indicating a practically insignificant effect. This implies policy that targets consumption through a sin tax would effectively raise revenue, but may do little at the individual level to curb consumption among these entrenched users.

The binging behavior of individuals in this sample provide support for educational intervention programs and their insensitivity to price changes could raise revenue for these programs. Examples of cost-effective policies could be further emphasizing that health care providers inform patients that smoking cessation will increase life expectancy

or treat an illness (Yu et al. [2004]) or focus on cognitive or interpersonal therapy (Stenberg et al. [2018]).

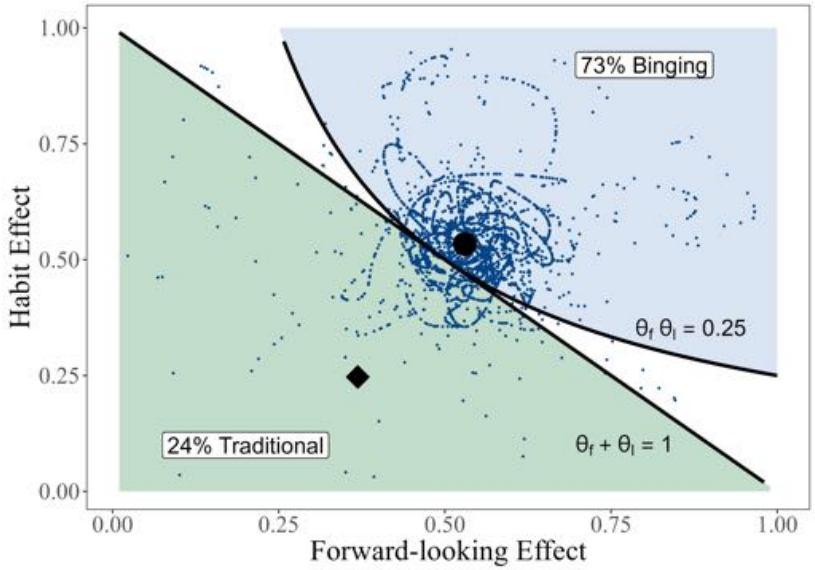


Figure 4.10: Blue dots: Estimates of the $N = 3,296$ type heterogeneous lag and lead coefficients using bandwidth $h = 0.128$. The blue region is associated to cyclical behavior, while the green is for saddle-point dynamics. Black diamond indicates two-way fixed effects estimate. Black dot indicates the average of $(\widehat{\theta}_f, \widehat{\theta}_l)$.

4.6 Concluding Remarks

The TFE RA model is a powerful tool to analyze consumption of addictive goods by controlling for dynamic-idiosyncratic differences across individuals. It was revealed that most individuals in the sample exhibit cyclical behavior that aligns with the pharmacology of nicotine. Furthermore, cigarette use is driven mostly by habitual and forward-looking behavior rather than prices, income, etc.

Extension to nonlinear models is a future direction for the type fixed effects GMM framework. With this it would be possible to study demand for nicotine products and

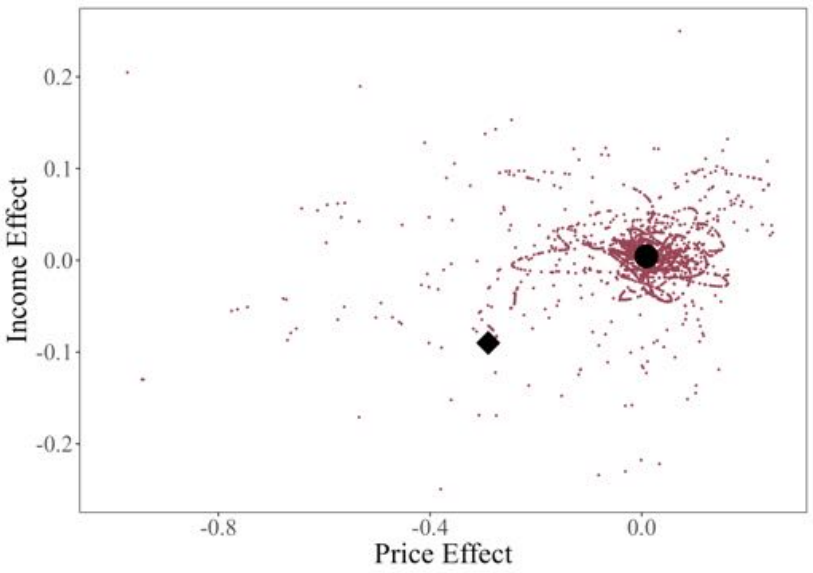


Figure 4.11: Estimates of the $N = 3,296$ type heterogeneous price and income coefficients using bandwidth $h = 0.128$. Only 95% of the sample is shown. Values appear to be practically concentrated near the origin indicating weak sensitivity to changes in price and earnings. Black diamond indicates two way fixed effects estimate. Black dot indicates the median of the TFE-GMM estimates.

the decision to quit or relapse under type heterogeneity. Another direction might be to replace local estimation techniques by global estimation through the method of sieves. I leave these topics for future work.

BIBLIOGRAPHY

R. Aaberge and A. Brandolini. Multidimensional poverty and inequality. Discussion Papers, No. 792, Statistics Norway, Research Department, Oslo, 2014.

J. M. Abowd and H. S. Farber. Job queues and the union status of workers. *ILR Review*, 35(3):354–367, 1982.

D. Acemoglu, S. Johnson, J. A. Robinson, and P. Yared. Income and democracy. *The American Economic Review*, 98(3):808–842, 2008.

M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

J. Aguilar Loyo and T. Boot. Grouped heterogeneity in linear panel data models with heterogeneous error variances. *Journal of Business & Economic Statistics*, pages 1–25, 2024.

J. S. Ahlquist and E. Wibbels. Riding the wave: World trade and factor-based models of democratization. *American Journal of Political Science*, 56(2):447–464, 2012.

M. Ahmed, R. Seraj, and S. M. S. Islam. The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 9(8):1295, 2020.

K. Ahuja, A. Dhurandhar, and K. R. Varshney. Learning to initialize gradient descent using gradient descent, 2020.

D. Aloise, A. Deshpande, P. Hansen, and P. Popat. NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning*, 75:245–248, 2008.

D. Aloise, P. Hansen, and L. Liberti. An improved column generation algorithm for minimum sum-of-squares clustering. *Mathematical Programming*, 131:195–220, 2012.

P. C. Álvarez Esteban, E. del Barrio, J. Cuesta-Albertos, and C. Matrón. A fixed-point approach to barycenters in wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744 – 762, 2016.

T. Ando and J. Bai. Panel data models with grouped factor structure under unknown group membership. *Journal of Applied Econometrics*, 31(1):163–191, 2016.

F. Andreoli and C. Zoli. From unidimensional to multidimensional inequality: a review. *Metron*, 78:5–42, 2020.

J. D. Angrist. Treatment effect heterogeneity in theory and practice. *The Economic Journal*, 114(494):C52–C83, 2004.

J. D. Angrist, K. Graddy, and G. W. Imbens. The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. *The Review of Economic Studies*, 67(3):499–527, 2000.

M. Arellano. Practitioners corner: Computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics and Statistics*, 49(4):431–434, 1987.

M. Arellano and S. Bonhomme. Identifying distributional characteristics in random coefficients panel data models. *The Review of Economic Studies*, 79(3):987–1020, 2012.

M. Arellano and J. Hahn. *Understanding Bias in Nonlinear Panel Models: Some Recent Developments*, volume 3 of *Econometric Society Monographs*. Cambridge University Press, 2007a.

M. Arellano and J. Hahn. Understanding bias in nonlinear panel models: Some recent developments. *Econometric Society Monographs*, 43:381, 2007b.

B. Arnold. The Lorenz curve: Evergreen after 110 years. In *Advances on Income Inequality and Concentration Measures*. Routledge, 2008.

B. C. Arnold. *Pareto Distributions*. International Co-operative Publishing House, 1983.

B. C. Arnold. *Majorization and the Lorenz order: A brief introduction*. Springer Science & Business Media, 2012.

B. C. Arnold and J. M. Sarabia. *Majorization and the Lorenz order with Applications in Applied Mathematics and Economics*. Springer, 2018.

A. Atkinson and F. Bourguignon. The comparison of multi-dimensioned distributions of economic status. *Review of Economic Studies*, 49:183–201, 1982.

M. Auld and P. Grootendorst. An empirical analysis of milk addiction. *Journal of Health Economics*, 23(6):1117–1133, 2004.

F. Aurenhammer, F. Hoffmann, and B. Aronov. Minkowski-type theorems and least-squares clustering. *Algorithmica*, 20:61–76, 1998.

J. Bai. Panel data models with interactive fixed effects. *Econometrica*, 77(4):1229–1279, 2009.

J. Bai and S. Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

B. Baltagi and J. M. Griffin. The econometrics of rational addiction: The case of cigarettes. *Journal of Business & Economic Statistics*, 19(4):449–54, 2001.

A. Banerjee. Multidimensional Gini index. *Mathematical Social Sciences*, 60:87–93, 2010.

A. Banerjee. Multidimensional Lorenz dominance: a definition and an example. *Keio Economic Studies*, 52:65–80, 2016.

G. Becker and K. Murphy. A theory of rational addiction. *Journal of Political Economy*, 96 (4):675–700, 1988.

G. Becker, M. Grossman, and K. Murphy. An empirical analysis of cigarette addiction. *American Economic Review*, 84(3):396–418, 1994.

N. L. Benowitz. Nicotine addiction. *New England Journal of Medicine*, 362(24):2295–2303, 2010.

C. A. Bester and C. B. Hansen. Grouped effects estimators in fixed effects models. *Journal of Econometrics*, 190(1):197–208, 2016.

P. Bickel and E. Lehmann. Descriptive statistics for nonparametric models, iii. dispersion. *Annals of Statistics*, 4:1139–1158, 1976.

S. Bonhomme and E. Manresa. Grouped patterns of heterogeneity in panel data. *Econometrica*, 83(3):1147–1184, 2015.

S. Bonhomme, T. Lamadon, and E. Manresa. Discretizing unobserved heterogeneity. *Econometrica*, 90(2):625–643, 2022.

J. F. Bonnans and A. Shapiro. Perturbation analysis of optimization problems. In *Springer Series in Operations Research*, 2000.

L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010: 19th International Conference on Computational Statistics, Paris France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, pages 177–186. Springer, 2010.

Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 4:375–417, 1991.

J. Bricker, L. J. Dettling, A. Henriques, J. W. Hsu, L. Jacobs, K. B. Moore, S. Pack, J. Sabelhaus, J. Thompson, and R. A. Windle. Changes in US family finances from 2013 to 2016: Evidence from the survey of consumer finances. *Fed. Res. Bull.*, 103:1, 2017.

W. Brock and R. Thomson. Convex solutions of implicit relations. *Mathematics Magazine*, 39:208–111, 1966.

M. J. Brusco. A repetitive branch-and-bound procedure for minimum within-cluster sums of squares partitioning. *Psychometrika*, 71:347–363, 2006.

M. J. Brusco and D. Steinley. A comparison of heuristic procedures for minimum within-cluster sums of squares partitioning. *Psychometrika*, 72(4):583–600, 2007.

D. Card. The effect of unions on the structure of wages: A longitudinal analysis. *Econometrica: Journal of the Econometric Society*, pages 957–979, 1996.

G. Carlier, A. Galichon, and F. Santambrogio. From Knothe's transport to Brenier's map and a continuation method for optimal transport. *SIAM Journal on Mathematical Analysis*, 41:2554–2576, 2010.

G. Carlier, V. Chernozhukov, and A. Galichon. Vector quantile regression: An optimal transport approach. *Annals of Statistics*, 44:1165–1192, 2016.

K. M. Carroll. The profound heterogeneity of substance use disorders: Implications for treatment development. *Current Directions in Psychological Science*, 30(4):358–364, 2021.

J. Cawley and C. J. Ruhm. Chapter three - the economics of risky health behaviors. In M. V. Pauly, T. G. Mcguire, and P. P. Barros, editors, *Handbook of Health Economics*, volume 2 of *Handbook of Health Economics*, pages 95–199. Elsevier, 2011.

F. Chaloupka. Rational addictive behavior and cigarette smoking. *Journal of Political Economy*, 99(4):722–42, 1991.

G. Chamberlain. Analysis of covariance with qualitative data. *The Review of Economic Studies*, 47(1):225–238, 1980.

L. Chang-Ching and N. Serena. Estimation of Panel Data Models with Parameter Heterogeneity when Group Membership is Unknown. *Journal of Econometric Methods*, 1(1): 1–14, August 2012.

A. Charpentier, A. Galichon, and M. Henry. Local utility and multivariate risk aversion. *Mathematics of Operations Research*, 41:266–276, 2016.

X. Cheng, F. Schorfheide, and P. Shao. Clustering for multi-dimensional heterogeneity. 2019.

V. Chernozhukov, A. Galichon, M. Hallin, and M. Henry. Monge-Kantorovich depth, quantiles, ranks and signs. *Annals of Statistics*, 45:223–256, 2017.

D. Chetverikov and E. Manresa. Spectral and post-spectral estimators for grouped panel data models. *arXiv preprint arXiv:2212.13324*, 2022.

P.-A. Chiappori, R. McCann, and B. Pass. Multi to one-dimensional transport. *Communications on Pure and Applied Mathematics*, 70:2405–2444, 2017.

J. Clausen. Branch and bound algorithms-principles and examples. 2003.

V. Dardanoni. Measuring social mobility. *Journal of Economic Theory*, 61:372–394, 1993.

S. Dasgupta. The hardness of k-means clustering. 2008.

R. Davidson and J.-Y. Duclos. Statistical inference for stochastic dominance and for the measurement of poverty and inequality. *Econometrica*, 68:1435–1464, 2000.

R. B. Davies. Algorithm as 155: The distribution of a linear combination of χ^2 random variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 29(3):323–333, 1980.

N. Deb and B. Sen. Multivariate rank-based distribution-free nonparametric testing using measure transportation. *Journal of the American Statistical Association*, 118:192–207, 2023.

P. Deb and P. K. Trivedi. Finite mixture for panels with fixed effects. *Journal of Econometric Methods*, 2(1):35–51, 2013.

K. Decancq and M. Lugo. Inequality of wellbeing: A multidimensional approach. *Economica*, 79:721–746, 2012.

E. J. Dockner and G. Feichtinger. Cyclical consumption patterns and rational addiction. *The American Economic Review*, 83(1):256–263, 1993.

D. Donaldson and J. Weymark. Single parameter generalizations of Gini indices of inequality. *Journal of Economic Theory*, 22:67–86, 1980.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7), 2011.

R. Dudley. *Real Analysis and Probability*. Cambridge University Press, 2002.

H. Edelsbrunner and N. R. Shah. Incremental topological flipping works for regular triangulations. *Algorithmica*, 15:223–241, 1996.

I. Ekeland. Notes on Optimal Transportation. *Economic Theory*, 42(2):437–459, 2010.

I. Ekeland, A. Galichon, and M. Henry. Comonotone measures of multivariate risks. *Mathematical Finance*, 22:109–132, 2012.

J. Fan and J.-T. Zhang. Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 62(2):303–322, 2000.

Y. Fan and H. Park. Minimum sliced distance estimation in structural models. 2022.

Y. Fan, M. Henry, B. Pass, and J. A. Rivero. Lorenz map, inequality ordering and curves based on multidimensional rearrangements. *arXiv preprint arXiv:2203.09000*, 2024.

O. Faugeras and Rüschendorf. Markov morphisms: a combined copula and mass transportation approach to multivariate quantiles. *Mathematica Applicanda*, 45:3–45, 2017.

M. Faure and N. Gravel. Reducing inequalities among unequals. *International Economic Review*, 62:357–404, 2021.

B. S. Ferguson. Interpreting the rational addiction model. *Health Economics*, 9(7):587–598, 2000.

I. Fernández-Val and J. Lee. Panel data models with nonadditive unobserved heterogeneity: Estimation and inference. *Quantitative Economics*, 4(3):453–481, 2013.

A. Figalli. On the continuity of center-outward distribution and quantile functions. *Nonlinear Analysis*, 177:413–429, 2018.

F. Fisher. Income distribution, value judgments and welfare. *Quarterly Journal of Economics*, 70:380–424, 1956.

J. D. Fisher, D. S. Johnson, T. M. Smeeding, and J. P. Thompson. Inequality in 3-d: Income, consumption, and wealth. *Review of Income and Wealth*, 68(1):16–42, 2022.

R. A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh and London, 1st edition, 1925. Reprinted in *Statistical Methods, Experimental Design, and Scientific Inference*, Oxford University Press.

E. W. Forgy. Cluster analysis of multivariate data : efficiency versus interpretability of classifications. *Biometrics*, 21:768–769, 1965.

R. B. Freeman. Longitudinal analyses of the effects of trade unions. *Journal of labor Economics*, 2(1):1–26, 1984.

S. Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models*. Springer Series in Statistics. Springer New York, 2006.

T. Gajdos and J. Weymark. Multidimensional generalized Gini indices. *Economic Theory*, 26:471–496, 2005.

A. Galichon. *Optimal Transport Methods in Economics*. Princeton University Press, 2018.

A. Galichon and M. Henry. Dual theory of choice under multivariate risks. *Journal of Economic Theory*, 147:1501–1516, 2012.

J. L. Gastwirth. A general definition of the Lorenz curve. *Econometrica*, 39:1037–1039, 1971.

P. Ghosal and B. Sen. Multivariate ranks and quantiles using optimal transport: Consistency, rates, and nonparametric testing. *Annals of Statistics*, 2022.

K. S. Gleditsch and M. D. Ward. Diffusion and the international context of democratization. *International Organization*, 60(4):911–933, 2006.

C. Gottschlich and D. Schuhmacher. The shortlist method for fast computation of the earth mover's distance and finding optimal solutions to transportation problems. *PLoS ONE*, 9, 2014.

B. S. Graham and J. L. Powell. Identification and estimation of average partial effects in 'irregular' correlated random coefficient panel data models. *Econometrica*, 80(5): 2105–2152, 2012.

J. E. Grant, M. N. Potenza, A. Weinstein, and D. A. Gorelick. Introduction to behavioral addictions. *The American Journal of Drug and Alcohol Abuse*, 36(5):233–241, 2010.

N. Gravel and P. Moyes. Ethically robust comparisons of bi-dimensionnal distributions with an ordinal attribute. *Journal of Economic Theory*, 147:1384–1426, 2012.

M. Grossman, F. Chaloupka, and I. Sirtalan. An empirical analysis of alcohol addiction: Results from the monitoring the future panels. *Economic Inquiry*, 36(1):39–48, 1998.

O. Grothe, F. Kächele, and F. Schmid. A multivariate extension of the Lorenz curve based on copulas and a related multivariate gini coefficient. *Journal of Economic Inequality*, 20: 727–748, 2022.

J. Hahn and H. R. Moon. Panel data models with finite number of multiple equilibria. *Econometric Theory*, 26(3):863–881, 2010.

M. Hallin. Measure transportation and statistical decision theory. *Annual Review of Statistics and Its Application*, 9:410–424, 2022.

M. Hallin and G. Mordant. Center-outward multiple-output Lorenz curves and Gini indices a measure transportation approach. Unpublished manuscript, 2022.

S. D. Hanna, K. T. Kim, and S. Lindamood. Behind the numbers: Understanding the survey of consumer finances. *Journal of Financial Counseling and Planning*, 29:410–418, 2018.

L. P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054, 1982.

P. Hansen, N. Mladenović, and J. A. Moreno Pérez. Variable neighbourhood search: methods and applications. *Annals of Operations Research*, 175(1):367–407, 2010.

G. Hardy, J. Littlewood, and G. Pólya. *Inequalities*. Cambridge University Press, 1934.

T. Hastie and R. Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(4):757–796, 1993.

J. Heckman and B. Singer. A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica*, 52(2):271–320, 1984.

J. Heckman and E. Vytlacil. Instrumental variables methods for the correlated random coefficient model: Estimating the average rate of return to schooling when the return is correlated with schooling. *Journal of Human Resources*, pages 974–987, 1998.

D. R. Hoover, J. A. Rice, C. O. Wu, and L.-P. Yang. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85(4):809–822, 1998.

C. Hsiao. *Analysis of Panel Data*. Cambridge University Press, 4th edition, 2022.

X. Hu, R. Shonkwiler, and M. C. Spruill. Random restarts in global optimization. 2009.

H. Joe. *Multivariate Models and Multivariate Dependence Concepts*. Chapman and Hall, 1997.

L. V. Kantorovich. Mathematical methods of organizing and planning production. *Management Science*, 6:366–422, 1960.

L. V. Kantorovich. On the translocation of masses. *Journal of Mathematical Sciences*, 133 (4):1381–1382, 2006a.

L. V. Kantorovich. On a problem of Monge. *Journal of Mathematical Sciences*, 133(4): 1383–1383, 2006b.

H. Kasahara and K. Shimotsu. Nonparametric identification of finite mixture models of dynamic discrete choices. *Econometrica*, 77(1):135–175, 2009.

A. Kennickell. Multiple imputation in the survey of consumer finances. *Statistical Journal of the IAOS*, 33:143–151, 2017a.

A. B. Kennickell. The bitter end? the close of the 2007 SCF field period. *Statistical Journal of the IAOS*, 33:93–99, 2017b.

A. B. Kennickell and R. L. Woodburn. Consistent weight design for the 1989, 1992 and 1995 SCFs, and the distribution of wealth. *Review of Income and Wealth*, 45:193–215, 1999.

J. Kim and L. Wang. Hidden group patterns in democracy developments: Bayesian inference for grouped heterogeneity. *Journal of Applied Econometrics*, 34(6):1016–1028, 2019.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

K. C. Kiwiel. Convergence and efficiency of subgradient methods for quasiconvex minimization. *Mathematical Programming*, 90:1–25, 2001.

H. Knothe. Contributions to the theory of convex bodies. *Michigan Mathematical Journal*, 4:39–52, 1957.

S.-C. Kolm. Multidimensional egalitarianisms. *The Quarterly Journal of Economics*, 91: 1–13, 1977.

S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE signal processing magazine*, 34(4):43–59, 2017.

G. Koshevoy. Multivariate Lorenz majorization. *Social Choice and Welfare*, 12:93–102, 1995.

G. Koshevoy and K. Mosler. The Lorenz zonoid of a multivariate distribution. *Journal of the American Statistical Association*, 91:873–882, 1996.

G. Koshevoy and K. Mosler. Multivariate Gini indices. *Journal of Multivariate Analysis*, 60:252–276, 1997.

G. Koshevoy and K. Mosler. Price majorization and the inverse Lorenz function. Discussion Papers in Statistics and Econometrics 3/99, University of Cologne, 1999.

G. Koshevoy and K. Mosler. Multivariate Lorenz dominance based on zonoids. *AStA*, 91:57–76, 2007.

R. Krishnapuram and J. Kim. A note on the Gustafson-Kessel and adaptive fuzzy clustering algorithms. *IEEE Transactions on Fuzzy systems*, 7(4):453–461, 1999.

L. Laage. A correlated random coefficient panel model with time-varying endogeneity. *arXiv preprint arXiv:2003.09367*, 2020.

A. Laporte, A. R. Dass, and B. S. Ferguson. Is the Rational Addiction model inherently impossible to estimate? *Journal of Health Economics*, 54(C):161–175, 2017.

S. Lee, Y. Liao, M. H. Seo, and Y. Shin. Fast inference for quantile regression with tens of millions of observations, 2023.

E. L. Lehmann. Some concepts of dependence. *The Annals of Mathematical Statistics*, 37: 1137–1153, 1966.

B. Lévy. A numerical algorithm for L2 semi-discrete optimal transport in 3D. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49:1693–1715, 2015.

Q. Li and J. S. Racine. *Nonparametric Econometrics: Theory and Practice*. Princeton University Press, 2007.

X. L. Liangjun Su. Determining the number of groups in latent panel structures with an application to income and democracy. *Quantitative Economics*, 8:729–760, 11 2017.

R. J. Little and D. B. Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28 (2):129–137, 1982.

M. Lorenz. Methods of measuring the concentration of wealth. *Publication of the American Statistical Association*, 9:209–219, 1905.

E. Maasoumi. The measurement and decomposition of multi-dimensional inequality. *Econometrica*, 54:991–997, 1986.

E. Maasoumi and J. Racine. A solution to aggregation and an application to multidimensional 'well-being' frontiers. *Journal of Econometrics*, 191:374–383, 2016.

J. MacQueen. Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pages 281–297, 1967.

A. W. Marshall, I. Olkin, and B. C. Arnold. *Inequalities: Theory of Majorization and Its Applications*. Springer, 2nd edition, 2011.

R. Martí, J. Lozano, A. Mendiburu, and L. Hernando. Multi-start methods. 2016.

R. J. McCann. Existence and uniqueness of monotone measure-preserving maps. *Duke Mathematical Journal*, 80(2):309–323, 1995.

S. L. McElroy, A. I. Guerdjikova, N. Mori, M. R. Munoz, and P. E. Keck. Overview of the treatment of binge eating disorder. *CNS Spectrums*, 20(6):546–556, 2015.

G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley Series in Probability and Statistics. Wiley, 2004.

A. Mehrabani. Estimation and identification of latent group structures in panel data. *Journal of Econometrics*, 2022.

H. O. Melberg. Rational addiction theory: a survey of opinions. HERO Online Working Paper Series 2008:7, University of Oslo, Health Economics Research Programme, 2009.

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

F. Merlevède, M. Peligrad, and E. Rio. A Bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields*, 151(3): 435–474, 2011.

G. Monge. Memoire sur la theorie des deblais et des remblais. *Histoire de l'Academie Royale des Sciences de Paris*, pages 666–704, 1781.

M. Mugnier. Unobserved clusters of time-varying heterogeneity in nonlinear panel data models. *Job Market Paper*, 2022a.

M. Mugnier. A simple and computationally trivial estimator for grouped fixed effects models. *Working Paper*, 2022b.

A. Müller and D. Stoyan. *Comparison Methods for Stochastic Models and Risks*. Wiley, 2002.

C. Nelson and R. Startz. The distribution of the instrumental variables estimator and its t-ratio when the instrument is a poor one. *The Journal of Business*, 63(1), 1990.

W. Newey and D. McFadden. Large sample estimation and hypothesis testing. In R. F. Engle and D. McFadden, editors, *Handbook of Econometrics*, volume 4, chapter 36, pages 2111–2245. Elsevier, 1 edition, 1986.

W. K. Newey and R. J. Smith. Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255, 2004.

J. Neyman and E. L. Scott. Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society*, pages 1–32, 1948.

S. Nickell. Biases in dynamic models with fixed effects. *Econometrica: Journal of the Econometric Society*, pages 1417–1426, 1981.

N. Olekalns and P. Bardsley. Rational addiction to caffeine: An analysis of coffee consumption. *Journal of Political Economy*, 104(5):1100–1104, 1996.

J. Pacheco and O. Valencia. Design of hybrids for the minimum sum-of-squares clustering problem. *Computational Statistics & Data Analysis*, 43(2):235–248, 2003.

M. Pesaran and R. Smith. Estimating long-run relationships from dynamic heterogeneous panels. *Journal of Econometrics*, 68(1):79–113, 1995.

G. Peyré and M. Cuturi. *Computational Optimal Transport*. arXiv:1803.00567, 2018.

D. Pollard. Strong consistency of k-means clustering. *Ann. Statist.*, 9(1):135–140, 01 1981.

D. Pollard. A central limit theorem for k-means clustering. *Ann. Probab.*, 10(4):919–926, 11 1982.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

L. Prechelt. Automatic early stopping using cross validation: quantifying the criteria. *Neural Networks*, 11(4):761–767, 1998.

G. Puccetti and M. Scarsini. Multivariate comonotonicity. *Journal of Multivariate Analysis*, 101:291–304, 2010.

J. Quiggin. Increasing risk: another definition. In A. Chikan, editor, *Progress in Decision, Utility and Risk Theory*, pages 239–248. Kluwer: Dordrecht, 1992.

S. Rachev and L. Rüschendorf. A characterization of random variables with minimal L^2 distance. *Journal of Multivariate Analysis*, 32:48–54, 1990.

R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239, 1984.

E. Rio. *Asymptotic Theory of Weakly Dependent Random Processes*, volume 80. Springer, 2017.

J. A. Rivero. Unobserved grouped heteroskedasticity and fixed effects, 2023.

H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

P. M. Robinson. Root-N-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 931–954, 1988.

R. T. Rockafellar. Characterization of the subdifferentials of convex functions. *Pacific Journal of Mathematics*, 17:497–510, 1966.

R. T. Rockafellar. *Convex Analysis*, volume 18. Princeton University Press, 1970.

M. Rosenblatt. Remarks on a multivariate transformation. *Annals of Mathematical Statistics*, 23:470–472, 1952.

M. Rosenblatt. A central limit theorem and a strong mixing condition. *Proceedings of the national Academy of Sciences*, 42(1):43–47, 1956.

D. B. Rubin. Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91:473–489, 1996.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

D. Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

F. Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Springer, 2015.

J.-M. Sarabia and V. Jorda. Bivariate Lorenz curves: a review of recent proposals. In *XXII Jornadas de ASEPUMA*, 2014.

J.-M. Sarabia and V. Jorda. Lorenz surfaces based on the Sarmanov–Lee distribution with applications to multidimensional inequality in well-being. *Mathematics*, 8:1–17, 2020.

G. R. Saunders, X. Wang, F. Chen, S.-K. Jang, M. Liu, C. Wang, S. Gao, Y. Jiang, C. Khunsriraksakul, J. M. Otto, et al. Genetic diversity fuels gene discovery for tobacco and alcohol use. *Nature*, 612(7941):720–724, 2022.

E. Savaglio. *Three approaches to the analysis of multidimensional inequality*, chapter 10. Routledge, 2006.

M. Shaked and J. G. Shanthikumar. *Stochastic Orders*. Springer, 2007.

H. Shi, M. Drton, and F. Han. Distribution-free consistent independence tests via center-outward ranks and signs. *Journal of the American Statistical Association*, 117:395–410, 2022.

A. F. Shorrocks. Ranking income distributions. *Economica*, 197:3–17, 1983.

B. W. Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

J. C. Spall. *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. John Wiley & Sons, 2005.

U. Stenberg, A. Vågan, M. Flink, V. Lynggaard, K. Fredriksen, K. F. Westermann, and F. Gallefoss. Health economic evaluations of patient education interventions a scoping review of the literature. *Patient Education and Counseling*, 101(6):1006–1035, 2018.

J. E. Stiglitz, A. Sen, J.-P. Fitoussi, et al. Report by the commission on the measurement of economic performance and social progress, 2009.

J. H. Stock and M. W. Watson. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179, 2002.

L. Su and G. Ju. Identifying latent grouped patterns in panel data models with interactive fixed effects. *Journal of Econometrics*, 206(2):554–573, 2018.

L. Su, Z. Shi, and P. C. B. Phillips. Identifying latent structures in panel data. *Econometrica*, 84(6):2215–2264, 2016.

Y. Sun. Estimation and inference in panel structure models. *Econometrics eJournal*, 2005.

T. Taguchi. On the two-dimensional concentration surface and extensions of concentration coefficient and Pareto distribution to the two-dimensional case-i. *Annals of the Institute of Statistical Mathematics*, 24:355–382, 1972a.

T. Taguchi. On the two-dimensional concentration surface and extensions of concentration coefficient and Pareto distribution to the two-dimensional case-ii. *Annals of the Institute of Statistical Mathematics*, 24:599–619, 1972b.

C. Villani. *Topics in Optimal Tranportation*. American Mathematical Society, 2003.

C. Villani. *Optimal Transport: Old and New*, volume 338. Springer, 2009.

J. Weymark. Generalized Gini inequality indices. *Mathematical Social Sciences*, 1:409–430, 1981.

E. N. Wolff. Household wealth trends in the United States, 1962 to 2019: Median wealth rebounds... but not enough. Working paper, National Bureau of Economic Research, 2021.

J. M. Wooldridge. *Econometric analysis of cross section and panel data*. MIT Press, 2010.

X. Xiao. Penalized stochastic gradient methods for stochastic convex optimization with expectation constraints. *Optimization-online*, 2019.

M. E. Yaari. The dual theory of choice under risk. *Econometrica*, 55:95–115, 1987.

H. Yang and E. G. Tabak. Conditional density estimation, latent variable discovery, and optimal transport. *Communications on Pure and Applied Mathematics*, 75(3):610–663, 2022.

S. Yitzhaki. On using linear regressions in welfare economics. *Journal of Business & Economic Statistics*, 14(4):478–486, 1996.

C.-M. Yu, C.-P. Lau, J. Chau, S. McGhee, S.-L. Kong, B. M.-Y. Cheung, and L. S.-W. Li. A short course of cardiac rehabilitation program is highly cost effective in improving long-term quality of life in patients with recent myocardial infarction or percutaneous coronary intervention. *Archives of Physical Medicine and Rehabilitation*, 85(12):1915–1922, 2004.

B. Zhang. Forecasting with bayesian grouped random effects in panel data. *arXiv preprint arXiv:2007.02435*, 2020.

X. Zhao, Y. Li, and Q. Zhao. Mahalanobis distance based on fuzzy clustering algorithm for image segmentation. *Digital Signal Processing*, 43:8–16, 2015.

Appendix A

LORENZ MAP, INEQUALITY ORDERING AND CURVES BASED ON MULTIDIMENSIONAL REARRANGEMENTS

A.1 User's implementation guide

In this section, we will point to specific computational routines the reader may use to accomplish each step in Section 2.2.2. All of the figures in this paper were generated via implementations in the `R` language, however we emphasize that they are standard and likely to be found in other languages and packages. Algorithm 1 is therefore intended to guide the reader across the key steps of the implementation. We specialize to the case of $d = 2$ and leave $d > 2$ for the online supplement.

For the vector quantile, the `transport` package for `R` provides various implementations to solve (2.8) that are found in the function `semidiscrete`. Both the standard descent approach of Aurenhammer et al. [1998] and, our preferred, multiscale initialization and L-BGFS approach of Mérigot [2011] are supported methods. Alternatively and for all of our calculations, we used the `Rgeogram` package that is a wrapper of the C++ Geogram library implementation of Mérigot [2011]. Both packages provide the optimal weight vector h required for the next steps.

The Lorenz map requires solving for the convex cells defined in (2.10). The optimal h from the vector quantile calculation can be used as an input for the `power_diagram` function in the `transport` package. It will provide as output the vertices of the convex cells. With a desired rank $r = (r_1, r_2)$, the next step is to find the area of the intersection of the convex cells with the rectangle $[0, r]$. The `sf` package provides these tools: first, the function `st_polygon` transforms the vertices of the cells and rectangle into a "polygon" object that can be then read into `st_intersection` and `st_area`, which calculate the intersection

and the area, respectively. These two functions can be used to calculate $\lambda(W_i^h \cap [0, r])$ in (2.11) for all $i = 1, \ldots, n$.

Calculation of the ILF is a standard application of the empirical distribution function. However, to facilitate drawing α-Lorenz curves, it is recommended to form a uniform grid of values and calculate the ecdf at each value, e.g., $\{0.01, \ldots, 0.99, 1\}^2$. Each pair should correspond to a row and column in a matrix of ecdf values. Then, pass this matrix as input into any function that plots contours of three-dimensional surfaces such as the base `R` function `contour` or `geom_contour` as part of the `ggplot2` package. Finally, with a sample of Lorenz map values, one can compute the Gini index (2.22) by plug-in.

A.2 Additional Results: 3-D Inequality Analysis

For $d = 2$, visualization is possible and smoothness of the quantities is critical for the most precise visual assessments. However, the calculation of the Lorenz map becomes more resource intensive as d grows larger. For instance, in $d = 3$, the power diagrams are formed by convex polytopes, in which the calculation of their volume is more demanding. To avoid this and still retain the quantitative properties of Lorenz mappings, one can consider a discretization of the vector quantile of Definition 1 that maps a uniform distribution over a discrete subset of $[0, 1]^d$ to the empirical distribution[1].

Consider arbitrary $d \geq 1$. In the discrete-discrete setting with a weighted sample, a transport map may not exist, but a unique optimal coupling π among the joint mass functions between the uniform vector and empirical distribution will exist. Generate a random sample $\{U_k\}_{k=1}^m \sim U[0, 1]^d$ with $m \geq n$ and calculate the optimal coupling π between $\{u_k\}_{k=1}^m$ and the empirical distribution. Then, the following alternative estimators

[1]Alternatively, the semidiscrete formulation lends itself to entropic regularization, which has shown to vastly improve computational speed and implementation, see Chapter 4 of Peyré and Cuturi [2018].

leverage equations (2.4) and (2.23) as expectations:

$$\widehat{\mathcal{L}}(r_1,\ldots,r_d) = \sum_{i=1}^{n}\sum_{k=1}^{m} X_i \mathbb{1}\left\{u_{k1} \le r_1,\ldots,u_{kd} \le r_d\right\}\, \pi(X_i,u_k) \tag{A.1}$$

$$\widehat{G} = 1 - \frac{2^d}{d}\sum_{i=1}^{n}\sum_{k=1}^{m} \phi(u_k)' X_i\, \pi(X_i,u_k) \tag{A.2}$$

where $\phi(u_k) = \prod_{j=1}^{d}(1-u_{kj})(1,\ldots,1) \in \mathbb{R}^d$. Calculating the optimal coupling π amounts to a linear program with equality constraints, which there are many implementations and algorithms available; see Gottschlich and Schuhmacher [2014] for the shortlist method employed herein.

The discretized version advantageously allows for quick and convenient calculations of resource shares and Gini indices of attribute vectors of larger dimension. As an example in analyzing inequality in income, wealth, and consumption jointly, we borrow the supplemented version of the SCF between 1989-2016 from Fisher et al. [2022]. Figure A.1 shows the 3-D Gini along with the marginal Gini's of income, wealth, and consumption. It is less skewed towards the wealth attribute indicating a potential dampening from the less unequal consumption attribute. Overall, inequality has been on the rise mirroring the rise in inequality in marginal attributes.

Table A.1 shows all possible Gini indices that can be formed with income, wealth, and consumption data. Comparing our multivariate Gini with the average of marginal Gini's, we see that the attributes are not mutually independent.

A.3 Vector quantiles as solutions to the semi-discrete transport problem

Let $\nu = \sum_{i=1}^{N} w_i \delta_i$ denote the empirical mass function and let λ denote the distribution of U. The vector quantile Q_X is closely related to power diagrams: a useful tool in computational geometry that partitions the ambient space ($[0,1]^2$ in our case) into a finite number of convex cells, one for each point in the weighted sample (Aurenhammer et al. [1998]). Let $h = (h_1,\ldots,h_n) \in \mathbb{R}^n$ be a vector of weights. A set $\{W_i^h\}_{i=1^n}$ constitutes

Algorithm 1 Vector Quantiles, Lorenz Maps, and ILFs

Input
  (X, w) Weighted sample with normalized points X_i and weights w_i
procedure VECTOR-QUANTILE(X, w)
  Set convergence tolerance δ, step size η.
  Set $s \leftarrow 0$, initialize weight vector h^0. ▷ e.g. multiscale approach
  repeat ▷ Begin gradient descent
    $s \leftarrow s + 1$
    $h^{(s)} \leftarrow h^{(s-1)} - \eta[w_i - \lambda(W_i^{h^{(s-1)}})]$ ▷ Modifiable, e.g., L-BFGS
  until $\left\| h^{(s)} - h^{(s-1)} \right\| < \delta$
  $h := h^{(s)}$ solution to (2.8) with (X_i, w_i)
  $\{W_i^h\}_{i=1}^n \leftarrow$ (2.9) ▷ e.g., computing power diagram using h
  return $\{W_i^h\}$ defined by their vertices
procedure LORENZ-MAP($r, X, \{W_i^h\}_{i=1}^n$)
  Input
    r Vector of ranks of interest from $\mathbb{R}^2$
    W_i^h Vertices of cells that define the vector quantile of (X, w)
  Output
    $\mathcal{L}_X$ Lorenz map evaluated at r, a vector
  $\mathcal{L}_X(r) \leftarrow 0$
  for $i = 1$ to n **do**
    Find vertices of $A_i := W_i^h \cap [0, r_1] \times [0, r_2]$ ▷ A_i is convex
    $\lambda_i := \lambda(A_i)$: ordinary area of A_i ▷ Many equate to 0 or w_i
    $\mathcal{L}_X(r) \leftarrow \mathcal{L}_X(r) + X_i \lambda_i$
  return $\mathcal{L}_X(r)$
procedure INVERSE-LORENZ-FUNCTION($X, \{W_i^h\}_{i=1}^n, m$)
  Input
    W_i^h Vertices of cells that define the vector quantile of (X, w)
    m Size of pseudo sample from $U([0,1]^2)$
  Output
    l_X Matrix of cumulative probabilities: rows and columns are coordinates
  Generate $Z :=$ evenly-spaced lattice in $[0,1]^2$ ▷ e.g., $Z = \{0.01, \ldots, 0.99, 1\}^2$
  for $j = 1$ to m **do**
    Draw a single $R_j \sim U[0,1]^2$
    $\mathcal{L}_j :=$ LORENZ-MAP($R_j, X, \{W_i^h\}_{i=1}^n$)
  for $z_{ij} \in Z$ **do**
    $(l_X)_{ij} := m^{-1} \sum_{j=1}^m \mathbb{1}\{\mathcal{L}_j \leq z_{ij}\}$ ▷ empirical cdf of the $\mathcal{L}_j$
  return l_X ▷ Matrix of function values is usual input for contour plots

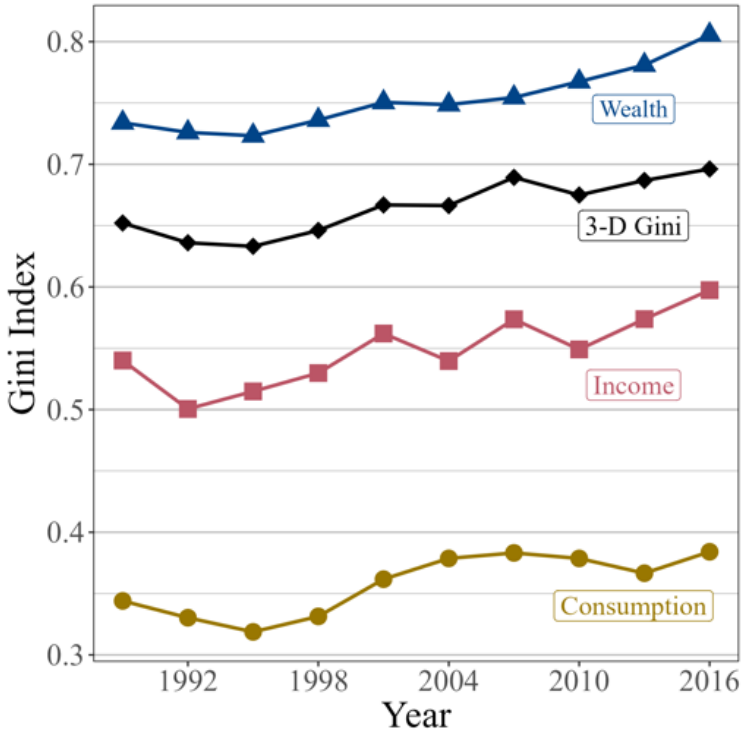


Figure A.1: Marginal Gini indices for income, wealth, and consumption along with the 3 dimensional Gini index, between 1989 and 2016.

a power diagram of $\{(X_i, w_i)\}_{i=1^n}$ if and only if, for any i,

$$W_i^h \quad = \quad \{u \in [0,1]^2 : u'x^i - h^i \geq u'x^j - h^j, 1 \leq j \leq n\}, \tag{A.3}$$

and the weights h satisfy $\lambda(W_i^h) = w_i$ for all $i = 1, \ldots, n$. Therefore, constructing the power diagram revolves around finding these weights.

Theorem 2 from Mérigot [2011] states that these weights are the unique solution to the dual formulation of the optimal transport problem between λ and ν. In other words, h solves the unconstrained convex optimization program (2.8). Consequently, with optimal h, the vector quantile Q_X associates points in any cell W_i^h to X_i as the solution to the semi-discrete optimal transport problem.

Let $L(h)$ denote the objective function of (2.8). The gradient of L at h is available.

	$d=3$		$d=2$			$d=1$		
Year	3-D Gini	Average	I-W	I-C	W-C	I	W	C
1989	0.652	0.539	0.708	0.518	0.608	0.540	0.734	0.344
1992	0.636	0.519	0.687	0.499	0.588	0.501	0.726	0.330
1995	0.633	0.519	0.689	0.493	0.585	0.515	0.723	0.319
1998	0.646	0.532	0.705	0.496	0.597	0.530	0.736	0.331
2001	0.669	0.558	0.723	0.537	0.625	0.562	0.750	0.362
2004	0.666	0.556	0.713	0.538	0.624	0.540	0.749	0.379
2007	0.689	0.570	0.729	0.558	0.632	0.574	0.754	0.383
2010	0.675	0.565	0.727	0.534	0.635	0.549	0.767	0.379
2013	0.687	0.574	0.743	0.543	0.630	0.574	0.781	0.367
2016	0.696	0.596	0.762	0.562	0.649	0.597	0.806	0.384

Table A.1: Various kinds of Gini indices depending on dimension $d = 1, 2, 3$ over the years $1989 - 2016$ for I-income, W-wealth, C-consumption. Our multivariate (3-D) Gini is compared against a simple average of the three marginal Gini's. If the resources were mutually independent, then these would coincide.

First, notice that

$$\int_{[0,1]^d} \max_{k=1,\dots,n} \{u'x^k - h^k\} du = \sum_{i=1}^{n} \int_{W_i^h} \left(u'X_i - h_i\right) \, d\lambda(u) \tag{A.4}$$

by the definition of W_i^h. Therefore, $L(h)$ can be rewritten as

$$\sum_{i=1}^{n} \left[w_i h_i + \int_{W_i^h} \left(u'X_i - h_i\right) \, d\lambda(u) \right]. \tag{A.5}$$

Then, the partial derivatives with respect to h_i for any $i = 1, \ldots, n$ are simply

$$\frac{\partial L(h)}{\partial h_i} = w_i - \int_{W_i^h} d\lambda(u) \tag{A.6}$$

which suggests that the optimal choice of h must satisfy the pushforward condition $\lambda(W_i^h) = w_i$ for all $i = 1, \ldots, n$ in order for the partial derivatives to vanish. Note that calculating this gradient requires calculating the ordinary area $\lambda(W_i^h)$ for a convex polygon W_i^h, for which there are fast algorithms available.

With $\nabla_h L$, a standard gradient descent with a user specified initialization of the weight vector would suffice. The Hessian is not available so a quasi-Newton approach such as the popular L-BGFS algorithm should be used, see Lévy [2015] for a description. Initialization is usually taken to be the origin, however this choice impacts performance. The following section establishes the computational approach towards vector quantiles we have taken in our paper.

A.3.1 Multiscale approach for minimizing L

Mérigot [2011] points out that initialization is a key challenge and, from our initial experiences using the SCF data, a standard first-order gradient descent approach with random initialization often fails to recover weights that satisfy the pushforward constraints, resulting in empty cells. In their paper, they propose a "multiscale" approach that considers a sequence of increasingly complicated descents based on initializations from the preceding simpler problem. These are defined by "decompositions" of the empirical measure.

A decomposition of the empirical measure ν is a sequence of discrete probability measures $\{\nu_\ell\}_{\ell=0}^L$ for which $L > 0$ is chosen, $\nu_0 = \nu$, and ν_ℓ is supported on a set $\mathcal{N}_\ell \subseteq \{1, \ldots, N\}$ such that

$$\nu_\ell = \sum_{i \in \mathcal{N}_\ell} w_{i,\ell} \delta_i \tag{A.7}$$

and $\mathcal{N}_\ell \supset \mathcal{N}_{\ell+1}$.

The goal of this decomposition is to optimally select an initial weight vector by solving the transport problem between λ and ν_l using the optimal weight vector from transporting λ and ν_{l+1} as the choice for initialization. As ℓ decreases to zero, the optimal weight vector for the original problem is constructed from this sequence. In fact, Theorem 3 of Mérigot [2011] shows that taking ν_ℓ and $\nu_{\ell+1}$ close in the Wasserstein sense for all ℓ converges to the original problem.

Call $\ell = 1, \ldots, L$ levels and for any arbitrary level ℓ we are given a transport map π_ℓ between ν_ℓ and $\nu_{\ell+1}$ such that, for any $i \in \mathcal{N}_{\ell+1}$:

$$w_{i,\ell+1} = \sum_{k \in \pi_\ell^{-1}(i)} w_{k,\ell}. \tag{A.8}$$

Solving for the transport map π_ℓ are shown to be equivalent to solving a weighted k-means problem and associates any point x_i for $i \in \mathcal{N}_\ell$ to its nearest neighbor x_j for $j \in \mathcal{N}_{\ell+1}$. In practice, the size $n(\ell)$ of $\mathcal{N}_\ell$ is chosen to be $n(\ell) = n/k^\ell$ with $k = 5$. With this, at each level ℓ, $n(\ell+1)$ points are drawn randomly from ν_ℓ and weighted k-means is performed and ν_ℓ is gathered. With ν_ℓ, we extract the optimal weight vector h_ℓ from minimizing L_ℓ based on (2.8) modified to be the transport problem between λ and ν_ℓ, which uses initialization $h_{\ell+1}$. This vector is used in the initialization of the transport problem between λ and $\nu_{\ell-1}$. See Algorithm 2 for this multiscale approach and note that despite improvements in convergence, the user may still be required to repeat the algorithm to obtain weights that satisfy the pushforward constraint (2.9).

A.4 Specific features and issues with the data source

We review some known issues with the data set that impact our analysis. See Hanna et al. [2018] for a more in-depth account.

Algorithm 2 Calculating Vector Quantiles (Power Diagram)

Input

  X sample of normalized resources (Income, wealth, etc.)

  w vector of sample weights

Output

  $\{W_i^h\}_{i=1}^n$ Power diagram cells via weights h

procedure MULTISCALE MINIMIZATION OF $L := L_0$ (Mérigot [2011])

  Output

    $h = h_0$ Solution to (2.8)

  Set convergence tolerance ε

  Initialize $h_L \leftarrow 0$

  for $\ell = L - 1$ to 0 **do**

    $h^i_{\ell,0} \leftarrow h^{\pi_\ell(i)}_{\ell+1}$ for all $i \in \mathcal{N}_\ell$

    $k \leftarrow 0$

    repeat

      $h_{\ell,k+1} \leftarrow \operatorname{argmin} L_\ell(h)$ (e.g., L-BFGS initialized with $h_{\ell,k}$)

      $v_{k+1} \leftarrow \nabla_h L_\ell(h_{\ell,k+1})$

      $k \leftarrow k + 1$

    until $\|v_k\| < \varepsilon$

    $h_\ell \leftarrow h_{\ell,k}$

Construct W_i^h via (A.3) (e.g., Edelsbrunner and Shah [1996])

Sampling strategy

The over sampling of high income and wealthy households is achieved by applying two distinct sampling techniques. The first sample is the core representative sample selected by a standard multi-stage area-probability design. The second is the high income supplement from statistical records derived from tax data by the Statistics of Income (SOI) division of the U.S. Internal Revenue Service. The stages sample disproportionately–usually one-third of the final sample is from the high income supplement. Sampling in this way retains characteristic information of the population while also addressing the known selection biases of the wealthy not responding to surveys. In order to represent the population with this sample, weights must be constructed for each unit of observation. For more details on the construction of weights and their implications on the distribution of wealth, see Kennickell and Woodburn [1999].

Unit of observation and timing of interviews

The observations in this data set are not households, but rather a subset called the *primary economic unit* (PEU) that may be individuals or couples and their financial dependents. For example in the 2016 data set 13% of PEUs were in a household that contained one or more members not in their PEU. Additionally, the respondent is not necessarily the head of the household, so special care must be taken if analyzing attitudes in relation to some demographic characteristics such as age. The interviews start in May of the survey year, after most income taxes are filed and usually finish by the end of the calendar year, see Kennickell [2017b] for challenges at the end of the interview period. Questions also may change over time so it is important to review the codebook each year when making comparisons across time.

Multiple Imputation

During interviews, respondents may omit answers or provide a range of values for which their response belongs. This missing data impacts analysis and so the SCF contains 5 imputed values for each PEU, creating a sample 5 times larger than the actual number of respondents and forms 5 data sets called implicates. Imputation is done by the Federal Reserve Imputation Technique Zeta model (FRITZ), details can be found in Kennickell [2017a] based upon the ideas of Little and Rubin [2019]. Multiple imputation for missing data provide multiple probable values. Each of these form a data set from which sample statistics can be found. The technique of Repeated Imputation Inference (RII) is applied in our analysis. For each implicate $\ell = 1, 2, 3, 4, 5$, the empirical Inverse Lorenz Function $\widehat{l}_{\ell *}$ is calculated using the appropriate quantile map estimator taking into account sample weights. Then the repeated-imputation estimate of l is

$$\widehat{l}(z) = \frac{1}{5} \sum_{\ell=1}^{5} \widehat{l}_{\ell *}(z).$$

Calculation of the Gini index follows a similar procedure. Accounting for the multiple imputation in the calculation of standard errors is an important issue, but is not revelant to our visualization technique. For more information on multiple imputation and inference with imputed values, see Rubin [1996].

Definition of Wealth

In the literature, there is no consensus on what factors should be included in wealth measurement. Wolff [2021] defines wealth as marketable weath, which is the sum of marketable or fungible assets less the current value of all debts. Bricker et al. [2017] define wealth as net worth including those assets which are not readily transformed into consumption: properties, vehicles, etc. In our analysis we consider all assets, including financial, as our wealth variable.

A.5 Additional details and results

A.5.1 Vector ranks and quantiles

Proposition 9 below, a seminal result in the theory of measure transportation (see Villani (2003, 2009)), states essential uniqueness of the gradient of a convex function (hence cyclically monotone map) that pushes the uniform distribution on $[0,1]^d$ into the distribution of an allocation X.

Following Villani [2003], we let $g_{\#}\nu$ denote the *image measure* (or *push-forward*) of a measure ν by a measurable map $g : \mathbb{R}^d \to \mathbb{R}^d$. Explicitly, for any Borel set A, $g_{\#}\nu(A) := \nu(g^{-1}(A))$. The symbol ∇ denotes the gradient, and D the Jacobian. The convex conjugate of a convex lower semicontinuous function ψ is denoted ψ^*.

Proposition 9 (McCann 1995)**.** Let P and λ be two distributions on $\mathbb{R}^d$. (*i*) If λ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$, with support contained in a convex set $\mathcal{U}$, the following statements hold: there exists a convex function $\psi : \mathcal{U} \to \mathbb{R} \cup \{+\infty\}$ such that $\nabla\psi_{\#}\lambda = P$. The function $\nabla\psi$ exists and is unique, λ-almost everywhere. (*ii*) If, in addition, P is absolutely continuous on $\mathbb{R}^d$ with support contained in a convex set $\mathcal{X}$, the following holds: there exists a convex function $\psi^* : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ such that $\nabla\psi^*_{\#}P = \lambda$. The function $\nabla\psi^*$ exists, is unique and equal to $(\nabla\psi)^{-1}$, P-almost everywhere.

Proposition 9 is an extension of Brenier [1991] (see also Rachev and Rüschendorf [1990]). It removes the finite variance requirement, which is undesirable in our context. Proposition 9 is the basis for the definition of vector quantiles and ranks in Chernozhukov et al. [2017]. In our context, it is applied with uniform reference measure.[2]

In case $d = 1$, gradients of convex functions are nondecreasing functions, hence vector quantiles and ranks reduce to classical quantile and cumulative distribution functions.

[2]This vector quantile notion was introduced in Galichon and Henry [2012] and Ekeland et al. [2012] and called μ-quantile.

As the notation indicates, the function ψ^* of Proposition 9 is the convex conjugate of ψ. In case of absolutely continuous distributions P on $\mathbb{R}^d$ with finite variance, the vector rank function solves a quadratic optimal transport problem, i.e., vector rank R minimizes, among all functions T such that $T(X)$ is uniform on $[0,1]^d$, the quantity $\mathbb{E}\|X - T(X)\|^2$, where $X \sim P$.

Proposition 9 is the basis for Definition 1. In the proofs, we shall use the notation $Q_X = \nabla\psi_X$ for the vector quantile of a random vector X and call convex function ψ_X the transport potential associated with the distribution of X.

A.5.2 Egalitarian multi-attribute allocations

Identical allocations: additional details

In this section, we consider bivariate allocations only. A sufficient condition for Assumption 1 is supermodularity of the potential function ψ_X of allocation X, as shown in Lemma 2 below. We also show in Lemma 2, that supermodularity of the potential function ψ_X also implies positive quadrant dependence of the two components X_1 and X_2 of X, i.e., $\mathbb{P}(X_1 \leq x_1, X_2 \leq x_2) \geq \mathbb{P}(X_1 \leq x_1)\mathbb{P}(X_2 \leq x_2)$, for all $x_1, x_2 \in \mathcal{X}$, see Lehmann [1966].

Lemma 2 (Supermodular potential)**.** *Suppose X has a supermodular potential function, i.e.,*

$$\mathbb{P}(\partial^2\psi_X(U)/\partial u_1\partial u_2 \geq 0) = 1, \text{ with } U \sim U[0,1]^2.$$

Then, Assumption 1 holds, and X_1 and X_2 are positive quadrant dependent.

For allocations satisfying Assumption 1, we show that Lorenz map and Inverse Lorenz Function of the identical allocation serve as upper and lower bounds, respectively. Without Assumption 1, some allocations may have a Lorenz map that is component-wise larger than the Lorenz map of the identical allocation for some ranks. To illustrate the point, consider the potential $\psi_X(u) = (u_1 - u_2)^2/2 + u_1 + u_2$. It corresponds to an allocation X, whose distribution is supported on the line $X_1 + X_2 = 2$. Calculating the

Lorenz map, we obtain

$$\mathcal{L}(r) \quad = \quad \begin{bmatrix} r_1 r_2 (r_1 - r_2)/2 + r_1 r_2 \\ r_1 r_2 (r_2 - r_1)/2 + r_1 r_2 \end{bmatrix}.$$

Notice, in particular, that $\mathcal{L}_1(r) > r_1 r_2$ in the region where $r_1 > r_2$. If the implicit relative price of resource 2 is 1, allocation X is an egalitarian allocation, since all individuals have equal budgets. However, this allocation does not satisfy 1 and its Lorenz map is not dominated by $(r_1 r_2, r_1 r_2)$ as we have shown. This apparent departure from properties of the scalar Lorenz curve is due to the fact that an allocation with $X_1 + X_2 = 2$ a.s. can also be considered egalitarian, as we discuss in the following section.

Egalitarian allocations

The identical allocation with Lorenz map $(r_1 r_2, r_1 r_2)$ is a very special instance of egalitarian allocation. We extend this narrow notion of egalitarian allocation to include income egalitarianism, in the terminology of Kolm [1977]. In the special case where the two resources are transferable with relative price p of the second resource, an allocation is deemed egalitarian if all agents have the same budget endowment, i.e., if $X_1 + pX_2 = 1 + p$ (where the constant value $1 + p$ is derived from the normalization $\mathbb{E}X_1 = \mathbb{E}X_2 = 1$). In the general case of non (or imperfectly) transferable resources, we call egalitarian the allocations with equalized shadow budgets.

Definition 9 (Egalitarian allocation)**.** An allocation X such that $X_1 + pX_2 = 1 + p$ a.s., for some $p > 0$, is called egalitarian.

Another way to interpret egalitarianism of such an allocation, beyond shadow budget equality, is through the perfect compensation of inequality in the marginal resource allocations by perfect negative correlation between resource allocations. The vector quantile and Lorenz map of egalitarian allocations can be characterized in the following way.

Proposition 10. Let (U_1, U_2) be a random vector with distribution $U[0,1]^2$. (i) An egalitarian allocation X such that $X_1 + pX_2 = 1 + p$, admits potential $\psi_X(u_1, u_2) = u_1 + u_2 + v(pu_1 - u_2)$ for some convex function v such that $\int_0^1 v(p-z)dz = \int_0^1 v(z)dz$ and allocation X is equal in distribution to $(1 + pv'(pU_1 - U_2), 1 - v'(pU_1 - U_2))$; (ii) The Lorenz map is given by

$$\mathcal{L}_X(r) \;=\; \begin{bmatrix} r_1 r_2 + \int_0^{r_2} [v(pr_1 - u_2) - v(-u_2)]\, du_2 \\ r_1 r_2 - \frac{1}{p}\int_0^{r_2} [v(pr_1 - u_2) - v(-u_2)]\, du_2 \end{bmatrix};$$

(iii) If, in addition, F_1^{-1} denotes the quantile function of X_1, then

$$v(z) \;=\; \frac{1}{p}\int_0^z \left(F_1^{-1}(H_p(y)) - 1\right) dy,$$

where H_p is the cdf of the random variable $pU_1 - U_2$; see Lemma 3 below for an explicit expression for $H_p(z)$.

Lemma 3 (Explicit formula for $H_p(z)$)**.** *The cumulative distribution function of $Z = pU_1 - U_2$ with $(U_1, U_2) \sim U[0,1]^2$ is given by the following.*

$$H_p(z) \;=\; \begin{cases} 1 & \text{if} \quad p < z, \\ 1 - \frac{p}{2} + z - \frac{z^2}{2p} & \text{if} \quad \max\{p-1, 0\} < z \le p, \\ \frac{1+2z}{2p} & \text{if} \quad 0 < z \le \max\{p-1, 0\}, \\ 1 - \frac{p}{2} + z & \text{if} \quad \min\{p-1, 0\} < z \le 0, \\ \frac{1}{p}\left(\frac{1}{2} + z + \frac{z^2}{2}\right) & \text{if} \quad -1 < z \le \min\{p-1, 0\}, \\ 0 & \text{if} \quad z \le -1. \end{cases}$$

We see in Proposition 10 that the distribution of the egalitarian allocation X is entirely determined by the convex function v, which is itself determined by the distribution

of one of the marginals of X. This follows from the deterministic linear relationship between the two resource allocations. The perfect negative correlation compensates any inequality in the marginal allocations.

With this definition of egalitarian allocations, we show that a large class of allocations are dominated in the Lorenz order by egalitarian allocations, and that egalitarian allocations are maximal in the Lorenz order of Definition 4.

Assumption 10. *For some $p > 0$, the potential ψ_X of allocation X satisfies for all $z \in [-1, p]$:*

$$\sup_{pu_1-u_2=z} \left\{ -\frac{1}{p} \frac{\partial^2 \psi_X(u_1,u_2)}{\partial u_1 \partial u_2} \right\} \;\leq\; \inf_{pu_1-u_2=z} \min \left\{ \frac{1}{p^2} \frac{\partial^2 \psi_X(u_1,u_2)}{\partial u_1^2}, \frac{\partial^2 \psi_X(u_1,u_2)}{\partial u_2^2} \right\}.$$

Before stating the main result of this section, which is an extension of property (7) in Section 2.2.4, we discuss sufficient conditions for Assumption 10 and examples of classes of allocations that satisfy Assumption 10. The following lemma provides sets of sufficient conditions based on a suitable choice of p.

Lemma 4 (Sufficient condition for Assumption 10)**.** *An allocation with potential ψ_X satisfies Assumption 10 if any of the following conditions hold.*

(*i*) *The potential ψ_X is supermodular.*

(*ii*) *The potential ψ_X satisfies:*

$$\sqrt{\inf_{u_1,u_2} \frac{\partial^2 \psi_X(u_1,u_2)}{\partial u_1^2} \times \inf_{u_1,u_2} \frac{\partial^2 \psi_X(u_1,u_2)}{\partial u_2^2}} + \inf_{u_1,u_2} \frac{\partial^2 \psi_X(u_1,u_2)}{\partial u_1 \partial u_2} \geq 0. \tag{A.9}$$

(*iii*) *The function*

$$p(u_1,u_2) \;:=\; \sqrt{\frac{\partial^2 \psi_X(u_1,u_2)}{\partial u_1^2} \Big/ \frac{\partial^2 \psi_X(u_1,u_2)}{\partial u_2^2}}$$

is positive and constant equal to p over $[0,1]^2$ and, for all $z \in [-1, p]$, the Hessian of ψ_X is constant over $pu_1 - u_2 = z$.

The first sufficient condition in Lemma 4, i.e., supermodularity of the potential ψ_X, imposes a form of positive dependence between the two resources, which implies Assumption 10 (and 1). However, Assumption 10 also accommodates allocations that do not exhibit positive dependence. For instance, the mixture of an egalitarian allocation with a positively dependent one satisfies Assumption 10.

Example 7. An allocation with potential $\psi(u_1, u_2) = u_1 + u_2 + v(pu_1 - u_2) + \tilde{\psi}(u_1, u_2)$, with v convex, $p > 0$ and $\tilde{\psi}$ ultramodular, satisfies Assumption 10. It mixes a perfectly negatively correlated allocation with a positively dependent one.

Aspecial case of condition (2) in Lemma 4 is the case, where ψ_X is a quadratic function, hence has a constant Hessian. Indeed, in that case, convexity of ψ_X immediately yields (A.9).

Example 8. All allocations with quadratic potential $\psi_X(u_1, u_2) = a_1u_1 + a_2u_2 + a_{11}u_1^2 + a_{12}u_1u_2 + a_{22}u_2^2$ with $a_1, a_{11}, a_2, a_{22}, a_{12} \in \mathbb{R}$, i.e., allocations of the form $X = (a_1 + 2a_{11}U_1 + a_{12}U_2, a_2 + 2a_{22}U_2 + a_{12}U_1)$, with $(U_1, U_2) \sim U[0,1]^2$, satisfy Assumption 10.

Sufficient condition (2) in Lemma 4 can also be used to show that allocations where the two marginal resource allocations are independent also satisfy Assumption 10. More generally, a large class of allocations defined as deviations from independence satisfy Assumption 10 as formalized in the following example.

Example 9. An allocation X with potential $\psi_X(u_1, u_2) = \phi_1(u_1) + \phi_2(u_2) + \rho\phi(u_1, u_2)$ satisfies Assumption 10 if $\phi_1'' \geq B_1, \phi_2'' \geq B_2, \frac{\partial^2\phi}{\partial u_1^2} \geq B_{11}, \frac{\partial^2\phi}{\partial u_2^2} \geq B_{22}$, $\frac{\partial^2\phi}{\partial u_1 u_2} \geq B_{12}$, and $-\rho B_{12} \leq \sqrt{(B_1 + \rho B_{11})(B_2 + \rho B_{22})}$ with $B_1, B_{11}, B_2, B_{22}, B_{12} \in \mathbb{R}$. The case $\rho = 0$ is the case of independent marginal allocations.

Assumption 10 is not satisfied, however, in case X_1 and X_2 are perfectly negatively dependent, i.e., $X_2 = -\phi(X_1)$ with increasing ϕ, when ϕ is nonlinear.

Under Assumption 10, we can complement property (7) in Section 2.2.4 and emulate the traditional property of Lorenz curves, which are maximal at perfect equality. Here we

show that egalitarian allocations dominate all allocations that satisfy Assumption 10, and are themselves undominated thereby forming a class of distributions that are maximal under the Lorenz order.

Property (7) continued [Lorenz map maximal at egalitarian allocations] For any allocation X satisfying Assumption 10, there is an egalitarian allocation $\tilde{X}$ such that $X \preccurlyeq_{\mathcal{L}} \tilde{X}$, i.e., $\mathcal{L}_X(r) \leq \mathcal{L}_{\tilde{X}}(r)$ for all $r \in [0,1]^2$. In addition, if two egalitarian allocations are ranked in the Lorenz order, then they are equal.

Proofs for Section A.5.2

Proof of Lemma 2. Let $U \sim U[0,1]^2$. Let $\tilde{X}_j := \frac{\partial \psi}{\partial u_j}(U_1, U_2)$, $j = 1, 2$. Then $(\tilde{X}_1, \tilde{X}_2)$ is distributed identically to (X_1, X_2). Since $\tilde{X}_2 = \frac{\partial \psi}{\partial u_2}(U_1, U_2)$ is monotonically increasing in U_2, we have $U_2 = \left(\frac{\partial \psi}{\partial u_2}\right)^{-1}(U_1, \tilde{X}_2)$. Hence

$$\tilde{X}_1 \;=\; \frac{\partial \psi}{\partial u_1}\left(U_1, \left(\frac{\partial \psi}{\partial u_2}\right)^{-1}(U_1, \tilde{X}_2)\right).$$

Under the stated Assumption, $\tilde{X}_1$ is increasing in U_1 and $\tilde{X}_2$. Since $\left(\tilde{X}_1, \tilde{X}_2\right) \overset{d}{=} (X_1, X_2)$, we have

$$\begin{aligned}
F_X(x_1, x_2) &= \mathbb{P}\left(\tilde{X}_1 \leq x_1, \tilde{X}_2 \leq x_2\right) \\
&= \mathbb{E}\left[\mathbb{P}\left(\frac{\partial \psi}{\partial u_1}\left(U_1, \left(\frac{\partial \psi}{\partial u_2}\right)^{-1}(U_1, \tilde{X}_2)\right) \leq x_1, \tilde{X}_2 \leq x_2\right) \,\middle|\, U_1\right] \\
&= \mathbb{E}\left[\min\left\{F_1(x_1|U_1), F_2(x_2|U_1)\right\}\right] \\
&\geq \mathbb{E}\left[F_1(x_1|U_1)\, F_2(x_2|U_1)\right],
\end{aligned}$$

where $F_i(\cdot|U_j)$ denotes the cumulative distribution function of X_i conditional on U_j. Now $F_1(x_1|U_1)$ is increasing in U_1, since

$$\begin{aligned} F_1(x_1|U_1) &= \mathbb{P}\left(\frac{\partial\psi}{\partial u_1}(U_1,U_2) \leq x_1 \middle| U_1\right) \\ &= \mathbb{P}\left(U_2 \leq \left(\frac{\partial\psi}{\partial u_1}\right)^{-1}(U_1,x_1) \middle| U_1\right) \\ &= \left(\frac{\partial\psi}{\partial u_1}\right)^{-1}(U_1,x_1). \end{aligned}$$

Similarly $F_2(x_2|U_1)$ is increasing in U_1. We conclude that $F_X(x_1,x_2) \geq F_1(x_1)F_2(x_2)$, see e.g. Joe [1997]. □

Proof of Proposition 10. The potential ψ of an egalitarian allocation satisfies $\partial\psi/\partial u_1 + p\partial\psi/\partial u_2 = 1+p$. Solutions are of the form

$$\psi_{(v,p)}(u_1,u_2) = u_1 + u_2 + v(pu_1 - u_2).$$

Convexity of ψ implies convexity of v. The normalization

$$\int_0^1\int_0^1 \nabla\psi_{(v,p)}(u_1,u_2)du_1du_2 = (1,1)$$

implies

$$\int_0^1\int_0^1 v'(pu_1-u_2)du_1du_2 = 0.$$

The latter, in turn, implies

$$\int_0^1 v(p-x)dx = \int_0^1 v(x)dx.$$

Call H_p the cdf of $Z = pU_1 - U_2$, where $(U_1,U_2) \sim U[0,1]^2$. Call F_1 the cdf of $\nabla_1\psi_{(p,v)} := 1 + pv'(Z)$, which is the first marginal of allocation X. Then

$$\begin{aligned} F_1(x) &= \mathbb{P}\left(v'(Z) \leq \frac{x-1}{p}\right) \\ &= \mathbb{P}\left(Z \leq (v')^{-1}\left(\frac{x-1}{p}\right)\right) \\ &= H_p\left((v')^{-1}\left(\frac{x-1}{p}\right)\right). \end{aligned}$$

Now

$$\begin{aligned} F_1(x) = H_p\left((v')^{-1}\left(\frac{x-1}{p}\right)\right) &\Rightarrow (v')^{-1}\left(\frac{x-1}{p}\right) = H_p^{-1}\left(F_1(x)\right) \\ &\Rightarrow \frac{x-1}{p} = v'\left(H_p^{-1}\left(F_1(x)\right)\right) \\ &\Rightarrow v'(z) = \frac{F_1^{-1}(H_p(z))-1}{p}. \end{aligned}$$

Hence

$$v(z) = \int_0^x \frac{F_1^{-1}(H_p(y))-1}{p} dy,$$

as desired. □

Proof of Lemma 4. A sufficient condition for Assumption 10 is

$$-\inf_{u_1,u_2} \frac{1}{p}\frac{\partial^2\psi_X(u_1,u_2)}{\partial u_1 \partial u_2} \leq \min\left\{\inf_{u_1,u_2}\frac{1}{p^2}\frac{\partial^2\psi_X(u_1,u_2)}{\partial u_1^2}, \inf_{u_1,u_2}\frac{\partial^2\psi_X(u_1,u_2)}{\partial u_2^2}\right\}.$$

If we choose the optimal value of p, i.e.,

$$p^2 = \frac{\inf_{u_1,u_2}\frac{\partial^2\psi_X(u_1,u_2)}{\partial u_1^2}}{\inf_{u_1,u_2}\frac{\partial^2\psi_X(u_1,u_2)}{\partial u_2^2}},$$

we get the sufficient inequality

$$-\inf_{u_1,u_2}\frac{\partial^2\psi_X(u_1,u_2)}{\partial u_1 \partial u_2} \leq \sqrt{\inf_{u_1,u_2}\frac{\partial^2\psi_X(u_1,u_2)}{\partial u_1^2} \times \inf_{u_1,u_2}\frac{\partial^2\psi_X(u_1,u_2)}{\partial u_2^2}}$$

as desired. □

Proof of Example 2. Let $\psi(u_1,u_2) = u_1 + u_2 + v(pu_1 - u_2) + \tilde{\psi}(u_1,u_2)$, with v convex and twice continuously differentiable, and ψ ultramodular. We have, for $j = 1, 2$,

$$\frac{\partial^2\psi(u_1,u_2)}{\partial u_1^2} = p^2 v''(pu_1 - u_2) + \frac{\partial^2\tilde{\psi}(u_1,u_2)}{\partial u_2^2},$$

$$\frac{\partial^2\psi(u_1,u_2)}{\partial u_1^2} = v''(pu_1 - u_2) + \frac{\partial^2\tilde{\psi}(u_1,u_2)}{\partial u_2^2}.$$

Also,

$$\frac{\partial^2\psi(u_1,u_2)}{\partial u_1\partial u_2} = -pv''(pu_1-u_2)+\frac{\partial^2\tilde{\psi}(u_1,u_2)}{\partial u_1\partial u_2}.$$

Therefore

$$\begin{aligned}
\sup_{pu_1-u_2=z}\left\{-\frac{1}{p}\frac{\partial^2\psi(u_1,u_2)}{\partial u_1\partial u_2}\right\} &= \sup_{pu_1-u_2=z}\left\{v''(pu_1-u_2)-\frac{1}{p}\frac{\partial^2\tilde{\psi}(u_1,u_2)}{\partial u_1\partial u_2}\right\} \\
&= v''(z)-\inf_{pu_1-u_2=z}\left\{\frac{1}{p}\frac{\partial^2\tilde{\psi}(u_1,u_2)}{\partial u_1\partial u_2}\right\} \\
\leq v''(z) &+ \min\left\{\inf_{pu_1-u_2=z}\frac{1}{p^2}\frac{\partial^2\tilde{\psi}(u_1,u_2)}{\partial u_1^2},\inf_{pu_1-u_2=z}\frac{\partial^2\tilde{\psi}(u_1,u_2)}{\partial u_2^2}\right\} \\
&= \inf_{pu_1-u_2=z}\min\left\{\frac{1}{p^2}\frac{\partial^2\psi(u_1,u_2)}{\partial u_1^2},\frac{\partial^2\psi(u_1,u_2)}{\partial u_2^2}\right\}.
\end{aligned}$$

□

Proof of "property (7) continued" in A.5.2. Define

$$v''(z) := \inf_{pu_1-u_2=z}\min\left\{\frac{1}{p^2}\frac{\partial^2\psi_X(u_1,u_2)}{\partial u_1^2},\frac{\partial^2\psi_X(u_1,u_2)}{\partial u_2^2}\right\}.$$

Under Assumption 10,

$$v''(z) \geq \sup_{pu_1-u_2=z}\left\{-\frac{1}{p}\frac{\partial^2\psi_X(u_1,u_2)}{\partial u_1\partial u_2}\right\}.$$

Hence $\psi := \psi_X(u_1,u_2)-v(pu_1-u_2)$ is an ultramodular function. Applying the proof of property (7) in Section 2.2.4, we find that for all $(r_1,r_2)\in[0,1]^2$,

$$\int_0^{r_1}\int_0^{r_2}\nabla\psi(u_1,u_2)du_1du_2 \leq (r_1r_2,r_1r_2).$$

Hence

$$\int_0^{r_1}\int_0^{r_2}\nabla\psi_X(u_1,u_2)du_1du_2 \leq \int_0^{r_1}\int_0^{r_2}\nabla\psi_{(v,p)}(u_1,u_2)du_1du_2,$$

where $\psi_{(v,p)}(u_1,u_2) := v(pu_1-u_2)+u_1+u_2$ as desired.

We now show that egalitarian allocations do not dominate each other. Suppose an egalitarian allocation $X_{(v,p)}$ with potential $v(pu_1 - u_2) + u_1 + u_2$ dominates an allocation $X_{(\tilde{v},\tilde{p})}$ with potential $\tilde{v}(\tilde{p}u_1 - u_2) + u_1 + u_2$. Then

$$\begin{bmatrix} r_1 r_2 + \int_0^{r_1}\int_0^{r_2} [1 + p\tilde{v}'(\tilde{p}r_1 - u_2)]\, du_1 du_2 \\ r_1 r_2 + \int_0^{r_1}\int_0^{r_2} [1 - \tilde{v}'(\tilde{p}r_1 - u_2)]\, du_1 du_2 \end{bmatrix} \leq \begin{bmatrix} r_1 r_2 + \int_0^{r_1}\int_0^{r_2} [1 + pv'(pr_1 - u_2)]\, du_1 du_2 \\ r_1 r_2 + \int_0^{r_1}\int_0^{r_2} [1 - v'(pr_1 - u_2)]\, du_1 du_2 \end{bmatrix}.$$

Hence, for all $(r_1, r_2) \in [0,1]^2$,

$$\int_0^{r_1}\int_0^{r_2} \tilde{v}'(\tilde{p}u_1 - u_2) \;=\; \int_0^{r_1}\int_0^{r_2} v'(pu_1 - u_2),$$

so that both allocations have the same Lorenz map, hence are equally distributed. □

A.5.3 Uniform Majorization

In this section, we show the undesirable feature of uniform majorization detailed in Section 2.3.1. Galichon and Henry [2012] show that the only rank dependent social evaluation functional (up to an affine transformation) that satisfies the uniform majorization principle of Kolm [1977] is given in (2.19). We now show that it is unsuitable as a tool to measure multivariate inequality with the following two observations about bivariate allocations X. First, the following expression shows that S_{UM} only depends on $\mathcal{L}_1(p,1) + \mathcal{L}_2(1,p)$, hence, on very specific features of the dependence between the two components X_1 and X_2 of allocation X.

$$S_{UM}(X) \;=\; \int_0^1 \Big([\mathcal{L}_1(p,1) - p] + [\mathcal{L}_2(1,p) - p] \Big)\, dp. \tag{A.10}$$

Second, and more troubling still, for any given fixed marginals for X_1 and X_2, S_{UM} is minimized when X_1 and X_2 are independent. Indeed, we show below that S_{UM} is always larger than minus the average of univariate Gini coefficients, which is its value

when the marginals are independent.

$$S_{UM}(X) \;\geq\; -\frac{1}{2}\left[\, G\left(X_1\right) + G\left(X_2\right) \right], \tag{A.11}$$

where $G(X_1)$ and $G(X_2)$ denote the classical scalar Gini index of components X_1 and X_2 respectively.

Proof of (A.10). Let U be uniformly distributed on $[0,1]^2$. Note that

$$S_{UM}(X) \;=\; 1 - \mathbb{E}\left[U_1 \frac{\partial \psi}{\partial u_1}(U_1, U_2)\right] - \mathbb{E}\left[U_1 \frac{\partial \psi}{\partial u_1}(U_1, U_2)\right].$$

Now,

$$\begin{aligned}
\mathbb{E}\left[U_1 \frac{\partial \psi}{\partial u_1}(U_1, U_2)\right] &= \int_0^1 \left[\int_0^1 u_1 \frac{\partial \psi}{\partial u_1}(u_1, u_2)\, du_1\right] du_2 \\
&= \int_0^1 \left[\int_0^1 u_1\, d\left(\frac{\partial \mathcal{L}_1}{\partial u_2}(u_1, u_2)\right)\right] du_2,
\end{aligned}$$

where the last equality follows from interchangeability of the order of integration and $\mathcal{L}_1$ is the first component of the Lorenz map. Note that

$$\frac{\partial \mathcal{L}_1}{\partial r_2}(r_1, r_2) \;=\; \int_0^{r_1} \frac{\partial \psi}{\partial u_1}(u_1, u_2)\, du_1$$

and

$$\frac{\partial}{\partial r_1}\left(\frac{\partial \mathcal{L}_1(r_1, r_2)}{\partial r_2}\right) \;=\; \frac{\partial \psi}{\partial u_1}(r_1, r_2).$$

Therefore

$$\begin{aligned}
\mathbb{E}\left[U_1 \frac{\partial \psi}{\partial u_1}(U_1, U_2)\right] &= \int_0^1 \left(u_1 \frac{\partial \mathcal{L}_1}{\partial u_2}(u_1, u_2)\Big|_0^1 - \int_0^1 \frac{\partial \mathcal{L}_1}{\partial u_2}(u_1, u_2)\, du_1\right) du_2 \\
&= \int_0^1 \left(\frac{\partial \mathcal{L}_1}{\partial u_2}(1, u_2) - \int_0^1 \frac{\partial \mathcal{L}_1}{\partial u_2}(u_1, u_2)\, du_1\right) du_2 \\
&= \int_0^1 \frac{\partial \mathcal{L}_1}{\partial u_2}(1, u_2)\, du_2 - \int_0^1 \int_0^1 \frac{\partial \mathcal{L}_1}{\partial u_2}(u_1, u_2)\, du_2\, du_1 \\
&= \mathcal{L}(1,1) - \mathcal{L}(1,0) - \int_0^1 \mathcal{L}_1(u_1, 1) - \mathcal{L}_1(u_1, 0)\, du_1 \\
&= 1 - \int_0^1 \mathcal{L}_1(u_1, 1)\, du_1.
\end{aligned}$$

Similarly, we have $\mathbb{E}\left[U_2 \frac{\partial \psi}{\partial u_2}(U_1, U_2)\right] = 1 - \int_0^1 \mathcal{L}_2(1, u_2)\, du_1$, as desired. □

Proof of (A.11). The inequality follows from $\mathcal{L}_1(p,1) \geq L_1(p)$ and $\mathcal{L}_2(1,p) \geq L_2(p)$. We now prove the latter. Letting $\nabla\psi$ be the vector quantile function of (X_1, X_2), note that since $\frac{\partial\psi}{\partial u_1}$ pushes uniform measure on $[0,1]^2$ forward to law(X_1), we can write

$$L_1(r_1) = \int_{\{u:\frac{\partial\psi}{\partial u_1}(u)\leq z_1(r_1)\}} \frac{\partial\psi}{\partial u_1}(u)du$$

where $z_1(r_1)$ is the quantile of the random variable X_1. Note that the area of the domain $\{u : \frac{\partial\psi}{\partial u_1}(u) \leq z_1(r_1)\}$ of integration must be r_1. On the other hand,

$$\mathcal{L}_1(r_1,1) = \int_0^{r_1}\int_0^1 \frac{\partial\psi}{\partial u_1}(u)du_1du_2$$

is an integral of the same function over a region with the same area. Writing $A := \left\{u : \frac{\partial\psi}{\partial u_1}(u) \leq z_1(r_1)\right\}$, we have $A = B \cup C$, where $B = A \cap ([0,r_1]\times[0,1])$ and $C = A \cap ((r_1,1]\times[0,1])$ and the union is disjoint. Similarly, $[0,r_1]\times[0,1] = B \cup D$ where $D = ([0,r_1]\times[0,1]) \cap A^c$. Note that the areas of C and D must be the same, $|C| = |D|$, and $\frac{\partial\psi}{\partial u_1}(u) \leq z_1(r_1)$ throughout C while $\frac{\partial\psi}{\partial u_1}(u) > z_1(r_1)$ throughout D. We have

$$\begin{aligned}
L_1(r_1) &= \int_B \frac{\partial\psi}{\partial u_1}(u)du + \int_C \frac{\partial\psi}{\partial u_1}(u)du \\
&\leq \int_B \frac{\partial\psi}{\partial u_1}(u)du + z_1(r_1)|C| \\
&= \int_B \frac{\partial\psi}{\partial u_1}(u)du + z_1(r_1)|D| \\
&\leq \int_B \frac{\partial\psi}{\partial u_1}(u)du + \int_D \frac{\partial\psi}{\partial u_1}(u)du \\
&= \mathcal{L}_1(r_1,1).
\end{aligned}$$

Note that this inequality holds for any dependence structure between X_1 and X_2. □

A.6 Inequality Dominance based on the Inverse Lorenz Function

We can also define an increasing inequality order based on the Inverse Lorenz Functions. Consider two allocations X and X', with respective Inverse Lorenz Functions l_X and $l_{X'}$. If $l_X(z) \leq l_{X'}(z)$ for some vector of shares z, a larger proportion of the population

commands the same share of resources in allocation X' than in allocation X. If this is true for any vector z of resource shares in $[0,1]^d$, then, we say that allocation X' is more unequal than allocation X.

Definition 10. An allocation X' is said to be more unequal in the weak Lorenz order than an allocation X if $l_{X'}(z) \geq l_X(z)$ for all $z \in [0,1]^d$. We denote this $X \succcurlyeq_l X'$.

The relation $X \succcurlyeq_l X'$ is equivalent to lower orthant dominance of the random vector $\mathcal{L}_X(U)$, with $U \sim U[0,1]^d$, over $\mathcal{L}_{X'}(U)$ (see Section 3.8 of Müller and Stoyan [2002]). In the scalar case, the orderings of Definitions 4 and 10 both coincide with the traditional Lorenz ordering. In higher dimensions, however, the equivalence may not hold[3]. Nonetheless, as the name indicates, the weak Lorenz inequality order of Definition 10 is weaker than the Lorenz order of Definition 4, as we show in Proposition 11.

Proposition 11. An allocation X' is more unequal in the weak Lorenz order than an allocation X, i.e., $X \succcurlyeq_l X'$ (Definition 10) if X' is more unequal in the Lorenz order, i.e., $X \preccurlyeq_{\mathcal{L}} X'$ (Definition 4).

Proof of Proposition 11. $\tilde{X} \succcurlyeq_{\mathcal{L}} X$ is equivalent to first order stochastic dominance of $\mathcal{L}_X(U)$ over $\mathcal{L}_{\tilde{X}}(U)$, where $U \sim U[0,1]^d$ (see Section 6.B page 266 of Shaked and Shanthikumar [2007]). Hence, $\tilde{X} \succcurlyeq_{\mathcal{L}} X$ implies $\mathbb{P}(\mathcal{L}_X(U) \in S) \leq \mathbb{P}(\mathcal{L}_{\tilde{X}}(U) \in S)$ for any lower set S, so that $\tilde{X} \succcurlyeq_{\mathcal{L}} X$ implies $\tilde{X} \succcurlyeq_l X$, given that the sets $[0,z]$ are lower sets. □

A.7 Proofs of the main results

Recall that in the proofs, we shall use the notation $Q_X = \nabla \psi_X$ for the vector quantile of a random vector X and call convex function ψ_X the transport potential associated with the distribution of X. See Section A.5.1 for details. In this section, we omit the X subscript of ψ_X for notational compactness.

[3]We have not been able to either prove the equivalence or find a counterexample.

Proof of Proposition 1. In case $d = 2$, the off diagonal elements of the Jacobian of $\mathcal{L}(r)$ are $\psi(r_1, r_2) - \psi(r_1, 0)$ and $\psi(r_1, r_2) - \psi(0, r_2)$. From the latter, by differentiation, we can recover $\nabla\psi(r_1, r_2)$. The result then follows from the fact that $\nabla\psi$ characterizes P_X, see for instance Chernozhukov et al. [2017]. The result extends straightforwardly to $d > 2$. □

Proof of Proposition 2. We only need to show the result for one component of the Lorenz map and the others follow with similar reasoning. We have for the first component

$$\begin{aligned}
\mathcal{L}_1(r) &= \int_0^{r_d} \cdots \int_0^{r_2} \int_0^{r_1} \frac{\partial \psi}{\partial u_1}(u_1, u_2, ..., u_d) du_1 du_2 ... du_d \\
&= \int_0^{r_d} \cdots \int_0^{r_2} [\psi(r_1, u_2, ..., u_d) - \psi(0, u_2, ..., u_d)] du_2 ... du_d \\
&\leq \int_0^{r_d} \cdots \int_0^{r_2} r_1 [\psi(1, u_2, ..., u_d) - \psi(0, u_2, ..., u_d)] du_2 ... du_d, \text{ by convexity} \\
&= r_1 \int_0^{r_d} \cdots \int_0^{r_2} \int_0^{1} \frac{\partial \psi}{\partial u_1}(u_1, u_2, ..., u_d) du_1 du_2 ... du_d
\end{aligned}$$

Now define $H(r_2, ..., r_d) = \int_0^{r_d} \cdots \int_0^{r_2} \int_0^1 \frac{\partial \psi}{\partial u_1}(u_1, u_2, ..., u_d) du_1 du_2 ... du_d$. Then

$$\frac{\partial H}{\partial r_2}(r_2, ..., r_d) = \int_0^{r_d} \cdots \int_0^{r_3} \int_0^{1} \frac{\partial \psi}{\partial u_1}(u_1, r_2, u_3, ..., u_d) du_1 du_3 ... du_d$$

is monotone increasing in r_2, since it is the integral of the functions $r_2 \mapsto \int_0^1 \frac{\partial \psi}{\partial u_1}(u_1, r_2, u_3, ..., u_d) du_1$, which are monotonically increasing by assumption. Therefore, $r_2 \mapsto H(r_2, ...r_d)$ is convex and so $H(r_2, r_3, ..., r_d) \leq H(0, r_3, ..., r_d) + r_2(H(1, r_3, ..., r_d) - H(0, r_3, ..., r_d))$. Note that $H(0, r_3, ..., r_d) = 0$ as an integral over a degenerate interval, so $H(r_2, r_3, ..., r_d) \leq r_2 H(1, r_3, ..., r_d)$. A very similar argument yields $H(1, r_3, ..., r_d) \leq r_3 H(1, 1, r_4, ..., r_d)$, so that $H(r_2, r_3, ..., r_d) \leq r_2 r_3 H(1, 1, r_4, ..., r_d)$, and, iterating in this way, we eventually obtain,

$$H(r_2, r_3, ..., r_d) \leq r_2 r_3 \cdots r_d H(1, 1, ..., 1).$$

We then conclude

$$\mathcal{L}_1(r) \leq r_1 r_2 r_3 \cdots r_d \int_0^1 \cdots \int_0^1 \frac{\partial \psi}{\partial u_1}(u_1, u_2, ..., u_d) du_1 du_2 du_3 ... du_d.$$

The integral $\int_0^1 \cdots \int_0^1 \frac{\partial \psi}{\partial u_1}(u_1, u_2, ..., u_d) du_1 du_2 du_3 ... du_d$ is 1, as the expected value of the normalized X_1, and so we obtain the desired result. □

Proof of Proposition 3. We need to show that a social evaluation functional S of the form (2.15) satisfies

$$X \succcurlyeq_{\mathcal{L}} X' \Rightarrow S(X) \geq S(X')$$

if and only if ϕ is of the form (2.16). To show this, we note that

$$\begin{aligned}
S(X) &= \int_{[0,1]^d} \phi_m(u)' \nabla \psi_{\tilde{X}}(u)\, du \\
&= \sum_{j=1}^{d} \int_{[0,1]^d} \phi_{m,j}(u) \nabla \psi_{\tilde{X},j}(u)\, du \\
&= \sum_{j=1}^{d} \int_{[0,1]^d} \left[\int_{[0,1]^d} \mathbb{1}\{u \leq r\}\, dm_j(r) \right] \nabla \psi_{\tilde{X},j}(u)\, du \\
&= \sum_{j=1}^{d} \int_{[0,1]^d} \left[\int_{[0,1]^d} \mathbb{1}\{u \leq r\} \nabla \psi_{\tilde{X},j}(u)\, du \right] dm_j(r) \\
&= \sum_{j=1}^{d} \int_{[0,1]^d} \mathcal{L}_j(r)\, dm_j(r).
\end{aligned}$$

So sufficiency holds for all non-negative measures. Necessity follows by taking one of the measures as a degenerate measure and the other measures as trivial measures. □

Proof of Proposition 4. It suffices to see that $\mathcal{L}_X(r) \geq \mathcal{L}_{X'}(r)$ if and only if

$$\int_{[0,1]^d} \mathbb{1}\{u \leq r\} \left(\nabla \psi_{\tilde{X}}(u) - \nabla \psi_{\tilde{X}'}(u) \right)\, du \geq 0.$$

□

Proof of Proposition 5. Suppose X' is obtained from X through an MRT, so that there is a supermodular and component-wise convex function ψ, such that $\nabla \psi_{\tilde{X}'}(u) = \nabla \psi_{\tilde{X}}(u) + \nabla \psi(u)$ holds for all $u \in [0,1]^d$. We want to show $\mathcal{L}_{X'} \leq \mathcal{L}_X$, i.e., $\iint^r \nabla \psi(u) du \leq 0$.

Consider the first component:

$$\begin{aligned}\int_0^{r_1}\int_0^{r_2}...\int_0^{r_d}\frac{\partial\psi(u)}{\partial u_1}du &= \int_0^{r_2}...\int_0^{r_d}[\psi(r_1,u_2,...,u_d)-\psi(0,u_2,...,u_d)]du_2...du_d \\ &\leq r_1\int_0^{r_2}...\int_0^{r_d}[\psi(1,u_2,...,u_d)-\psi(0,u_2,...,u_d)]du_2...du_d \\ &= r_1\int_0^{1}\int_0^{r_2}...\int_0^{r_d}\frac{\partial\psi(u)}{\partial u_1}du_1u_2....du_d\end{aligned}$$

by convexity.

Now, note that supermodularity implies the function $\int_0^{r_2}...\int_0^{r_d}\frac{\partial\psi(u)}{\partial u_1}du_2....du_d$ is convex with respect to r_2, and it is clearly 0 when $r_2=0$. Therefore,

$$\int_0^{r_2}...\int_0^{r_d}\frac{\partial\psi(u)}{\partial u_1}du_2....du_d \leq r_2\int_0^{1}\int_0^{r_3}...\int_0^{r_d}\frac{\partial\psi(u)}{\partial u_1}du_2....du_d.$$

Similarly,

$$\int_0^{r_3}...\int_0^{r_d}\frac{\partial\psi(u)}{\partial u_1}du_3....du_d \leq r_3\int_0^{1}\int_0^{r_4}...\int_0^{r_d}\frac{\partial\psi(u)}{\partial u_1}du_3....du_d.$$

Continuing iteratively in this way, we eventually obtain:

$$\begin{aligned}\int_0^{r_1}\int_0^{r_2}...\int_0^{r_d}\frac{\partial\psi(u)}{\partial u_1}du &\leq r_1r_2...r_d\int_0^{1}\int_0^{1}...\int_0^{1}\frac{\partial\psi(u)}{\partial u_1}du \\ &= r_1r_2...r_d[(L_{X'})_1(1,1,...1)-(L_X)_1(1,1,...1)]=0\end{aligned}$$

Similar reasoning applies to the other components, and the result follows. □

Proof of Proposition 6. (1) Let $l_X(z_1,z_2)=\alpha$. Since l_X is a cdf, hence non decreasing in both arguments, then $z_2:=l_X^{-1}(z_1;\alpha)=\inf\{\zeta;\alpha\leq l_X(z_1,\zeta)\}$ is non increasing in z_1 and non decreasing in α. (2) See Claim 1 in Brock and Thomson [1966]. □

Proposition 12. (Rawlesian limits of social welfare functions) For a random variable X in $\mathbb{R}$, define, as in Section 2.3.3, the social welfare function:

$$S_\delta(X):=\delta(\delta-1)\int_{[0,1]}(1-r)^{\delta-2}\mathcal{L}_X(r)\,dr.$$

Then $\lim_{\delta\to\infty}S_\delta(X)=Q_X(0+):=\lim_{u\to0^+}Q_X(u)$.

Proof of Proposition 12. Integrating by parts implies

$$S_\delta(X) = \delta \int_0^1 (1-r)^{\delta-1} Q_X(r) dr = \delta \sum_{i=1}^{N} \int_{(i-1)/N}^{i/N} (1-r)^{\delta-1} Q_X(r) dr$$

for any positive integer N. As the quantile $Q_X(\cdot)$ is increasing, we then have

$$\begin{aligned} S_\delta(X) &\leq \delta \sum_{i=1}^{N} Q_X(i/N) \int_{(i-1)/N}^{i/N} (1-r)^{\delta-1} dr \\ &= \sum_{i=1}^{N} Q_X(i/N)[-(1-r)^\delta]|_{(i-1)/N}^{i/N} \\ &= Q_X(1/N)(1-(1-1/N)^\delta) + \sum_{i=2}^{N} Q_X(i/N)[(1-\frac{i-1}{N})^\delta - (1-\frac{i}{N})^\delta] \\ &= Q_X(1/N) + M(\delta), \end{aligned}$$

where the function $M(\delta) := -Q_X(1/N)(1-1/N)^\delta + \sum_{i=2}^{N} Q_X(i/N)[(1-\frac{i-1}{N})^\delta - (1-\frac{i}{N})^\delta]$ tends to 0 as $\delta \to \infty$.

Similarly, we have

$$\begin{aligned} S_\delta(X) &\geq \delta \sum_{i=1}^{N} Q_X((i-1)/N) \int_{(i-1)/N}^{i/N} (1-r)^{\delta-1} dr \\ &= Q_X(0+) + m(\delta) \end{aligned}$$

where $m(\delta) := -Q_X(0+)(1-1/N)^\delta + \sum_{i=2}^{N} Q_X((i-1)/N)[(1-\frac{i-1}{N})^\delta - (1-\frac{i}{N})^\delta]$ tends to 0 as $\delta \to \infty$. We therefore have

$$Q_X(0+) \leq \liminf_{\delta\to\infty} S_\delta(X) \leq \limsup_{\delta\to\infty} S_\delta(X) \leq Q_X(1/N).$$

As this holds for every integer N, and $\lim_{N\to\infty} Q_X(1/N) = Q_X(0+)$, the result follows.

□

The following is a multivariate extension:

Proposition 13. Let X be a d dimensional random variable and Q_X its multivariate quantile. Consider the jth term in the sum defining the multivariate S-Gini in Section 2.3.3:

$$S^j_\delta(X) := \delta_j(\delta_j - 1)\int_{[0,1]^d}(1-r_j)^{\delta_j-2}\mathcal{L}_j(r)\,dr.$$

Then

$$\lim_{\delta_j\to\infty} S^j_\delta(X) = \int_{[0,1]^{d-1}} \Pi^d_{i=1,i\neq j}(1-u_i)Q_j(u_1,u_2,...u_{j-1},0+,u_{j+1},...u_d)du_1du_2...du_{j-1}du_{j+1}...du_d,$$

where $Q_j(u_1,u_2,...u_{j-1},0+,u_{j+1},...u_d) = \lim_{u_j\to 0^+} Q_j(u)$.

Note that the expression is a measure of inequality among agents with lowest rank in the jth component, with weighting functions in the other rank variables as in the standard Gini index.

Proof. The proof is similar to the argument above. Without loss of generality, we assume $j = 1$. Integrating by parts, we have, for any positive integer N:

$$\begin{aligned} S^1_\delta(X) &= \delta_1\int_{[0,1]^d}(1-r_1)^{\delta_1-1}\Big[\int_0^{r_2}\cdots\int_0^{r_d} Q_1(r_1,u_2,...,u_d)du_2...du_d\Big]dr \\ &= \delta_1\int_{[0,1]^{d-1}}\Big(\sum_{i=1}^N\int_{(i-1)/N}^{i/N}(1-r_1)^{\delta_1-1}\Big[\int_0^{r_2}\cdots\int_0^{r_d} Q_1(r_1,u_2,...,u_d)du_2...du_d\Big]dr_1\Big)dr_2...dr_d \end{aligned}$$

Using monotonicity of Q in the first argument, we then have:

$$\begin{aligned} &\delta_1\textstyle\sum_{i=1}^N\int_{(i-1)/N}^{i/N}(1-r_1)^{\delta_1-1}\Big[\int_0^{r_2}\cdots\int_0^{r_d} Q_1(r_1,u_2,...,u_d)du_2...du_d\Big]dr_1 \\ \leq\ &\delta_1\textstyle\sum_{i=1}^N\Big[\int_0^{r_2}\cdots\int_0^{r_d} Q_1(i/N,u_2,...,u_d)du_2...du_d\Big]\int_{(i-1)/N}^{i/N}(1-r_1)^{\delta_1-1}dr_1 \\ =\ &\textstyle\sum_{i=1}^N\Big[\int_0^{r_2}\cdots\int_0^{r_d} Q_1(i/N,u_2,...,u_d)du_2...du_d\Big]\Big[(1-\frac{i-1}{N})^{\delta_1}-(1-\frac{i}{N})^{\delta_1}\Big] \\ &= \int_0^{r_2}\cdots\int_0^{r_d} Q_1(1/N,u_2,...,u_d)du_2...du_d + M(\delta_1,r_2,...,r_d) \end{aligned} \tag{A.12}$$

where $M(\delta_1,r_2,...,r_d)$ converges monotonically to 0 as $\delta_1\to\infty$.

Therefore, using the monotone convergence theorem, we have

$$\begin{aligned} \limsup_{\delta_1\to\infty} S^1_\delta(X) &\leq \int_{[0,1]^{d-1}}\Big(\int_0^{r_2}\cdots\int_0^{r_d} Q_1(1/N,u_2,...,u_d)du_2...du_d\Big)dr_2...dr_d \\ &= \int_{[0,1]^{d-1}}\Pi^d_{i=2}(1-u_j)Q_1(1/N,u_2,...,u_d)du_2...du_d \end{aligned}$$

As this holds for all N, and Q_1 is monotone in the first argument, we can take the limit as $N \to \infty$ and apply the monotone convergence theorem again to obtain

$$\limsup_{\delta_1 \to \infty} S^1_\delta(X) \leq \int_{[0,1]^{d-1}} \Pi_{i=2}^d (1-u_j) Q_1(0+, u_2, ..., u_d) du_2 ... du_d$$

A very similar argument to the one used to derive (A.12) then implies:

$$\begin{aligned} &\delta_1 \textstyle\sum_{i=1}^N \int_{(i-1)/N}^{i/N} (1-r_1)^{\delta_1 - 1} \Big[\int_0^{r_2} \cdots \int_0^{r_d} Q_1(r_1, u_2, ..., u_d) du_2 ... du_d \Big] dr_1 \\ &\quad \geq \int_0^{r_2} \cdots \int_0^{r_d} Q_1(0+, u_2, ..., u_d) du_2 ... du_d + m(\delta_1, r_2, ..., r_d) \end{aligned}$$

where $m(\delta)$ converges monotonically to 0 as $\delta_1 \to \infty$. Proceeding as above gives

$$\liminf_{\delta_1 \to \infty} S^1_\delta(X) \geq \int_{[0,1]^{d-1}} \Pi_{i=2}^d (1-u_j) Q_1(0+, u_2, ..., u_d) du_2 ... du_d$$

which implies the desired conclusion. □

Appendix B

UNOBSERVED GROUPED HETEROSKEDASTICITY AND FIXED EFFECTS

B.1 Additional Results

DGP# 1		Missclassified		St. Error	
	G	WGFE	GFE	WGFE	GFE
	2	10.38%	10.28%	3.95%	3.90%
	3	6.55%	6.56%	3.09%	3.21%
	4	4.74%	4.69%	2.99%	2.99%
	5	4.13%	4.19%	3.17%	3.28%
DGP# 2					
	2	3.71%	18.51%	2.70%	6.38%
	3	4.85%	16.93%	2.93%	4.84%
	4	8.86%	20.41%	3.83%	3.89%
	5	7.78%	13%	4%	4.30%

Table B.1: Average rate of misclassification across 1,000 simulations.

WGFE	High	Low
P_g	32	58
σ_g	0.28	0.29

GFE	High	Low
P_g	49	41
σ_g	0.21	0.20

(a) $G = 2$

WGFE	High	Transition	Low
P_g	29	39	22
σ_g	0.22	0.31	0.24

GFE	High	Transition	Low
P_g	38	24	28
σ_g	0.26	0.29	0.24

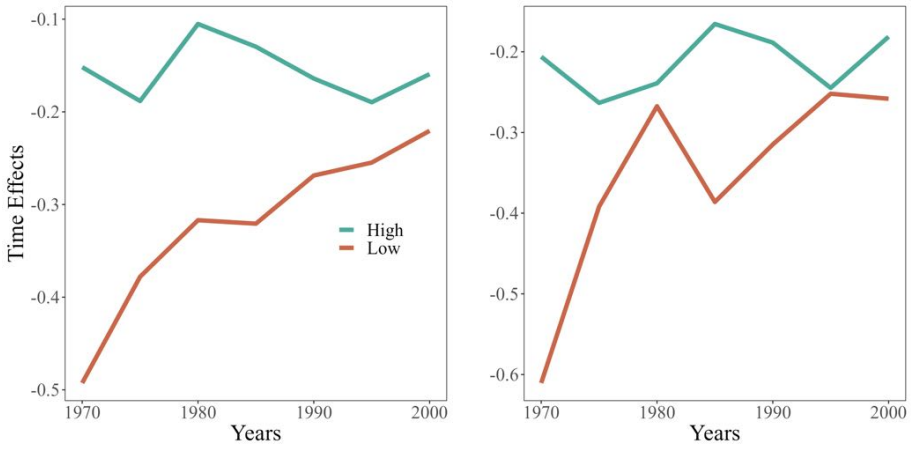
(b) $G = 3$

WGFE	High	Early	Middle	Late	Low
P_g	27	18	12	11	22
σ_g	0.19	0.27	0.27	0.20	0.20

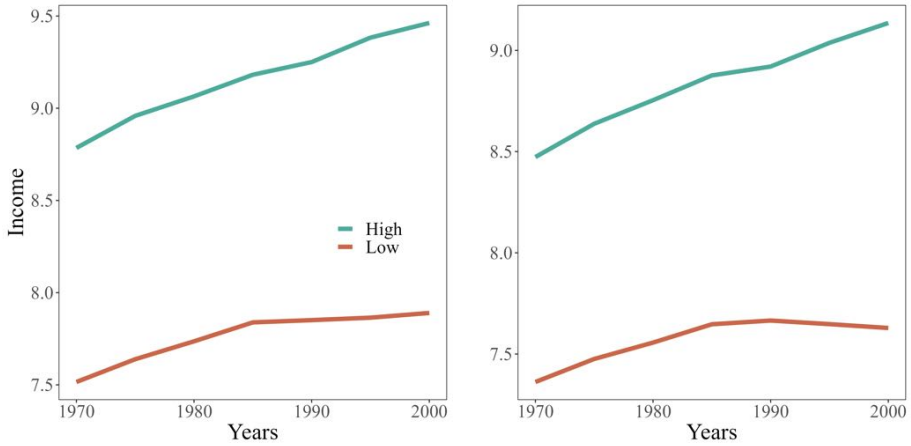
GFE	High	Early	Middle	Late	Low
P_g	30	12	14	13	21
σ_g	0.09	0.21	0.19	0.18	0.14

(c) $G = 5$

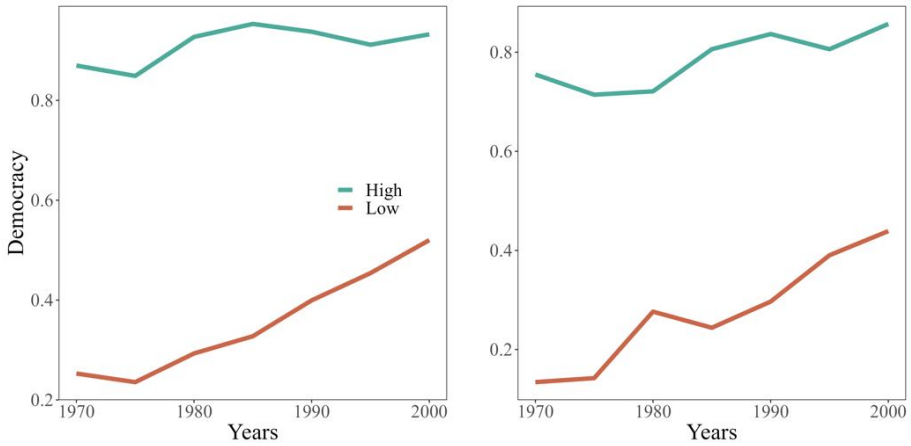
Figure B.1: WGFE and GFE estimates of group sizes and variances for $G = 2, 3, 5$.



(a) $G = 2$. Time effects α estimated via Left: WGFE, Right: GFE

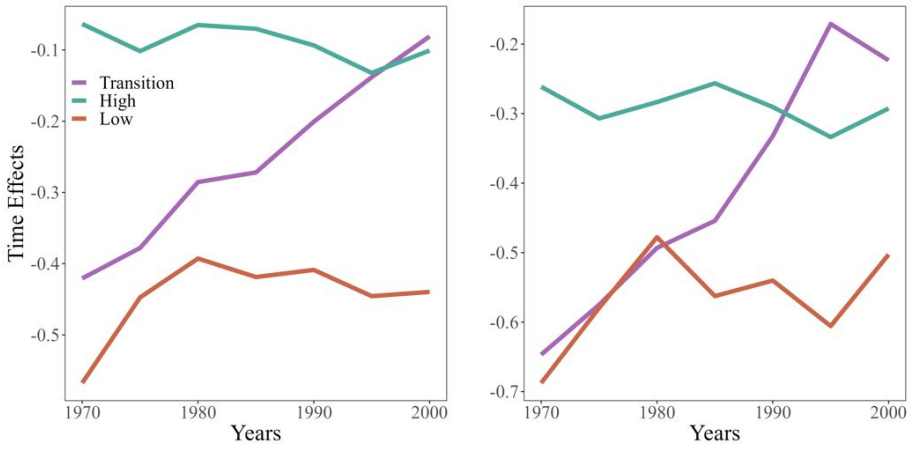


(b) $G = 2$. Average income within groups estimated via Left: WGFE, Right: GFE

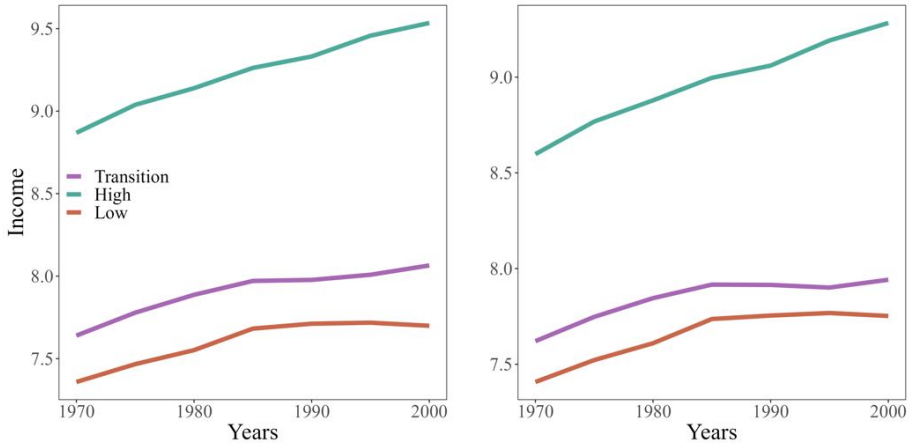


(c) $G = 2$. Average democracy index within groups estimated via Left: WGFE, Right: GFE

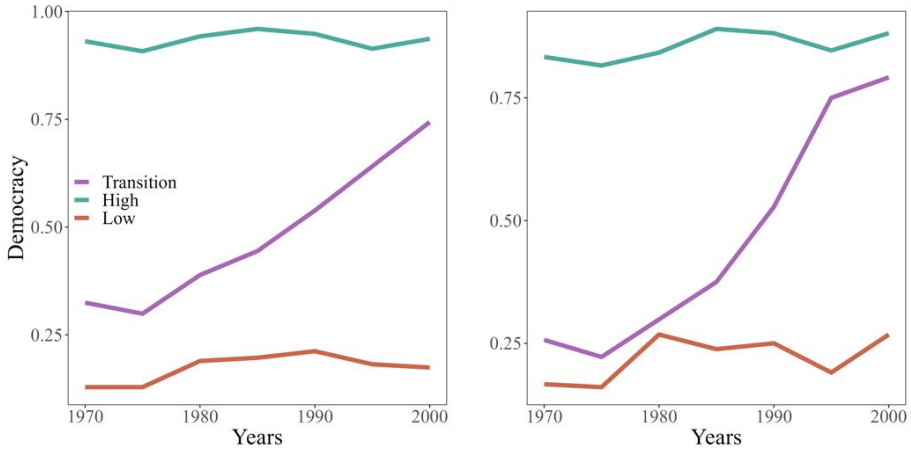
Figure B.2



(a) $G = 3$. Time effects α estimated via Left: WGFE, Right: GFE

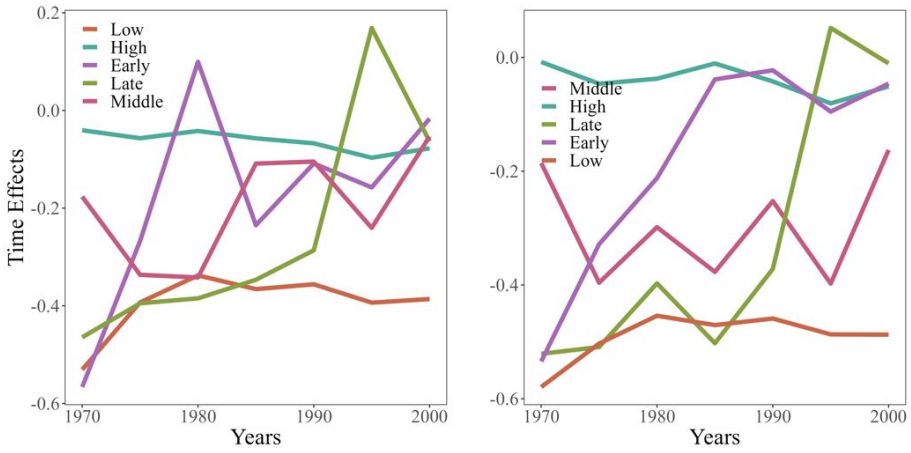


(b) $G = 3$. Average income within groups estimated via Left: WGFE, Right: GFE

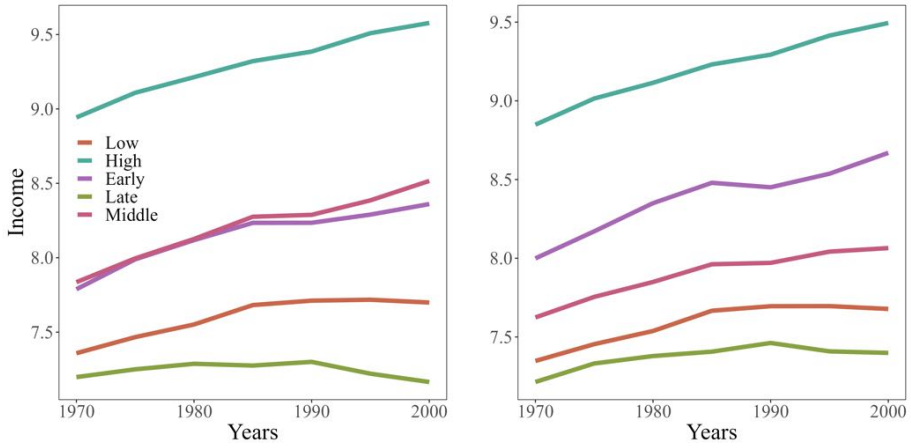


(c) $G = 3$. Average democracy index within groups estimated via Left: WGFE, Right: GFE

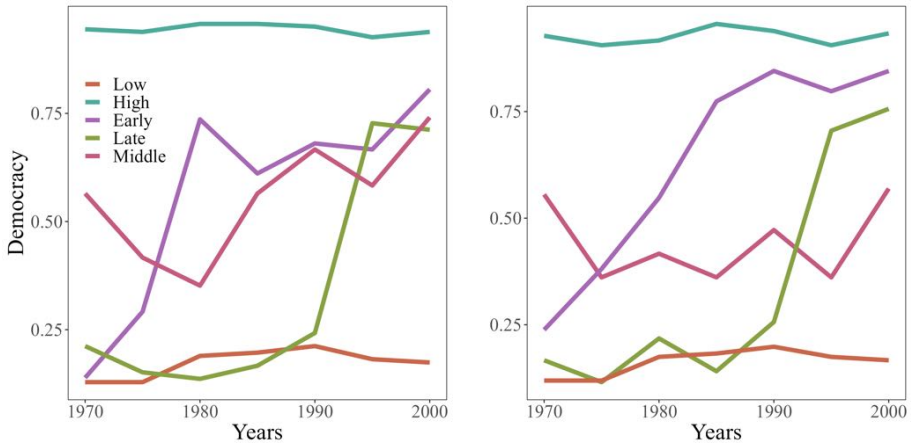
Figure B.3



(a) $G = 5$. Time effects α estimated via Left: WGFE, Right: GFE



(b) $G = 5$. Average income within groups estimated via Left: WGFE, Right: GFE



(c) $G = 5$. Average democracy index within groups estimated via Left: WGFE, Right: GFE

Figure B.4

High	Early	Late	Low
Australia	Argentina	Benin	Algeria
Austria	Bolivia	Central African Republic	Burundi
Belgium	Brazil	Madagascar	Cameroon
Canada	Burkina Faso	Malawi	Chad
Colombia	Chile	Mali	China
Costa Rica	Cyprus	Niger	Congo, Dem. Rep.
Denmark	Dominican Republic	Panama	Congo, Rep.
Finland	Ecuador	Philippines	Cote d'Ivoire
France	El Salvador	Romania	Egypt, Arab Rep.
Iceland	Ghana	South Africa	Gabon
India	Greece	Tanzania	Guinea
Ireland	Guatemala	Zambia	Iran
Israel	Honduras		Jordan
Italy	Indonesia		Kenya
Jamaica	Korea, Rep.		Mauritania
Japan	Malaysia		Morocco
Luxembourg	Mexico		Rwanda
Netherlands	Nepal		Singapore
New Zealand	Nicaragua		Syrian Arab Rep.
Norway	Nigeria		Togo
Sri Lanka	Paraguay		Tunisia
Sweden	Peru		Uganda
Switzerland	Portugal		
Trinidad and Tobago	Sierra Leone		
United Kingdom	Spain		
United State	Taiwan		
Venezuela	Thailand		
	Turkey		
	Uruguay		

Table B.2: List of countries in each group determined by WGFE assignment ($G = 4$).

High	Early	Late	Low
Australia	Argentina	Benin	Algeria
Austria	Bolivia	Burkina Faso	Burundi
Belgium	Brazil	Central African Republic	Cameroon
Canada	Ecuador	Chile	Chad
Colombia	Greece	Ghana	China
Costa Rica	Honduras	Madagascar	Congo, Dem. Rep.
Cyprus	Korea, Rep.	Malawi	Congo, Rep.
Denmark	Nepal	Mali	Cote d'Ivoire
Dominican Republic	Peru	Mexico	Egypt, Arab Rep.
El Salvador	Portugal	Nicaragua	Gabon
Finland	Spain	Niger	Guinea
France	Thailand	Panama	Indonesia
Guatemala	Uruguay	Philippines	Iran
Iceland		Romania	Jordan
India		South Africa	Kenya
Ireland		Taiwan	Mauritania
Israel		Tanzania	Morocco
Italy		Zambia	Nigeria
Jamaica			Paraguay
Japan			Rwanda
Luxembourg			Sierra Leone
Malaysia			Singapore
Netherlands			Syrian Arab Rep.
New Zealand			Togo
Norway			Tunisia
Sri Lanka			Uganda
Sweden			
Switzerland			
Trinidad and Tobago			
Turkey			
United Kingdom			
United States			
Venezuela			

Table B.3: List of countries in each group determined by GFE assignment ($G = 4$).

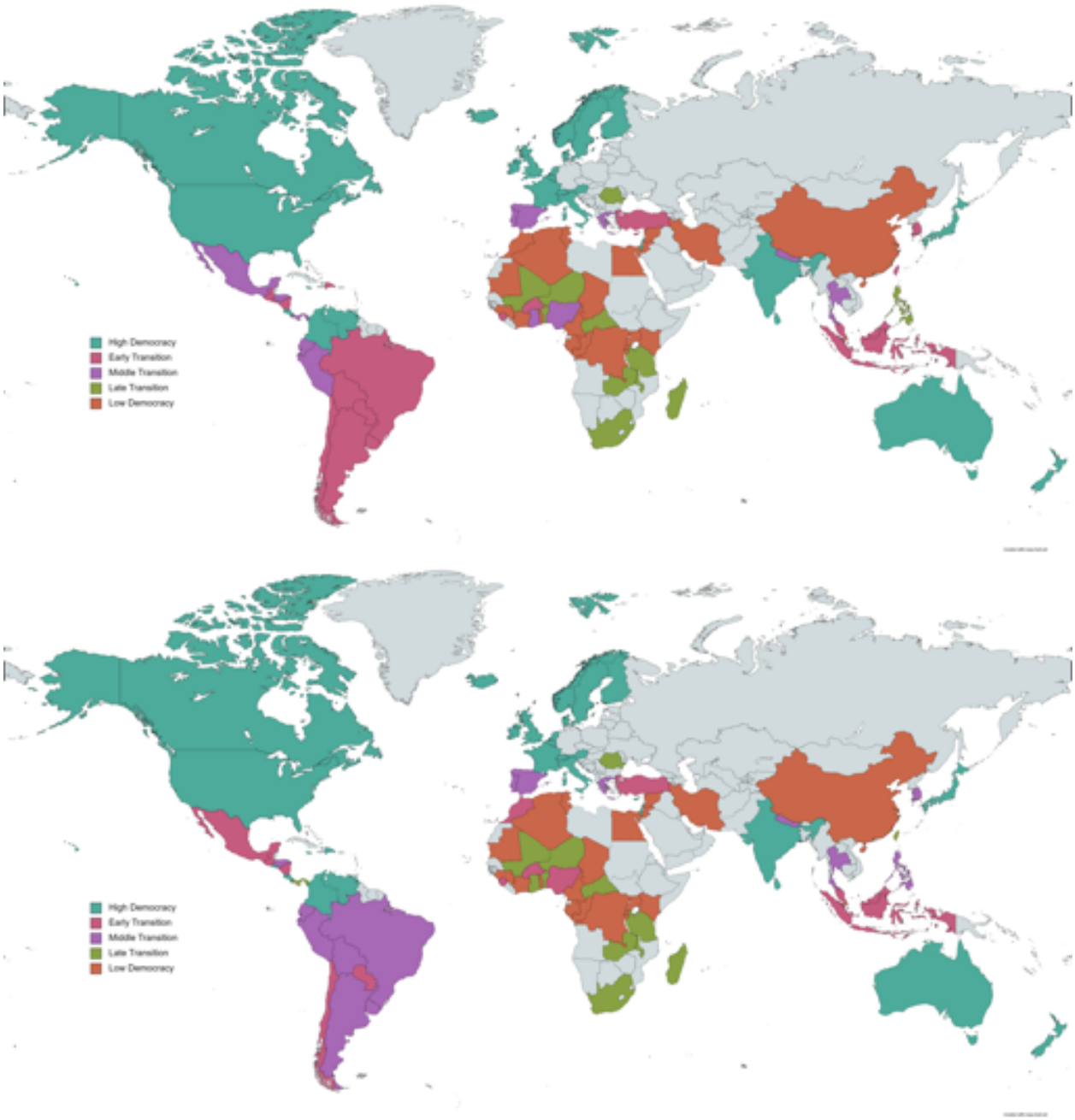


Figure B.5: $G = 5$ map with Top: WGFE groups, and Bottom: GFE groups

B.2 Variance Estimation

For all $g = 1, \ldots, G$, a White estimator for the group g variance is

$$\widehat{\sigma}_g^2 = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} \mathbb{1}\{\widehat{g}_i = g\}\widehat{u}_{it}^2}{T\sum_{j=1}^{N} \mathbb{1}\{\widehat{g}_i = g\}} \tag{B.1}$$

where $\widehat{u}_{it} = y_{it} - x_{it}'\widehat{\theta} - \widehat{\alpha}_{\widehat{g}_i t}$ are the WGFE residuals and, for any $t = 1, \ldots, T$, the White estimator of the variance of the group-specific time effects is

$$\widehat{\text{Var}}(\widehat{\alpha}_{gt}) = \frac{\sum_{i=1}^{N} \mathbb{1}\{\widehat{g}_i = g\}\widehat{u}_{it}^2}{\left(\sum_{j=1}^{N} \mathbb{1}\{\widehat{g}_i = g\}\right)^2}. \tag{B.2}$$

Theorem 8 suggests an estimator for the variance of $\widehat{\theta}$ as

$$\widehat{\text{Var}}\left(\widehat{\theta}\right) = \widehat{B}_\theta^{-1}\widehat{V}_\theta\widehat{B}_\theta^{-1}/NT \tag{B.3}$$

where, once again denoting $\overline{x}_{\widehat{g}_i t}$ as the average of x_{jt}, $j = 1 \ldots, N$ in group $\widehat{g}_i$ and denoting $\widehat{\sigma}_{\widehat{g}_i}$ as in (B.1):

$$\widehat{B}_\theta^{-1} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \widehat{\sigma}_{\widehat{g}_i}^{-1}(x_{it} - \overline{x}_{\widehat{g}_i t})(x_{it} - \overline{x}_{\widehat{g}_i t})' \tag{B.4}$$

The estimator $\widehat{V}_\theta$ of V_θ can be based on the estimator of Arellano [1987] clustered at the individual level:

$$\widehat{V}_\theta = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T} \widehat{\sigma}_{\widehat{g}_i}^{-2}(x_{it} - \overline{x}_{\widehat{g}_i t})(x_{is} - \overline{x}_{\widehat{g}_i s})'\widehat{u}_{it}\widehat{u}_{is}, \tag{B.5}$$

with a possible bias-correction of degrees of freedom $NT - p$.

B.3 Generalized Grouped Fixed Effects Estimation

The *generalized grouped fixed-effects* (GGFE) estimator for model (3.1) is the solution to the minimization problem

$$\widehat{\delta} = (\widehat{\theta}, \widehat{\alpha}, \widehat{\gamma}) = \underset{\delta \in \Theta \times \mathcal{A}^{GT} \times \Gamma_G^N}{\operatorname{argmin}} \; tr[\mathbf{\Omega}_\delta] \tag{B.6}$$

subject to

$$\mathbf{\Omega}_\delta = \sum_{g=1}^{G} \left(\mathbf{\Omega}_\delta^{1/2} \widehat{\mathbf{\Sigma}}_g(\delta) \mathbf{\Omega}_\delta^{1/2} \right)^{1/2} P_g(\delta) \tag{B.7}$$

$$\widehat{\mathbf{\Sigma}}_g(\delta) = \sum_{i=1}^{N} \frac{\mathbb{1}\{g_i = g\}}{\sum_{j=1}^{N} \mathbb{1}\{g_j = g\}} \left(y_i - x_i'\theta - \alpha_{g_i}\right) \cdot \left(y_i - x_i'\theta - \alpha_{g_i}\right)' \tag{B.8}$$

$$P_g(\delta) = \frac{1}{N} \sum_{j=1}^{N} \mathbb{1}\{g_j = g\}. \tag{B.9}$$

where the square root of a matrix is understood as the principle square root. The objective function contains $\mathbf{\Omega}_\delta$ as the solution to a nonlinear matrix equation (B.7). To solve this equation, a fixed-point algorithm can be used and is guaranteed to converge to the unique fixed-point when the group covariances (B.8) are non-singular (Álvarez Esteban et al. [2016]). The following is the iteration scheme:

$$\mathbf{\Omega}_\delta(s+1) \leftarrow \mathbf{\Omega}_\delta^{-1/2}(s) \left(\sum_{g=1}^{G} P_g \mathbf{\Omega}_\delta^{1/2}(s) \widehat{\mathbf{\Sigma}}_g \mathbf{\Omega}_\delta^{1/2}(s) \right)^2 \mathbf{\Omega}_\delta^{-1/2}(s) \tag{B.10}$$

using an arbitrary symmetric positive definite initialization matrix $\mathbf{\Omega}_\delta^{-1/2}(0)$, e.g. the identity matrix.

B.3.1 Computation

For the form of the gradient $\partial_\gamma tr[\mathbf{\Omega}_\delta]$ with respect to γ in the probability simplex Δ^G, see the following section in Appendix B.3.2. With this, I can construct an algorithm to solve (B.6).

Algorithm 3 (Lloyd's Algorithm for GGFE)**.**

1. *Initialize $\theta^{(0)}$; set the $\alpha^{(0)}$ as G randomly chosen $v_i = y_i - x_i\theta^{(0)}$; create initial assignments $g_i^{(0)}$ by assigning v_i to the closest $\{\alpha_{g_i}^{(0)}\}$; and calculate the $\{\widehat{\boldsymbol{\Sigma}}_g^{(0)}(\delta^{(0)})\}$. Set $s = 1$.*

2. *(Assignment Step). Calculate $\partial_{P_g^i} tr[\boldsymbol{\Omega}_\delta]$ for all i and g. Assign according to*

$$g_i^{(s+1)} \leftarrow \underset{g}{\operatorname{argmin}}\, \partial_{P_g^i} tr[\boldsymbol{\Omega}_\delta]\Big|_{\delta=\delta^{(s)}}$$

   and collect them in $\gamma^{(s+1)} = (g_1^{(s+1)}, \ldots, g_N^{(s+1)})$.

3. *(Update Step). Update $\alpha^{(s+1)}$ and $\theta^{(s+1)}$ according to*

$$(\theta^{(s+1)}, \alpha^{(s+1)}) = \underset{(\theta,\alpha)}{\operatorname{argmin}}\, tr[\boldsymbol{\Omega}_{(\theta,\alpha,\gamma^{(s+1)})}],$$

   then update $\{\widehat{\boldsymbol{\Sigma}}_g^{(s+1)}(\delta^{(s+1)})\}$.

4. *If $g_i^{(s)} = g_i^{(s+1)}$ for all i, stop. Otherwise, set $s \leftarrow s + 1$ and go back to step 2.*

B.3.2 Gradient of the criterion function with respect to assignments

Calculating the parameter values and labeling matrix that solve (B.6) is done through a gradient descent algorithm. For given parameter values θ and α, the existence of the gradient with respect to class assignments P^i of the objective function for $i = 1, \ldots, N$ is shown in Appendix B and C within Yang and Tabak [2022] and is of the form:

$$\nabla_{P^i} \mathrm{tr}[\widehat{\boldsymbol{\Omega}}_\delta] = \sum_{g=1}^{G} \mathrm{vec}(\mathbf{I})' \cdot \boldsymbol{W}_g \cdot \mathrm{vec}\Big((y_i - x_i\theta - \alpha_{g_i}) \cdot (y_i - x_i\theta - \alpha_{g_i})' + \widehat{\boldsymbol{\Sigma}}_g \Big) \vec{e}_g \tag{B.11}$$

where $\vec{e}_g$ is a vector of zeros except with a one in the gth entry and the $\boldsymbol{W}_g$ matrices are

$$\begin{aligned}(\boldsymbol{\Omega}_\delta^{1/2} \otimes \boldsymbol{\Omega}_\delta^{1/2}) \times \Big[\sum_{h=1}^{G} P_h (\boldsymbol{U}_h \otimes \boldsymbol{U}_h)(\boldsymbol{D}_h^{1/2} \otimes \mathbf{I} + \mathbf{I} \otimes \boldsymbol{D}_h^{1/2})^{-1} (\boldsymbol{D}_h^{1/2} \otimes \boldsymbol{D}_h^{1/2})(\boldsymbol{U}_h' \otimes \boldsymbol{U}_h') \Big]^{-1} \\ \times \big[(\boldsymbol{U}_g \otimes \boldsymbol{U}_g)(\boldsymbol{D}_g^{1/2} \otimes \mathbf{I} + \mathbf{I} \otimes \boldsymbol{D}_g^{1/2})^{-1} (\boldsymbol{U}_g' \otimes \boldsymbol{U}_g') \big] \times (\boldsymbol{\Sigma}_\delta^{1/2} \otimes \boldsymbol{\Omega}_\delta^{1/2})\end{aligned} \tag{B.12}$$

where $\boldsymbol{I}$ is the $T \times T$ identity matrix and $(\boldsymbol{U}_g, \boldsymbol{D}_g)$ are the corresponding orthonormal and diagonal matrices found from the eigendecompositions

$$\boldsymbol{\Omega}_\delta^{1/2} \widehat{\boldsymbol{\Sigma}}_g \boldsymbol{\Omega}_\delta^{1/2} = \boldsymbol{U}_g \boldsymbol{D}_g \boldsymbol{U}_g'$$

for $g \in \Gamma_G$. Note that $\boldsymbol{D}$ and $\boldsymbol{U}$ depend on the parameter values.

B.4 Computation

B.4.1 A Variable Neighborhood Search (VNS) algorithm

Algorithm 4. *(Variable Neighborhood Search (VNS)).*

1. *Set $iter_{\max}$ and $neigh_{\max}$ to desired values (usually both 10).*
2. *Initialize $(\theta^{(0)}, \alpha^{(0)})$ and perform one assignment step to get an initial g_{init}. Set $g^* = g_{init}$.*
3. *Set $n = 1$.*
4. *(Neighborhood jump). Relocate n random individuals to n other random groups and obtain a new grouping g' and perform one update step to obtain (θ', α').*
5. *Set $(\theta^{(0)}, \alpha^{(0)}) = (\theta', \alpha')$ and apply Algorithm 1 to get a grouping g.*
6. *(Local search). With grouping g from step 5, systematically check all re-assignments of individuals $i \in \{1, \ldots, N\}$ to groups $g \in \{1, \ldots, G\}$, updating g_i whenever the objective function decreases. Denote the resulting grouping by g''.*
7. *If the objective function with g'' improves relative to the one with g', then set $g^* = g''$ and go to step 3, otherwise set $n = n + 1$ and go to step 8.*
8. *If $n \leq neigh_{\max}$, then go to step 4; otherwise go to step 9.*
9. *Set $j \leftarrow j + 1$. If $j > iter_{\max}$, then stop; otherwise go to step 3.*

B.4.2 Initialization

Initialization of Algorithm 1, 3 and 4 is an important consideration in the search for global minima. Using pooled OLS, two-way fixed effects estimates or a convex combination of them for θ^0 seems to perform well, although no formal test of performance has been done and it hasn't been proposed in the literature. The justification is that the estimates should be nearby to the WGFE estimate. Other approaches are to randomly draw these parameters from some distribution or to perform WGFE estimation on a subsample to collect estimates that are then used in the full sample estimation. One could also obtain GFE estimates with the full sample and use those as initial parameters and groupings.

B.5 Proofs of the main results

Proof of Proposition 8. By the asymptotic equivalence of $\widehat{\theta}$ and the infeasible $\widetilde{\theta}$ given by Theorem 4, and the asymptotic normality of $\widetilde{\theta}$:

$$\sqrt{NT}\left(\widehat{\theta}-\theta^0\right)=\sqrt{NT}\left(\widetilde{\theta}-\theta^0\right)+\sqrt{NT}\left(\widehat{\theta}-\widetilde{\theta}\right)=\sqrt{NT}\left(\widetilde{\theta}-\theta^0\right)+o_p\left(\frac{\sqrt{N}}{T^\nu}\right).$$

where I have chosen δ such that $1/2-\delta=\nu$. Similarly for $\widehat{\alpha}_{gt}$ and $\widetilde{\alpha}_{gt}$, for any g and t.

□

B.5.1 Proof of Theorem 1

To show consistency of the infeasible version of WGFE (3.21), I show that the hypothesis of the standard theorem for consistency of M-estimators holds: 1. uniform weak convergence of the sample criterion to the population criterion, 2. the true values uniquely minimize the population criterion. I prove this in a series of lemmas.

The following is easily seen with a uniform law of large numbers due to compactness of the parameter space and the moment restrictions of Assumption 2.

Lemma 5. *Suppose Assumption 2 holds. For all $g = 1, \ldots, G$,*

$$\sup_{(\theta,\alpha)\in\Theta\times\mathcal{A}^{GT}} \left|\widetilde{Q}_g(\theta,\alpha) - Q_g(\theta,\alpha)\right| \to_p 0$$

as $N \to \infty$.

The following shows uniform convergence of the square roots

Lemma 6. *Suppose Assumption 2 holds. For all $g = 1, \ldots, G$, as $N \to \infty$*

$$\sup_{(\theta,\alpha)\in\Theta\times\mathcal{A}^{GT}} \left|\sqrt{\widetilde{Q}_g(\theta,\alpha)} - \sqrt{Q_g(\theta,\alpha)}\right| \to_p 0.$$

Proof.

$$\begin{aligned}\left|\sqrt{\widetilde{Q}_g(\theta,\alpha)} - \sqrt{Q_g(\theta,\alpha)}\right|^2 &\leq \left|\sqrt{\widetilde{Q}_g(\theta,\alpha)} - \sqrt{Q_g(\theta,\alpha)}\right|\left|\sqrt{\widetilde{Q}_g(\theta,\alpha)} + \sqrt{Q_g(\theta,\alpha)}\right| \\ &= \left|\widetilde{Q}_g(\theta,\alpha) - Q_g(\theta,\alpha)\right|\end{aligned}$$

Then

$$\left|\sqrt{\widetilde{Q}_g(\theta,\alpha)} - \sqrt{Q_g(\theta,\alpha)}\right| \leq \sqrt{\left|\widetilde{Q}_g(\theta,\alpha) - Q_g(\theta,\alpha)\right|}$$

so

$$\begin{aligned}\sup_{(\theta,\alpha)\in\Theta\times\mathcal{A}^{GT}} \left|\sqrt{\widetilde{Q}_g(\theta,\alpha)} - \sqrt{Q_g(\theta,\alpha)}\right| &\leq \sup_{(\theta,\alpha)\in\Theta\times\mathcal{A}^{GT}} \sqrt{\left|\widetilde{Q}_g(\theta,\alpha) - Q_g(\theta,\alpha)\right|} \\ &= \sqrt{\sup_{(\theta,\alpha)\in\Theta\times\mathcal{A}^{GT}} \left|\widetilde{Q}_g(\theta,\alpha) - Q_g(\theta,\alpha)\right|}\end{aligned}$$

Taking the probability limit as $N \to \infty$ and applying the continuous mapping theorem with Lemma 5 and we are done. □

The following shows uniform convergence of the sample criterion function to its population counterpart.

Lemma 7. *Suppose Assumption 1 holds. Then, as $N \to \infty$,*

$$\sup_{(\theta,\alpha)\in\Theta\times\mathcal{A}^{GT}} \left|\widetilde{Q}(\theta,\alpha) - Q(\theta,\alpha)\right| \to_p 0.$$

Proof. First, note that by the weak law of large numbers I have $P_g = P_g^0 + o_p(1)$. Then

$$\begin{aligned}\sup_{(\theta,\alpha)\in\Theta\times\mathcal{A}^{GT}} |\widetilde{Q}(\theta,\alpha) - Q(\theta,\alpha)| &= \sup_{(\theta,\alpha)\in\Theta\times\mathcal{A}^{GT}} \left| \sum_{g=1}^{G} P_g^0 \left(\sqrt{\widetilde{Q}_g(\theta,\alpha)} - \sqrt{Q_g(\theta,\alpha)} \right) + o_p(1) \right| \\ &\leq \sum_{g=1}^{G} P_g^0 \sup_{(\theta,\alpha)\in\Theta\times\mathcal{A}^{GT}} \left| \sqrt{\widetilde{Q}_g(\theta,\alpha)} - \sqrt{Q_g(\theta,\alpha)} \right| + o_p(1)\end{aligned}$$

and the result follows by Lemma 6.

□

Lemma 8. *Suppose Assumption 2 holds. Then the true values (θ^0, α^0) uniquely minimize $Q(\theta, \alpha)$.*

Proof. Using the data generating process and the identity:

$$\sqrt{a} - \sqrt{b} = \frac{a-b}{\sqrt{a}+\sqrt{b}}$$

I can write

$$Q(\theta,\alpha) - Q(\theta^0,\alpha^0) \tag{B.13}$$

$$\begin{aligned}&= \sum_{g=1}^{G} P_g^0 \left(\sqrt{\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{g_i^0 = g\} \left(u_{it} + x_{it}'(\theta^0 - \theta) + \left(\alpha_{gt}^0 - \alpha_{gt} \right) \right)^2 \right]} - \sqrt{\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{g_i^0 = g\} u_{it}^2 \right]} \right) \\ &= \sum_{g=1}^{G} P_g^0 \frac{\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{g_i^0 = g\} \left(u_{it} + x_{it}'(\theta^0 - \theta) + \left(\alpha_{gt}^0 - \alpha_{gt} \right) \right)^2 \right] - \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{g_i^0 = g\} u_{it}^2 \right]}{\sqrt{\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{g_i^0 = g\} \left(u_{it} + x_{it}'(\theta^0 - \theta) + \left(\alpha_{gt}^0 - \alpha_{gt} \right) \right)^2 \right]} + \sqrt{\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{g_i^0 = g\} u_{it}^2 \right]}}.\end{aligned} \tag{B.14}$$

For any $g = 1, \ldots, G$, there is an upper bound for the denominator independent of parameters. Consider the first term and note that each term in the summand is non negative:

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{g_i^0 = g\} \left(u_{it} + x_{it}'(\theta^0 - \theta) + \left(\alpha_{gt}^0 - \alpha_{gt} \right) \right)^2 \right] \\ =&\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{g_i^0 = g\} \left(u_{it}^2 + 2u_{it} \left(x_{it}'(\theta^0 - \theta) + \left(\alpha_{gt}^0 - \alpha_{gt} \right) \right) + \left(x_{it}'(\theta^0 - \theta) + \left(\alpha_{gt}^0 - \alpha_{gt} \right) \right)^2 \right) \right].\end{aligned} \tag{B.15}$$

Since $\mathbb{E}\left[u_{it}|x_{it}, g_i^0\right] = 0$, I know

$$\begin{aligned} &\mathbb{E}\left[\mathbb{1}\{g_i^0 = g\}u_{it}\left(x_{it}'(\theta^0 - \theta) + \left(\alpha_{gt}^0 - \alpha_{gt}\right)\right)\right] \\ =&\mathbb{E}\left[\mathbb{1}\{g_i^0 = g\}\left(x_{it}'(\theta^0 - \theta) + \left(\alpha_{gt}^0 - \alpha_{gt}\right)\right)\mathbb{E}\left[u_{it}|x_{it}, g_i^0\right]\right] \\ =&0. \end{aligned} \tag{B.16}$$

Therefore, (B.15) can be written as

$$\begin{aligned} &\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\{g_i^0 = g\}u_{it}^2\right] + \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\{g_i^0 = g\}\left(x_{it}'(\theta^0 - \theta) + \left(\alpha_{gt}^0 - \alpha_{gt}\right)\right)^2\right] \\ \leq&\mathbb{E}\left[\sum_{t=1}^{T}u_{it}^2\right] + \mathbb{E}\left[\sum_{t=1}^{T}\left(x_{it}'(\theta^0 - \theta) + \left(\alpha_{gt}^0 - \alpha_{gt}\right)\right)^2\right] \\ \leq&\sum_{t=1}^{T}\mathbb{E}\left[u_{it}^2\right] + \sum_{t=1}^{T}\mathbb{E}\left[\left(x_{it}'(\theta^0 - \theta) + \left(\alpha_{gt}^0 - \alpha_{gt}\right)\right)^2\right] \\ \leq&T\sqrt{M} + \sum_{t=1}^{T}\mathbb{E}\left[\left(x_{it}'(\theta^0 - \theta) + \left(\alpha_{gt}^0 - \alpha_{gt}\right)\right)^2\right]. \end{aligned}$$

where the last inequality follows by Assumption 2(b). Set $X_{it} = (1, x_{it})$ as a row vector and $\pi_{gt} = (\alpha_{gt}^0 - \alpha_{gt}, \theta^0 - \theta)'$. Then continuing with the previous inequality

$$\begin{aligned} &T\sqrt{M} + \sum_{t=1}^{T}\mathbb{E}\left[\left(x_{it}'(\theta^0 - \theta) + \left(\alpha_{gt}^0 - \alpha_{gt}\right)\right)^2\right] \\ =&T\sqrt{M} + \sum_{t=1}^{T}\mathbb{E}\left[\left(X_{it}'\pi_{gt}\right)^2\right] \\ \leq&T\sqrt{M} + \sum_{t=1}^{T}\|X_{it}\|^2\left\|\pi_{gt}\right\| \\ \leq&T\sqrt{M} + K \end{aligned}$$

due to the Cauchy-Schwarz inequality, compactness of the parameter space and Assumption 2(b), which results in a bound of constant $K > 0$. Note that the second term of the denominators for any g is bounded since it was a step in finding the bounds of the first term. Denote the bounding constant as C^{-1} and return to the original difference

(B.14). Expanding the quadratic cancels the sum of squared error terms and the center term is zero due to (B.16) so I have

$$\begin{aligned}
Q(\theta,\alpha) - Q(\theta^0,\alpha^0) \geq & C \sum_{g=1}^{G} P_g^0 \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{g_i^0 = g\}\left(x_{it}'(\theta^0-\theta) + \left(\alpha_{gt}^0 - \alpha_{gt}\right)\right)^2\right] \\
=& C \sum_{g=1}^{G} (P_g^0)^2 \sum_{t=1}^{T} \mathbb{E}\left[\left(x_{it}'(\theta^0-\theta) + \left(\alpha_{gt}^0 - \alpha_{gt}\right)\right)^2 \middle| g_i^0 = g\right] \\
\geq & C \sum_{g=1}^{G} (P_g^0)^2 \sum_{t=1}^{T} \mathbb{E}\left[\left(x_{it}'(\theta^0-\theta) - \mathbb{E}\left[x_{it}|g_i^0=g\right]'(\theta^0-\theta)\right)^2 \middle| g_i^0 = g\right]
\end{aligned}$$

where the last inequality is equality if and only if $\left(\alpha_{gt}^0 - \alpha_{gt}\right) = \mathbb{E}\left[x_{it}|g_i^0=g\right]'(\theta^0-\theta)$ since it minimizes the mean-squared error. Then,

$$\begin{aligned}
& Q(\theta,\alpha) - Q(\theta^0,\alpha^0) \\
\geq & C \sum_{g=1}^{G} (P_g^0)^2 \sum_{t=1}^{T} \mathbb{E}\left[\left(x_{it}'(\theta^0-\theta) - \mathbb{E}\left[x_{it}|g_i^0=g\right]'(\theta^0-\theta)\right)^2 \middle| g_i^0 = g\right] \\
\geq & C \min_{g=1,\ldots,G}\{(P_g^0)^2\}(\theta^0-\theta)' \left(\sum_{g=1}^{G} \mathbb{E}\left[\sum_{t=1}^{T}\left(x_{it} - \mathbb{E}\left[x_{it}|g_i^0=g\right]\right)\left(x_{it} - \mathbb{E}\left[x_{it}|g_i^0=g\right]\right)' \middle| g_i^0 = g\right]\right) \\
& \times (\theta^0-\theta) \\
\geq & C \min_{g=1,\ldots,G}\{(P_g^0)^2\}\rho^0 \left\|\theta^0-\theta\right\| \\
\geq & 0
\end{aligned}$$

by Assumption 2(e) where $\rho^0 > 0$ is the minimum eigenvalue. Equality occurs if and only if $\theta^0 = \theta$ by definiteness of the standard Euclidean norm. Therefore, $\left(\alpha_{gt}^0 - \alpha_{gt}\right) = \mathbb{E}\left[x_{it}|g_i^0=g\right]'(\theta^0-\theta) = 0$ and we are done.

With Lemma 7 and Lemma 8, I invoke consistency of M-estimators and therefore I have $\sqrt{N}$-consistency of (3.21).

□

B.5.2 *Proof of Asymptotic Normality of the infeasible WGFE estimator*

Proof of Asymptotic Normality of $\widetilde{\theta}$. First I will show asymptotic normality of $\widetilde{\theta}$. Consider

$$\sqrt{NT}\left(\widetilde{\theta}-\theta^0\right) = \left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{\sigma}^{-1}_{g_i^0}(\widetilde{\theta},\widetilde{\alpha})(x_{it}-\overline{x}_{g_i^0 t})(x_{it}-\overline{x}_{g_i^0 t})'\right]^{-1} \times\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{\sigma}^{-1}_{g_i^0}(\widetilde{\theta},\widetilde{\alpha})(x_{it}-\overline{x}_{g_i^0 t})u_{it}$$

where the true model was substituted into the estimator definition.

First I will show that the weight terms converge to the true group variances. Indeed,

$$\begin{aligned}\widetilde{\sigma}^2_g(\widetilde{\theta},\widetilde{\alpha}) &= \frac{1}{T\sum_{j=1}^{N}\mathbb{1}\{g_j^0=g\}}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{1}\{g_i^0=g\}\left(u_{it}+x_{it}'(\theta^0-\widetilde{\theta})+(\alpha^0_{g_i^0 t}-\widetilde{\alpha}_{gt})\right)^2\\ &= \mathbb{E}\left[\mathbb{1}\{g_i^0=g\}u_{it}^2\right]+o_p(1)\\ &= \sigma_g^2+o_p(1)\end{aligned}$$

by a weak law of large numbers and since the parameter estimators are consistent.

Therefore,

$$\sqrt{NT}\left(\widetilde{\theta}-\theta^0\right) = \left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sigma^{-1}_{g_i^0}(x_{it}-\overline{x}_{g_i^0 t})(x_{it}-\overline{x}_{g_i^0 t})'\right]^{-1}\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\sigma^{-1}_{g_i^0}(x_{it}-\overline{x}_{g_i^0 t})u_{it}+o_p(1).$$

Then, the variance conditional on covariates is

$$\begin{aligned}\mathrm{Var}\left(\sqrt{NT}\left(\widetilde{\theta}-\theta^0\right)\right) &= \left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sigma^{-1}_{g_i^0}(x_{it}-\overline{x}_{g_i^0 t})(x_{it}-\overline{x}_{g_i^0 t})'\right]^{-1}\\ \times\mathrm{Var}\left(\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\sigma^{-1}_{g_i^0}(x_{it}-\overline{x}_{g_i^0 t})u_{it}\right)&\times\left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sigma^{-1}_{g_i^0}(x_{it}-\overline{x}_{g_i^0 t})(x_{it}-\overline{x}_{g_i^0 t})'\right]^{-1}+o_p(1).\end{aligned}$$

The term in the middle can be rewritten as

$$\begin{aligned}&\mathrm{Var}\left(\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\sigma^{-1}_{g_i^0}(x_{it}-\overline{x}_{g_i^0 t})u_{it}\right)\\ &= \frac{1}{NT}\mathbb{E}\left[\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\sigma^{-1}_{g_i^0}(x_{it}-\overline{x}_{g_i^0 t})u_{it}\right)\left(\sum_{j=1}^{N}\sum_{s=1}^{T}\sigma^{-1}_{g_j^0}(x_{js}-\overline{x}_{g_j^0 s})u_{js}\right)'\right]\\ &-\frac{1}{NT}\mathbb{E}\left[\sum_{i=1}^{N}\sum_{t=1}^{T}\sigma^{-1}_{g_i^0}(x_{it}-\overline{x}_{g_i^0 t})u_{it}\right]\mathbb{E}\left[\sum_{j=1}^{N}\sum_{s=1}^{T}\sigma^{-1}_{g_j^0}(x_{js}-\overline{x}_{g_j^0 s})u_{js}\right]'\end{aligned}$$

where the second term is zero since $\mathbb{E}\left[u_{it}|x_{it}\right]=0$. Then,

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\sigma_{g_i^0}^{-1}(x_{it}-\overline{x}_{g_i^0 t})u_{it}\right)\left(\sum_{j=1}^{N}\sum_{s=1}^{T}\sigma_{g_j^0}^{-1}(x_{js}-\overline{x}_{g_j^0 s})u_{js}\right)'\right] \\
&=\mathbb{E}\left[\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\sigma_{g_i^0}^{-1}(x_{it}-\overline{x}_{g_i^0 t})u_{it}\right)\left(\sum_{j=1}^{N}\sum_{s=1}^{T}\sigma_{g_j^0}^{-1}(x_{js}-\overline{x}_{g_j^0 s})'u_{js}\right)\right] \\
&=\mathbb{E}\left[\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\left[\sigma_{g_i^0}\sigma_{g_j^0}\right]^{-1}(x_{it}-\overline{x}_{g_i^0 t})(x_{js}-\overline{x}_{g_j^0 s})'u_{it}u_{js}\right] \\
&=\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\left[\sigma_{g_i^0}\sigma_{g_j^0}\right]^{-1}\mathbb{E}\left[(x_{it}-\overline{x}_{g_i^0 t})(x_{js}-\overline{x}_{g_j^0 s})'u_{it}u_{js}\right].
\end{aligned}$$

Therefore,

$$\begin{aligned}
\text{Var}\left(\sqrt{NT}\left(\widetilde{\theta}-\theta^0\right)\right)=&\left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sigma_{g_i^0}^{-1}(x_{it}-\overline{x}_{g_i^0 t})(x_{it}-\overline{x}_{g_i^0 t})'\right]^{-1} \\
&\times\frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\left[\sigma_{g_i^0}\sigma_{g_j^0}\right]^{-1}\mathbb{E}\left[(x_{it}-\overline{x}_{g_i^0 t})(x_{js}-\overline{x}_{g_j^0 s})'u_{it}u_{js}\right] \\
&\times\left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sigma_{g_i^0}^{-1}(x_{it}-\overline{x}_{g_i^0 t})(x_{it}-\overline{x}_{g_i^0 t})'\right]^{-1}+o_p(1).
\end{aligned}$$

and, by Assumption 3 (*b*),

$$\text{Var}\left(\sqrt{NT}\left(\widetilde{\theta}-\theta^0\right)\right)\longrightarrow_p B_\theta^{-1}V_\theta B_\theta^{-1}.$$

Finally, by Assumption 3 (*c*),

$$\begin{aligned}
\sqrt{NT}\left(\widetilde{\theta}-\theta^0\right)&=\left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{\sigma}_{g_i^0}^{-1}(\widetilde{\theta},\widetilde{\alpha})(x_{it}-\overline{x}_{g_i^0 t})(x_{it}-\overline{x}_{g_i^0 t})'\right]^{-1}\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{\sigma}_{g_i^0}^{-1}(\widetilde{\theta},\widetilde{\alpha})(x_{it}-\overline{x}_{g_i^0 t})u_{it} \\
&=\left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sigma_{g_i^0}^{-1}(x_{it}-\overline{x}_{g_i^0 t})(x_{it}-\overline{x}_{g_i^0 t})'\right]^{-1}\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\sigma_{g_i^0}^{-1}(x_{it}-\overline{x}_{g_i^0 t})u_{it}+o_p(1) \\
&\longrightarrow_d B_\theta^{-1}Z
\end{aligned}$$

where $Z\sim N(0,V_\theta)$.

□

Proof of Asymptotic Normality of $\widetilde{\alpha}$. Consider the following for any $g = 1, \ldots, G$ and $t \in \mathcal{T}$:

$$\sqrt{N}\left(\widetilde{\alpha}_{gt} - \alpha^0_{g^0_i}\right) = \frac{\frac{1}{\sqrt{N}}\sum_{i=1}^N \mathbb{1}\{g^0_i = g\}\left(u_{it} - x'_{it}\left(\theta^0 - \widetilde{\theta}\right)\right)}{\frac{1}{N}\sum_{j=1}^N \mathbb{1}\{g^0_j = g\}}$$

which is obtained by substituting the population model for y_{it}.

Then,

$$\sqrt{N}\left(\widetilde{\alpha}_{gt} - \alpha^0_{g^0_i}\right) = \frac{\frac{1}{\sqrt{N}}\sum_{i=1}^N \mathbb{1}\{g^0_i = g\}u_{it}}{\frac{1}{N}\sum_{j=1}^N \mathbb{1}\{g^0_j = g\}} + \left[\frac{\frac{1}{\sqrt{N}}\sum_{i=1}^N \mathbb{1}\{g^0_i = g\}x_{it}}{\frac{1}{N}\sum_{j=1}^N \mathbb{1}\{g^0_j = g\}}\right]'\left(\theta^0 - \widetilde{\theta}\right).$$

By Assumption 2 (a, c) I know that

$$\frac{\sum_{i=1}^N \mathbb{1}\{g^0_i = g\}x_{it}}{\sum_{j=1}^N \mathbb{1}\{g^0_j = g\}} = O_p(1/\sqrt{NT}).$$

Therefore, by the consistency of $\widetilde{\theta}$, and for any $g = 1, \ldots, G$ and $t = 1, \ldots, T$

$$\sqrt{N}\left(\widetilde{\alpha}_{gt} - \alpha^0_{g^0_i}\right) = \frac{\frac{1}{\sqrt{N}}\sum_{i=1}^N \mathbb{1}\{g^0_i = g\}u_{it}}{\frac{1}{N}\sum_{j=1}^N \mathbb{1}\{g^0_j = g\}} + o_p(1).$$

The conditional variance for all g and t is easily calculated as

$$\text{Var}\left(\sqrt{N}\left(\widetilde{\alpha}_{gt} - \alpha^0_{g^0_i}\right)\right) = \left[\frac{1}{N}\sum_{j=1}^N \mathbb{1}\{g^0_j = g\}\right]^{-2} \times \frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[\sum_{j=1}^N \mathbb{1}\{g^0_i = g\}\mathbb{1}\{g^0_j = g\}u_{it}u_{jt}\right]$$

hence

$$\sqrt{N}\left(\widetilde{\alpha}_{gt} - \alpha^0_{g^0_i}\right) \longrightarrow_p \left[\mathbb{P}(g^0_i = g)\right]^{-2} Z_\alpha$$

where $Z_\alpha \sim N(0, v_{gt})$

□

B.5.3 Proof of Theorem 3 (Consistency of the WGFE Estimator)

The argument for consistency of $\widehat{\theta}$ largely follows that of Bonhomme and Manresa [2015]. First, define an auxiliary criterion function

$$\begin{aligned}\widetilde{Q}(\theta,\alpha,\gamma) &= \sum_{g=1}^{G} P_g \sqrt{\frac{1}{T\sum_{j=1}^{N} P_g^j}\sum_{i=1}^{N}\sum_{t=1}^{T} P_g^i \left(x_{it}'(\theta^0-\theta)+\alpha^0_{g_i^0 t}-\alpha_{g_i t}\right)^2 + \frac{1}{T\sum_{j=1}^{N} P_g^j}\sum_{i=1}^{N}\sum_{t=1}^{T} P_g^i u_{it}^2} \\ &= \sum_{g=1}^{G} \sqrt{P_g \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} P_g^i \left(x_{it}'(\theta^0-\theta)+\alpha^0_{g_i^0 t}-\alpha_{g_i t}\right)^2 + P_g \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} P_g^i u_{it}^2}\end{aligned} \tag{B.17}$$

where I have taken $P_g(\gamma) = P_g$. Note that the minimizers of this function are the true values so I want to show that this auxiliary criterion function uniformly converges to the WGFE objective function (3.10). To do this I first prove a few lemmas.

The following lemma shows that the variance of the sample barycenter of an arbitrary grouping of error terms will converge to its population counterpart.

Lemma 9. *Suppose Assumption 4 holds. Let $\widetilde{\gamma}$ be a random variable with support $\{1,\dots,G\}$ and let $\{\widetilde{g}_i\}$ be an i.i.d. sample from $\widetilde{\gamma}$. Then*

$$\sum_{g=1}^{G} P_g(\widetilde{\gamma}) \sqrt{\frac{1}{T\sum_{j=1}^{N}\mathbb{1}\{\widetilde{g}_i = g\}}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{1}\{\widetilde{g}_i = g\}u_{it}^2} \longrightarrow_p \sum_{g=1}^{G} \mathbb{P}(\widetilde{g}_i = g)\sqrt{\mathbb{E}\left[\mathbb{1}\{\widetilde{g}_i = g\}u_{it}^2\right]}.$$

Proof. First, for $g = 1, \ldots, G$, let $\mu_g = \mathbb{E}\left[\mathbb{1}\{\widetilde{g}_i = g\} u_{it}^2\right]$. Then, by Chebyshev's inequality,

$$
\begin{aligned}
&\mathbb{P}\left(\left| \frac{1}{T \sum_{j=1}^{N} \mathbb{1}\{\widetilde{g}_i = g\}} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{1}\{\widetilde{g}_i = g\} u_{it}^2 - \mu_g \right| \geq \varepsilon \right) \\
&\leq \frac{1}{\varepsilon^2} \text{Var}\left(\frac{1}{T \sum_{j=1}^{N} \mathbb{1}\{\widetilde{g}_i = g\}} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{1}\{\widetilde{g}_i = g\} u_{it}^2 \right) \\
&\leq \frac{1}{T^2 \varepsilon^2 \left(\sum_{j=1}^{N} \mathbb{1}\{\widetilde{g}_i = g\} \right)^2} \text{Var}\left(\sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{1}\{\widetilde{g}_i = g\} u_{it}^2 \right) \\
&= \frac{1}{T^2 \varepsilon^2 \left(\sum_{j=1}^{N} \mathbb{1}\{\widetilde{g}_i = g\} \right)} \text{Var}\left(\sum_{t=1}^{T} \mathbb{1}\{\widetilde{g}_i = g\} u_{it}^2 \right), \quad \text{independent across } i \\
&\leq \frac{2}{T^2 \varepsilon^2 \left(\sum_{j=1}^{N} \mathbb{1}\{\widetilde{g}_i = g\} \right)} \sum_{t=1}^{T} \sum_{s=1}^{T} \text{Cov}\left(\mathbb{1}\{\widetilde{g}_i = g\} u_{it}^2, \mathbb{1}\{\widetilde{g}_i = g\} u_{is}^2 \right) \\
&\leq \frac{1}{T^2 \varepsilon^2 \left(\sum_{j=1}^{N} \mathbb{1}\{\widetilde{g}_i = g\} \right)} \sum_{t=1}^{T} \sum_{s=1}^{T} \sqrt{\text{Var}(\mathbb{1}\{\widetilde{g}_i = g\} u_{it}^2)} \sqrt{\text{Var}(\mathbb{1}\{\widetilde{g}_i = g\} u_{is}^2)} \\
&\leq \frac{1}{T^2 \varepsilon^2 \sum_{j=1}^{N} \mathbb{1}\{\widetilde{g}_i = g\}} \sum_{t=1}^{T} \sum_{s=1}^{T} \sqrt{\mathbb{E}\left[u_{it}^4\right]} \sqrt{\mathbb{E}\left[u_{is}^4\right]} \\
&\leq \frac{M}{\varepsilon^2 \sum_{j=1}^{N} \mathbb{1}\{\widetilde{g}_i = g\}}
\end{aligned}
$$

where the bound on the fourth moments by $M > 0$ is due to Assumption 4(b) and we see that this probability converges to zero as N approaches infinity. Therefore, since $\widetilde{\gamma}$ was an arbitrarily chosen random variable and

$$
P_g(\widetilde{\gamma}) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{\widetilde{g}_i = g\} \longrightarrow_p \mathbb{E}\left[\mathbb{1}\{\widetilde{g}_i = g\}\right] = \mathbb{P}(\widetilde{g}_i = g)
$$

by the weak law of large numbers. Applying the continuous mapping theorem finishes the proof. □

The following lemma shows that the sample barycenter variance of any arbitrary grouping of error terms has an asymptotic lower bound as the barycenter variance with the true grouping.

Lemma 10. *Suppose Assumption 4 holds. Let $\widetilde{\gamma}$ be a random variable with support $\{1,\dots,G\}$ and let $\{\widetilde{g}_i\}$ be an i.i.d. sample from $\widetilde{\gamma}$. Then there exists $C \geq 0$ such that*

$$\widehat{Q}\left(\theta^0,\alpha^0,\widetilde{\gamma}\right) - \widehat{Q}\left(\theta^0,\alpha^0,\gamma^0\right) = C + o_p(1)$$

Proof. By Lemma 9 and the Slutsky theorem, I have

$$\begin{aligned}
&\sum_{g=1}^{G} P_g(\widetilde{\gamma})\sqrt{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{1}\{\widetilde{g}_i = g\}u_{it}^2} - \sum_{g=1}^{G} P_g(\gamma^0)\sqrt{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{1}\{g_i^0 = g\}u_{it}^2} \\
&= \sum_{g=1}^{G}\mathbb{P}(\widetilde{g}_i = g)\sqrt{\mathbb{E}\left[\mathbb{1}\{\widetilde{g}_i = g\}u_{it}^2\right]} - \sum_{g=1}^{G}\mathbb{P}(g_i^0 = g)\sqrt{\mathbb{E}\left[\mathbb{1}\{g_i^0 = g\}u_{it}^2\right]} + o_p(1).
\end{aligned}$$

as N, T approach infinity.

By definition,

$$Q(\theta^0,\alpha^0,\gamma^0) = \sum_{g=1}^{G}\mathbb{P}(g_i^0 = g)\sqrt{\mathbb{E}\left[\mathbb{1}\{g_i^0 = g\}u_{it}^2\right]}$$

which is a minimized value among (θ,α,γ) so

$$\sum_{g=1}^{G}\mathbb{P}(\widetilde{g}_i = g)\sqrt{\mathbb{E}\left[\mathbb{1}\{\widetilde{g}_i = g\}u_{it}^2\right]} = Q(\theta^0,\alpha^0,\widetilde{\gamma}) \geq Q(\theta^0,\alpha^0,\gamma^0).$$

Therefore, there exists $C \geq 0$ such that

$$\sum_{g=1}^{G}\mathbb{P}(\widetilde{g}_i = g)\sqrt{\mathbb{E}\left[\mathbb{1}\{\widetilde{g}_i = g\}u_{it}^2\right]} - \sum_{g=1}^{G}\mathbb{P}(g_i^0 = g)\sqrt{\mathbb{E}\left[\mathbb{1}\{g_i^0 = g\}u_{it}^2\right]} = C + o_p(1).$$

□

Lemma 11. *Suppose Assumption 4 holds. Then*

$$\sup_{(\theta,\alpha,\gamma)\in\Theta\times\mathcal{A}^{GT}\times\Gamma_G^N}\left|\widehat{Q}(\theta,\alpha,\gamma) - \widetilde{Q}(\theta,\alpha,\gamma)\right| \longrightarrow_p 0 \tag{B.18}$$

as N, T approach infinity.

Proof. Let $(\theta, \alpha, \gamma) \in \Theta \times \mathcal{A}^{GT} \times \Gamma_G^N$. Then

$$\widehat{Q}(\theta,\alpha,\gamma) - \widetilde{Q}(\theta,\alpha,\gamma)$$

$$= \sum_{g=1}^{G} \sqrt{P_g \frac{1}{NT} \sum_{i=1}^{N}\sum_{t=1}^{T} P_g^i \left(x_{it}'(\theta^0 - \theta) + \alpha^0_{g_i^0 t} - \alpha_{g_i t} \right)^2 + P_g^i u_{it}^2 + 2P_g^i u_{it} \left(x_{it}'(\theta^0 - \theta) + \alpha^0_{g_i^0 t} - \alpha_{g_i t} \right)} \tag{B.19}$$

$$- \sum_{g=1}^{G} \sqrt{P_g \frac{1}{NT} \sum_{i=1}^{N}\sum_{t=1}^{T} P_g^i \left(x_{it}'(\theta^0 - \theta) + \alpha^0_{g_i^0 t} - \alpha_{g_i t} \right)^2 + P_g^i u_{it}^2}$$

For $a, b > 0$ I have the identity

$$\sqrt{a} - \sqrt{b} = \frac{\sqrt{a} - \sqrt{b}}{1} \left(\frac{\sqrt{a} + \sqrt{b}}{\sqrt{a} + \sqrt{b}} \right) = \frac{a - b}{\sqrt{a} + \sqrt{b}} \tag{B.20}$$

which can be applied to (B.19) as

$$\sum_{g=1}^{G} \varphi_g(\theta,\alpha,\gamma) \left(P_g \frac{2}{NT} \sum_{i=1}^{N}\sum_{t=1}^{T} P_g^i u_{it} \left(x_{it}'(\theta^0 - \theta) + \alpha^0_{g_i^0 t} - \alpha_{g_i t} \right) \right) \tag{B.21}$$

where

$$\varphi_g(\theta,\alpha,\gamma) = \left(\sqrt{\widehat{Q}_g(\theta,\alpha,\gamma)} + \sqrt{\widetilde{Q}_g(\theta,\alpha,\gamma)} \right)^{-1} \tag{B.22}$$

with the $\widehat{Q}_g$ and $\widetilde{Q}_g$ denoting the terms within the weighted sum of square roots of $\widehat{Q}$ and $\widetilde{Q}$, respectively.

To show that the difference between the sample criterion function and auxiliary criterion is $o_p(1)$, we need to show the following for all $g = 1, \ldots, G$:

i. $\frac{2}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} P_g^i u_{it} \left(x_{it}'(\theta^0 - \theta) + \alpha^0_{g_i^0 t} - \alpha_{g_i t} \right) = o_p(1)$; and

ii. $\varphi_g(\theta,\alpha,\gamma) = O_p(1)$.

Proof of i. Expanding the sum reveals

$$\left(\frac{2}{NT} \sum_{i=1}^{N}\sum_{t=1}^{T} P_g^i u_{it} x_{it} \right)' (\theta^0 - \theta) + \frac{2}{NT} \sum_{i=1}^{N}\sum_{t=1}^{T} P_g^i \alpha^0_{g_i^0 t} u_{it} - \frac{2}{NT} \sum_{i=1}^{N}\sum_{t=1}^{T} P_g^i \alpha_{g_i t} u_{it}$$

which I will show each term is $o_p(1)$.

Define $\omega_g^i = P_g^i/N \leq 1/N$ so $\sum_i \omega_g^i = P_g$. Then for the first term:

$$\mathbb{E}\left[\left\|\sum_{i=1}^{N}\omega_g^i\frac{1}{T}\sum_{t=1}^{T}u_{it}x_{it}\right\|^2\right] \leq \mathbb{E}\left[\left(\sum_{i=1}^{N}\omega_g^i\left\|\frac{1}{T}\sum_{t=1}^{T}u_{it}x_{it}\right\|\right)^2\right] \tag{B.23}$$

$$\leq \mathbb{E}\left[\left(\sum_{i=1}^{N}\frac{1}{N}\left\|\frac{1}{T}\sum_{t=1}^{T}u_{it}x_{it}\right\|\right)^2\right] \tag{B.24}$$

$$\leq \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{1}{T}\sum_{t=1}^{T}u_{it}x_{it}\right\|^2\right] \tag{B.25}$$

$$= \mathbb{E}\left[\frac{1}{NT^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}u_{it}u_{is}x_{it}'x_{is}\right] \tag{B.26}$$

$$= \mathbb{E}\left[\frac{1}{NT^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}u_{it}u_{is}x_{it}'x_{is}\right] \tag{B.27}$$

$$\leq \frac{M}{T} \tag{B.28}$$

where the inequalities are due to the triangle inequality and then Jensen's inequality and Assumption 4 (c). Hence, the first term is $o_p(1)$ due to this inequality and Assumption 4 (a) that the parameter space is a compact subset of $\mathbb{R}^p$ so $\left\|\theta^0 - \theta\right\|$ is bounded for any θ.

For the last two terms, it is enough to show the third is $o_p(1)$. For every $g = 1, \ldots, G$,

$$\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}P_g^i\alpha_{g_it}u_{it}\right)^2 = \left(\frac{1}{T}\sum_{t=1}^{T}\alpha_{gt}\left(\frac{1}{N}\sum_{i=1}^{N}P_g^iu_{it}\right)\right)^2 \leq \left(\frac{1}{T}\sum_{t=1}^{T}\alpha_{gt}^2\right)\left(\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{N}\sum_{i=1}^{N}P_g^iu_{it}\right)^2\right) \tag{B.29}$$

where the left term is uniformly bounded because of compactness of the parameter space Assumption 4 (a). Then,

$$\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{N}\sum_{i=1}^{N}P_g^iu_{it}\right)^2 = \frac{1}{TN^2}\sum_{i=1}^{N}\sum_{j=1}^{N}P_g^iP_g^j\sum_{t=1}^{T}u_{it}u_{jt} \leq \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\frac{1}{T}\sum_{t=1}^{T}u_{it}u_{jt}\right|$$

$$\leq \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[u_{it}u_{jt}\right]\right| + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\frac{1}{T}\sum_{t=1}^{T}u_{it}u_{jt} - \mathbb{E}\left[u_{it}u_{jt}\right]\right|.$$

By Assumption 4 (*d*), $\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[u_{it}u_{jt}\right]\right| \leq M/N$. Also, by the Cauchy-Schwarz inequality,

$$\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\frac{1}{T}\sum_{t=1}^{T}u_{it}u_{jt}-\mathbb{E}\left[u_{it}u_{jt}\right]\right|\right)^2 \leq \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}u_{it}u_{jt}-\mathbb{E}\left[u_{it}u_{jt}\right]\right)^2$$

which is bounded in expectation by Assumption 4 (*e*).

Therefore, point one is $o_p(1)$ and it remains to show that point two is $O_p(1)$. However, by compactness of the parameter space and Lemma 9 it is easily seen that it is $O_p(1)$.

Therefore, since g was arbitrarily chosen, I have that

$$\widehat{Q}(\theta,\alpha,\gamma)-\widetilde{Q}(\theta,\alpha,\gamma)=o_p(1).$$

□

The following lemma shows that the true values are unique minimizers of the auxiliary criterion function in the probability limit.

Lemma 12. *Suppose Assumption 4 holds. There exists $K > 0$ such that for all $(\theta,\alpha,\gamma) \in \Theta \times \mathcal{A}^{GT} \times \Gamma_G^N$,*

$$\widetilde{Q}(\theta,\alpha,\gamma)-\widetilde{Q}(\theta^0,\alpha^0,\gamma^0) \geq K\left\|\theta^0-\theta\right\|^2+o_p(1)$$

Proof. For shorthand, take $P_g^0 = P_g(\gamma^0)$. I know $P_g \neq 0$ for some $g = 1,\ldots,G$, so denote

$$\begin{aligned}
&\left[\widetilde{\varphi}_g(\theta,\alpha,\gamma)\right]^{-1}\\
&=\sqrt{P_g\frac{1}{TN}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{1}\{g_i=g\}\left(x_{it}'(\theta^0-\theta)+\alpha^0_{g_i^0 t}-\alpha_{g_i t}\right)^2+P_g\frac{1}{TN}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{1}\{g_i=g\}u_{it}^2}\\
&+\sqrt{P_g\frac{1}{TN}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{1}\{g_i=g\}u_{it}^2}.
\end{aligned}$$

Adding by zero to

$$\widetilde{Q}(\theta,\alpha,\gamma)-\widetilde{Q}(\theta^0,\alpha^0,\gamma^0)$$
$$=\sum_{g=1}^{G}\sqrt{P_g\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{1}\{g_i=g\}\left(x_{it}'(\theta^0-\theta)+\alpha^0_{g_i^0t}-\alpha_{g_it}\right)^2+P_g\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{1}\{g_i=g\}u_{it}^2}$$
$$\underbrace{-\sum_{g=1}^{G}\sqrt{P_g\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{1}\{g_i=g\}u_{it}^2}+\sum_{g=1}^{G}\sqrt{P_g\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{1}\{g_i=g\}u_{it}^2}}_{=0}$$
$$-\sum_{g=1}^{G}\sqrt{P_g^0\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{1}\{g_i^0=g\}u_{it}^2}$$

then applying the identity (B.20) and Lemma 10 gives

$$\widetilde{Q}(\theta,\alpha,\gamma)-\widetilde{Q}(\theta^0,\alpha^0,\gamma^0)$$
$$=\sum_{g=1}^{G}\sqrt{P_g\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}P_g^i\left(x_{it}'(\theta^0-\theta)+\alpha^0_{g_i^0t}-\alpha_{g_it}\right)^2+P_g\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}P_g^iu_{it}^2}-\sqrt{P_g\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}P_g^iu_{it}^2}$$
$$+C+o_p(1)$$
$$\geq\sum_{g=1}^{G}\widetilde{\varphi}_g(\theta,\alpha,\gamma)P_g\left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}P_g^i\left(x_{it}'(\theta^0-\theta)+\alpha^0_{g_i^0t}-\alpha_{g_it}\right)^2\right]+o_p(1)$$

where this inequality is over those $g=1,\ldots,G$ such that $\mathbb{P}(g_i=g)>0$, in other words

those groups that are non empty and also

$$
\begin{aligned}
\left[\widetilde{\varphi}_g(\theta,\alpha,\gamma)\right]^{-1} &= \sqrt{P_g \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} P_g^i \left(x_{it}'(\theta^0-\theta)+\alpha^0_{g_i^0 t}-\alpha_{g_i t}\right)^2 + P_g \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} P_g^i u_{it}^2} \\
&+ \sqrt{P_g \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} P_g^i u_{it}^2} \\
&\leq \sqrt{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \left(x_{it}'(\theta^0-\theta)+\alpha^0_{g_i^0 t}-\alpha_{g_i t}\right)^2} + 2\sqrt{\frac{1}{TN}\sum_{i=1}^{N}\sum_{t=1}^{T} u_{it}^2} \\
&\leq \sqrt{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \|x_{it}\|^2 \|\theta^0-\theta\|^2} + \sqrt{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \|x_{it}\| \|\theta^0-\theta\| \left|\alpha^0_{g_i^0 t}-\alpha_{g_i t}\right|} \\
&+ \sqrt{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \left(\alpha^0_{g_i^0 t}-\alpha_{g_i t}\right)^2} + 2\sqrt{\frac{1}{TN}\sum_{i=1}^{N}\sum_{t=1}^{T} u_{it}^2}
\end{aligned}
$$

where all the terms can be bounded due to Assumption 4 (*a*,*b*) and Assumption 1 (*c*) so there exists a constant J for any g such that $\widetilde{\varphi}_g(\theta,\alpha,\gamma) > J$.

Then, since the (within-group) mean minimizes the sum of these squared deviations and I have assumed groups are non empty, the first term is bounded below as

$$
\begin{aligned}
&\widetilde{Q}(\theta,\alpha,\gamma) - \widetilde{Q}(\theta^0,\alpha^0,\gamma^0) \\
\geq & J \min_{g=1,\ldots,G}\{P_g\} \sum_{g=1}^{G} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} P_g^i \left(x_{it}'(\theta^0-\theta)+\alpha^0_{g_i^0 t}-\alpha_{g_i t}\right)^2 + o_p(1) \\
= & J \min_{g=1,\ldots,G}\{P_g\} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \left(x_{it}'(\theta^0-\theta)+\alpha^0_{g_i^0 t}-\alpha_{g_i t}\right)^2 + o_p(1) \\
\geq & J \min_{g=1,\ldots,G}\{P_g\} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \left(x_{it}'(\theta^0-\theta)-\overline{x}'_{g_i \wedge g_i^0 t}(\theta^0-\theta)\right)^2 + o_p(1) \\
\geq & J \min_{g=1,\ldots,G}\{P_g\} \min_{\gamma\in\Gamma}(\theta^0-\theta)' \left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \left(x_{it}-\overline{x}_{g_i \wedge g_i^0 t}\right)\left(x_{it}-\overline{x}_{g_i \wedge g_i^0 t}\right)'\right)(\theta^0-\theta) + o_p(1) \\
\geq & J \min_{\gamma\in\Gamma}\{\widehat{\rho}(\gamma)\} \left\|\theta^0-\theta\right\|^2 + o_p(1) \\
= & K \left\|\theta^0-\theta\right\|^2 + o_p(1)
\end{aligned}
$$

where the last equality is due to the convergence of probability of $\min_\gamma\{\widehat{\rho}(\gamma)\}$ to a nonzero constant by Assumption 4(f). □

Proof of consistency of the WGFE estimator. I now show consistency of the WGFE estimator $\widehat{\theta}$ of θ^0. By Lemma 11 and the definition of the WGFE estimator, I have

$$\widetilde{Q}(\widehat{\theta},\widehat{\alpha},\widehat{\gamma}) = \widehat{Q}(\widehat{\theta},\widehat{\alpha},\widehat{\gamma}) + o_p(1) \leq \widehat{Q}(\theta^0,\alpha^0,\gamma^0) + o_p(1) = \widetilde{Q}(\theta^0,\alpha^0,\gamma^0) + o_p(1). \quad \text{(B.30)}$$

Hence,

$$\widetilde{Q}(\widehat{\theta},\widehat{\alpha},\widehat{\gamma}) - \widetilde{Q}(\theta^0,\alpha^0,\gamma^0) \leq o_p(1).$$

Then, by Lemma 12,

$$o_p(1) \leq C\left\|\theta^0 - \widehat{\theta}\right\|^2 + o_p(1) \leq \widetilde{Q}(\widehat{\theta},\widehat{\alpha},\widehat{\gamma}) - \widetilde{Q}(\theta^0,\alpha^0,\gamma^0) \leq o_p(1).$$

Hence,

$$\left\|\widehat{\theta} - \theta^0\right\|^2 = o_p(1).$$

Next, I show the second part of the proposition:

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\widehat{\alpha}_{\widehat{g}_i t} - \alpha^0_{g_i^0 t}\right)^2 \longrightarrow_p 0$$

as $N, T \to \infty$.

Start with

$$\begin{aligned}
&\left|\widetilde{Q}(\widehat{\theta},\widehat{\alpha},\widehat{\gamma}) - \widetilde{Q}(\theta^0,\widehat{\alpha},\widehat{\gamma})\right| \\
&= \Bigg|\sum_{g=1}^{G}\sqrt{P_g\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}P_g^i\left(\left(x_{it}'(\theta^0-\widehat{\theta}) + \alpha^0_{g_i^0 t} - \widehat{\alpha}_{\widehat{g}_i t}\right)^2 + u_{it}^2\right)} \\
&- \sum_{g=1}^{G}\sqrt{P_g\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}P_g^i\left(\left(\alpha^0_{g_i^0 t} - \widehat{\alpha}_{\widehat{g}_i t}\right)^2 + u_{it}^2\right)}\Bigg| \\
&= \left|\sum_{g=1}^{G}\varphi_g(\widehat{\theta},\widehat{\alpha},\widehat{\gamma})P_g\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}P_g^i\left(\left(x_{it}'(\theta^0-\widehat{\theta}) + \alpha^0_{g_i^0 t} - \widehat{\alpha}_{\widehat{g}_i t}\right)^2 - \left(\alpha^0_{g_i^0 t} - \widehat{\alpha}_{\widehat{g}_i t}\right)^2\right)\right|
\end{aligned}$$

where

$$[\varphi_g(\widehat{\theta},\widehat{\alpha},\widehat{\gamma})]^{-1} = \sqrt{P_g \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} P_g^i \left(x_{it}'(\theta^0-\widehat{\theta}) + \alpha^0_{g_i^0 t} - \widehat{\alpha}_{\widehat{g}_i t}\right)^2 + P_g \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} P_g^i u_{it}^2} + \sqrt{P_g \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} P_g^i \left((\alpha^0_{g_i^0 t} - \widehat{\alpha}_{\widehat{g}_i t})^2 + u_{it}^2\right)}$$

is the sum of the square roots similar to what is found on the previous pages that is bounded above by a constant. Then

$$\begin{aligned}
&\left|\widetilde{Q}(\widehat{\theta},\widehat{\alpha},\widehat{\gamma}) - \widetilde{Q}(\theta^0,\widehat{\alpha},\widehat{\gamma})\right| \\
&\leq \sum_{g=1}^{G} \varphi_g(\widehat{\theta},\widehat{\alpha},\widehat{\gamma}) \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \left|\left(x_{it}'(\theta^0-\widehat{\theta})\right)\left(x_{it}'(\theta^0-\widehat{\theta}) + 2\left(\alpha^0_{g_i^0 t} - \widehat{\alpha}_{\widehat{g}_i t}\right)\right)\right| \\
&\leq C \times \left\|\theta^0-\widehat{\theta}\right\|^2 \times \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \|x_{it}\|^2 + C \times 4 \sup_{\alpha_t \in \mathcal{A}} |\alpha_t| \times \left\|\theta^0-\widehat{\theta}\right\| \times \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \|x_{it}\|
\end{aligned}$$

which is $o_p(1)$ because of the consistency of $\widehat{\theta}$.

Therefore,

$$o_p(1) = \widetilde{Q}(\widehat{\theta},\widehat{\alpha},\widehat{\gamma}) - \widetilde{Q}(\theta^0,\widehat{\alpha},\widehat{\gamma}) \leq \widetilde{Q}(\theta^0,\alpha^0,\gamma^0) - \widetilde{Q}(\theta^0,\widehat{\alpha},\widehat{\gamma}) + o_p(1).$$

Then

$$\begin{aligned}
&\widetilde{Q}(\theta^0,\alpha^0,\gamma^0) - \widetilde{Q}(\theta^0,\widehat{\alpha},\widehat{\gamma}) \\
&= \sum_{g=1}^{G} \sqrt{P_{g^0} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} P_{g^0}^i u_{it}^2} - \sqrt{P_g \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} P_g^i \left(\left(\alpha^0_{g_i^0 t} - \widehat{\alpha}_{\widehat{g}_i t}\right)^2 + u_{it}^2\right)} \\
&= \sum_{g=1}^{G} \sqrt{P_{g^0} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} P_{g^0}^i u_{it}^2} \underbrace{- \sqrt{P_g \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} P_g^i u_{it}^2} + \sqrt{P_g \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} P_g^i u_{it}^2}}_{=0} \\
&- \sqrt{P_g \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} P_g^i \left(\left(\alpha^0_{g_i^0 t} - \widehat{\alpha}_{\widehat{g}_i t}\right)^2 + u_{it}^2\right)}
\end{aligned}$$

Then, by Lemma 10,

$$
\begin{aligned}
&-\left(\sum_{g=1}^{G}\sqrt{P_g\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}P_g^i u_{it}^2}-\sum_{g=1}^{G}\sqrt{P_{g^0}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}P_{g^0}^i u_{it}^2}\right)-\sum_{g=1}^{G}P_g\varphi_g\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}P_g^i\left(\alpha_{g_i^0 t}^0-\widehat{\alpha}_{\widehat{g}_i t}\right)^2\\
&\leq -C+o_p(1)-\sum_{g=1}^{G}P_g\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}P_g^i\left(\alpha_{g_i^0 t}^0-\widehat{\alpha}_{\widehat{g}_i t}\right)^2\\
&\leq -\sum_{g=1}^{G}P_g\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}P_g^i\left(\alpha_{g_i^0 t}^0-\widehat{\alpha}_{\widehat{g}_i t}\right)^2+o_p(1)\\
&\leq o_p(1)
\end{aligned}
$$

Then,

$$o_p(1)=\sum_{g=1}^{G}P_g\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}P_g^i\left(\alpha_{g_i^0 t}^0-\widehat{\alpha}_{\widehat{g}_i t}\right)^2$$

which implies for each $g=1,\ldots,G$ (since these are all non negative terms),

$$o_p(1)=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}P_g^i\left(\alpha_{g_i^0 t}^0-\widehat{\alpha}_{\widehat{g}_i t}\right)^2 \tag{B.31}$$

so that

$$o_p(1)=\sum_{g=1}^{G}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}P_g^i\left(\alpha_{g_i^0 t}^0-\widehat{\alpha}_{\widehat{g}_i t}\right)^2 \tag{B.32}$$

$$=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\alpha_{g_i^0 t}^0-\widehat{\alpha}_{\widehat{g}_i t}\right)^2 \tag{B.33}$$

□

B.5.4 Proof of Theorem 4 (Consistency of group assignments)

Firstly I establish that WGFE group-specific effects are consistent with respect to the Hausdorff distance d_H in $\mathbb{R}^{GT}$, defined by

$$d_h(a,b)=\max\left\{\max_{g\in\{1,\ldots,G\}}\left(\min_{\widetilde{g}\in\{1,\ldots,G\}}\frac{1}{T}\sum_{t=1}^{T}(a_{\widetilde{g}t}-b_{gt})^2\right),\max_{g\in\{1,\ldots,G\}}\left(\max_{\widetilde{g}\in\{1,\ldots,G\}}\frac{1}{T}\sum_{t=1}^{T}(a_{\widetilde{g}t}-b_{gt})^2\right)\right\}.$$

Lemma 13. *Let Assumption 4 and Assumption 5(a, b) hold. Then, as $N, T \to \infty$,*

$$d_H(\widehat{\alpha}, \alpha^0) \to_p 0.$$

Proof. The proof is identical to that of Lemma B3 in Bonhomme and Manresa [2015]. I work on the terms in the maximum. I first show for any $g = 1, \ldots, G$ that

$$\min_{\widetilde{g} \in \{1,\ldots,G\}} \frac{1}{T} \sum_{t=1}^{T} (\widehat{\alpha}_{\widetilde{g}t} - \alpha^0_{gt})^2 \to_p 0. \tag{B.34}$$

Let $g \in \{1, \ldots, G\}$. I have

$$\frac{1}{NT} \sum_{i=1}^{N} \left(\min_{\widetilde{g} \in \{1,\ldots,G\}} \sum_{t=1}^{T} \mathbb{1}\{g_i^0 = g\} (\widehat{\alpha}_{\widetilde{g}t} - \alpha^0_{gt})^2 \right) = \left(\frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{g_i^0 = g\} \right) \left(\min_{\widetilde{g} \in \{1,\ldots,G\}} \frac{1}{T} \sum_{t=1}^{T} (\widehat{\alpha}_{\widetilde{g}t} - \alpha^0_{gt})^2 \right).$$

By Assumption 5(a) (group probabilities are non zero) it is enough to show that for all g, as N and T approach infinity

$$\frac{1}{NT} \sum_{i=1}^{N} \left(\min_{\widetilde{g} \in \{1,\ldots,G\}} \sum_{t=1}^{T} \mathbb{1}\{g_i^0 = g\} (\widehat{\alpha}_{\widetilde{g}t} - \alpha^0_{gt})^2 \right) \to_p 0$$

therefore to this end,

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{g_i^0 = g\} \left(\min_{\widetilde{g} \in \{1,\ldots,G\}} \frac{1}{T} \sum_{t=1}^{T} (\widehat{\alpha}_{\widetilde{g}t} - \alpha^0_{gt})^2 \right) &\leq \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{g_i^0 = g\} \left(\frac{1}{T} \sum_{t=1}^{T} (\widehat{\alpha}_{\widehat{g}_i t} - \alpha^0_{gt})^2 \right) \\ &\leq \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\widehat{\alpha}_{\widehat{g}_i t} - \alpha^0_{g_i^0 t} \right)^2 \end{aligned}$$

which is $o_p(1)$ by Proposition 3, which shows (B.34).

Now, for the second entry in the maximum, I first define the mapping

$$s(g) = \operatorname*{argmin}_{\widetilde{g} \in \{1,\ldots,G\}} \frac{1}{T} \sum_{t=1}^{T} (\widehat{\alpha}_{\widetilde{g}t} - \alpha^0_{gt})^2.$$

and show that it is one-to-one, with probability approaching one as T approaches infinity.

Let $g \neq \widetilde{g}$. By applying the reverse triangle inequality twice I have

$$\begin{aligned} \left\| \widehat{\alpha}_{s(g)} - \widehat{\alpha}_{s(\widetilde{g})} \right\| &= \left\| (\alpha^0_g - \alpha^0_{\widetilde{g}}) - (\alpha^0_g - \widehat{\alpha}_{s(g)}) - (\widehat{\alpha}_{s(\widetilde{g})} - \alpha^0_{\widetilde{g}}) \right\| \\ &\geq \left\| \alpha^0_g - \alpha^0_{\widetilde{g}} \right\| - \left\| \alpha^0_g - \widehat{\alpha}_{s(g)} \right\| - \left\| \widehat{\alpha}_{s(\widetilde{g})} - \alpha^0_{\widetilde{g}} \right\|. \end{aligned}$$

Therefore,

$$\frac{1}{\sqrt{T}}\left\|\widehat{\alpha}_{s(g)}-\widehat{\alpha}_{s(\widetilde{g})}\right\| = \left(\frac{1}{T}\sum_{t=1}^{T}(\widehat{\alpha}_{s(g)t}-\widehat{\alpha}_{s(\widetilde{g})t})^2\right)$$
$$\geq \left(\frac{1}{T}\sum_{t=1}^{T}(\alpha^0_{g}-\alpha^0_{\widetilde{g}})^2\right)^{1/2} - \left(\frac{1}{T}\sum_{t=1}^{T}(\alpha^0_{g}-\widehat{\alpha}_{s(g)})^2\right)^{1/2} - \left(\frac{1}{T}\sum_{t=1}^{T}(\widehat{\alpha}_{s(\widetilde{g})}-\alpha^0_{\widetilde{g}})^2\right)^{1/2}$$

is bounded away from zero as T grows large by Assumption 5(b) and (B.34). It follows that, with probability approaching one, $s(g) \neq s(\widetilde{g})$ for all $g \neq \widetilde{g}$ i.e. $s(g)$ is one-to-one. In particular, there exists an inverse mapping s^{-1} of s and with probability approaching one I have for all $\widetilde{g} \in \{1,\dots,G\}$:

$$\min_{g\in\{1,\dots,G\}}\frac{1}{T}\sum_{t=1}^{T}(\widehat{\alpha}_{\widetilde{g}t}-\alpha^0_{gt})^2 \leq \frac{1}{T}\sum_{t=1}^{T}\left(\widehat{\alpha}_{\widetilde{g}t}-\alpha^0_{s^{-1}(\widetilde{g})t}\right)^2 = \min_{h\in\{1,\dots,G\}}\frac{1}{T}\sum_{t=1}^{T}(\widehat{\alpha}_{ht}-\alpha^0_{s^{-1}(\widetilde{g})t})^2 \to_p 0$$

where I used (B.34) and the fact that $\widetilde{g} = s(s^{-1}(\widetilde{g}))$. This along with (B.34) completes the proof. □

The proof of Lemma 13 shows that there exists a permutation $s : \{1,\dots,G\} \to \{1,\dots,G\}$ such that

$$\frac{1}{T}\sum_{t=1}^{T}(\widehat{\alpha}_{s(g)t}-\alpha^0_{gt})^2 \to_p 0.$$

By relabeling the elements of $\widehat{\alpha}$ I can take $s(g) = g$ and this is a convention that is adopted for what remains.

The group membership profiles $\gamma_N = (g_1,\dots,g_N) \in \Gamma_G^N$ characterize a histogram

$$\lambda(\gamma_N) = N^{-1}\left(\sum_{i=1}^{N}\mathbb{1}\{g_i=1\},\dots,\sum_{i=1}^{N}\mathbb{1}\{g_i=G\}\right) \in \Delta^G$$

which converge to a probability mass function given the $\gamma_N = \{g_i\}_{i=1}^N$. In particular, the WGFE estimator $\widehat{\gamma}_N$ defines a sample mass function that will converge in probability to a probability mass function. Now, define

$$\widehat{\sigma}^2_g(\theta,\alpha,\gamma_N) = \frac{1}{T\sum_{j=1}^{N}\mathbb{1}\{g_j=g\}}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{1}\{g_i=g\}(u_{it}+x'_{it}(\theta^0-\theta)+\alpha^0_{gt}-\alpha_{gt})^2 \quad \text{(B.35)}$$

for any $(\theta, \alpha, \gamma_N) \in \Theta \times \mathcal{A}^{GT} \times \Gamma_G^N$.

By a weak law of large numbers, I have for any g as $N \to \infty$,

$$\widehat{\sigma}_g^2(\theta, \alpha, \gamma_N) \longrightarrow_p \sigma_g^2(\theta, \alpha, \gamma) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}\{g_i = g\}(u_{it} + x_{it}'(\theta^0 - \theta) + \alpha_{gt}^0 - \alpha_{gt})^2\right] \quad \text{(B.36)}$$

where γ is defined by $\lambda(\gamma^N) \to_p \lambda(\gamma) \in \Delta^G$. When $\theta = \theta^0$ and $\alpha = \alpha^0$ then this is a fixed-T variance of a subpopulation of u_{it} defined by the random partition γ.

For any $\eta > 0$ I define the subset of parameters $(\theta, \alpha) \in \mathcal{N}_\eta \subset \Theta \times \mathcal{A}^{GT}$ that satisfy $\left\|\theta - \theta^0\right\|^2 < \eta$, $\frac{1}{T}\left\|\alpha_g - \alpha_g^0\right\|^2 < \eta$ and

$$\left|\frac{\widehat{\sigma}_g(\theta, \alpha, \widehat{\gamma}_N)}{\widehat{\sigma}_{\widetilde{g}}(\theta, \alpha, \widehat{\gamma}_N)} - \frac{\widehat{\sigma}_g(\theta^0, \alpha^0, \widehat{\gamma}_N)}{\widehat{\sigma}_{\widetilde{g}}(\theta^0, \alpha^0, \widehat{\gamma}_N)}\right| < \eta$$

for all $g, \widetilde{g} = 1, \ldots, G$. The first step to showing consistency of group assignments is showing the sample probability of misassignment of individuals converges at exponential rate to zero with respect to T over all parameter values within a small enough neighborhood of the true values. Denote for fixed (θ, α) the optimal assignment of individual $i = 1, \ldots, N$ according to criterion (3.10) as $\widehat{g}_i(\theta, \alpha, \widehat{\gamma}_N)$ where $\widehat{\gamma}_N$ is the collection of optimal assignments used to determine sample group variances, which is also a function of (θ, α).

Lemma 14. *For $\eta > 0$ small enough we have, for all $\delta > 0$,*

$$\sup_{(\theta,\alpha) \in \mathcal{N}_\eta} \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{\widehat{g}_i(\theta, \alpha, \widehat{\gamma}_N) \neq g_i^0\} = o_p(T^{-\delta})$$

as $N, T \to \infty$.

Proof. For this proof I follow Bonhomme & Manresa (2015) wherever possible, however adjustments need to be made with the inclusion of second moment information. I suppress the $\widehat{\gamma}_N$ notation in the group variances until it is convenient to include it once again i.e. $\widehat{\sigma}_g(\theta, \alpha, \widehat{\gamma}_N) = \widehat{\sigma}_g(\theta, \alpha)$. By definition, I have, for all $g = 1, \ldots, G$:

$$\begin{aligned}
&\mathbb{1}\{\widehat{g}_i(\theta, \alpha, \widehat{\gamma}_N) = g\} \\
\leq &\mathbb{1}\left\{\frac{1}{\widehat{\sigma}_g(\theta, \alpha)} \sum_{t=1}^{T} (y_{it} - x_{it}'\theta - \alpha_{gt})^2 + \widehat{\sigma}_g(\theta, \alpha) \leq \frac{1}{\widehat{\sigma}_{g_i^0}(\theta, \alpha)} \sum_{t=1}^{T} (y_{it} - x_{it}'\theta - \alpha_{g_i^0 t})^2 + \widehat{\sigma}_{g_i^0}(\theta, \alpha)\right\}.
\end{aligned}$$

Then,

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\{\widehat{g}_i(\theta,\alpha,\widehat{\gamma}_N)\neq g_i^0\}=\sum_{g=1}^{G}\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\{g_i^0\neq g\}\mathbb{1}\{\widehat{g}_i(\theta,\alpha,\widehat{\gamma}_N)=g\}\\
\leq&\sum_{g=1}^{G}\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\{g_i^0\neq g\}\mathbb{1}\Bigg\{\frac{1}{\widehat{\sigma}_g(\theta,\alpha)}\sum_{t=1}^{T}(y_{it}-x_{it}'\theta-\alpha_{gt})^2+\widehat{\sigma}_g(\theta,\alpha)\\
\leq&\frac{1}{\widehat{\sigma}_{g_i^0}(\theta,\alpha)}\sum_{t=1}^{T}(y_{it}-x_{it}'\theta-\alpha_{g_i^0t})^2+\widehat{\sigma}_{g_i^0}(\theta,\alpha)\Bigg\}
\end{aligned}$$

Define $Z_{ig}(\theta,\alpha,\widehat{\gamma}_N)$ for $(\theta,\alpha)\in\mathcal{N}_\eta$ as the inner-most summand. I will bound this term by a quantity that does not depend on the parameters. Note first that

$$\begin{aligned}
Z_{ig}(\theta,\alpha,\widehat{\gamma}_N)&=\mathbb{1}\{g_i^0\neq g\}\\
\times\mathbb{1}\Bigg\{&\frac{1}{\widehat{\sigma}_g(\theta,\alpha)}\sum_{t=1}^{T}(u_{it}+x_{it}'(\theta^0-\theta)+\alpha_{g_i^0t}^0-\alpha_{gt})^2+\widehat{\sigma}_g(\theta,\alpha)\\
\leq&\frac{1}{\widehat{\sigma}_{g_i^0}(\theta,\alpha)}\sum_{t=1}^{T}(u_{it}+x_{it}'(\theta^0-\theta)+\alpha_{g_i^0t}^0-\alpha_{g_i^0t})^2+\widehat{\sigma}_{g_i^0}(\theta,\alpha)\Bigg\}\\
\leq\max_{g\neq\widetilde{g}}\mathbb{1}\Bigg\{&\frac{1}{\widehat{\sigma}_g(\theta,\alpha)}\sum_{t=1}^{T}(u_{it}+x_{it}'(\theta^0-\theta)+\alpha_{\widetilde{g}t}^0-\alpha_{gt})^2+\widehat{\sigma}_g(\theta,\alpha)\\
\leq&\frac{1}{\widehat{\sigma}_{\widetilde{g}}(\theta,\alpha)}\sum_{t=1}^{T}(u_{it}+x_{it}'(\theta^0-\theta)+\alpha_{\widetilde{g}t}^0-\alpha_{\widetilde{g}t})^2+\widehat{\sigma}_{\widetilde{g}}(\theta,\alpha)\Bigg\}
\end{aligned}$$

Before establishing a bound, I will rewrite the inequality condition through simple algebra. I can rewrite the expression within the indicator function of the previous inequality as

$$\begin{aligned}
&\sum_{t=1}^{T}(\alpha_{\widetilde{g}t}-\alpha_{gt})\left(u_{it}+x_{it}'(\theta^0-\theta)+\alpha_{\widetilde{g}t}^0-\frac{\alpha_{\widetilde{g}t}+\alpha_{gt}}{2}\right)\\
\leq&\frac{1}{2}\left(\frac{\widehat{\sigma}_g(\theta,\alpha)}{\widehat{\sigma}_{\widetilde{g}}(\theta,\alpha)}-1\right)\left(\sum_{t=1}^{T}(u_{it}+x_{it}'(\theta^0-\theta))^2+\sum_{t=1}^{T}(\alpha_{\widetilde{g}t}^0-\alpha_{\widetilde{g}t})^2\right)\\
&+\frac{\widehat{\sigma}_g(\theta,\alpha)}{\widehat{\sigma}_{\widetilde{g}}(\theta,\alpha)}\sum_{t=1}^{T}(\alpha_{\widetilde{g}t}^0-\alpha_{\widetilde{g}t})(u_{it}+x_{it}'\left(\theta^0-\theta\right)\Big)\\
&+\widehat{\sigma}_g(\theta,\alpha)\left(\frac{\widehat{\sigma}_{\widetilde{g}}(\theta,\alpha)-\widehat{\sigma}_g(\theta,\alpha)}{2}\right)
\end{aligned}$$

Make an addition on both sides by

$$A_T = \sum_{t=1}^{T}(\alpha^0_{\widetilde{g}t} - \alpha^0_{gt})\left(u_{it} + \alpha^0_{\widetilde{g}t} - \frac{\alpha^0_{\widetilde{g}t} + \alpha^0_{gt}}{2}\right) - \sum_{t=1}^{T}(\alpha_{\widetilde{g}t} - \alpha_{gt})\left(u_{it} + x'_{it}(\theta^0 - \theta) + \alpha^0_{\widetilde{g}t} - \frac{\alpha_{\widetilde{g}t} + \alpha_{gt}}{2}\right).$$

Then, the inequality becomes

$$\begin{aligned}
&\sum_{t=1}^{T}(\alpha^0_{\widetilde{g}t} - \alpha^0_{gt})\left(u_{it} + \alpha^0_{\widetilde{g}t} - \left(\frac{\alpha^0_{\widetilde{g}t} + \alpha^0_{gt}}{2}\right)\right) \\
&\leq A_T + \frac{1}{2}\left(\frac{\widehat{\sigma}_g(\theta,\alpha)}{\widehat{\sigma}_{\widetilde{g}}(\theta,\alpha)} - 1\right)\left(\sum_{t=1}^{T}(u_{it} + x'_{it}(\theta^0 - \theta))^2 + \sum_{t=1}^{T}(\alpha^0_{\widetilde{g}t} - \alpha_{\widetilde{g}t})^2\right) \\
&\quad + \frac{\widehat{\sigma}_g(\theta,\alpha)}{\widehat{\sigma}_{\widetilde{g}}(\theta,\alpha)}\sum_{t=1}^{T}(\alpha^0_{\widetilde{g}t} - \alpha_{\widetilde{g}t})(u_{it} + x'_{it}\left(\theta^0 - \theta\right)\Big) \\
&\quad + \widehat{\sigma}_g(\theta,\alpha)\left(\frac{\widehat{\sigma}_{\widetilde{g}}(\theta,\alpha) - \widehat{\sigma}_g(\theta,\alpha)}{2}\right).
\end{aligned}$$

Using this form, consider another bound of Z_{ig} through absolute values

$$\begin{aligned}
Z_{ig}(\theta,\alpha,\widehat{\gamma}_N) &\leq \max_{g\neq\widetilde{g}} \mathbb{1}\left\{\sum_{t=1}^{T}(\alpha^0_{\widetilde{g}t} - \alpha^0_{gt})\left(u_{it} + \alpha^0_{\widetilde{g}t} - \left(\frac{\alpha^0_{\widetilde{g}t} + \alpha^0_{gt}}{2}\right)\right)\right. \\
\leq |A_T| + \frac{1}{2}\left|\frac{\widehat{\sigma}_g(\theta,\alpha)}{\widehat{\sigma}_{\widetilde{g}}(\theta,\alpha)} - 1\right|&\left(\sum_{t=1}^{T}u^2_{it} + \sum_{t=1}^{T}|x'_{it}(\theta^0 - \theta)|^2 + 2\sum_{t=1}^{T}u_{it}x'_{it}(\theta^0 - \theta) + \sum_{t=1}^{T}(\alpha^0_{\widetilde{g}t} - \alpha_{\widetilde{g}t})^2\right) \qquad \text{(B.37)} \\
&+ \frac{\widehat{\sigma}_g(\theta,\alpha)}{\widehat{\sigma}_{\widetilde{g}}(\theta,\alpha)}\sum_{t=1}^{T}(\alpha^0_{\widetilde{g}t} - \alpha_{\widetilde{g}t})(u_{it} + x'_{it}\left(\theta^0 - \theta\right)\Big) \qquad \text{(B.38)} \\
&+ \frac{\widehat{\sigma}_g(\theta,\alpha)\widehat{\sigma}_{\widetilde{g}}(\theta,\alpha)}{2}\left|\frac{\widehat{\sigma}_g(\theta,\alpha)}{\widehat{\sigma}_{\widetilde{g}}(\theta,\alpha)} - 1\right|\Bigg\} \qquad \text{(B.39)}
\end{aligned}$$

Next, I will show a bound that does not depend on (θ,α) by bounding each individual term on the right-hand side of the inner inequality.

From Bonhomme and Manresa [2015], I have the bound

$$|A_T| \leq TC_1\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}u^2_{it}\right)^{1/2} + TC_2\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|\right) + TC_3\sqrt{\eta}$$

where C_1, C_2, C_3 are constants that are independent of η and T, which come from compactness of the parameter space.

The terms in the first pair of parenthesis of (B.37) are bounded by

$$\sum_{t=1}^{T} u_{it}^2 + \sum_{t=1}^{T} |x_{it}'(\theta^0 - \theta)|^2 + 2\sum_{t=1}^{T} u_{it} x_{it}'(\theta^0 - \theta) + \sum_{t=1}^{T} (\alpha_{\widetilde{g}t}^0 - \alpha_{\widetilde{g}t})^2$$
$$\leq T\left(\frac{1}{T}\sum_{t=1}^{T} u_{it}^2\right) + T\eta\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|^2\right) + 2T\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T} u_{it}^2\right)^{1/2}\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|^2\right)^{1/2} + T\eta$$

from the Cauchy-Schwarz inequality on the second and third terms along with the definition of $\mathcal{N}_\eta$ and the definition of $\mathcal{N}_\eta$ for the last term.

The terms in (B.38) are once again bounded using the Cauchy-Schwarz inequality

$$\sum_{t=1}^{T} (\alpha_{\widetilde{g}t}^0 - \alpha_{\widetilde{g}t}) u_{it} + \sum_{t=1}^{T} (\alpha_{\widetilde{g}t}^0 - \alpha_{\widetilde{g}t}) \left(x_{it}'\left(\theta^0 - \theta\right)\right)$$
$$\leq T\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T} u_{it}^2\right)^{1/2} + T\eta\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|^2\right)^{1/2}.$$

As for the variance terms, I bring back to notation indicating the collection of group assignments and bound as:

$$\begin{aligned}
\left|\frac{\widehat{\sigma}_g(\theta,\alpha,\widehat{\gamma}_N)}{\widehat{\sigma}_{\widetilde{g}}(\theta,\alpha,\widehat{\gamma}_N)} - 1\right| &\leq \left|\frac{\widehat{\sigma}_g(\theta,\alpha,\widehat{\gamma}_N)}{\widehat{\sigma}_{\widetilde{g}}(\theta,\alpha,\widehat{\gamma}_N)} - \frac{\widehat{\sigma}_g(\theta^0,\alpha^0,\widehat{\gamma}_N)}{\widehat{\sigma}_{\widetilde{g}}(\theta^0,\alpha^0,\widehat{\gamma}_N)} + \frac{\widehat{\sigma}_g(\theta^0,\alpha^0,\widehat{\gamma}_N)}{\widehat{\sigma}_{\widetilde{g}}(\theta^0,\alpha^0,\widehat{\gamma}_N)} - 1\right| \\
&\leq \left|\frac{\widehat{\sigma}_g(\theta,\alpha,\widehat{\gamma}_N)}{\widehat{\sigma}_{\widetilde{g}}(\theta,\alpha,\widehat{\gamma}_N)} - \frac{\widehat{\sigma}_g(\theta^0,\alpha^0,\widehat{\gamma}_N)}{\widehat{\sigma}_{\widetilde{g}}(\theta^0,\alpha^0,\widehat{\gamma}_N)}\right| + \left|\frac{\widehat{\sigma}_g(\theta^0,\alpha^0,\widehat{\gamma}_N)}{\widehat{\sigma}_{\widetilde{g}}(\theta^0,\alpha^0,\widehat{\gamma}_N)} - 1\right| \\
&< \eta + \sup_{\lambda(\gamma_N)\in\Delta^G}\left|\frac{\widehat{\sigma}_g(\theta^0,\alpha^0,\gamma_N)}{\widehat{\sigma}_{\widetilde{g}}(\theta^0,\alpha^0,\gamma_N)} - 1\right|
\end{aligned}$$

Next, I bound the ratio. I apply the triangle inequality after two additions by zero:

$$\frac{\widehat{\sigma}_g(\theta,\alpha,\widehat{\gamma}_N)}{\widehat{\sigma}_{\widetilde{g}}(\theta,\alpha,\widehat{\gamma}_N)} \leq \left|\frac{\widehat{\sigma}_g(\theta,\alpha,\widehat{\gamma}_N)}{\widehat{\sigma}_{\widetilde{g}}(\theta,\alpha,\widehat{\gamma}_N)} - \frac{\widehat{\sigma}_g(\theta^0,\alpha^0,\widehat{\gamma}_N)}{\widehat{\sigma}_{\widetilde{g}}(\theta^0,\alpha^0,\widehat{\gamma}_N)}\right| + \left|\frac{\widehat{\sigma}_g(\theta^0,\alpha^0,\widehat{\gamma}_N)}{\widehat{\sigma}_{\widetilde{g}}(\theta^0,\alpha^0,\widehat{\gamma}_N)} - 1\right| + 1 \tag{B.40}$$

$$\leq \eta + \sup_{\lambda(\gamma_N)\in\Delta^G}\left|\frac{\widehat{\sigma}_g(\theta^0,\alpha^0,\gamma_N)}{\widehat{\sigma}_{\widetilde{g}}(\theta^0,\alpha^0,\gamma_N)} - 1\right| + 1 \tag{B.41}$$

Next, I can easily bound $\widehat{\sigma}_g(\theta, \alpha, \widehat{\gamma}_N)$ independent of $(\theta, \alpha, \widehat{\gamma}_N)$ using compactness of the parameter space. To see this,

$$\begin{aligned}
\widehat{\sigma}_g^2(\theta, \alpha, \widehat{\gamma}_N) = \frac{1}{\sum_{i=1}^N \mathbb{1}\{\widehat{g}_i = g\}} \sum_{i=1}^N \mathbb{1}\{\widehat{g}_i = g\} \Bigg(\frac{1}{T}\sum_{t=1}^T u_{it}^2 + \frac{1}{T}\sum_{t=1}^T |x_{it}'(\theta^0 - \theta)|^2 + \frac{1}{T}\sum_{t=1}^T (\alpha_{\widetilde{g}t}^0 - \alpha_{gt})^2 \\
+ 2\frac{1}{T}\sum_{t=1}^T x_{it}'(\theta^0 - \theta)(\alpha_{\widetilde{g}t}^0 - \alpha_{gt}) + 2\frac{1}{T}\sum_{t=1}^T u_{it} x_{it}'(\theta^0 - \theta) \\
+ 2\frac{1}{T}\sum_{t=1}^T u_{it}(\alpha_{\widetilde{g}t}^0 - \alpha_{gt}) \Bigg) \\
\leq \max_{\gamma \in \Gamma_G^N} \sum_{i=1}^N \frac{\mathbb{1}\{g_i = g\}}{\sum_{i=1}^N \mathbb{1}\{g_i = g\}} \Bigg[\left(\frac{1}{T}\sum_{t=1}^T u_{it}^2 \right) + \eta \left(\frac{1}{T}\sum_{t=1}^T \|x_{it}\|^2 \right) + C_\alpha + 2C_\alpha \sqrt{\eta} \left(\frac{1}{T}\sum_{t=1}^T \|x_{it}\| \right) \\
+ 2\sqrt{\eta} \left(\frac{1}{T}\sum_{t=1}^T u_{it}^2 \right)^{1/2} \left(\frac{1}{T}\sum_{t=1}^T \|x_{it}\|^2 \right)^{1/2} + 2\sqrt{\eta} C_\alpha \left(\frac{1}{T}\sum_{t=1}^T u_{it}^2 \right)^{1/2} \Bigg]
\end{aligned}$$

where C_α is the constant that bounds $|\alpha_{\widetilde{g}t}^0 - \alpha_{gt}|$ for all $g, \widetilde{g}, t$ due to compactness of $\mathcal{A}$. Set this bound as $\widehat{\sigma}(N, T)$ so that

$$\widehat{\sigma}_g(\theta, \alpha, \widehat{\gamma}_N)\widehat{\sigma}_{\widetilde{g}}(\theta, \alpha, \widehat{\gamma}_N) \leq \widehat{\sigma}^2(N, T). \tag{B.42}$$

For (B.37),

$$\begin{aligned}
\sum_{t=1}^T u_{it}^2 + 2\sum_{t=1}^T u_{it} x_{it}'(\theta^0 - \theta) + \sum_{t=1}^T |x_{it}'(\theta^0 - \theta)|^2 + \sum_{t=1}^T (\alpha_{\widetilde{g}t}^0 - \alpha_{\widetilde{g}t})^2 \\
\leq T\left(\frac{1}{T}\sum_{t=1}^T u_{it}^2 \right) + 2T\sqrt{\eta} \left(\frac{1}{T}\sum_{t=1}^T u_{it}^2 \right)^{1/2} \left(\frac{1}{T}\sum_{t=1}^T \|x_{it}\|^2 \right)^{1/2} + \eta T \left(\frac{1}{T}\sum_{t=1}^T \|x_{it}\|^2 \right) + T\eta.
\end{aligned}$$

For (B.38),

$$\begin{aligned}
\sum_{t=1}^T (\alpha_{\widetilde{g}t}^0 - \alpha_{\widetilde{g}t}) u_{it} + \sum_{t=1}^T (\alpha_{\widetilde{g}t}^0 - \alpha_{\widetilde{g}t}) x_{it}'(\theta^0 - \theta) \\
\leq T\sqrt{\eta} \left(\frac{1}{T}\sum_{t=1}^T u_{it}^2 \right)^{1/2} + T\eta \left(\frac{1}{T}\sum_{t=1}^T \|x_{it}\|^2 \right)
\end{aligned}$$

Therefore, setting $\widehat{C}_{g,\widetilde{g}} = \sup_{\lambda(\gamma_N)\in\Delta^G} \left| \frac{\widehat{\sigma}_g(\theta^0,\alpha^0,\gamma_N)}{\widehat{\sigma}_{\widetilde{g}}(\theta^0,\alpha^0,\gamma_N)} - 1 \right|$, I have

$$Z_{ig}(\theta,\alpha,\widehat{\gamma}_N) \leq \max_{g\neq\widetilde{g}} \mathbb{1}\left\{ \sum_{t=1}^{T} (\alpha^0_{\widetilde{g}t} - \alpha^0_{gt}) \left(u_{it} + \alpha^0_{\widetilde{g}t} - \left(\frac{\alpha^0_{\widetilde{g}t} + \alpha^0_{gt}}{2} \right) \right) \right. \tag{B.43}$$

$$\leq TC_1\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T} u_{it}^2\right)^{1/2} + TC_2\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|\right) + TC_3\sqrt{\eta} \tag{B.44}$$

$$+\frac{1}{2}\left(\eta + \widehat{C}_{g,\widetilde{g}}\right)\left(T\left(\frac{1}{T}\sum_{t=1}^{T} u_{it}^2\right) + 2T\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T} u_{it}^2\right)^{1/2}\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|^2\right)^{1/2} + \eta T\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|^2\right) + T\eta\right) \tag{B.45}$$

$$+\left(\eta + \widehat{C}_{g,\widetilde{g}} + 1\right)\left(T\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T} u_{it}^2\right)^{1/2} + T\eta\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|^2\right)\right) \tag{B.46}$$

$$\left. +\frac{1}{2}\widehat{\sigma}^2(N,T)\widehat{C}_{g,\widetilde{g}} \right\} \tag{B.47}$$

Therefore, the right-hand side of the inequality does not depend on $(\theta,\alpha,\widehat{\gamma}_N)$ so, setting it as $\widetilde{Z}_{ig}$, I see that $\sup_{(\theta,\alpha)\in\mathcal{N}_\eta} Z_{ig}(\theta,\alpha,\widehat{\gamma}_N) \leq \widetilde{Z}_{ig}$. Hence,

$$\sup_{(\theta,\alpha)\in\mathcal{N}_\eta} \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{\widehat{g}_i(\theta,\alpha) \neq g_i^0\} \leq \frac{1}{N}\sum_{i=1}^{N}\sum_{g=1}^{G} \widetilde{Z}_{ig}.$$

Before moving on I will rearrange terms to get a cleaner bound by letting $\eta^* \geq \max\{\eta,\eta^2,\eta\sqrt{\eta}\}$: starting from the first term,

$$T\sqrt{\eta}\left(C_1\left(\frac{1}{T}\sum_{t=1}^{T} u_{it}^2\right)^{1/2} + C_2\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|\right) + C_3\right)$$

$$\leq T\eta^*\left(C_1\left(\frac{1}{T}\sum_{t=1}^{T} u_{it}^2\right)^{1/2} + C_2\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|\right) + C_3\right).$$

The second line is bounded by

$$T\widehat{C}_{g,\widetilde{g}}\left(\frac{1}{T}\sum_{t=1}^{T}u_{it}^2\right)+T\eta^*(1+\widehat{C}_{g,\widetilde{g}})$$
$$\times\left[\left(\frac{1}{T}\sum_{t=1}^{T}u_{it}^2\right)+2\left(\frac{1}{T}\sum_{t=1}^{T}u_{it}^2\right)^{1/2}\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|^2\right)^{1/2}+\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|^2\right)+1\right]$$

where I take special note that the first term is independent of η^*.

The third line is bounded by

$$T\eta^*(2+\widehat{C}_{g,\widetilde{g}})\left[\left(\frac{1}{T}\sum_{t=1}^{T}u_{it}^2\right)^{1/2}+\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|^2\right)\right].$$

The fourth term will end up negligible and so we'll come back to it shortly in the proof.

Now, fix $\widetilde{M}>\max\{\sqrt{M},M^*\}$, where these are constants given in Assumption 2(b) and Assumption 5(e). Note by Jensen's inequality I have $\mathbb{E}\left[u_{it}^2\right]\leq\sqrt{M}$ and by Cauchy-Schwarz $M^*\leq\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|\leq\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|^2\right)^{1/2}$. Therefore using standard probability algebra (union rule) and for all $g=1,\ldots,G$:

$$\mathbb{P}(\widetilde{Z}_{ig}=1)$$
$$\leq\sum_{\widetilde{g}\neq g}\mathbb{P}\left[\sum_{t=1}^{T}(\alpha_{\widetilde{g}t}^0-\alpha_{gt}^0)u_{it}\leq-\frac{1}{2}\sum_{t=1}^{T}(\alpha_{\widetilde{g}t}^0-\alpha_{gt}^0)^2+\frac{1}{2}T\widehat{C}_{g,\widetilde{g}}\left(\frac{1}{T}\sum_{t=1}^{T}u_{it}^2\right)\right.$$
$$+T\eta^*(1+\widehat{C}_{g,\widetilde{g}})\left(\left(\frac{1}{T}\sum_{t=1}^{T}u_{it}^2\right)+2\left(\frac{1}{T}\sum_{t=1}^{T}u_{it}^2\right)^{1/2}\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|^2\right)^{1/2}+\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|^2\right)+1\right)$$
$$+T\eta^*\left(C_1\left(\frac{1}{T}\sum_{t=1}^{T}u_{it}^2\right)^{1/2}+C_2\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|\right)+C_3\right)$$
$$+T\eta^*(2+\widehat{C}_{g,\widetilde{g}})\left(\left(\frac{1}{T}\sum_{t=1}^{T}u_{it}^2\right)^{1/2}+\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|^2\right)\right)$$
$$\left.+\frac{1}{2}\widehat{\sigma}^2(N,T)\widehat{C}_{g,\widetilde{g}}\right]$$

which is bounded by

$$\mathbb{P}(\widetilde{Z}_{ig}(\widehat{\gamma}^N)=1) \leq \sum_{g\neq\widetilde{g}} \mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\| \geq \widetilde{M}\right) + \mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T} u_{it}^2 \geq \widetilde{M}\right)$$
$$+\mathbb{P}\left(\sup_{\lambda(\gamma_N)\in\Delta^G}\left|\frac{\widehat{\sigma}_g(\theta^0,\alpha^0,\gamma_N)}{\widehat{\sigma}_{\widetilde{g}}(\theta^0,\alpha^0,\gamma_N)}-1\right| \geq \sup_{\lambda(\gamma)\in\Delta^G}\left|\frac{\sigma_g(\theta^0,\alpha^0,\gamma)}{\sigma_{\widetilde{g}}(\theta^0,\alpha^0,\gamma)}-1\right| \equiv C_{g,\widetilde{g}}\right)$$
$$+\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}(\alpha^0_{\widetilde{g}t}-\alpha^0_{gt})^2 \leq \frac{c_{g,\widetilde{g}}}{2} + \sup_{\lambda(\gamma)\in\Delta^G}\left|\frac{\sigma_g(\theta^0,\alpha^0,\gamma)}{\sigma_{\widetilde{g}}(\theta^0,\alpha^0,\gamma)}-1\right| \widetilde{M}\right)$$
$$+\sum_{g\neq\widetilde{g}}\mathbb{P}\left(\sum_{t=1}^{T}(\alpha^0_{\widetilde{g}t}-\alpha^0_{gt})u_{it} \leq -T\frac{c_{g,\widetilde{g}}}{4}\right.$$
$$\left.+T\eta^*\left((5C_{g,\widetilde{g}}+C_2+6)\widetilde{M}+(C_1+C_{g,\widetilde{g}}+2)\sqrt{\widetilde{M}}+(C_3+C_{g,\widetilde{g}}+1)\right)+C_0\right).$$

where the term independent of η^* has been eliminated and the constant comes from the union rule:

$$\frac{C_{g,\widetilde{g}}}{2}\widehat{\sigma}^2(N,T) = C_0 = \frac{C_{g,\widetilde{g}}}{2}\left(\widetilde{M}+\eta^*\left((3+2C_\alpha)\widetilde{M}+2C_\alpha\sqrt{\widetilde{M}}\right)+C_\alpha\right)\max_{\gamma\in\Gamma_G^N}\sum_{i=1}^{N}\frac{\mathbb{1}\{g_i=g\}}{\sum_{i=1}^{N}\mathbb{1}\{g_i=g\}}$$
$$= \frac{C_{g,\widetilde{g}}}{2}\left(\widetilde{M}+\eta^*\left((3+2C_\alpha)\widetilde{M}+2C_\alpha\sqrt{\widetilde{M}}\right)+C_\alpha\right)\times 1$$

given C_α some constant from the boundedness property of the compact subspace $\mathcal{A}$.

I will show the exponential rates for each term in the sum. But first, a lemma from Bonhomme and Manresa [2015].

Lemma 15 (Lemma B5 of Bonhomme and Manresa [2015])**.** *Let z_t be a strongly mixing process with zero mean, with strong mixing coefficients $\alpha[t] \leq e^{-at^{d_1}}$, and with tail probabilities $\mathbb{P}(|z_t| > z) \leq e^{1-(z/b)^{d_2}}$, where $a, b, d_1, d_2 > 0$ are constants. Then, for all $z > 0$ we have, for all $\delta > 0$, as $T \to \infty$,*

$$T^\delta \mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} z_t\right| \geq z\right) \to 0.$$

The second term is $o(T^{-\delta})$ by Lemma B5 of Bonhomme and Manresa [2015] and Assumption 2(b). To see this, set $z_t = u_{it}^2 - \mathbb{E}\left[u_{it}^2\right]$, which is necessarily strongly mixing

by Assumption 5(b), and taking $z = \widetilde{M} - \sqrt{M} > 0$ then for any $\delta > 0$

$$
\begin{aligned}
& o_p(T^{-\delta}) \\
& = \mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T}\left(u_{it}^2 - \mathbb{E}\left[u_{it}^2\right]\right)\right| \geq z\right) \\
& = \mathbb{P}\left(\left\{\frac{1}{T}\sum_{t=1}^{T}\left(u_{it}^2 - \mathbb{E}\left[u_{it}^2\right]\right) \leq -z\right\} \bigcup \left\{\frac{1}{T}\sum_{t=1}^{T}\left(u_{it}^2 - \mathbb{E}\left[u_{it}^2\right]\right) \geq z\right\}\right) \\
& \geq \mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T} u_{it}^2 \geq \widetilde{M} - (\sqrt{M} - \mathbb{E}\left[u_{it}^2\right])\right) \\
& \geq \mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T} u_{it}^2 \geq \widetilde{M}\right)
\end{aligned}
$$

where the last inequality is due to $\mathbb{E}\left[u_{it}^2\right] \leq \sqrt{M}$ by Assumption 2(b). Therefore,

$$\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T} u_{it}^2 \geq \widetilde{M}\right) = o(T^{-\delta}).$$

The third term is $o(1)$ as $N \to \infty$.

The fourth term is seen as $o(T^{-\delta})$ from a modification of the argument from Bonhomme and Manresa [2015]. We have for T large enough,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[(\alpha_{\widetilde{g}t}^0 - \alpha_{gt}^0)^2\right] \geq 0.$$

Then, applying Lemma B5 of Bonhomme and Manresa [2015] with $z_t = (\alpha_{\widetilde{g}t}^0 - \alpha_{gt}^0)^2 - \mathbb{E}\left[(\alpha_{\widetilde{g}t}^0 - \alpha_{gt}^0)^2\right]$, which satisfies appropriate mixing and tail conditions by Assumptions

2(a) and 5(b), and setting $z = c_{g,\widetilde{g}}/2 + C_{g,\widetilde{g}}\widetilde{M}$ yields

$$\begin{aligned}
o_p(T^{-\delta}) &= \mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T}(\alpha^0_{\widetilde{g}t} - \alpha^0_{gt})^2 - \mathbb{E}\left[(\alpha^0_{\widetilde{g}t} - \alpha^0_{gt})^2\right]\right| \geq z\right) \\
&= \mathbb{P}\Bigg(\left\{\frac{1}{T}\sum_{t=1}^{T}(\alpha^0_{\widetilde{g}t} - \alpha^0_{gt})^2 - \mathbb{E}\left[(\alpha^0_{\widetilde{g}t} - \alpha^0_{gt})^2\right] \leq -z\right\} \\
&\quad \bigcup \left\{\frac{1}{T}\sum_{t=1}^{T}(\alpha^0_{\widetilde{g}t} - \alpha^0_{gt})^2 - \mathbb{E}\left[(\alpha^0_{\widetilde{g}t} - \alpha^0_{gt})^2\right] \geq z\right\}\Bigg) \\
&\geq \mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}(\alpha^0_{\widetilde{g}t} - \alpha^0_{gt})^2 \leq -c_{g,\widetilde{g}}/2 + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[(\alpha^0_{\widetilde{g}t} - \alpha^0_{gt})^2\right])\right) \\
&\geq \mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}(\alpha^0_{\widetilde{g}t} - \alpha^0_{gt})^2 \leq c_{g,\widetilde{g}}/2 + C_{g,\widetilde{g}}\widetilde{M})\right)
\end{aligned}$$

therefore,

$$\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}(\alpha^0_{\widetilde{g}t} - \alpha^0_{gt})^2 \leq \frac{c_{g,\widetilde{g}}}{2} + C_{g,\widetilde{g}}\widetilde{M}\right) = o_p(T^{-\delta})$$

for any $\delta > 0$.

The last term is also $o(T^{-\delta})$. First, denote c as the minimum of the collection of $c_{g,\widetilde{g}}$ over all $g \neq \widetilde{g}$. Then, choose

$$\eta^* \leq \frac{c}{8\left((5C_{g,\widetilde{g}} + C_2 + 6)\widetilde{M} + (C_1 + C_{g,\widetilde{g}} + 2)\sqrt{\widetilde{M}} + (C_3 + C_{g,\widetilde{g}} + 1)\right)} \tag{B.48}$$

which does not depend on T.

Then for such η^* I have for all $g \neq \widetilde{g}$ the bound

$$\begin{aligned}
&\mathbb{P}\Bigg(\sum_{t=1}^{T}(\alpha^0_{\widetilde{g}t} - \alpha^0_{gt})u_{it} \leq -\frac{c_{g,\widetilde{g}}}{4} \\
&+\eta^*\left((5C_{g,\widetilde{g}} + C_2 + 6)\widetilde{M} + (C_1 + C_{g,\widetilde{g}} + 2)\sqrt{\widetilde{M}} + (C_3 + C_{g,\widetilde{g}} + 1)\right) + \frac{1}{T}C_0\Bigg) \\
&\leq \mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}(\alpha^0_{\widetilde{g}t} - \alpha^0_{gt})u_{it} \leq -\frac{c_{g,\widetilde{g}}}{8} + \frac{1}{T}C_0\right).
\end{aligned}$$

By Assumption 5(b), the process $\{(\alpha^0_{\widetilde{g}t} - \alpha^0_{gt})u_{it}\}$ has zero mean and is strongly mixing with faster-than-polynomial decay rate. Moreover, for all i, t, and $m > 0$,

$$\mathbb{P}(|(\alpha^0_{\widetilde{g}t} - \alpha^0_{gt})u_{it}| > m) \leq \mathbb{P}\left(|u_{it}| > \frac{m}{2\sup_{\alpha_t \in \mathcal{A}}}\right),$$

so $\{(\alpha^0_{\widetilde{g}t} - \alpha^0_{gt})u_{it}\}$ also satisfies the tail condition of Assumption 5(c) with a different constant $\widetilde{b} > 0$ rather than $b > 0$.

I then apply Lemma 15 with $z_t = (\alpha^0_{\widetilde{g}t} - \alpha^0_{gt})u_{it}$ and $z = z_{T^*} = \frac{c_{g,\widetilde{g}}}{8} - \frac{1}{T^*}C_0 > 0$ for T^* large enough. Therefore, for T large enough I see that

$$\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}(\alpha^0_{\widetilde{g}t} - \alpha^0_{gt})u_{it} \leq -\frac{c_{g,\widetilde{g}}}{8} + \frac{1}{T}C_0\right) = o(T^{-\delta})$$

Then, combining all results I have

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{g=1}^{G}\mathbb{P}\left(\widetilde{Z}_{ig} = 1\right) \leq G(G-1)\sup_{i=1,\ldots,N}\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N}\|x_{it}\| \geq \widetilde{M}\right) + o(T^{-\delta}) = o(T^{-\delta}).$$

To complete the proof, for η small enough (satisfying (B.48)) and T large enough, I have for all $\delta > 0$ and for all $\varepsilon > 0$,

$$\begin{aligned}\mathbb{P}\left(\sup_{(\theta,\alpha)\in\mathcal{N}_\eta}\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\{\widehat{g}_i(\theta,\alpha,\widehat{\gamma}_N) \neq g_i^0\} > \varepsilon T^{-\delta}\right) &\leq \mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{G}\sum_{g=1}^{G}\widetilde{Z}_{ig} > \varepsilon T^{-\delta}\right) \\ &\leq \frac{\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{1}{G}\sum_{g=1}^{G}\widetilde{Z}_{ig}\right]}{\varepsilon T^{-\delta}} = o(1)\end{aligned}$$

where I used the Markov inequality. Therefore,

$$\sup_{(\theta,\alpha)\in\mathcal{N}_\eta}\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\{\widehat{g}_i(\theta,\alpha,\widehat{\gamma}_N) \neq g_i^0\} = o_p(T^{-\delta}).$$

□

With this lemma, I can show the asymptotic properties claimed in the proof for $\left(\widehat{\theta},\widehat{\alpha},\widehat{\gamma}\right)$.

Define

$$\widehat{Q}(\theta,\alpha) = \sum_{g=1}^{G}\left(\frac{1}{N}\sum_{j=1}^{N}\mathbb{1}\{\widehat{g}_j(\theta,\alpha)=g\}\right)\sqrt{\widehat{Q}_g(\theta,\alpha)}$$

$$\widehat{Q}_g(\theta,\alpha) = \frac{1}{T\sum_{j=1}^{N}\mathbb{1}\{\widehat{g}_j(\theta,\alpha)=g\}}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{1}\{\widehat{g}_i(\theta,\alpha)=g\}(y_{it}-x_{it}'\theta-\alpha_{gt})^2$$

$$\widetilde{Q}(\theta,\alpha) = \sum_{g=1}^{G}\left(\frac{1}{N}\sum_{j=1}^{N}\mathbb{1}\{g_j^0=g\}\right)\sqrt{\widetilde{Q}_g(\theta,\alpha)}$$

$$\widetilde{Q}_g(\theta,\alpha) = \frac{1}{T\sum_{j=1}^{N}\mathbb{1}\{g_j^0=g\}}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{1}\{g_i^0=g\}(y_{it}-x_{it}'\theta-\alpha_{gt})^2$$

which are the objective functions from the proof of Theorem 3, fixed at the WGFE groupings and true groupings, respectively. Denote $(\widehat{\theta},\widehat{\alpha})$ and $(\widetilde{\theta},\widetilde{\alpha})$ as the minimizers of these functions, respectively.

Lemma 16. *For $\eta > 0$ small enough and for all $\delta > 0$,*

$$\sup_{(\theta,\alpha)\in\mathcal{N}_\eta}\left|\widehat{Q}(\theta,\alpha)-\widetilde{Q}(\theta,\alpha)\right| = o_p(T^{-\delta}).$$

Proof. Let $\eta > 0$ and consider $(\theta,\alpha)\in\mathcal{N}_\eta$. Note that $P_g(\widehat{\gamma}(\theta,\alpha)) = \frac{1}{N}\sum_{j=1}^{N}\mathbb{1}\{\widehat{g}_i(\theta,\alpha) = g\}$. I will suppress the $\widehat{\gamma}_N$ notation in the optimal group assignment. Consider the difference

$$\widehat{Q}(\theta,\alpha)-\widetilde{Q}(\theta,\alpha) = \sum_{g=1}^{G}\frac{\left(\frac{1}{N}\sum_{j=1}^{N}\mathbb{1}\{\widehat{g}_j(\theta,\alpha)=g\}\right)^2\widehat{Q}_g(\theta,\alpha)-\left(\frac{1}{N}\sum_{j=1}^{N}\mathbb{1}\{g_j^0=g\}\right)^2\widetilde{Q}_g(\theta,\alpha)}{\left(\frac{1}{N}\sum_{j=1}^{N}\mathbb{1}\{\widehat{g}_j(\theta,\alpha)=g\}\right)\sqrt{\widehat{Q}_g(\theta,\alpha)}+\left(\frac{1}{N}\sum_{j=1}^{N}\mathbb{1}\{g_j^0=g\}\right)\sqrt{\widetilde{Q}_g(\theta,\alpha)}}$$

where I have used the identity (B.20). Since the denominators are all $O_p(1)$ for each $g = 1,\ldots,G$, I focus on the numerator and note that

$$\left(\frac{1}{N}\sum_{j=1}^{N}\mathbb{1}\{\widehat{g}_j(\theta,\alpha)=g\}\right)^2\times\frac{1}{T\sum_{j=1}^{N}\mathbb{1}\{\widehat{g}_j(\theta,\alpha)=g\}} = \frac{1}{N}\sum_{j=1}^{N}\mathbb{1}\{\widehat{g}_j(\theta,\alpha)=g\}\times\frac{1}{NT}$$

so, for all $g = 1, \ldots, G$,

$$\left(\frac{1}{N}\sum_{j=1}^{N}\mathbb{1}\{\widehat{g}_j(\theta,\alpha)=g\}\right)^2\widehat{Q}_g(\theta,\alpha)-\left(\frac{1}{N}\sum_{j=1}^{N}\mathbb{1}\{g_j^0=g\}\right)^2\widetilde{Q}_g(\theta,\alpha)$$

$$=\frac{1}{N}\sum_{i=1}^{N}\left[\left(\frac{1}{N}\sum_{j=1}^{N}\mathbb{1}\{\widehat{g}_j(\theta,\alpha)=g\}\right)\mathbb{1}\{\widehat{g}_i(\theta,\alpha)=g\}-\left(\frac{1}{N}\sum_{j=1}^{N}\mathbb{1}\{g_j^0=g\}\right)\mathbb{1}\{g_i^0=g\}\right]$$

$$\times\frac{1}{T}\sum_{t=1}^{T}(y_{it}-x_{it}'\theta-\alpha_{gt})^2$$

$$\leq\frac{1}{N}\sum_{i=1}^{N}\underbrace{\left|\left(\frac{1}{N}\sum_{j=1}^{N}\mathbb{1}\{\widehat{g}_j(\theta,\alpha)=g\}\right)\mathbb{1}\{\widehat{g}_i(\theta,\alpha)=g\}-\left(\frac{1}{N}\sum_{j=1}^{N}\mathbb{1}\{g_j^0=g\}\right)\mathbb{1}\{g_i^0=g\}\right|}_{A_{gi}}$$

$$\times\frac{1}{T}\sum_{t=1}^{T}(y_{it}-x_{it}'\theta-\alpha_{gt})^2.$$

Then

$$A_{gi}\leq\mathbb{1}\{\widehat{g}_i(\theta,\alpha)=g\}\mathbb{1}\{g_i^0=g\}\frac{1}{N}\sum_{j=1}^{N}\underbrace{\left|\mathbb{1}\{\widehat{g}_j(\theta,\alpha)=g\}-\mathbb{1}\{g_j^0=g\}\right|}_{B_g}$$

$$+\mathbb{1}\{\widehat{g}_i(\theta,\alpha)\neq g\}\mathbb{1}\{g_i^0=g\}\frac{1}{N}\sum_{j=1}^{N}\mathbb{1}\{g_j^0=g\}$$

$$+\mathbb{1}\{\widehat{g}_i(\theta,\alpha)=g\}\mathbb{1}\{g_i^0\neq g\}\frac{1}{N}\sum_{j=1}^{N}\mathbb{1}\{\widehat{g}_j(\theta,\alpha)=g\}.$$

Note that

- $\mathbb{1}\{\widehat{g}_i(\theta,\alpha)\neq g\}\mathbb{1}\{g_i^0=g\}=\mathbb{1}\{\widehat{g}_i(\theta,\alpha)\neq g_i^0\}\mathbb{1}\{g_i^0=g\}\leq\mathbb{1}\{\widehat{g}_i(\theta,\alpha)\neq g_i^0\}$
- $\mathbb{1}\{\widehat{g}_i(\theta,\alpha)=g\}\mathbb{1}\{g_i^0=g\}=\mathbb{1}\{\widehat{g}_i(\theta,\alpha)=g_i^0\}\mathbb{1}\{g_i^0=g\}\leq 1$
- $B_g=\begin{cases}0 & \text{if } (\widehat{g}_j(\theta,\alpha)=g) \text{ and } g_j^0=g \text{ or } (\widehat{g}_j(\theta,\alpha)\neq g \text{ and } g_j^0\neq g)\\ 1 & \text{if } (\widehat{g}_j(\theta,\alpha)=g \text{ and } g_j^0\neq g) \text{ or } (\widehat{g}_j(\theta,\alpha)\neq g \text{ and } g_j^0=g)\end{cases}$

which gives us

$$\begin{aligned} B_g &= \mathbb{1}\{\widehat{g}_j(\theta,\alpha) = g\}\mathbb{1}\{g_j^0 \neq g\} + \mathbb{1}\{\widehat{g}_j(\theta,\alpha) \neq g_j^0\}\mathbb{1}\{g_j^0 = g\} \\ &\leq \mathbb{1}\{\widehat{g}_j(\theta,\alpha) = g\}\mathbb{1}\{g_j^0 \neq g\} + \mathbb{1}\{\widehat{g}_j(\theta,\alpha) \neq g_j^0\} \end{aligned}$$

where I deliberately find bounds that leverage Lemma 14. Then,

$$\begin{aligned} A_{gi} \leq & \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}\{\widehat{g}_j(\theta,\alpha) = g\}\mathbb{1}\{g_j^0 \neq g\} + \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}\{\widehat{g}_j(\theta,\alpha) \neq g_j^0\} \\ & + \mathbb{1}\{\widehat{g}_i(\theta,\alpha) \neq g_i^0\} \times 1 \\ & + \mathbb{1}\{\widehat{g}_i(\theta,\alpha) = g\}\mathbb{1}\{g_i^0 \neq g\} \times 1 \\ \leq & \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}\{\widehat{g}_j(\theta,\alpha) = g\}\mathbb{1}\{g_j^0 \neq g\} + \sup_{(\theta,\alpha)\in\mathcal{N}_\eta} \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}\{\widehat{g}_j(\theta,\alpha) \neq g_j^0\} \\ & + \mathbb{1}\{\widehat{g}_i(\theta,\alpha) \neq g_i^0\} + \mathbb{1}\{\widehat{g}_i(\theta,\alpha) = g\}\mathbb{1}\{g_i^0 \neq g\}. \end{aligned}$$

Now, by our assumptions, I have for any $(\theta,\alpha) \in \Theta \times \mathcal{A}^{GT}$, as $T \to \infty$

$$\frac{1}{T}\sum_{t=1}^{T}(y_{it} - x_{it}'\theta - \alpha_{gt})^2 = O_p(1)$$

for any $g = 1, \ldots, G$. Hence,

$$\begin{aligned} & \frac{1}{N}\sum_{i=1}^{N} A_{gi}\frac{1}{T}\sum_{t=1}^{T}(y_{it} - x_{it}'\theta - \alpha_{gt})^2 \\ \leq & O_p(1)\left[\frac{1}{N}\sum_{j=1}^{N} \mathbb{1}\{\widehat{g}_j(\theta,\alpha) = g\}\mathbb{1}\{g_j^0 \neq g\} + \sup_{(\theta,\alpha)\in\mathcal{N}_\eta} \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}\{\widehat{g}_j(\theta,\alpha) \neq g_j^0\}\right. \\ & \left. + \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{\widehat{g}_i(\theta,\alpha) \neq g_i^0\} + \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{\widehat{g}_i(\theta,\alpha) = g\}\mathbb{1}\{g_i^0 \neq g\}\right]. \end{aligned}$$

Then, by Lemma 14,

$$\begin{aligned} \left|\widehat{Q}(\theta,\alpha) - \widetilde{Q}(\theta,\alpha)\right| \leq & O_p(1)\left[\sum_{g=1}^{G}\frac{1}{N}\sum_{j=1}^{N} \mathbb{1}\{\widehat{g}_j(\theta,\alpha) = g\}\mathbb{1}\{g_j^0 \neq g\} + G\sup_{(\theta,\alpha)\in\mathcal{N}_\eta} \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}\{\widehat{g}_j(\theta,\alpha) \neq g_j^0\}\right. \\ & \left. + \sum_{g=1}^{G}\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{\widehat{g}_i(\theta,\alpha) \neq g_i^0\} + \sum_{g=1}^{G}\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{\widehat{g}_i(\theta,\alpha) = g\}\mathbb{1}\{g_i^0 \neq g\}\right] \\ = & o_p(T^{-\delta}). \end{aligned}$$

□

Proof of Theorem 4. For any $\eta > 0$, the consistency of $(\widehat{\theta}, \widehat{\alpha})$ and $(\widetilde{\theta}, \widetilde{\alpha})$ imply as $N, T \to \infty$,

$$\mathbb{P}\left(\left(\widehat{\theta}, \widehat{\alpha}\right) \notin \mathcal{N}_\eta\right) \to 0, \tag{B.49}$$

$$\mathbb{P}\left(\left(\widetilde{\theta}, \widetilde{\alpha}\right) \notin \mathcal{N}_\eta\right) \to 0. \tag{B.50}$$

Then, for any $\varepsilon > 0$,

$$\begin{aligned}
&\mathbb{P}\left[\left|\widehat{Q}\left(\widehat{\theta}, \widehat{\alpha}\right) - \widetilde{Q}\left(\widehat{\theta}, \widehat{\alpha}\right)\right| > \varepsilon T^{-\delta}\right] \\
&\leq \mathbb{P}\left(\left(\widehat{\theta}, \widehat{\alpha}\right) \notin \mathcal{N}_\eta\right) + \mathbb{P}\left[\sup_{(\theta,\alpha)\in\mathcal{N}_\eta}\left|\widehat{Q}(\theta, \alpha) - \widetilde{Q}(\theta, \alpha)\right| > \varepsilon T^{-\delta}\right] = o(1)
\end{aligned}$$

hence

$$\widehat{Q}\left(\widehat{\theta}, \widehat{\alpha}\right) - \widetilde{Q}\left(\widehat{\theta}, \widehat{\alpha}\right) = o_p(T^{-\delta})$$

and similarly for

$$\widehat{Q}\left(\widetilde{\theta}, \widetilde{\alpha}\right) - \widetilde{Q}\left(\widetilde{\theta}, \widetilde{\alpha}\right) = o_p(T^{-\delta}).$$

Then, using these and the minimizer definition of the feasible and infeasible WGFE estimators,

$$0 \leq \widetilde{Q}\left(\widehat{\theta}, \widehat{\alpha}\right) - \widetilde{Q}\left(\widetilde{\theta}, \widetilde{\alpha}\right) = \widetilde{Q}\left(\widehat{\theta}, \widehat{\alpha}\right) - \widehat{Q}\left(\widetilde{\theta}, \widetilde{\alpha}\right) + o_p(T^{-\delta}) \leq o_p(T^{-\delta}).$$

Therefore,

$$\widetilde{Q}\left(\widehat{\theta}, \widehat{\alpha}\right) - \widetilde{Q}\left(\widetilde{\theta}, \widetilde{\alpha}\right) = o_p(T^{-\delta}).$$

To show consistency of the parameter estimators, denote $\widetilde{u}_{it}^2 = y_{it} - x_{it}'\widetilde{\theta} - \widetilde{\alpha}_{g_i^0 t}$ and con-

sider the following

$$\widetilde{Q}\left(\widehat{\theta},\widehat{\alpha}\right)-\widetilde{Q}\left(\widetilde{\theta},\widetilde{\alpha}\right)=\sum_{g=1}^{G}\left(\frac{1}{N}\sum_{j=1}^{N}\mathbb{1}\{g_j^0=g\}\right)\left[\sqrt{\widetilde{Q}_g\left(\widehat{\theta},\widehat{\alpha}\right)}-\sqrt{\widetilde{Q}_g\left(\widetilde{\theta},\widetilde{\alpha}\right)}\right]$$

$$=\sum_{g=1}^{G}\left(\frac{1}{N}\sum_{j=1}^{N}\mathbb{1}\{g_j^0=g\}\right)\frac{\widetilde{Q}_g\left(\widehat{\theta},\widehat{\alpha}\right)-\widetilde{Q}_g\left(\widetilde{\theta},\widetilde{\alpha}\right)}{\sqrt{\widetilde{Q}_g\left(\widehat{\theta},\widehat{\alpha}\right)}+\sqrt{\widetilde{Q}_g\left(\widetilde{\theta},\widetilde{\alpha}\right)}}$$

$$=\sum_{g=1}^{G}\left(\frac{1}{N}\sum_{j=1}^{N}\mathbb{1}\{g_j^0=g\}\right)\frac{\widetilde{Q}_g\left(\widehat{\theta},\widehat{\alpha}\right)-\widetilde{Q}_g\left(\widetilde{\theta},\widetilde{\alpha}\right)}{2\sqrt{\widetilde{Q}_g\left(\widetilde{\theta},\widetilde{\alpha}\right)}+o_p(T^{-\delta})}$$

$$=\sum_{g=1}^{G}\frac{\frac{1}{NT}\sum_{i=1}^{N}\mathbb{1}\{g_i^0=g\}\sum_{t=1}^{T}\left(y_{it}-x_{it}'\widehat{\theta}-\widehat{\alpha}_{gt}\right)^2-\left(y_{it}-x_{it}'\widetilde{\theta}-\widetilde{\alpha}_{gt}\right)^2}{2\sqrt{\widetilde{Q}_g\left(\widetilde{\theta},\widetilde{\alpha}\right)}+o_p(T^{-\delta})}$$

$$=\sum_{g=1}^{G}\frac{\frac{1}{NT}\sum_{i=1}^{N}\mathbb{1}\{g_i^0=g\}\sum_{t=1}^{T}\left(\widetilde{u}_{it}+x_{it}'\left(\widetilde{\theta}-\widehat{\theta}\right)+\left(\widetilde{\alpha}_{gt}-\widehat{\alpha}_{gt}\right)\right)^2-\widetilde{u}_{it}^2}{2\sqrt{\widetilde{Q}_g\left(\widetilde{\theta},\widetilde{\alpha}\right)}}+o_p(T^{-\delta})$$

$$=\sum_{g=1}^{G}\frac{\frac{1}{NT}\sum_{i=1}^{N}\mathbb{1}\{g_i^0=g\}\sum_{t=1}^{T}\left(x_{it}'\left(\widetilde{\theta}-\widehat{\theta}\right)+\left(\widetilde{\alpha}_{gt}-\widehat{\alpha}_{gt}\right)\right)^2+2\widetilde{u}_{it}\left(x_{it}'\left(\widetilde{\theta}-\widehat{\theta}\right)+\left(\widetilde{\alpha}_{gt}-\widehat{\alpha}_{gt}\right)\right)}{2\sqrt{\widetilde{Q}_g\left(\widetilde{\theta},\widetilde{\alpha}\right)}}$$

$$+o_p(T^{-\delta})$$

$$=\sum_{g=1}^{G}\frac{\frac{1}{NT}\sum_{i=1}^{N}\mathbb{1}\{g_i^0=g\}\sum_{t=1}^{T}\left(x_{it}'\left(\widetilde{\theta}-\widehat{\theta}\right)+\left(\widetilde{\alpha}_{gt}-\widehat{\alpha}_{gt}\right)\right)^2}{2\sqrt{\widetilde{Q}_g\left(\widetilde{\theta},\widetilde{\alpha}\right)}}$$

$$+\sum_{g=1}^{G}\frac{1}{\sqrt{\widetilde{Q}_g\left(\widetilde{\theta},\widetilde{\alpha}\right)}}\frac{1}{NT}\sum_{i=1}^{N}\mathbb{1}\{g_i^0=g\}\sum_{t=1}^{T}\widetilde{u}_{it}\left(x_{it}'\left(\widetilde{\theta}-\widehat{\theta}\right)+\left(\widetilde{\alpha}_{gt}-\widehat{\alpha}_{gt}\right)\right)+o_p(T^{-\delta}).$$

The second term is zero and to see that, recall the first-order conditions for $\gamma=\gamma^0$ fixed:

$$0=\sum_{g=1}^{G}\frac{1}{\sqrt{\widetilde{Q}_g(\widetilde{\theta},\widetilde{\alpha})}}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{1}\{g_i^0=g\}x_{it}\left(y_{it}-x_{it}'\widetilde{\theta}-\widetilde{\alpha}_{gt}\right)$$

$$0=\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\{g_i^0=g\}\left(y_{it}-x_{it}'\widetilde{\theta}-\widetilde{\alpha}_{gt}\right),\quad\text{for}\quad t=1,\ldots,T$$

so that the second term is equal to

$$\sum_{g=1}^{G} \frac{1}{\sqrt{\widetilde{Q}_g\left(\widetilde{\theta},\widetilde{\alpha}\right)}} \left[\frac{1}{NT} \sum_{i=1}^{N} \mathbb{1}\{g_i^0 = g\} \sum_{t=1}^{T} \widetilde{u}_{it} x_{it}' \left(\widetilde{\theta} - \widehat{\theta}\right) + \frac{1}{T} \sum_{t=1}^{T} \left(\widetilde{\alpha}_{gt} - \widehat{\alpha}_{gt}\right) \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{g_i^0 = g\} \widetilde{u}_{it} \right]$$

$$= \left(\widetilde{\theta} - \widehat{\theta}\right)' \left[\sum_{g=1}^{G} \frac{1}{\sqrt{\widetilde{Q}_g\left(\widetilde{\theta},\widetilde{\alpha}\right)}} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{1}\{g_i^0 = g\} x_{it} \widetilde{u}_{it} \right] + 0$$

$$= 0$$

by using the first-order conditions of $\widetilde{Q}$.

Therefore,

$$\widetilde{Q}\left(\widehat{\theta},\widehat{\alpha}\right) - \widetilde{Q}\left(\widetilde{\theta},\widetilde{\alpha}\right)$$

$$= \sum_{g=1}^{G} \frac{1}{2\sqrt{\widetilde{Q}_g\left(\widetilde{\theta},\widetilde{\alpha}\right)}} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{1}\{g_i^0 = g\} \left(x_{it}' \left(\widetilde{\theta} - \widehat{\theta}\right) + \left(\widetilde{\alpha}_{gt} - \widehat{\alpha}_{gt}\right)\right)^2 + o_p(T^{-\delta})$$

$$\geq \widetilde{C} \sum_{g=1}^{G} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{1}\{g_i^0 = g\} \left(x_{it}' \left(\widetilde{\theta} - \widehat{\theta}\right) + \left(\widetilde{\alpha}_{gt} - \widehat{\alpha}_{gt}\right)\right)^2 + o_p(T^{-\delta})$$

$$\geq \widetilde{C} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(x_{it}' \left(\widetilde{\theta} - \widehat{\theta}\right) + \left(\widetilde{\alpha}_{g_i^0 t} - \widehat{\alpha}_{g_i^0 t}\right)\right)^2 + o_p(T^{-\delta})$$

$$\geq \widetilde{C} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(x_{it}' \left(\widetilde{\theta} - \widehat{\theta}\right) + \overline{x}_{g_i^0 t}' \left(\widetilde{\theta} - \widehat{\theta}\right)\right)^2 + o_p(T^{-\delta})$$

$$= \widetilde{C} \left(\widetilde{\theta} - \widehat{\theta}\right)' \left[\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(x_{it} - \overline{x}_{g_i^0 t}\right) \left(x_{it} - \overline{x}_{g_i^0 t}\right)' \right] \left(\widetilde{\theta} - \widehat{\theta}\right) + o_p(T^{-\delta})$$

$$\geq \widetilde{C} \widehat{\rho} \left\| \widetilde{\theta} - \widehat{\theta} \right\|^2 + o_p(1) + o_p(T^{-\delta})$$

where $\max_g \sqrt{\widetilde{Q}_g(\widetilde{\theta},\widetilde{\alpha})} < 1/\widetilde{C}$. Therefore,

$$\widehat{\theta} = \widetilde{\theta} + o_p(T^{-\delta}).$$

To show $\widehat{\alpha} \to_p \alpha$, I use the fact that $\widehat{\theta} = \widetilde{\theta} + o_p(T^{-\delta})$:

$$\widetilde{Q}\left(\widehat{\theta}, \widehat{\alpha}\right) - \widetilde{Q}\left(\widetilde{\theta}, \widetilde{\alpha}\right)$$

$$= \sum_{g=1}^{G} \frac{1}{2\sqrt{\widetilde{Q}_g\left(\widetilde{\theta}, \widetilde{\alpha}\right)}} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{1}\{g_i^0 = g\} \left(x_{it}'\left(\widetilde{\theta} - \widehat{\theta}\right) + \left(\widetilde{\alpha}_{gt} - \widehat{\alpha}_{gt}\right)\right)^2 + o_p(T^{-\delta})$$

$$\geq O_p(1) \times \sum_{g=1}^{G} \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{g_i^0 = g\} \frac{1}{T} \sum_{t=1}^{T} \left(\widetilde{\alpha}_{gt} - \widehat{\alpha}_{gt}\right)^2 + o_p(T^{-\delta})$$

$$= O_p(1) \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\widetilde{\alpha}_{g_i^0 t} - \widehat{\alpha}_{g_i^0 t}\right)^2 + o_p(T^{-\delta})$$

so $\frac{1}{T}\left\|\widetilde{\alpha}_g - \widehat{\alpha}_g\right\|^2 = o_p(T^{-\delta})$ for any $g = 1, \ldots, G$ therefore

$$\left\|\widetilde{\alpha}_g - \widehat{\alpha}_g\right\|^2 = o_p(T^{1-\delta}).$$

To show consistency of group assignments, by a union bound we have

$$\mathbb{P}\left(\sup_{i=1,\ldots,N} \left|\widehat{g}_i\left(\widehat{\theta}, \widehat{\alpha}\right) - g_i^0\right| > 0\right) \leq \mathbb{P}\left(\left(\widehat{\theta}, \widehat{\alpha}\right) \notin \mathcal{N}_\eta\right) + N \sup_{i=1,\ldots,N} \mathbb{P}\left(\left(\widehat{\theta}, \widehat{\alpha}\right) \in \mathcal{N}_\eta, \widehat{g}_i\left(\widehat{\theta}, \widehat{\alpha}\right) \neq g_i^0\right).$$

Now, taking η satisfying (B.48) gives us $\mathbb{P}\left(\left(\widehat{\theta}, \widehat{\alpha}\right) \notin \mathcal{N}_\eta\right) = o(1)$. Then, we know

$$\sup_{(\theta,\alpha) \in \mathcal{N}_\eta} \mathbb{1}\{\widehat{g}(\theta, \alpha) \neq g_i^0\} \leq \sum_{g=1}^{G} \widetilde{Z}_{ig}$$

where $\widetilde{Z}_{ig}$ is given by (B.43). Then, for all $\delta > 0$,

$$\begin{aligned}
N \sup_{i=1,\ldots,N} \mathbb{P}\left(\left(\widehat{\theta}, \widehat{\alpha}\right) \in \mathcal{N}_\eta, \widehat{g}_i\left(\widehat{\theta}, \widehat{\alpha}\right) \neq g_i^0\right) &= N \sup_{i=1,\ldots,N} \mathbb{E}\left[\mathbb{1}\{\left(\widehat{\theta}, \widehat{\alpha}\right) \in \mathcal{N}_\eta\} \mathbb{1}\{\widehat{g}(\theta, \alpha) \neq g_i^0\}\right] \\
&\leq N \sup_{i=1,\ldots,N} \mathbb{E}\left[\mathbb{1}\{\left(\widehat{\theta}, \widehat{\alpha}\right) \in \mathcal{N}_\eta\} \sum_{g=1}^{G} \widetilde{Z}_{ig}\right] \\
&\leq N \sup_{i=1,\ldots,N} \sum_{g=1}^{G} \mathbb{P}\left(\widetilde{Z}_{ig} = 1\right) \\
&\leq N\left[G(G-1) \sup_{i=1,\ldots,N} \mathbb{P}\left(\frac{1}{T} \sum_{t=1}^{T} \|x_{it}\| \geq \widetilde{M}\right) + o(T^{-\delta})\right] \\
&= o(NT^{-\delta}).
\end{aligned}$$

□

Appendix C

TYPE FIXED EFFECTS AND RATIONAL ADDICTION: A GMM FRAMEWORK FOR LATENT TYPE HETEROGENEITY

C.1 Plots & Additional Results

Bias	$\mathbb{E}[\theta_k(\xi_i)]$		$\text{Var}(\theta_k(\xi_i))$	
Specification	$\theta_1(\xi_i)$	$\theta_2(\xi_i)$	$\theta_1(\xi_i)$	$\theta_2(\xi_i)$
$\theta_1 = \theta_2$ & FE	0.073	0.047	-0.024	-0.024
$\theta_1 = \theta_2$ & TWFE	0.073	0.050	-0.023	-0.025
$\theta_1 = \theta_2$ & IFE	0.058	0.056	-0.025	-0.025
$\theta_1 = \theta_2$ & AR(1), $\rho = 0.9$	0.066	0.051	-0.009	-0.007
$\log(\theta_1) = \theta_2$ & AR(1), $\rho = 0.75$	0.062	0.065	-0.006	-0.083

Table C.1: Bias of the mean and variance estimators of the type-heterogeneous coefficients over 100 simulations across different specifications. Bandwidth chosen is $h = 0.15$. Larger bandwidth than the simulation presented in the body of the paper is in line with oversmoothing: estimates are more biased over most of the categories.

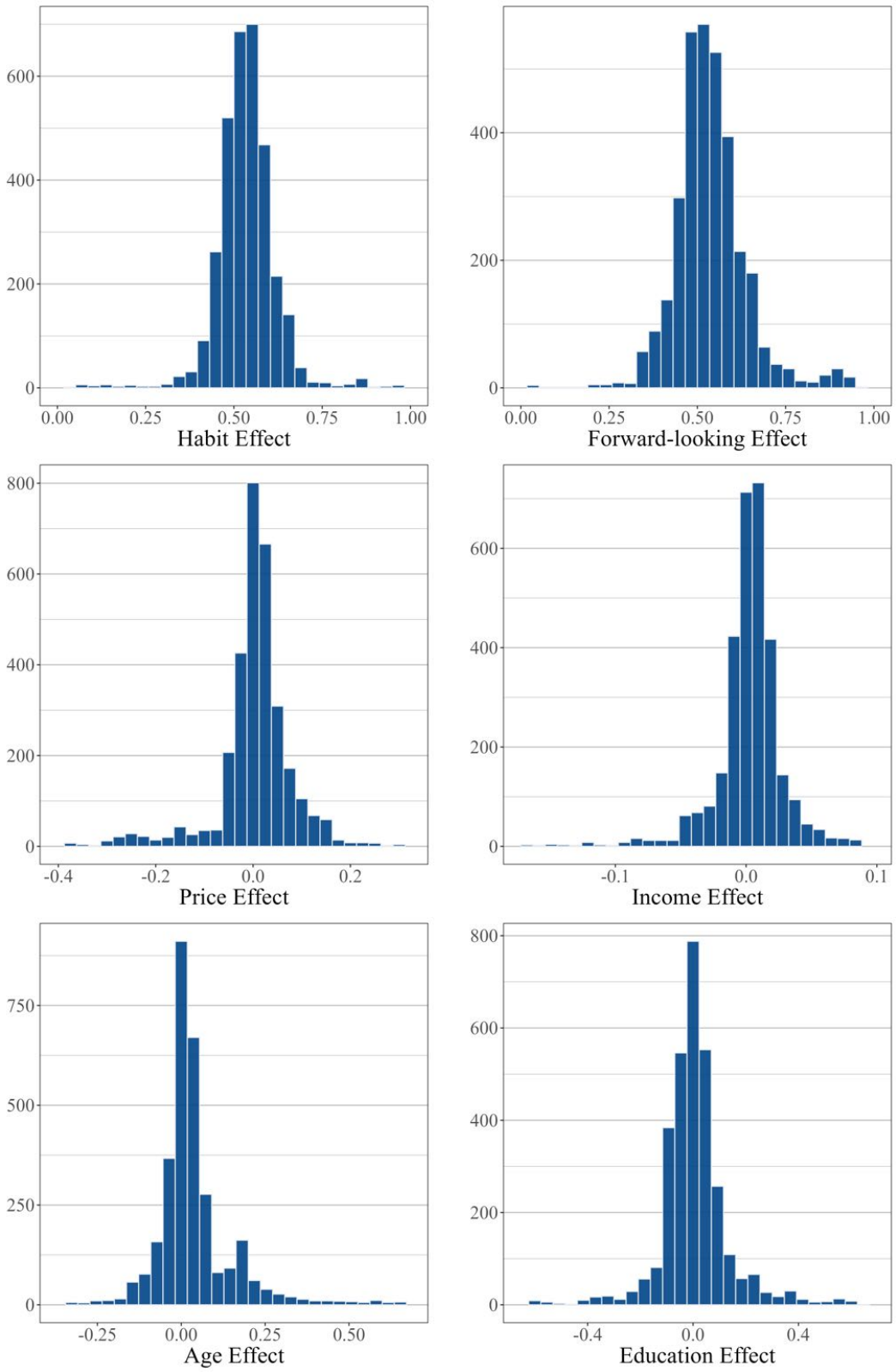


Figure C.1: Histograms of TFE-GMM estimates.

Bias	$\mathbb{E}[\theta_k(\xi_i)]$		$\text{Var}(\theta_k(\xi_i))$	
Specification	$\theta_1(\xi_i)$	$\theta_2(\xi_i)$	$\theta_1(\xi_i)$	$\theta_2(\xi_i)$
$\theta_1 = \theta_2$ & FE	0.072	0.049	0.014	0.015
$\theta_1 = \theta_2$ & TWFE	0.074	0.049	0.018	0.015
$\theta_1 = \theta_2$ & IFE	0.059	0.055	0.012	0.011

Table C.2: Bias of the mean and variance estimators of the type-heterogeneous coefficients over 100 simulations across different specifications. Bandwidth chosen is $h = 0.01$. The smaller bandwidth changes the direction of the bias on the variance estimator and in magnitude it is the smallest among the rest of the results.

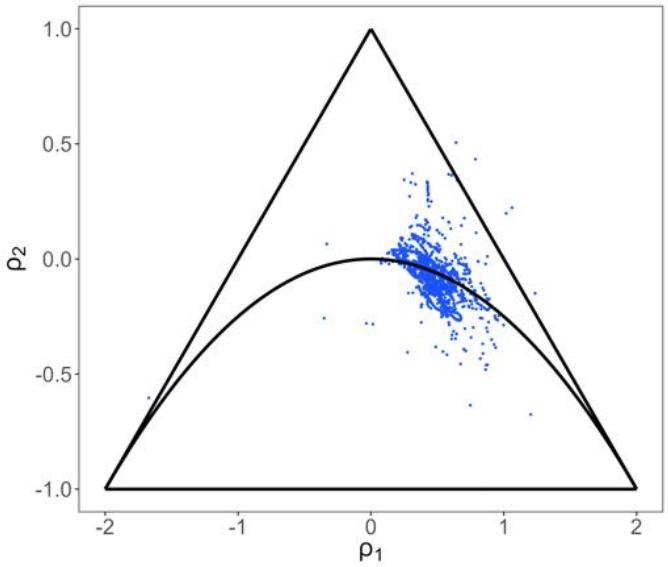


Figure C.2: Showing consumption paths are stationary AR(2).

C.2 Proofs for the main results

C.2.1 Proof of Theorem 5

Recall the function

$$\xi_i^0 = F(w_i; \theta^0, \alpha^0) = \underset{\xi \in \Xi}{\operatorname{argmin}} \left\| \mathbb{E}\left[g(w_{it}; \theta^0(\xi), \alpha_t^0(\xi)) \right] \right\|^2 \tag{C.1}$$

where the expectation is taken over the time-series dimension throughout.

The conclusion of the theorem requires that a change-of-variables applies to $F(\cdot; \theta^0 \alpha^0)$ in order to move from the space of observables to the unobserved types. Therefore we will need to show that F is differentiable as a function of w; see Bonnans and Shapiro [2000] for general arguments of functions of this form. Since g is a continuous function and w_{it} is well-behaved (Assumptions 7 (c) and (e)), differentiation and integration can be interchanged.

Let

$$G(w, \xi) = \left\| \mathbb{E}\left[g(w_{it}; \theta^0(\xi), \alpha_t^0(\xi)) \right] \right\|^2 \tag{C.2}$$

and, by definition of F, a necessary condition for a minimum is

$$\frac{\partial G(w, F(w; \theta^0, \alpha^0))}{\partial \xi} = 0 \tag{C.3}$$

where the derivative exists by Assumption 7(b, c), $F(w; \theta^0, \alpha^0) \in \mathring{\Xi}$ by Assumption 7 (a), and Assumption 7(d) will ensure the first-order condition produces the unique absolute minimum in the interior for any $w \in \mathcal{W}$.

Along the lines of the implicit function theorem, taking the partial derivative with respect to w of this first-order condition will yield a system of equations we can solve for $\frac{\partial F(w; \theta^0, \alpha^0)}{\partial w}$ provided differentiability conditions are met.

The derivative of G is given as

$$\begin{aligned} \frac{\partial G(w, \xi)}{\partial \xi} =& 2\mathbb{E}\left[\frac{\partial g(w_{it}; \theta^0(\xi), \alpha_t(\xi))}{\partial \theta^0} \cdot \frac{\partial \theta^0(\xi)}{\partial \xi} \right]' \mathbb{E}\left[g(w_{it}; \theta^0(\xi), \alpha_t^0(\xi)) \right] \\ &+ 2\mathbb{E}\left[\frac{\partial g(w_{it}; \theta^0(\xi), \alpha_t(\xi))}{\partial \alpha_t^0} \cdot \frac{\partial \alpha_t^0(\xi)}{\partial \xi} \right]' \mathbb{E}\left[g(w_{it}; \theta^0(\xi), \alpha_t^0(\xi)) \right]. \end{aligned}$$

Assuming for a moment that we are allowed to take the derivative of (C.3), we get

$$\frac{\partial^2 G(w, F(w;\theta^0,\alpha^0))}{\partial \xi^2} \cdot \frac{\partial F(w;\theta^0,\alpha^0)}{\partial w} + \frac{\partial^2 G(w, F(w;\theta^0,\alpha^0))}{\partial \xi \partial w} = 0. \tag{C.4}$$

All that remains to show is that the second partial exists and is non zero, and the cross-partial derivative exists and, by Assumption (b, c), they must exist since g is twice differentiable in its arguments and both θ^0 and α^0 are twice-differentiable and bounded. Evaluated at $\xi = F(w;\theta^0,\alpha^0)$ it is

$$\begin{aligned} &\frac{\partial^2 G(w, F(w;\theta^0,\alpha^0))}{\partial \xi^2} \\ &= 2 \left\| \mathbb{E}\left[\frac{\partial g(w_{it};\theta^0(\xi),\alpha_t(\xi))}{\partial \theta^0} \cdot \frac{\partial \theta^0(\xi)}{\partial \xi} + \frac{\partial g(w_{it};\theta^0(\xi),\alpha_t(\xi))}{\partial \alpha_t^0} \cdot \frac{\partial \alpha_t^0(\xi)}{\partial \xi} \right] \right\|^2 \Bigg|_{\xi = F(w;\theta^0,\alpha^0)} \end{aligned} \tag{C.5}$$

since the term with second order derivatives will retain the moment condition (4.5) and, by definition, it is zero at $\xi = F(w;\theta^0,\alpha^0)$. By Assumption 7 (d), (C.2.1) is positive so that the gradient of F with respect to w is well-defined by the implicit function theorem. By Assumption 7 (a) and convexity of G at $\xi = F(w;\theta^0,\alpha^0)$, $w \mapsto F(\cdot;\theta^0,\alpha^0) \in \mathring{\Xi}$ is injective: for any $w \in \mathcal{W}$ a unique ξ is produced by F as an argmin function by convexity. Therefore, the conditions for a change-of-variables is satisfied.

C.2.2 Proof of consistency of $\left(\widehat{\theta}, \widehat{\alpha}\right)$ *in endogenous linear model* (4.6)

This proof follows the strategy of Bonhomme and Manresa [2015] in their Appendix covering the group heterogeneous coefficients case. Let $\mu = (\mu_1, \mu_2, \ldots, \mu_N) \in \Xi^N$ denote a vector of types assigned to each individual in the sample. Denote $\xi^0 = (\xi_1^0, \ldots, \xi_N^0) \in \Xi^N$ as the population types. Let $\{\widehat{W}_i\}_{i\in\mathcal{N}}$ be a collection of positive definite matrices. Recall that for any positive definite matrix W, there exists a unique positive definite matrix $W^{1/2}$ such that $W = W^{1/2}W^{1/2}$ so that there exists a collection $\{\widehat{W}_i^{1/2}\}_{i\in\mathcal{N}}$ such that $\widehat{W}_i = \widehat{W}_i^{1/2}\widehat{W}_i^{1/2}$ for all $i \in \mathcal{N}$. Therefore for any $h > 0$ we can rewrite the TFE-GMM

objective function as

$$\widehat{Q}(\theta,\alpha,\mu) = \frac{1}{N}\sum_{i=1}^{N}\int_{\Xi}\widehat{Q}_i(\xi,\theta,\alpha,\mu)K_h(\xi-\mu_i)\,d\xi \tag{C.6}$$

where, for any $i \in \mathcal{N}$, the individual's GMM criterion is

$$\widehat{Q}_i(\xi,\theta,\alpha,\mu) = \left\|\widehat{W}_i^{1/2}\frac{1}{T}\sum_{t=1}^{T} z_{it}\left(y_{it} - x_{it}'\theta(\xi) - \alpha_t(\xi)\right)\right\|^2. \tag{C.7}$$

Using the true DGP (4.6) for y_{it} we can rewrite this as

$$\widehat{Q}_i(\xi,\theta,\alpha,\mu) = \left\|\widehat{W}_i^{1/2}\frac{1}{T}\sum_{t=1}^{T} z_{it}\left(u_{it} + x_{it}'\left(\theta^0(\xi_i^0) - \theta(\xi)\right) + \left(\alpha_t^0(\xi_i^0) - \alpha_t(\xi)\right)\right)\right\|^2. \tag{C.8}$$

Then, define an auxiliary objective function as

$$\widetilde{Q}(\theta,\alpha,\mu) = \frac{1}{N}\sum_{i=1}^{N}\int_{\Xi}\widetilde{Q}_i(\xi,\theta,\alpha,\mu)K_h(\xi-\mu_i)\,d\xi \tag{C.9}$$

where, for any $i \in \mathcal{N}$,

$$\widetilde{Q}_i(\xi,\theta,\alpha,\mu) = \left\|\widehat{W}_i^{1/2}\frac{1}{T}\sum_{t=1}^{T} z_{it}\left(x_{it}'\left(\theta^0(\xi_i^0) - \theta(\xi)\right) + \left(\alpha_t^0(\xi_i^0) - \alpha_t(\xi)\right)\right)\right\|^2 + \left\|\widehat{W}_i^{1/2}\frac{1}{T}\sum_{t=1}^{T} z_{it}u_{it}\right\|^2. \tag{C.10}$$

I show that $\widehat{Q}$ is uniformly convergent to $\widetilde{Q}$ as N, T tend to infinity and h tends to zero.

Lemma 17. *Let Assumptions 7 and 9 hold and suppose $h \to 0$ as $N, T \to \infty$ and $Nh \to \infty$. Then,*

$$\operatorname*{plim}_{N,T\to\infty}\sup_{(\theta,\alpha,\mu)\in\Theta\times\mathcal{A}\times\Xi^N}|\widehat{Q}(\theta,\alpha,\mu) - \widetilde{Q}(\theta,\alpha,\mu)| = 0 \tag{C.11}$$

Proof. Expanding the $\widehat{Q}_i$ for any $i \in \mathcal{N}$ using bilinearity of the inner product gives

$$\widehat{Q}_i(\xi,\theta,\alpha,\mu) = \left\|\widehat{W}_i^{1/2}\frac{1}{T}\sum_{t=1}^{T} z_{it}\left(x_{it}'\left(\theta^0(\xi_i^0) - \theta(\xi)\right) + \left(\alpha_t^0(\xi_i^0) - \alpha_t(\xi)\right)\right)\right\|^2 \tag{C.12}$$

$$+2\left(\frac{1}{T}\sum_{t=1}^{T} z_{it}u_{it}\right)'\widehat{W}_i\left(\frac{1}{T}\sum_{t=1}^{T} z_{it}\left(x_{it}'\left(\theta^0(\xi_i^0) - \theta(\xi)\right) + \left(\alpha_t^0(\xi_i^0) - \alpha_t(\xi)\right)\right)\right) \tag{C.13}$$

$$+\left\|\widehat{W}_i^{1/2}\frac{1}{T}\sum_{t=1}^{T} z_{it}u_{it}\right\|^2 \tag{C.14}$$

so that the difference for each individual GMM criterions for any parameter values is (C.13). Therefore, the difference in (C.11) is

$$\frac{2}{N}\sum_{i=1}^{N}\int_{\Xi}\left(\frac{1}{T}\sum_{t=1}^{T}z_{it}u_{it}\right)'\widehat{W}_i\left(\frac{1}{T}\sum_{t=1}^{T}z_{it}\left(x_{it}'\left(\theta^0(\xi_i^0)-\theta(\xi)\right)+\left(\alpha_t^0(\xi_i^0)-\alpha_t(\xi)\right)\right)\right)K_h(\xi-\mu_i)\,d\xi. \tag{C.15}$$

Next the integral is reduced using the mean value theorem where I assume without loss of generality that the Lebesgue measure of Ξ is 1. By Assumption 7(a), Ξ must be an interval and by Assumption 7(c) the integrand must be continuous as a function of ξ so there exists $\overline{\xi}\in\mathring{\Xi}$ such that (C.15) is

$$\frac{2}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}z_{it}u_{it}\right)'\widehat{W}_i\left(\frac{1}{T}\sum_{t=1}^{T}z_{it}\left(x_{it}'\left(\theta^0(\xi_i^0)-\theta(\overline{\xi})\right)+\left(\alpha_t^0(\xi_i^0)-\alpha_t(\overline{\xi})\right)\right)\right)K_h(\overline{\xi}-\mu_i). \tag{C.16}$$

Since the sample is iid over the cross-sectional dimension, we can study the limiting behavior for each term in the inner product. For the first, consider

$$\mathbb{E}\left[\left\|\frac{1}{T}\sum_{t=1}^{T}z_{it}u_{it}\right\|^2\right]=\frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}\left[z_{it}'z_{is}u_{it}u_{is}\right]\leq\frac{M}{T}\to 0 \tag{C.17}$$

as $T\to\infty$ by Assumption 9(d) so by Jensen's inequality the first term vanishes asymptotically. Next, we need to ensure that the other term is bounded. Consider it in two parts:

$$\left\|\frac{1}{T}\sum_{t=1}^{T}z_{it}x_{it}'\left(\theta^0(\xi_i^0)-\theta(\overline{\xi})\right)\right\|\leq\left(\frac{1}{T}\sum_{t=1}^{T}\left\|z_{it}x_{it}'\right\|\right)\left\|\theta^0(\xi_i^0)-\theta(\overline{\xi})\right\| \tag{C.18}$$

where Jensen's and Cauchy-Schwarz were applied successively. Since θ^0 and θ are both functions in Θ, by Assumption 9(b), it must be that $\left\|\theta^0(\xi_i^0)-\theta(\overline{\xi})\right\|\leq\eta$ for some scalar $\eta>0$. Therefore, by Assumption 9(c),

$$\left(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|z_{it}x_{it}'\right\|\right]\right)\left\|\theta^0(\xi_i^0)-\theta(\overline{\xi})\right\|\leq\frac{\eta}{T}\sum_{i=1}^{T}M=\eta M. \tag{C.19}$$

Lastly,

$$\left\|\frac{1}{T}\sum_{t=1}^{T}z_{it}\left(\alpha_t^0(\xi_i^0)-\alpha_t(\overline{\xi})\right)\right\|^2\leq\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_t^0(\xi_i^0)-\alpha_t(\overline{\xi})\right)^2\right)\left(\frac{1}{T}\sum_{t=1}^{T}\left\|z_{it}\right\|^2\right)$$

where $\mathcal{A}$ is a space of bounded functions by Assumption 9(b) so

$$\mathbb{E}\left\|\frac{1}{T}\sum_{t=1}^{T} z_{it}\left(\alpha_t^0(\xi_i^0)-\alpha_t(\overline{\xi})\right)\right\|^2 \leq \eta M. \tag{C.20}$$

Then, by applying Cauchy-Schwarz inequality on (C.16), by Assumption 9(e) on the kernel function and using the above derived bounds finishes the proof. □

The next lemma shows that the auxiliary objective function is uniquely minimized at the true values.

Lemma 18. *Suppose that Assumption 7 and 9 holds. Then, there exists a $C > 0$ for any $(\theta, \alpha, \mu) \in \Theta \times \mathcal{A} \times \Xi^N$,*

$$\widetilde{Q}(\theta,\alpha,\mu)-\widetilde{Q}(\theta^0,\alpha^0,\xi^0) \geq C\left[\left\|\theta^0-\theta\right\|_2^2+\frac{1}{T}\sum_{t=1}^{T}\left\|\alpha_t^0-\alpha_t\right\|_2^2\right]+o_p(1), \tag{C.21}$$

where the vector-function norm is defined as (4.34).

Proof. I begin with arguing that the auxiliary objective function vanishes at the true values asymptotically as $N, T \to \infty$ and $h \to 0$. Consider the following:

$$\begin{aligned}\widetilde{Q}(\theta^0,\alpha^0,\xi^0) = \frac{1}{N}\sum_{i=1}^{N}\int_{\Xi}\left\|\widehat{W}_i^{1/2}\frac{1}{T}\sum_{t=1}^{T} z_{it}\left(x_{it}'\left(\theta^0(\xi_i^0)-\theta^0(\xi)\right)+\left(\alpha_t^0(\xi_i^0)-\alpha_t^0(\xi)\right)\right)\right\|^2 K_h(\xi-\xi_i^0)\,d\xi \\ +\frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{W}_i^{1/2}\frac{1}{T}\sum_{t=1}^{T} z_{it}u_{it}\right\|^2\int_{\Xi}K_h(\xi-\xi_i^0)\,d\xi\end{aligned}$$

The second term in the sum is $o_p(1)$ by Assumption 9(d) as argued in the proof of Lemma 17 and the fact that the kernel is a bounded function on Ξ. For N and T sufficiently large, h will be small and, since the integrand is a continuous function on compact Ξ, by Assumption 9(e) the kernel function will weakly converge to the Dirac delta function. Therefore, the limit will be the integrand evaluated at $\xi = \xi_i^0$ so $\widetilde{Q}(\theta^0, \alpha^0, \xi^0) \to_p 0$ as $N, T \to \infty$.

Shift attention to the first term in the sum. Then, denoting $\widehat{c}_i$ as the minimum eigenvalue of $\widehat{W}_i$ for all $i \in \mathcal{N}$ and $\widehat{c} = \min_{i\in\mathcal{N}} \widehat{c}_i$, the difference can be written as

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\int_{\Xi}\left\|\widehat{W}_i^{1/2}\frac{1}{T}\sum_{t=1}^{T} z_{it}\left(x_{it}'\left(\theta^0(\xi_i^0)-\theta(\xi)\right)+\left(\alpha_t^0(\xi_i^0)-\alpha_t(\xi)\right)\right)\right\|^2 K_h(\xi-\mu_i)\,d\xi \\
\geq&\widehat{c}\frac{1}{N}\sum_{i=1}^{N}\int_{\Xi}\left\|\frac{1}{T}\sum_{t=1}^{T} z_{it}\left(x_{it}'\left(\theta^0(\xi_i^0)-\theta(\xi)\right)+\left(\alpha_t^0(\xi_i^0)-\alpha_t(\xi)\right)\right)\right\|^2 K_h(\xi-\mu_i)\,d\xi \\
=&\lim_{b\to 0}\frac{1}{N}\sum_{i=1}^{N}\widehat{c}_i\int_{\Xi\times\Xi}\left\|\frac{1}{T}\sum_{t=1}^{T} z_{it}\left(x_{it}'\left(\theta^0(\widetilde{\xi})-\theta(\xi)\right)+\left(\alpha_t^0(\widetilde{\xi})-\alpha_t(\xi)\right)\right)\right\|^2 K_h(\xi-\mu_i)K_b(\widetilde{\xi}-\xi_i^0)\,d\xi\,d\widetilde{\xi} \\
\geq&\lim_{b\to 0}\frac{1}{N}\sum_{i=1}^{N}\widehat{c}_i\int_{\Xi}\left\|\frac{1}{T}\sum_{t=1}^{T} z_{it}\left(x_{it}'\left(\theta^0(\xi)-\theta(\xi)\right)+\left(\alpha_t^0(\xi)-\alpha_t(\xi)\right)\right)\right\|^2 K_h(\xi-\mu_i)K_b(\xi-\xi_i^0)\,d\xi
\end{aligned}$$

using the fact of weak convergence of the kernel to the Dirac delta function and that $\Xi\times\Xi \supset \{(\xi,\widetilde{\xi}) \in \Xi\times\Xi : \xi = \widetilde{\xi}\}$ so one of the variables of integration is eliminated. Now, using Jensen's inequality:

$$\begin{aligned}
&\lim_{b\to 0}\frac{1}{N}\sum_{i=1}^{N}\widehat{c}_i\int_{\Xi}\left\|\frac{1}{T}\sum_{t=1}^{T} z_{it}\left(x_{it}'\left(\theta^0(\xi)-\theta(\xi)\right)+\left(\alpha_t^0(\xi)-\alpha_t(\xi)\right)\right)\right\|^2 K_h(\xi-\mu_i)K_b(\xi-\xi_i^0)\,d\xi \\
\geq&\lim_{b\to 0}\widehat{c}\,\widehat{p}(\mu)\int_{\Xi}\left\|\sum_{i=1}^{N}\frac{1}{T}\sum_{t=1}^{T} z_{it}\left(x_{it}'\left(\theta^0(\xi)-\theta(\xi)\right)+\left(\alpha_t^0(\xi)-\alpha_t(\xi)\right)\right)\frac{K_h(\xi-\mu_i)K_b(\xi-\xi_i^0)}{\sum_{j=1}^{N}K_h(\xi-\mu_j)K_b(\xi-\xi_j^0)}\right\|^2 d\xi
\end{aligned}$$

where

$$\frac{1}{N}\sum_{j=1}^{N}K_h(\xi-\mu_j)K_b(\xi-\xi_j^0) \geq \min_{\xi\in\Xi}\widehat{P}(\xi,\mu) = \widehat{p}(\mu) \geq 0.$$

Let $Z_{it} \in \mathbb{R}^k$ denote the non constant elements of z_{it} and let $\mathcal{M}(\mu,\xi)$ be the $(p+T)\times(p+T)$ matrix

$$\sum_{i=1}^{N}\frac{K_h(\xi-\mu_i)K_b(\xi-\xi_i^0)}{\sum_{j=1}^{N}K_h(\xi-\mu_j)K_b(\xi-\xi_j^0)}\begin{bmatrix}\frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}x_{it}Z_{it}'Z_{is}x_{is}' & \frac{1}{T^{3/2}}\sum_{t=1}^{T}\sum_{s=1}^{T}x_{it}Z_{it}'Z_{is}\delta_s' \\ \frac{1}{T^{3/2}}\sum_{t=1}^{T}\sum_{s=1}^{T}\delta_t Z_{it}'Z_{is}x_{is}' & \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\delta_t Z_{it}'Z_{is}\delta_s'\end{bmatrix}, \tag{C.22}$$

Then,

$$\lim_{b\to 0}\widehat{c}\,\widehat{p}(\mu)\int_{\Xi}\left\|\sum_{i=1}^{N}\frac{1}{T}\sum_{t=1}^{T}z_{it}\left(x_{it}'\left(\theta^0(\xi)-\theta(\xi)\right)+\left(\alpha_t^0(\xi)-\alpha_t(\xi)\right)\right)\frac{K_h(\xi-\mu_i)K_b(\xi-\xi_i^0)}{\sum_{j=1}^{N}K_h(\xi-\mu_j)K_b(\xi-\xi_j^0)}\right\|^2 d\xi$$

$$=\lim_{b\to 0}\widehat{c}\,\widehat{p}(\mu)\int_{\Xi}\begin{bmatrix}\theta^0(\xi)-\theta(\xi)\\ \frac{1}{\sqrt{T}}\left(\alpha^0(\xi)-\alpha(\xi)\right)\end{bmatrix}'\mathcal{M}(\mu,\xi)\begin{bmatrix}\theta^0(\xi)-\theta(\xi)\\ \frac{1}{\sqrt{T}}\left(\alpha^0(\xi)-\alpha(\xi)\right)\end{bmatrix}d\xi$$

$$\geq\lim_{b\to 0}\widehat{c}\min_{\mu\in\Xi^N}\widehat{p}(\mu)\min_{\mu\in\Xi^N,\xi\in\Xi}\widehat{\rho}(\mu,\xi)\int_{\Xi}\left[\left\|\theta^0(\xi)-\theta(\xi)\right\|^2+\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_t^0(\xi)-\alpha_t(\xi)\right)^2\right]d\xi$$

$$=C\left[\left\|\theta^0-\theta\right\|^2+\frac{1}{T}\sum_{t=1}^{T}\left\|\alpha_t^0-\alpha_t\right\|^2\right]$$

where $C=\lim_{b\to 0}\widehat{c}\min_{\mu\in\Xi^N}\widehat{p}(\mu)\min_{\mu\in\Xi^N,\xi\in\Xi}\widehat{\rho}(\mu,\xi)>0$ by Assumption 9 (a,f), thus completing the proof. □

To show consistency of the parameters, by Lemma 17 and 18 and the definition of the TFE-GMM estimator (4.32) as the minimizer of $\widehat{Q}$, we have that

$$\widetilde{Q}(\widehat{\theta},\widehat{\alpha},\widehat{\mu})=\widehat{Q}(\widehat{\theta},\widehat{\alpha},\widehat{\mu})+o_p(1)\leq\widehat{Q}(\theta^0,\alpha^0,\mu^0)+o_p(1)=\widetilde{Q}(\theta^0,\alpha^0,\mu^0)+o_p(1) \quad \text{(C.23)}$$

so, because $C>0$,

$$o_p(1)\leq C\left[\left\|\theta^0-\widehat{\theta}\right\|^2+\frac{1}{T}\sum_{t=1}^{T}\left\|\alpha_t^0-\widehat{\alpha}_t\right\|^2\right]+o_p(1)\leq\widetilde{Q}(\widehat{\theta},\widehat{\alpha},\widehat{\mu})-\widetilde{Q}(\theta^0,\alpha^0,\mu^0)\leq o_p(1). \quad \text{(C.24)}$$

Therefore it must be that $\left\|\theta^0-\widehat{\theta}\right\|^2=o_p(1)$ and $\frac{1}{T}\sum_{t=1}^{T}\left\|\alpha_t^0-\widehat{\alpha}_t\right\|^2=o_p(1)$. □