University of California
Los Angeles

Estimating Unobserved Group Effects

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Economics

by

Benjamin David Shapiro

2024

Abstract of the Dissertation

Estimating Unobserved Group Effects

by

Benjamin David Shapiro
Doctor of Philosophy in Economics
University of California, Los Angeles, 2024
Professor Denis Chetverikov, Chair

This dissertation introduces a time-varying unobserved group-period fixed effect estimator designed to address specific challenges in causal inference. The proposed estimator accommodates scenarios where treated individuals can transition between unobserved groups following treatment. Developed within a difference-in-differences framework, it is particularly valuable for controlling violations of parallel trends arising from unobserved group changes. For example, when estimating the impact of job loss on health outcomes without observing insurance status, the estimator helps account for the confounding effect of losing insurance (due to job loss) on health. Additionally, the approach proves useful for estimating the average treatment effect on the treated (ATT) when treatment compliance is unobserved.

The second chapter introduces a mixed integer optimization (MIO) procedure for estimating individual group assignments. While prior literature has often relied on K-means clustering for identifying unobserved group membership, this approach lacks asymptotic guarantees, and finite sample performance in the presence of non-spherical distributions and outliers. The MIO formulation, by contrast, provides global optimality and asymptotic guarantees, ensuring accurate estimation of group membership and convergence to our theoretical characterization of the estimator's distribution. However, due to the NP-Hard nature of the problem, the MIO approach becomes infeasible for datasets with more than 200 entities.

To address MIO's computational limitations, the third chapter presents a novel branch-and-bound algorithm leveraging proof that our estimators decision boundary is linear. Instead of directly searching over individual group memberships, the algorithm searches for the linear decision boundary that determines group assignments. This method significantly improves computational efficiency, allowing it to handle large-scale problems. For instance, while the MIO formulation may take months to solve a problem with 1,000 entities, the branch-and-bound algorithm can solve it within seconds. We show that this optimization procedure can offer significant improvements in accuracy over the K-means algorithm. Although the current implementation is limited to low-dimensional settings with two unobserved groups, the framework holds promise for extension to high-dimensional settings involving multiple groups.

The dissertation of Benjamin David Shapiro is approved.

Alper Atamtürk

Rosa Matzkin

Andres Santos

Denis Chetverikov, Committee Chair

University of California, Los Angeles

2024

To my mother and father, I am eternally grateful for your unwavering belief in me and for the endless support you have given throughout my journey. Your faith in my potential, even when I doubted it, has been my anchor. I will never forget how you applied to Penn State on my behalf without my knowledge—without that pivotal moment and your guidance, I would not be where I am today. Thank you for everything.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my deepest gratitude to my advisors, Denis Chetverikov, Andres Santos, Alper Atamtürk, and Rosa Matzkin, for their invaluable guidance, unwavering support, and insightful feedback throughout the development of this dissertation. Their mentorship has been instrumental in shaping my research and intellectual growth, and I am incredibly fortunate to have worked under their direction.

I also extend my sincere thanks to the members of the Econometric faculty and Adriana Lleras-Muney for their thoughtful comments, constructive criticism, and encouragement throughout this process. Their expertise and suggestions have significantly enriched this work.

I am especially grateful to my family and friends, whose love and support have sustained me throughout the highs and lows of this journey. I am profoundly grateful to my spouse, Shu Li, thank you for your patience, encouragement, and belief in me, even when the road seemed long and uncertain.

Special thanks to my colleagues and fellow graduate students for their camaraderie, collaboration, and insightful discussions that have contributed to my academic and personal development.

Vita

2008–2012	Cryptologist, United States Navy, Everett Washington
2012–2016	B.S. in Mathematic, B.A in Economics, B.A in Political science, *Schreyers Honors College*, The Pennsylvania State University
2017–2019	Research Assistant, Federal Reserve Bank of San Francisco
2019–2021	M.A. in Economics, Department of Economics, UCLA
2023–2024	Economist Intern, Amazon, Seattle Washington

CHAPTER 1

Time Varying Unobserved Group Period Fixed Effects Estimator

1.1 Introduction

In many applications, researchers encounter challenges when individuals transition between unobserved groups, which can confound the estimation of effects of interest. For instance, consider studying the impact of job loss on mental health outcomes. When someone loses their job, they may also lose their health insurance. Failing to account for this change in health insurance status could bias estimates of the true effect of job loss on mental health.

This study aims to develop a novel estimator to consistently estimate individual group choices across time, enabling precise estimation of the effect of interest. We develop this estimator within the difference-in-differences framework where we assume treated are changing groups due to treatment and controls do not change groups. This assumption will allow us to use the controls as an identification mechanism to connect treated group labels between pre/post treatment helping us identify group change effects.

This chapter extends the time-varying unobserved group fixed effects estimator introduced by Bonhomme and Manresa (2015) [BM15], who established the specific asymptotic assumptions required for consistency and asymptotic normality in the panel data setting. Our contribution is to relax their framework allowing for individuals to transition between groups over time using a difference-in-difference setting. We also develop a global optimization procedure for estimating group membership but that will be discussed more in Chapter

2 and 3.

This study was motivated by the rise of quasi-experimental methods, the continued reliance on traditional fixed effects, and the need for a more flexible fixed effect estimator. Currie, Kleven, and Zwiers (2020) [CKZ20] document these trends by estimating the frequency different methodologies are mentioned in NBER working papers and top 5 journals. They found that the use of the difference-in-differences method increased from near zero to 25% of working papers between 1980 and 2016. They also found that the use of fixed effects increased from 20% to 60% between 1980 and 2016. Suggesting modern economists heavily rely on difference-in-difference and are increasingly facing unobserved heterogeneity.

To contextualize modern endogeneity issues within quasi-experimental studies, we consider several examples. In Ayyagari and Shane (2015) [AS15], the authors estimate the impact of medicaid drug coverage expansion on mental health using longitudinal data. They observe Medicaid eligibility, partially observe private insurance status, and do not observe Medicaid use.

As a result, they can only estimate the intent-to-treat (ITT) effect, rather than the average treatment effect on the treated (ATT). If they fully observed all the information they could estimate the effect of Medicaid drug coverage expansion given a person was previously insured or uninsured. This would provide a more complete picture of the policies effect.

Similarly, Gebel and Vobemer (2014) [GV14] estimate the impact of job loss on health outcomes using longitudinal data, but they do not account for health insurance status, which introduces a confounding factor into their analysis. By observing health insurance status, they could disentangle the effect of job loss from the effect of losing insurance, thereby providing a clearer understanding of how employment loss affects health outcomes. If the loss of insurance is found to be the primary driver of negative health effects, this would lend strong support to policies aimed at helping the unemployed maintain health coverage. These and other examples demonstrate the confounding effects caused by unobserved group changes in the quasi-experimental settings. Our estimator addresses these challenges by estimating individuals group choices over time allowing us to control for these confounding effects.

1.2 Literature Review

Estimation of latent groups in panel data has been a focus of extensive research, with various methods proposed, including mixture models, differencing techniques, and factor models. In this chapter we extend Bonhomme and Manresa (2015) who introduced the Grouped Fixed Effects (GFE) estimator, a seminal approach for modeling unobserved group structures by clustering individuals into groups with shared fixed effects through minimizing the sum of squared residuals. Extensions to the GFE framework include Rivero (2023) [Riv23], who developed the Weighted Grouped Fixed Effects (WGFE) estimator, addressing group-specific heteroskedasticity to improve classification accuracy and estimation efficiency. The GFE estimator is also closely related to the interactive fixed effects (IFE) model proposed by Bai (2009) [Bai09], which models unobserved heterogeneity as a linear combination of time-varying factors and individual-specific loadings. While the IFE framework is flexible, its restrictive assumptions, such as factor orthogonality, limit its practical applicability. Ando and Bai (2016) [AB16] extended the IFE framework to explicitly account for latent group structures, combining the strengths of interactive effects and grouped fixed effects for a more nuanced understanding of unobserved heterogeneity.

An alternative perspective is provided by Bester and Hansen (2016) [BH16], who developed a latent group fixed effects estimator allowing individual-specific effects to vary within groups. They identified two primary sources of bias, incidental parameter bias and misspecification bias, and proposed grouping individuals with similar effects or reducing group sizes to address these. While innovative, practical challenges arise when individual effects can vary within group.

In a related strand, finite mixture models have been used to estimate latent classes, with Deb and Trivedi (1997) [DT97] applying a finite mixture negative binomial model to healthcare demand and Sun (2005) [Sun05] using multinomial logistic regression to estimate latent group choices. A key limitation of finite mixture models is the need to specify the error term distribution, adding complexity to estimation.

Theoretical insights into the incidental parameter problem also play a crucial role in understanding biases in panel data models with latent groups. Hahn and Moon (2010) [HM10] demonstrated that in game-theoretic models with finite equilibria, the incidental parameter problem is minimal even when the cross-sectional dimension grows exponentially with time. Additionally, Hahn and Newey (2004) [HN04] introduced a jackknife bias correction method, providing broader implications for addressing incidental parameter bias.

In this chapter we will assume the number of unknown groups is known. However, when the number of latent groups is unknown, Bonhomme and Manresa (2015) and Bai (2009) suggested using Bayesian Information Criteria (BIC) for group estimation. Su, Shi, and Phillips (2016) [SSP16] advanced this by developing the Classifier-Lasso (C-Lasso) method, which simultaneously identifies group memberships and estimates model parameters. Huang, Jin, and Su (2020) [HJS20] later extended C-Lasso to nonstationary panels, enabling applications in dynamic settings.

In summary, these developments highlight the diversity of approaches for modeling latent groups in panel data, ranging from fixed effects to mixture models and penalized estimation methods. Extensions of the GFE framework, advancements in factor models, and tools for estimating unknown group counts provide robust methodologies for addressing unobserved heterogeneity. The integration of these approaches to handle more complex settings, including dynamic group membership and unknown group structures, represents a promising direction for future research.

1.3 Difference-in-Difference Framework

In this section, we introduce our Difference-in-Differences (DiD) model, which is designed to accommodate multiple time periods and account for effects that vary based on unobserved groups. The DiD model is not only widely used in Economics due to its popularity, but it also provides a powerful framework for leveraging control variables as an identification mechanism, enabling us to link treated group labels over time. To illustrate, consider the following data generating process:

$$y_{it} = \lambda^0_{g^0_{it},t} + \zeta^0_{g^0_{it}} 1\{D_{iT'} = 1\} + \delta^0_{g^0_{ib},g^0_{ia}} D_{it} + x'_{it}\beta^0 + \epsilon_{it} \tag{1.1}$$

In equation 1.1, g^0_{it} represents which group individual i belongs to at time t where ^{0} is used to represent the "true" group or "true" parameter. We assume that treated individuals can only change groups at the treatment period, denoted T', and there is only one treatment period. Thus, we can denote the before treatment group as $g^0_{ib} \in \{1, ..., G\}$ and the after treatment group as $g^0_{ia} \in \{1, ..., G\}$ as described in equation 1.2.

$$g^0_{it} = \begin{cases} g^0_{ib}, & \text{if } t < T' \\ g^0_{ia}, & \text{if } t \geq T' \end{cases} \tag{1.2}$$

Equation 1.1 presents a standard Difference-in-Difference framework with multiple time periods. For clarity, the components of the model are defined as follows. First, the group-specific time trends, $\lambda^0_{g^0_{it},t} \in \Lambda^{GT}$, account for time-varying trends across different groups. The treatment status of individual i at time t is captured by D_{it}, and the intercept for treated individuals in group g^0_{it} is represented by $\zeta^0_{g^0_{it}} \in Z^G$. This specification allows for not only a difference in average outcomes between treated and control groups, but also for the possibility that the within-group differences among treated individuals may differ from those among controls. If needed, control variables, $x'_{it}\beta^0$ where $\beta^0 \in B^p$, can be incorporated to address potential violations of the parallel trends assumption. Lastly, the treatment effect parameter, $\delta^0_{g^0_{ib},g^0_{ia}} \in \Delta^{G^2}$, captures the treatment effect while accounting for the individual's group membership both before (g^0_{ib}) and during (g^0_{ia}) the treatment period. This ensures that the effect of treatment can vary not only across time but also depending on an individual's group affiliation at different stages. To simplify notation we will denote $1_{D_i} = 1\{D_{iT'} = 1\}$

thus our outcome model becomes.

$$y_{it} = \quad \lambda^0_{g^0_{it},t} + \zeta^0_{g^0_{it}} 1_{D_i} + \delta^0_{g^0_{ib},g^0_{ia}} D_{it} + x'_{it}\beta^0 + \epsilon_{it} \tag{1.3}$$

1.4 Time Varying Unobserved Group Period Estimator

In this section we develop our time varying unobserved group period fixed effects estimator (TV-GPFE). We estimate our parameters using a least squares framework where we not only minimize over our parameters but also over all possible individual group combinations $\gamma = \{(g_{1b}, g_{1a}), \ldots, (g_{Nb}, g_{Na})\} \in \Gamma^{G^2}$. It is important to note that an exhaustive search, even for relatively small problems, would take an impractically long time to compute. This poses a significant computational challenge. In the next section, we will propose global and local estimation procedures based on sample size to address this issue.

$$(\hat{\lambda}, \hat{\zeta}, \hat{\delta}, \hat{\beta}, \hat{\gamma}) = \underset{(\lambda,\zeta,\delta,\beta,\gamma)\in(\Lambda^{GT}, Z^G, \Delta^{G^2}, B^P, \Gamma^{G^2})}{\text{argmin}} \textstyle\sum_{i=1}^N \sum_{t=1}^T (y_{it} - \lambda_{g_{it},t} - \zeta_{g_{it}} 1_{D_i} - \delta_{g_{ib},g_{ia}} D_{it} - x'_{it}\beta)^2 \tag{1.4}$$

lets also define the infeasible estimator where we know individual group choices for each period.

$$(\tilde{\lambda}, \tilde{\zeta}, \tilde{\delta}, \tilde{\beta}) = \underset{(\lambda,\zeta,\delta,\beta)\in(\Lambda^{GT}, Z^G, \Delta^G, B^P)}{\text{argmin}} \textstyle\sum_{i=1}^N \sum_{t=1}^T (y_{it} - \lambda_{g^0_{it},t} - \zeta_{g^0_{it}} 1_{D_i} - \delta_{g^0_{ib},g^0_{ia}} D_{it} - x'_{it}\beta)^2 \tag{1.5}$$

In the subsequent sections, we will demonstrate that our TV-GPFE estimator converges in probability to our infeasible estimator. This implies that we can perform inference around the distribution of our infeasible estimator, which can be conceptualized as a panel data model where T approaches infinity.

There are numerous potential extensions to this model and estimator. For instance, if there is a positive probability of observing an individuals group choice for every possible group then you no longer need group identification assumptions and optimization will become faster. Below, we present several other possible extensions for the model.

1.4.1 Extension 1: Staggered Difference-in-Difference

A widely used approach in Economics is the Staggered Difference-in-Differences (DiD) model, where treatment is implemented at different times for different individuals. To account for this staggered treatment, we modify the treatment period indicator to be individual-specific, denoted as T'_i. Consequently, the treatment indicator D_{it} is redefined as follows:

$$D_{it} = \begin{cases} 1 & \text{if } t \geq T'_i \\ 0 & \text{if } t < T'_i \end{cases}$$

If we assume that the treatment periods for individuals are compact, we can consistently estimate the individuals' group selection both before and after treatment, under the same assumptions as in our current model. This framework enables us to model situations where treatment timing varies across individuals. Formally, our staggered treatment model is defined as follows:

$$y_{it} = \quad \lambda^0_{g^0_{it},t} + \zeta^0_{g^0_{it}} 1\{D_{iT'_i} = 1\} + \delta^0_{g^0_{ib},g^0_{ia}} D_{it} + x'_{it}\beta^0 + \epsilon_{it}$$

1.4.2 Extension 2: Intent To Treat

In many applications, we do not directly observe whether individuals were treated (D_{it}); rather, we only observe their eligibility for treatment, denoted as Z_{it}. Consequently, we cannot calculate the Average Treatment on the Treated (ATT) and must instead estimate the Intent to Treat (ITT) effect. The ITT is essentially the ATT biased towards zero. Ideally, we aim to estimate the ATT.

Our unobserved groups consist of individuals who comply with the treatment and those who do not. We hypothesize that if an individual complies with the treatment, they receive the treatment effect δ; otherwise, they receive no effect. Let us define $\delta_{g_{ia}}$ accordingly.

$$\delta_{g_{ia}} = \begin{cases} \delta & \text{if } D_{it} = 1 \\ 0 & \text{if } D_{it} = 0 \end{cases}$$

Now we set up our model to allow for differences in behavior over time between compliers and non-compliers, as well as distinct intercepts for these groups.

$$y_{it} = \quad \lambda^0_{g^0_{it},t} + \zeta^0_{g^0_{it}} 1_{D_i} + \delta^0_{g^0_{ia}} Z_{it} + x'_{it}\beta^0 + \epsilon_{it}$$

By estimating which group each individual is in we can back out ATT when treating treatment effects as fixed effects.

1.4.3 Extension 3: Multiple Treatments

In certain applications, it is important to evaluate the effectiveness of multiple treatments. To test multiple experiments we can randomly assign individuals to several groups. Let D'_{it} represent a vector of P treatment group indicators $D^1_{it}, ..., D^P_{it}$ and let $\delta'_{g^0_{ib},g^0_{ia}}$ denote a vector of P treatment effects $\delta^1_{g^0_{ib},g^0_{ia}}, ..., \delta^P_{g^0_{ib},g^0_{ia}}$. We will assume individuals receive the treatment simultaneously but you could staggered treatment over time. As long as the number of treatment groups and treatment times are compact then using the assumptions laid out in our asymptotic section we will be able to estimate the following model consistently.

$$y_{it} = \quad \lambda^0_{g^0_{it},t} + \zeta^0_{g^0_{it}} 1_{D_i} + D'_{it}\delta_{g^0_{ib},g^0_{ia}} + x'_{it}\beta^0 + \epsilon_{it}$$

1.4.4 Extension 4: Random Effects

In the current framework, we assume that unobserved group effects are fixed. However, in some cases, it may be more appropriate to model these effects as originating from an unknown group-specific distribution. This adjustment introduces significant challenges in group identification, particularly when the distributions overlap, making it difficult to distinguish between groups. Nonetheless, when one group follows a distribution and the other is fixed, identification becomes more feasible. This situation arises in the compliance literature, where it is uncertain whether an individual received treatment. If an individual received the treatment, their outcomes are drawn from the treated distribution; if they did not, they exhibit a fixed zero treatment effect.

In future research, we aim to demonstrate that by constructing a shrinking window around the fixed point (zero), it is possible to ensure that the probability of a treated individual being in this window approaches zero, while the probability of a non-treated individual being in the window converges to one, as both the number of individuals (N) and the number of time periods (T) increase. Thus, the model's success will depend on the relative growth rates of N and T.

1.5 Optimization

As can be seen in equation 1.4, conducting an exhaustive search over individual group choices becomes computationally infeasible as the problem size increases. Bonhomme and Manresa (2015) introduced a K-means procedure that efficiently estimates individual group choices, though it lacks asymptotic guarantees of correct classification. In this section, we propose a novel mixed integer optimization approach that ensures the accurate identification of individuals' true group memberships, providing a more robust solution in the asymptotic limit. This procedure was built upon a previous paper on mixed integer optimization for estimating group choice in a panel data setting. We discuss this in more detail in chapter 2.

1.5.1 Mixed Integer Optimization

Bertsimas, King, and Mazumder (2016) [BKM16] demonstrated that advancements in integer optimization, coupled with hardware improvements, have led to a remarkable 200 billion-fold increase in computational efficiency. They showed that mixed-integer optimization (MIO) could solve the best subset selection problem for instances with N in the thousands and P in the hundreds within minutes. Moreover, MIO could achieve near-optimal solutions for instances with N in the hundreds and P in the thousands within minutes. Building on this work, we aim to develop an MIO formulation to determine group membership for individuals.

Consider the following optimization procedure. Let $p \in P$ be the period, $g \in G$ be the group, N denote the number of individuals, and M be arbitrarily large. $z_{ipg} \in \{0, 1\}$ will be an indicator that determines which group g individual i belongs to at period p. Specifically, if z_{ipg} is 0 for group 1 this enforces $\tilde{\zeta}_{ip} = \zeta_1$ effectively assigning individual i to group 1 for the respective period.

$$\begin{aligned}
\min_{(\lambda,\zeta,\delta,\beta,z)} \textstyle\sum_{i=1}^{N} \sum_{t=1}^{T} \quad & \left(y_{it} - \tilde{\lambda}_{ipt} - \tilde{\zeta}_{ip} 1_{D_i} - \tilde{\delta}_i D_{it} - X'\beta\right)^2 & \\
& \text{subject to} & \\
& |\tilde{\zeta}_{ip} - \zeta_g| \leq z_{ipg} M & \forall i, p, g \\
& |\tilde{\lambda}_{ipt} - \lambda_{gt}| \leq z_{ipg} M & \forall i, p, g, t \\
& |\tilde{\delta}_i - \delta_{g,g'}| \leq z_{i1g} z_{i2g'} M & \forall i \\
& \textstyle\sum_{g=1}^{G} z_{ipg} = G - 1 & \forall i, p
\end{aligned} \tag{1.6}$$

The computational efficiency stems from the way we search over z_{ipg}. Bertsimas et. al. (2016) provide an excellent summary of the significant computational advancements in mixed-integer optimization (MIO) over the past two decades. Essentially, these advancements include sophisticated branch-and-bound techniques and effective pruning methods, which systematically divide and reduce the parameter space during the optimization process. By leveraging these methods, we can significantly enhance the speed and accuracy of our search algorithm, enabling us to handle complex problems more efficiently. To estimate these equations, several computational software options are available, including Gurobi, CPLEX, and MOSEK.

To further enhance computational efficiency, we can employ several strategies. A straightforward approach is to pre-order the groups by ζ_g, thereby eliminating the need for the algorithm to establish its own ordering. For instance, arranging them such that $\zeta_1 < \zeta_2 < \cdots < \zeta_G$ can streamline the computation process. We could also develop a novel pruning method that further reduces the parameter space by leveraging the underlying theoretical properties of the problem. We have developed such a method in the Chapter 3.

$$
\begin{aligned}
\min_{(\lambda,\zeta,\delta,\beta,z)} \textstyle\sum_{i=1}^{N} \sum_{t=1}^{T} \quad & \left(y_{it} - \tilde{\lambda}_{ipt} - \tilde{\zeta}_{ip} 1_{D_i} - \tilde{\delta}_i D_{it} - X'\beta\right)^2 \\
& \text{subject to} \\
& |\tilde{\zeta}_{ip} - \zeta_g| \leq z_{ipg} M && \forall i, p, g \\
& |\tilde{\lambda}_{ipt} - \lambda_{gt}| \leq z_{ipg} M && \forall i, p, g, t \\
& |\tilde{\delta}_i - \delta_{g,g'}| \leq z_{i1g} z_{i2g'} M && \forall i \\
& \textstyle\sum_{g=1}^{G} z_{ipg} = G - 1 && \forall i, p \\
& \zeta_{g-1} < \zeta_g && \forall g \in \{2, ..., G\}
\end{aligned}
\tag{1.7}
$$

1.5.2 K-Means

The integer optimization problem remains NP-hard, making it computationally infeasible for large sample sizes. In such cases, we adopt the approach proposed by Bonhomme and Manresa (2015), which includes leveraging K-Means clustering as a practical alternative.

In the K-Means approach, initial parameters are chosen, after which each individual is assigned to the group that minimizes their Sum of Squared Errors (SSE). The group memberships are then fixed, and the parameters are re-optimized. This iterative process continues until the change in SSE between iterations falls below a pre-defined epsilon threshold. For very large datasets, Bonhomme and Manresa incorporate advances in clustering methods, offering a more efficient variant of the K-Means algorithm. For further technical details, see the appendix of their work.

1. Let $(\lambda^s, \zeta^s, \delta^s, \beta^s)$ be some initial values where we set $s = 0$.

2. Compute for all $i = \{1, \ldots, N\}$:

$$(g_{ib}^{s+1}, g_{ia}^{s+1}) = \underset{g_b, g_a}{\text{argmin}} \sum_{t=1}^{T} (y_{it} - \lambda_{g_{it},t}^{s} - \zeta_{g_{it}}^{s} 1_{D_i} - \delta_{g_b,g_a}^{s} D_{it} - X'\beta^s)^2$$

3. Set $(\lambda^{s+1}, \zeta^{s+1}, \delta^{s+1}, \beta^{s+1})$ equal to

$$\underset{(\lambda,\zeta,\delta,\beta)}{\text{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - \lambda_{g_{it}^{s+1},t} - \zeta_{g_{it}^{s+1}} 1_{D_i} - \delta_{g_{ib}^{s+1},g_{ia}^{s+1}} D_{it} - X'\beta)^2$$

4. If $SSE_s - SSE_{s-1} < \epsilon$ where ϵ is some pre-defined constant then stop. Otherwise set $s = s + 1$ and go back to step (2).

1.6 Asymptotic Properties

In this section, we will establish the consistency and asymptotic normality of our parameters. To achieve this, we will use the following strategy. In Theorem 1, we will demonstrate the consistency of the non-group dependent parameters and the linear combination of group-dependent parameters. In Theorem 2, we will leverage Theorem 1 to show uniform consistency in estimating individual group choices over different periods of time. With the ability to uniformly estimate each individual's group choice, we can then prove the consistency of all our TV-GPFE estimators with respect to their infeasible counterparts. Since TV-GPFE converges in probability to the infeasible estimator, it also converges in distribution. Given that the infeasible estimator is simply a panel data estimator, we can easily characterize the asymptotic distribution. To begin, consider the following data generating process:

$$y_{it} = \lambda_{g_{it}^0,t}^0 + \zeta_{g_{it}^0}^0 1_{D_i} + \delta_{g_{ib}^0,g_{ia}^0}^0 D_{it} + x_{it}'\beta^0 + \epsilon_{it} \tag{1.8}$$

Let g_{it}^0 denote the group of individual i at time t, where ^{0} indicates the true value or true group. We assume that the number of groups, G, is known and fixed. Importantly, we allow treated individuals to change groups at the point of treatment, denoted by time T'. This extension goes beyond the current theory proposed by Bonhomme and Manresa (2015) who assumes individuals remain in the same groups. We also extend their theory by bringing it to the difference-in-difference model. Where we will use the controls to connect treated group labels across different periods helping us identify unobserved group change effects. To establish the consistency of the TV-GPFE estimator, we will consider the following assumptions.

Assumption 1. *Assume there exists a constant $M > 0$ such that:*

- *(a)* A, Λ, Z, Δ, B *are compact subsets of* $R^G, R^{GT}, R^G, R^{G^2}, R^P$ *respectfully.*

- *(b)* $E[||x_{it}||^2] \leq M$ *where* $|| \cdot ||$ *denotes the Euclidean Norm.*

- *(c)* $E[\epsilon_{it}] = 0$ $E[\epsilon_{it}^4] \leq M$

- *(d)* $|\frac{1}{NT} \sum_{i=1}^{N} \sum_{s=1}^{T} \sum_{t=1}^{T} E[\epsilon_{it}\epsilon_{is}x_{it}x_{is}]| \leq M$

- *(e)* $\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} | \sum_{t=1}^{T} \frac{1}{T} E[\epsilon_{it}\epsilon_{jt}]| \leq M$

- *(f)* $|\frac{1}{N^2T} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=1}^{T} \sum_{s=1}^{T} Cov[\epsilon_{js}\epsilon_{is}, \epsilon_{jt}\epsilon_{it}]| \leq M$

- *(g) Let* $\bar{x}_{g \wedge \tilde{g}, t}$ *denote the mean of* x_{it} *in the intersection of groups* $g_{it}^0 = g$*, and* $g_{it} = \tilde{g}$*. For all groupings* $\gamma = \{(g_{1b}, g_{1a}), ..., (g_{Nb}, g_{Na})\} \in \Gamma^{G^2}$ *we define* $\hat{\rho}(\gamma)$ *as the minimum eigenvalue of the following matrix:*

$$\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it} - \bar{x}_{g_{it}^0 \wedge g_{it}, t})(x_{it} - \bar{x}_{g_{it}^0 \wedge g_{it}, t})'$$

Then the $\underset{N,T \to \infty}{plim} \hat{\rho}(\gamma) = \rho > 0$

For Assumption 1.a, we assume that our parameter space is compact. Additionally, in Assumption 1.b, we assume that our data is compact. Assumption 1.c posits that our errors are centered at zero and bounded in the fourth moment. Assumption 1.d requires weak dependency over time between the product of errors and data. In Assumption 1.e, we assume weak dependency among individuals for the errors. Assumption 1.f similarly assumes weak dependency over time for the product of errors. Finally, Assumption 1.g is requiring our covariate matrix to have full rank condition. Using these assumptions we can now introduce Theorem 1.

Theorem 1. *Given that Assumption 1 is satisfied, as both N and T tend towards infinity, we observe the following:*

$\hat{\beta} \xrightarrow{p} \beta^0$

$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\lambda^0_{g^0_{it},t} - \hat{\lambda}_{\hat{g}_{it},t} + (\zeta^0_{g^0_{it}} - \hat{\zeta}_{\hat{g}_{it}})1_{D_i} + (\delta^0_{g^0_{ib},g^0_{ia}} - \hat{\delta}_{\hat{g}_{ib},\hat{g}_{ia}})D_{it})^2 \xrightarrow{p} 0$

In this theorem, we demonstrate that we can consistently estimate the non-group-dependent parameters as well as linear combinations of the group-dependent parameters. However, to consistently estimate the individual group-dependent parameters, we must first establish that we can consistently estimate individuals' group assignments over time. Leveraging this result, we further show that individual group assignments can be uniformly estimated for both the pre-treatment and post-treatment periods. To formalize this, we introduce the following assumptions First, define $\theta_{g_{ib},g_{ia},t}(1_{D_i}, D_{it}) = \lambda_{g_{it},t} + 1_{D_i}\zeta_{g_{it}} + D_{it}\delta_{g_{ib},g_{ia}}$.

Assumption 2 (2)**.** *Assume that as T goes to infinity T' grows at some constant rate w.r.t. T.*

-*(a_1)* $\forall\ (g, g') \in \{1, ..., G\}^2$:

$lim\ inf_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N} 1\{D_{iT'} = 1\}1\{g_{ib} = g, g_{ia} = g'\} > 0.$

-*(a_2)* $\forall\ g \in \{1, ..., G\}$:

$lim\ inf_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N} 1\{D_{iT'} = 0\}1\{g_{ib} = g, g_{ia} = g\} > 0.$

-*(a_3)* $\forall (g, g') \in \{1, ..., G\}^2 \quad s.t. \quad (g \neq g')$:

$lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N} 1\{D_{iT'} = 0\}1\{g_{ib} = g, g_{ia} = g'\} = 0.$

-*(b_1) For all $(g, \tilde{g}) \in \{1, \dots, G\}^2$ such that $g \neq \tilde{g}$:*

$lim\ inf_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} 1\{D_{iT'} = 0\}(\theta^0_{g,g,t} - \theta^0_{\tilde{g},\tilde{g},t})^2 > c_{g\tilde{g}} > 0$

-*(b_2) For all $(g_b, g_a, \tilde{g}_b, \tilde{g}_a) \in \{1, \dots, G\}^4$ such that $(g_b \neq \tilde{g}_b)$ or $(g_a \neq \tilde{g}_a)$:*

$lim\ inf_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} 1\{D_{iT'} = 1\}(\theta^0_{g_b,g_a,t} - \theta^0_{\tilde{g}_b,\tilde{g}_a,t})^2 > c_{g_a g_b \tilde{g}_a \tilde{g}_b} > 0$

- *(c) There exist constants $a > 0$ and $d_1 > 0$ and a sequence $a[t] \leq e^{-at^{d_1}}$ such that, for all $i \in \{1, \dots, N\}$ and $(g_b, g_a, \tilde{g}_b, \tilde{g}_a) \in \{1, \dots, G\}^4$ such that $g_b \neq \tilde{g}_b \vee g_a \neq \tilde{g}_a, \{\epsilon_{it}\}_t$, $\{\theta^0_{g_b,g_a,t} - \theta^0_{\tilde{g}_b,\tilde{g}_a,t}\}_t$, and $\{(\theta^0_{g_b,g_a,t} - \theta^0_{\tilde{g}_b,\tilde{g}_a,t})\epsilon_{it}\}_t$ are strongly mixing processes with mixing co-*

efficients $a[t]$. *Moreover,* $\mathbb{E}[(\theta^0_{g_b,g_a,t} - \theta^0_{\tilde{g_b},\tilde{g_a},t})\epsilon_{it}] = 0$.

- *(d) There exist constants* $b > 0$ *and* $d_2 > 0$ *such that* $\Pr(|\epsilon_{it}| > m) \leq e^{1-(m/b)^{d_2}}$ *for all* $i, t,$ *and* $m > 0$.

- *(e) There exists a constant* M^* *such that as* N, T *go to infinity*

$$\sup_{i \in \{1,\ldots,N\}} \Pr\left(\frac{1}{T}\sum_{t=1}^{T} \|x_{it}\| \geq M^*\right) = o(T^{-\xi}) \textit{ for all } \xi > 0.$$

Assumptions 2.a and 2.b outline the partitioning of the data by treatment group and treatment period. Assumption 2.a requires that the probability of an individual being assigned to a group remains positive, ensuring that groups are well-defined and identifiable. Additionally, this assumption holds that controls cannot switch groups across periods. By restricting group switching for controls, we can utilize their time trends to help identify treated individuals who do switch groups, thus allowing us to estimate the effect of group switching due to treatment. Assumption 2.b, meanwhile, ensures that the average squared distance between group fixed effects remains strictly positive as both the sample size N and the time horizon T increase. This condition is necessary because if the fixed points of the groups were to overlap, it would become impossible to distinguish between them. Importantly, we require this condition to hold in both the pre-treatment and post-treatment periods since the group memberships of treated individuals are estimated separately for each period.

In addition to these structural conditions, Assumptions 2.c, 2.d, and 2.e impose restrictions on the dependency structure, error tail behavior, and covariate compactness. Assumption 2.c ensures a decay in dependency structures, which limits the extent to which observations are dependent on each other over time. Assumption 2.d places bounds on the tails of the error distribution, preventing extreme outliers from disproportionately influencing the results. Finally, Assumption 2.e imposes compactness constraints on the covariate space, ensuring that covariates remain well-behaved and bounded over time. Together, these assumptions allow for dynamic variation in group fixed effects while still maintaining the conditions necessary for consistent estimation. With these assumptions, along with the results of Theorem 1, we establish the framework needed for consistent estimation of group choice across different time periods.

Theorem 2. ***Consistent Estimation of Group Choice:*** *Lets assumptions 1 and 2 hold. Then we can show for all $\xi > 0$, $g \in \{1, ..., G\}$ and as N and T tend to infinity:*

$$\begin{aligned} &\Pr\left(\sup_{i\in\{1,\ldots,N\}} |\hat{g}_{ib} - g^0_{ib}| > \xi\right) = o(1) + o\left(NT^{-\xi}\right), \\ &\Pr\left(\sup_{i\in\{1,\ldots,N\}} |\hat{g}_{ia} - g^0_{ia}| > \xi\right) = o(1) + o\left(NT^{-\xi}\right), \end{aligned} \tag{1.9}$$

Now that we have established consistency in the estimation of group choices, it follows naturally that our Time-Varying Group-Period Fixed Effects (TV-GPFE) estimators also converge to their corresponding infeasible estimators. Specifically, these estimators converge at the same rate as the convergence of group choices over time. This occurs because the identification of each individual's group membership is based on an averaging process over time, which ensures that as time increases, the group membership estimation improves, leading to our TV-GPFE parameters converging in probability to our infeasible parameters.

Assumption 3. -

-(a) For all $(g, \tilde{g}) \in \{1, \ldots, G\}^2$ such that $g \neq \tilde{g}$ and For any $q \in R$:

$$lim\ inf_{T\to\infty}\tfrac{1}{T}\textstyle\sum_{t=1}^{T} 1\{D_{iT'} = 0\}(\lambda^0_{gt} - \lambda^0_{\tilde{g}t} - q)^2 > c_{g\tilde{g}} > 0$$

-(b) For all $(g, \tilde{g}) \in \{1, \ldots, G\}^2$ such that $g \neq \tilde{g}$ and For any $q \in R$:

$$lim\ inf_{T\to\infty}\tfrac{1}{T}\textstyle\sum_{t=1}^{T} 1\{D_{iT'} = 1\}(\lambda^0_{gt} - \lambda^0_{\tilde{g}t} - q)^2 > c_{g\tilde{g}} > 0$$

To ensure consistent group matching between treated and control units across pre- and post-treatment periods, an additional condition is required. Specifically, it is necessary to guarantee that group-specific time trends are not simple mean shifts of one another. If this condition is violated, it becomes possible to misalign group time trends when associating treated group intercepts with control group trends. For instance, one could inadvertently match the time trend of group g in the control group with the treated group intercept for $g' \neq g$. This would allow for consistent estimation of both treated and control equations but would result in mislabeling and misestimating the group-specific parameters. This issue is elaborated further in the Identification section. Assumption 3 addresses this concern by ensuring that the time trends exhibit distinct variational differences, enabling correct matching of groups between treated and control units.

Theorem 3. ***Asymptotic Equivalency:*** *Let assumptions 1, 2, and 3 hold. Leverage the results from Theorem 2. Then we can show for all $\xi > 0$, $(g, \tilde{g}) \in \{1, ..., G\}^2$ and as N and T tend to infinity:*

$$\begin{aligned}
\hat{\beta} &= \tilde{\beta} + o(T^{-\xi}) \\
\hat{\zeta}_g &= \tilde{\zeta}_g + o(T^{-\xi}) \\
\hat{\lambda}_{gt} &= \tilde{\lambda}_{gt} + o(T^{-\xi}) \\
\hat{\delta}_{g,\tilde{g}} &= \tilde{\delta}_{g,\tilde{g}} + o(T^{-\xi})
\end{aligned} \tag{1.10}$$

Finally, we would like to uncover the asymptotic distribution of our infeasible estimator. To do this lets define $\mathcal{X}$ as a stacked vector of corresponding indicators and covariates with respect to our parameters. Furthermore lets define $\mathcal{B}$ as the stacked vector of parameters. Then we can describe our model as $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - \mathcal{X}_{it}'\tilde{\mathcal{B}})^2$. Now we will prove that our infeasible estimator $\mathcal{B}$ is consistent and asymptotically normal.

Assumption 4. -

- *(a) For all i, j, and t: $\mathbb{E}(\mathcal{X}_{jt}\epsilon_{it}) = 0$.*

- *(b) There exist positive definite matrices Σ_θ and Ω_θ such that*

$$\Sigma_\theta = \text{plim}_{N,T\to\infty} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathcal{X}_{it}\mathcal{X}_{it}',$$

$$\Omega_\theta = \lim_{N,T\to\infty} \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}[\epsilon_{it}\epsilon_{js}\mathcal{X}_{it}\mathcal{X}_{it}'].$$

- *(c) As N and T go to infinity:*

$$\sqrt{\frac{1}{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathcal{X}_{it}\epsilon_{it}\xrightarrow{d}\mathcal{N}(0,\Omega_{\theta}).$$

Using assumptions 1,2, 3, and 4 we can prove our TV-GPFE parameters are consistent and asymptotically normal.

Theorem 4. *Lets assumptions 1,2, 3, and 4 hold, then as N and T go to infinity at a rate such that for some $v > 0$ we have $\frac{N}{T^{v}} \to 0$.*

$$\sqrt{NT}(\hat{\mathcal{B}}-\mathcal{B}^{0})\xrightarrow{d}N(0,\Sigma_{\beta}^{-1}\Omega_{\beta}\Sigma_{\beta}^{-1}) \tag{1.11}$$

In this section, we extend the framework of Bonhomme and Manresa (2015) to incorporate group-specific parameters that vary across periods and subpopulations. Notably, we allow individuals to transition between groups over time, necessitating modifications to assumptions 2.a and 2.b in their original framework. Specifically, we now assume that groups are defined within each period and subpopulation, with group separation preserved within these contexts. This ensures consistent estimation of group membership within each period and subpopulation, even if the group indices differ across these contexts. We also introduce a new assumption, Assumption 3, which plays a crucial role in aligning group indices across periods and subpopulations to ensure accurate matching of groups. Notably, Assumption 3 suggests that group time trends exhibit persistent variational differences, preventing any mean shifts from overlapping with them. In the following section, we will delve into why this assumption is essential for the proper labeling of groups. A notable advantage of Assumption 3 is that it provides group separation through the group time trends, eliminating the need to assume that group treatment effects are non-zero.

1.7 Treatment Effects and Group Labels

In the previous section, we demonstrated that the parameters of our model can be consistently estimated and provided a characterization of their asymptotic distribution. In the following section, we establish that our treatment parameter can be interpreted as the causal treatment effect. Additionally, we outline the necessary assumptions required for correctly labeling and matching group indices.

1.7.1 Heterogeneous Treatment Effects

We adopt the Rubin potential outcomes framework for our analysis. Specifically, consider the following data generating process, where $Y_{it}(D)$ represents the potential outcome conditional on treatment status D, with $D \in \{0, 1\}$. For instance, $Y_{it}(0)$ denotes the potential outcome for individual i at time t in the absence of treatment. Consistent with our previous assumptions, we will continue to model the observed outcome Y_{it} following equation 1.3.

$$y_{it} = Y_{it}(0) + D_{it}(Y_{it}(1) - Y_{it}(0)) \tag{1.12}$$

We begin by establishing the following assumptions which are standard to the Rubins framework. Assumption 5.a assumes the data is generated according to equation 1.12. Assumption 5.b posits that the probability of receiving treatment prior to the actual treatment period is zero. Meanwhile, Assumption 5.c states that the probability of receiving treatment after the treatment period is a non-negative value.

Assumption 5. -

- *(a)* $\forall i, t$, *the pairs* (y_{it}, D_{it}) *are generated according to equation 1.12.*
- *(b)* $\forall\ i \wedge t < T'$, *the probability* $P(D_{it} = 0) = 1$
- *(c)* $\forall\ i \wedge t \geq T'$, *the probability* $P(D_{it} = 1) \in (0, 1)$.

Next, we impose the standard difference-in-differences assumption of parallel trends. However, our assumption is more specific, we require parallel trends only for treated and control individuals who do not switch groups. This refinement is necessary due to our assumption that control individuals remain in the same group.

Assumption 6. *Within Group Parallel Trends* $\forall g \in \{1, ..., G\}$.

$$E[Y_{\boldsymbol{it}}(0) - Y_{\boldsymbol{it\text{-}1}}(0)|D_{it} = 1, g^0_{it} = g, g^0_{it-1} = g] =$$
$$E[Y_{\boldsymbol{it}}(0) - Y_{\boldsymbol{it\text{-}1}}(0)|D_{it} = 0, g^0_{it} = g, g^0_{it-1} = g]$$

While this assumption is sufficient for estimating the treatment effect for individuals who do not change groups, it does not provide a theoretical framework for determining where treated individuals would have been had they not switched groups. To address this, we introduce an additional assumption. This assumption attributes the entire group-switching effect to the treatment effect parameter for individuals who change groups. Specifically, it restricts control group parameters from being influenced by group-switching effects. One interpretation of this assumption is that treatment induces group changes, and therefore, any group-switching effects should be solely attributed to the treatment effect.

Although this assumption is sufficient for calculating the treatment effect of people who do not change groups, it does create a theoretical control outcome for treated individuals who do change group. For that we require an additional assumption. The following assumption contributes the entirety of the group change effect to the treatent effect parameter. Specifically, it does not allow for a parameter in the controls to partially explain the effect of group changes. One way to interpret this is treatment is causing individuals to change groups therefore the group change effect should only be contributed to the treatment effect. Alternatively, one could relax the treatment parameter to not rely on group changes and instead contribute the group change effect to the control model. In this scenario we could drop assumption 7.

Assumption 7. *Conditional Independence of Untreated Outcomes with Respect to Group*

Changes

$$E[Y_{i,\boldsymbol{t}}(0)|D_{i,t}=1, g^0_{it}=g, g^0_{it-1}=g'] =$$
$$E[Y_{i,\boldsymbol{t}}(0)|D_{i,t}=1, g^0_{it}=g, g^0_{it-1}=g]$$

Using assumptions 5,6,7 and 2.a we can show if groups are known our treatment parameter can be interpreted as the average treatment effect on treated (ATT).

$$E[Y_{i,t}(1)-Y_{i,t}(0)|D_{i,t}=1, g^0_{ia}=g', g^0_{ib}=g]$$

$$\begin{aligned}
&= \underbrace{E[Y_{i,\mathbf{t}}(1)-Y_{i,\textbf{t-1}}(0)|D_{i,t}=1, g^0_{ia}=g', g^0_{ib}=g]}_{\text{Total Difference}} - \\
&\underbrace{E[Y_{i,\mathbf{t}}(0)-Y_{i,\textbf{t-1}}(0)|D_{i,t}=0, g^0_{ia}=g', g^0_{ib}=g']}_{\text{Post Treatment Group Time Trend}} - \\
&(E[Y_{i,\textbf{t-1}}(0)|D_{i,t}=1, g^0_{ia}=g', g^0_{ib}=g'] \\
&\underbrace{E[Y_{i,\textbf{t-1}}(0)|D_{i,t}=1, g^0_{ia}=g, g^0_{ib}=g]}_{\text{Pre Treatment Group Difference}}) \\
&= \delta^0_{g^0_{ib},g^0_{ia}}
\end{aligned}$$

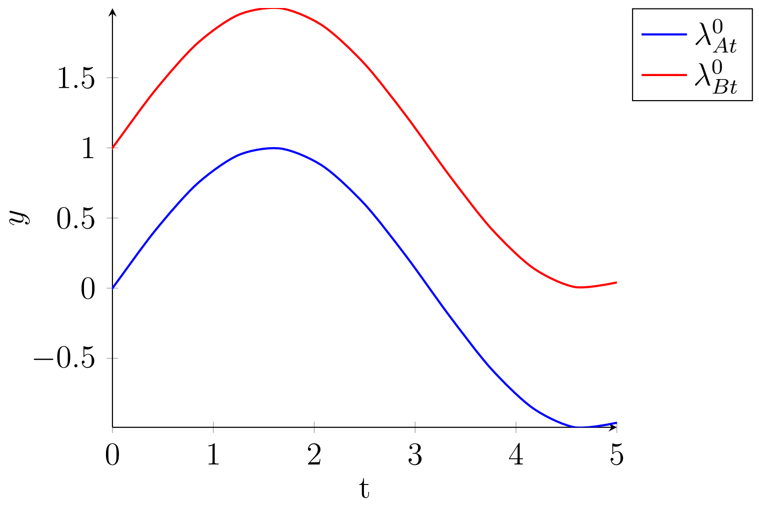
1.7.2 Labeling Groups

In many applications, unobserved group effects are not merely nuisance parameters but are instead the primary effect of interest. For example, if Ayyagari and Shane (2015) had access to complete health insurance data for all individuals, they could estimate the differential impact of Medicaid drug coverage expansion on those who were previously insured versus those who were uninsured. One of the strengths of using difference-in-differences (DiD) is that, by keeping the control group consistent, we can leverage group-specific time trends to separately identify treatment effects related to individuals switching between groups and those who remain in the same group. However, we will need to rely on some additional assumptions, this section outlines the assumptions needed for consistent estimation and identification of heterogeneous treatment effects. It also outlines additional assumptions needed for labeling groups.

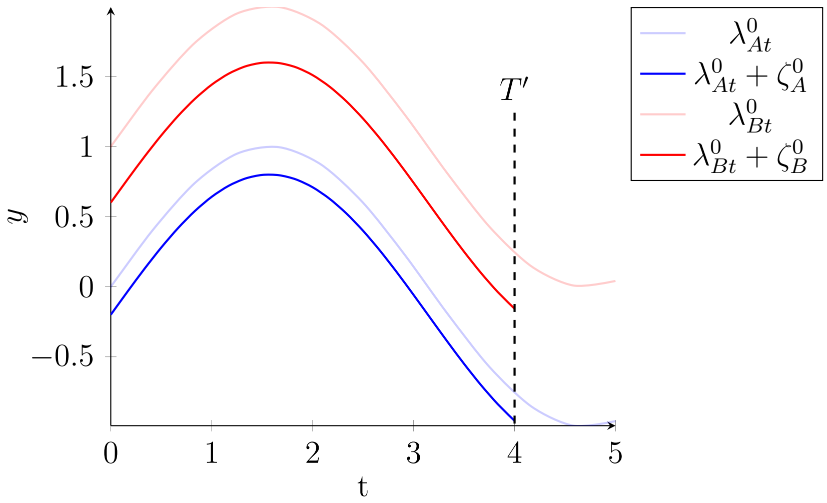
Potential Identification Issue

One of the key strengths of the difference-in-difference framework is its ability to leverage control groups to help identify treated groups. Specifically, we can detect treated individuals whose time trends are similar to those of the controls and group them accordingly. However, additional constraints are necessary to ensure accurate identification of these groups. To illustrate this, consider the following example, where we analyze the time trends of the control group over both the pre-treatment and post-treatment periods.

To illustrate this problem, let's start by considering two control groups, labeled as "Group A" (blue) and "Group B" (red), with distinct time trends. If the difference in time trends between these two groups is constant over time, we can assume that the separation between the groups is valid. This is visualized in the plot below, where the blue and red curves represent the time trends of Groups A and B.



Now, consider adding a treated group before the treatment period. The treated individuals may exhibit an intercept shift in their time trends. Below, we darken the treated group's trends while keeping the control trends lighter. Notice our group separation conditions are still satisfied so we can consistently estimate our parameters.



Here's where the identification problem becomes evident. Suppose we define the mean difference between Group A and Group B's time trends as R, such that $\lambda^0_{BT} = \lambda^0_{AT} + R$. Now, consider that the estimated treatment effect for Group A is $\zeta_B = \zeta^0_A - R$. In this case, it is possible to misclassify a treated individual from Group A as being in Group B due to the overlap in time trends. This misclassification can be demonstrated mathematically:

$$\lambda^0_{BT} + \zeta_B = \lambda^0_{AT} + R + \zeta^0_A - R = \lambda^0_{AT} + \zeta^0_A$$

As shown, the difference between Group A and Group B's post-treatment time trends can vanish, making it impossible to distinguish between the groups based purely on time trends. This highlights the need for additional constraints or information to ensure correct identification of treated individuals in the difference-in-difference framework. Assumption 3 ensures that group time trends differ in a way that prevents perfect alignment. Specifically, this implies that it is not possible to shift the time trends of control group B onto the group equation for treated group A and achieve consistent estimation. Any attempt to do so will result in a non-vanishing error, ensuring that groups cannot be mismatched asymptotically between periods and subpopulations.

1.7.3 Group Labeling

In the previous section, we ensured that the group indices remain consistent between treated and control units across both the pre-treatment and post-treatment periods. However, this consistency does not provide explicit labeling of the groups. To correctly label the groups, we must impose one of the following assumptions.

Monotonic Assumption

In health care research, we frequently encounter situations where insurance status is unobserved, which can introduce violations of the parallel trends assumption when studying outcomes that are affected by insurance status. For instance, consider a study examining the impact of job loss on mental health. If health insurance status is not observed before and after job loss, the loss of insurance could bias the estimated effect of job loss on mental health outcomes. In this context, it may be reasonable to assume that individuals with health insurance exhibit better average health outcomes. Thus, in equation 1.3, the group with the highest value of ζ_g in probability will converge to the insured population.

Location Assumption

Alternatively, if we have prior knowledge of the value associated with one of the groups, and group separation holds, we can leverage this information to identify the groups. For example, in cases where we do not observe who received treatment but know who was eligible, we can assume that individuals who did not receive treatment have a treatment effect of zero, while those who did have a non-zero effect. Consequently, the group with an estimated treatment effect $\hat{\delta}_g \neq 0$ can be identified as the treatment group.

1.8 Conclusion

This paper introduces a novel time-varying unobserved group-period fixed effect estimator (TV-GPFE) within a Difference-in-Difference framework. Our model allows for individuals to change groups over time due to treatment, addressing key limitations of previous models that do not account for these dynamics. By developing both global and local estimators, we offer solutions for both small and large sample sizes, leveraging mixed integer optimization and K-Means clustering.

Our framework addresses critical challenges in causal inference by correcting violations of the parallel trends assumption caused by latent group membership, ensuring more accurate estimation of treatment effects. Additionally, it enables the estimation of the average treatment effect on the treated (ATT) even when compliance with treatment is unobserved. To achieve these advances, we derive the identification assumptions necessary for handling heterogeneous treatment effects and detail the conditions required to consistently label groups across different periods and subpopulations with varying group selection problems. Furthermore, we develop a mixed integer optimization framework, elaborated in Chapters 2 and 3, which facilitates globally optimal estimation in small-sample settings, providing correct confidence intervals for complex estimation challenges.

In conclusion, the TV-GPFE model provides a robust approach for estimating treatment effects in the presence of unobserved, time-varying groups, offering significant improvements in flexibility and precision over traditional methods.

1.9 Appendix

1.9.1 Proof of Theorem 1

Let's begin by defining our time-varying unobserved group period fixed effect estimator. In this definition, $\gamma^0 = (g^0_{b1}, g^0_{a1}), ..., (g^0_{bN}, g^0_{aN}) \in \Gamma_G$ represents the true groupings for each individual for both before and after treatment. Furthermore, γ denotes any specific grouping.

$$\begin{aligned}
\widehat{Q}(\lambda, \zeta, \delta, \beta, \gamma) &= \tfrac{1}{NT} \textstyle\sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - \lambda_{g_{it},t} - \zeta_{g_{it}} 1_{D_i} \\
&\quad -\delta_{g_{ib},g_{ia}} D_{it} - x'_{it}\beta)^2 \\
&= \tfrac{1}{NT} \textstyle\sum_{i=1}^{N} \sum_{t=1}^{T} (\epsilon_{it} + (\lambda^0_{g^0_{it},t} - \lambda_{g_{it},t}) \\
&\quad +1_{D_i}(\zeta^0_{g^0_{it}} - \zeta_{g_{it}}) + D_{it}(\delta^0_{g^0_{ib},g^0_{ia}} - \delta_{g_{ib},g_{ia}}) \\
&\quad +x'_{it}(\beta^0 - \beta))^2
\end{aligned} \tag{1.13}$$

Let's also define an auxiliary estimator under the implicit assumption that our errors and data are independent.

$$\begin{aligned}
\widetilde{Q}(\lambda, \zeta, \delta, \beta, \gamma) &= \tfrac{1}{NT} \textstyle\sum_{i=1}^{N} \sum_{t=1}^{T} ((\lambda^0_{g^0_{it},t} - \lambda_{g_{it},t}) \\
&\quad +1_{D_i}(\zeta^0_{g^0_{it}} - \zeta_{g_{it}}) + D_{it}(\delta^0_{g^0_{ib},g^0_{ia}} - \delta_{g_{ib},g_{ia}}) \\
&\quad +x'_{it}(\beta^0 - \beta))^2 + \tfrac{1}{NT} \textstyle\sum_{i=1}^{N} \sum_{t=1}^{T} \epsilon^2_{it}
\end{aligned} \tag{1.14}$$

Lemma 1: Let assumption 1 hold, and assume that we know the true number of groups G. Additionally, note that individuals are allowed to change at the treatment period T'.

$$\underset{N,T\to\infty}{\text{plim}} \sup_{(\lambda,\zeta,\delta,\beta,\gamma)\in(\Lambda^{GT}, Z^G, \Delta^{G^2}, B^P, \Gamma^G)} |\widehat{Q}(\lambda, \zeta, \delta, \beta, \gamma) - \widetilde{Q}(\lambda, \zeta, \delta, \beta, \gamma)| = 0 \tag{1.15}$$

Proof:

$$\begin{aligned}
\widehat{Q}(\lambda, \zeta, \delta, \beta, \gamma) - \widetilde{Q}(\lambda, \zeta, \delta, \beta, \gamma) &= \tfrac{2}{NT} \textstyle\sum_{i=1}^{N} \sum_{t=1}^{T} \epsilon_{it}(\lambda^0_{g^0_{it},t} - \lambda_{g_{it},t}) \\
&\quad +\epsilon_{it} 1_{D_i}(\zeta^0_{g^0_{it}} - \zeta_{g_{it}}) \\
&\quad +\epsilon_{it} D_{it}(\delta^0_{g^0_{ib},g^0_{ia}} - \delta_{g_{ib},g_{ia}}) \\
&\quad +\epsilon_{it} x'_{it}(\beta^0 - \beta)
\end{aligned}$$

First lets show that $\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\epsilon_{it}x'_{it}(\beta^0-\beta)=o_p(1)$. Using Cauchy Swartz theorem we have

$$\begin{aligned}&(\tfrac{2}{NT}\textstyle\sum_{i=1}^{N}\sum_{t=1}^{T}\epsilon_{it}x'_{it}(\beta^0-\beta))^2\\ &=(\tfrac{2}{N}\textstyle\sum_{i=1}^{N}(\beta^0-\beta)(\frac{1}{T}\sum_{t=1}^{T}\epsilon_{it}x'_{it}))^2\\ &\leq \tfrac{4}{N^2}(\textstyle\sum_{i=1}^{N}||\beta^0-\beta||^2)\times(\sum_{i=1}^{N}(||\frac{1}{T}\sum_{t=1}^{T}\epsilon_{it}x'_{it}||^2)\\ &=(\tfrac{4}{N}\textstyle\sum_{i=1}^{N}||\beta^0-\beta||^2)\times(\frac{1}{N}\sum_{i=1}^{N}(||\frac{1}{T}\sum_{t=1}^{T}\epsilon_{it}x'_{it}||^2)\end{aligned}$$

By assumption 1.a we have $||\beta^0-\beta||^2$ is bounded in probability. By assumption 1.d we have $E[\frac{1}{N}\sum_{i=1}^{N}(||\frac{1}{T}\sum_{t=1}^{T}\epsilon_{it}x'_{it}||^2]\leq\frac{M}{T}$ so by markov inequality this converges in probability to 0 as T goes to infinity. Then by continuous mapping theorem and the properties of O operators we have $\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\epsilon_{it}x'_{it}(\beta^0-\beta)=o_p(1)$

Next lets show $\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\epsilon_{it}(\lambda^0_{g^0_{it},t}-\lambda_{g_{it},t})=o_p(1)$. to do this we will show that $\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\epsilon_{it}\lambda_{g_{it},t}=o_p(1)$ uniformly across our parameter space and this will imply $\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\epsilon_{it}\lambda^0_{g^0_{it},t}=o_p(1)$. lets specify the before and after treatment groups for notational convenience $\lambda_{g_{it},t}=\lambda_{g_{ib}g_{ia}t}$

$$\begin{aligned}&\tfrac{2}{NT}\textstyle\sum_{i=1}^{N}\sum_{t=1}^{T}\epsilon_{it}\lambda_{g_{it},t}\\ &=\tfrac{2}{NT}\textstyle\sum_{g=1}^{G}\sum_{g'=1}^{G}\sum_{i=1}^{N}\sum_{t=1}^{T}\epsilon_{it}\lambda_{gg't}1(g_{ib}=g\wedge g_{ia}=g')\\ &=\tfrac{2}{T}\textstyle\sum_{g=1}^{G}\sum_{g'=1}^{G}\sum_{t=1}^{T}\lambda_{gg't}\frac{1}{N}\sum_{i=1}^{N}\epsilon_{it}1(g_{ib}=g\wedge g_{ia}=g')\end{aligned}$$

Now using Cauchy Swartz inequality for each $(g,g')\in\{1,...,G\}^2$ we can show.

$$\begin{aligned}&(\tfrac{2}{T}\textstyle\sum_{t=1}^{T}\lambda_{gg't}\frac{1}{N}\sum_{i=1}^{N}\epsilon_{it}1(g_{ib}=g\wedge g_{ia}=g'))^2\\ &\leq(\tfrac{2}{T}\textstyle\sum_{t=1}^{T}\lambda^2_{gg't})\times(\frac{2}{T}\sum_{t=1}^{T}(\frac{1}{N}\sum_{i=1}^{N}\epsilon_{it}1(g_{ib}=g\wedge g_{ia}=g'))^2)\end{aligned}$$

By assumption 1a $\frac{2}{T}\sum_{t=1}^{T}\lambda^2_{gg't}$ is bounded. Next lets show the second term converges in probability to 0.

$$\begin{aligned}
&\tfrac{2}{T}\textstyle\sum_{t=1}^{T}(\tfrac{1}{N}\sum_{i=1}^{N}\epsilon_{it}1(g_{ib}=g\wedge g_{ia}=g'))^2\\
&=\tfrac{2}{T}\textstyle\sum_{t=1}^{T}\tfrac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\epsilon_{it}\epsilon_{jt}1(g_{ib}=g\wedge g_{ia}=g')1(g_{jb}=g\wedge g_{ja}=g')\\
&=\tfrac{2}{N^2}\textstyle\sum_{i=1}^{N}\sum_{j=1}^{N}1(g_{ib}=g\wedge g_{ia}=g')1(g_{jb}=g\wedge g_{ja}=g')\tfrac{1}{T}\sum_{t=1}^{T}\epsilon_{it}\epsilon_{jt}\\
&\leq\tfrac{2}{N^2}\textstyle\sum_{i=1}^{N}\sum_{j=1}^{N}|\tfrac{1}{T}\sum_{t=1}^{T}\epsilon_{it}\epsilon_{jt}|\\
&=\tfrac{2}{N^2}\textstyle\sum_{i=1}^{N}\sum_{j=1}^{N}|\tfrac{1}{T}\sum_{t=1}^{T}\epsilon_{it}\epsilon_{jt}-E[\epsilon_{it}\epsilon_{jt}]+E[\epsilon_{it}\epsilon_{jt}]|\\
&\leq\tfrac{1}{N^2}\textstyle\sum_{i=1}^{N}\sum_{j=1}^{N}|\tfrac{1}{T}\sum_{t=1}^{T}\epsilon_{it}\epsilon_{jt}-E[\epsilon_{it}\epsilon_{jt}]|+|\tfrac{1}{T}\sum_{t=1}^{T}E[\epsilon_{it}\epsilon_{jt}]|
\end{aligned}$$

By 1e we have $\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{T}\sum_{t=1}^{T}|E[\epsilon_{it}\epsilon_{jt}]|\leq M$ therefore for our second term above have $\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{T}|\sum_{t=1}^{T}E[\epsilon_{it}\epsilon_{jt}]|\leq\frac{M}{N}$. Thus converges in probability as N goes to infinity. For the first term, by Cauchy Swartz inequality we can show.

$$\begin{aligned}
&(\tfrac{1}{N^2}\textstyle\sum_{i=1}^{N}\sum_{j=1}^{N}|\tfrac{1}{T}\sum_{t=1}^{T}\epsilon_{it}\epsilon_{jt}-E[\epsilon_{it}\epsilon_{jt}]|)^2\\
&\leq\tfrac{1}{N^2}\textstyle\sum_{i=1}^{N}\sum_{j=1}^{N}(\tfrac{1}{T}\sum_{t=1}^{T}\epsilon_{it}\epsilon_{jt}-E[\epsilon_{it}\epsilon_{jt}])^2\\
&=\tfrac{1}{N^2}\textstyle\sum_{i=1}^{N}\sum_{j=1}^{N}\tfrac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}(\epsilon_{it}\epsilon_{jt}-E[\epsilon_{it}\epsilon_{jt}])(\epsilon_{is}\epsilon_{js}-E[\epsilon_{is}\epsilon_{js}])\\
&\leq|\tfrac{1}{T^2N^2}\textstyle\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}cov(\epsilon_{it}\epsilon_{jt},\epsilon_{is}\epsilon_{js})|
\end{aligned}$$

By Assumption 1.g, we have $\left|\frac{1}{TN^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathrm{cov}(\epsilon_{it}\epsilon_{jt},\epsilon_{is}\epsilon_{js})\right|\leq M$. Thus, we can show for the corresponding term that $\left|\frac{1}{T^2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathrm{cov}(\epsilon_{it}\epsilon_{jt},\epsilon_{is}\epsilon_{js})\right|\leq\frac{M}{T}$. Therefore, the first term converges in probability to 0 as $T\to\infty$. Hence, $\frac{2}{T}\sum_{t=1}^{T}\left(\frac{1}{N}\sum_{i=1}^{N}\epsilon_{it}\mathbf{1}(g_{ib}=g\wedge g_{ia}=g')\right)^2=o_p(1)$. Since $\frac{2}{T}\sum_{t=1}^{T}\lambda_{gg't}^2$ is bounded, we also have $\left(\frac{2}{T}\sum_{t=1}^{T}\lambda_{gg't}\frac{1}{N}\sum_{i=1}^{N}\epsilon_{it}\mathbf{1}(g_{ib}=g\wedge g_{ia}=g')\right)^2=o_p(1)$. Thus, we get our desired result $\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\epsilon_{it}\lambda_{g_{it},t}=o_p(1)$. Since the rate of convergence does not depend on the value of λ, the result holds uniformly over all λ, implying $\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\epsilon_{it}\lambda^0_{g^0_{it},t}=o_p(1)$. Combining the previous results, we can conclude that $\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\epsilon_{it}\left(\lambda^0_{g^0_{it},t}-\lambda_{g_{it},t}\right)=o_p(1)$.

Next we will show $\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\epsilon_{it}1_{D_i}(\zeta^0_{g^0_{it}}-\zeta_{g_{it}})=o_p(1)$. To do this we will show $\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\epsilon_{it}1_{D_i}\zeta_{g_{it}}=o_p(1)$ uniformly on the parameter space which implies

$\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\epsilon_{it}1_{D_i}\zeta^0_{g^0_{it}} = o_p(1)$. Lets first explicitly define the before and after groups for notational convenience $\zeta_{g_{it}} = \zeta_{g_{ib},g_{ia}}$.

$$\begin{aligned}&\tfrac{2}{NT}\textstyle\sum_{i=1}^{N}\sum_{t=1}^{T}\epsilon_{it}1_{D_i}\zeta_{g_{it}}\\ &= \textstyle\sum_{g=1}^{G}\sum_{g'=1}^{G}\frac{1}{T}\sum_{t=1}^{T}\zeta_{gg't}\frac{2}{N}\sum_{i=1}^{N}1(g_{ib}=g\wedge g_{ia}=g')1_{D_i}\epsilon_{it}\end{aligned}$$

Then for each $(g,g') \in \{1,...,G\}^2$ by Cauchy Swartz theorem we have

$$\begin{aligned}&(\tfrac{1}{T}\textstyle\sum_{t=1}^{T}\zeta_{gg't}\frac{2}{N}\sum_{i=1}^{N}1(g_{ib}=g\wedge g_{ia}=g')1_{D_i}\epsilon_{it})^2\\ &\leq (\tfrac{1}{T}\textstyle\sum_{t=1}^{T}\zeta^2_{gg't})\times(\frac{1}{T}\sum_{t=1}^{T}(\frac{2}{N}\sum_{i=1}^{N}1(g_{ib}=g\wedge g_{ia}=g')1_{D_i}\epsilon_{it})^2)\end{aligned}$$

By assumption 1.a $(\frac{1}{T}\sum_{t=1}^{T}\zeta^2_{gg't})$ is bounded. So next lets show $(\frac{1}{T}\sum_{t=1}^{T}(\frac{2}{N}\sum_{i=1}^{N}1(g_{ib} = g\wedge g_{ia}=g')1_{D_i}\epsilon_{it})^2)$ converges in probability to 0.

$$\begin{aligned}&(\tfrac{1}{T}\textstyle\sum_{t=1}^{T}(\frac{2}{N}\sum_{i=1}^{N}1(g_{ib}=g\wedge g_{ia}=g')1_{D_i}\epsilon_{it})^2)\\ &= \tfrac{4}{N^2}\textstyle\sum_{i=1}^{N}\sum_{j=1}^{N}1(g_{ib}=g\wedge g_{ia}=g')1(g_{jb}=g\wedge g_{ja}=g')1_{D_i}1_{D_j}\frac{1}{T}\sum_{t=1}^{T}\epsilon_{it}\epsilon_{jt}\\ &\leq \tfrac{4}{N^2}\textstyle\sum_{i=1}^{N}\sum_{j=1}^{N}|\frac{1}{T}\sum_{t=1}^{T}\epsilon_{it}\epsilon_{jt}|\end{aligned}$$

But we already showed in our previous analysis that converges in probability to 0 uniformly. Therefore leveraging our previous analysis we have $\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\epsilon_{it}1_{D_i}(\zeta^0_{g^0_{it}} - \zeta_{g_{it}}) = o_p(1)$

Next lets show $\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\epsilon_{it}D_{it}(\delta^0_{g^0_{ib},g^0_{ia}} - \delta_{g_{ib},g_{ia}})$. To do this we will first show that $\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\epsilon_{it}D_{it}\delta_{g_{ib},g_{ia}} = o_p(1)$ uniformly on the parameter space which will imply $\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\epsilon_{it}D_{it}\delta^0_{g^0_{ib},g^0_{ia}} = o_p(1)$. We assume once your treated your always treated and prior to treatment the treatment effect is 0.

$$\begin{aligned}&\tfrac{2}{NT}\textstyle\sum_{i=1}^{N}\sum_{t=1}^{T}\epsilon_{it}D_{it}\delta_{g_{ib},g_{ia}}\\ &= \textstyle\sum_{g=1}^{G}\sum_{g'=1}^{G}\frac{2}{NT}\sum_{i=1}^{N}1(g_{ib}=g\wedge g_{ia}=g')1_{D_i}\sum_{t=t'}^{T}\epsilon_{it}\delta_{gg't}\end{aligned}$$

From here we see that the proof follows exactly as the previous results. This implies $\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\epsilon_{it}D_{it}(\delta^0_{g^0_{ib},g^0_{ia}} - \delta_{g_{ib},g_{ia}}) = o_p(1)$

□

We have just showed the time-varying unobserved group period fixed effect estimator converges in probability to its auxiliary version where errors explicitly treated as independent from the rest of the model. Next we want to prove that the parameters who do not depend on unobserved groups converge in probability.

Lemma 2 for all $(\lambda, \zeta, \delta, \beta, \gamma) \in (\Lambda^{GT}, Z^G, \Delta^{G^2}, B^P, \Gamma^G)$

$$\widetilde{Q}(\lambda, \zeta, \delta, \beta, \gamma) - \widetilde{Q}(\lambda^0, \zeta^0, \delta^0, \beta^0, \gamma^0) \geq \hat{\rho}||\beta - \beta^0||^2$$

where $\text{plim}_{N,T\rightarrow\infty}\hat{\rho} = \rho > 0$

Proof: Let us denote for every grouping $\gamma = \{(g_{1b}, g_{1a}), ..., (g_{Nb}, g_{Na})\}$.

$$\Sigma(\gamma) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{x}_{g^0_{it}\wedge g_{it},t})(x_{it} - \bar{x}_{g^0_{it}\wedge g_{it},t})'$$

We now arrive at the following result, where the first inequality is derived from the fact that the squared deviations are minimized when calculated from the sample mean.

$$\begin{aligned}
\widetilde{Q}(\lambda, \zeta, \delta, \beta, \gamma) - \widetilde{Q}(\lambda^0, \zeta^0, \delta^0, \beta^0, \gamma^0) \quad &= \tfrac{1}{NT}\textstyle\sum_{i=1}^{N}\sum_{t=1}^{T}((\lambda^0_{g^0_{it},t} - \lambda_{g_{it},t}) \\
&+1_{D_i}(\zeta^0_{g^0_{it}} - \zeta_{g_{it}}) + D_{it}(\delta^0_{g^0_{ib},g^0_{ia}} - \delta_{g_{ib},g_{ia}}) \\
&+x'_{it}(\beta^0 - \beta))^2 \\
&\geq \tfrac{1}{NT}\textstyle\sum_{i=1}^{N}\sum_{t=1}^{T}(\beta^0 - \beta)'\Sigma(\gamma)(\beta^0 - \beta) \\
&\geq \min_{\gamma\in\Gamma_G}(\beta^0 - \beta)'\Sigma(\gamma)(\beta^0 - \beta) \\
&\geq \min_{\gamma\in\Gamma_G}\hat{\rho}(\gamma))||\beta^0 - \beta||^2
\end{aligned}$$

□

To show that $\hat{\beta}$ is consitent for β^0 we have from lemma 1 and the definition of GPFE.

$$\begin{aligned}
\widetilde{Q}(\hat{\lambda}, \hat{\zeta}, \hat{\delta}, \hat{\beta}, \hat{\gamma}) \quad &= \widehat{Q}(\hat{\lambda}, \hat{\zeta}, \hat{\delta}, \hat{\beta}, \hat{\gamma}) + o_p(1) \leq \widehat{Q}(\lambda^0, \zeta^0, \delta^0, \beta^0, \gamma^0) + o_p(1) \\
&= \widetilde{Q}(\lambda^0, \zeta^0, \delta^0, \beta^0, \gamma^0) + o_p(1)
\end{aligned}$$

So since since $\tilde{Q}$ converges to the truth in probability, by lemma 2 and assumption 1g we have $||\beta^0 - \beta||^2 = o_p(1)$ since the minimum eigenvalue $\hat{\rho}(\gamma)$ is assumed to be positive in probability. Lastly we will show the convergence of the quadratic mean.

$$\begin{aligned}
|\widetilde{Q}(\hat{\lambda}, \hat{\zeta}, \hat{\delta}, \hat{\beta}, \hat{\gamma}) - \widetilde{Q}(\hat{\lambda}, \hat{\zeta}, \hat{\delta}, \beta^0, \hat{\gamma})| \quad &= |\tfrac{1}{NT} \textstyle\sum_{i=1}^N \sum_{t=1}^T x'_{it}(\beta^0 - \hat{\beta})(2(\lambda^0_{g^0_{it},t} - \hat{\lambda_{g_{it},t}}) \\
&+21_{D_i}(\zeta^0_{g^0_{it}} - \hat{\zeta_{g_{it}}}) + 2D_{it}(\delta^0_{g^0_{ib},g^0_{ia}} - \hat{\delta_{g_{ib},g_{ia}}}) \\
&+x'_{it}(\beta^0 - \hat{\beta}))| \\
&\leq \tfrac{1}{NT} \textstyle\sum_{i=1}^N \sum_{t=1}^T ||x_{it}||^2 \times ||\beta^0 - \hat{\beta}||^2 \\
&+(4 \sup_{\lambda_t \in \Lambda^G} |\lambda_t|) \times ||x_{it}|| \times ||\beta^0 - \hat{\beta}|| \\
&+(4 \sup_{\zeta \in Z^G} |\zeta|) \times ||x_{it}|| \times ||\beta^0 - \hat{\beta}|| \\
&+(4 \sup_{\delta \in \Delta^{G^2}} |\delta|) \times ||x_{it}|| \times ||\beta^0 - \hat{\beta}||
\end{aligned}$$

which is $o_p(1)$ by assumptions 1a, 1b and consistency of $\hat{\beta}$. Combining this with Lemma 1 and 2 we get

$$\begin{aligned}
\widetilde{Q}(\hat{\lambda}, \hat{\zeta}, \hat{\delta}, \hat{\beta}, \hat{\gamma}) \quad &= \widetilde{Q}(\hat{\lambda}, \hat{\zeta}, \hat{\delta}, \beta^0, \hat{\gamma}) + o_p(1) = \widehat{Q}(\hat{\lambda}, \hat{\zeta}, \hat{\delta}, \beta^0, \hat{\gamma}) + o_p(1) \leq \widehat{Q}(\lambda^0, \zeta^0, \delta^0, \beta^0, \gamma^0) + o_p(1) \\
&= \widetilde{Q}(\lambda^0, \zeta^0, \delta^0, \beta^0, \gamma^0) + o_p(1)
\end{aligned}$$

Therefore $\widetilde{Q}(\hat{\lambda}, \hat{\zeta}, \hat{\delta}, \beta^0, \hat{\gamma}) - \widetilde{Q}(\lambda^0, \zeta^0, \delta^0, \beta^0, \gamma^0) = o_p(1)$ which shows.

$$\begin{aligned}
\widetilde{Q}(\hat{\lambda}, \hat{\zeta}, \hat{\delta}, \beta^0, \hat{\gamma}) - \widetilde{Q}(\lambda^0, \zeta^0, \delta^0, \beta^0, \gamma^0) \quad &= \tfrac{1}{NT} \textstyle\sum_{i=1}^N \sum_{t=1}^T ((\lambda^0_{g^0_{it},t} - \hat{\lambda}_{\widehat{g_{it}},t}) \\
&+1_{D_i}(\zeta^0_{g^0_{it}} - \hat{\zeta}_{\widehat{g_{it}}}) + D_{it}(\delta^0_{g^0_{ib},g^0_{ia}} - \hat{\delta}_{\hat{g}_{ib},\hat{g}_{ia}}))^2 = o_p(1)
\end{aligned} \tag{1.16}$$

This ends the proof of Theorem 1. □

1.9.2 Proof of Theorem 2

In Theorem 2, we begin by demonstrating the consistent estimation of pre- and post-treatment group choices. Building on this result, we then establish the consistency of the Group-Period Fixed Effect (GPFE) estimator in relation to the infeasible estimator. Finally we derive the asymptotic distribution of our infeasible estimator. To formalize this, let us define for notational simplicity:

$$\theta_{g_{ib},g_{ia},t}(1_{D_i}, D_{it}) = \lambda_{g_{it},t} + 1_{D_i}\zeta_{g_{it}} + D_{it}\delta_{g_{ib},g_{ia}},$$

Where the parameters correspond to both treated and untreated groups. Our first step is to establish consistency for the estimated parameters, $\hat{\theta}$, in relation to their true values, θ^0, for both the treated individuals ($D_{iT'} = 1$) and untreated groups. Given that the objective function is invariant to the relabeling of groups, we demonstrate consistency using the Hausdorff distance, d_H, in $\mathbb{R}^{G^2T}$, defined as follows:

$$\begin{aligned} d_H(a,b)^2 = \quad & \max\{\max\nolimits_{(g,g')\in\{1,...,G\}^2}\left(\min\nolimits_{(\tilde{g},\tilde{g}')\in\{1,...,G\}^2}\tfrac{1}{T}\textstyle\sum_{t=1}^{T}(a_{\tilde{g}\tilde{g}'t} - b_{gg't})^2\right), \\ & \max\nolimits_{(\tilde{g},\tilde{g}')\in\{1,...,G\}^2}\left(\min\nolimits_{(g,g')\in\{1,...,G\}^2}\tfrac{1}{T}\textstyle\sum_{t=1}^{T}(a_{\tilde{g}\tilde{g}'t} - b_{gg't})^2\right)\} \end{aligned}$$

Lemma 3: Lets assume 1a-1g and 2a - 2b hold. We will show as N,T go to infinity we have.

$$d_H(\theta^0, \hat{\theta})^2 \xrightarrow{P} 0$$

Proof:

Let's study the terms in the maximum iteratively. We will begin with the first term and demonstrate it for all $(g, g') \in \{1, ..., G\}^2$ in the treated and all $(g, g) \in \{1, ..., G\}$ in the controls.

$$\min_{(\tilde{g},\tilde{g}')\in\{1,...,G\}^2}\frac{1}{T}\sum_{t=1}^{T}(\hat{\theta}_{\tilde{g}\tilde{g}'t}(1_{D_i}, D_{it}) - \theta^0_{gg't}(1_{D_i}, D_{it}))^2 \xrightarrow{p} 0 \qquad (1.17)$$

to prove this we can sum over all individuals segmenting it into treated and control groups.

$$\min_{(\tilde{g},\tilde{g}')\in\{1,\ldots,G\}^2} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\hat{\theta}_{\tilde{g}\tilde{g}'t}(1_{D_i}, D_{it}) - \theta^0_{gg't}(1_{D_i}, D_{it}))^2$$
$$= \sum_{d\in\{0,1\}} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \min_{(\tilde{g},\tilde{g}')\in\{1,\ldots,G\}^2} 1(g^0_{ib} = g \wedge g^0_{ia} = g' \wedge D_{iT'} = d)(\hat{\theta}_{\tilde{g}\tilde{g}'t}(1_{D_i}, D_{it}) - \theta^0_{gg't}(1_{D_i}, D_{it}))^2$$
$$= \sum_{d\in\{0,1\}} \left(\frac{1}{N}\sum_{i=1}^{N} 1(g^0_{ib} = g \wedge g^0_{ia} = g' \wedge D_{iT'} = d)\right)\left(\min_{(\tilde{g},\tilde{g}')\in\{1,\ldots,G\}^2} \frac{1}{T}\sum_{t=1}^{T}(\hat{\theta}_{\tilde{g}\tilde{g}'t}(1_{D_i}, D_{it}) - \theta^0_{gg't}(1_{D_i}, D_{it}))^2\right)$$

By Assumption 2.a, we can disregard the term associated with individuals in the control group who switch groups. This allows us to decompose the problem into two distinct components: treated and control groups.

$$\left(\frac{1}{N}\sum_{i=1}^{N} 1(g^0_{ib} = g \wedge g^0_{ia} = g' \wedge D_{iT'} = 1)\right)\left(\min_{(\tilde{g},\tilde{g}')\in\{1,\ldots,G\}^2} \frac{1}{T}\sum_{t=1}^{T}(\hat{\theta}_{\tilde{g}\tilde{g}'t}(1_{D_i}, D_{it}) - \theta^0_{gg't}(1_{D_i}, D_{it}))^2\right)$$
$$+ \left(\frac{1}{N}\sum_{i=1}^{N} 1(g^0_{ib} = g \wedge g^0_{ia} = g \wedge D_{iT'} = 0)\right)\left(\min_{(\tilde{g},\tilde{g})\in\{1,\ldots,G\}} \frac{1}{T}\sum_{t=1}^{T}(\hat{\theta}_{\tilde{g}\tilde{g}t}(1_{D_i}, D_{it}) - \theta^0_{ggt}(1_{D_i}, D_{it}))^2\right) \tag{1.18}$$

Also by assumption 2.a, for all $(g, g') \in \{1, ..., G\}^2$ we know there is a positive probability for each group combination to exist for both controls and treated so the first term must be positive. Thus its suffice to show for treated individuals.

$$\frac{1}{NT}\sum_{i=1}^{N} \min_{(\tilde{g},\tilde{g}')\in\{1,\ldots,G\}^2} \sum_{t=1}^{T} 1(g^0_{ib} = g \wedge g^0_{ia} = g' \wedge D_{iT'} = 1)(\hat{\theta}_{\tilde{g}\tilde{g}'t}(1_{D_i}, D_{it}) - \theta^0_{gg't}(1_{D_i}, D_{it}))^2 \xrightarrow{p} 0$$

Similarly, by assumption 2.a we know for all $(g, g) \in \{1, ..., G\}$ in the controls the first term is positive so its suffice to show for control individuals.

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \min_{(\tilde{g},\tilde{g})\in\{1,\ldots,G\}} 1(g^0_{ib} = g \wedge g^0_{ia} = g \wedge D_{iT'} = 0)(\hat{\theta}_{\tilde{g}\tilde{g}t}(1_{D_i}, D_{it}) - \theta^0_{ggt}(1_{D_i}, D_{it}))^2 \xrightarrow{p} 0$$

So aggregating 1.18 we can bound our minimization with our least squares group predictions which we showed by theorem 1 to converge in probability to 0.

$$
\begin{aligned}
&\sum_{d\in\{0,1\}} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \min_{(\tilde{g},\tilde{g})\in\{1,\ldots,G\}} 1(g_{ib}^0 = g \wedge g_{ia}^0 = g \wedge D_{iT'} = d)(\hat{\theta}_{\tilde{g}\tilde{g}t}(1_{D_i}, D_{it}) - \theta_{ggt}^0(1_{D_i}, D_{it}))^2 \\
&= \frac{1}{NT}\sum_{i=1}^{N}\left(\min_{(\tilde{g},\tilde{g}')\in\{1,\ldots,G\}^2} \sum_{t=1}^{T} 1(g_{ib}^0 = g \wedge g_{ia}^0 = g')(\hat{\theta}_{\tilde{g}\tilde{g}'t} - \theta_{gg't}^0)^2 \right) \\
&\leq \frac{1}{NT}\sum_{i=1}^{N}\left(\sum_{t=1}^{T} 1(g_{ib}^0 = g \wedge g_{ia}^0 = g')(\hat{\theta}_{\hat{g}\hat{g}'t} - \theta_{gg't}^0)^2 \right) \\
&\leq \frac{1}{NT}\sum_{i=1}^{N}\left(\sum_{t=1}^{T} (\hat{\theta}_{\hat{g}\hat{g}'t} - \theta_{gg't}^0)^2 \right) \xrightarrow{p} 0
\end{aligned}
\tag{1.19}
$$

For the second term of our Hausdorff distance minimization equation let us first define:

$$\sigma(g, g') = \min_{(\tilde{g},\tilde{g}')\in\{1,\ldots,G\}^2} \frac{1}{T}\sum_{t=1}^{T}(\hat{\theta}_{\tilde{g}\tilde{g}'t}(1_{D_i}, D_{it}) - \theta_{gg't}^0(1_{D_i}, D_{it}))^2$$

First we will show $\sigma(g, g')$ is one-to-one with probability approaching 1 as T goes to infinity. let $g \neq \tilde{g}$ or $g' \neq \tilde{g}'$ then by triangle inequality.

$$
\begin{aligned}
\left(\tfrac{1}{T}\textstyle\sum_{t=1}^{T}(\hat{\theta}_{\sigma(g,g')t}(1_{D_i}, D_{it}) - \hat{\theta}_{\sigma(\tilde{g},\tilde{g}')t}(1_{D_i}, D_{it}))^2\right)^{\frac{1}{2}} \geq \quad & \left(\tfrac{1}{T}\textstyle\sum_{t=1}^{T}(\theta_{gg't}^0(1_{D_i}, D_{it}) - \theta_{\tilde{g}\tilde{g}'t}^0(1_{D_i}, D_{it}))^2\right)^{\frac{1}{2}} \\
& - \left(\tfrac{1}{T}\textstyle\sum_{t=1}^{T}(\hat{\theta}_{\sigma(g,g')t}(1_{D_i}, D_{it}) - \theta_{gg't}^0(1_{D_i}, D_{it}))^2\right)^{\frac{1}{2}} \\
& - \left(\tfrac{1}{T}\textstyle\sum_{t=1}^{T}(\hat{\theta}_{\sigma(\tilde{g},\tilde{g}')t}(1_{D_i}, D_{it}) - \theta_{\tilde{g}\tilde{g}'t}^0(1_{D_i}, D_{it}))^2\right)^{\frac{1}{2}}
\end{aligned}
$$

By assumption 2.b lim $\inf_{T\to\infty}\left(\frac{1}{T}\sum_{t=1}^{T}(\theta_{gg't}^0(1_{D_i}, D_{it}) - \theta_{\tilde{g}\tilde{g}'t}^0(1_{D_i}, D_{it}))^2\right)^{\frac{1}{2}} > c_{gg'\tilde{g}\tilde{g}'}$ for controls and treated. This implies By the first part of the proof and the definition of σ we know the latter two terms converge in probability to 0. Hence, $\sigma(g, g') \neq \sigma(\tilde{g}, \tilde{g}')$ as T goes to infinity. Therefore, $\sigma(g, g')$ is one-to-one and admits a well defined inverse. Where the inverse finds the group effect that is closest to the inputted true group effect. Now lets show the second term in d_H converges in probability to zero.

$$\min_{(g,g')\in\{1,\ldots,G\}^2} \frac{1}{T}\sum_{t=1}^{T}(\hat{\theta}_{\tilde{g}\tilde{g}'t}(1_{D_i}, D_{it}) - \theta_{gg't}^0(1_{D_i}, D_{it}))^2 \xrightarrow{p} 0$$

Now with probability approaching 1 we have for all $(\tilde{g}, \tilde{g}') \in \{1, ..., G\}^2$ we have using the results above...

$$\begin{aligned}
&\min_{(g,g')\in\{1,\ldots,G\}^2} \frac{1}{T}\sum_{t=1}^{T}\left(\hat{\theta}_{\tilde{g}\tilde{g}'t}(1_{D_i}, D_{it}) - \theta^0_{gg't}(1_{D_i}, D_{it})\right)^2 \\
&\leq \frac{1}{T}\sum_{t=1}^{T}\left(\hat{\theta}_{\tilde{g}\tilde{g}'t}(1_{D_i}, D_{it}) - \theta^0_{\sigma^{-1}(\tilde{g},\tilde{g}')t}(1_{D_i}, D_{it})\right)^2 \qquad (1.20)\\
&= \min_{(\tilde{h},\tilde{h}')\in\{1,\ldots,G\}^2} \frac{1}{T}\sum_{t=1}^{T}\left(\hat{\theta}_{\tilde{h}\tilde{h}'t}(1_{D_i}, D_{it}) - \theta^0_{\sigma^{-1}(\tilde{g},\tilde{g}')t}(1_{D_i}, D_{it})\right)^2 \xrightarrow{p} 0
\end{aligned}$$

The Lemma shows that their exists a permutation σ such that $\frac{1}{T}\sum_{t=1}^{T}(\hat{\theta}_{\sigma(g,g')t}(1_{D_i}, D_{it}) - \theta^0_{gg't}(1_{D_i}, D_{it}))^2 \xrightarrow{p} 0$. So by relabeling the elements of $\hat{\theta}$ we can take $\sigma(g, g') = (g, g')$ for the rest of the proofs. For any $\eta > 0$ lets define N_η to represent the set of parameters (β, θ) that satisfy $||\beta - \beta^0||^2 < \eta$ and $\frac{1}{T}\sum_{t=1}^{T}(\theta_{gg't}(1_{D_i}, D_{it}) - \theta^0_{gg't}(1_{D_i}, D_{it}))^2 < \eta$ for all (g, g'). Now lets prove that we can consistent estimate group choice.

Lemma B.4: For a small enough $\eta > 0$, we have for all $\delta > 0$ and as N and T tend to infinity,

$$\sup_{(\theta,\beta)\in N_\eta} \frac{1}{N}\sum_{i=1}^{N} 1\left[\hat{g}_{ib}(\theta,\beta) \neq g^0_{ib} \vee \hat{g}_{ia}(\theta,\beta) \neq g^0_{ia}\right] = o_p(T^{-\delta}).$$

PROOF: Before we begin, notice to establish the consistency of our group estimates, we analyze the probability of correctly identifying the true grouping structure. Consider the probability that our least squares estimator selects a particular grouping, denoted by (g, g'). Observe that the probability of selecting this grouping is bounded above by the probability that this grouping exhibits a lower sum of squared errors (SSE) than the true grouping. Formally, we have:

$$P\left((g, g') \text{ is selected}\right) \leq P\left(\text{SSE}_{(g,g')} \leq \text{SSE}_{\text{true}}\right).$$

This inequality holds because the properties of the least squares estimator ensure that the chosen grouping will have at least as small an SSE as the true grouping. However, multiple groupings may achieve an SSE that is as low or lower than that of the true grouping.

$$\begin{aligned} 1\{\hat{g}_{ia}(\theta,\beta)=g \wedge \hat{g}_{ib}(\theta,\beta)=g'\} \quad &\leq 1\{\textstyle\sum_{t=1}^{T}(y_{it}-\theta_{gg't}(1_{D_i},D_{it})-x_{it}'\beta)^2 \\ &\leq \textstyle\sum_{t=1}^{T}(y_{it}-\theta_{g_{ib}^0 g_{ia}^0 t}(1_{D_i},D_{it})-x_{it}'\beta)^2\} \end{aligned}$$

Now we can formalize the probability of our estimator selecting an incorrect grouping, we start by expressing this probability in terms of the sum of squared errors (SSE) like above. Specifically, the probability that the estimator chooses an incorrect grouping is bounded above by the probability that an incorrect grouping has an SSE less than or equal to that of the true grouping. This can be represented as follows:

$$\begin{aligned} \tfrac{1}{N}\textstyle\sum_{i=1}^{N}\mathbf{1}\{\hat{g}_{ib}(\theta,\beta)\neq g_{ib}^0 \vee \hat{g}_{ia}(\theta,\beta)\neq g_{ia}^0\} \quad &= \textstyle\sum_{g=1}^{G}\sum_{g'=1}^{G}\frac{1}{N}\sum_{i=1}^{N} \quad \mathbf{1}\{g_{ib}^0\neq g' \vee g_{ia}^0\neq g\}\times \\ &\qquad \mathbf{1}\{\hat{g}_{ib}(\theta,\beta)=g' \wedge \hat{g}_{ia}(\theta,\beta)=g\} \\ &\leq \textstyle\sum_{g=1}^{G}\sum_{g'=1}^{G}\frac{1}{N}\sum_{i=1}^{N} \quad Z_{igg'}(\theta,\beta) \end{aligned}$$

where we define $Z_{igg'}(\theta,\beta)$ to represent the probability that an incorrect grouping exhibits an SSE less than or equal to that of the true grouping. Formally,

$$\begin{aligned} Z_{igg'}(\theta,\beta) = \quad &\mathbf{1}\{g_{ib}^0\neq g' \vee g_{ia}^0\neq g\}\times \\ &\mathbf{1}\{\textstyle\sum_{t=1}^{T}\left(y_{it}-\theta_{gg't}(1_{D_i},D_{it})-x_{it}'\beta\right)^2 \leq \\ &\textstyle\sum_{t=1}^{T}(y_{it}-\theta_{g_{ib}^0 g_{ia}^0 t}(1_{D_i},D_{it})-x_{it}'\beta)^2\}. \end{aligned}$$

To proceed, we aim to bound $Z_{igg'}(\theta,\beta)$ uniformly over all $(\theta,\beta)\in N_\eta$ by a quantity independent of these parameters. For simplicity of notation, we temporarily omit $(1_{D_i},D_{it})$ from θ in the following steps. Then, for any (θ,β) and for all i, we have...

$$\begin{aligned} Z_{igg'}(\theta,\beta) = \quad &1\{g_{ib}^0\neq g' \vee g_{ia}^0\neq g\}\times \\ &1\{\textstyle\sum_{t=1}^{T}(\theta_{g_{ib}^0 g_{ia}^0 t}-\theta_{gg't})\times(\epsilon_{it}+x_{it}'(\beta^0-\beta)+\theta_{g_{ib}^0 g_{ia}^0 t}^0-\frac{\theta_{g_{ib}^0 g_{ia}^0 t}+\theta_{gg't}}{2})\leq 0\} \\ &\leq \max_{h\neq g\vee h'\neq g'} 1\{\textstyle\sum_{t=1}^{T}(\theta_{hh't}-\theta_{gg't})\times(\epsilon_{it}+x_{it}'(\beta^0-\beta)+\theta_{hh't}^0-\frac{\theta_{hh't}+\theta_{gg't}}{2})\leq 0\} \end{aligned}$$

To analyze the convergence of the final term in probability, we introduce a sequence A_t that closely resembles this term. By constructing an appropriate bound on A_t, we can then rigorously establish its convergence properties. This approach will provide insight into the probabilistic behavior of the final term.

$$\begin{aligned}
A_T &= |\textstyle\sum_{t=1}^{T}(\theta_{hh't} - \theta_{gg't}) \times (\epsilon_{it} + x'_{it}(\beta^0 - \beta) + \theta^0_{hh't} - \frac{\theta_{hh't}+\theta_{gg't}}{2})- \\
&\quad (\theta^0_{hh't} - \theta^0_{gg't}) \times (\epsilon_{it} + \theta^0_{hh't} - \frac{\theta^0_{hh't}+\theta^0_{gg't}}{2})| \\
&\leq |\textstyle\sum_{t=1}^{T}(\theta_{hh't} - \theta_{gg't})\epsilon_{it} - (\theta^0_{hh't} - \theta^0_{gg't})\epsilon_{it}| \\
&\quad +|\textstyle\sum_{t=1}^{T}(\theta_{hh't} - \theta_{gg't})x'_{it}(\beta^0 - \beta)| \\
&\quad +|\textstyle\sum_{t=1}^{T}(\theta_{hh't} - \theta_{gg't})(\theta^0_{hh't} - \frac{\theta_{hh't}+\theta_{gg't}}{2}) \\
&\quad -(\theta^0_{hh't} - \theta^0_{gg't})(\theta^0_{hh't} - \frac{\theta^0_{hh't}+\theta^0_{gg't}}{2})|
\end{aligned} \tag{1.21}$$

Next, let's examine the convergence properties of the terms on the right-hand side of equation 1.21. Focusing on the first term in 1.21, we observe that, by applying the Cauchy-Schwarz Inequality and using our definition of N_η we have...

$$\begin{aligned}
&|\textstyle\sum_{t=1}^{T}(\theta_{hh't} - \theta_{gg't})\epsilon_{it} - (\theta^0_{hh't} - \theta^0_{gg't})\epsilon_{it})| \\
&= |\textstyle\sum_{t=1}^{T}(\theta_{hh't} - \theta^0_{hh't})\epsilon_{it} + (\theta_{gg't} - \theta^0_{gg't})\epsilon_{it})| \\
&\leq \sqrt{\textstyle\sum_{t=1}^{T}(\theta_{hh't} - \theta^0_{hh't})^2}\sqrt{\textstyle\sum_{t=1}^{T}\epsilon^2_{it}} + \sqrt{\textstyle\sum_{t=1}^{T}(\theta_{gg't} - \theta^0_{gg't})^2}\sqrt{\textstyle\sum_{t=1}^{T}\epsilon^2_{it}} \\
&\leq 2\sqrt{T\eta}\sqrt{\tfrac{1}{T}\textstyle\sum_{t=1}^{T}\epsilon^2_{it}} \\
&\leq 2T\sqrt{\eta}\sqrt{\tfrac{1}{T}\textstyle\sum_{t=1}^{T}\epsilon^2_{it}}
\end{aligned}$$

For the second term in 1.21 we can show using the same tricks as before with the knowledge our parameter space is compact $\sum_{t=1}^{T}(\theta_{hh't} - \theta_{gg't})^2 \leq C_1T$...

$$\begin{aligned}
&|\textstyle\sum_{t=1}^{T}(\theta_{hh't} - \theta_{gg't})x'_{it}(\beta^0 - \beta)| \\
&\leq \sqrt{\textstyle\sum_{t=1}^{T}(\theta_{hh't} - \theta_{gg't})^2}\sqrt{\textstyle\sum_{t=1}^{T}(x'_{it}(\beta^0 - \beta))^2} \\
&\leq \sqrt{C_1T}\sqrt{\tfrac{1}{T}\textstyle\sum_{t=1}^{T}(x'_{it}(\beta^0 - \beta))^2} \\
&\leq C_1T\sqrt{\tfrac{1}{T^2}\textstyle\sum_{t=1}^{T}||x_{it}||^2||\beta^0 - \beta||^2} \\
&\leq C_1T\sqrt{\eta}\sqrt{\tfrac{1}{T^2}\textstyle\sum_{t=1}^{T}||x_{it}||^2}
\end{aligned}$$

For the final term in 1.21 given all our parameters are bounded we can show for some constant $C_2 > 0$...

$$\begin{aligned}
&|\textstyle\sum_{t=1}^{T}(\theta_{hh't}-\theta_{gg't})(\theta^0_{hh't}-\frac{\theta_{hh't}+\theta_{gg't}}{2})-\\
&(\theta^0_{hh't}-\theta^0_{gg't})(\theta^0_{hh't}-\frac{\theta^0_{hh't}+\theta^0_{gg't}}{2})|\\
&=\textstyle\sum_{t=1}^{T}(\theta_{hh't}-\theta^0_{hh't}+\theta^0_{gg't}-\theta_{gg't})(\theta^0_{hh't})+\\
&(\theta_{hh't}-\theta^0_{hh't}+\theta^0_{gg't}-\theta_{gg't})(\frac{\theta_{hh't}+\theta_{gg't}}{2})+\\
&(\theta^0_{hh't}-\theta^0_{gg't})(\frac{\theta_{hh't}+\theta_{gg't}}{2}-\frac{\theta^0_{hh't}+\theta^0_{gg't}}{2})\\
&\le\sqrt{\textstyle\sum_{t=1}^{T}(\theta_{hh't}-\theta^0_{hh't})^2}\sqrt{\textstyle\sum_{t=1}^{T}(\theta^0_{hh't})^2}+\\
&\sqrt{\textstyle\sum_{t=1}^{T}(\theta_{gg't}-\theta^0_{gg't})^2}\sqrt{\textstyle\sum_{t=1}^{T}(\theta^0_{hh't})^2}+\\
&\sqrt{\textstyle\sum_{t=1}^{T}(\theta_{hh't}-\theta^0_{hh't})^2}\sqrt{\textstyle\sum_{t=1}^{T}(\frac{\theta_{hh't}+\theta_{gg't}}{2})^2}+\\
&\sqrt{\textstyle\sum_{t=1}^{T}(\theta_{gg't}-\theta^0_{gg't})^2}\sqrt{\textstyle\sum_{t=1}^{T}(\frac{\theta_{hh't}+\theta_{gg't}}{2})^2}+\\
&\sqrt{\textstyle\sum_{t=1}^{T}(\theta_{hh't}-\theta^0_{hh't})^2}\sqrt{\textstyle\sum_{t=1}^{T}(\frac{1}{2}(\theta^0_{hh't}-\theta^0_{gg't}))^2}+\\
&\sqrt{\textstyle\sum_{t=1}^{T}(\theta_{gg't}-\theta^0_{gg't})^2}\sqrt{\textstyle\sum_{t=1}^{T}(\frac{1}{2}(\theta^0_{hh't}-\theta^0_{gg't}))^2}\\
&\le\sqrt{T\eta}\sqrt{C_2T}\\
&\le\sqrt{\eta}C_2T
\end{aligned}$$

Combining these results we can show using CS inequality that for all $(\theta, \beta) \in N_\eta$.

$$A_T \le 2T\sqrt{\eta}\sqrt{\tfrac{1}{T}\textstyle\sum_{t=1}^{T}\epsilon_{it}^2} + C_1T\sqrt{\eta}\sqrt{\tfrac{1}{T^2}\textstyle\sum_{t=1}^{T}||x_{it}||^2} + TC_2\sqrt{\eta}$$

where C_1, C_2 are independent of η and T if we subtract A_T from both sides $Z_{igg'}$ we can show...

$$\begin{aligned}
Z_{igg'}(\theta,\beta) \le \max_{h\neq g\vee h'\neq g'} \quad & 1\{\textstyle\sum_{t=1}^{T}(\theta^0_{hh't}-\theta^0_{gg't})\times(\epsilon_{it}+\theta^0_{hh't}-\frac{\theta^0_{hh't}+\theta^0_{gg't}}{2}) \le \\
& T2\sqrt{\eta}(\tfrac{1}{T}\textstyle\sum_{t=1}^{T}\epsilon_{i,t})^{\frac{1}{2}}) + TC_1\sqrt{\eta}(\tfrac{1}{T}||x_{it}||) + TC_2\sqrt{\eta}\}
\end{aligned}$$

Note that the right hand side doesn't depend on (β, θ) thus it follows that $\sup\limits_{(\beta,\theta)\in N_\eta} Z_{igg'}(\theta, \beta) \le$

$\tilde{Z}_{igg'}$ removing the dependency of β, θ on the function Z.

$$\begin{aligned}\tilde{Z}_{igg'} = \max_{h\neq g \vee h' \neq g'} 1\{ \quad & \textstyle\sum_{t=1}^{T}(\theta^{0}_{hh't} - \theta^{0}_{gg't})\epsilon_{it} + \frac{\sum_{t=1}^{T}(\theta^{0}_{hh't}\theta^{0}_{gg't})^{2}}{2} \leq \\ & T2\sqrt{\eta}(\tfrac{1}{T}\textstyle\sum_{t=1}^{T}\epsilon_{i,t})^{\frac{1}{2}}) + TC_{1}\sqrt{\eta}(\tfrac{1}{T}\textstyle\sum_{t=1}^{T}||x_{it}||) + TC_{2}\sqrt{\eta}\}\end{aligned}$$

As a result

$$\sup_{(\theta,\beta)\in N_{\eta}} \tfrac{1}{N}\textstyle\sum_{i=1}^{N} 1\left[\hat{g}_{ib}(\theta,\beta) \neq g \vee \hat{g}_{ia}(\theta,\beta) \neq g'\right] \leq \tfrac{1}{N}\textstyle\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{Z}_{igg'}$$

Remember, $Z_{igg'}$, represents the probability we mislabel an individuals group choice either before treatment or after treatment. Next we will show the probability of mislabeling goes to 0 as N and T go to infinity. Fix $\tilde{M} > \max(\sqrt{M}, M^{*})$, where M and M^{*} are given by Assumptions 1 and 2e, respectively. Note $E[\epsilon^{2}_{it}] \leq \sqrt{M}$. We have the following.

$$\begin{aligned}Pr(\tilde{Z}_{igg'} = 1) \quad & \leq \textstyle\sum_{h\neq g \vee h' \neq g'} Pr(\quad \sum_{t=1}^{T}(\theta^{0}_{hh't} - \theta^{0}_{gg't})\epsilon_{it} \leq -\frac{\sum_{t=1}^{T}(\theta^{0}_{hh't}+\theta^{0}_{gg't})^{2}}{2} \\ & \qquad +T2\sqrt{\eta}(\tfrac{1}{T}\textstyle\sum_{t=1}^{T}\epsilon_{i,t})^{\frac{1}{2}} \\ & \qquad +TC_{1}\sqrt{\eta}(\tfrac{1}{T}\textstyle\sum_{t=1}^{T}||x_{it}||) \\ & \qquad +TC_{2}\sqrt{\eta}) \\ & \leq \textstyle\sum_{h\neq g \vee h' \neq g'} Pr(\quad \frac{1}{T}\sum_{t=1}^{T}||x_{it}||) \geq \tilde{M}) \\ & \qquad +Pr(\tfrac{1}{T}\textstyle\sum_{t=1}^{T}(\theta^{0}_{hh't} - \theta^{0}_{gg't})^{2} \leq \frac{c_{hh'gg'}}{2}) \\ & \qquad +Pr(\tfrac{1}{T}\textstyle\sum_{t=1}^{T}\epsilon^{2}_{i,t} \geq \tilde{M}) \\ & \qquad +Pr(\textstyle\sum_{t=1}^{T}(\theta^{0}_{hh't} - \theta^{0}_{gg't})\epsilon_{it} \leq -T\frac{c_{hh'gg'}}{4} + T2\sqrt{\eta}\sqrt{\tilde{M}} \\ & \qquad +TC_{1}\sqrt{\eta}\sqrt{\tilde{M}} + TC_{2}\sqrt{\eta})\end{aligned} \tag{1.22}$$

The first term on the right hand side will converge in probability to 0 due to assumption 1.b. To bound the last three terms we will rely on the following lemma. Specifically, we rely on Theorem 6.2 in Rio (2000) [Rio00] whose proof was also outlined by Bonhomme and Manresa (2015).

LEMMA B.5: 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 $\Pr(|z_t| \geq z) \leq e^{-z^b d_2}$, where a, b, d_1, and d_2 are positive constants. Then, for all $z \geq 0$, we have, for all $\delta > 0$,

$$T^\delta \Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T} z_t\right| \geq z\right) \xrightarrow[T\to\infty]{} 0.$$

PROOF: Let $s^2 = \sup_{t>1}(\sum_{s\geq 1} \mathbb{E}|z_t z_s|)$. Note that $s^2 < \infty$ under the condition of Lemma B.5. Let also $d = \frac{d_1 d_2}{d_1+d_2}$. By evaluating inequality (1.7) in Merlevède, Peligrad, and Rio (2011) [MPR11] at $\lambda = T^{\frac{z}{4}}$ and $r = T^{\frac{1}{2}}$, we obtain that there exists a constant $f > 0$ independent of T such that, for all $z > 0$ and $T \geq 1$,

$$\Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T} z_t\right| \geq z\right) \leq 4\left(1 + T^{\frac{1}{2}}\frac{z^2}{16s^2}\right)^{-(\frac{1}{2})T^{\frac{1}{2}}} + \frac{16f}{z}\exp\left(-a\left(\frac{T^{\frac{1}{2}}z}{4b}\right)^d\right).$$

Lemma B.5 directly follows. □

It is crucial to note that this needs to hold for every individual across all time periods. Therefore, the groups must be separated for each treatment group with unique group parameters in each period. Without this separation, we cannot guarantee that the probability will hold. This requirement marks a significant difference from Bonhomme and Manresa. Additionally, we relax their group separation condition by allowing the series to be bounded below in the limit, without requiring the existence of the limit. Specifically, we must establish new group separation and probability assumptions for each period and treatment group with unique group parameters. Next we must utilize results from exponential inequalities.

We now bound the last three terms in 1.22 using this inequality. From Assumptions 1a and 2b we can show where $h \neq g \vee h' \neq g'$ we have for *all treatment groups and periods.* In particular, we need to assume group separation for *each period and subpopulation* with its own unique group parameter.

We now bound the last three terms in 1.22 using Lemma B.5. Lets start with $Pr(\frac{1}{T}\sum_{t=1}^{T}(\theta^0_{hh't} - \theta^0_{gg't})^2 \leq \frac{c_{hh'gg'}}{2})$. Using Fatou's lemma and assumption 2.b we know the following is true for all $h \neq g \vee h' \neq g'$.

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[(\theta^0_{hh't} - \theta^0_{gg't})^2\right] > c_{hh'gg'}.$$

So when T is large enough

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[(\theta^0_{hh't} - \theta^0_{gg't})^2\right] > \frac{2c_{hh'gg'}}{3}.$$

By utilizing Lemma B.5 and defining z_t as $(\theta^0_{hh't} - \theta^0_{gg't})^2 - \mathbb{E}\left[(\theta^0_{hh't} - \theta^0_{gg't})^2\right]$, where z_t satisfies the assumption 1a that its expectation equals zero and also meets the tail conditions for a strongly mixing process as described in 2c. Additionally, by setting $z = \frac{c_{hh'gg'}}{6}$, it can be demonstrated that, for any $\delta > 0$, as T approaches infinity the term converges in probability to 0.

$$\begin{aligned}
&\Pr\left(\tfrac{1}{T}\textstyle\sum_{t=1}^{T}(\theta^0_{hh't} - \theta^0_{gg't})^2 \leq \tfrac{c_{hh'gg'}}{2}\right) \\
&= \Pr\left(\tfrac{1}{T}\textstyle\sum_{t=1}^{T}(\theta^0_{hh't} - \theta^0_{gg't})^2 - \mathbb{E}\left[(\theta^0_{hh't} - \theta^0_{gg't})^2\right] \leq \tfrac{c_{hh'gg'}}{2} - \tfrac{1}{T}\textstyle\sum_{t=1}^{T}\mathbb{E}\left[(\theta^0_{hh't} - \theta^0_{gg't})^2\right]\right) \\
&\leq \Pr\left(\tfrac{1}{T}\textstyle\sum_{t=1}^{T}(\theta^0_{hh't} - \theta^0_{gg't})^2 - \mathbb{E}\left[(\theta^0_{hh't} - \theta^0_{gg't})^2\right] \leq \tfrac{c_{hh'gg'}}{2} - \tfrac{2c_{hh'gg'}}{3}\right) \\
&\leq \Pr\left(\left|\tfrac{1}{T}\textstyle\sum_{t=1}^{T}(\theta^0_{hh't} - \theta^0_{gg't})^2 - \mathbb{E}\left[(\theta^0_{hh't} - \theta^0_{gg't})^2\right]\right| \leq \tfrac{c_{hh'gg'}}{6}\right) \\
&= \Pr\left(\left|\tfrac{1}{T}\textstyle\sum_{t=1}^{T} z_t\right| \leq z\right) \leq o(T^{-\delta})
\end{aligned}$$

Next lets examine $Pr(\frac{1}{T}\sum_{t=1}^{T}\epsilon^2_{i,t} \geq \tilde{M})$. We can apply the strongly mixing conditions to the third term in 1.22 by setting $z_t = \epsilon^2_{it} - \mathbb{E}(\epsilon^2_{it})$ and $z = \tilde{M} - \sqrt{M}$. Note this satisfies the strongly mixing conditions by assumption 1c and 2c yielding for all $\delta > 0$.

$$\Pr\left(\frac{1}{T}\sum_{t=1}^{T}\epsilon^2_{it} \geq \tilde{M}\right) = o(T^{-\delta})$$

Finally, we proceed to bound the remaining term in equation (1.22). To achieve this, let c be defined as the minimum of $c_{hh'gg'}$, from Assumption 2(b), over all cases where $\bar{h} \neq \bar{g}$ or $\tilde{h} \neq \tilde{g}$. Then, we obtain

$$\eta \leq \left(\frac{c}{8(2\sqrt{\tilde{M}} + C_1\sqrt{\tilde{M}} + C_2)} \right)^2 .$$

Consider $z_t = (\theta^0_{hh't} - \theta^0_{gg't})\epsilon_{it}$ and set $z = \frac{c_{hh'gg'}}{8}$. Now utilize assumption 2.c and for η that satisfies the above condition we establish the following for all $\bar{h} \neq \bar{g}$ or $\tilde{h} \neq \tilde{g}$.

$$\begin{aligned}
&\Pr\left(\tfrac{1}{T}\textstyle\sum_{t=1}^{T}(\theta^0_{hh't} - \theta^0_{gg't})\epsilon_{it} \leq -\tfrac{c_{hh'gg'}}{4} + 2\sqrt{\eta}\sqrt{\tilde{M}} + C_1\sqrt{\eta}\sqrt{\tilde{M}} + C_2\sqrt{\eta} \right) \\
&\leq \Pr\left(\tfrac{1}{T}\textstyle\sum_{t=1}^{T}(\theta^0_{hh't} - \theta^0_{gg't})\epsilon_{it} \leq -\tfrac{c_{hh'gg'}}{8} \right) \\
&\leq \Pr\left(\left| \tfrac{1}{T}\textstyle\sum_{t=1}^{T}(\theta^0_{hh't} - \theta^0_{gg't})\epsilon_{it} \right| \geq \tfrac{c_{hh'gg'}}{8} \right) \\
&= \Pr\left(\left| \tfrac{1}{T}\textstyle\sum_{t=1}^{T} z_t \right| \geq z \right) \leq o(T^{-\delta})
\end{aligned}$$

Note that $\{(\theta^0_{hh't} - \theta^0_{gg't})\epsilon_{it}\}_t$ satisfies the tail conditions in assumption 2.d. but we will have to change the b coefficients because $\Pr(|\epsilon_{it}| > \frac{m}{2\sup|\theta|}) \leq e^{1-(m/b)^{d_2}}$. Also note that this only holds for every individual because we are requiring group separation condition for treated and controls in pre- and post- treatment periods. Combining all these results and using assumption 2e we find for η satisfying out previous condition and all $\delta > 0$

$$\begin{aligned}
\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\sum_{g=1}^{G}\sum_{g'=1}^{G} Pr(\tilde{Z}_{igg'} = 1) \leq \quad & (G^4 + G^2) \sup_{i=1,..,N} Pr(\tfrac{1}{T}\textstyle\sum_{t=1}^{T}||x_{it}||) \geq \tilde{M}) \\
& + o(T^{-\delta}) = o(T^{-\delta})
\end{aligned}$$

Finally, we have for all η satisfying the condition above, and all $\delta > 0$, $\varepsilon > 0$. using the markov inequality and the results above we prove Lemma B.4.

$$\Pr\left(\sup_{(\theta,\beta)\in N_\eta} \frac{1}{N}\sum_{i=1}^{N} 1\left[\hat{g}_{ib}(\theta,\beta) \neq g^0_{ib} \vee \hat{g}_{ia}(\theta,\beta) \neq g^0_{ia} \right] > \varepsilon T^{-\delta} \right)$$

$$\leq \Pr\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{g=1}^{G}\sum_{g'=1}^{G} \tilde{Z}_{igg'} > \varepsilon T^{-\delta} \right) \leq \frac{\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{g=1}^{G}\sum_{g'=1}^{G} \tilde{Z}_{igg'} \right)}{\varepsilon T^{-\delta}} = o(1),$$

□

1.9.3 Proof of Theorem 3

Having established that group choice can be consistently estimated, we now proceed to demonstrate that our time-varying unobserved group-period fixed effect estimator converges to the infeasible estimator. We begin by proving this convergence for $\hat{\theta}_{g_{ib},g_{ia},t}(1_{D_i}, D_{it})$. Given that our group choice proof applies to both treated and control groups, specifically including treated units in both pre- and post-treatment periods, we can extend the convergence result to $\hat{\lambda}_{g_{it},t}$, $\hat{\zeta}_{g_{it}}$, and $\hat{\delta}_{g_{ib},g_{ia}}$.

$$\widehat{Q}(\theta,\beta) \quad = \quad \tfrac{1}{NT}\textstyle\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it}-\theta_{\hat{g}_{ib}(\lambda,\zeta,\delta,\beta),\hat{g}_{ia}(\lambda,\zeta,\delta,\beta),t}(1_{D_i},D_{it})-x_{it}'\beta)^2 \tag{1.23}$$

and

$$\tilde{Q}(\theta,\beta) \quad = \quad \tfrac{1}{NT}\textstyle\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it}-\theta_{g_{ib}^0,g_{ia}^0,t}(1_{D_i},D_{it})-x_{it}'\beta)^2 \tag{1.24}$$

Let $\eta > 0$ be small enought to satisfy Lemma B4 then using assumptions 1a-1c and lemma B4 we can see for all $\delta > 0$.

$$\sup_{(\theta,\beta)\in N_\eta}\left|\widehat{Q}(\theta,\beta)-\tilde{Q}(\theta,\beta)\right| = o_p(T^{-\delta}). \tag{1.25}$$

Now, by consistency of $\hat{\beta}$ (Theorem 1) and $\hat{\theta}$ (Lemma B.3), we have, as N and T tend to infinity for all periods and treatment groups,

$$Pr((\hat{\theta},\hat{\beta}) \notin N_\eta) \to 0. \tag{1.26}$$

Similiarly, $(\tilde{\theta},\tilde{\beta})$ are also consistent under the conditions of Theorem 1, we have for all periods and treatment groups.

$$Pr((\tilde{\theta},\tilde{\beta}) \notin N_\eta) \to 0. \tag{1.27}$$

By combining 1.25 and 1.26 we have for all $\delta > 0$ as N and T tend to infinity,

$$\hat{Q}(\hat{\theta}, \hat{\beta}) - \tilde{Q}(\hat{\theta}, \hat{\beta}) = o_p(T^{-\delta}). \tag{1.28}$$

We can prove it with the following...

$$\begin{aligned} Pr\left[\left|\hat{Q}(\hat{\theta}, \hat{\beta}) - \tilde{Q}(\hat{\theta}, \hat{\beta})\right| > \varepsilon T^{-\delta}\right] \; & \leq Pr\left((\hat{\theta}, \hat{\beta}) \notin N_\eta\right) \\ & + Pr\left[\sup_{(\theta,\beta)\in N_\eta}\left|\hat{Q}(\theta, \beta) - \tilde{Q}(\theta, \beta)\right| > \varepsilon T^{-\delta}\right], \end{aligned} \tag{1.29}$$

Which is $o(1)$ by 1.25 and 1.26. Analagously we can combine 1.25 and 1.27 to show

$$\hat{Q}(\tilde{\theta}, \tilde{\beta}) - \tilde{Q}(\tilde{\theta}, \tilde{\beta}) = o_p(T^{-\delta}). \tag{1.30}$$

By combining 1.28 and 1.30 with the the fact our hats consistently estimate our tildes we can show.

$$0 \leq \tilde{Q}(\hat{\theta}, \hat{\beta}) - \tilde{Q}(\tilde{\theta}, \tilde{\beta}) = \hat{Q}(\hat{\theta}, \hat{\beta}) - \hat{Q}(\tilde{\theta}, \tilde{\beta}) + o_p(T^{-\delta}) \leq o_p(T^{-\delta}).$$

It follows that for all periods and treatment groups.

$$\tilde{Q}(\hat{\theta}, \hat{\beta}) - \tilde{Q}(\tilde{\theta}, \tilde{\beta}) = o_p(T^{-\delta}). \tag{1.31}$$

Using 1.31 we can now prove consistency for our parameters. Lets start with β. We get the fourth equality from our first order conditions of least square estimators.

$$
\begin{aligned}
\tilde{Q}(\hat{\theta}, \hat{\beta}) - \tilde{Q}(\tilde{\theta}, \tilde{\beta}) \quad &= \tfrac{1}{NT} \textstyle\sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - \hat{\theta}_{g_{ib}^0, g_{ia}^0, t}(1_{D_i}, D_{it}) - x'_{it}\hat{\beta})^2 - \\
&\quad (y_{it} - \tilde{\theta}_{g_{ib}^0, g_{ia}^0, t}(1_{D_i}, D_{it}) - x'_{it}\tilde{\beta})^2 \\
\\
&= \tfrac{1}{NT} \textstyle\sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - \tilde{\theta}_{g_{ib}^0, g_{ia}^0, t}(1_{D_i}, D_{it}) \\
&\quad -(\hat{\theta}_{g_{ib}^0, g_{ia}^0, t}(1_{D_i}, D_{it}) - \tilde{\theta}_{g_{ib}^0, g_{ia}^0, t}(1_{D_i}, D_{it})) \\
&\quad -x'_{it}\tilde{\beta} - x'_{it}(\hat{\beta} - \tilde{\beta}))^2 \\
&\quad -(y_{it} - \tilde{\theta}_{g_{ib}^0, g_{ia}^0, t}(1_{D_i}, D_{it}) - x'_{it}\tilde{\beta})^2 \\
\\
&= \tfrac{1}{NT} \textstyle\sum_{i=1}^{N} \sum_{t=1}^{T} 2(y_{it} - \tilde{\theta}_{g_{ib}^0, g_{ia}^0, t}(1_{D_i}, D_{it}) - x'_{it}\tilde{\beta}) \times \\
&\quad (-(\hat{\theta}_{g_{ib}^0, g_{ia}^0, t}(1_{D_i}, D_{it}) - \tilde{\theta}_{g_{ib}^0, g_{ia}^0, t}(1_{D_i}, D_{it})) - x'_{it}(\hat{\beta} - \tilde{\beta})) + \\
&\quad (-(\hat{\theta}_{g_{ib}^0, g_{ia}^0, t}(1_{D_i}, D_{it}) - \tilde{\theta}_{g_{ib}^0, g_{ia}^0, t}(1_{D_i}, D_{it})) - x'_{it}(\hat{\beta} - \tilde{\beta}))^2 \\
\\
&= \tfrac{1}{NT} \textstyle\sum_{i=1}^{N} \sum_{t=1}^{T} (-(\hat{\theta}_{g_{ib}^0, g_{ia}^0, t}(1_{D_i}, D_{it}) - \tilde{\theta}_{g_{ib}^0, g_{ia}^0, t}(1_{D_i}, D_{it})) - x'_{it}(\hat{\beta} - \tilde{\beta}))^2 \\
\\
&\geq (\tilde{\beta} - \hat{\beta})' \tfrac{1}{NT} \textstyle\sum_{i=1}^{N} \sum_{t=1}^{T} ((x_{it} - \bar{x}_{g_{ib}^0, g_{ia}^0, t})(x_{it} - \bar{x}_{g_{ib}^0, g_{ia}^0, t})')(\tilde{\beta} - \hat{\beta}) \\
&\geq \hat{\rho} ||\tilde{\beta} - \hat{\beta}||^2
\end{aligned}
\tag{1.32}
$$

by assumption 1g $\hat{\rho} \xrightarrow{p} \rho$ and $\rho > 0$ implying $(\tilde{\beta} - \hat{\beta}) = o(T^{-\delta})$ for all $\delta > 0$. Next lets show this work for $\hat{\theta}_{g_{ib}^0, g_{ia}^0, t}(1_{D_i}, D_{it})$. Next we we show that $\frac{1}{T} \sum_{t=1}^{T} (\hat{\theta}_{g_{ib}^0, g_{ia}^0, t}(1_{D_i}, D_{it}) - \tilde{\theta}_{g_{ib}^0, g_{ia}^0, t}(1_{D_i}, D_{it}))^2 = o(T^{-\delta})$. Lets use 1.32 and 1.31 and assumption 1.b and we can show using the same first order condition logic we applied to the previous equation

$$
\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (-(\hat{\theta}_{g_{ib}^0, g_{ia}^0, t}(1_{D_i}, D_{it}) - \tilde{\theta}_{g_{ib}^0, g_{ia}^0, t}(1_{D_i}, D_{it})))^2 = o(T^{-\delta})
$$

which by 2a implies for all $(g, g') \in \{1, .., G\}^2$

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(-(\hat{\theta}_{gg't}(1_{D_i}, D_{it}) - \tilde{\theta}_{gg't}(1_{D_i}, D_{it})))^2 = o(T^{-\delta})$$

Because of our our tail assumption and lemma B.5 we see that our sequence converges for each t.

$$(\tilde{\theta}_{gg't}(1_{D_i}, D_{it}) - \hat{\theta}_{gg't}(1_{D_i}, D_{it})) = o(T^{1-\delta})$$

But since $\delta > 0$ we can generalize this to be...

$$(\tilde{\theta}_{gg't}(1_{D_i}, D_{it}) - \hat{\theta}_{gg't}(1_{D_i}, D_{it})) = o(T^{-\delta})$$

Recall we defined $\theta_{g^0_{ib},g^0_{ia},t}(1_{D_i}, D_{it}) = \lambda_{g^0_{it},t} + 1_{D_i}\zeta_{g^0_{it}} + D_{it}\delta_{g^0_{ib},g^0_{ia}}$. We know that our θ function must converge for each subgroup for both pre and post treatment periods. To estimate our group time trends lets use the controls where $D_{iT'} = 0$. Notice we are estimating each groups time trends for both pre and post treatment since controls do not change groups.

$$\tilde{\lambda}_{gt} - \hat{\lambda}_{gt} = o(T^{-\delta})$$

Given that our time trends and group selection are consistently estimated, we can proceed by using the treated units in the pre-treatment period to estimate the treated group intercepts (i.e., when $D_{iT'} = 1$ and $t < T'$). However, its possible to have an identification issue as there is the potential issue of mislabeling the groups. In particular, if for some $g \neq g'$ we have ${\lambda^0}_{gt} = c + {\lambda^0}_{g't}$ then its possible to match the group time trend to the wrong treatment group as explained in the identification section. By leveraging assumption 3 we ensure that $\lambda^0_{gt} - \zeta^0_g - \hat{\lambda}_{g't} - \hat{\zeta}_{g'} \neq o(T^{-\delta})$ by mis-estimating $\hat{\zeta}_{g'}$ thus guaranteeing:

$$\tilde{\lambda}_{gt} + \tilde{\zeta}_g - \hat{\lambda}_{gt} - \hat{\zeta}_g = o(T^{-\delta})$$

This approach allows for the consistent estimation of the treated group intercepts, ζ_g, for each group.

Theorem 2 ensures that individuals facing the same set of parameters are correctly grouped. However, it does not guarantee consistent labels across sub-groups with different unobserved group parameters. For example, it doesn't ensure that group indices align between treated and control groups, which could lead to misinterpreted treatment effects. To ensure consistent labeling between treated and control groups, we rely on an additional assumption: the time trends of the groups should not merely differ by a mean shift. As long as this assumption holds, we achieve consistent labeling using the least squares framework. Further discussion on this is provided in the identification section.

$$\tilde{\zeta}_g - \hat{\zeta}_g = o(T^{-\delta})$$

With the consistent estimation of the previous parameters for our group, we can estimate our group-dependent treatment effects using the post-treated group. We show that we can consistently estimate our treatment effect parameter (i.e., when $D_{iT'} = 1$ for $t \geq T'$).

$$\tilde{\lambda}_{gt} + \tilde{\zeta}_g + \tilde{\delta}_{gg'} - \hat{\lambda}_{gt} - \hat{\zeta}_g - \hat{\delta}_{gg'} = o(T^{-\delta})$$

$$\tilde{\delta}_{gg'} - \hat{\delta}_{gg'} = o(T^{-\delta})$$

□

1.9.4 Proof of Theorem 4

In Theorem 4 we showed our time varying group period fixed effect estimators parameters converges in distribution to their respective parameters in the infeasible estimator. Next lets characterize the distribution of our infeasible estimator which is a panel data model with multiple time periods. Let $1_{g_{i,t} \times D_i}$ be a $G \times 1$ vector indicating which group i is in at time t if i is treated otherwise its all zeros. Let $1_{g_{it},t}$ be a $GT \times 1$ vector indicating which group i is in at time t. Let $1_{g_i^2}$ be a $G \times G$ vector indicating which group i before and after treatment.

Then we have...

$$\begin{aligned}
&\tfrac{1}{NT}\textstyle\sum_{i=1}^{N}\sum_{t=1}^{T} && (y_{it} - (\tilde{\zeta}_{g_{it}^0} 1_{D_i} + \\
& && \tilde{\lambda}_{g_{it}^0,t} + \tilde{\delta}_{g_{ib}^0,g_{ia}^0} D_{it} + x_{it}'\tilde{\beta}))^2 \\
&= \tfrac{1}{NT}\textstyle\sum_{i=1}^{N}\sum_{t=1}^{T} && (y_{i,t} - \begin{pmatrix} 1_{g_{i,t}\times D} \\ 1_{g_{i,t},t} \\ 1_{g_i^2} \\ x_{it} \end{pmatrix}' \begin{pmatrix} \tilde{\zeta} \\ \tilde{\lambda} \\ \tilde{\delta} \\ \tilde{\beta} \end{pmatrix})^2 \\
&= \tfrac{1}{NT}\textstyle\sum_{i=1}^{N}\sum_{t=1}^{T} && (y_{i,t} - \mathcal{X}_{i,t}'\tilde{\mathcal{B}})^2
\end{aligned} \tag{1.33}$$

Using Assumption 4 we will show can show $\sqrt{NT}(\tilde{\mathcal{B}} - \mathcal{B}^0) \xrightarrow{d} N(0, \Sigma_\beta^{-1}\Omega_\beta\Sigma_\beta^{-1})$. First lets decompose...

$$\begin{aligned}
\sqrt{NT}(\tilde{\mathcal{B}} - \mathcal{B}^0) &= \sqrt{NT}((\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}'Y - \mathcal{B}^0) \\
&= (\tfrac{1}{NT}\mathcal{X}'\mathcal{X})^{-1}\tfrac{1}{\sqrt{NT}}\mathcal{X}'\epsilon
\end{aligned}$$

By assumption 4.c we know that $\frac{1}{\sqrt{NT}}\mathcal{X}'\epsilon = \sqrt{\frac{1}{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathcal{X}_{it}'\epsilon_{it} \xrightarrow{d} \mathcal{N}(0, \Omega_\theta)$. By assumption 4.b we can show $\Sigma_\theta = \text{plim}_{N,T\to\infty}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathcal{X}_{it}\mathcal{X}_{it}' = \frac{1}{NT}\mathcal{X}'\mathcal{X}$ and is invertible. Then by Slutsky theorem we have we can characterize the distribution of our infeasible parameters.

$$\sqrt{NT}(\tilde{\mathcal{B}} - \mathcal{B}^0) = (\tfrac{1}{NT}\mathcal{X}'\mathcal{X})^{-1}\tfrac{1}{\sqrt{NT}}\mathcal{X}'\epsilon \xrightarrow{d} N(0, \Sigma_\beta^{-1}\Omega_\beta\Sigma_\beta^{-1})$$

Since theorem 2 tells us our TV-GPFE parameters converge in probability, therefore distribution, to our infeasible parameters we have $(\hat{\mathcal{B}} - \tilde{\mathcal{B}}) = o(T^{-\delta})$. Therefore we can characterize the distribution of our TV-GPFE parameters using the same distribution.

$$\sqrt{NT}(\hat{\mathcal{B}} - \mathcal{B}^0) = (\tfrac{1}{NT}\mathcal{X}'\mathcal{X})^{-1}\tfrac{1}{\sqrt{NT}}\mathcal{X}'\epsilon \xrightarrow{d} N(0, \Sigma_\beta^{-1}\Omega_\beta\Sigma_\beta^{-1}) \tag{1.34}$$

In this formulation we are letting T go to infinity so the traditional estimator of Ω which treats T as fixed doesn't work. We can instead use Hansen (2006) who allowed for $T \to \infty$

to estimate Ω.

$$\hat{\Omega}_\beta = \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\hat{\epsilon}_{it}\hat{\epsilon}_{js}\mathcal{X}_{it}\mathcal{X}_{it}'$$

□

CHAPTER 2

Mixed Integer Optimization Formulation For Grouped Heterogeneity in Panel Data

2.1 Introduction

Integer optimization problems are widespread in economics, with applications ranging from subset selection in high-dimensional settings to portfolio selection and bundling decisions. Due to the NP-hard nature of these problems, economists have traditionally relied on local approximations, structural assumptions, or convex relaxations to render the problems tractable. However, recent advances in integer optimization algorithms have significantly improved solution efficiency, opening up new possibilities for tackling previously intractable problems. In this chapter, I develop a mixed integer optimization approach to address the issue of Grouped Heterogeneity in Panel Data. By leveraging these algorithmic advancements, I demonstrate that we can now estimate complex models in mere seconds.

2.2 Background

From 1991 to 2015, both hardware and algorithmic advancements contributed to a remarkable 450-billion-fold increase in solution efficiency for integer optimization problems [BKM16]. These improvements have continued with modern optimization platforms, such as Gurobi, with its 2024 version achieving nearly twice the speed compared to its 2016 version. Additionally, emerging technologies, including quantum computing, enhanced multi-threading capabilities, and novel algorithmic developments—such as those proposed in this chapter—suggest that this trend of increasing computational efficiency is likely to persist. These advancements hold substantial promise for integer optimization ability to tackle increasingly complex problems in the years to come.

Bertsimas et. al. (2016) demonstrated that integer optimization could effectively be used to solve subset selection problems for small to medium-sized problems with high precision, as formulated below:

$$\min_{\beta} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_0, \tag{2.1}$$

where the L_0-norm $\|\beta\|_0$ directly penalizes the number of non-zero coefficients, thus encouraging a sparse solution. However, in practical applications, the field has largely adopted the Lasso as an alternative approach for subset selection. The Lasso reformulates the problem by substituting the L_0-norm constraint with the L_1-norm constraint, transforming the problem into a convex optimization problem, which can be solved more efficiently, particularly in high-dimensional datasets:

$$\min_{\beta} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1. \tag{2.2}$$

This approximation is widely used due to its balance between computational efficiency and sparsity, despite its limitations in exact subset selection. Notably, Bertsimas et al. (2016) showed that modern advances in integer optimization not only make it feasible to solve small to medium-sized problems using the Best Subset Selector but also demonstrate that this approach outperforms the Lasso in terms of identifying the correct sparsity structure and predictive performance. This finding suggests that the field may benefit from a renewed focus on mixed integer optimization techniques rather than relying solely on relaxation-based methods.

In this chapter, we address the problem of unobserved group heterogeneity, specifically within the context of fixed effects models, where individuals make unobserved group choices, and these unobserved group effects introduce bias in parameter estimation. This problem was most recently developed by Bonhomme and Manresa (2015), who formulated it in terms of a panel dataset where both the model parameters and potential group memberships are simultaneously estimated. The objective function for this model is given by:

$$(\hat{\theta}, \hat{\alpha}, \hat{\gamma}) = \underset{(\theta,\alpha,\gamma)\in\Theta\times A^{GT}\times\Gamma_G}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T} \left(y_{it} - x_{it}'\theta - \alpha_{g_i t}\right)^2, \tag{2.3}$$

Where y_{it} represents the dependent variable, x_{it} the vector of covariates, θ the coefficients, and $\alpha_{g_{it}}$ the group-specific fixed effect. Due to the combinatorial nature of searching over all possible group assignments, solving this optimization problem through exhaustive search quickly becomes computationally infeasible. Bonhomme and Manresa (2015) addressed this challenge by using a K-means clustering algorithm to estimate individual group assignments iteratively. Their iterative procedure is summarized in Algorithm 1 below.

Algorithm 1:

1. Let $(\theta^{(0)}, \alpha^{(0)}) \in \Theta \times A^{GT}$ be some initial values, and set $s = 0$.

2. For each $i \in \{1, \dots, N\}$, compute:

$$g_i^{(s+1)} = \arg \min_{g\in\{1,\dots,G\}} \sum_{t=1}^{T} \left(y_{it} - x_{it}'\theta^{(s)} - \alpha_{gt}^{(s)}\right)^2. \tag{2.4}$$

3. Update the parameter estimates by solving:

$$\left(\theta^{(s+1)}, \alpha^{(s+1)}\right) = \arg \min_{(\theta,\alpha)\in\Theta\times A^{GT}} \sum_{i=1}^{N}\sum_{t=1}^{T} \left(y_{it} - x_{it}'\theta - \alpha_{g_i^{(s+1)}t}\right)^2. \tag{2.5}$$

4. Increment s by 1 and repeat from Step 2 until numerical convergence.

While this iterative procedure is computationally feasible, it is not guaranteed to converge to the global optimum, as it can get trapped in local minima. In this chapter, we develop a mixed integer optimization formulation for this problem that enables us to compute the global minimum, thereby improving the robustness of parameter estimates. We demonstrate that, under certain distributional assumptions, this global approach yields substantially better results, offering more accurate estimates of group heterogeneity and reducing bias in the estimation process.

2.3 Literature Review

Clustering algorithms have long been used to classify data based on minimizing a sum of squares objective function. One of the earliest formalizations of this approach was Lloyd's (1957) work on the continuous K-means algorithm [Llo82]. Since then, K-means has become a cornerstone in the clustering literature, with numerous extensions (see Bock 2008 for a comprehensive review [Boc08]).

Despite its popularity, K-means suffers from several well-documented limitations: it is sensitive to initialization, struggles with non-spherical cluster shapes, and is highly affected by outliers. Ahmed, Seraj, and Islam (2020) provide an extensive review of these shortcomings [ASI20]. Addressing these challenges requires alternative methods with better finite-sample performance for clustering group data.

Recent efforts to address the limitations of traditional clustering include the work of Chetverikov and Manresa (2021) [CM22], which extends Bonhomme and Manresa's framework. They propose a model where covariates are structured as:

$$x_{it} = \sum_{m=1}^{M} \rho_{im} \alpha^{m}_{g(i)t} + z_{it},$$

where z_{it} is a zero-mean component independent of group-time-specific effects. Under these assumptions, they develop a *Spectral Estimator*, which is computationally efficient and easy to implement. However, the method's reliance on structural constraints for the covariates may limit its applicability in more general settings.

Chu (2017) [Chu17] takes a different approach by introducing a *Composite Quasi-Likelihood (CQL) Estimator* for dynamic panel data models with unobserved group heterogeneity and spatially dependent errors. This method offers three key contributions: (1) it integrates parameter estimation with latent group classification to address unobserved heterogeneity, (2) it remains robust to misspecified group structures, avoiding reliance on instrumental variables, and (3) it achieves computational feasibility through iterative updates, even in the presence of non-convexity. These innovations enhance both estimation efficiency and classification accuracy, providing practical solutions for empirical applications in complex panel data settings.

However, these methods either require structural assumptions or are local approximations. Instead, we explore a *Mixed Integer Optimization (MIO)* approach which gives exact solutions without any additional structural assumptions. The use of MIO in least squares problems has a rich history, first outlined by Lazimy (1957) [Laz82], who formulated the problem as a mixed integer quadratic programming problem. While MIO problems are NP-hard, advances in algorithms such as *Branch-and-Bound* (see Lawler 1966 for a review [LW66]) make them computationally feasible for certain problems.

MIO is not new to economics. For example, Mansini, Ogryczak, and Speranza (2015) [MOS15] extensively discussed its applications in portfolio optimization, and Parisio and Glielmo (2011) [PG11] used MIO to solve economic scheduling problems. We will be focus on using MIO to cluster linear models.

In the context of clustering, Burgard, Pinheiro, and Schmidt (2023) [BPS24] employed MIO to solve support vector machinery problems under the assumption of linearly separable classes. While promising, this approach requires strong structural assumptions about the data-generating process to ensure the decision boundary is correct.

A similar approach to our work is the Mixed Integer Non-Linear Programming (MINLP) formulation of the Minimum-Sum-of-Squares Clustering (MSSC) problem. Burgard et al. (2023) [BMH23] extended this formulation with novel heuristics to improve computational efficiency. Our approach builds on this work but differs by leveraging a linear model for clustering rather than minimizing to centroids. This added structure allows for more efficient heuristics, which will be detailed in Chapter 3.

An additional consideration is clustering in high-dimensional spaces. Adding irrelevant covariates can degrade clustering performance due to the *curse of dimensionality.* While we focus on low-dimensional settings in this chapter, Chapter 3 will address the high-dimensional case and propose a potential solution in the appendix. There is a broad literature discussing alternative solutions to high-dimensional clustering problems (see Sim et al. 2013 [SGZ13] and Souvenir et al. 2005 [SP05]), which may be worth exploring further.

Our approach in Chapter 2 but in particular Chapter 3 will provide a computational feasible global optimization procedure for clustering groups in linear regressions even for large N. We believe this approach provides more attractive finite sample properties than the others currently adopted in the literature. This is primarily due to its ability to find the exact solution without additional structural constraints.

2.4 The Framework

Following the methodology of Bonhomme and Manresa (2015), we address the challenge of modeling unobserved grouped heterogeneity in panel data settings, where the objective is to capture the effect of latent group structures on observed outcomes. Specifically, we aim to estimate three sets of parameters through the following optimization problem:

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

In this framework, θ represents the vector of coefficients associated with the observed covariates x_{it}, while α denotes a set of group-specific fixed effects, with $\alpha_{g_i t}$ capturing unobserved heterogeneity across groups. Additionally, γ comprises the parameters that govern the assignment of each observation to one of G latent groups. This framework is particularly suited to settings where unobserved factors affect the outcome systematically across groups. By introducing group-level fixed effects, it provides a means to improve predictive accuracy and to address biases arising from omitted variables that vary by group, allowing us to differentiate within-group variation from across-group variation more precisely.

A critical aspect of this approach is that it requires consistent and asymptotically normal estimates of the parameters (θ, α, γ), which are theoretically achievable if the optimization algorithm finds the global minimum of the objective function. However, identifying the global minimum is computationally challenging due to the combinatorial complexity of assigning individual observations to groups. Bonhomme and Manresa (2015) addressed this challenge by employing a K-means clustering algorithm, which offers computational efficiency and demonstrated robustness in simulation settings. Nevertheless, the K-means-based approach does not guarantee convergence to the global minimum, which may introduce biases in inference, particularly in constructing confidence intervals for parameters.

In this chapter, we advance this methodology by developing a novel mixed integer optimization formulation that is guaranteed to find the global minimum. By ensuring convergence to the global minimum, this approach enables accurate estimation of confidence intervals, offering improved reliability and robustness in inference for models with unobserved grouped heterogeneity. This formulation enhances the methodological rigor of grouped fixed-effects models in panel data, ensuring both computational feasibility and theoretical consistency in parameter estimation.

2.5 Mixed Integer Optimization Formulation

In this section, we address the challenge of estimating unobserved group-period effects in a time-varying fixed effect framework by reformulating the estimation problem as a mixed integer optimization problem. The goal is to minimize the residual sum of squares while accurately capturing the grouping structure and the associated fixed effects. The problem is initially defined as:

$$(\hat{\theta}, \hat{\alpha}, \hat{\gamma}) = \underset{(\theta,\alpha,\gamma)\in\Theta\times A^{GT}\times\Gamma_G}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(y_{it} - x_{it}'\theta - \alpha_{g_i,t}\right)^2, \tag{2.6}$$

To transform this into a mixed integer optimization problem that explicitly handles discrete group assignments, we introduce binary indicators to represent group membership dynamically and reformulate the model as follows:

$$
\begin{aligned}
\min_{(\theta,\alpha,\tilde{\alpha},z)} \textstyle\sum_{i=1}^{N} \sum_{t=1}^{T} \quad & (y_{it} - x_{it}'\theta - \tilde{\alpha}_{it})^2 \\
& \text{subject to} \\
& \tilde{\alpha}_{it} - \alpha_{gt} \leq z_{ig} M \qquad \forall i, g \\
& \tilde{\alpha}_{it} - \alpha_{gt} \geq -z_{ig} M \quad \forall i, g \\
& \textstyle\sum_{g=1}^{G} z_{ig} = G - 1 \qquad \forall i,
\end{aligned}
\tag{2.7}
$$

where θ is the coefficient vector for the covariates, and $\tilde{\alpha}_{it}$ denotes the individual group-period fixed effect. We introduce the binary variable z_{ig}, which takes the value 1 if individual i is assigned to group g and 0 otherwise, to enforce discrete group assignments. The parameter M is a sufficiently large constant to ensure computational feasibility in enforcing group constraints.

In this setup, the constraint $\tilde{\alpha}_{it} - \alpha_{gt} \leq z_{ig}M$ and $\tilde{\alpha}_{it} - \alpha_{gt} \geq -z_{ig}M$ ensures that $\tilde{\alpha}_{it}$ takes on the group-period fixed effect α_{gt} only when $z_{ig} = 0$. When $z_{ig} = 1$, this constraint is non-binding due to the large constant M, satisfying the inequality trivially. To ensure that each individual is assigned to exactly one group, we impose the constraint $\sum_{g=1}^{G} z_{ig} = G - 1$, which acts as a selection mechanism for group membership.

This integer optimization formulation provides a structured approach to estimating time-varying group effects. By embedding group-period effects into a discrete optimization framework, we introduce flexibility in handling dynamic group memberships while minimizing the residual variance. The mixed-integer nature of the problem aligns well with branch-and-bound techniques, which facilitate efficient exploration of feasible solutions.

2.5.1 Branch and Bound

While an exhaustive search is theoretically possible, it quickly becomes impractical. For instance, if we had a problem with 50 individuals across 2 groups, your typical computer would take years to solve this problem. However, by employing a Branch and Bound approach, we can solve this problem within seconds.

Our method employs a Branch and Bound framework, which systematically explores group assignments by branching over individual group choices and bounding the problem using the current set of assignments. Each node in the tree represents the group assignment for an individual i, while each branch corresponds to a specific group choice. For instance, at a branch representing group g for individual i, we assign $z_{ig} = 1$.

Once an individual's group is fixed, we solve Equation 2.7 using the current set of constrained individuals (i.e., those assigned in earlier steps along the tree). Since the remaining individuals (those yet to be assigned) remain unconstrained, this formulation provides a lower bound for the sum of squared errors (SSE). This is because as we move further down the tree and impose additional constraints, the SSE can only increase.

Consequently, if the SSE at any node exceeds the SSE of the current best solution (leaf node), we can prune its branches. This ensures that the pruned node cannot lead to the globally optimal solution, improving computational efficiency while preserving optimality. The efficiency of the search is enhanced by using relaxation, initializing with an optimal leaf node, and pruning. Our approach consists of five refined steps:

1. **Individual Ordering in the Tree**: Strategically ordering individuals within the tree enhances search efficiency, allowing us to prioritize and reach potentially optimal branches more quickly.

2. **Initial Feasible Leaf Node**: We start by identifying an initial feasible leaf node to establish a preliminary *upper bound* on the Sum of Squared Errors (SSE). This initial solution provides a benchmark, enabling early pruning of branches that cannot offer

an improvement.

3. **Search Strategy**: A structured search strategy is employed to minimize the number of nodes we explore, optimizing our path to verification of the global optimal solution.

4. **Continuous Relaxation**: At each node in the tree traversal, we solve a continuous relaxation of the integer-constrained problem. This relaxation, where integer constraints are temporarily lifted, yields a *lower bound* for the SSE at that node.

5. **Branch Pruning**: Using both the relaxed bounds and the updated global best, we prune branches at nodes where the SSE lower bound exceeds the best known leaf node SSE. This indicates that further exploration down that path cannot yield the global optimal solution, thereby significantly reducing the search space.

6. **Updating the Global Best Leaf Node**: As we explore the tree, any newly encountered feasible leaf node with a lower SSE than our current best becomes the updated global best. This updated leaf node further tightens the SSE bound, enhancing our ability to prune suboptimal branches.

7. **Termination**: The search concludes when all branches are pruned or the SSE values of all frontier nodes exceed the current best SSE, confirming that the global minimum has been identified.

Each of these steps contributes to an efficient search for the optimal group assignments, ensuring that the algorithm quickly converges to the global minimum. Lets discuss a few in more detail.

2.5.1.1 Ordering Nodes

When ordering individuals in nodes, our objective is to initially select individuals in a way that minimizes the likelihood of misclassifying someone's group early on in the tree, allowing this risk to increase gradually as we progress. To achieve this, we first use a K-means algorithm to assign individuals to groups, providing a preliminary grouping structure. Under these group-level assignments, we run our regression model.

We construct our decision tree by alternating focus between groups, prioritizing individuals based on their distance from regression lines. Initially, we select individuals farthest from the regression lines of other groups, ensuring that early decisions in the tree have significant consequences for the sum of squared errors (SSE). This approach leverages the assumption that if K-means clustering performed well, early misclassifications will have a large impact, guiding the tree construction toward the correct structure. As a result, individuals closest to decision boundaries are addressed later in the process, allowing the latter stages of the algorithm to fine-tune the decision lines.

2.5.1.2 Search Strategy

In the branch-and-bound framework, search strategies are critical for structuring the algorithm's exploration and pruning of the search tree. These strategies directly impact computational efficiency and the speed of convergence to an optimal solution, which is especially important in combinatorial and integer optimization contexts. Common search strategies include breadth-first, depth-first, and best-first searches. However, our approach customizes the search based on the ordering of nodes, with a particular emphasis on navigating group decision boundaries accurately.

Our search strategy begins with a depth-first search (DFS). The primary motivation behind DFS is to reach a feasible solution quickly by traveling deeply down each branch, allowing us to establish an initial upper bound early in the search. This upper bound facilitates pruning by eliminating branches that cannot yield better solutions, reducing unnecessary calculations as we proceed. Once we identify the first feasible leaf node, we switch to a breadth-last strategy.

The breadth-last strategy revisits nodes at deeper levels but focuses on nodes closer to decision boundaries that might otherwise be overlooked in a pure depth-first framework. This allows the search to refine the solution further by examining boundary cases. This combined strategy reaches a solution similar to the K-means initial group assignment via DFS, followed by boundary refinement via breadth-last allows us to identify any misclassifications or boundary ambiguities, achieving a more precise solution.

This hybrid approach leverages the computational efficiency of DFS for rapid solution identification, then systematically broadens the search near boundaries where the K-means clustering assumptions may falter. As a result, this strategy focuses computational resources on refining the nonlinear boundary, leading to a more robust classification while minimizing the risk of misclassification near decision boundaries.

2.5.2 Simulation Results Comparing MIO and K-means Formulations

In this section, we present simulation studies to evaluate the performance of our Mixed-Integer Optimization (MIO) formulation relative to the K-means approach. We generate data from a normal distribution with two groups, separated by a mean difference. The degree of overlap between the groups is controlled by the standard errors, where 3 standard errors correspond to significant overlap and 1 standard error represents minimal to practically no overlap. Table 2.1 provides a summary of the results, showing that the MIO formulation consistently achieves superior group classification accuracy, particularly in scenarios with substantial group overlap.

In addition to accuracy improvements, computational efficiency was a critical consideration for this approach. Our findings indicate that the MIO formulation offers significant advantages in computational efficiency over exhaustive search. For example, we estimate that solving a scenario with 50 entities using exhaustive search would take approximately 356,820 years on our hardware[1]. Furthermore, when comparing two scenarios with 20 entities—one with extended time—the results suggest that the MIO approach scales effectively with time. However, challenges arose in cases with $N = 100$ and significant group overlap, indicating that the MIO formulation has its limitations in such high-complexity settings. We partially

[1] This estimate assumes a computational capacity of 10^{11} operations per second.

address this concern in the next section.

Table 2.1: Simulation Results Comparing MIO and K-means Formulations

	Entities = 20		Entities = 50		Entities = 20
	MIO	K-means	MIO	K-means	MIO
Time Horizon	24	24	20	20	50
Sum of Squared Errors (SSE)	3866.15	3972.36	983.97	983.97	8136.43
Group Accuracy (%)	88.00	80.00	100.00	100.00	97.25
Run Time (seconds)	125.00	0.0161	8474.00	0.0177	357.00
Standard Deviation (SD)	3	3	1	1	3

SSE denotes Sum of Squared Errors, and SD represents the standard deviation of the noise.

These results underscore the robustness of the MIO formulation in achieving higher group classification accuracy compared to K-means, particularly under conditions with substantial noise. Furthermore, the computational efficiency and scalability of the MIO approach provide significant advantages over exhaustive search, making it a practical solution for larger datasets.

2.5.3 Improving Efficiency of MIO Algorithm

If you know more about your problem, additional constraints can dramatically reduce the time complexity for finding the optimal solution. In this section we explore one general constraint. In 2.8, we present a comprehensive framework to identify the global minimum solution for the unobserved group heterogeneity problem. Given that the indexing of groups lacks inherent meaning, we impose an ordering on the group effect parameters by arranging them in ascending order based on their index values. This ordering effectively reduces redundant comparisons, as it eliminates the need for the algorithm to evaluate permutations such as whether group 1's effect is greater than group 2's or vice versa. This structured approach simplifies the search space and enhances algorithmic efficiency.

$$
\begin{aligned}
\min_{(\Theta,\alpha,\tilde{\alpha},z)} \textstyle\sum_{i=1}^{N} \sum_{t=1}^{T} \quad & (y_{it} - x_{it}'\theta - \tilde{\alpha}_{it})^2 \\
& \text{subject to} \\
& \tilde{\alpha}_{it} - \alpha_{gt} \leq z_{ig} M \quad \forall i, g \\
& \textstyle\sum_{g=1}^{G} z_{ig} = G - 1 \quad \forall i, \\
& \alpha_{g,t} \leq \alpha_{g+1,t} \qquad \forall g \in \{1, ..., G-1\}
\end{aligned}
\tag{2.8}
$$

2.5.4 Simulations Assessing Enhanced Efficiency in MIO Formulations

In this section, we compare formulation 2.8 with the additional constraint to the K-means solution, focusing on improvements in computational efficiency, particularly regarding runtime. We are using the same setup that we used in the previous simulations. In our previous simulations, solving the 50-entity problem required approximately 2.3 hours. With the new formulation, the same problem was completed in just 33 seconds—a dramatic improvement. While this advancement makes the 100-entity problem with overlap feasible, solving the $N = 200$ case remains challenging due to excessive computational time.

Table 2.2: Comparison of MIO and GFE Performance Metrics

	20 Entities		50 Entities		20 Entities
Metric	MIO	GFE	MIO	GFE	MIO
Time Periods	24	24	20	20	50
Sum of Squared Errors (SSE)	3934.4	4029.1	965.03	965.03	8088.88
Group Accuracy (%)	87.0	71.25	100.0	100.0	98.0
Runtime (sec)	6.9	0.01325	33.0	0.01575	42.35
Standard Deviation (SD)	3	3	1	1	3

SSE denotes Sum of Squared Errors, and SD represents the standard deviation of the noise.

2.6 Summary

In this chapter, we introduced a Mixed Integer Optimization (MIO) formulation to estimate the unobserved time-varying group choice fixed effect model in panel data. Our results demonstrate that the proposed MIO approach improves both the accuracy and precision of estimates compared to the commonly used K-means method. The formulation proved feasible in noisy environments with significant group overlap for problems up to $N = 100$. However, for larger datasets with substantial overlap, particularly when $N > 200$, the computational time becomes impractical.

To address these limitations, the next chapter presents a novel algorithm that combines Branch-and-Bound techniques with the structural properties of linear models in least squares. This new global optimization procedure aims to efficiently identify either the global optimum or a close approximation to it, even for large-scale problems where N can reach up to 1 million.

CHAPTER 3

Linear Search Algorithm for Unobserved Heterogeneity in Panel Data

While Mixed Integer Optimization over individual group choice offers a globally optimal solution to the challenge of unobserved group heterogeneity, it remains an NP-hard problem, making it computationally infeasible for large sample sizes. In this chapter, we harness properties of least squares estimators to significantly reduce the computational complexity. Specifically, We prove the decision boundary must be linear then re-formulate the problem to search for the decision boundary which implies individuals group choices. We motivate this approach by demonstrating that under strong assumptions about group membership the NP-hard group selection problem can be transformed into one with linear time complexity relative to N. Afterwards, we relax these assumptions to develop a Linear Search Algorithm. Through simulations, we show that this approach not only accommodates large sample sizes but also consistently outperforms K-means in terms of accuracy.

3.1 The Problem

To start, let us revisit the issue of unobserved group heterogeneity in panel data settings. The estimation problem can be framed as follows:

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

The challenge lies in accurately estimating θ, our parameter of interest, which suffers from bias due to our omitted unobserved group time trends $\alpha_{g_i t}$. This unobserved group heterogeneity complicates the optimization by introducing unknown group assignments that must be inferred simultaneously with the main parameters.

Mixed Integer Optimization (MIO) over individual group choices provides a framework for obtaining consistent estimates of these parameters, but the computational burden is significant; being NP-hard, MIO becomes impractical for datasets with large N. Similarly, K-means clustering, while faster, struggles to capture the complex, dynamic group structures in panel data, often leading to imprecise group assignments and inaccurate parameter estimates.

This chapter explores how leveraging properties of linear models in least squares estimation can reduce the computational complexity of group assignment, allowing for more scalable and efficient estimation procedures that retain accurate parameter estimation across large N.

3.2 Characterizing The Decision Boundary

To improve the MIO formulation, we aim to exploit certain properties of the least squares estimator to reduce computational complexity. Our first approach is to characterize the decision boundary for group assignment. Then we will bound the location the decision boundary must cross through. By identifying both the shape and location of this boundary, we can reduce number of group combinations we search over.

To achieve this, we take the least squares estimates, $\hat{\theta}$ and $\hat{\alpha}$, as given, and focus on the group assignment decision for a particular individual i. Specifically, the decision boundary for assigning individual i to group g versus an alternative group $\tilde{g}$ occurs when the Sum of Squared Errors (SSE) is equal for both assignments, rendering the algorithm indifferent between the two choices.

$$
\begin{aligned}
&\textstyle\sum_{t=1}^{T}(y_{it} - \hat{\alpha}_{g_{it}} - x'_{it}\hat{\theta})^2 = \sum_{t=1}^{T}(y_{it} - \hat{\alpha}_{\tilde{g}_{it}} - x'_{it}\hat{\theta})^2 \\
&\textstyle\sum_{t=1}^{T} y_{it}^2 - 2y_{it}(\hat{\alpha}_{g_{it}} - x'_{it}\hat{\theta}) + (\hat{\alpha}_{g_{it}} - x'_{it}\hat{\theta})^2 = \sum_{t=1}^{T} y_{it}^2 - 2y_{it}(\hat{\alpha}_{\tilde{g}_{it}} - x'_{it}\hat{\theta}) + (\hat{\alpha}_{\tilde{g}_{it}} - x'_{it}\hat{\theta})^2 \\
&\textstyle\sum_{t=1}^{T} -2y_{it}(\hat{\alpha}_{g_{it}} - x'_{it}\hat{\theta}) + (\hat{\alpha}_{g_{it}} - x'_{it}\hat{\theta})^2 = \sum_{t=1}^{T} -2y_{it}(\hat{\alpha}_{\tilde{g}_{it}} - x'_{it}\hat{\theta}) + (\hat{\alpha}_{\tilde{g}_{it}} - x'_{it}\hat{\theta})^2 \\
&\textstyle\sum_{t=1}^{T} -2y_{it}(\hat{\alpha}_{g_{it}} - \hat{\alpha}_{\tilde{g}_{it}}) = \sum_{t=1}^{T}(\hat{\alpha}_{\tilde{g}_{it}} - x'_{it}\hat{\theta})^2 - (\hat{\alpha}_{g_{it}} - x'_{it}\hat{\theta})^2 \\
&\textstyle\sum_{t=1}^{T} -2y_{it}(\hat{\alpha}_{g_{it}} - \hat{\alpha}_{\tilde{g}_{it}}) = \sum_{t=1}^{T}(\hat{\alpha}_{\tilde{g}_{it}} + 2x'_{it} + \hat{\alpha}_{g_{it}})(\hat{\alpha}_{\tilde{g}_{it}} - \hat{\alpha}_{g_{it}}) \\
&\textstyle\sum_{t=1}^{T} -2y_{it}(\hat{\alpha}_{g_{it}} - \hat{\alpha}_{\tilde{g}_{it}}) = \sum_{t=1}^{T}(\hat{\alpha}_{\tilde{g}_{it}} + \hat{\alpha}_{g_{it}})(\hat{\alpha}_{\tilde{g}_{it}} - \hat{\alpha}_{g_{it}}) + 2x'_{it}(\hat{\alpha}_{\tilde{g}_{it}} - \hat{\alpha}_{g_{it}}) \\
&\begin{bmatrix} y_i \\ x_i \end{bmatrix}^{\top} (\hat{\alpha}_{g_i} - \hat{\alpha}_{\tilde{g}_i}) = \frac{\sum_{t=1}^{T}\left(\hat{\alpha}_{\tilde{g}_{it}} + \hat{\alpha}_{g_{it}}\right)\left(\hat{\alpha}_{g_{it}} - \hat{\alpha}_{\tilde{g}_{it}}\right)}{2}
\end{aligned}
\tag{3.1}
$$

From this decomposition, it is apparent that the decision boundary is represented by a hyperplane in the data space $\mathbb{R}^{(1+P)T}$ where P is the number of covariates and T is the number of time periods. The linearity of this optimal decision boundary in both finite samples and asymptotically will provide crucial leverage in constructing a more efficient algorithm.

We can simplify this problem further by considering an alternative estimator. Bonhomme and Manresa (2015) assume that individuals remain in the same group over time. By averaging across time, the group assignment problem remains unchanged. While this approach sacrifices some time variation, it reduces the search space and minimizes errors around the regression line, both of which significantly decrease the computational burden of the optimization process. The time-averaged estimator is formulated as follows:

$$
(\hat{\theta}, \hat{\bar{\alpha}}, \hat{\gamma}) = \underset{(\theta,\alpha,\gamma)\in\Theta\times A^{GT}\times\Gamma_G}{\operatorname{argmin}} \sum_{i=1}^{N} \left(\bar{y}_i - \bar{x}'_i\theta - \bar{\alpha}_{g_i}\right)^2, \tag{3.2}
$$

where $\bar{y}_i$ and $\bar{x}'_i$ denote the time-averaged outcomes and covariates, respectively. Under this estimator, the decision boundary remains linear (see proof in Appendix). Specifically, the decision boundary lies equidistant between the regression lines corresponding to the two groups, as given by:

$$\bar{y}_i = \frac{\hat{\bar{\alpha}}_{\tilde{g}} + \hat{\bar{\alpha}}_g}{2} + \bar{x}_i'\hat{\theta} \tag{3.3}$$

This linear decision rule further facilitates computational efficiency, making it highly applicable to large-scale problems.

3.2.1 Locating The Decision Boundary: A Motivating Example

Having established that the decision boundary is linear, the next step is to identify its potential location. Under strong assumptions regarding group composition, we can pinpoint a single location where the decision boundary must necessarily lie. Furthermore, under these assumptions, we demonstrate that the computational complexity of estimating our parameters is linear in N in the two-dimensional setting. In this section, we will first present this motivating example based on these restrictive assumptions and then proceed to relax these assumptions to extend the framework to more general cases.

To develop an intuitive understanding of this problem, we turn to the estimator presented in Equation 3.2. By leveraging properties of least squares estimators, we can characterize the group intercepts. For a given group , the group-specific intercept can be expressed as:

$$\hat{\bar{\alpha}}_g = \bar{\bar{y}}_g - \bar{\bar{x}}'_g\hat{\theta} \tag{3.4}$$

where $\bar{\bar{y}}_g$ and $\bar{\bar{x}}_g$ represent the group-level means of the dependent variable and independent variables, respectively, and denotes the estimated coefficients. To proceed, let us redefine Equation 3.3 using our definition of group-specific intercept above and derive its implications:

$$\begin{aligned} \bar{y}_i &= \tfrac{\bar{\bar{y}}_{\tilde{g}} - \bar{\bar{x}}'_{\tilde{g}}\hat{\theta} + \bar{\bar{y}}_g - \bar{\bar{x}}'_g\hat{\theta}}{2} + \bar{x}_i'\hat{\theta} \\ \bar{y}_i &= \tfrac{\bar{\bar{y}}_{\tilde{g}} + \bar{\bar{y}}_g}{2} - \tfrac{\bar{\bar{x}}'_{\tilde{g}}\hat{\theta} + \bar{\bar{x}}'_g\hat{\theta}}{2} + \bar{x}_i'\hat{\theta} \end{aligned} \tag{3.5}$$

This reformulation enables us to make precise assumptions about the relationship between our observable data and the decision boundary's location. Specifically, under certain assumptions about the number of groups and their composition, we can identify a unique point where the decision boundary must reside. We now formalize the intuition with the following theorem:

Theorem 5. *Assume there are two groups with an equal number of individuals in each group. Under this symmetry, the decision boundary for the following problem must lie at the midpoint of the group-level means, denoted by* $(\bar{\bar{x}}', \bar{\bar{y}})$ *.*

$$(\hat{\theta}, \hat{\bar{\alpha}}, \hat{\gamma}) = \underset{(\theta,\alpha,\gamma)\in\Theta\times A^{GT}\times\Gamma_G}{argmin} \sum_{i=1}^{N} \left(\bar{y}_i - \bar{x}_i'\theta - \bar{\alpha}_{g_i}\right)^2 ,$$

Given that the linear decision boundary must pass through $(\bar{\bar{x}}', \bar{\bar{y}})$, we can use an efficient algorithm to identify the globally optimal decision boundary. In the two-dimensional case, the algorithm starts with an initial line and iteratively rotates it clockwise $N-1$ times to align just above the next closest vector. This will explore all possible group combinations.

However, the process becomes significantly more complex in higher-dimensional settings. In the case of two covariate dimensions, the algorithm may require up to $(N-1)^2$ iterations. This occurs because the plane first "snaps" to the nearest vector along one covariate dimension. Then, holding these two points fixed, it rotates by snapping to the closest vector along the other dimension, orthogonal to the first two points. This sequence involves $N-1$ rotations for each dimension, resulting in the iterative process. Thus, in low-dimensional settings, the algorithm operates with linear complexity in N, but in higher-dimensional cases, its complexity appears to scale polynomially with N depending on the number of covariates.

3.2.1.1 Linear Rotation Algorithm

In this section, we describe the current implementation of the algorithm. It will have a similar flavor as the Gomroy Cutting Plane Method [Gom10]. Throughout, we assume a two-group setting with an equal number of individuals in each group. We also assume the single covariate setting but it is easily extended to the multi covariate setting.

Step 1: Initialization

Initialize a two-dimensional vector $\mathbf{v}$ with any direction. For example, if you had a single covariate you could initialize at $[0, 1]$. For all vectors in the space (x, y), subtract their mean $(\bar{\bar{x}}, \bar{\bar{y}})$ to obtain the centered vectors $(\tilde{x}, \tilde{y})$, where:

$$\tilde{x} = x - \bar{\bar{x}}, \quad \tilde{y} = y - \bar{\bar{y}}.$$

Step 2: Group Classification

Classify each vector $(\tilde{x}, \tilde{y})$ into one of two groups using the perpendicular vector to $\mathbf{v}$, denoted as $\mathbf{v}'$. Note, it does not matter how the perpendicular vector is oriented. Compute the dot product of $\mathbf{v}'$ with each $(\tilde{x}, \tilde{y})$:

$$\text{Dot product} = \mathbf{v}' \cdot (\tilde{x}, \tilde{y}).$$

If the dot product is greater than 0, classify the vector into Group 1; otherwise, classify it into Group 0.

Step 3: Regression and Error Calculation

Using the current group classification and estimate the following regression model:

$$(\hat{\theta}, \hat{\bar{\alpha}}) = \underset{(\theta,\alpha)\in\Theta\times A^{GT}}{\operatorname{argmin}} \sum_{i=1}^{N} \left(\bar{y}_i - \bar{x}_i\theta - \bar{\alpha}_{g_i}\right)^2,$$

Compute the sum of squared errors (SSE) and save for later comparison.

Step 4: Rotation of the Vector

Identify the vector closest to the current $\mathbf{v}$. To do this:

1. Compute the angle of each $(\tilde{x}, \tilde{y})$ using the arctangent function:

$$\theta = \arctan(\tilde{x}, \tilde{y}).$$

2. Find the angle closest to $\mathbf{v}$ in the clockwise direction.

3. Update $\mathbf{v}$ to align with the selected $(\tilde{x}, \tilde{y})$ using a rotation matrix:

$$\mathbf{v}_{\text{new}} = \begin{bmatrix} \cos(\Delta\theta) & -\sin(\Delta\theta) \\ \sin(\Delta\theta) & \cos(\Delta\theta) \end{bmatrix} \mathbf{v},$$

   where $\Delta\theta$ is the angular difference between $\mathbf{v}$ and the selected vector $(\tilde{x}, \tilde{y})$. This rotation ensures the updated $\mathbf{v}$ points in the direction of the selected vector.

Rotate $\mathbf{v}$ by the difference in angles, which changes the group classification of the corresponding vector.

Step 5: Iterative Optimization

Repeat Steps 2 through 4 for $N-1$ rotations, where N is the number of vectors in the space. Track the group classifications and SSE values for each iteration. After completing all rotations, select the group classification that corresponds to the smallest SSE. This classification represents the globally optimal solution.

3.2.1.2 Linear Rotation Algorithm Simulations

This section presents a simulation study to evaluate the performance of the proposed Linear Rotation (LR) algorithm compared to K-means. The primary focus of this study is to illustrate the computational efficiency and robustness of the LR algorithm.

To simulate the data, we generate an independent variable $X \sim \mathcal{N}(0, 10)$ with $N = 1000$ observations and a noise term $\epsilon \sim \mathcal{N}(0, 3)$. The data is partitioned into two groups, with an equal proportion of individuals (50% per group). Each group is characterized by distinct intercepts (0 and 7, respectively), while the slope parameter (θ) is fixed at 7. This setup introduces sufficient overlap between the groups, creating a challenging estimation problem. Specifically, such overlap creates issues for optimization procedures like MIO as there are many points where placing them in either group creates similar lower bounded SSE's. For N equal to 1000 is could take our MIO formulation in Chapter 2 months or even years to solve. Figure 3.1 provides a visual illustration of the simulated data structure, highlighting the separation between the groups and the overall linear trend.

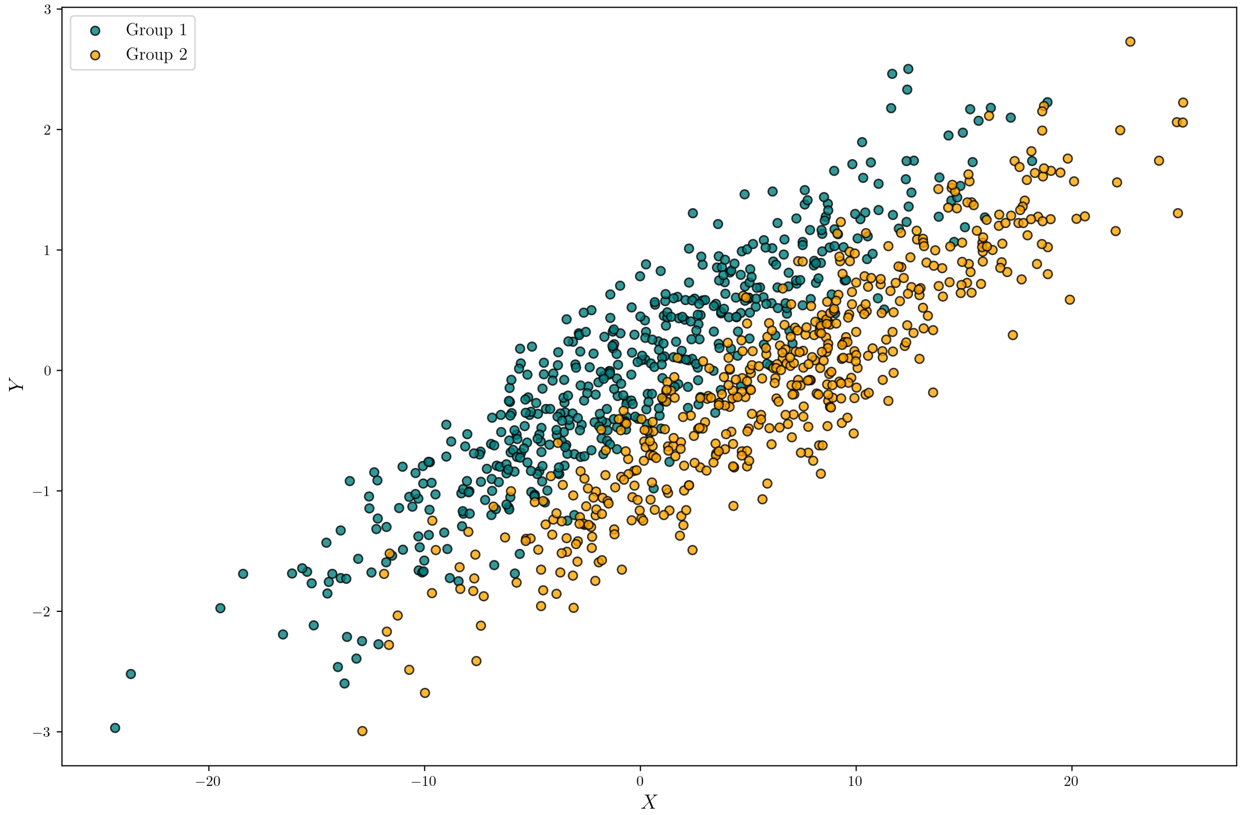


Figure 3.1: Example of Simulated Data with Two Overlapping Groups.

We will compare our LR algorithm with K-means the current suggestion for large sample environments. We initialize our K-means algorithm by randomly assigning individuals to one of two groups (50/50 split) and then running a linear regression to estimate the parameters. This initialization ensures favorable conditions for K-means, given the normality of the data and the equal group proportions. In subsequent sections, we will demonstrate a data-generating process in which the performance of the K-means algorithm significantly deteriorates, highlighting its sensitivity to initial conditions with respect to the underlying data distribution.

Table 3.1 summarizes the results comparing the LR algorithm with K-means over 1000 simulations. We find that the LR algorithm reduces SSE by 12.6% relative to the K-means approach but only increases group accuracy by 1%. We believe K-means is misclassifying outliers. We believe this because K-means is known to have trouble with outliers. Furthermore, if it misclassifies them, this would create leverage causing large increases in SSE. We also find that it only took 10 seconds for the LR algorithm to classify 1000 individuals group choices. This is a dramatic improvement over our MIO formulation which would have taken several months to solve this problem.

Table 3.1: Performance Comparison Between LR and the K-means GFE Models

Metric	**LR (Linear Rotation)**	**K-means (Group Fixed Effects)**
Sum of Squared Errors (SSE)	5541.93	6346.28
Group Accuracy (%)	88.6	87.7
Runtime (seconds)	10.1	3.4

Notes: For this analysis, we simulated data for 1000 individuals across 1000 simulations. Two groups with intercepts 7 units apart were generated, with a variance of 3. This setup introduced overlap between groups, making them not perfectly separable.

To better understand the performance of these algorithms we plot their θ estimates in Figure 3.2. Given the true value of θ is 7, we find that both estimators seem to be unbiased but the LR estimator has a tighter distribution.

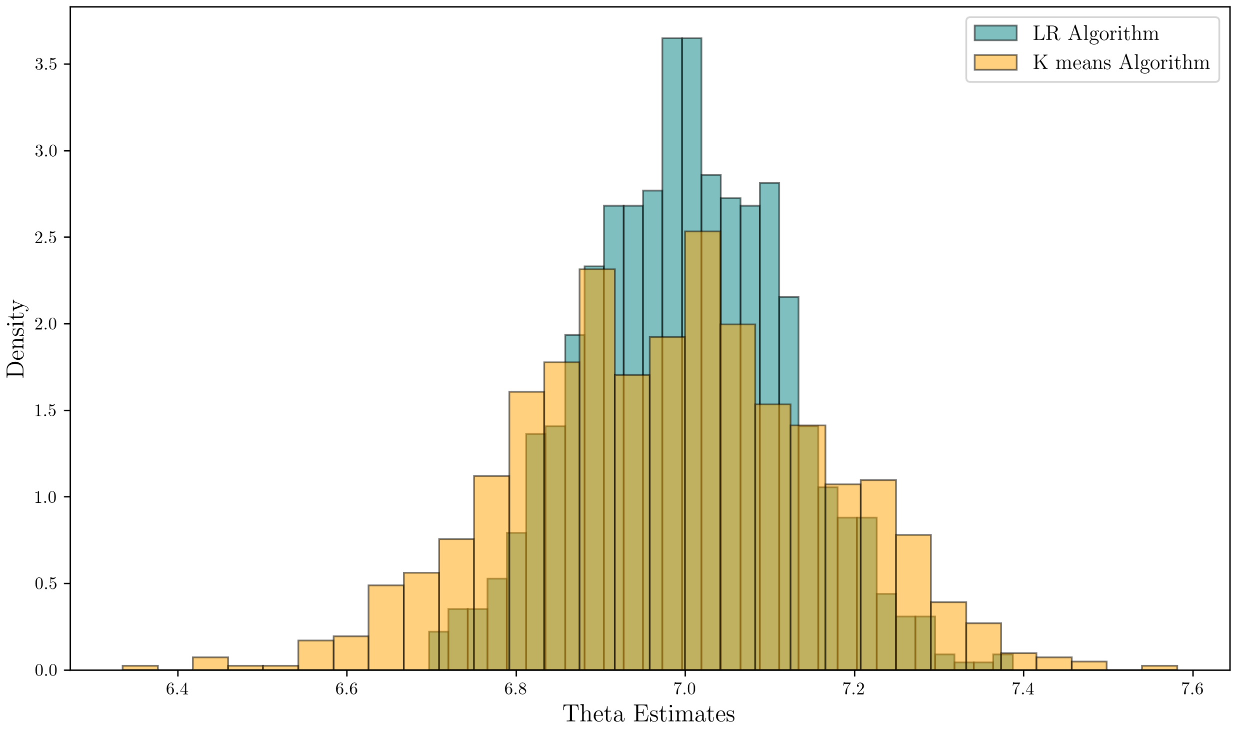


Figure 3.2: Distribution of Theta Estimates

3.3 Bounding the Decision Boundary

In the previous sections, we discussed a motivating example that transformed the NP-hard group choice problem into a tractable one. However, this was achieved under the assumptions that there must be two groups, the groups share the same slope, and the number of individuals in each group is equal. In this section, we relax the assumption that the groups must have an equal number of individuals and demonstrate how we can identify a region through which the decision boundary must pass.

To explore this, let us return to our simplified setting and revisit the problem of two unobserved groups. Consider the characterization of the decision boundary outlined in Equation (3.6). In the two-group setting (denoted as A and B), we can reformulate the boundary as follows:

$$\bar{y}_i = \tfrac{\bar{\bar{y}}_A+\bar{\bar{y}}_B}{2} - \tfrac{\bar{\bar{x}}_A\hat{\theta}+\bar{\bar{x}}_B\hat{\theta}}{2} + \bar{x}'_i\hat{\theta} \tag{3.6}$$

Notice, if we set $\bar{x}'_i = \frac{\bar{\bar{x}}'_A+\bar{\bar{x}}'_B}{2}$ this would lead to $\bar{y}_i = \frac{\bar{\bar{y}}_A+\bar{\bar{y}}_B}{2}$ the decision boundary must pass through $(\frac{\bar{\bar{x}}'_A+\bar{\bar{x}}'_B}{2}, \frac{\bar{\bar{y}}_A+\bar{\bar{y}}_B}{2})$. If we could bound bound this region, then we would know the area which the decision line must pass through. Specifically, we will create conservative bounds by bounding each individual dimension. To find the minimum of the weighted average $\frac{\bar{\bar{x}}'_A+\bar{\bar{x}}'_B}{2}$ do the following steps:

Determining the Minimum Point

1. For each dimension of X', denoted X^p, Assign the smallest value of X^p to Group A:

$$\bar{\bar{x}}^p_A = \min(X^p),$$

where $X^p = \{x_1, x_2, \dots, x_n\}$ represents the set of scalar values.
2. Distribute the remaining values of X^p to Group B, ensuring:

$$\bar{\bar{x}}^p_B = \frac{\sum X^p - \min(X^p)}{|X^p| - 1}.$$

Thus, the minimum point at which the decision boundary could cross the X-dimension is given by:

$$\bar{\bar{X}}^p_{\min} = \frac{\bar{\bar{x}}^p_B + \bar{\bar{x}}^p_A}{2}.$$

Determining the Maximum Point

1. For the same dimension of X', denoted X^p, Assign the largest value of X^p to Group B:

$$\bar{\bar{x}}^p_B = \max(X^p).$$

2. Distribute all other values of X^p to Group A, resulting in:

$$\bar{\bar{x}}^p_A = \frac{\sum X^p - \max(X^p)}{|X^p| - 1}.$$

The maximum point at which the decision boundary could cross the X^p-dimension is then:

$$\bar{\bar{X}}^p_{\max} = \frac{\bar{\bar{x}}^p_B + \bar{\bar{x}}^p_A}{2}.$$

Similarly, to bound $\frac{\bar{\bar{y}}_A + \bar{\bar{y}}_B}{2}$, the same procedure is applied to the scalar values in Y, yielding:

$$(\bar{\bar{Y}}_{\min}, \bar{\bar{Y}}_{\max}),$$

These operations define the bounds for the weighted averages of X and Y:

$$(\bar{\bar{X}}'_{\min}, \bar{\bar{X}}'_{\max}, \bar{\bar{Y}}_{\min}, \bar{\bar{Y}}_{\max}).$$

Since the problem is convex, we can guarantee that the decision boundary must cross a point that lies within the rectangle defined by these bounds. This region represents all feasible points for the decision boundary:

$$\text{Rectangle: } [\bar{\bar{X}}^1_{\min}, \bar{\bar{X}}^1_{\max}] \times \ldots \times [\bar{\bar{X}}^P_{\min}, \bar{\bar{X}}^P_{\max}] \times [\bar{\bar{Y}}_{\min}, \bar{\bar{Y}}_{\max}].$$

3.4 Optimization Algorithm for Least Squares Classification

Next, we will leverage the rectangular bounds and the linear decision boundary to create a relatively fast global optimization algorithm for our least squares problem. The algorithm is designed to systematically identify the optimal decision boundary within a bounded rectangular region. The primary objective is to classify data points into distinct groups while minimizing the Sum of Squared Errors (SSE) of a regression model fitted to the classified data. By leveraging a decision tree framework, the algorithm iteratively refines candidate decision boundaries by segmenting the rectangle, classifying points based on their spatial relationships, and evaluating potential solutions against a lower bound of the SSE.

This approach ensures computational efficiency through strategic pruning. Branches of the decision tree are terminated early if their lower bound SSE exceeds the current best solution. Additionally, the method handles unclassified points by either exploring all possible group assignments or further subdividing the decision segments. The iterative refinement process converges on the optimal segmentation, leveraging convexity to guarantee that the solution lies within the specified bounds.

In the analysis that follows, we focus on a simplified model with a single covariate, where the data is averaged over time. This approach is chosen because increasing the dimensionality of the covariates not only expands the dimensionality of the rectangle and decision boundary but also increases the number of segments that must be matched, adding to the computational burden. While we believe there are promising methods to effectively reduce the excess dimensionality of covariates and simplify the computational process (See Appendix), we leave this exploration to future work.

Step 1: Segment the Rectangle

Given the rectangle $[\bar{\bar{X}}_{\min}, \bar{\bar{X}}_{\max}] \times [\bar{\bar{Y}}_{\min}, \bar{\bar{Y}}_{\max}]$, each side, k, is divided into S equal segments. Segments on the X-axis and Y-axis are defined as follows:

$$X_{ks} = \bar{\bar{X}}_{\min} + s \cdot \Delta X, \quad \text{for } s = 0, \ldots, S, \quad \Delta X = \frac{\bar{\bar{X}}_{\max} - \bar{\bar{X}}_{\min}}{S},$$

$$Y_{ks} = \bar{\bar{Y}}_{\min} + s \cdot \Delta Y, \quad \text{for } s = 0, \ldots, S, \quad \Delta Y = \frac{\bar{\bar{Y}}_{\max} - \bar{\bar{Y}}_{\min}}{S}.$$

This process creates a grid of segments along the boundaries of the rectangle, represented by their endpoints. For example, $(X_{k0}, Y_{k0}, X_{k1}, Y_{k1})$ is the first segment for side k.

Step 2: Pair Segments to Create Initial Nodes

Pair each segment s on side k with every segment s' on any other side $k' \neq k$. Each pair of segments defines a potential area through which the decision line must pass. These pairs form the initial nodes in the decision tree. This process creates $6S^2$ pairs. Notice if we increased the dimensionality of our rectangle the number and nature of pairs would increase.

Step 3: Classify Data Points and Compute Lower Bound SSE

For each node, decision lines are constructed by connecting the endpoints of the paired segments. Four decision lines are defined using the four endpoints of the segment to create boundaries within which the true decision line must lie. For example, one decision line can be expressed as:

$$L_1 : y = m_1 x + c_1,$$

where:

$$m_1 = \frac{Y_{k'1} - Y_{k0}}{X_{k'1} - X_{k0}}, \quad c_1 = Y_{k0} - m_1 X_{k0}.$$

Once the decision lines are defined, data points are classified based on their position relative to these lines. Note relative depends on which sides our segments are on but we leave the details to the appendix. If the sides were left and right. Then for each data point (x_k, y_k), points above all decision lines are assigned to Group G_1, while points below all decision lines are assigned to Group G_2. For instance:

$$\begin{aligned} G_1 &: \text{points satisfying } y_k > m_l x_k + c_l, \quad \forall l \in \{1, 2, 3, 4\} \\ G_2 &: \text{points satisfying } y_k < m_l x_k + c_l, \quad \forall l \in \{1, 2, 3, 4\} \end{aligned}$$

Points that lie between the lines are labeled as *unclassified* and excluded from further analysis. Next, the linear regression problem is solved for the set of all classified points which we will denote with $\tilde{N}$.

$$(\hat{\theta}, \hat{\bar{\alpha}}) = \underset{(\theta,\alpha)\in\Theta\times A^2}{\operatorname{argmin}} \sum_{i=1}^{\tilde{N}} \sum_{t=1}^{T} \left(\bar{y}_i - \bar{x}_i'\theta - \bar{\alpha}_{g_i}\right)^2, \quad \text{where } g_i \in \{1, 2\}.$$

Finally, we compute The Sum of Squared Errors (SSE) for classified points. This SSE serves as a lower bound since unclassified points are removed, and their inclusion would only increase the SSE.

$$\text{SSE}_{\text{LB}} = \sum_i \left(y_i - \hat{y}_i\right)^2.$$

Step 4: Compare SSE to the Current Best SSE

The current best SSE is initialized using a baseline solution, currently we are using the K-Means clustering result, denoted as $\mathrm{SSE}_{\mathrm{Best}}$. For each node, if $\mathrm{SSE}_{\mathrm{LB}} > \mathrm{SSE}_{\mathrm{Best}}$, the branch is pruned. If $\mathrm{SSE}_{\mathrm{LB}} \leq \mathrm{SSE}_{\mathrm{Best}}$, the number of unclassified points, denoted u, is evaluated. If $u \leq 5$, all 2^u possible classifications of the unclassified points are enumerated. For each classification, the SSE is computed as:

$$\mathrm{SSE}_{\mathrm{new}} = \sum_i (y_i - \hat{y}_i)^2 .$$

If $\mathrm{SSE}_{\mathrm{new}} < \mathrm{SSE}_{\mathrm{Best}}$, the best SSE is updated. If $u > 5$, each segment is subdivided in half. These subdivided segments from one side are paired with those from the other side, generating four new branches for further exploration.

Step 5: Iterate Through the Tree

After completing one level of the decision tree, the new branches are taken, and the process is repeated until no branches remain. The classification corresponding to the smallest SSE, $\mathrm{SSE}_{\mathrm{Best}}$, represents the global minimum for the least squares problem.

3.5 Example of the Process

This figure illustrates an example of a single node in the decision tree, as described in the segmentation and classification process. The red rectangle represents the bounds of the current region under consideration, $[\bar{X}_{\mathrm{min}}, \bar{X}_{\mathrm{max}}] \times [\bar{Y}_{\mathrm{min}}, \bar{Y}_{\mathrm{max}}]$. At this node, the sides of the rectangle have been divided into segments, and a specific pairing of segments is being evaluated. The green lines on the edges of the rectangle correspond to the selected segment endpoints for this pairing.

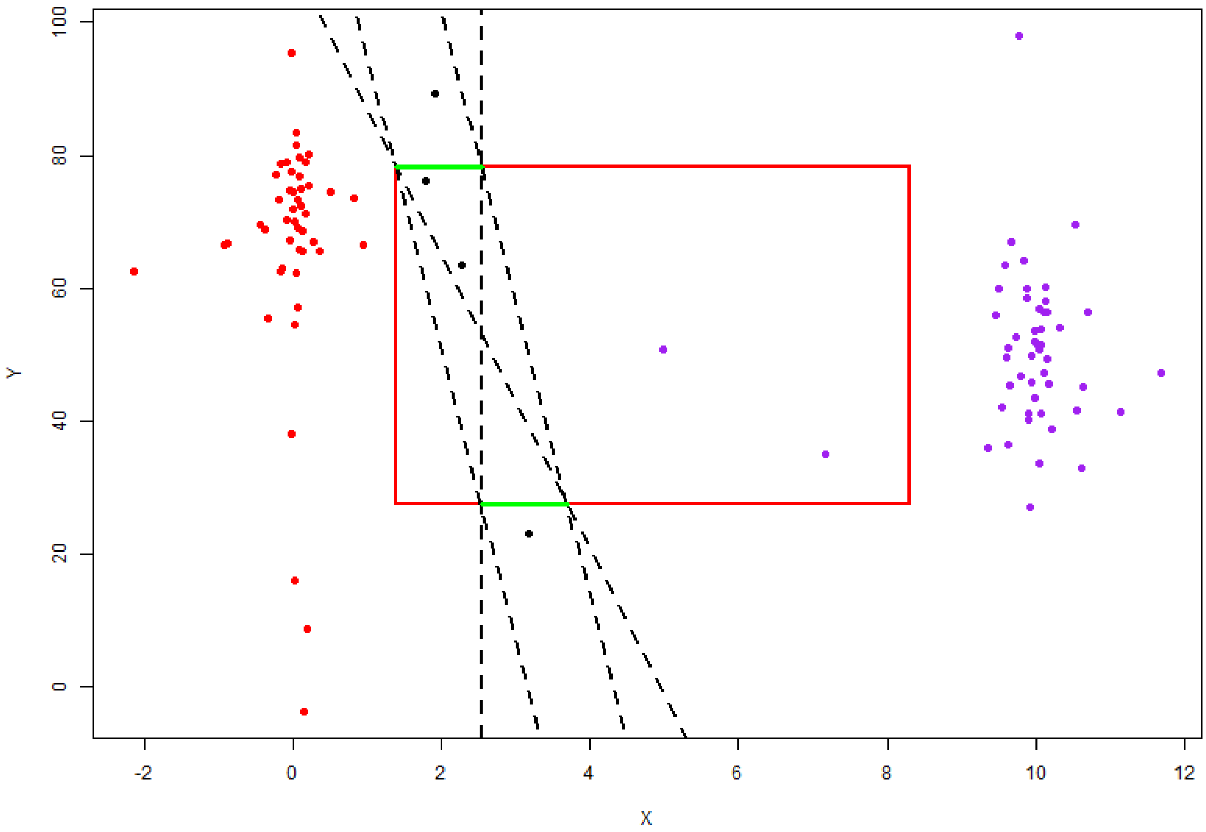


Figure 3.3: Example of Segmenting Data with Bounding Box

From these segments, four decision lines are constructed, shown as the dashed black lines. These lines are defined by connecting the endpoints of the paired segments, creating boundaries that divide the data points into distinct regions. The decision lines split the space into three parts: the area to the left of all the decision lines, the area to the right of all the decision lines, and the region in between. These regions determine the initial classification of the data points.

In this example, the red points are classified into G_1 (Group 1), as they lie on one side of the decision lines. The purple points are classified into G_2 (Group 2), lying on the opposite side of the decision lines. Points that fall between the two lines, within the region of uncertainty, remain unclassified at this step. These unclassified points are excluded from the regression analysis at this node.

At this stage, the classified points are used to fit the regression model and the Sum of Squared Errors (SSE) is calculated. This SSE represents the lower bound for the current node since unclassified points are ignored. The algorithm then compares this lower bound SSE with the current best SSE obtained from previous nodes or the initial K-Means solution. If the lower bound SSE is greater than the current best SSE, this branch of the decision tree is pruned, and no further exploration is conducted. Otherwise, if the lower bound SSE is promising, the algorithm continues by either resolving the classification of the unclassified points or further refining the segments by subdividing them into smaller sections.

3.6 Simulations

In this section, we present a simulation demonstrating not only the computational efficiency of our Linear Search algorithm but also an example of a distribution where K-means performance is highly dependent on initial conditions. We assume the following data generating process where $g(i)$ represents the unobserved individuals group choice.

$$Y_i = X_i\theta + \epsilon_i + \alpha_{g(i)}$$

We simulate a dataset of 100 individuals, where each individual's X variable is independently drawn with equal probability (50%) from one of two Cauchy distributions: $Cauchy(0, 0.35)$ or $Cauchy(10, 0.35)$. If an individual's X value is sampled from the first distribution, there is a 90% probability that they are assigned to Group A (denoted as the red group). Conversely, if their X value is sampled from the second distribution, the probability of being assigned to Group A is reduced to 10%.

Individuals assigned to Group A are given an intercept of 70, while those not in Group A are assigned an intercept of 10. This setup results in data with heavy-tailed characteristics.

The errors are modeled as independently and identically distributed normal random variables with mean 0 and standard deviation 5 ($\epsilon \sim N(0,5)$). The true value of the slope parameter is set to $\beta = 8$. A visual representation of the resulting distribution and corresponding group assignments is provided below.

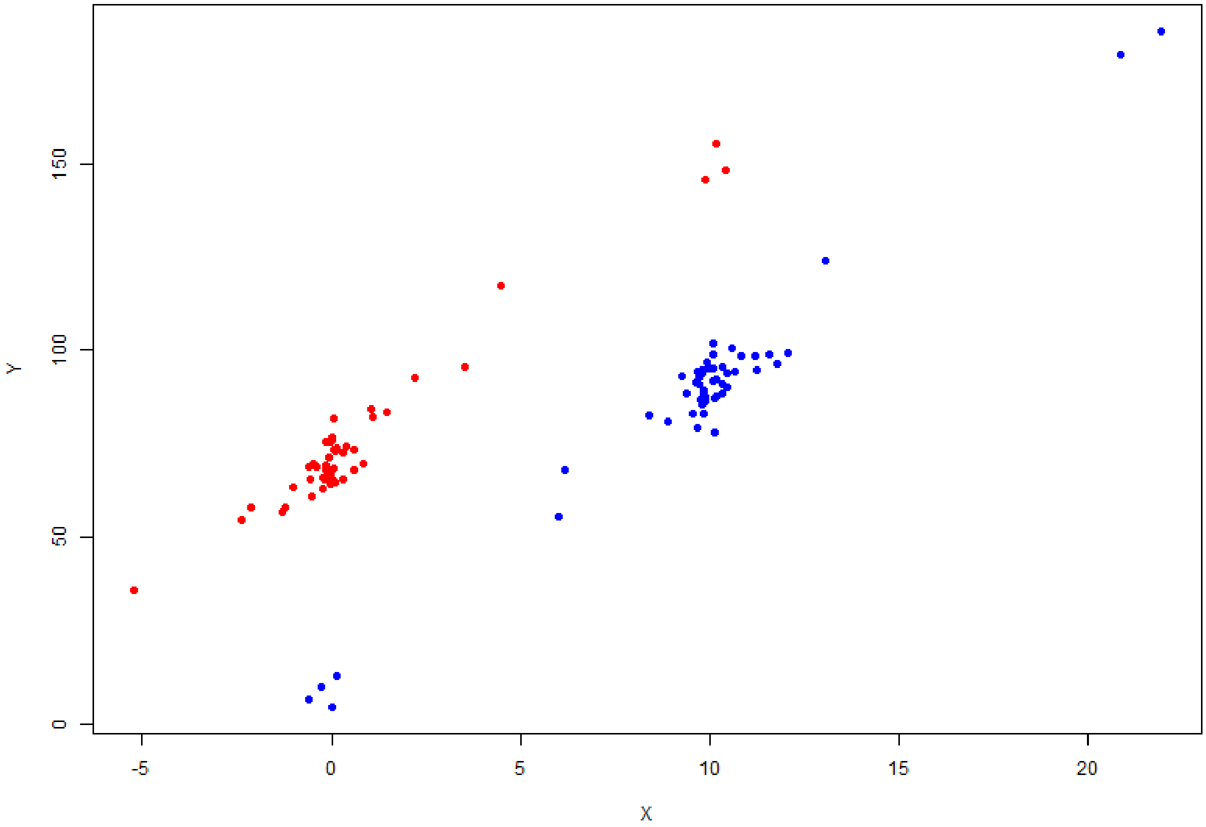


Figure 3.4: Example of Segmenting Data with Bounding Box

We selected this distribution specifically for its sparse tails, which can amplify the effects of misclassifications, leading to significant leverage and substantial swings in the estimated β values. K-means appears to struggle with correcting poor initializations under these conditions. Interestingly, we observed that reducing the variance of the error term exacerbated the performance of K-means rather than improving it. This behavior suggests that higher variance may provide K-means with additional flexibility to escape local minima, though the underlying mechanism for this phenomenon warrants further investigation.

After running the simulations, we observe the following results. Both K-means and Linear Search perform well in terms of group classification accuracy, with Linear Search achieving an impressive 99.9% accuracy compared to 98.8% for K-means. In terms of computational efficiency, both methods are relatively fast: Linear Search completes in an average of 2 seconds, while K-means is exceptionally quick, requiring less than a second on average.

However, a stark difference emerges in their ability to estimate parameters accurately. On average, Linear Search achieves a 24% reduction in the Sum of Squared Errors (SSE) relative to K-means, with reductions as high as 84% in certain cases. The impact of this improvement on the variance between the true effect and the estimated effect is presented in Table 3.2, highlighting the superior precision of the Linear Search approach.

Table 3.2: Performance Comparison Between LS and GFE Models

Metric	**LS (Linear Search)**	**K-means**
Variance of Beta	12.18	603.61
Group Accuracy (%)	99.97	98.93
Runtime (seconds)	2.101	0.008

Notes: reduced SSE by at most 84.795% and on average 24.686%

To better understand where this variances is coming from in figure 3.5 we plot all of the betas from our 1000 simulations. Notice that there is a serious bias towards 0 for the K-means estimator. Furthermore notice that our distribution for our global optimizer is approximately normal whereas you can see the K-means is unusual. Theoretically, our distribution should be normal given our errors are normal showing that there is problem with estimating confidence intervals in K-means.

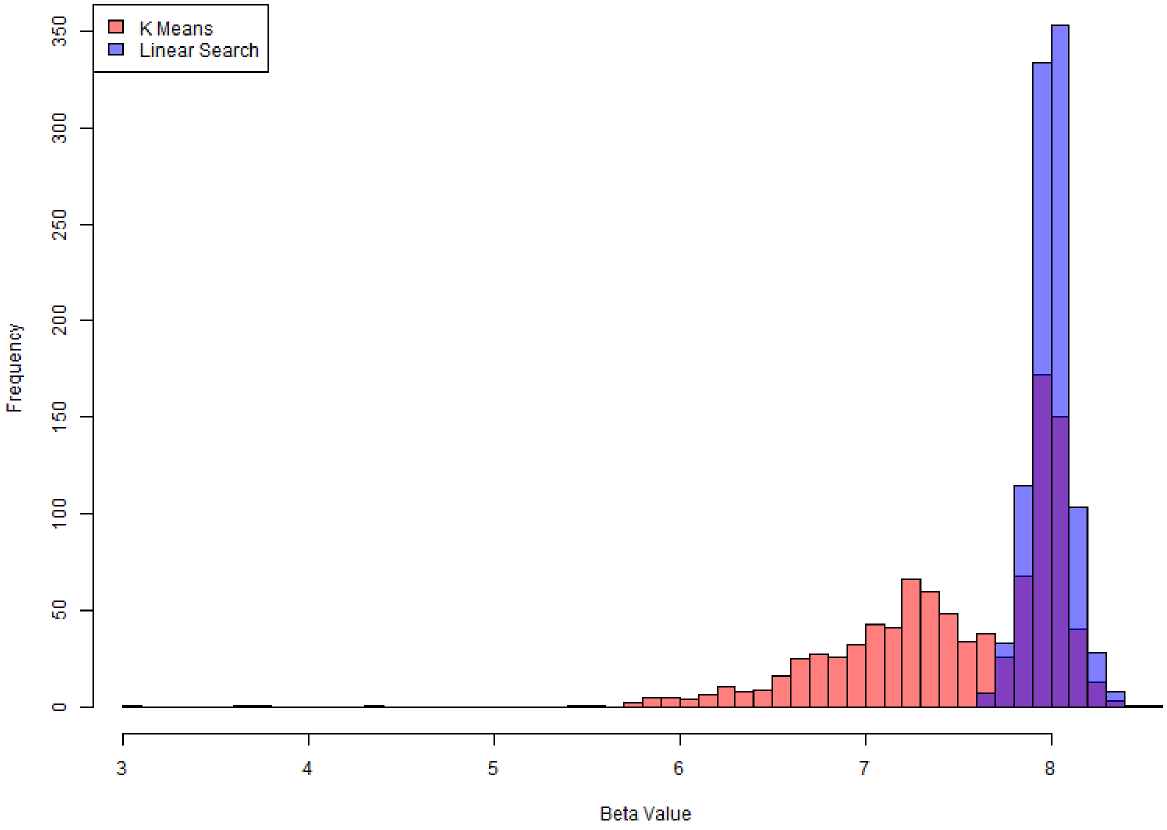


Figure 3.5: Example of Segmenting Data with Bounding Box

3.7 Early Stopping

As problem sizes increase, the computational burden of the optimization procedure also grows. Specifically, the presence of data points located near the true decision boundary significantly slows the optimization process. This phenomenon arises because these points make it challenging to separate groups, increasing the complexity of the search for the true slope. However, in practice, we observe that the algorithm efficiently identifies the general region where the true decision boundary, and thus the true slope lies. This suggests that stopping the algorithm early may be a practical approach if the region of uncertainty for the true slope becomes sufficiently narrow.

We simulate data for $N = 1,000,000$ entities, where the covariates x_i are drawn from a normal distribution with mean 5 and standard deviation 5. Random noise ϵ_i is added, drawn from a normal distribution with mean 0 and standard deviation 15. Entities are randomly assigned to one of two groups, with 500,000 entities in each group. Group membership affects the intercept: entities in Group 1 receive an intercept of 10, while entities in Group 2 receive an intercept of 70. The true slope β is set to 8. The outcome y_i is generated according to the model:

$$y_i = x_i\beta + \alpha_{g(i)} + \epsilon_i,$$

Where $\alpha_{g(i)}$ is the group-specific intercept. Due to the overlap in the covariates and random error, the data exhibit substantial overlap, with many points lying near the halfway point between the groups.

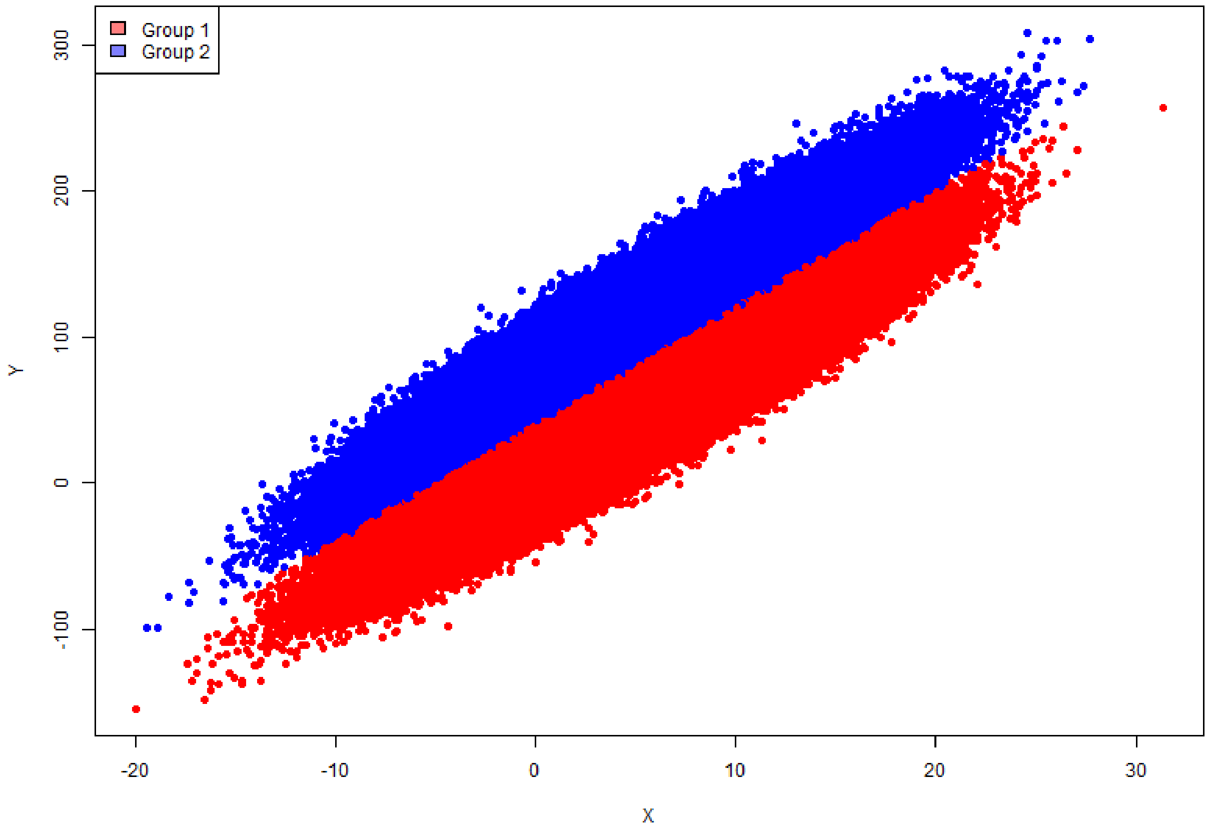


Figure 3.6: Data Generating Process

We run the optimization procedure for a full day, allowing the algorithm to refine the segmentation of the data. At the end of the run, we examine the remaining segments and calculate the maximum and minimum slopes (β) across all segments. This analysis reveals that the estimated slope lies within the range $[7.873, 8.107]$, while the true slope is $\beta = 8$.

The results demonstrate that, even for extremely large problems with significant group overlap, the optimization procedure can provide tight bounds on the true slope. These findings suggest that the method is robust and capable of delivering meaningful insights even under challenging conditions.

3.8 Summary

Addressing unobserved group heterogeneity in panel data presents significant challenges, as it requires simultaneous parameter estimation and group assignment. While Mixed Integer Optimization (MIO) provides a solution, in practice the NP-hard nature of the problem renders it impractical for large datasets. This section demonstrated how structurally constraining the regression line inherently constrains the shape and location of the decision boundary. By leveraging these constraints, we avoid directly searching over individual group assignments, focusing instead on the feasible region where the decision boundary must lie.

Simulations reveal that this approach is vastly more computationally efficient than MIO, solving problems in under a minute that would otherwise take years. Additionally, we highlighted the pitfalls of K-means clustering under certain distributions, where it can fail dramatically, underscoring the necessity of a global optimization framework for these problems. While there remains much to refine and extend within this framework, the initial results are promising.

3.9 Appendix

3.9.1 Total Number of Matched Segments

Let us divide each side into S segments. Our objective is to match the segments of each side with the segments of all other sides. First, observe that there are six possible pairings of the sides: Left-Right, Left-Bottom, Left-Top, Right-Bottom, Right-Top, and Bottom-Top.

For each of these six pairings, since each side has S segments, there are S^2 possible matches between the two sides. Therefore, the total number of matches across all pairings is:

$$6 \cdot S^2 = 6S^2.$$

3.9.2 High Dimensional Covariates

When the dimension of our covariate X is high, we increase the dimensions of both the rectangle and the decision boundary. Unfortunately, this also increases the number of matched segments that we need to search over. Beyond 4 or 5 dimensions, the curse of dimensionality creates a large search space, significantly slowing down the optimization procedure. This phenomenon poses challenges not only for our optimization procedure but also for any clustering method attempting to accurately identify clusters. In what follows, we provide a preliminary suggestion for addressing this issue. However, we believe further work is needed in this area.

Suppose we have the following model, where β is the effect of interest, and Z' is a covariate matrix of control variables:

$$y_{it} = x_{it}\beta + Z'_{it}\omega + \alpha_{g(i)t} + \epsilon_{it}.$$

The challenge arises because the inclusion of controls increases the dimensionality of the clustering problem, potentially degrading the optimization procedure's ability to identify the true clusters. To better understand this, let M be the orthogonal projection matrix associated with Z'. If we partial out Z', we obtain:

$$My = Mx\beta + M\alpha_{g(i)t} + M\epsilon_{it}.$$

This approach successfully reduces the problem to a two-dimensional space, where our optimization procedure has empirically performed well. However, if the grouping vector is correlated with any dimension of Z', this partialing out process may corrupt the grouping structure, violating the assumptions necessary for consistent estimation of group assignments.

Thus, our suggestion is to carefully decide which dimensions are likely to be correlated with the way individuals group themselves. By partialing out as many other dimensions as possible, the optimization procedure's ability to estimate groups can be enhanced, improving both accuracy and efficiency.

3.9.3 Classification of Groups

In step 3 in section 3.4 we mentioned group classification depended on its relative position of the decision lines. In what follows we describe exactly how we determine which group you are on given the sides of the segments you are matching between.

1. **Left-Right Segments:**
    - **group1:** Points are above all the decision line.
    - **group2:** Points are below all the decision line.
2. **Left-Top Segments:**
    - **group1:** Points are above the lines and to the left of a vertical line.
    - **group2:** Points are below the lines and to the right of a vertical line.
3. **Left-Bottom Segments:**

- **group1:** Points are below the lines and to the left of a vertical line.
- **group2:** Points are above the lines and to the right of a vertical line.

4. **Top-Bottom Segments:**
    - **group1:** Points are below a negatively sloped line, above a positively sloped line, and to the left of vertical lines.
    - **group2:** Points are above a negatively sloped line, below a positively sloped line, and to the right of vertical lines.
5. **Top-Right Segments:**
    - **group1:** Points are above the line and to the right of a vertical line.
    - **group2:** Points are below the line and to the left of a vertical line.
6. **Bottom-Right Segments:**
    - **group1:** Points are below the line and to the right of a vertical line.
    - **group2:** Points are above the line and to the left of a vertical line.

3.9.4 Aggregate Decision Boundary Proof

$$\begin{aligned}
&(\bar{y}_i - \hat{\hat{\alpha}}_g - \bar{x}_i'\hat{\theta})^2 = (\bar{y}_i - \hat{\hat{\alpha}}_{\tilde{g}} - \bar{x}_i'\hat{\theta})^2 \\
&\bar{y}_i^2 - 2\bar{y}_i(\hat{\hat{\alpha}}_g + \bar{x}_i'\hat{\theta}) + (\hat{\hat{\alpha}}_g + \bar{x}_i'\hat{\theta})^2 = \bar{y}_i^2 - 2\bar{y}_i(\hat{\hat{\alpha}}_{\tilde{g}} + \bar{x}_i'\hat{\theta}) + (\hat{\hat{\alpha}}_{\tilde{g}} + \bar{x}_i'\hat{\theta})^2 \\
&-2\bar{y}_i(\hat{\hat{\alpha}}_g + \bar{x}_i'\hat{\theta}) + (\hat{\hat{\alpha}}_g + \bar{x}_i'\hat{\theta})^2 = -2\bar{y}_i(\hat{\hat{\alpha}}_{\tilde{g}} + \bar{x}_i'\hat{\theta}) + (\hat{\hat{\alpha}}_{\tilde{g}} + \bar{x}_i'\hat{\theta})^2 \\
&-2\bar{y}_i(\hat{\hat{\alpha}}_g - \hat{\hat{\alpha}}_{\tilde{g}}) = (\hat{\hat{\alpha}}_{\tilde{g}} + \bar{x}_i'\hat{\theta})^2 - (\hat{\hat{\alpha}}_g + \bar{x}_i'\hat{\theta})^2 \\
&-2\bar{y}_i(\hat{\hat{\alpha}}_g - \hat{\hat{\alpha}}_{\tilde{g}}) = (\hat{\hat{\alpha}}_{\tilde{g}} + \bar{x}_i'\hat{\theta} + \hat{\hat{\alpha}}_g + \bar{x}_i'\hat{\theta})(\hat{\hat{\alpha}}_{\tilde{g}} - \hat{\hat{\alpha}}_g) \\
&2\bar{y}_i = (\hat{\hat{\alpha}}_{\tilde{g}} + \hat{\hat{\alpha}}_g + 2\bar{x}_i'\hat{\theta}) \\
&\bar{y}_i = \tfrac{\hat{\hat{\alpha}}_{\tilde{g}} + \hat{\hat{\alpha}}_g}{2} + \bar{x}_i'\hat{\theta}
\end{aligned} \tag{3.7}$$

3.9.5 Proof of Theorem 5

Previously we showed our estimator can take on the following form.

$$\bar{y}_i = \tfrac{\bar{\bar{y}}_{\tilde{g}}+\bar{\bar{y}}_g}{2} - \tfrac{\bar{\bar{x}}_{\tilde{g}}\hat{\theta}+\bar{\bar{x}}_g\hat{\theta}}{2} + \bar{x}_i'\hat{\theta} \tag{3.8}$$

If there are two groups $(g, \tilde{g})$ and the have equal number of individuals we can show averaging over the group averages is just the average of the data since $N_g = N_{\tilde{g}}$.

$$\begin{aligned}
&\tfrac{1}{2}\left(\textstyle\sum_{i\in G}\tfrac{\bar{y}_i}{N_g} + \sum_{i\in\tilde{G}}\tfrac{\bar{y}_i}{N_{\tilde{g}}}\right)\\
&= \tfrac{1}{2}\left(\textstyle\sum_{i\in G}\tfrac{\bar{y}_i}{N_g} + \sum_{i\in\tilde{G}}\tfrac{\bar{y}_i}{N_g}\right)\\
&= \textstyle\sum_{i=1}^{N}\tfrac{\bar{y}_i}{2N_g}\\
&= \bar{\bar{y}}
\end{aligned} \tag{3.9}$$

similarly we can show..

$$\begin{aligned}
&\tfrac{1}{2}\left(\textstyle\sum_{i\in G}\tfrac{\bar{x}_i'\hat{\theta}}{N_g} + \sum_{i\in\tilde{G}}\tfrac{\bar{x}_i'\hat{\theta}}{N_{\tilde{g}}}\right)\\
&= \tfrac{1}{2}\left(\textstyle\sum_{i\in G}\tfrac{\bar{x}_i'\hat{\theta}}{N_g} + \sum_{i\in\tilde{G}}\tfrac{\bar{x}_i'\hat{\theta}}{N_g}\right)\\
&= \textstyle\sum_{i=1}^{N}\tfrac{\bar{x}_i'\hat{\theta}}{2N_g}\\
&= \bar{\bar{x}}\hat{\theta}
\end{aligned} \tag{3.10}$$

Now we can transform 3.8 into the following.

$$\bar{y}_i = \bar{\bar{y}} - \bar{\bar{x}}\hat{\theta} + \bar{x}_i'\hat{\theta} \tag{3.11}$$

Notice if we set $\bar{x}_i' = \bar{\bar{x}}$ our answer is $\bar{\bar{y}}$. This makes sense, because if this was a 1 group regression it would run through the center of our data.

References

[AB16] Tomohiro Ando and Jushan Bai. "Panel data models with grouped factor structure under unknown group membership." *Journal of Applied Econometrics*, **31**(1):163–191, 2016.

[AS15] Padmaja Ayyagari and Dan M Shane. "Does prescription drug coverage improve mental health? Evidence from Medicare Part D." *Journal of health economics*, **41**:46–58, 2015.

[ASI20] Mohiuddin Ahmed, Raihan Seraj, and Syed Mohammed Shamsul Islam. "The k-means algorithm: A comprehensive survey and performance evaluation." *Electronics*, **9**(8):1295, 2020.

[Bai09] Jushan Bai. "Panel data models with interactive fixed effects." *Econometrica*, **77**(4):1229–1279, 2009.

[BH16] C Alan Bester and Christian B Hansen. "Grouped effects estimators in fixed effects models." *Journal of Econometrics*, **190**(1):197–208, 2016.

[BKM16] Dimitris Bertsimas, Angela King, and Rahul Mazumder. "Best Subset Selection VIA a Modern Optimization Lens." *The Annals of Statistics*, **44**(2):813–852, 2016.

[BM15] Stephane Bonhomme and Elena Manresa. "Grouped Patterns of Heterogeneity in Panel Data." *Econometrica*, **83**(3):1147–1184, 2015.

[BMH23] Jan Pablo Burgard, Carina Moreira Costa, Christopher Hojny, Thomas Kleinert, and Martin Schmidt. "Mixed-integer programming techniques for the minimum sum-of-squares clustering problem." *Journal of Global Optimization*, **87**(1):133–189, 2023.

[Boc08] Hans-Hermann Bock. "Origins and extensions of the k-means algorithm in cluster analysis." *Electronic journal for history of probability and statistics*, **4**(2):1–18, 2008.

[BPS24] Jan Pablo Burgard, Maria Eduarda Pinheiro, and Martin Schmidt. "Mixed-integer quadratic optimization and iterative clustering techniques for semi-supervised support vector machines." *TOP*, pp. 1–38, 2024.

[Chu17] Ba Chu. "Composite Quasi-Maximum Likelihood Estimation of Dynamic Panels with Group-Specific Heterogeneity and Spatially Dependent Errors." 2017.

[CKZ20] Janet Currie, Henrik Kleven, and Esmée Zwiers. "Technology and big data are changing economics: Mining text to track methods." In *AEA Papers and Proceedings*, volume 110, pp. 42–48. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, 2020.

[CM22] Denis Chetverikov and Elena Manresa. "Spectral and post-spectral estimators for grouped panel data models." *arXiv preprint arXiv:2212.13324*, 2022.

[DT97] Partha Deb and Pravin K Trivedi. "Demand for medical care by the elderly: a finite mixture approach." *Journal of applied Econometrics*, **12**(3):313–336, 1997.

[Gom10] Ralph E Gomory. *Outline of an algorithm for integer solutions to linear programs and an algorithm for the mixed integer problem.* Springer, 2010.

[GV14] Michael Gebel and Jonas Voßemer. "The impact of employment transitions on health in Germany. A difference-in-differences propensity score matching approach." *Social science & medicine*, **108**:128–136, 2014.

[HJS20] Wenxin Huang, Sainan Jin, and Liangjun Su. "Identifying latent grouped patterns in cointegrated panels." *Econometric Theory*, **36**(3):410–456, 2020.

[HM10] Jinyong Hahn and Hyungsik Roger Moon. "Panel data models with finite number of multiple equilibria." *Econometric Theory*, **26**(3):863–881, 2010.

[HN04] Jinyong Hahn and Whitney Newey. "Jackknife and analytical bias reduction for nonlinear panel models." *Econometrica*, **72**(4):1295–1319, 2004.

[Laz82] Rafael Lazimy. "Mixed-integer quadratic programming." *Mathematical Programming*, **22**:332–349, 1982.

[Llo82] Stuart Lloyd. "Least squares quantization in PCM." *IEEE transactions on information theory*, **28**(2):129–137, 1982.

[LW66] Eugene L Lawler and David E Wood. "Branch-and-bound methods: A survey." *Operations research*, **14**(4):699–719, 1966.

[MOS15] Renata Mansini, Wlodzimierz Ogryczak, and M Grazia Speranza. *Linear and mixed integer programming for portfolio optimization*, volume 21. Springer, 2015.

[MPR11] Florence Merlevède, Magda Peligrad, and Emmanuel Rio. "A Bernstein type inequality and moderate deviations for weakly dependent sequences." *Probability Theory and Related Fields*, **151**:435–474, 2011.

[PG11] Alessandra Parisio and Luigi Glielmo. "A mixed integer linear formulation for microgrid economic scheduling." In *2011 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, pp. 505–510. IEEE, 2011.

[Rio00] Emmanuel Rio. *Théorie asymptotique des processus aléatoires faiblement dépendants*, volume 31. Springer Science & Business Media, 2000.

[Riv23] Jorge A Rivero. "Unobserved Grouped Heteroskedasticity and Fixed Effects." *arXiv preprint arXiv:2310.14068*, 2023.

[SGZ13] Kelvin Sim, Vivekanand Gopalkrishnan, Arthur Zimek, and Gao Cong. "A survey on enhanced subspace clustering." *Data mining and knowledge discovery*, **26**:332–397, 2013.

[SP05] Richard Souvenir and Robert Pless. "Manifold clustering." In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pp. 648–653. IEEE, 2005.

[SSP16] Liangjun Su, Zhentao Shi, and Peter CB Phillips. "Identifying latent structures in panel data." *Econometrica*, **84**(6):2215–2264, 2016.

[Sun05] Yixiao X Sun. "Estimation and inference in panel structure models." 2005.