

UCLA

UCLA Electronic Theses and Dissertations

Title

Three Essays on Unobserved Heterogeneity in Panel and Network Data Models

Permalink

<https://escholarship.org/uc/item/9x23540z>

Author

Shang, Hualei

Publication Date

2020

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Three Essays on Unobserved Heterogeneity
in Panel and Network Data Models

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

by

Hualei Shang

2020

© Copyright by
Hualei Shang
2020

ABSTRACT OF THE DISSERTATION

Three Essays on Unobserved Heterogeneity
in Panel and Network Data Models

by

Hualei Shang

Doctor of Philosophy in Economics

University of California, Los Angeles, 2020

Professor Rosa Liliana Matzkin, Chair

This dissertation consists of three chapters that study unobserved heterogeneity in panel and network data models. In Chapter 1, I propose a semi-nonparametric panel data model with a latent group structure. I assume that individual parameters are heterogeneous across groups but homogeneous within a group while the group membership is unknown. I first approximate the infinite-dimensional function with a sieve expansion; then, I propose a Classifier-Lasso(C-Lasso) procedure to simultaneously identify the individuals' membership and estimate the group-specific parameters. I show that: (i) the classification exhibits uniform consistency; (ii) C-Lasso and post-Lasso estimators achieve oracle properties so that they are asymptotically equivalent to infeasible estimators as if the group membership is known; and (iii) the estimators are consistent and asymptotically normally distributed. Simulations demonstrate an excellent finite sample performance of this approach in both classification and estimation.

In Chapter 2 (joint with Wenyu Zhou), we study a nonparametric additive panel regression model with grouped heterogeneity. The model can be regarded as a natural extension to the heterogeneous panel model studied in Su, Shi, and Phillips (2016). We propose to estimate the nonparametric components using a sieve-approximation-based Classifier-Lasso

method. We establish the asymptotic properties of the estimator and show that they enjoy the so-called oracle property. In addition, we present the decision rule for group classification and establish its consistency. Then, a BIC-type information criterion is developed to determine the group pattern of each nonparametric component. We further investigate the finite sample performance of the estimation method and the information criterion through Monte Carlo simulations. Results show that both work well. Finally, we apply the model and the estimation method to study the demand for cigarettes in the United States using panel data of 46 states from 1963 to 1992.

In Chapter 3, I study a network sample selection model in which 1) bilateral fixed effects enter the pairwise outcome equation additively; 2) link formation depends on latent variables from both sides nonparametrically. I first propose a four-cycle structure to difference out the fixed effects; next, utilizing the idea proposed in Auerbach (2019), I manage to use the kernel function to control for the selection bias. I then introduce estimators for the parameters of interest and characterize their asymptotic properties.

The dissertation of Hualei Shang is approved.

Ying Nian Wu

Denis Nikolaye Chetverikov

Shuyang Sheng

Zhipeng Liao

Rosa Liliana Matzkin, Committee Chair

University of California, Los Angeles

2020

DEDICATIONS

To my parents and lovely niece

Contents

- 1 Semi-Nonparametric Panel Data Models with Latent Structures** **1**
- 1.1 Introduction 2
- 1.2 Penalized Sieve Estimation 7
 - 1.2.1 Semi-Nonparametric Panel Data Structure Models 7
 - 1.2.2 Sieve Approximation 8
 - 1.2.3 Penalized Estimation of α and f 10
- 1.3 Asymptotic Properties 11
 - 1.3.1 Assumptions 11
 - 1.3.2 Preliminary Rates of Convergence 17
 - 1.3.3 Classification Consistency 18
 - 1.3.4 The Oracle Property and Asymptotic Distributions 19
 - 1.3.5 Determination of Number of Groups 22
- 1.4 Simulation 24
 - 1.4.1 Data Generating Process 24
 - 1.4.2 Main Result 25
 - 1.4.3 Comparison with Complete Homogeneity and Heterogeneity 29
 - 1.4.4 Comparison with Misspecified Parametric model 31
- 1.5 Conclusion 33
- 1.A Proofs of the Main Results 34
- 1.B Proofs of Technical Lemmas 49

2	Nonparametric Additive Panel Regression Models with Grouped Heterogeneity	64
2.1	Introduction	65
2.2	Model	68
2.3	Estimation	69
2.3.1	Sieve Approximation	70
2.3.2	Penalized Estimation of h and f	71
2.4	Asymptotic Properties	73
2.4.1	Preliminary Rates of Convergence	73
2.4.2	Classification Consistency	78
2.4.3	The Oracle Property and Asymptotic Distributions	79
2.4.4	Determination of Number of Groups	81
2.5	Simulation	83
2.5.1	Data Generating Process	83
2.5.2	Simulation Results	86
2.6	Empirical Illustration	88
2.7	Conclusion	92
2.A	Proofs of the Main Results	93
2.B	Proofs of Technical Lemmas	104
3	A Network Sample Selection Model	114
3.1	Motivation	115
3.2	Model Setup	118
3.2.1	Explanation of the Link Formation Process	119
3.3	Estimation Strategy	121
3.4	Asymptotic Properties	125
3.4.1	Consistency	125
3.4.2	Asymptotic Distribution when ω_i has Finite Support	127

3.4.3	Asymptotic Distribution when ω_i is Continuous	128
3.5	Extension to Directed Networks	129
3.6	Conclusion	132

List of Figures

2.1	Estimated Functions of h_1	90
2.2	Estimated Functions of h_2	91
3.1	Trade Flows	117
3.2	Trade with Same Countries	120
3.3	Trade with Same Countries with Same Probabilities	120
3.4	Countries of the Same Type	120
3.5	Countries of Different Types	121
3.6	Four Cycle	122
3.7	Directed Four Cycle	131

List of Tables

1.1	Literature Review on Classification	6
1.2	RMSE (Maximum) of C-Lasso and post-Lasso Estimators in DGP 1	27
1.3	RMSE of C-Lasso and post-Lasso Estimators in DGP 1	28
1.4	Comparison with Complete Homogeneity and Heterogeneity in DGP 1	30
1.5	Comparison with Misspecified Parametric Model in DGP 1	31
1.6	Comparison with Misspecified Parametric Model in DGP 2	32
2.1	Simulation Results for Group-specific Parameters in DGP 1	87
2.2	Simulation Results for Group-specific Parameters in DGP 2	87
2.3	Simulation Results for Group-specific Parameters in DGP 3	88

ACKNOWLEDGMENT

I am grateful to Rosa Matzkin, Zhipeng Liao, Shuyang Sheng, Denis Chetverikov, and Ying Nian Wu for invaluable guidance and continuous support. I thank Jinyong Hahn, Andres Santos, and participants at UCLA econometrics proseminars for helpful comments and discussions. All errors, of course, are mine.

VITA

Education

University of California, Los Angeles	Los Angeles, USA
C.Phil., Economics, Department of Economics	2016
M.A., Economics, Department of Economics	2015
Peking University	Beijing, China
M.A., Economics, National School of Development	2014
Tsinghua University	Beijing, China
B.E., Engineering Physics, Department of Engineering Physics	2008

Fellowships and Awards

UCLA Dissertation Year Fellowship, 2019-2020
UCLA Economic Departmental Teaching Assistantship, 2015-2019
UCLA Economic Departmental Fellowship, 2014-2015
Graduate Dean's Scholar Award, UCLA, 2014-2015
Best Teaching Assistant Award, Peking University, 2011-2012
National Scholarship (Top 3 students), Peking University, 2011-2012

Teaching Experience

Instructor

Microeconomic Theory I, UCLA, Summer 2018

Teaching Assistant

Econometrics Laboratory, UCLA, Summer 2019, Spring 2017

Microeconomic Theory I, UCLA, Spring 2018, Spring 2016

Statistics for Economists, UCLA, Winter 2018, Fall 2018, Winter 2017

Microeconomic Theory II, UCLA, Fall 2016

Principles of Economics II, UCLA, Winter 2016

Principles of Economics I, UCLA, Fall 2015

Intermediate Macroeconomics, Peking University, Fall 2013, Spring 2012

Advanced Econometrics (Masters), Peking University, Fall 2012

Intermediate Econometrics, Peking University, Fall 2012

Chapter 1

Semi-Nonparametric Panel Data

Models with Latent Structures

1.1 Introduction

In semi-nonparametric panel data models, it is almost universal to assume that the regression parameters are the same across individuals, while unobserved heterogeneity is merely modeled through individual-specific effects. However, since most panel data cover cross-sectional units with different characteristics, to control for individual heterogeneity remains a challenge. One important task is how to model the influence of heterogeneity on the individual regression parameters. To tackle the problem while preserving the power of cross-sectional averaging, I propose a semi-nonparametric panel data model with a latent group structure.

I assume that individuals belong to different groups while the group identity is unknown a priori. Individual regression parameters are the same within the group but differ across groups. In Economics, the groups could be understood as different convergence clubs in the economic growth studies (Phillips and Sul (2007)), stock returns in different sectors in financial markets (Ke, Fan, and Wu (2015)), spatial geographic groupings in economic geography (Bester and Hansen (2016); Fan, Lv, and Qi (2011)) or multiplicity of Nash equilibria in game theory or Macroeconomics models (Hahn and Moon (2010)). Several important examples and policy implications will be discussed at the end of this section.

This group structure modeling reaches a good balance between its two alternatives: complete parameter homogeneity or complete parameter heterogeneity. Traditional panel data models always assume that individuals share the same parameters. Although this approach is easy to implement and achieves good convergence rate, homogeneity assumption has been frequently rejected in empirical studies; see Hsiao and Tahmiscioglu (1997), Lee, Pesaran, and Smith (1997), Durlauf, Kourtellos, and Minkin (2001), Phillips and Sul (2007), Browning and Carro (2007), Browning and Carro (2010), Su and Chen (2013) and Browning and Carro (2014). To the other extreme, if we allow for complete parameter heterogeneity, the key advantage of working with panel data is lost. If the time dimension is short, estimation could be very imprecise. See survey papers by Baltagi, Bresson, and Pirotte (2008) and Hsiao and Pesaran (2008). Compared with the two pieces of literature above, the group

structure approach simultaneously alleviates the misspecification problem common in the first one and preserves the power of cross-section averaging lost in the second one.

In the literature of panel structure modeling, there are two dimensions to consider. First, whether the parameters of interest are finite or infinite-dimensional; Second, what approach to use. Please see Table 1.1 for a summary. I discuss the literature mainly according to the approaches they implement but will also mention the parameters of interest in the process.

First, the k-means algorithm or its variants are commonly used to classify individuals into different groups. Lin and Ng (2012) and Sarafidis and Weber (2015) studied linear panel data models with finite dimensional coefficients following some group structure. Bonhomme and Manresa (2015) focuses on the grouped patterns of time-varying fixed effects. Ando and Bai (2014), Ando and Bai (2016) and Ando and Bai (2017) generalized Bonhomme and Manresa (2015) and studied panel data models where interactive fixed effects exhibit some group structure. Abraham et al. (2003), Luan and Li (2003), Chiou and Li (2007) and Tarpey (2007) applied the k-means algorithm or its variants to different realizations of random curves that depend on a deterministic index $t \in \mathcal{T}$.

Another approach, called classifier-Lasso (C-lasso), proposed by Su, Shi, and Phillips (2016), treated clustering as a process of shrinking individual-specific coefficients into some group-specific parameters. They imposed the group structure on finite dimensional parameters. Su and Ju (2018) extended this method to include interactive fixed effects. Su, Wang, and Jin (2019) assumed that time-varying coefficients follow some group structures.

There also exist some other classifying methods. Ke, Fan, and Wu (2015) proposed a clustering algorithm in regression via data-driven segmentation (CARDS). Wang, Phillips, and Su (2018) further generalized it into the panel data. Vogt and Linton (2017) implemented a thresholding method combining with kernel estimation to classify nonparametric functions into different groups. Vogt and Linton (2020) further developed a clustering method that does not rely on any smooth parameters, like the bandwidth or number of basis functions.

This paper follows the C-Lasso approach (Su, Shi, and Phillips (2016)) but considers

semi-nonparametric panel data models instead. C-lasso enjoys several significant advantages over the k-means algorithm and other alternatives. First, it allows some individuals left unclassified, adding more flexibility to the model. Second, the k-means method relies heavily on the initial values of the group identity, while C-Lasso is not sensitive to that. Third, the computation burden of k-means is more significant than that of C-lasso. Finally, C-lasso could be easily combined with some other methods.

Practically, my method could be separated into two steps. I first approximate the infinite-dimensional functions with a sieve expansion and then use C-Lasso to shrink individual-specific coefficients of basis functions into some group-specific parameters.

The main contribution of this paper is that I generalize the latent group structures from parametric to semi-nonparametric panel data models. Thus, further exploration beyond the parametric specification of the unobserved heterogeneity in response mechanisms becomes possible. Although Su, Wang, and Jin (2019), Vogt and Linton (2017) and Vogt and Linton (2020) also considered clustering of functions, in Su, Wang, and Jin (2019), the regressor is one-dimensional deterministic (t/T) while in my paper, they could be multiple-dimensional general random variables. The approaches in Vogt and Linton (2017) and Vogt and Linton (2020) are difficult to be applied to partially linear models; however, in my research, partially linear and nonparametric models are of no significant difference. So far as I know, my paper is the first one in the literature to impose group structures to semi-nonparametric panel data models flexibly.

I also contribute to the extensive literature of estimation in semi-nonparametric panel data models, including, but not limited to, partially linear and nonparametric panel data models. In addition to the estimation, my approach simultaneously identifies individuals' membership. However, this doesn't affect the asymptotic properties of the estimators, which are equivalent to those of the oracle estimators that use individual group identity information. The latter are well studied in the literature. For detailed discussions, I direct readers to survey papers by Su and Ullah (2011), Ai and Li (2008) and Ullah and Roy (1998).

To further illustrate applications of my method, I discuss the following three examples:

Example 1 (Learning Curve): In Atkin, Khandelwal, and Osman (2017), the authors conducted a random experiment for rug producers in Egypt. They generated exogenous variation in access to foreign markets and studied the impact of exporting on firm performance.

The most crucial step is to estimate how the quality changes as the volume of production increases, i.e., the learning curves. They assumed that different firms share the same learning curve.

However, due to unobserved heterogeneity (for example, the management levels of owners or proficiencies of workers in different firms might differ.), it might not be appropriate to make such a homogeneity assumption. My approach then would complement their study to further explore the heterogeneity of different firms in terms of learning.

Example 2 (Trade Cost): Atkin and Donaldson (2015) used newly collected CPI microdata from Ethiopia and Nigeria to study how cost-shifting characteristics (such as distance) affect the spatial price gaps.

However, distance is only an imperfect proxy for measuring transportation costs. The origin-destination paths exhibit considerable unobserved heterogeneity (for example, the quality of the roads is unobserved.). Although the authors also tried the quickest-route travel time measure as a more plausible alternative for the geographic distance, the same concern remains.

My method, on the other hand, would help to capture the heterogeneity of routes by merely imposing a group structure (high quality and low quality roads) on the effect of distance on price gaps.

Example 3 (Policy Analysis): Clemens, Lewis, and Postel (2018) evaluated the labor market effects of abrogation of the manual laborer (Bracero) agreements between the United States and Mexico. They estimated how the exclusion of Mexican farmworkers affect the employment and wages of domestic workers.

To study the heterogeneity of the effects, they split the states of the US into three groups using Bracero fraction (B/L , the ratio of Bracero workers and the whole labor force) as a criterion: no exposure with $B/L = 0$, low exposure with $0 < B/L < 0.2$ and high exposure with $B/L > 0.2$.

Even though this criterion might capture some heterogeneity of the influence of the policy on different states, it would be useful to use my approach at least as a robustness check. I could automatically accomplish the classification and estimation with an additional harmless assumption that the effect could be expressed as a time-varying function. The advantage, however, is to avoid any subjective judgment which might be arbitrary.

Table 1.1: Literature Review on Classification

Approaches Parameters of Interest	k-means Or its Variants	Classifier-Lasso	Other Approaches
Finite Dimensional	Lin and Ng (2012) Sarafidis and Weber (2015) Bester and Hansen (2016)	Su, Shi, and Phillips (2016) Su and Ju (2018)	Ke, Fan, and Wu (2015) Wang, Phillips, and Su (2018)
Infinite Dimensional	Bonhomme and Manresa (2015) Ando and Bai (2014) Ando and Bai (2016) Ando and Bai (2017) Abraham et al. (2003) Luan and Li (2003) Chiou and Li (2007) Tarpey (2007)	Su and Ju (2018) Su, Wang, and Jin (2019)	Vogt and Linton (2017) Vogt and Linton (2020)

The rest of the paper is organized as follows. Section 1.2 discusses the model. Section 1.3 presents the estimation and inference results. Section 1.4 reports Monte Carlo simulation findings. Section 1.5 concludes. All proofs of the main results are given in the Appendix.

Notation: Throughout the paper, I consider the case that (N, T) pass jointly to infinity, which is denoted as $(N, T) \rightarrow \infty$. For any real value matrix A , I write the transpose A' , the Frobenius norm $\|A\|_F \equiv (\text{tr}(AA'))^{\frac{1}{2}}$ and the Moore-Penrose inverse A^- . When A is symmetric, I denote $\mu_{\max}(A)$ and $\mu_{\min}(A)$ as its largest and lowest eigenvalues, respectively. For a square integrable function f defined on the support Ω , $\|f\|_2$ denotes its L^2 norm: $\|f\|_2 \equiv \left\{ \int_{\Omega} |f(x)|^2 dx \right\}^{\frac{1}{2}}$. The operator \xrightarrow{P} means convergence in probability, \xrightarrow{D} convergence in distribution. $\alpha \asymp \beta$ denotes that α and β are of the same magnitude, i.e., $\alpha = O(\beta)$ and

$\beta = O(\alpha)$. I use superscript 0 to denote the true values of parameters.

1.2 Penalized Sieve Estimation

In this section, I assume that the number of groups K^0 is known and will discuss in Section 1.3.5 how to determine it.

1.2.1 Semi-Nonparametric Panel Data Structure Models

I mainly focus on the partially linear model, since 1) all the results hold for nonparametric models as long as the conditions of finite-dimensional parameters are excluded. 2) it is more involving to develop the theory for partially linear models. I will briefly mention how to apply the method into nonparametric models when necessary.

A partially linear model in panel data takes the following form:

$$y_{it} = \mu_i + \omega'_{it}\beta_i + h_i(x_{it}) + u_{it} \quad u_{it} = \sigma_i(\omega_{it}, x_{it})\varepsilon_{it} \quad (1.1)$$

where $i = 1, 2, \dots, N$, $t = 1, 2, \dots, T$. ω_{it} is a $p \times 1$ vector of regressors. x_{it} is a $d \times 1$ vector of controls that affect the outcome through $h_i(x_{it})$. μ_i 's represent the unobserved individual fixed effects which might be correlated with ω_{it} and x_{it} . ε_{it} has mean 0 and variance 1 and is independent of $\{\omega_{it}, x_{it}\}$, so u_{it} is the error term with mean 0 and variance $\sigma_i^2(\omega_{it}, x_{it})$ conditional on $\{\omega_{it}, x_{it}\}$.

I denote the true value of β_i as β_i^0 , and $h_i(x_{it})$ as $h_i^0(x_{it})$ with a compact support \mathcal{X} . I assume that the finite-dimensional parameters β_i 's and infinite-dimensional functions h_i 's exhibit the following group pattern

$$\beta_i^0 = \sum_{k=1}^{K^0} \alpha_k^0 \mathbf{1}\{i \in G_k^0\} \quad (1.2)$$

$$h_i^0(x_{it}) = \sum_{k=1}^{K^0} f_k^0(x_{it}) \mathbf{1}\{i \in G_k^0\} \quad \text{for any } x_{it} \in \mathcal{X} \quad (1.3)$$

which means that individuals within group k share the same parameter α_k^0 and same function f_k^0 . $\{G_k^0, k = 1, 2, \dots, K^0\}$ are mutually exclusive, meaning that $\cup_{k=1}^{K^0} G_k^0 = \{1, 2, \dots, N\}$, and $G_k^0 \cap G_j^0 = \emptyset$ if $j \neq k$. $N_k = \#G_k^0$ denotes the cardinality of G_k^0 , and obviously $\sum_{k=1}^{K^0} N_k = N$. The notations I use are consistent with Su, Shi, and Phillips (2016).

Following Sun (2005), Lin and Ng (2012), Bonhomme and Manresa (2015) and Su, Shi, and Phillips (2016), I assume that individual group identity doesn't change over time. Let $\alpha = (\alpha_1, \dots, \alpha_{K^0})'$, $f = (f_1, \dots, f_{K^0})'$ and denote the corresponding true values as α^0 and f^0 , respectively.

The goal is to determine individuals' group identities and to estimate the group-specific parameters α and f .

Remark. For nonparametric panel data models, equation 1.1 becomes

$$y_{it} = \mu_i + h_i(x_{it}) + u_{it} \quad u_{it} = \sigma_i(\omega_{it}, x_{it})\varepsilon_{it} \quad (1.4)$$

I no longer have β_i and only need to focus on h_i , $i = 1, \dots, N$, and f_k , $k = 1, \dots, K^0$. The group structure is shown in equation 1.3 and the parameter of interest is group-specific f .

1.2.2 Sieve Approximation

I propose first to approximate h_i , $i = 1, \dots, N$ and f_k , $k = 1, \dots, K^0$ by a linear combination of a tensor-product linear sieve basis. A tensor product linear sieve is the product of univariate sieves. In this paper, I focus on univariate B-splines of order κ (or degree $\kappa - 1$).

I assume that $f_k(x_{it})$, $k = 1, \dots, K^0$ share the same compact support, which is, with loss of generality, normalized to $[0, 1]^d$. Following Chen (2007) and Ai and Chen (2003), I consider the Hölder space $\Lambda^r([0, 1]^d)$ of order $r > 0$. Let \underline{r} denote the largest integer satisfying $\underline{r} < r$. The Hölder space is a space of functions $f : [0, 1]^d \rightarrow \mathcal{R}$ such that the first \underline{r} derivatives are bounded, and the \underline{r} -th derivatives are Hölder continuous with the exponent $r - \underline{r} \in (0, 1]$.

The Hölder space becomes a Banach space when endowed with the Hölder norm:

$$\|f\|_{\Lambda^r} = \sup_x |f(x)| + \max_{a_1+a_2+\dots+a_d=r} \sup_{x \neq x'} \frac{|\nabla^a f(x) - \nabla^a f(x')|}{(\|x - x'\|_F)^{r-|a|}} < \infty$$

where for any $d \times 1$ nonnegative vector $a = (a_1, \dots, a_d)'$, I write $|a| = a_1 + \dots + a_d$ and denote the $|a|$ th derivative of function g as

$$\nabla^a f(x) = \frac{\partial^{|a|}}{\partial x_1^{a_1} \dots \partial x_d^{a_d}} f(x)$$

A Hölder ball with radius c is defined as $\Lambda_c^r([0, 1]^d) \equiv \{f \in \Lambda^r([0, 1]^d) : \|f\|_{\Lambda^r} \leq c < \infty\}$. It is known that functions in $\Lambda_c^r([0, 1]^d)$ could be uniformly well approximated by B-splines of order $\kappa \geq r + 1$. Let $B^J(x_{it})$ denote $J \times 1$ basis functions, then I could approximate $h_i(x_{it})$ and $f_k(x_{it})$ by $B^J(x_{it})'\gamma_i$ and $B^J(x_{it})'\pi_k$, respectively, where γ_i and π_k are $J \times 1$ vectors:

$$\begin{aligned} h_i(x_{it}) &= B^J(x_{it})'\gamma_i + \delta_{h_i}(x_{it}) \quad i = 1, \dots, N \\ f_k(x_{it}) &= B^J(x_{it})'\pi_k + \delta_{f_k}(x_{it}) \quad k = 1, \dots, K^0 \end{aligned}$$

where $\delta_{h_i}(x_{it})$ and $\delta_{f_k}(x_{it})$ are the corresponding approximation errors.

Then I could rewrite 1.1 as

$$y_{it} = \mu_i + \omega'_{it}\beta_i + B^J(x_{it})'\gamma_i + e_{it} \tag{1.5}$$

where $e_{it} = \delta_{h_i}(x_{it}) + u_{it}$.

Define $z_{it} \equiv (\omega'_{it}, \sqrt{J}B^J(x_{it})')'$ and $\theta_i \equiv (\beta'_i, \frac{1}{\sqrt{J}}\gamma'_i)'$, $i = 1, \dots, N$, it could be expressed as

$$y_{it} = \mu_i + z'_{it}\theta_i + e_{it} \tag{1.6}$$

where $\frac{1}{\sqrt{J}}$ is the normalization parameter.

At the same time, 1.3 becomes

$$\gamma_i^0 = \sum_{k=1}^{K^0} \pi_k^0 \mathbf{1}\{i \in G_k^0\}$$

Let $\eta_k = \left(\alpha'_k, \frac{1}{\sqrt{J}}\pi'_k\right)'$, 1.2 and 1.3 could be compressed as

$$\theta_i^0 = \sum_{k=1}^{K^0} \eta_k^0 \mathbf{1}\{i \in G_k^0\} \quad (1.7)$$

Remark. For nonparametric panel data models, equation 1.7 becomes

$$\frac{1}{\sqrt{J}}\gamma_i^0 = \sum_{k=1}^{K^0} \frac{1}{\sqrt{J}}\pi_k^0 \mathbf{1}\{i \in G_k^0\}$$

Furthermore, I need to change θ and η to $\frac{1}{\sqrt{J}}\gamma$ and $\frac{1}{\sqrt{J}}\pi$ respectively whenever possible.

Note that I keep the normalization factor $\frac{1}{\sqrt{J}}$ to emphasize that I focus on the normalized parameters for simplicity.

1.2.3 Penalized Estimation of α and f

Given the model specified in 1.6, I first take the deviation from the mean across individuals to concentrate out the individual effects μ_i 's and obtain

$$y_{it} - \bar{y}_i = (z_{it} - \bar{z}_i)' \theta_i + e_{it} - \bar{e}_i \quad (1.8)$$

where $\bar{y}_i = \frac{1}{T} \sum_{t=1}^T y_{it}$, with similar definitions for \bar{z}_i and \bar{e}_i .

For simplicity, I further define $\tilde{y}_{it} = y_{it} - \bar{y}_i$ and similarly for \tilde{z}_{it} , \tilde{e}_{it} , then 1.8 could be compressed as

$$\tilde{y}_{it} = \tilde{z}'_{it} \theta_i + \tilde{e}_{it} \quad (1.9)$$

To estimate θ_i , I minimize the following least square criterion function:

$$Q_{NT}(\theta) = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T (\tilde{y}_{it} - \tilde{z}'_{it} \theta_i)^2 \quad (1.10)$$

where $\theta = (\theta_1, \dots, \theta_N)$.

To include the latent group structure in my model, I propose to estimate θ and η by minimizing the following criterion function:

$$Q_{NT,\lambda}(\theta, \eta) = Q_{NT}(\theta) + \frac{\lambda}{N} \sum_{i=1}^N \prod_{k=1}^{K^0} \|\theta_i - \eta_k\|_F \quad (1.11)$$

where λ is the tuning parameter. The additional penalty item is used to shrink the individual parameters θ_i , $i = 1, \dots, N$ to particular unknown group-specific parameters η_k , $k = 1, \dots, K^0$ while at the same time to classify individuals into a priori unknown groups.

1.3 Asymptotic Properties

This section include 5 subsections. They are organized as follows: in Subsection 1.3.1, I make general assumptions about the model. Based on that, I characterize the preliminary convergence rates for individual coefficients θ_i , $i = 1, \dots, N$ and group-specific parameters η_k , $k = 1, \dots, K^0$ in Subsection 1.3.2. Subsection 1.3.3 presents the results of classification consistency. After that, Subsection 1.3.4 reports the asymptotically distribution of group-specific parameters α_k and f_k , $k = 1, \dots, K^0$. Subsection 1.3.5 discusses how to determine the number of groups.

1.3.1 Assumptions

Assumption 1.1. (i) For each $i = 1, \dots, N$, $\{\omega_{it}, x_{it}, \varepsilon_{it}\}$ is stationary strong mixing with mixing coefficient $\alpha_i(\cdot)$. $\alpha(\cdot) \equiv \max_{i \leq i \leq N} \alpha_i(\cdot)$ satisfies $\alpha(j) \leq c_\alpha \exp(-\rho j)$ for some $0 < c_\alpha < \infty$, $0 < \rho < \infty$. $\{\omega_{it}, x_{it}, \varepsilon_{it}\}$ are independent across i .

(ii) There exists positive \bar{c} such that $\max_{1 \leq i \leq N} \mathbf{E} \|\omega_{it}\|_F^q < \bar{c} < \infty$ and $\max_{1 \leq i \leq N} \mathbf{E} \|u_{it}\|_F^q < \bar{c} < \infty$ for some $q > 6$.

(iii) For the parametric component,

(i) ω_{it} does not contain 1.

(ii) Let \mathcal{B} denote the parameter space for β_i . \mathcal{B} is compact and convex subset of \mathcal{R}^p such that β_i^0 lies in the interior of \mathcal{B} for each i .

(iv) For the nonparametric component,

(i) For $k = 1, \dots, K^0$, $\mathbf{E}[f_k(x_{it})] = 0$.

(ii) For $k = 1, \dots, K^0$, $f_k^0 \in \mathcal{F} = \Lambda_c^{r_1}([0, 1]^d)$ with $r_1 > 0$.

(iii) For each $i = 1, \dots, N$, denote the marginal density function of $\{x_{it}\}$ as $f(x_{i.})$, then there exist positive constants \underline{c} and \bar{c} such that

$$0 < \underline{c} < \min_{1 \leq i \leq N} \inf_{x_{i.} \in [0, 1]^d} \{f(x_{i.})\} \leq \max_{1 \leq i \leq N} \sup_{x_{i.} \in [0, 1]^d} \{f(x_{i.})\} < \bar{c} < \infty$$

(v) There exist $\underline{c} > 0$ such that

$$\min_{1 \leq j \neq k \leq K^0} \left\{ \|\alpha_j^0 - \alpha_k^0\|_F^2 + \|f_j^0 - f_k^0\|_2^2 \right\} > \underline{c}$$

(vi) For $j = 1, \dots, p$, $\mathbf{E}[\omega_{it}^j | x_{it}] \in \mathcal{F} = \Lambda_c^{r_2}([0, 1]^d)$ with $r_2 > 0$.

(vii) There exist positive constants \underline{c} and \bar{c} such that

$$0 < \underline{c} < \min_{1 \leq i \leq N} \mu_{\min}(\text{Var}(z_{it})) \leq \max_{1 \leq i \leq N} \mu_{\max}(\text{Var}(z_{it})) < \bar{c} < \infty$$

$$0 < \underline{c} < \min_{1 \leq i \leq N} \mu_{\min}(\text{Var}(\omega_{it})) \leq \max_{1 \leq i \leq N} \mu_{\max}(\text{Var}(\omega_{it})) < \bar{c} < \infty$$

(viii) $\frac{N_k}{N} \rightarrow \tau_k$ for each $k = 1, \dots, K^0$ as $N \rightarrow \infty$. There exists positive constants \underline{c} and \bar{c} such that $0 < \underline{c} < \min_{1 \leq k \leq K^0} \{\tau_k\} \leq \max_{1 \leq k \leq K^0} \{\tau_k\} < \bar{c} < 1$

Assumption 1.1(i) implies that the strong mixing coefficients $\alpha(l)$ decay exponentially fast to 0 as $l \rightarrow \infty$ uniformly. Similar conditions are assumed in Su, Shi, and Phillips (2016), Su, Wang, and Jin (2019), Vogt and Linton (2017), etc. For more discussions on this, I refer readers to Su, Wang, and Jin (2019). Assumption 1.1(ii) imposes the moment condition restrictions for ω_{it} and u_{it} . Assumption 1.1(iii) specifies restrictions on the parametric component. The first part means that I do not include the intercept in the parametric component. The second part imposes restrictions on the finite dimensional parameter space, which is commonly assumed in the literature.

Assumption 1.1(iv) imposes restrictions on the nonparametric component. The first part is a harmless normalization. The second one is the smooth condition such that I could approximate any function $f_k \in \mathcal{F}$ well using the tensor-product of univariate B-splines. By the approximation theory, there exists $\pi_k \in \mathcal{R}^J$ such that

$$\sup_{x \in [0,1]^d} \|f_k(x) - B^{J'} \pi_k\|_{\infty} = O(J^{-\frac{\tau_1}{d}})$$

Similarly, for each individual, there exists γ_i such that

$$\sup_{x \in [0,1]^d} \|h_i(x) - B^{J'} \gamma_i\|_{\infty} = O(J^{-\frac{\tau_1}{d}})$$

Then, after controlling for the approximation error, the difference between $f_k(x)$ and $h_i(x)$ is reflected by the difference between π_k and γ_i . The third part is also assumed in Vogt and Linton (2017). First, it makes the functions $h_i(x_{it})$ comparable across individuals. Second, it guarantees that $h_i(x_{it})$ could be estimated uniformly well.

Assumption 1.1(v) specifies that the group-specific parameters are well separated from each other. This condition considers the parametric and nonparametric parameters simultaneously. Most importantly, it implies that the group-specific vectors are well separated from

each other. Consider $\|f_j^0 - f_k^0\|_2$ first,

$$\begin{aligned}
& \|f_j^0 - f_k^0\|_2 \\
& \leq \|f_j^0 - B^{J'}\pi_j\|_2 + \|f_k^0 - B^{J'}\pi_k\|_2 + \left\| \sqrt{J}B^{J'} \left(\frac{1}{\sqrt{J}}(\pi_j - \pi_k) \right) \right\|_2 \\
& = O(J^{-\frac{r_1}{d}}) + \left\{ \left(\frac{1}{\sqrt{J}}(\pi_j - \pi_k) \right)' \int_{[0,1]^d} JB^J(x)B^J(x)' dx \left(\frac{1}{\sqrt{J}}(\pi_j - \pi_k) \right) \right\}^{\frac{1}{2}} \\
& \asymp \left\| \frac{1}{\sqrt{J}}(\pi_j - \pi_k) \right\|_F
\end{aligned}$$

where the last equation holds because the eigenvalues of $\int_{[0,1]^d} JB^J(x)B^J(x)' dx$ are bounded above and away from 0.

Similarly,

$$\begin{aligned}
& \left\| \frac{1}{\sqrt{J}}(\pi_j - \pi_k) \right\|_F \\
& \asymp \left\| \sqrt{J}B^{J'} \left(\frac{1}{\sqrt{J}}(\pi_j - \pi_k) \right) \right\|_2 \\
& \leq \|f_j^0 - f_k^0\|_2 + \|f_j^0 - B^{J'}\pi_j\|_2 + \|f_k^0 - B^{J'}\pi_k\|_2 \\
& = \|f_j^0 - f_k^0\|_2 + O(J^{-\frac{r_1}{d}}) \\
& \asymp \|f_j^0 - f_k^0\|_2
\end{aligned}$$

Thus $\|f_j^0 - f_k^0\|_2^2 \asymp \left\| \frac{1}{\sqrt{J}}(\pi_j - \pi_k) \right\|_F^2$, consequently

$$\begin{aligned}
& \|\alpha_j^0 - \alpha_k^0\|_F^2 + \|f_j^0 - f_k^0\|_2^2 \\
& \asymp \|\alpha_j^0 - \alpha_k^0\|_F^2 + \left\| \frac{1}{\sqrt{J}}(\pi_j - \pi_k) \right\|_F^2 \\
& = \|\eta_j^0 - \eta_k^0\|_F^2
\end{aligned}$$

where $\eta_k = \left(\alpha_k', \frac{1}{\sqrt{J}}\pi_k' \right)'$. I have transformed the difference between two groups into Euclidean

distance between two vectors. Similarly I could get that

$$\begin{aligned} & \|\beta_i - \alpha_k\|_F^2 + \|h_i - f_k\|_2^2 \\ & \asymp \|\theta_i - \eta_k\|_F^2 \end{aligned}$$

if $i \notin G_k^0$. This result guarantees that the penalty item in 1.11 could shrink the individual coefficients to some group-specific parameters.

Assumption 1.1(vi) imposes smooth conditions on the conditional expectation of ω_{it} given x_{it} . Similarly as the second part of Assumption 1.1(iv), this condition guarantees that I could approximate $\mathbf{E}[\omega_{it}|x_{it}]$ well with B-splines. There are two approximation errors involved in the semiparametric model if I aim to estimate the parametric parameters. For an excellent illustration, I refer to Chernozhukov et al. (2018).

Assumption 1.1(vii) is the identification condition with sieve approximation. As demonstrated in Section 1.2.3, I take the demean approach to get rid of the individual fixed effect, consequently requiring that $\mathbf{E}[\tilde{z}_{it}\tilde{z}'_{it}]$ is positive definite to identify the coefficients. The corresponding population value is $\text{Var}(z_{it})$. It is better to understand this condition by thinking of the partitioned matrix

$$\text{Var}(z_{it}) = \begin{bmatrix} \text{Var}(\omega_{it}) & \text{Cov}(\omega_{it}, \sqrt{J}B^J(x_{it})) \\ \text{Cov}(\sqrt{J}B^J(x_{it}), \omega_{it}) & \text{Var}(\sqrt{J}B^J(x_{it})) \end{bmatrix}$$

Consider $\text{Var}(\sqrt{J}B^J(x_{it}))$ first. Define $\check{B}^J(x) \equiv B^J(x) - \int_{[0,1]^d} B^J(x)dx$ and $\tilde{B}^J(x) \equiv B^J(x) - \mathbf{E}[B^J(x)]$. By the properties of B-splines, eigenvalues of $J \int_{[0,1]^d} \check{B}^J(x)\check{B}^J(x)'dx$ are bounded above and away from certain constant numbers. Combining the third part of Assumption 1.1(iv) and more properties of B-splines, I could get that eigenvalues of $J \int_{[0,1]^d} \tilde{B}^J(x)\tilde{B}^J(x)'dx$ are also bounded above and away from some constant numbers, say

$\bar{\mu}$ and $\underline{\mu}$, respectively. Furthermore, I could conclude that

$$\max_{1 \leq i \leq N} \mu_{\max} \left(\text{Var}(\sqrt{J}B^J(x_{it})) \right) \leq \bar{\mu} \bar{c}$$

and

$$\min_{1 \leq i \leq N} \mu_{\min} \left(\text{Var}(\sqrt{J}B^J(x_{it})) \right) \geq \underline{\mu} \underline{c}$$

Define $\tilde{S}pl(\kappa) \equiv \left\{ \tilde{B}^J(x)' a, x \in [0, 1]^d, a \in \mathcal{R}^J \right\}$ as the demeaned polynomial spline sieve of order κ (I choose the same order for all univariate B-splines). Define $p(x_{it})$ as the projection of $\mathbf{E}[\tilde{\omega}_{it}|x_{it}]$ onto $\tilde{S}pl(\kappa)$. For each $i = 1, \dots, N$, one sufficient condition for positive definiteness of $\text{Var}(z_{it})$ is that $\mathbf{E} \left[(\tilde{\omega}_{it} - p(x_{it})) (\tilde{\omega}_{it} - p(x_{it}))' \right]$ is positive definite. However, it is tedious to give lower-level conditions for the uniform positive definiteness of $\text{Var}(z_{it})$ for $i = 1, \dots, N$.

Assumption 1.1(viii) is commonly assumed in the classification literature, which implies that each group would include an asymptotically non-negligible number of individuals.

Assumption 1.2. As $(N, T) \rightarrow \infty$, $\lambda \rightarrow 0$, $J \rightarrow \infty$, $J^2(\ln T)^3 T^{-1} \rightarrow 0$, $N^2 T^{1-\frac{q}{2}} (\ln T)^{\frac{3q}{2}} \rightarrow 0$.

Assumption 1.2 specifies several restrictions on J , N and T . The condition $J^2(\ln T)^3 T^{-1} \rightarrow 0$ is very similar to Assumption 2 in Newey (1997) on independent observations, only up to a small logarithmic factor $(\ln T)^3$. The last condition requires that T cannot increase too slow compared with N . The intuition is clear: as T grows, more and more information of each individual is revealed, and it becomes easier to tell different observations from different groups apart. The q is the moment restriction I make in Assumption 1.1(ii), which is set to be larger than 6 to allow that N and T increase at the same rate.

Remark. For nonparametric panel data models, I could simply 1) exclude all the assumptions solely involving α and ω_{it} , e.g., Assumption (iii) and (vi) are no longer needed; 2) delete

the part with α and ω_{it} for assumptions with both α and f , e.g., Assumption (v) becomes:
 There exist $\underline{c} > 0$ such that

$$\min_{1 \leq j \neq k \leq K^0} \|f_j^0 - f_k^0\|_2^2 > \underline{c}.$$

Most of the changes are trivial, so I don't bother to list all of them.

1.3.2 Preliminary Rates of Convergence

The following result gives the preliminary rates of convergence for θ_i , $i = 1, \dots, N$ and η_k , $k = 1, \dots, K^0$.

Theorem 1.1. *Suppose Assumption 1.1, 1.2 hold, then*

(i) $\|\hat{\theta}_i - \theta_i^0\|_F = O_p(J^{-\frac{r_1}{d}} + J^{\frac{1}{2}}T^{-\frac{1}{2}} + \lambda)$ for $i = 1, 2, \dots, N$

(ii) $\frac{1}{N} \sum_{i=1}^N \|\hat{\theta}_i - \theta_i^0\|_F^2 = O_p(J^{-2\frac{r_1}{d}} + JT^{-1})$

(iii) $\|\hat{\eta}_{(k)} - \eta_k^0\|_F = O_p(J^{-\frac{r_1}{d}} + J^{\frac{1}{2}}T^{-\frac{1}{2}})$, for $k = 1, \dots, K^0$, where $(\hat{\eta}_{(1)}, \dots, \hat{\eta}_{(K^0)})$ is a suitable permutation of $(\hat{\eta}_1, \dots, \hat{\eta}_{K^0})$

Theorem 1.1(i) and (ii) give the pointwise and mean square convergence rate of $\hat{\theta}_i$. In Theorem 1.1(i), the first item, $J^{-\frac{r_1}{d}}$, comes from the approximation error. The second one, $J^{\frac{1}{2}}T^{-\frac{1}{2}}$, demonstrates the contribution of interaction between B-splines and the error term. Similar as other Lasso-like estimators, the penalty item is reflected by λ . However, in Theorem 1.1(ii), the penalty item disappears. I direct interested readers to the details in the proof. The convergence rate of η_k , similarly, does not depend on λ . It is worth emphasizing that the convergence rate of η_k depends on the mean square instead of the pointwise convergence rate of θ_i .

By Assumption 1.2, it is clear that $\hat{\theta}_i$ and $\hat{\eta}_{(k)}$ converges in probability to θ_i^0 and η_k^0 ,

respectively. For simplicity, I denote $\hat{\eta}_k$ as $\hat{\eta}_{(k)}$. I further define

$$\hat{G}_k = \{i \in \{1, \dots, N\} : \hat{\beta}_i = \hat{\alpha}_k\} \quad k = 1, \dots, K^0$$

which denote the set of individuals that are classified into group k .

1.3.3 Classification Consistency

To ensure the consistency of classification, I require more assumptions.

Assumption 1.3. *As $(N, T) \rightarrow \infty$, $\lambda T^{\frac{1}{2}} J^{-\frac{1}{2}} (\ln T)^{-3-v} \rightarrow \infty$, $\lambda J^{\frac{\tau_1}{d}} (\ln T)^{-v} \rightarrow \infty$, $T^{\frac{1}{2}} J^{-\frac{1}{2}} (\ln T)^{-3-v} \rightarrow \infty$ and $\lambda (\ln T)^v \rightarrow 0$ for some $v > 0$.*

Assumption 1.3 imposes restrictions on λ and some further ones on J . Intuitively, I require that λ dominates all the other errors from approximation or u_{it} such that the penalty item will take effect and shrink the individual coefficients to some group-specific parameters.

Following Su et al. (2016), I define

$$\begin{aligned} \hat{E}_{kNT,i} &\equiv \{i \notin \hat{G}_k | i \in G_k^0\} \\ \hat{F}_{kNT,i} &\equiv \{i \notin G_k^0 | i \in \hat{G}_k\} \end{aligned}$$

where $i = 1, \dots, N$ and $k = 1, \dots, K^0$. And $\hat{E}_{kNT} = \cup_{i \in G_k^0} \hat{E}_{kNT,i}$, $\hat{F}_{kNT} = \cup_{i \in \hat{G}_k} \hat{F}_{kNT,i}$. \hat{E}_{kNT} denotes the event of classifying individuals that belong to G_k^0 into groups other than \hat{G}_k ; and \hat{F}_{kNT} denotes the event of classifying individuals into \hat{G}_k but it turns out that they don't belong to G_k^0 .

The following theorem demonstrates that I achieve consistent classification.

Theorem 1.2. *Suppose Assumption 1.1, 1.2 and 1.3 hold, then*

$$(i) P(\cup_{k=1}^{K^0} \hat{E}_{kNT}) \leq \sum_{k=1}^{K^0} P(\hat{E}_{kNT}) \rightarrow 0 \text{ as } (N, T) \rightarrow \infty$$

(ii) $P(\cup_{k=1}^{K^0} \hat{F}_{kNT}) \leq \sum_{k=1}^{K^0} P(\hat{F}_{kNT}) \rightarrow 0$ as $(N, T) \rightarrow \infty$

Theorem 1.2 guarantees that with probability approaching 1, I correctly classify individuals in the same group, say G_k^0 , into one group \hat{G}_k , and those classified into the same group, \hat{G}_k , belong to one correct group G_k^0 .

There might exist some individuals that are not classified into any group \hat{G}_k , $k = 1, \dots, K^0$. However, as well explained in Su, Shi, and Phillips (2016), empirically, I could modify the classifier and classify individuals into the closest group, while theoretically, I can ignore the problem in the large sample.

In the simulation, since the sample size is small, I force every individual classified into some group. For every individual i , I classify it into \hat{G}_k if

$$k = \arg \min_{1 \leq j \leq K^0} \left\{ \left\| \hat{\theta}_i - \hat{\eta}_j \right\|_F \right\}$$

1.3.4 The Oracle Property and Asymptotic Distributions

The C-lasso method simultaneously accomplishes two tasks: to classify individuals into different groups and to estimate θ_i , $i = 1, \dots, N$, and η_k , $k = 1, \dots, K^0$. Given the estimated coefficients, I could conduct inference for the estimators I am interested in: $\hat{\alpha}_k$ and $\hat{f}_k(x)$, where $\hat{\alpha}_k$ is part of $\hat{\eta}_k$ and $\hat{f}_k(x)$ could be constructed by $\hat{f}_k(x) = \sqrt{J}B^J(x)' \hat{\eta}_k$.

An alternative strategy would be to implement the post-Lasso approach. Given the estimated groups \hat{G}_k , $k = 1, \dots, K^0$, I could pool the observations classified into the same group together and estimate group-specific parameters. I denote the post-Lasso estimators as $\hat{\alpha}_{\hat{G}_k}$ and $\hat{f}_{\hat{G}_k}(x)$.

My goal is to show that the C-lasso and post-Lasso estimators exhibit the oracle property, i.e., they are asymptotically equivalent to the infeasible estimators as if the group membership is known. Before I give precise results, more definitions and assumptions are required.

Let $u_i = (u_{i1}, u_{i2}, \dots, u_{iT})$. $\text{Var}(u_i|\omega_i, x_i) = \Sigma_i^{\frac{1}{2}} V_i \Sigma_i^{\frac{1}{2}}$, where

$$\Sigma_i = \text{diag}(\sigma_i^2(\omega_{i1}, x_{i1}), \dots, \sigma_i^2(\omega_{iT}, x_{iT}))$$

$$V_i = \mathbf{E}[\varepsilon_i \varepsilon_i']$$

Assumption 1.4. (i) For $k = 1, \dots, K^0$, there exists two positive constants \underline{c}_v and \bar{c}_v such that

$$0 < \underline{c}_v \leq \lim_{N, T \rightarrow \infty} \min_{i \in G_k^0} \mu_{\min}(V_i) \leq \lim_{N, T \rightarrow \infty} \max_{i \in G_k^0} \mu_{\max}(V_i) \leq \bar{c}_v \delta_{NT}$$

for some nondecreasing sequence δ_{NT} which satisfies $\delta_{NT} N^{-1} \rightarrow 0$ as $N, T \rightarrow \infty$.

(ii) There exists positive \bar{c} such that $\max_{1 \leq i \leq N} \mathbf{E} \|\omega_{it} \sigma_i(\omega_{it}, x_{it})\|_F^q < \bar{c} < \infty$ for $q > 6$.

(iii) Let $z_{it, \sigma} \equiv z_{it} \sigma_i(\omega_{it}, x_{it})$, $\omega_{it, \sigma} \equiv \omega_{it} \sigma_i(\omega_{it}, x_{it})$ and $B_{it, \sigma} \equiv \sqrt{J} B_{it}^J(x_{it}) \sigma_i(\omega_{it}, x_{it})$. There exist positive constants \underline{c} and \bar{c} such that

$$0 < \underline{c} < \min_{1 \leq i \leq N} \mu_{\min} \left(\text{Var} \left(z_{it, \sigma} \right) \right) \leq \max_{1 \leq i \leq N} \mu_{\max} \left(\text{Var} \left(z_{it, \sigma} \right) \right) < \bar{c} < \infty$$

$$0 < \underline{c} < \min_{1 \leq i \leq N} \mu_{\min} \left(\text{Var} \left(\omega_{it, \sigma} \right) \right) \leq \max_{1 \leq i \leq N} \mu_{\max} \left(\text{Var} \left(\omega_{it, \sigma} \right) \right) < \bar{c} < \infty$$

$$0 < \underline{c} < \min_{1 \leq i \leq N} \mu_{\min} \left(\text{Var} \left(B_{it, \sigma} \right) \right) \leq \max_{1 \leq i \leq N} \mu_{\max} \left(\text{Var} \left(B_{it, \sigma} \right) \right) < \bar{c} < \infty$$

The Assumptions are analogous to Assumption A.3 in Su, Wang, and Jin (2019). Assumption 1.4(i) imposes restrictions on the covariance matrix of ε_i . Assumption 1.4(ii) specifies more moment conditions. The first condition in Assumption 1.4(iii) assures that the eigenvalues of the interactive items of z_{it} and the error term are bounded above and away from 0 uniformly. Moreover, since I am interested in α_k and $f_k(x)$ instead of η_k , the other two conditions are required.

Assumption 1.5. (i) As $(N, T) \rightarrow \infty$, $NTJ^{-2\frac{r_1}{d}} J^{-2\frac{r_2}{d}} \rightarrow 0$.

(ii) As $(N, T) \rightarrow \infty$, $NTJ^{-2\frac{r_1}{d}} \rightarrow 0$.

Assumption 1.5(i) is used to guarantee that the group-specific finite-dimensional estimators, $\hat{\alpha}_k$ and $\hat{\alpha}_{\hat{G}_k}$, achieves \sqrt{NT} convergence rate. Assumption 1.5(ii), on the other hand, is used to establish the pointwise convergence rate of the group-specific infinite-dimensional estimators $\hat{f}_k(x)$ and $\hat{f}_{\hat{G}_k}(x)$.

The following theorem establishes the asymptotic distribution of α_k .

Theorem 1.3. *Suppose Assumption 1.1, 1.2, 1.3, 1.4 and 1.5(i) hold. Then for any $k \in \{1, \dots, K^0\}$,*

(i)

$$\sqrt{N_k T} V_{k, \omega}^{-\frac{1}{2}} \left(\hat{\alpha}_k - \alpha_k^0 \right) \xrightarrow{D} N(0, 1)$$

(ii)

$$\sqrt{N_k T} V_{k, \omega}^{-\frac{1}{2}} \left(\hat{\alpha}_{\hat{G}_k} - \alpha_k^0 \right) \xrightarrow{D} N(0, 1)$$

where

$$V_{k, \omega} = \left(\hat{Q}_{G_k^0, \tilde{\omega} \setminus \tilde{B}} \right)^{-1} \frac{1}{N_k} \sum_{i \in G_k^0} \frac{1}{T} W'_{i, \tilde{\omega} \setminus \tilde{B}} \Sigma_i^{\frac{1}{2}} V_i \Sigma_i^{\frac{1}{2}} W_{i, \tilde{\omega} \setminus \tilde{B}} \left(\hat{Q}_{G_k^0, \tilde{\omega} \setminus \tilde{B}} \right)^{-1}$$

in which

$$\hat{Q}_{G_k^0, \tilde{\omega} \setminus \tilde{B}} = \hat{Q}_{G_k^0, \tilde{\omega} \tilde{\omega}} - \hat{Q}_{G_k^0, \tilde{\omega} \tilde{B}} \hat{Q}_{G_k^0, \tilde{B} \tilde{B}}^{-1} \hat{Q}'_{G_k^0, \tilde{\omega} \tilde{B}}$$

$$W_{it, \tilde{\omega} \setminus \tilde{B}} = \tilde{\omega}_{it} - \hat{Q}_{G_k^0, \tilde{\omega} \tilde{B}} \hat{Q}_{G_k^0, \tilde{B} \tilde{B}}^{-1} \sqrt{J} \tilde{B}_{it}^J$$

$$W_{i, \tilde{\omega} \setminus \tilde{B}} = \left(W_{i1, \tilde{\omega} \setminus \tilde{B}}, W_{i2, \tilde{\omega} \setminus \tilde{B}}, \dots, W_{iT, \tilde{\omega} \setminus \tilde{B}} \right)'$$

and $\hat{Q}_{G_k^0, \tilde{\omega} \tilde{\omega}} \equiv \frac{1}{N_k T} \sum_{t=1}^T \sum_{i \in G_k^0} \tilde{\omega}_{it} \tilde{\omega}'_{it}$. $\hat{Q}_{G_k^0, \tilde{B} \tilde{B}}$ and $\hat{Q}_{G_k^0, \tilde{\omega} \tilde{B}}$ are similarly defined.

Theorem 1.4. *Suppose Assumption 1.1, 1.2, 1.3, 1.4 and 1.5(ii) hold. Then for any $k \in \{1, \dots, K^0\}$,*

(i)

$$\sqrt{N_k T / J V_{k,B}^{-\frac{1}{2}}} \left(\hat{f}_k(x) - f_k^0(x) \right) \xrightarrow{D} N(0, 1)$$

(ii)

$$\sqrt{N_k T / J V_{k,B}^{-\frac{1}{2}}} \left(\hat{f}_{\hat{G}_k}(x) - f_k^0(x) \right) \xrightarrow{D} N(0, 1)$$

where

$$V_{k,B} = B^J(x)' \left(\hat{Q}_{G_k^0, \tilde{B} \setminus \tilde{\omega}} \right)^{-1} \frac{1}{N_k} \sum_{i \in G_k^0} \frac{1}{T} W'_{i, \tilde{B} \setminus \tilde{\omega}} \Sigma_i^{\frac{1}{2}} V_i \Sigma_i^{\frac{1}{2}} W_{i, \tilde{B} \setminus \tilde{\omega}} \left(\hat{Q}_{G_k^0, \tilde{B} \setminus \tilde{\omega}} \right)^{-1} B^J(x)$$

in which the different components are similarly defined as those in Theorem 1.3.

Theorems 1.3 and 1.4 indicate that the C-Lasso and post-Lasso estimators of both α_k and $f_k(x)$ are asymptotically equivalent to the infeasible estimators, which are denoted as $\hat{\alpha}_{G_k^0}$ and $\hat{f}_{G_k^0}$. Thus both C-Lasso and post-Lasso estimators exhibit oracle properties.

In my simulation results, the C-Lasso and post-Lasso estimators are of no much difference.

Remark. *For nonparametric panel data models, Theorem 1.3 no longer exists and the statement of Theorem 1.4 needs minor modifications.*

1.3.5 Determination of Number of Groups

In this section, I discuss how to use the Information Criterion(IC) to decide the number of groups K^0 . As is common in the literature, I need to assume that K^0 is bounded above from a finite integer K_{\max} . I make the dependence of $\hat{\theta}_i$ and $\hat{\eta}_k$ on K and λ explicit by denoting them as $\hat{\theta}_i(K, \lambda)$ and $\hat{\eta}_k(K, \lambda)$.

Using the post-Lasso estimator $\hat{\eta}_{\hat{G}_k}(K, \lambda)$, I could calculate

$$\hat{\sigma}_{\hat{G}(K, \lambda)}^2 = \frac{1}{NT} \sum_{k=1}^K \sum_{i \in \hat{G}_k(K, \lambda)} \sum_{t=1}^T \left(\tilde{y}_{it} - \tilde{z}'_{it} \hat{\eta}_{\hat{G}_k}(K, \lambda) \right)^2$$

Then I choose K to minimize the following information criterion

$$\text{IC}(K, \lambda) = \ln \left(\hat{\sigma}_{\hat{G}(K, \lambda)}^2 \right) + \rho_{NT}(p + J)K$$

where ρ_{NT} is another tuning parameter. Let $\hat{K}(\lambda) \equiv \arg \min_{1 \leq K \leq K_{\max}} \text{IC}(K, \lambda)$.

Let $G^{(K)} \equiv \{G_{K,1}, \dots, G_{K,K}\}$ be any K -partition of $\{1, \dots, N\}$ and \mathcal{G}_K a collection of all such partitions. Further define

$$\hat{\sigma}_{G^{(K)}}^2 \equiv \frac{1}{NT} \sum_{k=1}^K \sum_{i \in \hat{G}_{K,k}} \sum_{t=1}^T \left(\tilde{y}_{it} - \tilde{z}'_{it} \hat{\eta}_{\hat{G}_{K,k}} \right)^2$$

Some more assumptions are required.

Assumption 1.6. As $(N, T) \rightarrow \infty$, $\min_{1 \leq K < K^0} \inf_{G^{(K)} \in \mathcal{G}_K} \hat{\sigma}_{G^{(K)}}^2 \xrightarrow{P} \underline{\sigma}^2 > \sigma_0^2$, where $\sigma_0^2 = \text{plim}_{(N, T) \rightarrow \infty} \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T u_{it}^2$.

Assumption 1.7. As $(N, T) \rightarrow \infty$, $\rho_{NT}J \rightarrow 0$ and $\rho_{NT}NT \rightarrow \infty$.

When to decide the correct number of groups, there are three different situations to consider: $K < K^0$, $K = K^0$, and $K > K^0$, corresponding to under-fitted, correct, and over-fitted models respectively. Assumption 1.6 is used to guarantee that in the under-fitted models, the first item in the IC criterion is more significant than that in the correct model. As long as the second item is dominated, which is imposed in Assumption 1.7, I will not choose under-fitted models with probability approaching 1. Assumption 1.7 further implies that the over-fitted models will not be picked out with probability approaching one as well.

The following theorem formally summarizes this intuition.

Theorem 1.5. *Suppose Assumptions 1.1, 1.2, 1.3, 1.4, 1.5, 1.6 and 1.7 hold. Then $P(\hat{K}(\lambda) = K^0) \rightarrow 1$ as $(N, T) \rightarrow \infty$.*

Theorem 1.5 shows that the IC criterion is useful in deciding the correct number of groups asymptotically. However, in finite samples, I suggest that readers use it with caution. There is always a positive probability that misspecified models are selected. Thus I recommend readers try different numbers of groups, compare the results, and discuss possible implications.

1.4 Simulation

In this section, I evaluate the finite sample performance of the classification and estimation procedure.

1.4.1 Data Generating Process

Restate the model: $y_{it} = \mu_i + \omega'_{it}\beta_i + h_i(x_{it}) + u_{it}$. The data generating process(DGP) I consider has the following settings:

- (i) There are 3 different groups with equal group size $N/3$.
- (ii) The B-splines are of order 4(degree 3) and the number of interior points, J_0 , is set to be the closest integer to $(NT)^{\frac{1}{5}}$. Note that $J = J_0 + d$.
- (iii) The penalty parameter λ is chosen to be $(NT)^{-\frac{1}{8}}$. Note the settings are consistent with all the assumptions under the situation that N and T grow at the same speed.
- (iv) The individual fixed effects, μ_i , are independently drawn from a uniform $[0, 1]$ distribution. Since they are demeaned away anyway, this is a harmless setting.
- (v) The regressors, ω_{it} and x_{it} , are independently drawn from Uniform $[0, 1]$.
- (vi) The error terms, u_{it} , are independently distributed and $u_{it} \sim N(0, 1)$.

DGP 1: For different groups, the finite dimensional coefficients and the infinite-dimensional functions are set to be

$$\beta_i^0 = \begin{cases} 1 & \text{if } i \in G_1^0 \\ 2 & \text{if } i \in G_2^0 \\ 3 & \text{if } i \in G_3^0 \end{cases} \text{ and } h_i^0(x) = \begin{cases} \sin(2\pi x) & \text{if } i \in G_1^0 \\ \sin(4\pi x) & \text{if } i \in G_2^0 \\ \sin(6\pi x) & \text{if } i \in G_3^0 \end{cases}$$

I consider different combinations of N and T . For each combination, I simulate 200 times.

1.4.2 Main Result

For C-lasso estimators, since there are three different groups each involving parametric and nonparametric estimators, I report both the maximum RMSE of $\hat{\alpha}_k$ and \hat{f}_k , and RMSE of $\hat{\alpha}$ and \hat{f} , where $\hat{\alpha} \equiv (\hat{\alpha}_1, \hat{\alpha}_2, \dots, \hat{\alpha}_{K^0})$ and $\hat{f} \equiv (\hat{f}_1, \hat{f}_2, \dots, \hat{f}_{K^0})$. Denote the number of repetitions as M . The maximum RMSE of $\hat{\alpha}$ is defined as

$$\max\{\text{RMSE}\}_{\hat{\alpha}} \equiv \frac{1}{M} \sum_{m=1}^M \sqrt{\max_{1 \leq k \leq K^0} \|\hat{\alpha}_{k,m} - \alpha_k^0\|_F^2}$$

where $\hat{\alpha}_{k,m}$ denotes the estimated parametric parameters of k th group in m th repetition and α_k^0 is the corresponding true value. Similarly, the maximum RMSE of \hat{f} is

$$\max\{\text{RMSE}\}_{\hat{f}} \equiv \frac{1}{M} \sum_{m=1}^M \sqrt{\max_{1 \leq k \leq K^0} \|\hat{f}_{k,m} - f_k^0\|_2^2}$$

where $\hat{f}_{k,m}$ and f_k^0 are defined similarly. We further define RMSE of $\hat{\alpha}$ and \hat{f} as

$$\{\text{RMSE}\}_{\hat{\alpha}} \equiv \frac{1}{M} \sum_{m=1}^M \sqrt{\sum_{k=1}^{K^0} \|\hat{\alpha}_{k,m} - \alpha_k^0\|_F^2}$$

$$\{\text{RMSE}\}_{\hat{f}} \equiv \frac{1}{M} \sum_{m=1}^M \sqrt{\sum_{k=1}^{K^0} \|\hat{f}_{k,m} - f_k^0\|_2^2}$$

For post-Lasso estimators $\hat{\alpha}_{\hat{G}}$ and $\hat{f}_{\hat{G}}$, and oracle estimators $\hat{\alpha}_{G^0}$ and \hat{f}_{G^0} , I similarly define maximum RMSE and RMSE, where $\hat{\alpha}_{\hat{G}} \equiv (\hat{\alpha}_{\hat{G}_1}, \hat{\alpha}_{\hat{G}_2}, \dots, \hat{\alpha}_{\hat{G}_{K^0}})$, $\hat{f}_{\hat{G}} \equiv (\hat{f}_{\hat{G}_1}, \hat{f}_{\hat{G}_2}, \dots, \hat{f}_{\hat{G}_{K^0}})$ and $\hat{\alpha}_{G^0} \equiv (\hat{\alpha}_{G_1^0}, \hat{\alpha}_{G_2^0}, \dots, \hat{\alpha}_{G_{K^0}^0})$, $\hat{f}_{G^0} \equiv (\hat{f}_{G_1^0}, \hat{f}_{G_2^0}, \dots, \hat{f}_{G_{K^0}^0})$.

The main results are reported in Table 1.2 and 1.3. I discuss Table 1.2 first. When T is relatively small ($T = 60$), the classification error is comparatively large. Around 25% ($N = 90$) or 20% ($N = 180$) of individuals are classified into wrong groups. Consequently, the maximum RMSE of $\hat{\alpha}$, \hat{f} and $\hat{\alpha}_{\hat{G}}$, $\hat{f}_{\hat{G}}$ are considerable compared with that of the oracle estimators. However, as T increases, the classification error shrinks quickly. For the case $N = 90, T = 90$, $N = 180, T = 90$ and $N = 270, T = 90$, more than 90% of individuals are assigned the correct group identity. As a result, the maximum RMSE of C-lasso and post-lasso estimators decrease. When I further consider $N = 180, T = 180$ and $N = 270, T = 180$, the classification errors are only 1.2% and 0.2% respectively, and the RMSE of C-lasso and post-lasso estimators are almost the same as that of the oracle estimators. If I increase T to 270 and consider $N = 270$, I achieve almost 100% correct classification. Consequently, the RMSE of C-lasso, post-Lasso and oracle estimators are of no difference. In Table 1.3, I get similar results.

By carefully comparing the results in Table 1.2 and 1.3, I further find that most of RMSE of C-lasso and post-lasso estimators could be attributed to the maximum RMSE of them.

Table 1.2: RMSE (Maximum) of C-Lasso and post-Lasso Estimators in DGP 1

N	T	% of correct Classification	C-Lasso			Post-Lasso			Oracle		
			Maximum RMSE of $\hat{\alpha}$	Maximum RMSE of f	Maximum RMSE of $\hat{\alpha}_{\hat{G}}$	Maximum RMSE of $\hat{\alpha}_{\hat{G}}$	Maximum RMSE of $f_{\hat{G}}$	Maximum RMSE of $\hat{\alpha}_{G^0}$	Maximum RMSE of f_{G^0}		
DGP	90	60	0.557	0.384	0.556	0.384	0.108	0.112	0.090	0.104	
	90	90.7	0.258	0.235	0.260	0.236	0.090	0.100	0.080	0.100	
	60	80.8	0.462	0.344	0.462	0.341	0.080	0.100	0.085	0.085	
	180	90	0.182	0.153	0.182	0.153	0.065	0.051	0.051	0.051	
	180	98.8	0.062	0.062	0.062	0.062	0.045	0.052	0.052	0.054	
	270	90	0.191	0.148	0.190	0.147	0.037	0.036	0.036	0.036	
	270	180	0.038	0.037	0.038	0.037	0.036	0.036	0.036	0.036	
	270	99.99	0.030	0.032	0.030	0.032	0.030	0.032	0.030	0.032	

Table 1.3: RMSE of C-Lasso and post-Lasso Estimators in DGP 1

DGP	N	T	% of correct Classification	C-Lasso			Post-Lasso			Oracle		
				RMSE of $\hat{\alpha}$	RMSE of f	RMSE of $\hat{\alpha}_{\hat{C}}$	RMSE of $f_{\hat{C}}$	RMSE of $\hat{\alpha}_{\hat{C}^0}$	RMSE of $f_{\hat{C}^0}$	RMSE of $\hat{\alpha}_{\hat{C}^0}$	RMSE of $f_{\hat{C}^0}$	
	90	60	76.9	0.598	0.417	0.597	0.416	0.131	0.152			
	90	90	90.7	0.290	0.267	0.292	0.267	0.110	0.134			
	180	60	80.8	0.491	0.366	0.491	0.365	0.097	0.125			
	180	90	94.1	0.199	0.170	0.198	0.170	0.078	0.106			
	180	180	98.8	0.079	0.082	0.080	0.082	0.055	0.068			
	270	90	94.0	0.205	0.164	0.205	0.162	0.062	0.075			
	270	180	99.8	0.045	0.053	0.045	0.053	0.043	0.052			
	270	270	99.99	0.036	0.045	0.036	0.045	0.036	0.045			

1.4.3 Comparison with Complete Homogeneity and Heterogeneity

To further illustrate the advantages of C-lasso and post-Lasso estimators over complete parameter homogeneity or complete parameter heterogeneity, I compare the results of the three different approaches.

To make the approaches comparable, I define RMSE of C-lasso estimators in a different way.

$$\begin{aligned} \{\text{RMSE}\}_{\hat{\beta}}^{\text{ind}} &\equiv \frac{1}{M} \sum_{m=1}^M \sqrt{\frac{1}{N} \sum_{i=1}^N \|\hat{\beta}_{i,m} - \beta_i^0\|_F^2} \\ \{\text{RMSE}\}_{\hat{h}}^{\text{ind}} &\equiv \frac{1}{M} \sum_{m=1}^M \sqrt{\frac{1}{N} \sum_{i=1}^N \|\hat{h}_{i,m} - h_i^0\|_2^2} \end{aligned}$$

where $\hat{\beta}_{i,m}$ and $\hat{h}_{i,m}$ denotes the estimated parametric and nonparametric parameters of individual i in m th repetition using C-lasso.

For post-lasso and oracle estimators, I similarly define $\{\text{RMSE}\}_{\hat{\beta}_{\hat{G}}}^{\text{ind}}$, $\{\text{RMSE}\}_{\hat{h}_{\hat{G}}}^{\text{ind}}$ and $\{\text{RMSE}\}_{\hat{\beta}_{G^0}}^{\text{ind}}$, $\{\text{RMSE}\}_{\hat{h}_{G^0}}^{\text{ind}}$, respectively.

If we assume individual share the same parameters, I denote the corresponding defined RMSE of parametric and nonparametric estimators as $\{\text{RMSE}\}_{\hat{\beta}_{ho}}^{\text{ind}}$ and $\{\text{RMSE}\}_{\hat{f}_{ho}}^{\text{ind}}$. If we allow for complete parameter heterogeneity, I use $\{\text{RMSE}\}_{\hat{\beta}_{he}}^{\text{ind}}$ and $\{\text{RMSE}\}_{\hat{f}_{he}}^{\text{ind}}$.

The results are reported in Table 1.4. Under complete parameter homogeneity, the model is misspecified. $\{\text{RMSE}\}_{\hat{\beta}_{ho}}^{\text{ind}}$ and $\{\text{RMSE}\}_{\hat{f}_{ho}}^{\text{ind}}$ don't change much as N and T vary. While under complete parameter homogeneity, we fail to account for the group structure. When T is comparatively small ($T = 60$), C-lasso and post-lasso estimators don't necessarily outperform those under complete parameter homogeneity and heterogeneity. However, as long as T is large enough ($T = 90, 180, 270$), C-lasso and post-lasso estimators perform much better.

Table 1.4: Comparison with Complete Homogeneity and Heterogeneity in DGP 1

DGP	N	T	% of correct Classification	C-Lasso		Post-Lasso		Oracle		Homogeneity		Heterogeneity	
				RMSE of $\hat{\beta}$	RMSE of \hat{h}	RMSE of $\hat{\beta}_{\hat{G}}$	RMSE of $\hat{h}_{\hat{G}}$	RMSE of $\hat{\beta}_{G^0}$	RMSE of \hat{h}_{G^0}	RMSE of $\hat{\beta}_{ho}$	RMSE of \hat{h}_{ho}	RMSE of $\hat{\beta}_{he}$	RMSE of \hat{h}_{he}
	90	60	76.9	0.381	9.419	0.551	0.396	0.075	0.088	0.818	0.580	0.500	14.099
	90	90	90.7	0.286	0.288	0.330	0.257	0.063	0.077	0.818	0.579	0.388	0.403
	180	60	80.8	0.374	13.155	0.503	0.367	0.056	0.072	0.818	0.579	0.497	23.875
	180	90	94.1	0.284	0.366	0.300	0.223	0.045	0.061	0.817	0.579	0.394	0.531
	180	180	98.8	0.184	0.183	0.065	0.062	0.032	0.039	0.817	0.578	0.268	0.268
	270	90	94.0	0.286	4.159	0.323	0.221	0.036	0.044	0.817	0.578	0.394	6.230
	270	180	99.8	0.181	0.190	0.054	0.047	0.025	0.030	0.817	0.578	0.267	0.282
	270	270	99.99	0.146	0.149	0.024	0.027	0.021	0.026	0.817	0.578	0.217	0.221

1.4.4 Comparison with Misspecified Parametric model

In terms of classification, there is a concern that it might not be necessary to use semi-nonparametric models, because we might still achieve good classification even the model is misspecified as fully parametric.

To address this concern, I compare the classification errors of two different models: the true model and misspecified parametric model.

We first use DGP 1 as before. The results are shown in Table 1.5. The classification errors of the true model are always smaller than those of the misspecified model.

Table 1.5: Comparison with Misspecified Parametric Model in DGP 1

	N	T	% of Correct Classification	
			True Model	Misspecified Model
DGP	90	60	76.9	48.9
	90	90	90.7	56.5
	180	60	80.8	51.1
	180	90	94.1	62.0
	180	180	98.8	83.2
	270	90	94.0	65.9
	270	180	99.8	87.3
	270	270	99.99	94.2

As T increases, the classification error of the misspecified model decreases, so one might conclude that it is still plausible to do classification using the misspecified model. However, under certain circumstances, the classification error of the misspecified model is large and does not improve even as T increases. To illustrate this idea, I consider a new model and DGP 2:

$$y_{it} = \mu_i + h_i(x_{it}) + u_{it}$$

where

$$h_i^0(x) = \begin{cases} \cos(2\pi x) & \text{if } i \in G_1^0 \\ \cos(4\pi x) & \text{if } i \in G_2^0 \\ \cos(6\pi x) & \text{if } i \in G_3^0 \end{cases}$$

The other setting are the same as DGP 1.

Then the misspecified parametric model is

$$y_{it} = \mu_i + x_{it}\phi_i + u_{it}$$

After simple calculation, we could see that for individuals from different groups, the parameters are the same under the misspecified model, thus it is theoretically impossible to classify individuals into correct groups. The simulation results are shown in Table 1.6. With the misspecified model, the percentage of correct classification is at most 40.6% and doesn't increase as T increases. Considering that with three equally-sized groups, there is at least 33.3% correct classification under suitable permutation, the error almost achieves its upper bound. On the contrary, with nonparametric model, I could still achieve good classification and the classification error shrinks as T increases.

Table 1.6: Comparison with Misspecified Parametric Model in DGP 2

	N	T	% of Correct Classification	
			True Model	Misspecified Model
DGP	90	60	81.4	40.6
	90	90	94.5	40.2
	180	60	84.4	38.2
	180	90	95.6	37.9
	180	180	99.8	38.3
	270	90	95.1	37.2
	270	180	99.8	37.1
	270	270	99.99	37.2

1.5 Conclusion

I propose a semi-nonparametric panel data model with latent group structures. I first approximate the infinite-dimensional parameters with a sieve expansion. Then using the C-Lasso method, I simultaneously classify individuals into different groups and estimate the group-specific parameters. The C-Lasso and post-Lasso estimators achieve uniform classification consistency and exhibit the oracle property. Simulations demonstrate an excellent finite sample performance of this approach.

It is possible to extend this research in several different directions. First, what if I consider high-dimensional panel data models where the response mechanisms exhibit heterogeneity? Although it seems plausible, it is not trivial at all to apply my method to high-dimensional data. Certain highly-mathematical techniques are required. Thus I leave it for future research. Second, it is natural to generalize my approach to unbalanced panels with some minor changes. Third, cross-sectional dependence could also be introduced into my framework, although much more technical details need to be taken care of.

Appendix

1.A Proofs of the Main Results

I use $\|\cdot\|$ to denote Frobenius norm in the Appendix for simplicity.

Proof of Theorem 1.1

Proof. (i) For each individual, I define

$$Q_i(\theta_i) \equiv \frac{1}{T} \sum_{t=1}^T (\tilde{y}_{it} - \tilde{z}'_{it} \theta_i)^2$$

and

$$Q_i(\theta_i, \eta) \equiv Q_i(\theta_i) + \lambda \prod_{k=1}^{K^0} \|\theta_i - \eta_k\|$$

Since $\hat{\theta}_i$ minimizes $Q_i(\theta_i, \hat{\eta})$, I have $Q_i(\hat{\theta}_i, \hat{\eta}) \leq Q_i(\theta_i^0, \hat{\eta})$, which is equivalent to

$$\left(Q_i(\hat{\theta}_i) - Q_i(\theta_i^0) \right) + \lambda \left(\prod_{k=1}^{K^0} \|\hat{\theta}_i - \hat{\eta}_k\| - \prod_{k=1}^{K^0} \|\theta_i^0 - \hat{\eta}_k\| \right) \leq 0$$

- Consider the first part:

$$\begin{aligned} & Q_i(\hat{\theta}_i) - Q_i(\theta_i^0) \\ &= \frac{1}{T} \sum_{t=1}^T (\tilde{y}_{it} - \tilde{z}'_{it} \hat{\theta}_i)^2 - \frac{1}{T} \sum_{t=1}^T (\tilde{y}_{it} - \tilde{z}'_{it} \theta_i^0)^2 \\ &= (\hat{\theta}_i - \theta_i^0)' \hat{Q}_{i, \tilde{z}\tilde{z}} (\hat{\theta}_i - \theta_i^0) - 2(\hat{\theta}_i - \theta_i^0)' \hat{Q}_{i, \tilde{z}\tilde{e}} \end{aligned}$$

where $\hat{Q}_{i, \tilde{z}\tilde{z}} = \frac{1}{T} \sum_{t=1}^T \tilde{z}_{it} \tilde{z}'_{it}$, $\hat{Q}_{i, \tilde{z}\tilde{e}} = \frac{1}{T} \sum_{t=1}^T \tilde{z}_{it} \tilde{e}_{it}$, $\tilde{e}_{it} = \tilde{\delta}_{h_i}(x_{it}) + \tilde{u}_{it}$.

- Consider the second part, I have

$$\begin{aligned}
& \left| \prod_{k=1}^{K^0} \|\hat{\theta}_i - \hat{\eta}_k\| - \prod_{k=1}^{K^0} \|\theta_i^0 - \hat{\eta}_k\| \right| \\
& \leq \left| \prod_{k=1}^{K^0-1} \|\hat{\theta}_i - \hat{\eta}_k\| \left(\|\hat{\theta}_i - \hat{\eta}_{K^0}\| - \|\theta_i^0 - \hat{\eta}_{K^0}\| \right) \right| \\
& \quad + \left| \prod_{k=1}^{K^0-2} \|\hat{\theta}_i - \hat{\eta}_k\| \|\theta_i^0 - \hat{\eta}_{K^0}\| \left(\|\hat{\theta}_i - \hat{\eta}_{K^0-1}\| - \|\theta_i^0 - \hat{\eta}_{K^0-1}\| \right) \right| \\
& \quad + \dots \\
& \quad + \left| \prod_{k=2}^{K^0} \|\theta_i^0 - \hat{\eta}_k\| \left(\|\hat{\theta}_i - \hat{\eta}_1\| - \|\theta_i^0 - \hat{\eta}_1\| \right) \right| \\
& \leq c_{1i,NT}(\hat{\theta}, \theta^0, \hat{\eta}) \|\hat{\theta}_i - \theta_i^0\|
\end{aligned}$$

$$\text{where } c_{1i,NT}(\hat{\theta}, \theta^0, \hat{\eta}) \equiv \prod_{k=1}^{K^0-1} \|\hat{\theta}_i - \hat{\eta}_k\| + \prod_{k=1}^{K^0-2} \|\hat{\theta}_i - \hat{\eta}_k\| \|\theta_i^0 - \hat{\eta}_{K^0}\| + \dots + \prod_{k=2}^{K^0} \|\theta_i^0 - \hat{\eta}_k\|.$$

Together I have

$$\begin{aligned}
& (\hat{\theta}_i - \theta_i^0)' \hat{Q}_{i,\bar{z}\bar{z}} (\hat{\theta}_i - \theta_i^0) \\
& \leq \left| 2(\hat{\theta}_i - \theta_i^0)' \hat{Q}_{i,\bar{z}\bar{e}} \right| + \lambda c_{1i,NT}(\hat{\theta}, \theta^0, \hat{\eta}) \|\hat{\theta}_i - \theta_i^0\| \\
& \leq 2 \|\hat{\theta}_i - \theta_i^0\| \|\hat{Q}_{i,\bar{z}\bar{e}}\| + \lambda c_{1i,NT}(\hat{\theta}, \theta^0, \hat{\eta}) \|\hat{\theta}_i - \theta_i^0\|
\end{aligned}$$

By Lemma 1.3, $\mu_{\min}(\hat{Q}_{i,\bar{z}\bar{z}}) > \underline{c} > 0$ w.p.a. 1, then I have w.p.a. 1,

$$\|\hat{\theta}_i - \theta_i^0\| \leq \underline{c}^{-1} \left(2 \|\hat{Q}_{i,\bar{z}\bar{e}}\| + \lambda c_{1i,NT}(\hat{\theta}, \theta^0, \hat{\eta}) \right)$$

By Lemma 1.3, $\|\hat{Q}_{i,\bar{z}\bar{e}}\| = O_p(J^{-\frac{r_1}{d}} + J^{\frac{1}{2}} T^{-\frac{1}{2}})$, thus

$$\|\hat{\theta}_i - \theta_i^0\| = O_p(J^{-\frac{r_1}{d}} + J^{\frac{1}{2}} T^{-\frac{1}{2}} + \lambda)$$

Remark. The argument depends on the condition that $c_{1i,NT}(\hat{\theta}, \theta^0, \hat{\eta}) = O_p(1)$.

We show this by considering a constrained optimization problem.

Define

$$\begin{aligned}\mathcal{B}_b &\equiv \left\{ \beta : \|\beta_i\|_F \leq c < \infty, i = 1, \dots, N \right\} \\ \mathcal{R}_b &\equiv \left\{ \gamma : |\gamma_{ij}| \leq c < \infty, i = 1, \dots, N, j = 1, \dots, J \right\} \\ \mathcal{A}_b &\equiv \left\{ \alpha : \|\alpha_k\|_F \leq c < \infty, k = 1, \dots, K^0 \right\} \\ \Pi_b &\equiv \left\{ \pi : |\pi_{kj}| \leq c < \infty, k = 1, \dots, K^0, j = 1, \dots, J \right\}\end{aligned}$$

where c is a generic constant which varies.

Further define $\Theta_b \equiv \{\theta : \beta \in \mathcal{B}_b, \gamma \in \mathcal{R}_b\}$, $\mathcal{H}_b \equiv \{\eta : \alpha \in \mathcal{A}_b, \pi \in \Pi_b\}$. Remember that $\theta = (\theta_1, \dots, \theta_N)$, where $\theta_i \equiv \left(\beta'_i, \frac{1}{\sqrt{J}}\gamma'_i\right)'$, $i = 1, \dots, N$, and $\eta = (\eta_1, \dots, \eta_{K^0})$, where $\eta_k \equiv \left(\alpha'_k, \frac{1}{\sqrt{J}}\pi'_k\right)'$, $k = 1, \dots, K^0$.

If c is large enough, 1) by Assumption 1.1(iii), I could imply that β^0 and α^0 lie in the interior of \mathcal{B}_b and \mathcal{A}_b respectively; 2) Similarly, by Assumption 1.1(iv), I could get that γ^0 and π^0 lie in the interior of \mathcal{R}_b and Π_b respectively, thus $\theta^0 \in \Theta_b$ and $\eta^0 \in \mathcal{H}_b$.

Then I search over Θ_b and \mathcal{H}_b to minimize the objective function 1.11, namely

$$(\hat{\theta}, \hat{\eta}) = \arg \min_{\theta \in \Theta_b, \eta \in \mathcal{H}_b} \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T (\tilde{y}_{it} - \tilde{z}'_{it}\theta_i)^2 + \frac{\lambda}{N} \sum_{i=1}^N \prod_{k=1}^{K^0} \|\theta_i - \eta_k\|_F$$

The restrictions guarantee that $c_{1i,NT}(\hat{\theta}, \hat{\eta}) = O(1)$.

Practically, I set c large enough and conduct the constrained optimization, which works well in my simulations.

- (ii) Let $m_{JT} = J^{-\frac{r_1}{d}} + J^{\frac{1}{2}}T^{-\frac{1}{2}}$ and v denotes a $(p+J) \times N$ matrix. In order to show that $\frac{1}{N} \sum_{i=1}^N \|\hat{\theta}_i - \theta_i^0\|^2 = O_p(J^{-2\frac{r_1}{d}} + JT^{-1})$, I just need to prove that for any ε , there exists

a constant $M = M(\varepsilon)$ such that, for sufficiently large N and T ,

$$P \left\{ \inf_{\frac{1}{N} \sum_{i=1}^N \|v_i\|^2 = M} Q_{NT}(\theta^0 + m_{JT}v, \hat{\eta}) > Q_{NT}(\theta^0, \eta^0) \right\} \geq 1 - \varepsilon$$

This implies that w.p.a.1 there exists a local minimum $\{\hat{\theta}, \hat{\eta}\}$ such that $\frac{1}{N} \sum_{i=1}^N \|\hat{\theta}_i - \theta_i^0\|^2 = O_p(J^{-2\frac{r_1}{d}} + JT^{-1})$ holds.

$$\begin{aligned} & m_{JT}^{-2} \left(Q_{NT}(\theta^0 + m_{JT}v, \hat{\eta}) - Q_{NT}(\theta^0, \eta^0) \right) \\ &= \frac{1}{N} \sum_{i=1}^N v_i' \hat{Q}_{i, \tilde{z}\tilde{z}} v_i - \frac{2}{N} m_{JT}^{-1} \sum_{i=1}^N v_i' \hat{Q}_{i, \tilde{z}\tilde{e}} + \frac{\lambda}{N} \sum_{i=1}^N \prod_{k=1}^{K^0} \|\theta_i^0 + m_{JT}v_i - \hat{\eta}_k\| \\ &\geq c \frac{1}{N} \sum_{i=1}^N \|v_i\|^2 - 2 \left\{ \frac{1}{N} \sum_{i=1}^N \|v_i\|^2 \right\}^{\frac{1}{2}} \left\{ \frac{m_{JT}^{-2}}{N} \sum_{i=1}^N \|\hat{Q}_{i, \tilde{z}\tilde{e}}\|^2 \right\}^{\frac{1}{2}} \end{aligned}$$

where the last inequality holds w.p.a 1 by Lemma 1.3.

By Lemma 1.3, $\frac{1}{N} \sum_{i=1}^N \|\hat{Q}_{i, \tilde{z}\tilde{e}}\|^2 = O_p(J^{-2\frac{r_1}{d}} + JT^{-1})$, then $\frac{m_{JT}^{-2}}{N} \sum_{i=1}^N \|\hat{Q}_{i, \tilde{z}\tilde{e}}\|^2 = O_p(1)$, thus for sufficiently large M , I have $m_{JT}^{-2} \left(Q_{NT}(\theta^0 + m_{JT}v, \hat{\eta}) - Q_{NT}(\theta_0, \eta_0) \right) > 0$ w.p.a.1.

(iii) Further consider $c_{1i, NT}(\hat{\theta}, \theta^0, \eta)$, where $\hat{\theta}$ and η lie in the interior of Θ_b and \mathcal{H}_b respec-

tively.

$$\begin{aligned}
& c_{1i,NT}(\hat{\theta}, \theta^0, \eta) \\
&= \prod_{k=1}^{K^0-1} \|\hat{\theta}_i - \eta_k\| + \prod_{k=1}^{K^0-2} \|\hat{\theta}_i - \eta_k\| \|\theta_i^0 - \eta_{K^0}\| + \cdots + \prod_{k=2}^{K^0} \|\theta_i^0 - \eta_k\| \\
&\leq \prod_{k=1}^{K^0-1} \left(\|\hat{\theta}_i - \theta_i^0\| + \|\theta_i^0 - \eta_k\| \right) + \prod_{k=1}^{K^0-2} \left(\|\hat{\theta}_i - \theta_i^0\| + \|\theta_i^0 - \eta_k\| \right) \|\theta_i^0 - \eta_{K^0}\| \\
&\quad + \cdots + \prod_{k=2}^{K^0} \|\theta_i^0 - \eta_k\| \\
&\leq \sum_{s=0}^{K^0-1} c_{1si,NT}(\theta^0, \eta) \|\hat{\theta}_i - \theta_i^0\|^s + \sum_{s=0}^{K^0-2} c_{2si,NT}(\theta^0, \eta) \|\hat{\theta}_i - \theta_i^0\|^s \\
&\quad + \cdots + \sum_{s=0}^0 c_{K^0si,NT}(\theta^0, \eta) \|\hat{\theta}_i - \theta_i^0\|^s \\
&\leq \sum_{s=0}^{K^0-1} c_{si,NT}(\theta^0, \eta) \|\hat{\theta}_i - \theta_i^0\|^s \\
&\leq c_{2i,NT}(\theta^0, \eta) \sum_{s=0}^{K^0-1} \|\hat{\theta}_i - \theta_i^0\|^s \\
&\leq c_{2i,NT}(\theta^0, \eta) \left(1 + 2 \|\hat{\theta}_i - \theta_i^0\| \right)
\end{aligned}$$

where $c_{2i,NT}(\theta^0, \eta) = \max_{1 \leq s \leq K^0} c_{si,NT}(\theta^0, \eta)$ and $c_{si,NT}(\theta^0, \eta) = \sum_{k=1}^{K^0} c_{k si,NT}(\theta^0, \eta)$.

The last inequality holds w.p.a 1.

Define $p_{NT}(\theta, \eta) \equiv \frac{1}{N} \sum_{i=1}^N \prod_{k=1}^{K^0} \|\theta_i - \eta_k\|$, then

$$\begin{aligned}
& \left| p_{NT}(\hat{\theta}, \eta) - p_{NT}(\theta^0, \eta) \right| \\
&\leq \frac{1}{N} \sum_{i=1}^N c_{1i,NT}(\hat{\theta}, \theta^0, \eta) \|\hat{\theta}_i - \theta_i^0\| \\
&\leq \frac{c_{2i,NT}(\theta^0, \eta)}{N} \sum_{i=1}^N \left(\|\hat{\theta}_i - \theta_i^0\| + 2 \|\hat{\theta}_i - \theta_i^0\|^2 \right) \\
&\leq c_{2i,NT}(\theta^0, \eta) \left(\frac{1}{N} \sum_{i=1}^N \|\hat{\theta}_i - \theta_i^0\|^2 \right)^{\frac{1}{2}} + c_{2i,NT}(\theta^0, \eta) \frac{1}{N} \sum_{i=1}^N \|\hat{\theta}_i - \theta_i^0\|^2 \\
&= O_p(J^{-\frac{r-1}{d}} + J^{\frac{1}{2}} T^{-\frac{1}{2}})
\end{aligned}$$

where I use $c_{2i,NT}(\theta^0, \eta) = O(1)$, which is implied by a similar argument as that in the proof of Theorem 1.1(i).

Since $p_{NT}(\hat{\theta}, \hat{\eta}) \leq p_{NT}(\hat{\theta}, \eta^0)$, note that $p_{NT}(\theta^0, \eta^0) = 0$,

$$\begin{aligned}
0 &\geq p_{NT}(\hat{\theta}, \hat{\eta}) - p_{NT}(\hat{\theta}, \eta^0) \\
&= \left(p_{NT}(\hat{\theta}, \hat{\eta}) - p_{NT}(\theta^0, \hat{\eta}) \right) + \left(p_{NT}(\theta^0, \hat{\eta}) - p_{NT}(\theta^0, \eta^0) \right) - \left(p_{NT}(\hat{\theta}, \eta^0) - p_{NT}(\theta^0, \eta^0) \right) \\
&= O_p(J^{-\frac{\tau_1}{d}} + J^{\frac{1}{2}}T^{-\frac{1}{2}}) + p_{NT}(\theta^0, \hat{\eta}) \\
&= O_p(J^{-\frac{\tau_1}{d}} + J^{\frac{1}{2}}T^{-\frac{1}{2}}) + \sum_{j=1}^{K^0} \frac{N_j}{N} \prod_{k=1}^{K^0} \|\eta_j^0 - \hat{\eta}_k\|
\end{aligned}$$

Then there exists a permutation of $\{1, \dots, K^0\}$ such that $\|\hat{\eta}_k - \eta_k^0\| = O_p(J^{-\frac{\tau_1}{d}} + J^{\frac{1}{2}}T^{-\frac{1}{2}})$. □

Proof of Theorem 1.2

Proof. (i) For any $i \in G_k^0$ and $l \neq k$, by Theorem 1.1, $\|\hat{\theta}_i - \hat{\eta}_l\| \xrightarrow{p} \|\eta_k^0 - \eta_l^0\| \neq 0$. Suppose that $\|\hat{\theta}_i - \hat{\eta}_k\| \neq 0$ for some $i \in G_k^0$, which means that $i \notin \hat{G}_k$, then the first order condition with respect to θ_i is

$$\begin{aligned}
0_{p+J} &= -2\hat{Q}_{i,\bar{z}\bar{u}} + \left(2\hat{Q}_{i,\bar{z}\bar{z}} + \frac{\lambda}{\|\hat{\theta}_i - \hat{\eta}_k\|} \prod_{l=1, l \neq k}^{K^0} \|\hat{\theta}_i - \hat{\eta}_l\| \right) (\hat{\theta}_i - \hat{\eta}_k) \\
&\quad - 2\hat{Q}_{i,\bar{z}\bar{\delta}} + 2\hat{Q}_{i,\bar{z}\bar{z}} (\hat{\eta}_k - \theta_i^0) + \lambda \sum_{j=1, j \neq k}^{K^0} \hat{e}_{ij} \prod_{l=1, l \neq j}^{K^0} \|\hat{\theta}_i - \hat{\eta}_l\| \\
&\equiv \hat{A}_{i1} + \hat{A}_{i2} + \hat{A}_{i3} + \hat{A}_{i4} + \hat{A}_{i5}
\end{aligned}$$

where $\hat{e}_{ij} = \frac{\hat{\theta}_i - \hat{\eta}_j}{\|\hat{\theta}_i - \hat{\eta}_j\|}$ if $\|\hat{\theta}_i - \hat{\eta}_j\| \neq 0$ and $\|\hat{e}_{ij}\|_F \leq 1$ otherwise.

From the proof of Theorem 1.1, I have that

$$\|\hat{\theta}_i - \theta_i^0\| \leq \underline{c}^{-1} \left(2\|\hat{Q}_{i,\bar{z}\bar{e}}\| + \lambda c_{1i,NT}(\hat{\theta}, \theta^0, \hat{\eta}) \right)$$

Let $\mu_{1,JT} = \left(J^{-\frac{r_1}{d}} + J^{\frac{1}{2}} T^{-\frac{1}{2}} (\ln T)^3 + \lambda \right) (\ln T)^v$ and $\mu_{2,JT} = \left(J^{-\frac{r_1}{d}} + J^{\frac{1}{2}} T^{-\frac{1}{2}} (\ln T)^3 \right) (\ln T)^v$ for some $v > 0$. By Lemma 1.3, I could show that

$$\begin{aligned} P \left(\max_{1 \leq i \leq N} \|\hat{\theta}_i - \theta_i^0\| \geq c\mu_{1,JT} \right) &= o(N^{-1}) \\ P \left(\|\hat{\eta}_k - \eta_k^0\| \geq c\mu_{2,JT} \right) &= o(N^{-1}) \end{aligned}$$

for any $c > 0$.

Let $\hat{c}_{ik} = \prod_{l=1, l \neq k}^{K^0} \|\hat{\theta}_i - \hat{\eta}_l\|$, then

$$\begin{aligned} \hat{c}_{ik} &= \prod_{l=1, l \neq k}^{K^0} \|\hat{\theta}_i - \hat{\eta}_l\| \\ &= \prod_{l=1, l \neq k}^{K^0} \left\| (\hat{\theta}_i - \eta_k^0) - (\hat{\eta}_l - \eta_l^0) + (\eta_k^0 - \eta_l^0) \right\| \\ &= \prod_{l=1, l \neq k}^{K^0} \left\| \eta_k^0 - \eta_l^0 + o_p(1) \right\| \\ &= O_p(1) \end{aligned}$$

Similarly let $c_{ik}^0 = \prod_{l=1, l \neq k}^{K^0} \|\theta_i^0 - \eta_l^0\|$. Define $\bar{c}_k^0 = \max_{i \in G_k^0} c_{ik}^0$ and $\underline{c}_k^0 = \min_{i \in G_k^0} c_{ik}^0$.

$$P \left(\frac{\underline{c}_k^0}{2} \leq \hat{c}_{ik} \leq 2\bar{c}_k^0 \right) = 1 - o(N^{-1})$$

And $P \left(\max_{i \in G_k^0} \|\hat{A}_{i5}\| \geq C\lambda\mu_{1,JT} \right) = o(N^{-1})$ for large enough $C > 0$.

Define

$$\begin{aligned} \Xi_{kNT} &\equiv \left\{ \frac{\underline{c}_k^0}{2} \leq \hat{c}_{ik} \leq 2\bar{c}_k^0 \right\} \cap \left\{ \|\hat{\eta}_k - \eta_k^0\| \leq c \left(J^{-\frac{r_1}{d}} + J^{\frac{1}{2}} T^{-\frac{1}{2}} (\ln T)^3 \right) (\ln T)^v \right\} \\ &\quad \cap \left\{ 0 < \underline{c} < \min_{0 \leq i \leq N} \mu_{\min}(\hat{Q}_{i,\hat{z}\hat{z}}) \leq \max_{0 \leq i \leq N} \mu_{\max}(\hat{Q}_{i,\hat{z}\hat{z}}) < \bar{c} < \infty \right\} \\ &\quad \cap \left\{ \max_{1 \leq i \leq N} \|\hat{Q}_{i,\hat{z}\hat{\delta}}\| \leq C\theta_{NT} \right\} \cap \left\{ \max_{1 \leq i \leq N} \|\hat{\theta}_i - \theta_i^0\| \leq c\mu_{1,JT} \right\} \end{aligned}$$

for some $C > 0$ and $c > 0$. $\theta_{NT} \equiv \max_{1 \leq k \leq K^0} \sup_{x \in [0,1]^d} \|f_k^0(x) - B^{J'} \pi_k^0\| = O(J^{-\frac{r_1}{d}})$.

Then $P(\Xi_{kNT}) = 1 - o(N^{-1})$.

Let $\phi_{ik} = \frac{\hat{\theta}_i - \hat{\eta}_k}{\|\hat{\theta}_i - \hat{\eta}_k\|}$. Conditional on Ξ_{kNT} , I have that uniformly in $i \in G_k^0$, with probability $1 - o(N^{-1})$,

$$\begin{aligned} |\phi'_{ik} \hat{A}_{i2}| &\geq 2\bar{c} \|\hat{\theta}_i - \hat{\eta}_k\| + \lambda \hat{c}_{ik} \geq \lambda \frac{c_k^0}{2} \\ |\phi'_{ik} \hat{A}_{i3}| &\leq 2 \|\hat{Q}_{i, \bar{z}\bar{\delta}}\| \leq 2C\theta_{NT} \\ |\phi'_{ik} \hat{A}_{i4}| &\leq 2\bar{c} \|\hat{\eta}_k - \eta_k^0\| \leq 2\bar{c}c \left(J^{-\frac{r_1}{d}} + J^{\frac{1}{2}} T^{-\frac{1}{2}} (\ln T)^3 \right) (\ln T)^v \\ |\phi'_{ik} \hat{A}_{i5}| &\leq \max_{i \in G_k^0} \|\hat{A}_{i5}\| \leq C\lambda\mu_{1,JT} \end{aligned}$$

Then

$$\begin{aligned} & \left| \phi'_{ik} (\hat{A}_{i2} + \hat{A}_{i3} + \hat{A}_{i4} + \hat{A}_{i5}) \right| \\ & \geq \phi'_{ik} \hat{A}_{i2} - \left| \phi'_{ik} (\hat{A}_{i3} + \hat{A}_{i4} + \hat{A}_{i5}) \right| \\ & \geq \lambda \frac{c_k^0}{2} - \left[2C\theta_{NT} + 2\bar{c}c \left(J^{-\frac{r_1}{d}} + J^{\frac{1}{2}} T^{-\frac{1}{2}} (\ln T)^3 \right) (\ln T)^v + C\lambda\mu_{1,JT} \right] \\ & \geq \lambda \frac{c_k^0}{4} \end{aligned}$$

where I use Assumption 1.2 and 1.3.

Thus

$$\begin{aligned}
& P(\hat{\mathbf{E}}_{kNT,i}) \\
&= P(i \notin \hat{G}_k | i \in G_k^0) \\
&= P(-\hat{A}_{i1} = \hat{A}_{i2} + \hat{A}_{i3} + \hat{A}_{i4} + \hat{A}_{i5}) \\
&\leq P\left(\left|\phi'_{ik}\hat{A}_{i1}\right| \geq \left|\phi'_{ik}(\hat{A}_{i2} + \hat{A}_{i3} + \hat{A}_{i4} + \hat{A}_{i5})\right|\right) \\
&\leq P\left(\left|\hat{A}_{i1}\right| \geq \lambda \frac{C_k^0}{4}, \Xi_{kNT}\right) + P(\Xi_{kNT}^c) \\
&= o(N^{-1})
\end{aligned}$$

Thus, with probability $1 - o(N^{-1})$ such that $\|\theta_i - \eta_k\|$ is not differentiable with respect to θ_i for some $i \in G_k^0$, which means that $P(\|\theta_i - \eta_k\| = 0 | i \in G_k^0) = 1 - o(N^{-1})$.

Then

$$\begin{aligned}
& P\left(\cup_{k=1}^{K^0} \hat{\mathbf{E}}_{kNT}\right) \\
&\leq \sum_{k=1}^{k^0} P(\hat{\mathbf{E}}_{kNT}) \\
&\leq \sum_{k=1}^{K^0} \sum_{i \in G_k^0} P(\hat{\mathbf{E}}_{kNT,i}) \\
&\leq \sum_{k=1}^{K^0} \sum_{i \in G_k^0} \left(P\left(\left|\hat{A}_{i1}\right| \geq \lambda \frac{C_k^0}{4}, \Xi_{kNT}\right) + P(\Xi_{kNT}^c) \right) \\
&\leq N \max_{1 \leq i \leq N} P\left(\left\|\hat{Q}_{i,\tilde{z}\tilde{u}}\right\| \geq \lambda \frac{C_k^0}{4}\right) + o(1) \\
&\leq NP\left(\max_{1 \leq i \leq N} \left\|\hat{Q}_{i,\tilde{z}\tilde{u}}\right\| \geq \lambda \frac{C_k^0}{4}\right) + o(1) \\
&= o(1)
\end{aligned}$$

where I use $\lambda T^{\frac{1}{2}} J^{-\frac{1}{2}} (\ln T)^{-3} \rightarrow \infty$.

(ii) The proof is similar to Su, Shi, and Phillips (2016) Theorem 2.2 (ii) and thus omitted.

□

Proof of Theorem 1.3

Proof. First, I will prove that $\sqrt{N_k T} (\hat{\alpha}_k - \hat{\alpha}_{\hat{G}_k}) = o_p(1)$. Then as long as I could prove (ii), consequently (i) holds as well.

- The first order conditions with respect to θ_i and η_k are

$$\begin{aligned} 0_{(p+J) \times 1} &= -\frac{2}{T} \sum_{t=1}^T \tilde{z}_{it} (\tilde{y}_{it} - \tilde{z}'_{it} \hat{\theta}_i) + \lambda \sum_{j=1}^{K^0} \hat{e}_{ij} \prod_{l=1, l \neq j}^{K^0} \|\hat{\theta}_i - \hat{\eta}_l\| \\ 0_{(p+J) \times 1} &= \lambda \sum_{i=1}^N \hat{e}_{ik} \prod_{l=1, l \neq k}^{K^0} \|\hat{\theta}_i - \hat{\eta}_l\| \end{aligned}$$

where $\hat{e}_{ij} = \frac{\hat{\theta}_i - \hat{\eta}_j}{\|\hat{\theta}_i - \hat{\eta}_j\|}$ if $\hat{\theta}_i \neq \hat{\eta}_j$ and $\hat{e}_{ij} \leq 1$ otherwise.

Note that (1) if $i \in \hat{G}_k$, $\|\hat{\theta}_i - \hat{\eta}_k\| = 0$. (2) if $i \in \hat{G}_k$ and $l \neq k$,

$$\|\hat{\theta}_i - \hat{\eta}_l\| = \left\| (\hat{\theta}_i - \eta_k^0) + (\eta_k^0 - \eta_l^0) - (\hat{\eta}_l - \eta_l^0) \right\| \asymp \|\eta_k^0 - \eta_l^0\|$$

Let \hat{G}_0 be the set of unclassified individuals.

Then I have

$$\begin{aligned} & \sum_{i \in \hat{G}_k} \sum_{j=1}^{K^0} \hat{e}_{ij} \prod_{l=1, l \neq j}^{K^0} \|\hat{\theta}_i - \hat{\eta}_l\| \\ &= \sum_{i \in \hat{G}_k} \hat{e}_{ik} \prod_{l=1, l \neq k}^{K^0} \|\hat{\theta}_i - \hat{\eta}_l\| + \sum_{i \in \hat{G}_k} \sum_{j=1, j \neq k}^{K^0} \hat{e}_{ij} \prod_{l=1, l \neq j}^{K^0} \|\hat{\theta}_i - \hat{\eta}_l\| \\ &= \sum_{i \in \hat{G}_k} \hat{e}_{ik} \prod_{l=1, l \neq k}^{K^0} \|\hat{\theta}_i - \hat{\eta}_l\| \end{aligned}$$

While

$$\begin{aligned}
& \sum_{i=1}^N \hat{e}_{ik} \prod_{l=1, l \neq k}^{K^0} \|\hat{\theta}_i - \hat{\eta}_l\| \\
&= \sum_{i \in \hat{G}_k} \hat{e}_{ik} \prod_{l=1, l \neq k}^{K^0} \|\hat{\theta}_i - \hat{\eta}_l\| + \sum_{i \in \hat{G}_0} \hat{e}_{ik} \prod_{l=1, l \neq k}^{K^0} \|\hat{\theta}_i - \hat{\eta}_l\| + \sum_{j=1, j \neq k}^{K^0} \sum_{i \in \hat{G}_j} \hat{e}_{ik} \prod_{l=1, l \neq k}^{K^0} \|\hat{\theta}_i - \hat{\eta}_l\| \\
&= \sum_{i \in \hat{G}_k} \hat{e}_{ik} \prod_{l=1, l \neq k}^{K^0} \|\hat{\theta}_i - \hat{\eta}_l\| + \sum_{i \in \hat{G}_0} \hat{e}_{ik} \prod_{l=1, l \neq k}^{K^0} \|\hat{\theta}_i - \hat{\eta}_l\|
\end{aligned}$$

Thus I have

$$\frac{2}{\hat{N}_k T} \sum_{t=1}^T \sum_{i \in \hat{G}_k} \tilde{z}_{it} (\tilde{y}_{it} - \tilde{z}'_{it} \hat{\eta}_k) + \frac{\lambda}{\hat{N}_k} \sum_{i \in \hat{G}_0} \hat{e}_{ik} \prod_{l=1, l \neq k}^{K^0} \|\hat{\eta}_k - \hat{\eta}_l\| = 0 \quad (1.12)$$

Let $\hat{Q}_{\hat{G}_k, \tilde{z}\tilde{z}} \equiv \frac{1}{\hat{N}_k T} \sum_{t=1}^T \sum_{i \in \hat{G}_k} \tilde{z}_{it} \tilde{z}'_{it}$, then

$$\begin{aligned}
\hat{\eta}_k &= \hat{Q}_{\hat{G}_k, \tilde{z}\tilde{z}}^{-1} \frac{1}{\hat{N}_k T} \sum_{t=1}^T \sum_{i \in \hat{G}_k} \tilde{z}_{it} \tilde{y}_{it} + \hat{Q}_{\hat{G}_k, \tilde{z}\tilde{z}}^{-1} \frac{\lambda}{2\hat{N}_k} \sum_{i \in \hat{G}_0} \hat{e}_{ik} \prod_{l=1, l \neq k}^{K^0} \|\hat{\eta}_k - \hat{\eta}_l\| \\
&= \hat{\eta}_{\hat{G}_k} + \hat{R}_k
\end{aligned}$$

where $\hat{R}_k = \hat{Q}_{\hat{G}_k, \tilde{z}\tilde{z}}^{-1} \frac{\lambda}{2\hat{N}_k} \sum_{i \in \hat{G}_0} \hat{e}_{ik} \prod_{l=1, l \neq k}^{K^0} \|\hat{\eta}_k - \hat{\eta}_l\|$.

For any $c > 0$,

$$\begin{aligned}
& P \left(\sqrt{\hat{N}_k T} \|\hat{\eta}_k - \hat{\eta}_{\hat{G}_k}\| \geq c \right) \\
&= P \left(\sqrt{\hat{N}_k T} \|\hat{R}_k\| \geq c \right) \\
&\leq \sum_{k=1}^{K^0} \sum_{i \in G_k^0} P \left(i \in \hat{G}_0 | i \in G_k^0 \right) \\
&\leq \sum_{k=1}^{K^0} \sum_{i \in G_k^0} P \left(i \notin \hat{G}_k | i \in G_k^0 \right) \\
&= o(1)
\end{aligned}$$

Thus $\sqrt{\hat{N}_k T} (\hat{\alpha}_k - \hat{\alpha}_{\hat{G}_k}) = o_p(1)$. By Theorem 1.2, similar to the proof in the first part, I could get that

$$\hat{N}_k \xrightarrow{P} N_k$$

Thus $\sqrt{\hat{N}_k T} (\hat{\alpha}_k - \hat{\alpha}_{\hat{G}_k}) = o_p(1)$.

- Now I focus on $\hat{\alpha}_{\hat{G}_k}$.

Let

$$\begin{aligned}\hat{Q}_{\hat{G}_k, \tilde{\omega} \setminus \tilde{B}} &= \hat{Q}_{\hat{G}_k, \tilde{\omega} \tilde{\omega}} - \hat{Q}_{\hat{G}_k, \tilde{\omega} \tilde{B}} \hat{Q}_{\hat{G}_k, \tilde{B} \tilde{B}}^{-1} \hat{Q}'_{\hat{G}_k, \tilde{\omega} \tilde{B}} \\ \hat{W}_{it, \tilde{\omega} \setminus \tilde{B}} &= \tilde{\omega}_{it} - \hat{Q}_{\hat{G}_k, \tilde{\omega} \tilde{B}} \hat{Q}_{\hat{G}_k, \tilde{B} \tilde{B}}^{-1} \sqrt{J} \tilde{B}_{it}^J \\ \hat{\Delta}_{it, \tilde{\omega} \setminus \tilde{B}} &= \mathbf{E}[\tilde{\omega}_{it} | x_{it}] - \hat{Q}_{\hat{G}_k, \tilde{\omega} \tilde{B}} \hat{Q}_{\hat{G}_k, \tilde{B} \tilde{B}}^{-1} \sqrt{J} \tilde{B}_{it}^J\end{aligned}$$

where $\hat{Q}_{\hat{G}_k, \tilde{\omega} \tilde{\omega}}, \hat{Q}_{\hat{G}_k, \tilde{\omega} \tilde{B}}$ are defined similar to $\hat{Q}_{\hat{G}_k, \tilde{z} \tilde{z}}$.

Then I have that

$$\begin{aligned}& \sqrt{\hat{N}_k T} \hat{\alpha}_{\hat{G}_k} \\ &= \left(\hat{Q}_{\hat{G}_k, \tilde{\omega} \setminus \tilde{B}} \right)^{-1} \frac{1}{\sqrt{\hat{N}_k T}} \sum_{t=1}^T \sum_{i \in \hat{G}_k} \hat{W}_{it, \tilde{\omega} \setminus \tilde{B}} \left(\tilde{\omega}'_{it} \theta_i^0 + \sqrt{J} \tilde{B}_{it}^{J'} \frac{1}{\sqrt{J}} \gamma_i^0 + \tilde{\delta}_{h_i, it} + \tilde{u}_{it} \right) \\ &= \left(\hat{Q}_{\hat{G}_k, \tilde{\omega} \setminus \tilde{B}} \right)^{-1} \frac{1}{\sqrt{\hat{N}_k T}} \sum_{t=1}^T \sum_{i \in \hat{G}_k} \hat{W}_{it, \tilde{\omega} \setminus \tilde{B}} \left(\tilde{\omega}'_{it} \theta_i^0 + \sqrt{J} \tilde{B}_{it}^{J'} \frac{1}{\sqrt{J}} \gamma_i^0 \right) \\ &\quad + \left(\hat{Q}_{\hat{G}_k, \tilde{\omega} \setminus \tilde{B}} \right)^{-1} \frac{1}{\sqrt{\hat{N}_k T}} \sum_{t=1}^T \sum_{i \in \hat{G}_k} \hat{\Delta}_{it, \tilde{\omega} \setminus \tilde{B}} \tilde{\delta}_{h_i, it} \\ &\quad + \left(\hat{Q}_{\hat{G}_k, \tilde{\omega} \setminus \tilde{B}} \right)^{-1} \frac{1}{\sqrt{\hat{N}_k T}} \sum_{t=1}^T \sum_{i \in \hat{G}_k} (\omega_{it} - \mathbf{E}[\omega_{it} | x_{it}]) \tilde{\delta}_{h_i, it} \\ &\quad + \left(\hat{Q}_{\hat{G}_k, \tilde{\omega} \setminus \tilde{B}} \right)^{-1} \frac{1}{\sqrt{\hat{N}_k T}} \sum_{t=1}^T \sum_{i \in \hat{G}_k} \hat{W}_{it, \tilde{\omega} \setminus \tilde{B}} \tilde{u}_{it} \\ &= \hat{B}_{k1} + \hat{B}_{k2} + \hat{B}_{k3} + \hat{B}_{k4}\end{aligned}$$

By Theorem 1.2, similar to the proof in the first part, I could get that

$$\begin{aligned}\hat{N}_k &\xrightarrow{P} N_k \\ \hat{Q}_{\hat{G}_k, \tilde{\omega} \setminus \tilde{B}} &\xrightarrow{P} \hat{Q}_{G_k^0, \tilde{\omega} \setminus \tilde{B}}\end{aligned}$$

Furthermore, I could prove that in the following analysis, whenever I encounter \hat{G}_k , I could safely replace it with G_k^0 , only up to $o_p(1)$ error.

(The proofs are similar to Corollary 2.3 in Su, Shi, and Phillips (2016) and Lemma A.6 in Su, Wang, and Jin (2019) and thus omitted.)

By the properties of the approximation error, $\hat{B}_{k2} = o_P(1)$, $\hat{B}_{k3} = o_p(1)$.

Now I consider \hat{B}_{k1} ,

$$\begin{aligned}& \left(\hat{Q}_{\hat{G}_k, \tilde{\omega} \setminus \tilde{B}} \right)^{-1} \frac{1}{\sqrt{\hat{N}_k T}} \sum_{t=1}^T \sum_{i \in \hat{G}_k} \hat{W}_{it, \tilde{\omega} \setminus \tilde{B}} \left(\tilde{\omega}'_{it} \theta_i^0 + \sqrt{J} \tilde{B}_{it}^{J'} \frac{1}{\sqrt{J}} \gamma_i^0 \right) \\ &= \left(\hat{Q}_{G_k^0, \tilde{\omega} \setminus \tilde{B}} \right)^{-1} \frac{1}{\sqrt{N_k T}} \sum_{t=1}^T \sum_{i \in G_k^0} W_{it, \tilde{\omega} \setminus \tilde{B}} \left(\omega'_{it} \theta_i^0 + \sqrt{J} \tilde{B}_{it}^{J'} \frac{1}{\sqrt{J}} \gamma_i^0 \right) + o_p(1) \\ &= \left(\hat{Q}_{G_k^0, \tilde{\omega} \setminus \tilde{B}} \right)^{-1} \frac{1}{\sqrt{N_k T}} \sum_{t=1}^T \sum_{i \in G_k^0} W_{it, \tilde{\omega} \setminus \tilde{B}} \left(\tilde{\omega}'_{it} \alpha_k^0 + \sqrt{J} \tilde{B}_{it}^{J'} \frac{1}{\sqrt{J}} \pi_k^0 \right) + o_p(1) \\ &= \sqrt{N_k T} \alpha_k^0 + o_p(1)\end{aligned}$$

I apply the similar procedure to \hat{B}_{k5} . Thus I have

$$\begin{aligned}& \sqrt{N_k T} \left(\hat{\alpha}_{\hat{G}_k} - \alpha_k^0 \right) \\ &= o_p(1) + \left(\hat{Q}_{G_k^0, \tilde{\omega} \setminus \tilde{B}} \right)^{-1} \frac{1}{\sqrt{N_k T}} \sum_{t=1}^T \sum_{i \in G_k^0} W_{it, \tilde{\omega} \setminus \tilde{B}} \tilde{u}_{it}\end{aligned}$$

Let $\sum_{t=1}^T W_{it, \tilde{\omega} \setminus \tilde{B}} \tilde{u}_{it} = W'_{i, \tilde{\omega} \setminus \tilde{B}} u_i = W'_{i, \tilde{\omega} \setminus \tilde{B}} \sum_i^{\frac{1}{2}} \varepsilon_i$, c be a $p \times 1$ nonrandom vector satisfying $\|c\| = 1$. Let

$$\begin{aligned}
B_k &= c' \left(\hat{Q}_{G_k^0, \tilde{\omega} \setminus \tilde{B}} \right)^{-1} \frac{1}{\sqrt{N_k T}} \sum_{i \in G_k^0} W'_{i, \tilde{\omega} \setminus \tilde{B}} \Sigma_i^{\frac{1}{2}} \varepsilon_i \\
&= \sum_{i \in G_k^0} a_i \xi_i
\end{aligned}$$

where $a_i = \left(\frac{1}{N_k T} c' \left(\hat{Q}_{G_k^0, \tilde{\omega} \setminus \tilde{B}} \right)^{-1} W'_{i, \tilde{\omega} \setminus \tilde{B}} \Sigma_i^{\frac{1}{2}} V_i \Sigma_i^{\frac{1}{2}} W_{i, \tilde{\omega} \setminus \tilde{B}} \left(\hat{Q}_{G_k^0, \tilde{\omega} \setminus \tilde{B}} \right)^{-1} c \right)^{\frac{1}{2}}$ and $\{\xi_i\}_{i=1}^N$ are independent with mean 0 and variance one conditional on $\{\omega_i, x_i\}_{i=1}^N$. Next, I just need to check the Lindeberg condition that if

$$\frac{\max_{i \in G_k^0} a_i^2}{\sum_{i \in G_k^0} a_i^2} = o_p(1)$$

then $\frac{\sum_{i \in G_k^0} a_i \xi_i}{\sum_{i \in G_k^0} a_i^2} \xrightarrow{D} N(0, 1)$.

Note that

$$\begin{aligned}
&\max_{i \in G_k^0} a_i^2 \\
&= \max_{i \in G_k^0} \frac{1}{N_k T} c' \left(\hat{Q}_{G_k^0, \tilde{\omega} \setminus \tilde{B}} \right)^{-1} W'_{i, \tilde{\omega} \setminus \tilde{B}} \Sigma_i^{\frac{1}{2}} V_i \Sigma_i^{\frac{1}{2}} W_{i, \tilde{\omega} \setminus \tilde{B}} \left(\hat{Q}_{G_k^0, \tilde{\omega} \setminus \tilde{B}} \right)^{-1} c \\
&\leq \frac{1}{N_k} \max_{i \in G_k^0} \mu_{\max}(V_i) \mu_{\max} \left(\frac{1}{T} W'_{i, \tilde{\omega} \setminus \tilde{B}} \Sigma_i W_{i, \tilde{\omega} \setminus \tilde{B}} \right) \left(\mu_{\min} \left(\hat{Q}_{G_k^0, \tilde{\omega} \setminus \tilde{B}} \right) \right)^{-2} \|c\| \\
&= o_p(1)
\end{aligned}$$

Then

$$\begin{aligned}
& \frac{\max_{i \in G_k^0} a_i^2}{\sum_{i \in G_k^0} a_i^2} \\
& \leq \frac{\frac{1}{N_k} \max_{i \in G_k^0} \mu_{\max}(V_i) \mu_{\max} \left(\frac{1}{T} W'_{i, \tilde{\omega} \setminus \tilde{B}} \sum_i W_{i, \tilde{\omega} \setminus \tilde{B}} \right) \left(\mu_{\min} \left(\hat{Q}_{G_k^0, \tilde{\omega} \setminus \tilde{B}} \right) \right)^{-2}}{\mu_{\min}(V_i) \mu_{\min} \left(\frac{1}{T} W'_{i, \tilde{\omega} \setminus \tilde{B}} \sum_i W_{i, \tilde{\omega} \setminus \tilde{B}} \right) \left(\mu_{\max} \left(\hat{Q}_{G_k^0, \tilde{\omega} \setminus \tilde{B}} \right) \right)^{-2}} \\
& = o_p(1)
\end{aligned}$$

Apply the central limit theorem, I have

$$\sqrt{\hat{N}_k T} \frac{c' \left(\hat{\alpha}_{\hat{G}_k} - \alpha_k^0 \right)}{\sum_{i \in G_k^0} a_i^2} \xrightarrow{D} N(0, 1)$$

Consequently,

$$\sqrt{\hat{N}_k T} V_{k, \omega}^{-\frac{1}{2}} \left(\hat{\alpha}_k - \alpha_k^0 \right) \xrightarrow{D} N(0, 1_p)$$

$$\text{where } V_{k, \omega} = \left(\hat{Q}_{G_k^0, \tilde{\omega} \setminus \tilde{B}} \right)^{-1} \frac{1}{N_k} \sum_{i \in G_k^0} \frac{1}{T} W'_{i, \tilde{\omega} \setminus \tilde{B}} \sum_i^{\frac{1}{2}} V_i \sum_i^{\frac{1}{2}} W_{i, \tilde{\omega} \setminus \tilde{B}} \left(\hat{Q}_{G_k^0, \tilde{\omega} \setminus \tilde{B}} \right)^{-1}.$$

□

Proof of Theorem 1.4

Proof. The proof of Theorem 1.4 is similar to that of Theorem 1.3 and thus omitted. □

1.B Proofs of Technical Lemmas

I use $\|\cdot\|$ to denote Frobenius norm in the Appendix for simplicity. I use C to indicate some generic constant, which varies.

Lemma 1.1. *Let ξ_{it} be a \mathcal{R}^{d_ξ} random variable and $\mathbf{E}[\xi_{it}] = 0$ for all i, t . For each $i = 1, \dots, N$, ξ_{it} is stationary strong mixing with mixing coefficient $\alpha_i(j)$. $\alpha(j) \equiv \max_{1 \leq i \leq N} \alpha_i(j)$ satisfies $\alpha(j) \leq c_\alpha \exp(-\rho j)$ for some $0 < c_\alpha < \infty$, $0 < \rho < \infty$. ξ_{it} are independent across i . Assume that $\mathbf{E}\|\xi_{it}\|^q < \infty$ for some $q \geq 3$, Then*

$$P \left(\max_{1 \leq i \leq N} \left\| \frac{1}{T} \sum_{t=1}^T \xi_{it} \right\| \geq CT^{-\frac{1}{2}} (\ln T)^3 \right) = o(N^{-1})$$

for large enough $C > 0$ if $N^2 T^{1-\frac{q}{2}} = O(1)$.

Proof. This lemma is adapted from Su, Shi, and Phillips (2016) Lemma S1.2 and could be derived using Theorem 2 of Merlevède et al. (2009). A slightly weaker version is

$$P \left(\max_{1 \leq i \leq N} \left\| \frac{1}{T} \sum_{t=1}^T \xi_{it} \right\| \geq c\lambda \right) = o(N^{-1})$$

for any $c > 0$ and λ satisfies that $T^{-\frac{1}{2}} (\ln T)^3 = o(\lambda)$. For convenience, I could choose $\lambda = T^{-\frac{1}{2}} (\ln T)^{3+v}$ for some $v > 0$. \square

Lemma 1.2. *Let ξ_{it} be a \mathcal{R}^{d_ξ} random variable and $\mathbf{E}[\xi_{it}] = 0$ for all i, t . For each $i = 1, \dots, N$, ξ_{it} is stationary strong mixing with mixing coefficient $\alpha_i(j)$. $\alpha(j) \equiv \max_{1 \leq i \leq N} \alpha_i(j)$ satisfies $\alpha(j) \leq c_\alpha \exp(-\rho j)$ for some $0 < c_\alpha < \infty$, $0 < \rho < \infty$. ξ_{it} are independent across i . Assume that $\max_{1 \leq i \leq N} \max_{1 \leq t \leq T} \mathbf{E}\|\xi_{it}\|^{\frac{q}{2}} < \infty$ for some $q > 6$ such that $N^2 T^{1-\frac{q}{2}} (\ln T)^{\frac{3q}{2}} \rightarrow 0$ as $N, T \rightarrow \infty$. Then*

$$P \left(\max_{1 \leq i \leq N} \left\| \frac{1}{T} \sum_{t=1}^T \xi_{it} \right\| \geq c \right) = o(N^{-1})$$

for any $c > 0$.

Proof. Let $\lambda_{NT} = N^2 T^{1-\frac{q}{2}} (\ln T)^{\frac{3q}{2}}$ and $\eta_{NT} = T (\ln T)^{-3} \lambda_{NT}^{\frac{2}{q}}$. Let τ_ξ be an arbitrary $d_\xi \times 1$ nonrandom vector with $\|\tau_\xi\| = 1$. Let $\mathbf{1}_{it} = \mathbf{1}\{\|\xi_{it}\| \leq \eta_{NT}\}$ and $\bar{\mathbf{1}}_{it} = \mathbf{1} - \mathbf{1}_{it}$. Define

$$\xi_{1,it} = \tau_\xi' \{\xi_{it} \mathbf{1}_{it} - \mathbf{E}[\xi_{it} \mathbf{1}_{it}]\}$$

$$\xi_{2,it} = \tau_\xi' \xi_{it} \bar{\mathbf{1}}_{it}$$

$$\xi_{3,it} = \tau_\xi' \mathbf{E}[\xi_{it} \bar{\mathbf{1}}_{it}]$$

Then $\xi_{1,it} + \xi_{2,it} - \xi_{3,it} = \tau_\xi' \xi_{it}$ since $\mathbf{E}[\xi_{it}] = 0$. I prove the lemma by showing that for any $c > 0$

$$(i) \quad NP \left(\max_{1 \leq i \leq N} \left\| \frac{1}{T} \sum_{t=1}^T \xi_{1,it} \right\| \geq c \right) = o(1)$$

$$(ii) \quad NP \left(\max_{1 \leq i \leq N} \left\| \frac{1}{T} \sum_{t=1}^T \xi_{2,it} \right\| \geq c \right) = o(1)$$

$$(iii) \quad \max_{1 \leq i \leq N} \left\| \frac{1}{T} \sum_{t=1}^T \xi_{3,it} \right\| = o(1)$$

To prove (i),

$$\begin{aligned} & NP \left(\max_{1 \leq i \leq N} \left\| \frac{1}{T} \sum_{t=1}^T \xi_{1,it} \right\| \geq c \right) \\ & \leq N \sum_{i=1}^N P \left(\left\| \frac{1}{T} \sum_{t=1}^T \xi_{1,it} \right\| \geq c \right) \\ & \leq N \sum_{i=1}^N \exp \left\{ - \frac{C_0 T^2 c^2}{T v_0^2 + \eta_{NT}^2 + T c \eta_{NT} (\ln T)^2} \right\} \\ & \leq N^2 \exp \left\{ - \frac{C_0 T^2 c^2}{T v_{0,\max}^2 + \eta_{NT}^2 + T c \eta_{NT} (\ln T)^2} \right\} \\ & \leq \exp \left\{ - \frac{C_0 T^2 c^2}{T v_{0,\max}^2 + T^2 (\ln T)^{-6} \lambda_{NT}^{\frac{4}{q}} + T c T (\ln T)^{-3} \lambda_{NT}^{\frac{2}{q}} (\ln T)^2} + 2 \ln N \right\} \\ & \rightarrow 0 \end{aligned}$$

To prove (ii),

$$\begin{aligned}
& NP \left(\max_{1 \leq i \leq N} \left\| \frac{1}{T} \sum_{t=1}^T \xi_{2,it} \right\| \geq c \right) \\
& \leq NP \left(\max_{1 \leq i \leq N} \max_{1 \leq t \leq T} \|\xi_{it}\| > \eta_{NT} \right) \\
& \leq N^2 T \max_{1 \leq i \leq N} \max_{1 \leq t \leq T} P(\|\xi_{it}\| > \eta_{NT}) \\
& \leq N^2 T \frac{1}{T^{\frac{q}{2}} (\ln T)^{-\frac{3q}{2}} \lambda_{NT}} \max_{1 \leq i \leq N} \max_{1 \leq t \leq T} \mathbf{E} \left[\|\xi_{it}\|^{\frac{q}{2}} \mathbf{1} \left\{ \|\xi_{it}\| > T (\ln T)^{-3} \lambda_{NT}^{\frac{2}{q}} \right\} \right] \\
& = o \left(N^2 T^{1-\frac{q}{2}} (\ln T)^{\frac{3q}{2}} \lambda_{NT}^{-1} \right) \\
& = o(1)
\end{aligned}$$

To prove (iii),

$$\begin{aligned}
& \max_{1 \leq i \leq N} \left\| \frac{1}{T} \sum_{t=1}^T \xi_{3,it} \right\| \\
& \leq \max_{1 \leq i \leq N} \max_{1 \leq t \leq T} \left\| \mathbf{E} [\xi_{it} \bar{\mathbf{1}}_{it}] \right\| \\
& \leq \max_{1 \leq i \leq N} \max_{1 \leq t \leq T} \left\{ \left(\mathbf{E} \|\xi_{it}\|^{\frac{q}{2}} \right)^{\frac{2}{q}} \left(P(\|\xi_{it}\| > \eta_{NT}) \right)^{\frac{q-2}{q}} \right\} \\
& \leq \max_{1 \leq i \leq N} \max_{1 \leq t \leq T} \left\{ \left(\mathbf{E} \|\xi_{it}\|^{\frac{q}{2}} \right)^{\frac{2}{q}} \right\} \times \max_{1 \leq i \leq N} \max_{1 \leq t \leq T} \left\{ \left(P(\|\xi_{it}\| > \eta_{NT}) \right)^{\frac{q-2}{q}} \right\} \\
& \leq c_{\xi} \max_{1 \leq i \leq N} \max_{1 \leq t \leq T} \left\{ \left(\eta_{NT}^{-\frac{q-2}{2}} \mathbf{E} \left[\|\xi_{it}\|^{\frac{q}{2}} \mathbf{1} \{ \|\xi_{it}\| > \eta_{NT} \} \right] \right)^{\frac{q-2}{q}} \right\} \\
& = o(1)
\end{aligned}$$

This completes the proof. □

Lemma 1.3. *Suppose that Assumption 1.1 and 1.2 hold, then*

(i)

$$P(0 < \underline{c} < \min_{0 \leq i \leq N} \mu_{\min}(\hat{Q}_{i,\bar{z}\bar{z}}) \leq \max_{0 \leq i \leq N} \mu_{\max}(\hat{Q}_{i,\bar{z}\bar{z}}) < \bar{c} < \infty) = 1 - o(N^{-1})$$

(ii)

$$\|\hat{Q}_{i,\bar{z}\bar{e}}\| = O_p(J^{-\frac{r_1}{d}} + J^{\frac{1}{2}}T^{-\frac{1}{2}})$$

(iii)

$$\frac{1}{N} \sum_{i=1}^N \|\hat{Q}_{i,\bar{z}\bar{e}}\|^2 = O_p(J^{-2\frac{r_1}{d}} + JT^{-1})$$

(iv)

$$P\left(\max_{0 \leq i \leq N} \|\hat{Q}_{i,\bar{z}\bar{e}}\| \geq c \left(J^{-\frac{r_1}{d}} + J^{\frac{1}{2}}T^{-\frac{1}{2}} (\ln T)^3\right) (\ln T)^v\right) = o(N^{-1})$$

for any $c > 0$ and some $v > 0$.

Proof. (i) Consider the difference between $\text{Var}(z_{it})$ and $\hat{Q}_{i,\bar{z}\bar{z}}$.

Let $\mu_k(A)$ be the k th largest eigenvalue of matrix A . Denote \mathbb{S}_{p+J} as the permutation group of $\{1, \dots, p+J\}$. By Hoffman-Wielandt inequality,

$$\min_{\sigma \in \mathbb{S}_{p+J}} \sum_{k=1}^{p+J} \left| \mu_k(\hat{Q}_{i,\bar{z}\bar{z}}) - \mu_{\sigma(k)}(\text{Var}(z_{it})) \right|^2 \leq \|\hat{Q}_{i,\bar{z}\bar{z}} - \text{Var}(z_{it})\|^2$$

Because

$$\begin{aligned} & \|\hat{Q}_{i,\bar{z}\bar{z}} - \text{Var}(z_{it})\|^2 \\ & \leq 2 \left\| \hat{Q}_{i,zz} - \mathbf{E}[z_{it}z'_{it}] \right\|^2 + 2 \left\| \frac{1}{T} \sum_{t=1}^T z_{it} \frac{1}{T} \sum_{t=1}^T z'_{it} - \mathbf{E}[z_{it}] \mathbf{E}[z'_{it}] \right\|^2 \end{aligned}$$

(i) Consider the first item, for any $c > 0$,

- By Lemma 1.2,

$$P\left(\max_{1 \leq i \leq N} \left| \frac{1}{T} \sum_{t=1}^T \omega_{it,j} \omega_{it,k} - \mathbf{E}[\omega_{it,j} \omega_{it,k}] \right| \geq c\right) = o(N^{-1})$$

- Similar as the proof in Lemma 1.2, I could get that

$$P \left(\max_{1 \leq i \leq N} \max_{1 \leq k \leq J} \left| \frac{1}{T} \sum_{t=1}^T \omega_{it,j} \sqrt{J} B_{it,k}^J - \mathbf{E} [\omega_{it,j} \sqrt{J} B_{it,k}^J] \right| \geq cJ^{-\frac{1}{2}} \right) = o(N^{-1})$$

where I use $\lambda_{NT} = N^2 T^{1-q} J^q (\ln T)^{3q} \rightarrow 0$ as $(N, T) \rightarrow \infty$, which could be derived by $J^2 (\ln T)^3 T^{-1} \rightarrow 0$ and $N^2 T^{1-\frac{q}{2}} (\ln T)^{\frac{3q}{2}} \rightarrow 0$. And I set $\eta_{NT} = T J^{-1} (\ln T)^{-3} \lambda_{NT}^{\frac{1}{q}}$ and $\xi_{it,jk} = \omega_{it,j} B_{it,k}^J$.

- Similar as the proof in Lemma 1.2 (only the first step is enough),

$$P \left(\max_{1 \leq i \leq N} \max_{1 \leq j \leq J} \max_{1 \leq k \leq J} \left| \frac{1}{T} \sum_{t=1}^T J B_{it,j}^J B_{it,k}^J - \mathbf{E} [J B_{it,j}^J B_{it,k}^J] \right| \geq cJ^{-\frac{1}{2}} \right) = o(N^{-1})$$

Note that there are only $O(J)$ nonzero elements in $B_{it}^J B_{it}^{J'} - \mathbf{E} [B_{it}^J B_{it}^{J'}]$.

Thus for any $c > 0$,

$$P \left(\max_{1 \leq i \leq N} \left\| \hat{Q}_{i,zz} - \mathbf{E} [z_{it} z'_{it}] \right\|^2 \geq c \right) = o(N^{-1})$$

(ii) Consider the second item, for any $c > 0$, similar as the proof in Lemma 1.2,

•

$$P \left(\max_{1 \leq i \leq N} \left| \frac{1}{T} \sum_{t=1}^T \omega_{it,j} - \mathbf{E} [\omega_{it,j}] \right| \geq cJ^{-\frac{1}{2}} \right) = o(N^{-1})$$

•

$$P \left(\max_{1 \leq i \leq N} \max_{1 \leq k \leq J} \left| \frac{1}{T} \sum_{t=1}^T \sqrt{J} B_{it,k}^J - \mathbf{E} [\sqrt{J} B_{it,k}^J] \right| \geq cJ^{-1} \right) = o(N^{-1})$$

Thus I could get

$$P \left(\max_{1 \leq i \leq N} \left\| \frac{1}{T} \sum_{t=1}^T z_{it} \frac{1}{T} \sum_{t=1}^T z'_{it} - \mathbf{E}[z_{it}] \mathbf{E}[z'_{it}] \right\|^2 \geq c \right) = o(N^{-1})$$

Combining part (i) and (ii) together, I have

$$P \left(\min_{\sigma \in \mathbb{S}_{p+J}} \sum_{k=1}^{p+J} \left| \mu_k(\hat{Q}_{i,\tilde{z}\tilde{z}}) - \mu_{\sigma(k)}(\text{Var}(z_{it})) \right|^2 \leq c \right) = 1 - o(N^{-1})$$

(ii) Let $\hat{Q}_{i,\tilde{z}\tilde{\delta}} = \frac{1}{T} \sum_{t=1}^T \tilde{z}_{it} \tilde{\delta}_{h_i,it}$ and $\hat{Q}_{i,\tilde{z}\tilde{u}} = \frac{1}{T} \sum_{t=1}^T \tilde{z}_{it} \tilde{u}_{it}$, where $\tilde{\delta}_{h_i,it} = \tilde{h}_{i,it}^0 - \tilde{B}_{it}^{J'} \gamma_i^0$, then I have $\|\hat{Q}_{i,\tilde{z}\tilde{e}}\| \leq \|\hat{Q}_{i,\tilde{z}\tilde{\delta}}\| + \|\hat{Q}_{i,\tilde{z}\tilde{u}}\|$.

For the first part, since

$$\begin{aligned} & \|\hat{Q}_{i,\tilde{z}\tilde{\delta}}\| \\ &= \left\| \hat{Q}_{i,z\delta} - \frac{1}{T} \sum_{t=1}^T z_{it} \frac{1}{T} \sum_{t=1}^T \delta_{h_i,it} \right\| \\ &\leq \|\hat{Q}_{i,z\delta}\| + \left\| \frac{1}{T} \sum_{t=1}^T z_{it} \frac{1}{T} \sum_{t=1}^T \delta_{h_i,it} \right\| \end{aligned}$$

- For the first item,

$$\mathbf{E} \left[\|\hat{Q}_{i,z\delta}\|^2 \right] = \mathbf{E} \left[\left\| \frac{1}{T} \sum_{t=1}^T \omega'_{it} \delta_{h_i,it} \right\|^2 \right] + \mathbf{E} \left[\left\| \frac{1}{T} \sum_{t=1}^T \sqrt{J} B_{it}^{J'} \delta_{h_i,it} \right\|^2 \right]$$

(i)

$$\begin{aligned}
& \mathbf{E} \left[\left\| \frac{1}{T} \sum_{t=1}^T \omega'_{it} \delta_{h_i, it} \right\|^2 \right] \\
&= \frac{1}{T^2} \sum_{t=1}^T \sum_{s=1}^T \mathbf{E} \left[\omega'_{it} \omega_{is} \delta_{h_i, it} \delta_{h_i, is} \right] \\
&\leq \theta_{NT}^2 \frac{1}{T^2} \sum_{t=1}^T \sum_{s=1}^T \mathbf{E} \left[|\omega'_{it} \omega_{is}| \right] \\
&\leq \theta_{NT}^2 \frac{1}{T^2} \sum_{t=1}^T \sum_{s=1}^T \sqrt{\mathbf{E} \left[\|\omega_{it}\|^2 \right]} \sqrt{\mathbf{E} \left[\|\omega_{is}\|^2 \right]} \\
&\leq \theta_{NT}^2 \max_{1 \leq i \leq N} \max_{1 \leq t \leq T} \mathbf{E} \left[\|\omega_{it}\|^2 \right] \\
&= O \left(J^{-2\frac{r_1}{d}} \right)
\end{aligned}$$

(ii)

$$\begin{aligned}
& \mathbf{E} \left[\left\| \frac{1}{T} \sum_{t=1}^T \sqrt{J} B_{it}^{J'} \delta_{h_i, it} \right\|^2 \right] \\
&= \frac{1}{T^2} \sum_{t=1}^T \sum_{s=1}^T \mathbf{E} \left[J B_{it}^{J'} B_{is}^J \delta_{h_i, it} \delta_{h_i, is} \right] \\
&\leq \theta_{NT}^2 J \frac{1}{T^2} \sum_{t=1}^T \sum_{s=1}^T \mathbf{E} \left[B_{it}^{J'} B_{is}^J \right] \\
&= \theta_{NT}^2 J \mathbf{E} \left[\frac{1}{T} \sum_{t=1}^T B_{it}^{J'} \frac{1}{T} \sum_{s=1}^T B_{it}^J \right] \\
&= \theta_{NT}^2 J \sum_{j=1}^J \mathbf{E} \left[\frac{1}{T} \sum_{t=1}^T B_{it, j}^J \frac{1}{T} \sum_{s=1}^T B_{it, j}^J \right] \\
&= O \left(J^{-2\frac{r_1}{d}} \right)
\end{aligned}$$

Thus $\mathbf{E} \left[\left\| \hat{Q}_{i, z\delta} \right\|^2 \right] = O \left(J^{-2\frac{r_1}{d}} \right)$.

- For the second item,

$$\begin{aligned} & \mathbf{E} \left[\left\| \frac{1}{T} \sum_{t=1}^T z_{it} \frac{1}{T} \sum_{t=1}^T \delta_{h_i, it} \right\|^2 \right] \\ = & \mathbf{E} \left[\left\| \frac{1}{T} \sum_{t=1}^T \omega_{it} \frac{1}{T} \sum_{t=1}^T \delta_{h_i, it} \right\|^2 \right] + \mathbf{E} \left[\left\| \frac{1}{T} \sum_{t=1}^T \sqrt{J} B_{it}^J \frac{1}{T} \sum_{t=1}^T \delta_{h_i, it} \right\|^2 \right] \end{aligned}$$

Similarly, I could get that

$$\mathbf{E} \left[\left\| \frac{1}{T} \sum_{t=1}^T z_{it} \frac{1}{T} \sum_{t=1}^T \delta_{h_i, it} \right\|^2 \right] = O\left(J^{-2\frac{r_1}{d}}\right)$$

For the second part, similarly

$$\begin{aligned} & \left\| \hat{Q}_{i, \bar{z}\bar{u}} \right\| \\ = & \left\| \hat{Q}_{i, zu} - \frac{1}{T} \sum_{t=1}^T z_{it} \frac{1}{T} \sum_{t=1}^T u_{it} \right\| \\ \leq & \left\| \hat{Q}_{i, zu} \right\| + \left\| \frac{1}{T} \sum_{t=1}^T z_{it} \frac{1}{T} \sum_{t=1}^T u_{it} \right\| \end{aligned}$$

- Consider the first item,

$$\mathbf{E} \left[\left\| \hat{Q}_{i, zu} \right\|^2 \right] = \mathbf{E} \left[\left\| \frac{1}{T} \sum_{t=1}^T \omega'_{it} u_{it} \right\|^2 \right] + \mathbf{E} \left[\left\| \frac{1}{T} \sum_{t=1}^T \sqrt{J} B_{it}^{J'} u_{it} \right\|^2 \right]$$

(i)

$$\begin{aligned}
& \mathbf{E} \left[\left\| \frac{1}{T} \sum_{t=1}^T \omega'_{it} u_{it} \right\|^2 \right] \\
&= \frac{1}{T^2} \sum_{t=1}^T \sum_{s=1}^T \mathbf{E} \left[\omega'_{it} \omega_{it} u_{it} u_{is} \right] \\
&= \frac{1}{T^2} \sum_{t=1}^T \sum_{s=1}^T \mathbf{E} \left[\|\omega_{it}\|^2 u_{it}^2 \right] + \frac{2}{T^2} \sum_{t=1}^{T-1} \sum_{s=t+1}^T \mathbf{E} \left[\omega'_{it} \omega_{is} u_{it} u_{is} \right] \\
&= O(T^{-1}) + \frac{2}{T^2} \sum_{j=1}^p \sum_{t=1}^{T-1} \sum_{s=t+1}^T \mathbf{E} \left[\omega_{it,j} \omega_{is,j} u_{it} u_{is} \right] \\
&\leq O(T^{-1}) + \frac{C}{T^2} \sum_{j=1}^p \max_{1 \leq i \leq N} \max_{1 \leq t \leq T} \left\{ \mathbf{E} \left[\left| \omega_{it,j} u_{it} \right|^{\frac{q}{2}} \right] \right\}^{\frac{4}{q}} \sum_{t=1}^T \sum_{l=1}^{\infty} (\alpha(l))^{\frac{q-4}{q}} \\
&= O(T^{-1})
\end{aligned}$$

(ii)

$$\begin{aligned}
& \mathbf{E} \left[\left\| \frac{1}{T} \sum_{t=1}^T \sqrt{J} B_{it}^{J'} u_{it} \right\|^2 \right] \\
&= \frac{1}{T^2} \sum_{t=1}^T \sum_{s=1}^T \mathbf{E} \left[J B_{it}^{J'} B_{is}^J u_{it} u_{is} \right] \\
&\leq \frac{CJ}{T^2} \sum_{t=1}^T \sum_{s=1}^T \mathbf{E} \left[u_{it} u_{is} \right] \\
&= O(T^{-1}J)
\end{aligned}$$

Thus $\mathbf{E} \left[\left\| \hat{Q}_{i,zu} \right\|^2 \right] = O(T^{-1}J)$.

- Consider the second item,

$$\begin{aligned}
& \mathbf{E} \left[\left\| \frac{1}{T} \sum_{t=1}^T z_{it} \frac{1}{T} \sum_{t=1}^T u_{it} \right\|^2 \right] \\
&= \mathbf{E} \left[\left\| \frac{1}{T} \sum_{t=1}^T \omega_{it} \frac{1}{T} \sum_{t=1}^T u_{it} \right\|^2 \right] + \mathbf{E} \left[\left\| \frac{1}{T} \sum_{t=1}^T \sqrt{J} B_{it}^J \frac{1}{T} \sum_{t=1}^T u_{it} \right\|^2 \right] \\
&= O(T^{-1})
\end{aligned}$$

Thus $\|\hat{Q}_{i,\bar{z}\bar{u}}\| = O_p(J^{\frac{1}{2}}T^{-\frac{1}{2}})$.

In sum, I have proved that

$$\|\hat{Q}_{i,ze}\| = O_p(J^{-\frac{r_1}{d}} + J^{\frac{1}{2}}T^{-\frac{1}{2}})$$

(iii) Consider

$$\begin{aligned}
& \mathbf{E} \left[\frac{1}{N} \sum_{i=1}^N \|\hat{Q}_{i,\bar{z}\bar{e}}\|^2 \right] \\
&= \frac{1}{N} \sum_{i=1}^N \mathbf{E} \left[\|\hat{Q}_{i,\bar{z}\bar{e}}\|^2 \right] \\
&\leq \frac{2}{N} \sum_{i=1}^N \left(\mathbf{E} \left[\|\hat{Q}_{i,\bar{z}\bar{\delta}}\|^2 \right] + \mathbf{E} \left[\|\hat{Q}_{i,\bar{z}\bar{u}}\|^2 \right] \right)
\end{aligned}$$

Note that from the proof of (ii), I could strengthen the results to

$$\begin{aligned}
\max_{1 \leq i \leq N} \mathbf{E} \left[\|\hat{Q}_{i,\bar{z}\bar{\delta}}\|^2 \right] &= O\left(J^{-2\frac{r_1}{d}}\right) \\
\max_{1 \leq i \leq N} \mathbf{E} \left[\|\hat{Q}_{i,\bar{z}\bar{u}}\|^2 \right] &= O(T^{-1}J)
\end{aligned}$$

Consequently,

$$\mathbf{E} \left[\frac{1}{N} \sum_{i=1}^N \|\hat{Q}_{i,ze}\|^2 \right] = O\left(J^{-2\frac{r_1}{d}} + T^{-1}J\right)$$

This completes the proof.

(iv) Note that $\|\hat{Q}_{i,\bar{z}\bar{e}}\| = \|\hat{Q}_{i,\bar{z}\bar{\delta}}\| + \|\hat{Q}_{i,\bar{z}\bar{u}}\|$. To prove (iv), I can show that for large enough $C > 0$, any $c > 0$ and any $v > 0$,

$$P\left(\max_{1 \leq i \leq N} \|\hat{Q}_{i,\bar{z}\bar{\delta}}\| \geq CJ^{-\frac{r_1}{d}}\right) = o(N^{-1})$$

$$P\left(\max_{1 \leq i \leq N} \|\hat{Q}_{i,\bar{z}\bar{u}}\| \geq cJ^{\frac{1}{2}}T^{-\frac{1}{2}}(\ln T)^{3+v}\right) = o(N^{-1})$$

(i) For the first part, consider $\|\hat{Q}_{i,z\delta}\|$ and $\left\|\frac{1}{T} \sum_{t=1}^T z_{it} \frac{1}{T} \sum_{t=1}^T \delta_{h_i, it}\right\|$ separately. First,

$$\begin{aligned} & \|\hat{Q}_{i,z\delta}\|^2 \\ &= \left\| \frac{1}{T} \sum_{t=1}^T \omega_{it} \delta_{h_i, it} \right\|^2 + \left\| \frac{1}{T} \sum_{t=1}^T \sqrt{J} B_{it}^J \delta_{h_i, it} \right\|^2 \\ &\leq \theta_{NT}^2 \frac{1}{T} \sum_{t=1}^T \|\omega_{it}\|^2 + \theta_{NT}^2 J \sum_{j=1}^J \left(\frac{1}{T} \sum_{t=1}^T B_{it,j}^J \right)^2 \end{aligned}$$

First consider $\frac{1}{T} \sum_{t=1}^T \|\omega_{it}\|^2$. By Lemma 1.2, for any $c > 0$,

$$P\left(\max_{1 \leq i \leq N} \left| \frac{1}{T} \sum_{t=1}^T \left(\|\omega_{it}\|^2 - \mathbf{E} [\|\omega_{it}\|^2] \right) \right| \geq c\right) = o(N^{-1})$$

Then for large enough $C > 0$, I could show that

$$\begin{aligned} & P\left(\max_{1 \leq i \leq N} \frac{1}{T} \sum_{t=1}^T \|\omega_{it}\|^2 \geq C\right) \\ &\leq P\left(\max_{1 \leq i \leq N} \frac{1}{T} \sum_{t=1}^T \mathbf{E} [\|\omega_{it}\|^2] + \max_{1 \leq i \leq N} \left| \frac{1}{T} \sum_{t=1}^T \left(\|\omega_{it}\|^2 - \mathbf{E} [\|\omega_{it}\|^2] \right) \right| \geq C\right) \\ &= o(N^{-1}) \end{aligned}$$

Next consider $\frac{1}{T} \sum_{t=1}^T B_{it,j}^J$, for any $c > 0$ and $1 \leq j \leq J$, I want to show

$$P \left(\max_{1 \leq j \leq J} \max_{1 \leq i \leq N} \left| \frac{1}{T} \sum_{t=1}^T B_{it,j}^J - \mathbf{E} [B_{it,j}^J] \right| \geq cJ^{-1} \right) = o(N^{-1})$$

Since

$$\begin{aligned} & NP \left(\max_{1 \leq j \leq J} \max_{1 \leq i \leq N} \left| \frac{1}{T} \sum_{t=1}^T B_{it,j}^J - \mathbf{E} [B_{it,j}^J] \right| \geq cJ^{-1} \right) \\ & \leq N \sum_{i=1}^N \sum_{j=1}^J P \left(\left| \frac{1}{T} \sum_{t=1}^T B_{it,j}^J - \mathbf{E} [B_{it,j}^J] \right| \geq cJ^{-1} \right) \\ & \leq N^2 J \exp \left(-\frac{C_0 c^2 T^2 J^{-2}}{T v_{0,\max} + 2 + 2cT J^{-1} (\ln T)^2} \right) \end{aligned}$$

As long as $(\ln T)^3 J T^{-1} = o(1)$, I could get the result. Then for large enough $C > 0$ and for any $1 \leq j \leq J$,

$$\begin{aligned} & P \left(\max_{1 \leq j \leq J} \max_{1 \leq i \leq N} \frac{1}{T} \sum_{t=1}^T B_{it,j}^J \geq C J^{-1} \right) \\ & \leq P \left(\max_{1 \leq j \leq J} \max_{1 \leq i \leq N} \frac{1}{T} \sum_{t=1}^T \mathbf{E} [B_{it,j}^J] \right. \\ & \quad \left. + \max_{1 \leq j \leq J} \max_{1 \leq i \leq N} \left| \frac{1}{T} \sum_{t=1}^T B_{it,j}^J - \mathbf{E} [B_{it,j}^J] \right| \geq C J^{-1} \right) \\ & = o(N^{-1}) \end{aligned}$$

Thus for large enough $C > 0$,

$$\begin{aligned}
& P \left(\max_{1 \leq i \leq N} J \sum_{j=1}^J \left(\frac{1}{T} \sum_{t=1}^T B_{it,j}^J \right)^2 \geq C^2 \right) \\
& \leq P \left(J^2 \max_{1 \leq j \leq J} \max_{1 \leq i \leq N} \left(\frac{1}{T} \sum_{t=1}^T B_{it,j}^J \right)^2 \geq C^2 \right) \\
& \leq P \left(\max_{1 \leq j \leq J} \max_{1 \leq i \leq N} \left(\frac{1}{T} \sum_{t=1}^T B_{it,j}^J \right)^2 \geq C^2 J^{-2} \right) \\
& \leq P \left(\left(\max_{1 \leq j \leq J} \max_{1 \leq i \leq N} \frac{1}{T} \sum_{t=1}^T B_{it,j}^J \right)^2 \geq C^2 J^{-2} \right) \\
& = o(N^{-1})
\end{aligned}$$

Combining the previous results, I have for large enough $C > 0$

$$\begin{aligned}
& P \left(\max_{1 \leq i \leq N} \|\hat{Q}_{i,z\delta}\| \geq C J^{-\frac{r_1}{d}} \right) \\
& \leq P \left(\max_{1 \leq i \leq N} \|\hat{Q}_{i,z\delta}\|^2 \geq C^2 J^{-2\frac{r_1}{d}} \right) \\
& \leq P \left(\max_{1 \leq i \leq N} \|\hat{Q}_{i,\omega\delta}\|^2 + \max_{1 \leq i \leq N} \|\hat{Q}_{i,B\delta}\|^2 \geq C^2 J^{-2\frac{r_1}{d}} \right) \\
& \leq P \left(\max_{1 \leq i \leq N} \|\hat{Q}_{i,\omega\delta}\|^2 \geq \frac{1}{2} C^2 J^{-2\frac{r_1}{d}} \right) + P \left(\max_{1 \leq i \leq N} \|\hat{Q}_{i,B\delta}\|^2 \geq \frac{1}{2} C^2 J^{-2\frac{r_1}{d}} \right) \\
& \leq P \left(\theta_{NT}^2 \max_{1 \leq i \leq N} \frac{1}{T} \sum_{t=1}^T \|\omega_{it}\|^2 \geq \frac{1}{2} C^2 J^{-2\frac{r_1}{d}} \right) \\
& \quad + P \left(\theta_{NT}^2 \max_{1 \leq i \leq N} J \sum_{j=1}^J \left(\frac{1}{T} \sum_{t=1}^T B_{it,j}^J \right)^2 \geq \frac{1}{2} C^2 J^{-2\frac{r_1}{d}} \right) \\
& \leq P \left(\max_{1 \leq i \leq N} \frac{1}{T} \sum_{t=1}^T \|\omega_{it}\|^2 \geq C \right) + P \left(\max_{1 \leq i \leq N} J \sum_{j=1}^J \left(\frac{1}{T} \sum_{t=1}^T B_{it,j}^J \right)^2 \geq C \right) \\
& = o(N^{-1})
\end{aligned}$$

Similarly, I could prove that for large enough $C > 0$

$$P \left(\max_{1 \leq i \leq N} \left\| \frac{1}{T} \sum_{t=1}^T z_{it} \frac{1}{T} \sum_{t=1}^T \delta_{h_i, it} \right\| \geq C J^{-\frac{r_1}{d}} \right) = o(N^{-1})$$

Thus $P \left(\max_{1 \leq i \leq N} \|\hat{Q}_{i, \tilde{z}\tilde{\delta}}\| \geq C J^{-\frac{r_1}{d}} \right) = o(N^{-1})$.

(ii) For the second part, since

$$\begin{aligned} & \|\hat{Q}_{i, \tilde{z}\tilde{u}}\| \\ & \leq \|\hat{Q}_{i, \tilde{\omega}\tilde{u}}\| + \|\hat{Q}_{i, \tilde{B}\tilde{u}}\| \\ & = \left\| \frac{1}{T} \sum_{t=1}^T \omega_{it} u_{it} \right\| + \left\| \frac{1}{T} \sum_{t=1}^T \omega_{it} \frac{1}{T} \sum_{t=1}^T u_{it} \right\| \\ & \quad + \left\| \frac{1}{T} \sum_{t=1}^T \sqrt{J} B_{it}^J u_{it} \right\| + \left\| \frac{1}{T} \sum_{t=1}^T \sqrt{J} B_{it}^J \frac{1}{T} \sum_{t=1}^T u_{it} \right\| \end{aligned}$$

First consider $\left\| \frac{1}{T} \sum_{t=1}^T \omega_{it} u_{it} \right\|$. By Lemma 1.1, for any $c > 0$ and $v > 0$,

$$P \left(\max_{1 \leq i \leq N} \left\| \frac{1}{T} \sum_{t=1}^T \omega_{it} u_{it} \right\| \geq c T^{-\frac{1}{2}} (\ln T)^{3+v} \right) = o(N^{-1})$$

Similarly, I could show that

$$P \left(\max_{1 \leq i \leq N} \left\| \frac{1}{T} \sum_{t=1}^T \omega_{it} \frac{1}{T} \sum_{t=1}^T u_{it} \right\| \geq c T^{-\frac{1}{2}} (\ln T)^{3+v} \right) = o(N^{-1})$$

Next consider $\left\| \frac{1}{T} \sum_{t=1}^T \sqrt{J} B_{it}^J u_{it} \right\|$, By Lemma 1.1, for any $c > 0$ and $v > 0$,

$$P \left(\max_{1 \leq i \leq N} \left\| \frac{1}{T} \sum_{t=1}^T \sqrt{J} B_{it}^J u_{it} \right\| \geq c J^{\frac{1}{2}} T^{-\frac{1}{2}} (\ln T)^{3+v} \right) = o(N^{-1})$$

Similarly,

$$P \left(\max_{1 \leq i \leq N} \left\| \frac{1}{T} \sum_{t=1}^T \sqrt{J} B_{it}^J \frac{1}{T} \sum_{t=1}^T u_{it} \right\| \geq c J^{\frac{1}{2}} T^{-\frac{1}{2}} (\ln T)^{3+v} \right) = o(N^{-1})$$

This completes the proof.

□

Chapter 2

Nonparametric Additive Panel

Regression Models with Grouped

Heterogeneity

2.1 Introduction

Panel regression models have attracted considerable attention in both theoretical and applied econometrics. They provide researchers a convenient way to tackle unobserved heterogeneity that plays an important role in panel data analysis. Over the past few decades, substantial progress has been made in terms of the identification and estimation of various panel regression models; see Arellano and Honoré (2001), Mátyás and Sevestre (2013) and Baltagi (2015) for a comprehensive review. However, most of the literature uses fixed effects to control for individual-specific heterogeneity. Even though such a modeling scheme facilitates technical analysis, it ignores the potential nonlinear effects of explanatory variables and non-additive heterogeneity, both of which have been emphasized by multiple empirical studies. For example, using panel data of listed firms in the Chinese stock market, Ni, Wang, and Xue (2015) found that investor sentiment has nonlinear effects on stock returns, and such effects are heterogeneous across different subgroups of stocks.

To address the problem of non-additive heterogeneity in the data, recent econometrics literature has studied panel regression models with grouped heterogeneity; see Su, Shi, and Phillips (2016), Vogt and Linton (2017), Miao, Su, and Wang (2020), among many others. There are two main features in the models: first, every individual is assumed to have a unique unobserved group membership; second, the functional relationship between the dependent and independent variables is homogeneous within the same group but heterogeneous across different groups. By introducing the grouped heterogeneity, such models can reach a good balance between flexibility and parsimony compared with panel regression models with fixed effects and classical random coefficients panel models. To our best knowledge, the current literature in this area mainly focuses on linear panel regression models, which has motivated us to fill such a gap by considering a nonparametric counterpart.

In this paper, we propose a nonparametric additive panel regression model with grouped heterogeneity, which can simultaneously consider both nonlinear effects of explanatory variables and non-additive heterogeneity. Additive regression models have a wide variety of

applications in economics, statistics and many other disciplines; see Sperlich, Tjøstheim, and Yang (2002), Profit and Sperlich (2004), Mammen, Støve, and Tjøstheim (2009) and Huang, Horowitz, and Wei (2010), etc. Therefore, this paper naturally contributes to the literature of additive regression models by incorporating grouped heterogeneity into consideration. It is worth noting that Vogt and Linton (2017) and Vogt and Linton (2020) also considered nonparametric panel regression models with grouped heterogeneity. The clustering methods developed in these two papers suffer from the curse of dimensionality. Also, their approach can not be easily generalized to additive regression models.

To estimate the proposed model, we adopt a sieve-approximation-based penalized estimation method, which can identify the latent group structure and estimate parameters of interest in a single step. Our estimation method evolves from the so-called *Classifier*-Lasso estimation method for panel regression models that was first proposed in Su, Shi, and Phillips (2016). Su, Wang, and Jin (2019) applied a similar sieve-approximation-based estimation method to estimate time-varying coefficients panel models. However, the time-varying coefficients considered in Su, Wang, and Jin (2019) are nonrandom; thus, the asymptotic properties derived in their paper do not directly apply to the nonparametric additive regression models considered here. More importantly, unlike previous literature on the *Classifier*-Lasso estimation method, which defines the group structure based on all the coefficients, we take a different approach by considering the subgroup structure of each additive component. This refinement allows us to handle models with a relatively large number of groups since it is the product of group numbers of each nonparametric component. In practice, these group numbers are usually unknown *ex ante* and have to be estimated from the observed data, so we further develop a BIC-type information criterion that can consistently determine group numbers for the model. We establish the convergence rate of the nonparametric components' estimators and their linear functionals' asymptotic normality under some regularity conditions. We also demonstrate the finite sample performance of the estimation method and the BIC-type information criterion through Monte Carlo simulations. The results show

that both perform well in general.

We illustrate the usefulness of the proposed model and estimation method by applying them to study the consumer demand for cigarettes in the United States using panel data of 46 states from 1963 to 1992. We find that group heterogeneity exists in the effect of the retail price of a pack of cigarettes on cigarette sales. More specifically, all 46 states can be classified into two groups according to their price elasticity of demand for cigarettes. There are 28 states in the first group and 18 states in the second group, and those in the first group are, on average, more sensitive to price. However, we do not find evidence indicating there exists grouped heterogeneity in the effect of real per capita disposable income on cigarette sales.

The rest of the paper is organized as follows. We introduce the nonparametric additive panel regression model with grouped heterogeneity in Section 2.2. In Section 2.3, we describe the proposed sieve-approximation-based *Classifier*-Lasso estimation method. Section 2.4 establishes the asymptotic properties of the proposed estimator. Section 2.5 reports the Monte Carlo simulation results. An empirical application is presented in Section 2.6. Finally, Section 2.7 concludes.

Notation: For any matrix A , we denote $\|A\|_F = (\text{tr}(AA'))^{1/2}$ as its Frobenius norm, A' as its transpose and A^{-1} as its Moore-Penrose generalized inverse. If A is also a squared matrix, we denote $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ as its largest and smallest eigenvalues, $\|A\|_S = (\lambda_{\max}(AA'))$ as its spectral norm. The L_q -norm of a p -dimensional vector v is denoted by $\|v\|_q$, where $\|v\|_q \equiv (\sum_{i=1}^p |v_i|^q)^{1/q}$ when $1 \leq q < \infty$ and $\|v\|_q \equiv \max_{i=1, \dots, p} |v_i|$ when $q = \infty$. For a vector-valued function $f(\cdot)$ defined on $[0, 1]$, we let $\|f\|_2$ to be its L_2 -norm, i.e., $\|f\|_2 = (\int_0^1 \|f(x)\|^2 dx)^{1/2}$. For a set G , its cardinality is denoted by $|G|$. For a set $[N]$, we define $[N] \equiv \{1, 2, \dots, N\}$. For functions $f(n)$ and $g(n)$, we let $f(n) \gtrsim g(n)$ and $g(n) \lesssim f(n)$ mean $f(n) \geq cg(n)$ for a generic constant $c > 0$, $f(n) \asymp g(n)$ denote both $f(n) \gtrsim g(n)$ and $f(n) \lesssim g(n)$ hold. We let $(N, T) \rightarrow \infty$ denote N and T diverging to infinity joint, \xrightarrow{P} convergence in probability, \xrightarrow{D} convergence in probability. As a general rule for this paper,

we write c as positive generic constants that are independent of n in different places.

2.2 Model

In this section, we introduce the nonparametric additive panel regression model with grouped heterogeneity. Suppose researchers observe panel data of N individuals for T periods, i.e., $\{\{y_{it}, x'_{it}\}_{i=1}^N\}_{t=1}^T$. The primary interest here is to study the effect of the explanatory variables x on the explained variable y . We assume y_{it} is generated according to the following econometric model:

$$y_{it} = \mu_i + \sum_{j=1}^p h_{i,j}(x_{it,j}) + u_{it}, \quad u_{it} = \sigma_i(x_{it})\varepsilon_{it}, \quad (2.1)$$

for $i = 1, \dots, N$ and $t = 1, \dots, T$, where $x_{it} = (x_{it,1}, \dots, x_{it,p})'$ is a $p \times 1$ vector of explanatory variables, μ_i denotes the unobserved individual fixed effect which can be correlated with x_{it} , ε_{it} is an error term which has mean zero and variance one and is uncorrelated with x_{it} and u_{it} is an error term with mean zero and variance $\sigma_i^2(x_{it})$ conditional on x_{it} . In addition, $h_{i,j}(x)$ is a smooth function defined on a compact support \mathcal{X}_j for $j = 1, \dots, p$, and we assume $\mathcal{X}_j = [0, 1]$ without loss of generality. Throughout this paper, we let $h_{i,j}^0(x)$ denote the true parameter of interest to be estimated.

To capture the non-additive unobserved heterogeneity that can affect the functional relationship directly, we impose the following group structure on the nonparametric components

$\{h_{i,1}^0, \dots, h_{i,p}^0\}_{i=1}^N$:

$$h_{i,j}^0(x) = \sum_{k=1}^{K_j^0} f_{k,j}^0(x) \mathbf{1}\{i \in G_{k,j}^0\} \quad \text{for any } x \in [0, 1] \text{ and } j = 1, \dots, p, \quad (2.2)$$

where $f_{k,j}^0(x)$ is some smooth function defined on $[0, 1]$, $G_{k,j}^0$ denote the k -th group of the nonparametric function of the j -th explanatory variable $x_{it,j}$, K_j^0 is the total number of groups of $h_{i,j}^0(x)$. We assume $\{G_{k,j}^0\}_{k=1}^{K_j^0}$ are mutually exclusive, i.e., $\cup_{k=1}^{K_j^0} G_{k,j}^0 = \{1, 2, \dots, N\}$ for all

$1 \leq j \leq p$, and $G_{m,j}^0 \cap G_{n,j}^0 = \emptyset$ if $m \neq n$. Furthermore, we let $N_{k,j}$ denote the cardinality of the set $G_{k,j}^0$, i.e., $N_{k,j} = |G_{k,j}^0|$, and we have $\sum_{k=1}^{K_j^0} N_{k,j} = N$ by definition. Finally, we let $f_j = (f_{1,j}, \dots, f_{K_j^0,j})'$ for $j = 1, \dots, p$, which is the vector of the j -th infinite-dimensional parameters to be estimated. Following the convention in the literature, we assume that the group memberships do not vary across different time periods.

Based on the above setup, our goals include (1) estimating $\{h_{i,1}(x), \dots, h_{i,p}(x)\}$ for $i = 1, \dots, N$; (2) estimating the group-level parameters $\{f_{1,j}(x), \dots, f_{K_j^0,j}(x)\}$ for $j = 1, \dots, p$; (3) identifying the group memberships $\{G_{1,j}^0, \dots, G_{K_j^0,j}^0\}$ for $j = 1, \dots, p$. It is worth noting that the nonparametric additive panel regression model given by equations 2.1 and 2.2 is fairly general since it takes account of both the additive heterogeneity represented by the individual fixed effect as well as the non-additive heterogeneity that directly affect the functional relationships. Such a model can be regarded as a natural extension of the linear panel regression models with grouped heterogeneity. Because of the additive structure, we can avoid the curse of dimensionality and still capture the nonlinearity in the marginal effects of explanatory variables. Therefore, our model can become an appealing choice for empirical studies in economics, sociology, and many other fields.

2.3 Estimation

In this section, we propose the sieve-approximation-based *Classifier*-Lasso estimation method. This section includes two subsections. In Subsection 2.3.1, we discuss the sieve approximation for nonparametric functions $h_{i,j}(x)$ and $f_{k,j}(x)$ for all $i = 1, \dots, N$, $j = 1, \dots, p$ and $k = 1, \dots, K_j$. In Subsection 2.3.2, we introduce the optimization problem and the related estimators.

2.3.1 Sieve Approximation

Since the infinite-dimensional parameters are unknown functions, we first approximate them using the sieve approximation method; see Ai and Chen (2003) and Chen (2007) for more details on sieve estimation. In this paper, we use the B-spline polynomials of order κ (or degree $\kappa - 1$) to form basis functions on $[0, 1]$ because it is well-known that the B-splines have good properties and are computationally easy.

We first use the B-spline basis functions to approximate $h_{i,j}$ and $f_{k,j}$, for $k = 1, \dots, K_j^0$, $j = 1, \dots, p$ and $i = 1, \dots, N$. We assume that these functions are contained in the Hölder space, which is defined as follows. We consider the Hölder space $\Lambda^r([0, 1])$ of order $r > 0$. Let \underline{r} denote the largest integer satisfying $\underline{r} < r$. The Hölder space is a space of functions $f : [0, 1] \rightarrow \mathcal{R}$ such that the first \underline{r} derivatives are bounded, and the \underline{r} -th derivatives are Hölder continuous with the exponent $r - \underline{r} \in (0, 1]$. The Hölder space becomes a Banach space when endowed with the Hölder norm:

$$\|f\|_{\Lambda^r} = \sup_x |f(x)| + \sup_{x \neq x'} \frac{|\nabla^{\underline{r}} f(x) - \nabla^{\underline{r}} f(x')|}{(\|x - x'\|_F)^{r - \underline{r}}} < \infty,$$

where for any nonnegative scalar a ,

$$\nabla^{\underline{r}} f(x) = \frac{\partial^{\underline{r}}}{\partial x^{\underline{r}}} f(x).$$

A Hölder ball with radius c is defined as $\Lambda_c^r([0, 1]) \equiv \{f \in \Lambda^r([0, 1]) : \|f\|_{\Lambda^r} \leq c < \infty\}$. It is known that functions in $\Lambda_c^r([0, 1])$ could be approximated sufficiently well by the B-spline polynomials of order $\kappa \geq \underline{r} + 1$. Let $B^J(x_{it,j})$ denote $J \times 1$ basis functions, then we could approximate $h_{i,j}(x_{it,j})$ and $f_{k,j}(x_{it,j})$ by $B^J(x_{it,j})' \gamma_{i,j}$ and $B^J(x_{it,j})' \pi_{k,j}$, respectively, where

$\gamma_{i,j}$ and $\pi_{k,j}$ are $J \times 1$ vectors:

$$\begin{aligned} h_{i,j}(x_{it,j}) &= B^J(x_{it,j})' \gamma_{i,j} + \delta_{h_{i,j}}(x_{it,j}), & i = 1, \dots, N, \quad j = 1, \dots, p, \\ f_{k,j}(x_{it,j}) &= B^J(x_{it,j})' \pi_{k,j} + \delta_{f_{k,j}}(x_{it,j}), & k = 1, \dots, K_j^0, \quad j = 1, \dots, p, \end{aligned}$$

where $\delta_{h_{i,j}}(x_{it,j})$ and $\delta_{f_{k,j}}(x_{it,j})$ are corresponding approximation errors.

Define $z_{it,j} \equiv \sqrt{J} B^J(x_{it,j})$ and $\theta_{i,j} \equiv \frac{1}{\sqrt{J}} \gamma_{i,j}$, $i = 1, \dots, N$, then equation 2.1 could be expressed as

$$y_{it} = \mu_i + \sum_{j=1}^p z'_{it,j} \theta_{i,j} + e_{it} \quad (2.3)$$

where $\frac{1}{\sqrt{J}}$ is the normalization term and $e_{it} = u_{it} + \sum_{j=1}^p \delta_{h_{i,j}}(x_{it,j})$.

At the same time, we let $\eta_{k,j} = \frac{1}{\sqrt{J}} \pi_{k,j}$, then equation 2.2 implies

$$\theta_{i,j}^0 = \sum_{k=1}^{K_j^0} \eta_{k,j}^0 \mathbf{1}\{i \in G_{k,j}^0\}. \quad (2.4)$$

Thus we have constructed the sieve approximations for $h_{i,j}(x)$ and $f_{k,j}(x)$, respectively.

2.3.2 Penalized Estimation of h and f

Since our main interest is to quantify the effect of different explanatory variables on the explained variable, we use standard transformation to eliminate the individual fixed effect μ_i and thus get rid of the potential incidental parameter problem caused by the individual fixed effects. We take the deviation from the mean across individuals, which gives the following equation

$$y_{it} - \bar{y}_i = \sum_{j=1}^P (z_{it,j} - \bar{z}_{i,j})' \theta_{i,j} + e_{it} - \bar{e}_i, \quad (2.5)$$

where $\bar{y}_i = \frac{1}{T} \sum_{t=1}^T y_{it}$, with similar definitions for $\bar{z}_{i,j}$ and \bar{e}_i .

For the sake of notational simplicity, we further define $\tilde{y}_{it} = y_{it} - \bar{y}_i$ and similarly for $\tilde{z}_{it,j}$,

\tilde{e}_{it} , then equation 2.5 could be written as

$$\tilde{y}_{it} = \sum_{j=1}^p \tilde{z}'_{it,j} \theta_{i,j} + \tilde{e}_{it}. \quad (2.6)$$

At this moment, we assume that K_j^0 is known in the estimation procedure. Later we will discuss how to use a BIC-type criterion to consistently estimate K_j^0 , for $j = 1, \dots, p$. Recall our goals are to estimate both $h_{i,j}(x)$, $f_{k,j}(x)$ and identify the latent group structure. To achieve these goals, we propose to minimize the following criterion function:

$$Q_{NT,\lambda}(\theta, \eta) = Q_{NT}(\theta) + \frac{\lambda}{N} \sum_{i=1}^N \sum_{j=1}^p \prod_{k=1}^{K_j^0} \|\theta_{i,j} - \eta_{k,j}\|_F, \quad (2.7)$$

where

$$Q_{NT}(\theta) = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \left(\tilde{y}_{it} - \sum_{j=1}^p \tilde{z}'_{it,j} \theta_{i,j} \right)^2. \quad (2.8)$$

In equations 2.7 and 2.8, we let $\theta = (\theta_1, \dots, \theta_N)$, in which $\theta_i = (\theta'_{i,1}, \dots, \theta'_{i,p})'$, and $\eta = (\eta'_1, \dots, \eta'_p)'$, in which $\eta_j = (\eta'_{1,j}, \dots, \eta'_{K_j^0,j})'$. λ is some positive tuning parameter which depends on N and T . The additional penalty is used to shrink the individual parameters $\theta_{i,j}$, $i = 1, \dots, N$ to a particular unknown group-specific parameters $\eta_{k,j}$ for some $k \in \{1, \dots, K_j^0\}$ while at the same time to classify individuals into a priori unknown groups.

Let $\hat{\theta}$ and $\hat{\eta}$ be the solution to the minimization problem given by equation 2.7. Then $\{\hat{h}_{i,1}(x), \dots, \hat{h}_{i,p}(x)\}$ for $i = 1, \dots, N$, and $\{\hat{f}_{1,j}(x), \dots, \hat{f}_{K_j^0,j}(x)\}$ for $j = 1, \dots, p$ are given by

$$\begin{aligned} \hat{h}_{i,j}(x) &= \sqrt{J} B^J(x)' \hat{\theta}_{i,j} & \text{for } j = 1, \dots, p, \\ \hat{f}_{k,j}(x) &= \sqrt{J} B^J(x)' \hat{\eta}_{k,j} & \text{for } k = 1, \dots, K_j^0, \quad j = 1, \dots, p. \end{aligned}$$

The latent group structure is identified using the following rule: $i \in \hat{G}_{k,j}$ if $\hat{h}_{i,j} = \hat{f}_{k,j}$. As pointed out in Su, Shi, and Phillips (2016), all individuals will be classified into certain groups asymptotically. However, in finite samples, it may be the case that some individuals

are left as unclassified if the tuning parameter is relatively small. When such situation appears, we can use another decision rule to determine the latent group structure: $i \in \hat{G}_{k,j}$ if $\|\hat{h}_{i,j} - \hat{f}_{k,j}\|_F \leq \|\hat{h}_{i,j} - \hat{f}_{l,j}\|_F$, for all $l = 1, \dots, K_j$.

2.4 Asymptotic Properties

In this section, we establish the asymptotic properties for the estimators proposed in Section 2.3. This section include four subsections. They are organized as follows: in Subsection 2.4.1, we characterize the preliminary convergence rates for individual coefficients $\hat{\theta}_{i,j}$, for $i = 1, \dots, N$ and $j = 1, \dots, p$ and the group-specific parameters $\hat{\eta}_{k,j}$, for $j = 1, \dots, p$ and $k = 1, \dots, K_j^0$. Subsection 2.4.2 presents the results of classification consistency. After that, Subsection 2.4.3 reports the asymptotic distribution of group-specific parameters $f_{k,j}$, for $j = 1, \dots, p$ and $k = 1, \dots, K_j^0$. Subsection 2.4.4 discusses how to determine the number of groups.

2.4.1 Preliminary Rates of Convergence

We first give the necessary assumptions for establishing the convergence rate of $\hat{\theta}$ and $\hat{\eta}$. Define $x_{it} \equiv (x_{it,1}, \dots, x_{it,p})'$ and $z_{it} \equiv (z'_{it,1}, \dots, z'_{it,p})'$.

Assumption 2.1. (i) For each $i = 1, \dots, N$, $\{x_{it}, \varepsilon_{it} : t \geq 1\}$ is stationary strong mixing with mixing coefficient $\alpha_i(j)$. $\alpha(j) \equiv \max_{1 \leq i \leq N} \alpha_i(j)$ satisfies $\alpha(j) \leq c_\alpha \exp(-\rho j)$ for some $0 < c_\alpha < \infty$, $0 < \rho < \infty$. $\{x_{it}, \varepsilon_{it}\}$ are independent across i .

(ii) There exists positive \bar{c} such that $\max_{i,t} \|u_{it}\|_F^q < \bar{c} < \infty$ for some $q > 6$.

(iii) For the nonparametric functions $\{f_{1,j}^0, \dots, f_{K_j^0,j}^0\}_{j=1}^p$, we have

(i) $\mathbf{E}[f_{k,j}^0(x_{it,j})] = 0$, for $j = 1, \dots, p$ and $k = 1, \dots, K_j^0$.

(ii) $f_{k,j}^0 \in \mathcal{F} = \Lambda_c^r([0, 1])$ with $r > 0$, for $j = 1, \dots, p$ and $k = 1, \dots, K_j^0$.

(iii) $\forall i \in \{1, \dots, N\}$, let $f_{it,j}(x)$ denote the marginal density function of $\{x_{it,j}\}$, we have $f_{it,j}(x) = f_{i,j}(x)$ for all $1 \leq t \leq T$ and $x \in [0, 1]$. Furthermore, there exist positive constants \underline{c} and \bar{c} such that

$$0 < \underline{c} < \min_{1 \leq i \leq N} \min_{1 \leq j \leq p} \inf_{x \in [0,1]} \{f_{i,j}(x)\} \leq \max_{1 \leq i \leq N} \max_{1 \leq j \leq p} \sup_{x \in [0,1]} \{f_{i,j}(x)\} < \bar{c} < \infty.$$

(iv) There exist $\underline{c} > 0$ such that for any $j = 1, \dots, p$,

$$\min_{1 \leq m \neq n \leq K_j^0} \|f_{m,j}^0 - f_{n,j}^0\|_2^2 > \underline{c}.$$

(v) There exist positive constants \underline{c} and \bar{c} such that

$$0 < \underline{c} < \min_{1 \leq i \leq N} \mu_{\min}(\text{Var}(z_{it})) \leq \max_{1 \leq i \leq N} \mu_{\max}(\text{Var}(z_{it})) < \bar{c} < \infty.$$

(vi) $\frac{N_{k,j}}{N} \rightarrow \tau_{k,j}$ for $j = 1, \dots, p$ and $k = 1, \dots, K_j^0$ as $N \rightarrow \infty$. There exists positive constants \underline{c} and \bar{c} such that

$$0 < \underline{c} < \min_{1 \leq j \leq p} \min_{1 \leq k \leq K_j^0} \{\tau_{k,j}\} \leq \max_{0 \leq j \leq p} \max_{1 \leq k \leq K_j^0} \{\tau_{k,j}\} < \bar{c} < 1$$

Assumption 2.1(i) implies that the strong mixing coefficients $\alpha(l)$ decay exponentially fast to 0 as $l \rightarrow \infty$ uniformly. Similar conditions are made in Su, Shi, and Phillips (2016), Su, Wang, and Jin (2019), Vogt and Linton (2017), etc. For more discussions on this, we refer readers to Su, Wang, and Jin (2019). Assumption 2.1(ii) imposes moment restrictions for u_{it} .

Assumption 2.1(iii) imposes restrictions on the nonparametric functions. The first part is a harmless normalization. The second one is the smooth condition which ensures we can approximate any function $f \in \mathcal{F}$ sufficiently well using the tensor-product of univariate

B-splines. By results from the approximation theory, there exists $\pi_{k,j} \in \mathcal{R}^J$ such that

$$\sup_{x \in [0,1]} \|f_{k,j}(x) - B^{J'} \pi_{k,j}\|_{\infty} = O(J^{-r})$$

Similarly, for each individual, there exists $\gamma_{i,j}$ such that

$$\sup_{x \in [0,1]} \|h_{i,j}(x) - B^{J'} \gamma_{i,j}\|_{\infty} = O(J^{-r}).$$

Then, after controlling for the approximation error, the difference between $f_{k,j}(x)$ and $h_{i,j}(x)$ is reflected by the difference between $\pi_{k,j}$ and $\gamma_{i,j}$. The third part is also assumed in Vogt and Linton (2017). First, this condition makes the functions $h_{i,j}(x_{it})$ comparable across individuals. Second, it guarantees that $h_{i,j}(x_{it})$ could be estimated uniformly well.

Assumption 2.1(iv) specifies that the group-specific parameters are well separated from each other. At the same time, it also implies that the group-specific vectors π and η are well separated. For $1 \leq m \neq n \leq K_j$, let's consider $\|f_{m,j}^0 - f_{n,j}^0\|_2$ first. Notice that

$$\begin{aligned} & \|f_{m,j}^0 - f_{n,j}^0\|_2 \\ & \leq \|f_{m,j}^0 - B^{J'} \pi_{m,j}\|_2 + \|f_{n,j}^0 - B^{J'} \pi_{n,j}\|_2 + \left\| \sqrt{J} B^{J'} \left(\frac{1}{\sqrt{J}} (\pi_{m,j} - \pi_{n,j}) \right) \right\|_2 \\ & = O(J^{-r}) + \left\{ \left(\frac{1}{\sqrt{J}} (\pi_{m,j} - \pi_{n,j}) \right)' \int_{[0,1]} J B^J(x) B^J(x)' dx \left(\frac{1}{\sqrt{J}} (\pi_{m,j} - \pi_{n,j}) \right) \right\}^{\frac{1}{2}} \\ & \asymp \left\| \frac{1}{\sqrt{J}} (\pi_{m,j} - \pi_{n,j}) \right\|_F, \end{aligned}$$

where the last equation holds because the eigenvalues of $\int_{[0,1]^d} J B^J(x) B^J(x)' dx$ are bounded above and away from zero.

Similarly, we have

$$\begin{aligned}
& \left\| \frac{1}{\sqrt{J}}(\pi_{m,j} - \pi_{n,j}) \right\|_F \\
& \asymp \left\| \sqrt{J} B^{J'} \left(\frac{1}{\sqrt{J}}(\pi_{m,j} - \pi_{n,j}) \right) \right\|_2 \\
& \leq \left\| f_{m,j}^0 - f_{n,j}^0 \right\|_2 + \left\| f_{m,j}^0 - B^{J'} \pi_{m,j} \right\|_2 + \left\| f_{n,j}^0 - B^{J'} \pi_{n,j} \right\|_2 \\
& = \left\| f_{m,j}^0 - f_{n,j}^0 \right\|_2 + O(J^{-r}) \\
& \asymp \left\| f_{m,j}^0 - f_{n,j}^0 \right\|_2
\end{aligned}$$

Therefore, we have $\left\| f_{m,j}^0 - f_{n,j}^0 \right\|_2^2 \asymp \left\| \frac{1}{\sqrt{J}}(\pi_{m,j} - \pi_{n,j}) \right\|_F^2 = \left\| \eta_{m,j}^0 - \eta_{n,j}^0 \right\|_F^2$. In a similar fashion, we can get

$$\left\| h_{i,j} - f_{k,j} \right\|_2^2 \asymp \left\| \theta_{i,j} - \eta_{k,j} \right\|_F^2.$$

if $i \notin G_{k,j}^0$. This result guarantees that the penalty item in the criterion function 2.7 could shrink the individual coefficients to some group-specific parameters.

Assumption 2.1(v) is a standard identification condition for sieve estimation. As demonstrated in Section 2.3.2, we take the demean approach to get rid of the individual fixed effects, which consequently requires that $\mathbf{E}[\tilde{z}_{it} \tilde{z}'_{it}]$ is positive definite to identify the coefficients. Then notice that the corresponding population value is $\text{Var}(z_{it})$. Assumption 2.1(vi) is commonly assumed in the classification literature, which implies that each group would include an asymptotically non-negligible number of individuals.

Assumption 2.2. *As $(N, T) \rightarrow \infty$, we have $\lambda \rightarrow 0$, $J \rightarrow \infty$, $J^{\frac{3}{2}}(\ln T)^3 T^{-1} \rightarrow 0$ and $N^2 T^{1-\frac{q}{2}} \rightarrow 0$.*

Assumption 2.2 specifies several restrictions on J , N and T . Let's first focus on the first part of the condition, i.e., $J^{\frac{3}{2}}(\ln T)^3 T^{-1} \rightarrow 0$. This condition is comparable to the Assumption 2 in Newey (1997) for independent observations. The last condition requires that T cannot increase too slow compared with N . The intuition is clear: as T grows,

more information of each individual is revealed, making it easier to identify the latent group structures. The q is the moment restriction we make in Assumption 2.1(ii), which is set to be larger than 6 to allow that N and T increase at the same rate.

We are now ready to establish the preliminary convergence rates for $\hat{\theta}$ and $\hat{\eta}$, which are given in Theorem 2.1.

Theorem 2.1. *Suppose Assumption 2.1, 2.2 hold, then*

- (i) $\|\hat{\theta}_i - \theta_i^0\|_F = O_p(J^{-r} + J^{\frac{1}{2}}T^{-\frac{1}{2}} + \lambda)$ and $\|\hat{\theta}_{i,j} - \theta_{i,j}^0\|_F = O_p(J^{-r} + J^{\frac{1}{2}}T^{-\frac{1}{2}} + \lambda)$ for $i = 1, 2, \dots, N, j = 1, \dots, p$.
- (ii) $\frac{1}{N} \sum_{i=1}^N \|\hat{\theta}_i - \theta_i^0\|_F^2 = O_p(J^{-2r} + JT^{-1})$ and $\frac{1}{N} \sum_{i=1}^N \|\hat{\theta}_{i,j} - \theta_{i,j}^0\|_F^2 = O_p(J^{-2r} + JT^{-1})$ for $j = 1, \dots, p$.
- (iii) $\|\hat{\eta}_{(k),j} - \eta_{k,j}^0\|_F = O_p(J^{-r} + J^{\frac{1}{2}}T^{-\frac{1}{2}})$, for $k = 1, \dots, K_j^0, j = 1, \dots, p$, where $(\hat{\eta}_{(1),j}, \dots, \hat{\eta}_{(K_j^0),j})$ is a suitable permutation of $(\hat{\eta}_{1,j}, \dots, \hat{\eta}_{K_j^0,j})$ for $j = 1, \dots, p$.

Theorem 2.1(i) and (ii) give the pointwise and mean square convergence rates of $\hat{\theta}_{i,j}$ for $j = 1, \dots, p$. In Theorem 2.1(i), the first term, J^{-r} , comes from the approximation error. The second term, $J^{\frac{1}{2}}T^{-\frac{1}{2}}$, demonstrates the contribution of the interaction between B-splines and the error term. Similar as other Lasso-like estimators, the penalty item is reflected by λ . However, in Theorem 2.1(ii), the penalty term disappears. We direct interested readers to the details in the proof. The convergence rate of $\hat{\eta}_{k,j}$, similarly, does not depend on λ .

By Assumption 2.2 and Theorem 2.1, it is clear that $\hat{\theta}_{i,j}$ and $\hat{\eta}_{(k),j}$ converges in probability to $\theta_{i,j}^0$ and $\eta_{k,j}^0$, respectively. For notational simplicity, we denote $\hat{\eta}_{(k),j}$ as $\hat{\eta}_{k,j}$ and further define

$$\hat{G}_{k,j} = \left\{ i \in \{1, \dots, N\} : \hat{\theta}_{i,j} = \hat{\eta}_{k,j} \right\} \quad \text{for } k = 1, \dots, K_j^0,$$

which denotes the set of individuals whose functions of the j -th explanatory variable are classified into the k -th group, for $1 \leq k \leq K_j^0$.

2.4.2 Classification Consistency

To ensure the group classification's consistency, we need to impose more assumptions, which are given in Assumption 2.3.

Assumption 2.3. *As $(N, T) \rightarrow \infty$, $\lambda T^{\frac{1}{2}} J^{-\frac{1}{2}} (\ln T)^{-3-v} \rightarrow \infty$, $\lambda J^r (\ln T)^{-v} \rightarrow \infty$, $T^{\frac{1}{2}} J^{-\frac{1}{2}} (\ln T)^{-3-v} \rightarrow \infty$ and $\lambda (\ln T)^v \rightarrow 0$ for some $v > 0$.*

Assumption 2.3 imposes restrictions on λ and some further ones on J . Intuitively, we require that λ dominates all other errors of approximation or u_{it} to make sure the penalty term can effectively shrink the individual coefficients to corresponding group-specific parameters.

Following Su, Shi, and Phillips (2016) and Su, Wang, and Jin (2019), we define

$$\begin{aligned}\hat{E}_{ik,j} &\equiv \{i \notin \hat{G}_{k,j} | i \in G_{k,j}^0\} \\ \hat{F}_{ik,j} &\equiv \{i \notin G_{k,j}^0 | i \in \hat{G}_{k,j}\}\end{aligned}$$

where $i = 1, \dots, N$, $j = 1, \dots, p$ and $k = 1, \dots, K_j^0$. We let $\hat{E}_{k,j} = \cup_{i \in G_{k,j}^0} \hat{E}_{ik,j}$, $\hat{F}_{k,j} = \cup_{i \in \hat{G}_{k,j}} \hat{F}_{ik,j}$. Here $\hat{E}_{k,j}$ denotes the event of classifying individuals that belong to $G_{k,j}^0$ into groups other than $\hat{G}_{k,j}$; and $\hat{F}_{k,j}$ denotes the event of classifying individuals who don't belong to $G_{k,j}^0$ into $\hat{G}_{k,j}$. These two events mimic the Type I and Type II errors in hypothesis testing literature, respectively.

The following theorem establishes the consistency of the group membership estimator.

Theorem 2.2. *Suppose Assumption 2.1, 2.2 and 2.3 hold, then*

- (i) $P(\cup_{j=1}^p \cup_{k=1}^{K_j^0} \hat{E}_{k,j}) \leq \sum_{j=1}^p \sum_{k=1}^{K_j^0} P(\hat{E}_{k,j}) \rightarrow 0$ as $(N, T) \rightarrow \infty$.
- (ii) $P(\cup_{j=1}^p \cup_{k=1}^{K_j^0} \hat{F}_{k,j}) \leq \sum_{j=1}^p \sum_{k=1}^{K_j^0} P(\hat{F}_{k,j}) \rightarrow 0$ as $(N, T) \rightarrow \infty$.

Theorem 2.2 guarantees that with probability approaching 1, we can correctly classify individuals in the same group, say $G_{k,j}^0$, into one group $\hat{G}_{k,j}$, and those classified into the same group, $\hat{G}_{k,j}$, belong to one correct group $G_{k,j}^0$ for $j = 1, \dots, p$ and $k = 1, \dots, K_j^0$.

2.4.3 The Oracle Property and Asymptotic Distributions

As mentioned previously, the *Classifier*-lasso estimation method can simultaneously accomplish two tasks: to classify individuals into different groups and to estimate $\theta_{i,j}$, for $i = 1, \dots, N$ and $j = 1, \dots, p$, and $\eta_{k,j}$, for $k = 1, \dots, K_j^0$ and $j = 1, \dots, p$. Given the estimated coefficients, we might want to conduct statistical inference on the functionals of the nonparametric components. For example, $\hat{f}_{k,j}(x)$, which is constructed by $\hat{f}_{k,j}(x) = \sqrt{J}B^J(x)'\hat{\eta}_{k,j}$.

An alternative strategy would be to implement the post-Lasso approach. Given the estimated groups $\hat{G}_{k,j}$, for $j = 1, \dots, p$ and $k = 1, \dots, K_j^0$, we could conduct a constrained optimization to estimate group-specific parameters. We denote the post-Lasso estimators as $\hat{f}_{\hat{G}_{k,j}}(x)$.

Our goal in this subsection is to show that both the C-lasso estimator and the post-Lasso estimator enjoy the oracle property, i.e., they are asymptotically equivalent to the infeasible estimators as if the group memberships are known *ex ante*. Before we move to the results, more definitions and assumptions are required.

Let $u_i = (u_{i1}, u_{i2}, \dots, u_{iT})$, $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \dots, \varepsilon_{iT})$ and $\text{Var}(u_i|x_i) = \Sigma_i^{\frac{1}{2}}V_i\Sigma_i^{\frac{1}{2}}$, where

$$\begin{aligned}\Sigma_i &= \text{diag}(\sigma_i^2(x_{i1}), \dots, \sigma_i^2(x_{iT})) \\ V_i &= \mathbf{E}[\varepsilon_i\varepsilon_i']\end{aligned}$$

We then formally demonstrate how to construct the oracle estimators. Given the correct group membership $G_{k,j}^0$ for $1 \leq k \leq K_j^0$ and $1 \leq j \leq p$, define $\tilde{z}_{it,G^0} \equiv (\tilde{z}'_{it,G_1^0}, \tilde{z}'_{it,G_2^0}, \dots, \tilde{z}'_{it,G_p^0})'$, where

$$\tilde{z}_{it,G_j^0} \equiv \underbrace{(0'_{J \times 1}, \dots, \overbrace{\tilde{z}'_{it,j}}^{G_{k,j}^0 \text{ th}}, \dots, 0'_{J \times 1})'}_{K_j^0 \text{ vectors}}$$

for $1 \leq j \leq p$. \tilde{z}_{it,G_j^0} is composed of K_j^0 column vectors of length J . All the vector are $0_{J \times 1}$ except for the $G_{k,j}^0$ th, which equals to $\tilde{z}_{it,j}$. Then \tilde{z}_{it,G^0} is a $(J \sum_{j=1}^p K_j^0) \times 1$ vector.

The regression equation is

$$\tilde{y}_{it} = \tilde{z}'_{it,G^0} \eta + \tilde{\varepsilon}_{it}$$

where η is a $(J \sum_{j=1}^p K_j^0) \times 1$ vector. Let $\eta \equiv (\eta'_1, \eta'_2, \dots, \eta'_p)'$, and $\eta_j \equiv (\eta'_{1,j}, \eta'_{2,j}, \dots, \eta'_{K_j^0,j})'$ for $1 \leq j \leq p$.

Denote the estimated η as $\hat{\eta}_{G^0}$ with all the components $\hat{\eta}_{G_{k,j}^0}$. Then construct the corresponding $\hat{f}_{G_{k,j}^0} \equiv z'_{it,j} \hat{\eta}_{G_{k,j}^0}$ for $1 \leq k \leq K_j^0$ and $1 \leq j \leq p$, which is the oracle estimator.

Define

$$V_{G^0} \equiv \left(\mathbf{E}[\tilde{z}_{it,G^0} \tilde{z}'_{it,G^0}] \right)^{-1} \mathbf{E} \left[\tilde{z}_{i.,G^0} \Sigma_i^{1/2} V_i \Sigma_i^{1/2} \tilde{z}'_{i.,G^0} \right] \left(\mathbf{E}[\tilde{z}_{it,G^0} \tilde{z}'_{it,G^0}] \right)^{-1}$$

where $\tilde{z}_{i.,G^0} = (\tilde{z}_{i1,G^0}, \tilde{z}_{i2,G^0}, \dots, \tilde{z}_{iT,G^0})$. We could divide V_{G^0} into different cells $V_{G_{k,j}^0}$ for $1 \leq k \leq K_j^0$ and $1 \leq j \leq p$ according to the true group structure.

Assumption 2.4. (i) For $j = 1, \dots, p$ and $k = 1, \dots, K_j^0$, there exists two positive constants \underline{c}_v and \bar{c}_v such that

$$0 < \underline{c}_v \leq \lim_{N,T \rightarrow \infty} \min_{i \in G_{k,j}^0} \mu_{\min}(V_i) \leq \lim_{N,T \rightarrow \infty} \max_{i \in G_{k,j}^0} \mu_{\max}(V_i) \leq \bar{c}_v \delta_{NT}$$

for some nondecreasing sequence δ_{NT} which satisfies $\delta_{NT} N^{-1} \rightarrow 0$ as $N, T \rightarrow \infty$.

(ii) Let $B_{it,\sigma} \equiv \sqrt{J} B_{it}^J(x_{it}) \sigma_i(x_{it})$. There exist positive constants \underline{c} and \bar{c} such that

$$0 < \underline{c} < \min_{1 \leq i \leq N} \mu_{\min}(\text{Var}(B_{it,\sigma})) \leq \max_{1 \leq i \leq N} \mu_{\max}(\text{Var}(B_{it,\sigma})) < \bar{c} < \infty$$

Assumptions 2.4 is analogous to Assumption A.3 in Su, Wang, and Jin (2019). Assumption 2.4(i) imposes restrictions on the covariance matrix of ε_i . Assumption 2.4(ii) assures that the eigenvalues of the interactive items of z_{it} and the error term are bounded above and away from zero uniformly.

Assumption 2.5. As $(N, T) \rightarrow \infty$, $NTJ^{-2r} \rightarrow 0$.

Assumption 2.5 is used to establish the pointwise convergence rate of the group-specific infinite-dimensional estimators $\hat{f}_{k,j}(x)$ and $\hat{f}_{\hat{G}_{k,j}}(x)$. The following Theorem 2.3 establishes the asymptotic distribution of the estimated functional of $f_{k,j}$.

Theorem 2.3. Suppose Assumption 2.1, 2.2, 2.3, 2.4 and 2.5 hold. Then for any $j \in \{1, \dots, p\}$, $k \in \{1, \dots, K_j^0\}$,

(i)

$$\sqrt{N_{k,j}T/JV_{k,j,B}^{-\frac{1}{2}}} \left(\hat{f}_{k,j}(x) - f_{k,j}^0(x) \right) \xrightarrow{D} N(0, 1)$$

(ii)

$$\sqrt{N_{k,j}T/JV_{k,j,B}^{-\frac{1}{2}}} \left(\hat{f}_{\hat{G}_{k,j}}(x) - f_{k,j}^0(x) \right) \xrightarrow{D} N(0, 1)$$

where

$$V_{k,j,B} = B^J(x)' V_{G_{k,j}^0} B^J(x)$$

and $V_{G_{k,j}^0}$ is the corresponding cell in V_{G^0} .

Theorems 2.3 indicates that the *Classifier*-lasso and post-Lasso estimators of $f_{k,j}(x)$ are asymptotically equivalent to the infeasible estimators, which are denoted as $f_{G_{k,j}^0}$. Thus both C-Lasso and post-Lasso estimators exhibit oracle properties.

2.4.4 Determination of Number of Groups

In this section, we discuss how to use a BIC-type information criterion to determine the number of groups K_j^0 , $j = 1, \dots, p$. Define $K^0 = (K_1^0, \dots, K_p^0)$. Following the literature, we assume that K_j^0 is bounded above from a finite integer K_{\max} for all $j = 1, \dots, p$. We make the dependence of $\hat{\theta}_{i,j}$ and $\hat{\eta}_{k,j}$ on K and λ explicit by denoting them as $\hat{\theta}_{i,j}(K, \lambda)$ and $\hat{\eta}_{k,j}(K, \lambda)$.

Using the post-Lasso estimator $\hat{\eta}_{\hat{G}}(K, \lambda)$, we could calculate

$$\hat{\sigma}_{\hat{G}(K, \lambda)}^2 = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \left(\tilde{y}_{it} - \tilde{z}'_{it} \hat{\eta}_{\hat{G}}(K, \lambda) \right)^2.$$

Then we choose $K = (K_1, \dots, K_p)$ to minimize the following information criterion

$$\text{IC}(K, \lambda) = \ln \left(\hat{\sigma}_{\hat{G}(K, \lambda)}^2 \right) + \rho_{NT} \cdot pJ \sum_{j=1}^p K_j$$

where ρ_{NT} is the tuning parameter. Let $\hat{K}(\lambda) \equiv \arg \min_{1 \leq K_j \leq K_{\max}, j=1, \dots, p} \text{IC}(K, \lambda)$. We next show that the above information criterion can consistently select the number of groups for each nonparametric component. Let $G_j^{(K)} \equiv \{G_{K,1,j}, \dots, G_{K,K,j}\}$ be any K -partition of $\{1, \dots, N\}$ for variable j , and \mathcal{G}_K a collection of all such partitions for all $1 \leq j \leq p$. Further define

$$\hat{\sigma}_{G^{(K)}}^2 \equiv \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \left(\tilde{y}_{it} - \tilde{z}'_{it} \hat{\eta}_{\hat{G}_{K,k}} \right)^2.$$

We first introduce some assumptions.

Assumption 2.6. As $(N, T) \rightarrow \infty$, $\min_{1 \leq K_j < K_j^0, 1 \leq j \leq p} \inf_{G^{(K)} \in \mathcal{G}_K} \hat{\sigma}_{G^{(K)}}^2 \xrightarrow{P} \underline{\sigma}^2 > \sigma_0^2$, where $\sigma_0^2 = \text{plim}_{(N, T) \rightarrow \infty} \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T u_{it}^2$.

Assumption 2.7. As $(N, T) \rightarrow \infty$, $\rho_{NT}J \rightarrow 0$ and $\rho_{NT}NT \rightarrow \infty$.

When to decide the correct number of groups, there are three different situations to consider: $K_j < K_j^0$, $K_j = K_j^0$, and $K_j > K_j^0$ for each $1 \leq j \leq p$, corresponding to under-fitted, correct, and over-fitted models, respectively. Assumption 2.6 is used to guarantee that in the under-fitted models, the first term in the IC criterion is always larger than in the correct model. It implies that we will not choose under-fitted models with probability approaching one as long as the second term in the IC criterion is dominated, which is ensured by Assumption 2.7. Similarly, Assumption 2.7 is a condition to ensure that the over-fitted

models will not be picked out with probability approaching one. The following theorem formally summarizes such intuition.

Theorem 2.4. *Suppose Assumptions 2.1, 2.2, 2.3, 2.4, 2.5, 2.6 and 2.7 hold. Then $P(\hat{K}(\lambda) = K^0) \rightarrow 1$ as $(N, T) \rightarrow \infty$.*

Theorem 2.4 shows that the IC criterion can consistently determine the correct number of groups for each nonparametric component. However, in finite samples, we suggest that readers use it with caution. There is always some probability, even though quite small that a misspecified model is selected. Thus we recommend that readers try different numbers of groups, compare the results, and discuss possible implications in empirical studies.

2.5 Simulation

In this section, we investigate the finite sample performance of the sieve-approximation-based *Classifier*-Lasso estimation method for nonparametric additive panel regression models.

2.5.1 Data Generating Process

We consider three different data generating processes (DGPs). In all three DGPs, we let $x_{it,s}$ follow a standard normal distribution across both i and t for $s = 1, \dots, p$, μ_i follows a standard normal distribution for all individuals i , and $u_{it} \sim_{i.i.d.} N(0, 1)$ across both i and t . For each DGP, we consider four different combinations of (N, T) to investigate their influence on the estimates. These four combinations are: (1) $(N, T) = (100, 40)$; (2) $(N, T) = (100, 80)$; (3) $(N, T) = (200, 80)$; (4) $(N, T) = (200, 160)$, which analogize various data structures in the real-world data sets. The three DGPs are detailed as follows.

DGP 1 In this data generating process, we assume y_{it} is given by the following specification

$$y_{it} = \mu_i + h_{i,1}(x_{it,1}) + h_{i,2}(x_{it,2}) + u_{it},$$

where

$$h_{i,1}(x) = \begin{cases} x - \frac{1}{2} & \text{if } i \in G_{1,1}^0, \\ 3x^2 - 1 & \text{if } i \in G_{2,1}^0, \end{cases}$$

and

$$h_{i,2}(x) = \begin{cases} \sin(2\pi x) & \text{if } i \in G_{1,2}^0, \\ \sin(4\pi x) & \text{if } i \in G_{2,2}^0. \end{cases}$$

Here $G_{k,j}^0$ denotes the set of individuals such that the individual-specific function $h_{i,j}$ is in the k -th group of the function of $x_{it,j}$. Furthermore, we assume $G_{1,1}^0 = \{1, 2, \dots, \frac{1}{2}N\}$ and $G_{1,2}^0 = \{1, 2, \dots, \frac{1}{2}N\}$.

DGP 2 In this data generating process, we assume y_{it} is given by the following specification

$$y_{it} = \mu_i + h_{i,1}(x_{it,1}) + h_{i,2}(x_{it,2}) + h_{i,3}(x_{it,3}) + u_{it},$$

where

$$h_{i,1}(x) = \begin{cases} \sin(2\pi x) & \text{if } i \in G_{1,1}^0, \\ \sin(4\pi x) & \text{if } i \in G_{2,1}^0, \end{cases}$$

and

$$h_{i,2}(x) = \begin{cases} \cos(2\pi x) & \text{if } i \in G_{1,2}^0, \\ \cos(4\pi x) & \text{if } i \in G_{2,2}^0, \end{cases}$$

and

$$h_{i,3}(x) = \begin{cases} x - \frac{1}{2} & \text{if } i \in G_{1,3}^0, \\ 3x^2 - 1 & \text{if } i \in G_{2,3}^0. \end{cases}$$

Here we let $G_{1,1}^0 = \{1, 2, \dots, \frac{N}{4}\}$, $G_{1,2}^0 = \{1, 2, \dots, \frac{N}{2}\}$ and $G_{1,3}^0 = \{1, 2, \dots, \frac{3}{4}N\}$.

DGP 3 In this data generating process, we assume y_{it} is given by the following specification

$$y_{it} = \mu_i + h_{i,1}(x_{it,1}) + h_{i,2}(x_{it,2}) + h_{i,3}(x_{it,3}) + u_{it},$$

where

$$h_{i,1}(x) = \begin{cases} \sin(2\pi x) & \text{if } i \in G_{1,1}^0, \\ \sin(4\pi x) & \text{if } i \in G_{2,1}^0, \end{cases}$$

and

$$h_{i,2}(x) = \begin{cases} \cos(2\pi x) & \text{if } i \in G_{1,2}^0, \\ \cos(4\pi x) & \text{if } i \in G_{2,2}^0, \end{cases}$$

and

$$h_{i,3}(x) = \begin{cases} x - \frac{1}{2} & \text{if } i \in G_{1,3}^0, \\ 3x^2 - 1 & \text{if } i \in G_{2,3}^0, \\ x^3 - 3x^2 + \frac{3}{4} & \text{if } i \in G_{3,3}^0. \end{cases}$$

Here we let $G_{1,1}^0 = \{1, 2, \dots, \frac{N}{4}\}$, $G_{1,2}^0 = \{1, 2, \dots, \frac{N}{2}\}$, $G_{1,3}^0 = \{1, 2, \dots, \frac{1}{4}N\}$ and $G_{2,3}^0 = \{\frac{1}{4}N + 1, \dots, \frac{3}{4}N\}$.

As the number of nonparametric functions and the number of groups for each nonparametric component increases from DGP 1 to DGP 3, grouped heterogeneity in each nonparametric component becomes stronger and stronger,

For a fixed DGP and a given combination of (N, T) , we estimate the model using the iterative procedure introduced in Su, Wang, and Jin (2019) and simulate with 100 repetitions. We let the tuning parameter $\lambda = (NT)^{-1/8}$, which satisfies all the related assumptions on λ given in Section 2.4 to ensure the consistency of the estimators. We use the cubic B-splines (B-splines of order 4) for sieve approximation, and we let the number of interior points J_0 to be the integer closest to $(NT)^{\frac{1}{5}}$.

To measure the accuracy of the estimation approach developed in this paper, we report the root mean square errors (RMSE) of both individual-specific and group-specific unknown functions as well as the rate of correct classification for each unknown function. More specifically, for the j -th nonparametric function, the RMSE of the group-specific estimates

are given by

$$RMSE = \frac{1}{R} \sum_{r=1}^R \sqrt{\sum_{k=1}^{K_j^0} \|\hat{h}_{k,j} - h_{k,j}^0\|_2^2},$$

respectively, where R is the number of repetitions which equals 100 in our setting. The correct classification rate for the j -th nonparametric component is given by

$$CC_j = \frac{1}{R} \sum_{r=1}^R \left\{ \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^{K_j^0} \mathbf{1}\{i \in \hat{G}_{k,j}, i \in G_{k,j}^0\} \right\}.$$

We report the RMSE for both C-Lasso and Post-Lasso estimates as well as the oracle estimates. Here the oracle estimates is estimated assuming the group memberships are known.

2.5.2 Simulation Results

Table 2.1, Table 2.2, and Table 2.3 report the simulation results for the group-specific parameters in DGP 1, DGP 2, and DGP 3, respectively. There are several interesting findings. First, we can see that the rate of correct classification (CC Rate) increases when both N and T increase. When $(N, T) = (100, 80)$, the rate of correct classification is larger than 98% in DGP 1 and DGP 2, and when $(N, T) = (200, 160)$, the misclassification error is almost zero in all DGPs, showing that the estimation method has satisfying performance. Second, the correct classification rate is higher in DGP 1 than in DGP 2 and DGP 3 when (N, T) is fixed, which shows that the complexity of the group structure will also affect the finite sample performance of the estimation method. Third, the RMSEs of the C-Lasso estimators are usually larger than the RMSEs of the post-Lasso estimators. In addition, the finite sample performance of the post-Lasso estimators is very close to that of oracle estimators when (N, T) is large, which is consistent with the theoretical justification in Section 2.4 and the simulation findings in Su, Shi, and Phillips (2016) and Su, Wang, and Jin (2019). Based on these findings, we recommend using the post-Lasso estimators in empirical studies.

Table 2.1: Simulation Results for Group-specific Parameters in DGP 1

Function	N	T	CC Rate	RMSE (C-Lasso)	RMSE (Post-Lasso)	RMSE (Oracle)
h_1^0	100	40	84.49%	0.1577	0.1557	0.0893
	100	80	98.71%	0.0708	0.0693	0.0674
	200	80	98.38%	0.0568	0.0530	0.0501
	200	160	99.96%	0.0372	0.0364	0.0364
h_2^0	100	40	94.65%	0.1356	0.1364	0.0965
	100	80	99.81%	0.0724	0.0718	0.0708
	200	80	99.72%	0.0535	0.0520	0.0512
	200	160	100.00%	0.0374	0.0372	0.0372

Table 2.2: Simulation Results for Group-specific Parameters in DGP 2

Function	N	T	CC Rate	RMSE (C-Lasso)	RMSE (Post-Lasso)	RMSE (Oracle)
h_1^0	100	40	96.11%	0.1356	0.1305	0.1088
	100	80	99.87%	0.0809	0.0764	0.0761
	200	80	99.81%	0.0632	0.0588	0.0580
	200	160	100.00%	0.0439	0.0425	0.0425
h_2^0	100	40	90.92%	0.1776	0.1753	0.0948
	100	80	99.76%	0.0750	0.0717	0.0707
	200	80	99.64%	0.0548	0.0514	0.0500
	200	160	100.00%	0.0383	0.0366	0.0366
h_3^0	100	40	74.17%	0.3233	0.2926	0.1023
	100	80	98.64%	0.0948	0.0801	0.0760
	200	80	97.98%	0.0808	0.0629	0.0568
	200	160	99.95%	0.0530	0.0415	0.0414

Table 2.3: Simulation Results for Group-specific Parameters in DGP 3

Function	N	T	CC Rate	RMSE (C-Lasso)	RMSE (Post-Lasso)	RMSE (Oracle)
h_1^0	100	40	96.87%	0.1409	0.1367	0.1108
	100	80	99.90%	0.0814	0.0800	0.0787
	200	80	99.87%	0.0608	0.0591	0.0578
	200	160	100.00%	0.0440	0.0434	0.0434
h_2^0	100	40	92.29%	0.1645	0.1603	0.0951
	100	80	99.83%	0.0733	0.0701	0.0687
	200	80	99.67%	0.0546	0.0511	0.0496
	200	160	99.99%	0.0374	0.0366	0.0366
h_3^0	100	40	63.73%	1.8325	1.4873	0.1441
	100	80	92.74%	0.1586	0.1479	0.1063
	200	80	90.48%	0.1526	0.1372	0.0783
	200	160	99.90%	0.0599	0.0589	0.0588

2.6 Empirical Illustration

In this section, we apply the model and the estimation method developed in this paper to analyze a textbook example: exploring the effects of different explanatory variables on cigarettes sales in the United States. The data set is from Baltagi, Griffin, and Xiong (2000), which covers 46 American states over the period 1963 - 1992. The explanatory variables included in this data set are the yearly per capita sales of cigarettes, the yearly average retail price of a pack of cigarettes measured at the price level in 1992, the yearly real per capita disposable income and the minimum real price of cigarettes in neighboring states. In Baltagi, Griffin, and Xiong (2000), they modeled the cigarettes sales using a dynamic linear panel regression model which is specified as

$$\ln y_{it} = \alpha + \beta_1 \ln y_{i,t-1} + \beta_2 \ln x_{it,1} + \beta_3 \ln x_{it,2} + \beta_4 \ln x_{it,3} + u_{it}, \quad (2.9)$$

where i represents the i -th state ($i = 1, \dots, 46$), t represents the t -th year ($t = 1, \dots, 29$), y_{it} denotes the yearly per capita sales of cigarettes, $x_{it,1}$ is the yearly average retail price of a pack of cigarettes measured at the price level in 1983, $x_{it,2}$ is the yearly real per capita

disposable income, $x_{it,3}$ is the minimum real price of cigarettes in neighboring states and u_{it} denotes the unobserved demand shock.

Baltagi, Griffin, and Xiong (2000) estimated the model 2.9 using various estimation techniques such as OLS, 2SLS, Shrinkage OLS, etc. However, the estimation results in 2.9 can give very different policy implications since the signs of β 's are opposite when using different estimation techniques. It might be caused by the parametric restriction of the linear panel regression model because the marginal effects of explanatory variables are restricted to be constant. It is well known that the consumer demand for many goods often exhibits diminishing returns to scale, i.e., consumer demand may depend on the absolute scale of certain explanatory variables. Therefore, using linear panel regression models to estimate the demand can also be problematic from consumer theory. To address this problem, we propose to estimate the consumer demand for cigarettes using the nonparametric additive panel regression model with grouped heterogeneity developed in this paper. The grouped heterogeneity of consumer demand may be induced by culture, customs, social norms, and many other latent factors shared by different states. It is worth noting that Mammen, Støve, and Tjøstheim (2009) used a similar additive panel regression model to analyze this data set. Compared with their work, our analysis takes account of the state-level unobserved heterogeneity in the consumer demand for cigarettes, which provides a more accurate picture of the consumer demand on cigarettes. We consider the following model

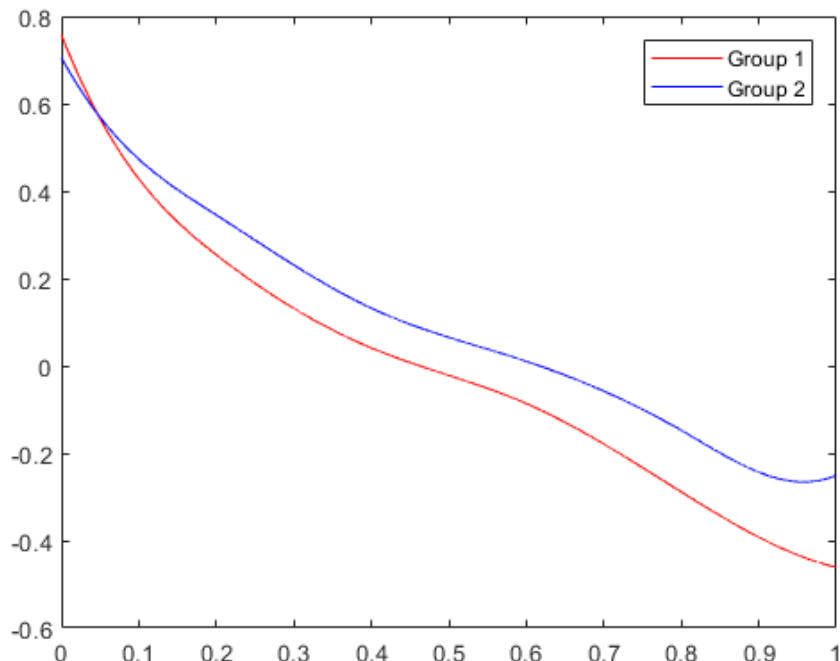
$$\ln y_{it} = \beta_1 \ln y_{i,t-1} + h_{i,1}(x_{it,1}) + h_{i,2}(x_{it,2}) + \alpha_i + u_{it}, \quad (2.10)$$

where $x_{1,it}$ is the yearly average retail price of a pack of cigarettes measured at the price level in 1983, $x_{2,it}$ is the yearly real per capita disposable income. We don't include the minimum real price of cigarettes in neighboring states in model 2.10 because the effect of this variable on the cigarette sales is negligible compared with other explanatory variables. Since our model is nonparametric, it requires a larger amount of observations to ensure the

accuracy of estimation, and thus we omit less relevant variables here.

We impose latent group structures on both $h_{i,1}(x_{it,1})$ and $h_{i,2}(x_{it,2})$ for all $i = 1, \dots, N$. The values of explanatory variables are normalized to $[0, 1]$. Using the information criterion and the estimation method proposed above, we find that there exist two groups of $h_{i,1}(x_{it,1})$. However, we do not find evidence indicating there is grouped heterogeneity in $h_{i,2}(x_{it,2})$. We use post-Lasso estimator to recover the estimated functions of $h_1(x)$ and $h_2(x)$, respectively. The estimated functions of $h_1(x)$ are shown in Figure 2.1.

Figure 2.1: Estimated Functions of h_1



For $h_1(x)$, there are 28 states in Group 1 and 18 states in Group 2. Group 1 includes Arizona, Arkansas, California, Connecticut, Florida, Georgia, Indiana, Iowa, Kansas, Kentucky, Maine, Michigan, Mississippi, Missouri, Nebraska, Nevada, New Hampshire, New Jersey, Ohio, Oklahoma, Pennsylvania, South Carolina, South Dakota, Texas, Utah, Vermont, Virginia, and Washington. On the other hand, Group 2 includes Alabama, Delaware, DC, Idaho, Illinois, Louisiana, Maryland, Massachusetts, Minnesota, Montana, New Mexico, New York, North Dakota, Rhode Island, Tennessee, West Virginia, Wisconsin, and Wyoming.

From Figure 2.1, we can see that consumers living in the states of Group 1 are, on average, more sensitive to the price of cigarettes, meaning that their price elasticity of demand is more considerable.

For $h_2(x)$, the estimation method indicates that only one group exists, and the estimated function is shown in Figure 2.2.

Figure 2.2: Estimated Functions of h_2

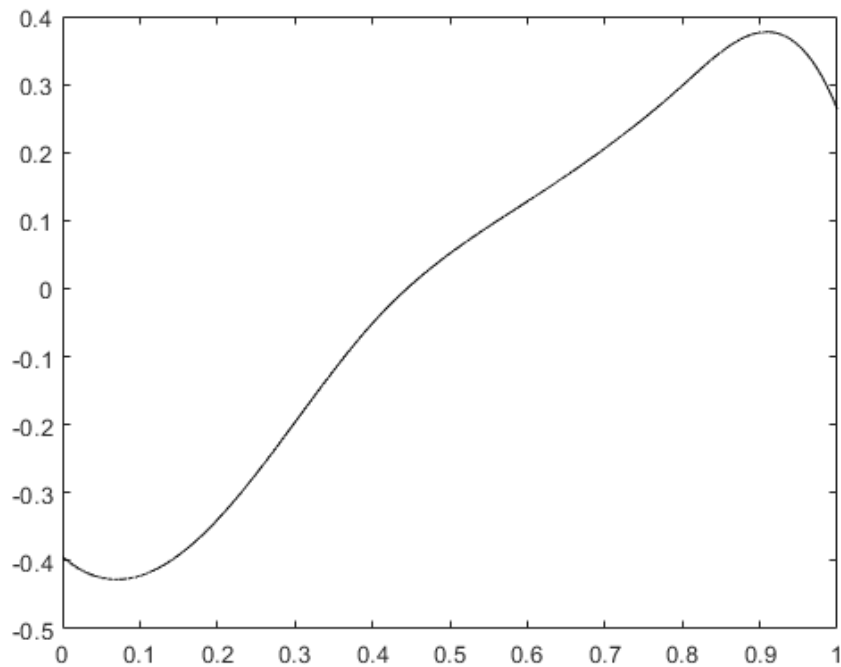


Figure 2.2 implies that states with a higher real per capita disposable income have larger amounts of cigarette sales. This is consistent with the findings in Baltagi and Levin (1992) and Mammen, Støve, and Tjøstheim (2009). It is worth noting that the estimated function of $h_2(x)$ also indicates that the real per capita disposable income will have a negative impact on cigarette sales if it exceeds some threshold. We conjecture that such reduction of cigarette sales is because people with higher income are usually more aware of the harms of smoking on health.

2.7 Conclusion

In this paper, we study a nonparametric additive panel regression model with grouped heterogeneity. This model contributes to the literature on both nonparametric panel regression models and panel models with grouped heterogeneity. The proposed model can handle both the nonlinear effects of explanatory variables and the non-additive heterogeneity at the same time, making it an appealing choice for empirical studies. To estimate the model, we develop a sieve-approximation-based *Classifier*-Lasso estimation method, which can simultaneously estimate the parameters of interest and identify the latent group structure. We successfully establish the asymptotic properties of the proposed estimator and the consistency of the group classification. In addition, we show that the proposed estimation method enjoys the so-called oracle property, which means that parameters are estimated as if the latent group structure is known in advance. Such finding is consistent with Su, Shi, and Phillips (2016) and Su, Wang, and Jin (2019). Since group numbers are usually unknown in general and have to be estimated from the observed data, we further develop a BIC-type information criterion to determine them. We show that this criterion can consistently estimate the number of groups for each nonparametric component under some regularity conditions. We investigate the finite sample performance of the proposed estimators and the information criterion through Monte Carlo simulations. Both work well. Finally, we apply the model and estimation method developed in this paper to estimate the demand for cigarettes in the United States using panel data of 46 American states from 1963 to 1992.

Appendix

2.A Proofs of the Main Results

We use $\|\cdot\|$ to denote Frobenius norm in the Appendix for simplicity.

Proof of Theorem 2.1

Proof. (i) For each individual, I define

$$Q_i(\theta_i) \equiv \frac{1}{T} \sum_{t=1}^T \left(\tilde{y}_{it} - \sum_{j=1}^p \tilde{z}'_{it,j} \theta_{i,j} \right)^2 = \frac{1}{T} \sum_{t=1}^T (\tilde{y}_{it} - \tilde{z}'_{it} \theta_i)^2$$

and

$$Q_i(\theta_i, \eta) \equiv Q_i(\theta_i) + \lambda \sum_{j=1}^p \prod_{k=1}^{K_j^0} \|\theta_{i,j} - \eta_{k,j}\|$$

Since $\hat{\theta}_i$ minimizes $Q_i(\theta_i, \hat{\eta})$, I have $Q_i(\hat{\theta}_i, \hat{\eta}) \leq Q_i(\theta_i^0, \hat{\eta})$, which is equivalent to

$$\left(Q_i(\hat{\theta}_i) - Q_i(\theta_i^0) \right) + \lambda \sum_{j=1}^p \left(\prod_{k=1}^{K_j^0} \|\hat{\theta}_{i,j} - \hat{\eta}_{k,j}\| - \prod_{k=1}^{K_j^0} \|\theta_{i,j}^0 - \hat{\eta}_{k,j}\| \right) \leq 0$$

- Consider the first part:

$$\begin{aligned} & Q_i(\hat{\theta}_i) - Q_i(\theta_i^0) \\ &= \frac{1}{T} \sum_{t=1}^T (\tilde{y}_{it} - \tilde{z}'_{it} \hat{\theta}_i)^2 - \frac{1}{T} \sum_{t=1}^T (\tilde{y}_{it} - \tilde{z}'_{it} \theta_i^0)^2 \\ &= (\hat{\theta}_i - \theta_i^0)' \hat{Q}_{i,\tilde{z}\tilde{z}} (\hat{\theta}_i - \theta_i^0) - 2(\hat{\theta}_i - \theta_i^0)' \hat{Q}_{i,\tilde{z}\tilde{e}} \end{aligned}$$

where $\hat{Q}_{i,\tilde{z}\tilde{z}} = \frac{1}{T} \sum_{t=1}^T \tilde{z}_{it} \tilde{z}'_{it}$, $\hat{Q}_{i,\tilde{z}\tilde{e}} = \frac{1}{T} \sum_{t=1}^T \tilde{z}_{it} \tilde{e}_{it}$, $\tilde{e}_{it} = \sum_{j=1}^p \tilde{\delta}_{h_{i,j}}(x_{it,j}) + \tilde{u}_{it}$.

- Consider the second part, I have

$$\begin{aligned}
& \left| \prod_{k=1}^{K_j^0} \|\hat{\theta}_{i,j} - \hat{\eta}_{k,j}\| - \prod_{k=1}^{K_j^0} \|\theta_{i,j}^0 - \hat{\eta}_{k,j}\| \right| \\
& \leq \left| \prod_{k=1}^{K_j^0-1} \|\hat{\theta}_{i,j} - \hat{\eta}_{k,j}\| \left(\|\hat{\theta}_{i,j} - \hat{\eta}_{K_j^0,j}\| - \|\theta_{i,j}^0 - \hat{\eta}_{K_j^0,j}\| \right) \right| \\
& \quad + \left| \prod_{k=1}^{K_j^0-2} \|\hat{\theta}_{i,j} - \hat{\eta}_{k,j}\| \|\theta_{i,j}^0 - \hat{\eta}_{K_j^0,j}\| \left(\|\hat{\theta}_{i,j} - \hat{\eta}_{K_j^0-1,j}\| - \|\theta_{i,j}^0 - \hat{\eta}_{K_j^0-1,j}\| \right) \right| \\
& \quad + \dots \\
& \quad + \left| \prod_{k=2}^{K_j^0} \|\theta_{i,j}^0 - \hat{\eta}_{k,j}\| \left(\|\hat{\theta}_{i,j} - \hat{\eta}_{1,j}\| - \|\theta_{i,j}^0 - \hat{\eta}_{1,j}\| \right) \right| \\
& \leq c_{1ji,NT}(\hat{\theta}, \theta^0, \hat{\eta}) \|\hat{\theta}_{i,j} - \theta_{i,j}^0\|
\end{aligned}$$

where $c_{1ji,NT}(\hat{\theta}, \theta^0, \hat{\eta}) \equiv \prod_{k=1}^{K_j^0-1} \|\hat{\theta}_{i,j} - \hat{\eta}_{k,j}\| + \prod_{k=1}^{K_j^0-2} \|\hat{\theta}_{i,j} - \hat{\eta}_{k,j}\| \|\theta_{i,j}^0 - \hat{\eta}_{K_j^0,j}\| + \dots + \prod_{k=2}^{K_j^0} \|\theta_{i,j}^0 - \hat{\eta}_{k,j}\|$.

Thus

$$\begin{aligned}
& \left| \sum_{j=1}^p \left(\prod_{k=1}^{K_j^0} \|\hat{\theta}_{i,j} - \hat{\eta}_{k,j}\| - \prod_{k=1}^{K_j^0} \|\theta_{i,j}^0 - \hat{\eta}_{k,j}\| \right) \right| \\
& \leq \sum_{j=1}^p c_{1ji,NT}(\hat{\theta}, \theta^0, \hat{\eta}) \|\hat{\theta}_{i,j} - \theta_{i,j}^0\| \\
& \leq p c_{1i,NT}(\hat{\theta}, \theta^0, \hat{\eta}) \|\hat{\theta}_i - \theta_i^0\|
\end{aligned}$$

where $c_{1i,NT}(\hat{\theta}, \theta^0, \hat{\eta}) = \max_{1 \leq j \leq p} c_{1ji,NT}(\hat{\theta}, \theta^0, \hat{\eta})$.

Together I have

$$\begin{aligned}
& (\hat{\theta}_i - \theta_i^0)' \hat{Q}_{i,\bar{z}\bar{z}} (\hat{\theta}_i - \theta_i^0) \\
& \leq \left| 2(\hat{\theta}_i - \theta_i^0)' \hat{Q}_{i,\bar{z}\bar{e}} \right| + \lambda p c_{1i,NT}(\hat{\theta}, \theta^0, \hat{\eta}) \|\hat{\theta}_i - \theta_i^0\| \\
& \leq 2 \|\hat{\theta}_i - \theta_i^0\| \|\hat{Q}_{i,\bar{z}\bar{e}}\| + \lambda p c_{1i,NT}(\hat{\theta}, \theta^0, \hat{\eta}) \|\hat{\theta}_i - \theta_i^0\|
\end{aligned}$$

By Lemma 2.3, $\mu_{\min}(\hat{Q}_{i,\bar{z}\bar{z}}) > \underline{c} > 0$ w.p.a. 1, then I have w.p.a. 1,

$$\|\hat{\theta}_i - \theta_i^0\| \leq \underline{c}^{-1} \left(2 \|\hat{Q}_{i,\bar{z}\bar{e}}\| + \lambda p c_{1i,NT}(\hat{\theta}, \theta^0, \hat{\eta}) \right)$$

By Lemma 2.3, $\|\hat{Q}_{i,\bar{z}\bar{e}}\| = O_p(J^{-r} + J^{\frac{1}{2}} T^{-\frac{1}{2}})$, thus

$$\|\hat{\theta}_i - \theta_i^0\| = O_p(J^{-r} + J^{\frac{1}{2}} T^{-\frac{1}{2}} + \lambda)$$

Consequently we could get

$$\|\hat{\theta}_{i,j} - \theta_{i,j}^0\| = O_p(J^{-r} + J^{\frac{1}{2}} T^{-\frac{1}{2}} + \lambda)$$

for $i = 1, 2, \dots, N$ and $j = 1, \dots, p$.

Remark. *The argument depends on the condition that $c_{1i,NT}(\hat{\theta}, \theta^0, \hat{\eta}) = Op(1)$.*

We show this by considering a constrained optimization problem.

Define

$$\begin{aligned}
\mathcal{R}_b & \equiv \left\{ \gamma : |\gamma_{i,jm}| \leq c < \infty, i = 1, \dots, N, j = 1, \dots, p, m = 1, \dots, J \right\} \\
\Pi_b & \equiv \left\{ \pi : |\pi_{k,jm}| \leq c < \infty, k = 1, \dots, K_j^0, j = 1, \dots, p, m = 1, \dots, J \right\}
\end{aligned}$$

where c is a generic constant, $\gamma = (\gamma_1, \dots, \gamma_N)$, $\gamma_i = (\gamma'_{i,1}, \dots, \gamma'_{i,p})'$ for $i = 1, \dots, N$, $\pi = (\pi'_1, \dots, \pi'_p)'$, $\pi_j = (\pi'_{1,j}, \dots, \pi'_{K_j^0,j})'$ for $j = 1, \dots, p$.

Further define $\Theta_b \equiv \{\theta : \gamma \in \mathcal{R}_b\}$, $\mathcal{H}_b \equiv \{\eta : \pi \in \Pi_b\}$. Remember that $\theta = (\theta_1, \dots, \theta_N)$, where $\theta_i \equiv \frac{1}{\sqrt{J}}\gamma_i$, $i = 1, \dots, N$, and $\eta = (\eta'_1, \dots, \eta'_p)'$, where $\eta_j \equiv \frac{1}{\sqrt{J}}\pi_j$, $j = 1, \dots, p$.

If c is large enough, by Assumption 2.1(iii), we could get that γ^0 and π^0 lie in the interior of \mathcal{R}_b and Π_b respectively, thus $\theta^0 \in \Theta_b$ and $\eta^0 \in \mathcal{H}_b$.

Then we search over Θ_b and \mathcal{H}_b to minimize the objective function 2.7, namely

$$(\hat{\theta}, \hat{\eta}) = \arg \min_{\theta \in \Theta_b, \eta \in \mathcal{H}_b} \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T (\tilde{y}_{it} - \tilde{z}'_{it}\theta_i)^2 + \frac{\lambda}{N} \sum_{i=1}^N \sum_{j=1}^p \prod_{k=1}^{K_j^0} \|\theta_{i,j} - \eta_{k,j}\|_F$$

The restrictions guarantee that $c_{1i,NT}(\hat{\theta}, \theta^0, \hat{\eta}) = O(1)$.

Practically, we set c large enough and conduct the constrained optimization, which works well in my simulations.

- (ii) Let $m_{JT} = J^{-r} + J^{\frac{1}{2}}T^{-\frac{1}{2}}$ and v denotes a $(pJ) \times N$ matrix. In order to show that $\frac{1}{N} \sum_{i=1}^N \|\hat{\theta}_i - \theta_i^0\|^2 = O_p(J^{-2r} + JT^{-1})$, I just need to prove that for any ε , there exists a constant $M = M(\varepsilon)$ such that, for sufficiently large N and T ,

$$P \left\{ \inf_{\frac{1}{N} \sum_{i=1}^N \|v_i\|^2 = M} Q_{NT}(\theta^0 + m_{JT}v, \hat{\eta}) > Q_{NT}(\theta^0, \eta^0) \right\} \geq 1 - \varepsilon$$

This implies that w.p.a.1 there exists a local minimum $\{\hat{\theta}, \hat{\eta}\}$ such that $\frac{1}{N} \sum_{i=1}^N \|\hat{\theta}_i - \theta_i^0\|^2 = O_p(J^{-2r} + JT^{-1})$ holds.

$$\begin{aligned} & m_{JT}^{-2} \left(Q_{NT}(\theta^0 + m_{JT}v, \hat{\eta}) - Q_{NT}(\theta^0, \eta^0) \right) \\ &= \frac{1}{N} \sum_{i=1}^N v'_i \hat{Q}_{i,\tilde{z}\tilde{z}} v_i - \frac{2}{N} m_{JT}^{-1} \sum_{i=1}^N v'_i \hat{Q}_{i,\tilde{z}\tilde{e}} + \frac{\lambda}{N} \sum_{i=1}^N \sum_{j=1}^p \prod_{k=1}^{K_j^0} \|\theta_{i,j}^0 + m_{JT}v_{i,j} - \hat{\eta}_{k,j}\| \\ &\geq c \frac{1}{N} \sum_{i=1}^N \|v_i\|^2 - 2 \left\{ \frac{1}{N} \sum_{i=1}^N \|v_i\|^2 \right\}^{\frac{1}{2}} \left\{ \frac{m_{JT}^{-2}}{N} \sum_{i=1}^N \|\hat{Q}_{i,\tilde{z}\tilde{e}}\|^2 \right\}^{\frac{1}{2}} \end{aligned}$$

where the last inequality holds w.p.a.1 by Lemma 2.3.

By Lemma 2.3, $\frac{1}{N} \sum_{i=1}^N \|\hat{Q}_{i,\bar{z}\bar{e}}\|^2 = O_p(J^{-2r} + JT^{-1})$, then $\frac{m_{JT}^{-2}}{N} \sum_{i=1}^N \|\hat{Q}_{i,\bar{z}\bar{e}}\|^2 = O_p(1)$, thus for sufficiently large M , I have $m_{JT}^{-2} (Q_{NT}(\theta^0 + m_{JT}v, \hat{\eta}) - Q_{NT}(\theta_0, \eta_0)) > 0$ w.p.a.1.

Since $\frac{1}{N} \sum_{i=1}^N \|\hat{\theta}_{i,j} - \theta_{i,j}^0\|^2 \leq \frac{1}{N} \sum_{i=1}^N \|\hat{\theta}_i - \theta_i^0\|^2$, we also have $\frac{1}{N} \sum_{i=1}^N \|\hat{\theta}_{i,j} - \theta_{i,j}^0\|^2 = O_p(J^{-2r} + JT^{-1})$.

(iii) Further consider $c_{1ji,NT}(\hat{\theta}, \theta^0, \eta)$, where $\hat{\theta}$ and η lie in the interior of Θ_b and \mathcal{H}_b respectively.

$$\begin{aligned}
& c_{1ji,NT}(\hat{\theta}, \theta^0, \eta) \\
&= \prod_{k=1}^{K_j^0-1} \|\hat{\theta}_{i,j} - \eta_{k,j}\| + \prod_{k=1}^{K_j^0-2} \|\hat{\theta}_{i,j} - \eta_{k,j}\| \|\theta_{i,j}^0 - \eta_{K_j^0,j}\| + \cdots + \prod_{k=2}^{K_j^0} \|\theta_{i,j}^0 - \eta_{k,j}\| \\
&\leq \prod_{k=1}^{K_j^0-1} \left(\|\hat{\theta}_{i,j} - \theta_{i,j}^0\| + \|\theta_{i,j}^0 - \eta_{k,j}\| \right) + \prod_{k=1}^{K_j^0-2} \left(\|\hat{\theta}_{i,j} - \theta_{i,j}^0\| + \|\theta_{i,j}^0 - \eta_{k,j}\| \right) \|\theta_{i,j}^0 - \eta_{K_j^0,j}\| \\
&\quad + \cdots + \prod_{k=2}^{K_j^0} \|\theta_{i,j}^0 - \eta_{k,j}\| \\
&\leq \sum_{s=0}^{K_j^0-1} c_{1jsi,NT}(\theta^0, \eta) \|\hat{\theta}_{i,j} - \theta_{i,j}^0\|^s + \sum_{s=0}^{K_j^0-2} c_{2jsi,NT}(\theta^0, \eta) \|\hat{\theta}_{i,j} - \theta_{i,j}^0\|^s \\
&\quad + \cdots + \sum_{s=0}^0 c_{K_j^0psi,NT}(\theta^0, \eta) \|\hat{\theta}_{i,j} - \theta_{i,j}^0\|^s \\
&\leq \sum_{s=0}^{K_j^0-1} c_{jsi,NT}(\theta^0, \eta) \|\hat{\theta}_{i,j} - \theta_{i,j}^0\|^s \\
&\leq c_{2ji,NT}(\theta^0, \eta) \sum_{s=0}^{K_j^0-1} \|\hat{\theta}_{i,j} - \theta_{i,j}^0\|^s \\
&\leq c_{2ji,NT}(\theta^0, \eta) \left(1 + 2 \|\hat{\theta}_{i,j} - \theta_{i,j}^0\| \right)
\end{aligned}$$

where $c_{2ji,NT}(\theta^0, \eta) = \max_{1 \leq s \leq K_j^0} c_{jsi,NT}(\theta^0, \eta)$ and $c_{jsi,NT}(\theta^0, \eta) = \sum_{k=1}^{K_j^0} c_{kjsi,NT}(\theta^0, \eta)$.

The last inequality holds w.p.a 1.

Define $p_{NT}(\theta, \eta) \equiv \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^p \prod_{k=1}^{K_j^0} \|\theta_{i,j} - \eta_{k,j}\|$, then

$$\begin{aligned}
& \left| p_{NT}(\hat{\theta}, \eta) - p_{NT}(\theta^0, \eta) \right| \\
& \leq \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^p c_{1ji,NT}(\hat{\theta}, \theta^0, \eta) \|\hat{\theta}_i - \theta_i^0\| \\
& \leq \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^p c_{2ji,NT}(\theta^0, \eta) \left(\|\hat{\theta}_i - \theta_i^0\| + 2\|\hat{\theta}_i - \theta_i^0\|^2 \right) \\
& \leq p c_{2i,NT}(\theta^0, \eta) \left(\frac{1}{N} \sum_{i=1}^N \|\hat{\theta}_i - \theta_i^0\|^2 \right)^{\frac{1}{2}} + p c_{2i,NT}(\theta^0, \eta) \frac{1}{N} \sum_{i=1}^N \|\hat{\theta}_i - \theta_i^0\|^2 \\
& = O_p(J^{-r} + J^{\frac{1}{2}} T^{-\frac{1}{2}})
\end{aligned}$$

where $c_{2i,NT}(\theta^0, \eta) = \max_{1 \leq j \leq p} c_{2ji,NT}(\theta^0, \eta)$ and we use $c_{2ji,NT}(\theta^0, \eta) = O(1)$, which is implied by a similar argument as that in the proof of Theorem 2.1(i).

Since $p_{NT}(\hat{\theta}, \hat{\eta}) \leq p_{NT}(\hat{\theta}, \eta^0)$, note that $p_{NT}(\theta^0, \eta^0) = 0$,

$$\begin{aligned}
0 & \geq p_{NT}(\hat{\theta}, \hat{\eta}) - p_{NT}(\hat{\theta}, \eta^0) \\
& = \left(p_{NT}(\hat{\theta}, \hat{\eta}) - p_{NT}(\theta^0, \hat{\eta}) \right) + \left(p_{NT}(\theta^0, \hat{\eta}) - p_{NT}(\theta^0, \eta^0) \right) - \left(p_{NT}(\hat{\theta}, \eta^0) - p_{NT}(\theta^0, \eta^0) \right) \\
& = O_p(J^{-r} + J^{\frac{1}{2}} T^{-\frac{1}{2}}) + p_{NT}(\theta^0, \hat{\eta}) \\
& = O_p(J^{-r} + J^{\frac{1}{2}} T^{-\frac{1}{2}}) + \sum_{j=1}^p \sum_{m=1}^{K_j^0} \frac{N_{m,j}}{N} \prod_{k=1}^{K_j^0} \|\eta_{m,j}^0 - \hat{\eta}_{k,j}\|
\end{aligned}$$

Then there exists a permutation of $\{1, \dots, K_j^0\}$ for $j = 1, \dots, p$ such that $\|\hat{\eta}_{k,j} - \eta_{k,j}^0\| = O_p(J^{-r} + J^{\frac{1}{2}} T^{-\frac{1}{2}})$.

□

Proof of Theorem 2.2

Proof. (i) For any $i \in G_{k_j,j}^0$, $j = 1, \dots, p$ and $l \neq k_j$, by Theorem 2.1, $\|\hat{\theta}_{i,j} - \hat{\eta}_{l,j}\| \xrightarrow{p} \|\eta_{k_j,j}^0 - \eta_{l,j}^0\| \neq 0$.

Define a set $\mathbb{S}_i = \left\{ j; \left\| \hat{\theta}_{i,j} - \hat{\eta}_{k_j,j} \right\| \neq 0, i \in G_{k_j,j}^0, \forall 1 \leq k_j \leq K_j^0 \right\}$, which means that $i \in G_{k_j,j}^0$ and $i \notin \hat{G}_{k_j,j}$ if and only if $j \in \mathbb{S}_i$, then the first order condition with respect to $\theta_{i,j}, j \in \mathbb{S}_i$ is

$$\begin{aligned}
0_J &= -2\hat{Q}_{i,\tilde{z}\tilde{u},j} - 2\hat{Q}_{i,\tilde{z}\tilde{\delta},j} + 2\sum_{r=1}^p \hat{Q}_{i,\tilde{z}\tilde{z},jr} \left(\hat{\theta}_{i,r} - \theta_{i,r}^0 \right) \\
&\quad + \frac{\lambda}{\left\| \hat{\theta}_{i,j} - \hat{\eta}_{k_j,j} \right\|} \prod_{l=1, l \neq k_j}^{K_j^0} \left\| \hat{\theta}_{i,j} - \hat{\eta}_{l,j} \right\| \left(\hat{\theta}_{i,j} - \hat{\eta}_{k_j,j} \right) \\
&\quad + \lambda \sum_{m=1, m \neq k_j}^{K_j^0} \hat{e}_{im,j} \prod_{l=1, l \neq m}^{K_j^0} \left\| \hat{\theta}_{i,j} - \hat{\eta}_{l,j} \right\| \\
&= -2\hat{Q}_{i,\tilde{z}\tilde{u},j} - 2\hat{Q}_{i,\tilde{z}\tilde{\delta},j} + 2\sum_{r \in \mathbb{S}_i} \hat{Q}_{i,\tilde{z}\tilde{z},jr} \left(\hat{\theta}_{i,r} - \hat{\eta}_{k_j,r} \right) \\
&\quad + \frac{\lambda}{\left\| \hat{\theta}_{i,j} - \hat{\eta}_{k_j,j} \right\|} \prod_{l=1, l \neq k_j}^{K_j^0} \left\| \hat{\theta}_{i,j} - \hat{\eta}_{l,j} \right\| \left(\hat{\theta}_{i,j} - \hat{\eta}_{k_j,j} \right) \\
&\quad + 2\sum_{r \in \mathbb{S}_i} \hat{Q}_{i,\tilde{z}\tilde{z},jr} \left(\hat{\eta}_{k_j,r} - \theta_{i,r}^0 \right) \\
&\quad + \lambda \sum_{m=1, m \neq k_j}^{K_j^0} \hat{e}_{im,j} \prod_{l=1, l \neq m}^{K_j^0} \left\| \hat{\theta}_{i,j} - \hat{\eta}_{l,j} \right\| \\
&\quad + 2\sum_{r \notin \mathbb{S}_i} \hat{Q}_{i,\tilde{z}\tilde{z},jr} \left(\hat{\theta}_{i,r} - \theta_{i,r}^0 \right) \\
&\equiv \hat{A}_{i1,j} + \hat{A}_{i2,j} + \hat{A}_{i3,j} + \hat{A}_{i4,j} + \hat{A}_{i5,j} + \hat{A}_{i6,j} + \hat{A}_{i7,j} \\
&\equiv \hat{A}_{i,j}
\end{aligned}$$

where $\hat{e}_{im,j} = \frac{\hat{\theta}_{i,j} - \hat{\eta}_{m,j}}{\left\| \hat{\theta}_{i,j} - \hat{\eta}_{m,j} \right\|}$ if $\left\| \hat{\theta}_{i,j} - \hat{\eta}_{m,j} \right\| \neq 0$ and $\hat{e}_{im,j} \leq 1$ otherwise, and $\hat{Q}_{i,\tilde{z}\tilde{u},j} = \frac{1}{T} \sum_{t=1}^T \tilde{z}_{it,j} \tilde{u}_{it}$, $\hat{Q}_{i,\tilde{z}\tilde{\delta},j} = \frac{1}{T} \sum_{t=1}^T \tilde{z}_{it,j} \tilde{\delta}_{h_i,it}$, $\tilde{\delta}_{h_i,it} = \sum_{j=1}^p \tilde{\delta}_{h_i,j,it}$, $\hat{Q}_{i,\tilde{z}\tilde{z},jr} = \frac{1}{T} \sum_{t=1}^T \tilde{z}_{it,j} \tilde{z}'_{it,r}$.

From the proof of Theorem 2.1, I have that

$$\left\| \hat{\theta}_i - \theta_i^0 \right\| \leq \underline{c}^{-1} \left(2 \left\| \hat{Q}_{i,\tilde{z}\tilde{e}} \right\| + \lambda p c_{1i,NT}(\hat{\theta}, \theta^0, \hat{\eta}) \right)$$

Let $\mu_{1,JT} = \left(J^{-r} + J^{\frac{1}{2}} T^{-\frac{1}{2}} (\ln T)^3 + \lambda \right) (\ln T)^v$ and $\mu_{2,JT} = \left(J^{-r} + J^{\frac{1}{2}} T^{-\frac{1}{2}} (\ln T)^3 \right) (\ln T)^v$

for some $v > 0$. By Lemma 2.3, I could show that

$$P\left(\max_{1 \leq i \leq N} \|\hat{\theta}_i - \theta_i^0\| \geq c\mu_{1,JT}\right) = o(N^{-1})$$

$$P\left(\|\hat{\eta}_k - \eta_k^0\| \geq c\mu_{2,JT}\right) = o(N^{-1})$$

for any $c > 0$.

Let $\hat{c}_{ik_j,j} = \prod_{l=1, l \neq k_j}^{K_j^0} \|\hat{\theta}_{i,j} - \hat{\eta}_{l,j}\|$, then

$$\begin{aligned} \hat{c}_{ik_j,j} &= \prod_{l=1, l \neq k_j}^{K_j^0} \|\hat{\theta}_{i,j} - \hat{\eta}_{l,j}\| \\ &= \prod_{l=1, l \neq k_j}^{K_j^0} \left\| (\hat{\theta}_{i,j} - \eta_{k_j,j}^0) - (\hat{\eta}_{l,j} - \eta_{l,j}^0) + (\eta_{k_j,j}^0 - \eta_{l,j}^0) \right\| \\ &= \prod_{l=1, l \neq k_j}^{K_j^0} \left\| \eta_{k_j,j}^0 - \eta_{l,j}^0 + o_p(1) \right\| \\ &= O_p(1) \end{aligned}$$

Similarly let $c_{ik_j,j}^0 = \prod_{l=1, l \neq k_j}^{K_j^0} \|\theta_{i,j}^0 - \eta_{l,j}^0\|$. Define $\bar{c}^0 = \max_{j \in \mathbb{S}_i} \max_{i \in G_{k_j,j}^0} c_{ik_j,j}^0$ and $\underline{c}^0 = \min_{j \in \mathbb{S}_i} \min_{i \in G_{k_j,j}^0} c_{ik_j,j}^0$, then for all $j \in \mathbb{S}_i$,

$$P\left(\frac{\underline{c}^0}{2} \leq \hat{c}_{ik_j,j} \leq 2\bar{c}^0\right) = 1 - o(N^{-1})$$

And $P\left(\max_{i \in G_{k_j,j}^0} \|\hat{A}_{i6,j}\| \geq C\lambda\mu_{1,JT}\right) = o(N^{-1})$ for large enough $C > 0$.

Define

$$\begin{aligned} \Xi_{kNT} &\equiv \left\{ \frac{\underline{c}^0}{2} \leq \hat{c}_{ik_j,j} \leq 2\bar{c}^0, \forall j \in \mathbb{S}_i \right\} \cap \left\{ \|\hat{\eta}_{k,j} - \eta_{k,j}^0\| \leq c\mu_{2,JT}, \forall 1 \leq k \leq K_j^0, \forall 1 \leq j \leq p \right\} \\ &\cap \left\{ 0 < \underline{c} < \min_{0 \leq i \leq N} \mu_{\min}(\hat{Q}_{i,\bar{z}\bar{z}}) \leq \max_{0 \leq i \leq N} \mu_{\max}(\hat{Q}_{i,\bar{z}\bar{z}}) < \bar{c} < \infty \right\} \\ &\cap \left\{ \max_{1 \leq i \leq N} \|\hat{Q}_{i,\bar{z}\bar{\delta},j}\| \leq C\theta_{NT}, \forall 1 \leq j \leq p \right\} \cap \left\{ \max_{1 \leq i \leq N} \|\hat{\theta}_i - \theta_i^0\| \leq c\mu_{1,JT} \right\} \end{aligned}$$

for some $C > 0$ and $c > 0$. $\theta_{NT} \equiv \max_{0 \leq j \leq p} \max_{1 \leq k \leq K_j^0} \sup_{x \in [0,1]} \|f_{k,j}^0(x) - B^{J'} \pi_{k,j}^0\| = O(J^{-r})$.

Then $P(\Xi_{kNT}) = 1 - o(N^{-1})$.

For all $j \in \mathbb{S}_i$, define $\psi_{ik_j} \equiv \left((\hat{\theta}_{i,j} - \hat{\eta}_{k_j,j})', j \in \mathbb{S}_i \right)'$ and $\phi_{ik_j} \equiv \|\psi_{ik_j}\|$. I multiply $\hat{A}_{i,j}$ from left by $\frac{(\hat{\theta}_{i,j} - \hat{\eta}_{k_j,j})'}{\phi_{ik_j}}$, and then take summation for all $j \in \mathbb{S}_i$, then I could get

$$\begin{aligned}
0 &= -2 \frac{1}{\phi_{ik_j}} \sum_{j \in \mathbb{S}_i} (\hat{\theta}_{i,j} - \hat{\eta}_{k_j,j})' \hat{Q}_{i,\bar{z}\bar{u},j} - 2 \frac{1}{\phi_{ik_j}} \sum_{j \in \mathbb{S}_i} (\hat{\theta}_{i,j} - \hat{\eta}_{k_j,j})' \hat{Q}_{i,\bar{z}\bar{\delta},j} \\
&\quad + 2 \frac{1}{\phi_{ik_j}} \sum_{j \in \mathbb{S}_i} \sum_{r \in \mathbb{S}_i} (\hat{\theta}_{i,j} - \hat{\eta}_{k_j,j})' \hat{Q}_{i,\bar{z}\bar{z},jr} (\hat{\theta}_{i,r} - \hat{\eta}_{k_j,r}) \\
&\quad + \frac{1}{\phi_{ik_j}} \sum_{j \in \mathbb{S}_i} (\hat{\theta}_{i,j} - \hat{\eta}_{k_j,j})' \frac{\lambda}{\|\hat{\theta}_{i,j} - \hat{\eta}_{k_j,j}\|} \prod_{l=1, l \neq k_j}^{K_j^0} \|\hat{\theta}_{i,j} - \hat{\eta}_{l,j}\| (\hat{\theta}_{i,j} - \hat{\eta}_{k_j,j}) \\
&\quad + 2 \frac{1}{\phi_{ik_j}} \sum_{j \in \mathbb{S}_i} \sum_{r \in \mathbb{S}_i} (\hat{\theta}_{i,j} - \hat{\eta}_{k_j,j})' \hat{Q}_{i,\bar{z}\bar{z},jr} (\hat{\eta}_{k_j,r} - \theta_{i,r}^0) \\
&\quad + \lambda \frac{1}{\phi_{ik_j}} \sum_{j \in \mathbb{S}_i} (\hat{\theta}_{i,j} - \hat{\eta}_{k_j,j})' \sum_{m=1, m \neq k_j}^{K_j^0} \hat{e}_{im,j} \prod_{l=1, l \neq m}^{K_j^0} \|\hat{\theta}_{i,j} - \hat{\eta}_{l,j}\| \\
&\quad + 2 \frac{1}{\phi_{ik_j}} \sum_{j \in \mathbb{S}_i} (\hat{\theta}_{i,j} - \hat{\eta}_{k_j,j})' \sum_{r \notin \mathbb{S}_i} \hat{Q}_{i,\bar{z}\bar{z},jr} (\hat{\theta}_{i,r} - \theta_{i,r}^0) \\
&\equiv \hat{A}_{i1} + \hat{A}_{i2} + \hat{A}_{i3} + \hat{A}_{i4} + \hat{A}_{i5} + \hat{A}_{i6} + \hat{A}_{i7}
\end{aligned}$$

Conditional on Ξ_{kNT} , I have that uniformly in $i \in G_{k_j,j}^0$, $j \in \mathbb{S}_i$, with probability $1 - o(N^{-1})$,

$$|\hat{A}_{i3} + \hat{A}_{i4}| \geq 2c\phi_{ik_j} + \lambda \min_{j \in \mathbb{S}_i} \hat{c}_{ik_j,j} \geq \lambda \frac{c^0}{2}$$

$$|\hat{A}_{i2}| \leq 2 \|\hat{Q}_{i,\bar{z}\bar{\delta},j}\| \leq 2C\theta_{NT}$$

$$|\hat{A}_{i5}| \leq 2C\mu_{2,JT}$$

$$|\hat{A}_{i6}| \leq C\lambda\mu_{1,JT}$$

$$|\hat{A}_{i7}| \leq C\lambda\mu_{1,JT}$$

Then

$$\begin{aligned}
& \left| \hat{A}_{i2} + \hat{A}_{i3} + \hat{A}_{i4} + \hat{A}_{i5} + \hat{A}_{i6} \right| \\
& \geq \left| \hat{A}_{i2} \right| - \left| \hat{A}_{i3} + \hat{A}_{i4} + \hat{A}_{i5} + \hat{A}_{i6} \right| \\
& \geq \lambda \frac{c^0}{2} - \left[2C\theta_{NT} + 2C\mu_{2,JT} + 2C\lambda\mu_{1,JT} \right] \\
& \geq \lambda \frac{c^0}{4}
\end{aligned}$$

where I use Assumption 2.2 and 2.3.

Thus

$$\begin{aligned}
& P \left(i \notin \hat{G}_{k,j} \mid i \in G_{k,j}^0, j \in \mathbb{S}_i \right) \\
& = P \left(-\hat{A}_{i1} = \hat{A}_{i2} + \hat{A}_{i3} + \hat{A}_{i4} + \hat{A}_{i5} + \hat{A}_{i6} + \hat{A}_{i7} \right) \\
& \leq P \left(\left| \hat{A}_{i1} \right| \geq \left| \hat{A}_{i2} + \hat{A}_{i3} + \hat{A}_{i4} + \hat{A}_{i5} + \hat{A}_{i6} + \hat{A}_{i7} \right| \right) \\
& \leq P \left(\left| \hat{A}_{i1} \right| \geq \lambda \frac{c^0}{4}, \Xi_{kNT} \right) + P \left(\Xi_{kNT}^c \right) \\
& = o(N^{-1})
\end{aligned}$$

For any $\mathbb{S}_i \subseteq \{1, \dots, p\}$, the result holds. So I could further get that $P(\hat{E}_{ik,j}) = P(i \notin \hat{G}_{k,j} \mid i \in G_{k,j}^0) = o(N^{-1})$.

Then

$$\begin{aligned}
& P(\cup_{j=1}^p \cup_{k=1}^{K_j^0} \hat{\mathbf{E}}_{k,j}) \\
& \leq \sum_{j=1}^p \sum_{k=1}^{k_j^0} P(\hat{\mathbf{E}}_{k,j}) \\
& \leq \sum_{j=1}^p \sum_{k=1}^{k_j^0} \sum_{i \in G_{k,j}^0} P(\hat{\mathbf{E}}_{ik,j}) \\
& \leq \sum_{j=1}^p \sum_{k=1}^{k_j^0} \sum_{i \in G_{k,j}^0} \left(P\left(\left|\hat{A}_{i1}\right| \geq \lambda \frac{c^0}{4}, \Xi_{kNT}\right) + P(\Xi_{kNT}^c) \right) \\
& \leq Np \max_{1 \leq i \leq N} P\left(\left\|\hat{Q}_{i,\bar{z}\bar{u}}\right\| \geq \lambda \frac{c^0}{4}\right) + o(1) \\
& \leq NpP\left(\max_{1 \leq i \leq N} \left\|\hat{Q}_{i,\bar{z}\bar{u}}\right\| \geq \lambda \frac{c^0}{4}\right) + o(1) \\
& = o(1)
\end{aligned}$$

where I use $\lambda T^{\frac{1}{2}} J^{-\frac{1}{2}} (\ln T)^{-3} \rightarrow \infty$.

(ii) The proof is similar to Su, Shi, and Phillips (2016) Theorem 2.2 (ii) and thus omitted. □

Proof of Theorem 2.3

Proof. The proof of Theorem 2.3 is similar to the one in Su, Wang, and Jin (2019) and thus is omitted. □

2.B Proofs of Technical Lemmas

We use $\|\cdot\|$ to denote Frobenius norm in the Appendix for simplicity and use C to indicate some generic constant, which varies.

Lemma 2.1. *Let ξ_{it} be a \mathcal{R}^{d_ξ} random variable and $\mathbf{E}[\xi_{it}] = 0$ for all i, t . For each $i = 1, \dots, N$, ξ_{it} is stationary strong mixing with mixing coefficient $\alpha_i(j)$. $\alpha(j) \equiv \max_{1 \leq i \leq N} \alpha_i(j)$ satisfies $\alpha(j) \leq c_\alpha \exp(-\rho j)$ for some $0 < c_\alpha < \infty$, $0 < \rho < \infty$. ξ_{it} are independent across i . Assume that $\mathbf{E}\|\xi_{it}\|^q < \infty$ for some $q \geq 3$, Then*

$$P\left(\max_{1 \leq i \leq N} \left\| \frac{1}{T} \sum_{t=1}^T \xi_{it} \right\| \geq CT^{-\frac{1}{2}} (\ln T)^3\right) = o(N^{-1})$$

for large enough $C > 0$ if $N^2 T^{1-\frac{q}{2}} = O(1)$.

Proof. It is the same as Lemma 1.1 and thus omitted. □

Lemma 2.2. *Let ξ_{it} be a \mathcal{R}^{d_ξ} random variable and $\mathbf{E}[\xi_{it}] = 0$ for all i, t . For each $i = 1, \dots, N$, ξ_{it} is stationary strong mixing with mixing coefficient $\alpha_i(j)$. $\alpha(j) \equiv \max_{1 \leq i \leq N} \alpha_i(j)$ satisfies $\alpha(j) \leq c_\alpha \exp(-\rho j)$ for some $0 < c_\alpha < \infty$, $0 < \rho < \infty$. ξ_{it} are independent across i . Assume that $\max_{1 \leq i \leq N} \max_{1 \leq t \leq T} \mathbf{E}\|\xi_{it}\|^{\frac{q}{2}} < \infty$ for some $q > 6$ such that $N^2 T^{1-\frac{q}{2}} (\ln T)^{\frac{3q}{2}} \rightarrow 0$ as $N, T \rightarrow \infty$. Then*

$$P\left(\max_{1 \leq i \leq N} \left\| \frac{1}{T} \sum_{t=1}^T \xi_{it} \right\| \geq c\right) = o(N^{-1})$$

for any $c > 0$.

Proof. It is the same as Lemma 1.2 and thus omitted. □

Lemma 2.3. *Suppose that Assumption 2.1 and 2.2 hold, then*

(i)

$$P(0 < \underline{c} < \min_{0 \leq i \leq N} \mu_{\min}(\hat{Q}_{i, \bar{z}\bar{z}}) \leq \max_{0 \leq i \leq N} \mu_{\max}(\hat{Q}_{i, \bar{z}\bar{z}}) < \bar{c} < \infty) = 1 - o(N^{-1})$$

(ii)

$$\|\hat{Q}_{i,\bar{z}\bar{e}}\| = O_p(J^{-r} + J^{\frac{1}{2}}T^{-\frac{1}{2}})$$

(iii)

$$\frac{1}{N} \sum_{i=1}^N \|\hat{Q}_{i,\bar{z}\bar{e}}\|^2 = O_p(J^{-2r} + JT^{-1})$$

(iv)

$$P \left(\max_{0 \leq i \leq N} \|\hat{Q}_{i,\bar{z}\bar{e}}\| \geq c \left(J^{-r} + J^{\frac{1}{2}}T^{-\frac{1}{2}} (\ln T)^3 \right) (\ln T)^v \right) = o(N^{-1})$$

for any $c > 0$ and some $v > 0$.

Proof. (i) Consider the difference between $\text{Var}(z_{it})$ and $\hat{Q}_{i,\bar{z}\bar{z}}$.

Let $\mu_k(A)$ be the k th largest eigenvalue of matrix A . Denote \mathbb{S}_{pJ} as the permutation group of $\{1, \dots, pJ\}$. By Hoffman-Wielandt inequality,

$$\min_{\sigma \in \mathbb{S}_{pJ}} \sum_{k=1}^{pJ} \left| \mu_k(\hat{Q}_{i,\bar{z}\bar{z}}) - \mu_{\sigma(k)}(\text{Var}(z_{it})) \right|^2 \leq \left\| \hat{Q}_{i,\bar{z}\bar{z}} - \text{Var}(z_{it}) \right\|^2$$

Because

$$\begin{aligned} & \left\| \hat{Q}_{i,\bar{z}\bar{z}} - \text{Var}(z_{it}) \right\|^2 \\ & \leq 2 \left\| \hat{Q}_{i,zz} - \mathbf{E}[z_{it}z'_{it}] \right\|^2 + 2 \left\| \frac{1}{T} \sum_{t=1}^T z_{it} \frac{1}{T} \sum_{t=1}^T z'_{it} - \mathbf{E}[z_{it}]\mathbf{E}[z'_{it}] \right\|^2 \end{aligned}$$

(i) Consider the first item, for any $c > 0$, $v > 0$,

- Similar as the proof in Lemma 2.2, we could get

$$\begin{aligned} & P \left(\max_{1 \leq r \leq p} \max_{1 \leq s \leq p} \max_{1 \leq i \leq N} \max_{1 \leq j \leq J} \max_{1 \leq k \leq J} \left| \frac{1}{T} \sum_{t=1}^T JB_{rit,j}^J B_{sit,k}^J - \mathbf{E} [JB_{rit,j}^J B_{sit,k}^J] \right| \geq cJ^{-\frac{1}{2}} \right) \\ & = o(N^{-1}) \end{aligned}$$

Note that there are only $O(J)$ nonzero elements in $B_{it}^J B_{it}^{J'} - \mathbf{E} [B_{it}^J B_{it}^{J'}]$.

Thus for any $c > 0$,

$$P \left(\max_{1 \leq i \leq N} \left\| \hat{Q}_{i,zz} - \mathbf{E}[z_{it}z'_{it}] \right\|^2 \geq c \right) = o(N^{-1})$$

(ii) Consider the second item, for any $c > 0$, similar as the proof in Lemma 2.2,

$$P \left(\max_{1 \leq r \leq p} \max_{1 \leq i \leq N} \max_{1 \leq j \leq J} \left| \frac{1}{T} \sqrt{J} B_{rit,j}^J - \mathbf{E} \left[\sqrt{J} B_{rit,j}^J \right] \right| \geq cJ^{-1} \right) = o(N^{-1})$$

Thus we could get

$$P \left(\max_{1 \leq i \leq N} \left\| \frac{1}{T} \sum_{t=1}^T z_{it} \frac{1}{T} \sum_{t=1}^T z'_{it} - \mathbf{E}[z_{it}] \mathbf{E}[z'_{it}] \right\|^2 \geq c \right) = o(N^{-1})$$

Combining part (i) and (ii) together, we have

$$P \left(\min_{\sigma \in \mathbb{S}_{pJ}} \sum_{k=1}^{pJ} \left| \mu_k(\hat{Q}_{i,\tilde{z}\tilde{z}}) - \mu_{\sigma(k)}(\text{Var}(z_{it})) \right|^2 \leq c \right) = 1 - o(N^{-1})$$

(ii) Let $\hat{Q}_{i,\tilde{z}\tilde{\delta}} = \frac{1}{T} \sum_{t=1}^T \tilde{z}_{it} \tilde{\delta}_{h_i,it}$, and $\hat{Q}_{i,\tilde{z}\tilde{u}} = \frac{1}{T} \sum_{t=1}^T \tilde{z}_{it} \tilde{u}_{it}$, where $\tilde{\delta}_{h_i,it} = \sum_{j=1}^p \tilde{\delta}_{h_i,j,it}$, then we have $\left\| \hat{Q}_{i,\tilde{z}\tilde{\delta}} \right\| \leq \left\| \hat{Q}_{i,\tilde{z}\tilde{\delta}} \right\| + \left\| \hat{Q}_{i,\tilde{z}\tilde{u}} \right\|$.

For the first part, since

$$\begin{aligned} & \left\| \hat{Q}_{i,\tilde{z}\tilde{\delta}} \right\| \\ &= \left\| \hat{Q}_{i,z\delta} - \frac{1}{T} \sum_{t=1}^T z_{it} \frac{1}{T} \sum_{t=1}^T \delta_{h_i,it} \right\| \\ &\leq \left\| \hat{Q}_{i,z\delta} \right\| + \left\| \frac{1}{T} \sum_{t=1}^T z_{it} \frac{1}{T} \sum_{t=1}^T \delta_{h_i,it} \right\| \end{aligned}$$

- For the first item,

$$\mathbf{E} \left[\left\| \hat{Q}_{i,z\delta} \right\|^2 \right] = \sum_{r=1}^p \mathbf{E} \left[\left\| \frac{1}{T} \sum_{t=1}^T \sqrt{J} B_{rit}^J{}' \delta_{h_i,it} \right\|^2 \right]$$

For any $1 \leq r \leq p$,

$$\begin{aligned} & \mathbf{E} \left[\left\| \frac{1}{T} \sum_{t=1}^T \sqrt{J} B_{rit}^J{}' \delta_{h_i,it} \right\|^2 \right] \\ &= \frac{1}{T^2} \sum_{t=1}^T \sum_{s=1}^T \mathbf{E} \left[J B_{rit}^J{}' B_{ris}^J \delta_{h_i,it} \delta_{h_i,is} \right] \\ &\leq \theta_{NT}^2 J \frac{1}{T^2} \sum_{t=1}^T \sum_{s=1}^T \mathbf{E} \left[B_{rit}^J{}' B_{ris}^J \right] \\ &= \theta_{NT}^2 J \mathbf{E} \left[\frac{1}{T} \sum_{t=1}^T B_{rit}^J{}' \frac{1}{T} \sum_{s=1}^T B_{rit}^J \right] \\ &= \theta_{NT}^2 J \sum_{j=1}^J \mathbf{E} \left[\frac{1}{T} \sum_{t=1}^T B_{rit,j}^J \frac{1}{T} \sum_{s=1}^T B_{rit,j}^J \right] \\ &= O \left(J^{-2r} \right) \end{aligned}$$

Thus $\mathbf{E} \left[\left\| \hat{Q}_{i,z\delta} \right\|^2 \right] = O \left(J^{-2r} \right)$.

- For the second item,

$$\begin{aligned} & \mathbf{E} \left[\left\| \frac{1}{T} \sum_{t=1}^T z_{it} \frac{1}{T} \sum_{t=1}^T \delta_{h_i,it} \right\|^2 \right] \\ &= \sum_{r=1}^p \mathbf{E} \left[\left\| \frac{1}{T} \sum_{t=1}^T \sqrt{J} B_{rit}^J \frac{1}{T} \sum_{t=1}^T \delta_{h_i,it} \right\|^2 \right] \end{aligned}$$

Similarly, we could get that

$$\mathbf{E} \left[\left\| \frac{1}{T} \sum_{t=1}^T z_{it} \frac{1}{T} \sum_{t=1}^T \delta_{h_i,it} \right\|^2 \right] = O \left(J^{-2r} \right)$$

For the second part, similarly

$$\begin{aligned}
& \|\hat{Q}_{i,\tilde{z}\tilde{u}}\| \\
&= \left\| \hat{Q}_{i,zu} - \frac{1}{T} \sum_{t=1}^T z_{it} \frac{1}{T} \sum_{t=1}^T u_{it} \right\| \\
&\leq \|\hat{Q}_{i,zu}\| + \left\| \frac{1}{T} \sum_{t=1}^T z_{it} \frac{1}{T} \sum_{t=1}^T u_{it} \right\|
\end{aligned}$$

- Consider the first item,

$$\mathbf{E} \left[\|\hat{Q}_{i,zu}\|^2 \right] = \sum_{r=1}^p \mathbf{E} \left[\left\| \frac{1}{T} \sum_{t=1}^T \sqrt{J} B_{rit}^{J'} u_{it} \right\|^2 \right]$$

For any $1 \leq r \leq p$,

$$\begin{aligned}
& \mathbf{E} \left[\left\| \frac{1}{T} \sum_{t=1}^T \sqrt{J} B_{rit}^{J'} u_{it} \right\|^2 \right] \\
&= \frac{1}{T^2} \sum_{t=1}^T \sum_{s=1}^T \mathbf{E} [J B_{rit}^{J'} B_{ris}^J u_{it} u_{is}] \\
&\leq \frac{CJ}{T^2} \sum_{t=1}^T \sum_{s=1}^T \mathbf{E} [u_{it} u_{is}] \\
&= O(T^{-1}J)
\end{aligned}$$

Thus $\mathbf{E} \left[\|\hat{Q}_{i,zu}\|^2 \right] = O(T^{-1}J)$.

- Consider the second item,

$$\begin{aligned}
& \mathbf{E} \left[\left\| \frac{1}{T} \sum_{t=1}^T z_{it} \frac{1}{T} \sum_{t=1}^T u_{it} \right\|^2 \right] \\
&= \sum_{r=1}^p \mathbf{E} \left[\left\| \frac{1}{T} \sum_{t=1}^T \sqrt{J} B_{rit}^J \frac{1}{T} \sum_{t=1}^T u_{it} \right\|^2 \right] \\
&= O(T^{-1})
\end{aligned}$$

Thus $\|\hat{Q}_{i,\bar{z}\bar{u}}\| = O_p(J^{\frac{1}{2}}T^{-\frac{1}{2}})$.

In sum, we have proved that

$$\|\hat{Q}_{i,ze}\| = O_p(J^{-r} + J^{\frac{1}{2}}T^{-\frac{1}{2}})$$

(iii) Consider

$$\begin{aligned}
& \mathbf{E} \left[\frac{1}{N} \sum_{i=1}^N \|\hat{Q}_{i,\bar{z}\bar{e}}\|^2 \right] \\
&= \frac{1}{N} \sum_{i=1}^N \mathbf{E} \left[\|\hat{Q}_{i,\bar{z}\bar{e}}\|^2 \right] \\
&\leq \frac{2}{N} \sum_{i=1}^N \left(\mathbf{E} \left[\|\hat{Q}_{i,\bar{z}\bar{\delta}}\|^2 \right] + \mathbf{E} \left[\|\hat{Q}_{i,\bar{z}\bar{u}}\|^2 \right] \right)
\end{aligned}$$

Note that from the proof of (ii), we could strengthen the results to

$$\begin{aligned}
\max_{1 \leq i \leq N} \mathbf{E} \left[\|\hat{Q}_{i,\bar{z}\bar{\delta}}\|^2 \right] &= O(J^{-2r}) \\
\max_{1 \leq i \leq N} \mathbf{E} \left[\|\hat{Q}_{i,\bar{z}\bar{u}}\|^2 \right] &= O(T^{-1}J)
\end{aligned}$$

Consequently,

$$\mathbf{E} \left[\frac{1}{N} \sum_{i=1}^N \|\hat{Q}_{i,ze}\|^2 \right] = O(J^{-2r} + T^{-1}J)$$

This completes the proof.

(iv) Note that $\|\hat{Q}_{i,\tilde{z}\tilde{e}}\| = \|\hat{Q}_{i,\tilde{z}\tilde{\delta}}\| + \|\hat{Q}_{i,\tilde{z}\tilde{u}}\|$. To prove (iv), we can show that for large enough $C > 0$, any $c > 0$ and any $v > 0$,

$$P\left(\max_{1 \leq i \leq N} \|\hat{Q}_{i,\tilde{z}\tilde{\delta}}\| \geq CJ^{-r}\right) = o(N^{-1})$$

$$P\left(\max_{1 \leq i \leq N} \|\hat{Q}_{i,\tilde{z}\tilde{u}}\| \geq cJ^{\frac{1}{2}}T^{-\frac{1}{2}}(\ln T)^{3+v}\right) = o(N^{-1})$$

(i) For the first part, consider $\|\hat{Q}_{i,z\delta}\|$ and $\left\|\frac{1}{T}\sum_{t=1}^T z_{it}\frac{1}{T}\sum_{t=1}^T \delta_{h_i,it}\right\|$ separately. First,

$$\begin{aligned} & \|\hat{Q}_{i,z\delta}\|^2 \\ &= \sum_{r=1}^p \left\| \frac{1}{T} \sum_{t=1}^T \sqrt{J} B_{rit}^J \delta_{h_i,it} \right\|^2 \\ &\leq \theta_{NT}^2 J \sum_{r=1}^p \sum_{j=1}^J \left(\frac{1}{T} \sum_{t=1}^T B_{rit,j}^J \right)^2 \end{aligned}$$

Consider $\frac{1}{T}\sum_{t=1}^T B_{rit,j}^J$, for any $c > 0$ and $1 \leq j \leq J$, we want to show

$$P\left(\max_{1 \leq r \leq p} \max_{1 \leq j \leq J} \max_{1 \leq i \leq N} \left| \frac{1}{T} \sum_{t=1}^T B_{rit,j}^J - \mathbf{E}[B_{rit,j}^J] \right| \geq cJ^{-1}\right) = o(N^{-1})$$

Since

$$\begin{aligned} & NP \left(\max_{1 \leq r \leq p} \max_{1 \leq j \leq J} \max_{1 \leq i \leq N} \left| \frac{1}{T} \sum_{t=1}^T B_{rit,j}^J - \mathbf{E}[B_{rit,j}^J] \right| \geq cJ^{-1} \right) \\ &\leq pN \sum_{i=1}^N \sum_{r=1}^p \sum_{j=1}^J P \left(\left| \frac{1}{T} \sum_{t=1}^T B_{rit,j}^J - \mathbf{E}[B_{rit,j}^J] \right| \geq cJ^{-1} \right) \\ &\leq pN^2 J \exp \left(- \frac{C_0 c^2 T^2 J^{-2}}{Tv_{0,\max} + 2 + 2cTJ^{-1}(\ln T)^2} \right) \end{aligned}$$

As long as $(\ln T)^3 JT^{-1} = o(1)$, we could get the result. Then for large enough

$C > 0$ and for any $1 \leq j \leq J$,

$$\begin{aligned}
& P \left(\max_{1 \leq r \leq p} \max_{1 \leq j \leq J} \max_{1 \leq i \leq N} \frac{1}{T} \sum_{t=1}^T B_{rit,j}^J \geq C J^{-1} \right) \\
& \leq P \left(\max_{1 \leq r \leq p} \max_{1 \leq j \leq J} \max_{1 \leq i \leq N} \frac{1}{T} \sum_{t=1}^T \mathbf{E} [B_{rit,j}^J] \right. \\
& \quad \left. + \max_{1 \leq r \leq p} \max_{1 \leq j \leq J} \max_{1 \leq i \leq N} \left| \frac{1}{T} \sum_{t=1}^T B_{rit,j}^J - \mathbf{E} [B_{rit,j}^J] \right| \geq C J^{-1} \right) \\
& = o(N^{-1})
\end{aligned}$$

Thus for large enough $C > 0$,

$$\begin{aligned}
& P \left(\max_{1 \leq i \leq N} J \sum_{r=1}^p \sum_{j=1}^J \left(\frac{1}{T} \sum_{t=1}^T B_{rit,j}^J \right)^2 \geq C^2 \right) \\
& \leq P \left(J^2 p \max_{1 \leq r \leq p} \max_{1 \leq j \leq J} \max_{1 \leq i \leq N} \left(\frac{1}{T} \sum_{t=1}^T B_{rit,j}^J \right)^2 \geq C^2 \right) \\
& \leq P \left(\max_{1 \leq r \leq p} \max_{1 \leq j \leq J} \max_{1 \leq i \leq N} \left(\frac{1}{T} \sum_{t=1}^T B_{rit,j}^J \right)^2 \geq C^2 J^{-2} p^{-1} \right) \\
& \leq P \left(\left(\max_{1 \leq r \leq p} \max_{1 \leq j \leq J} \max_{1 \leq i \leq N} \frac{1}{T} \sum_{t=1}^T B_{rit,j}^J \right)^2 \geq C^2 J^{-2} p^{-1} \right) \\
& = o(N^{-1})
\end{aligned}$$

Combining the previous results, we have for large enough $C > 0$

$$\begin{aligned}
& P \left(\max_{1 \leq i \leq N} \|\hat{Q}_{i,z\delta}\| \geq CJ^{-r} \right) \\
& \leq P \left(\max_{1 \leq i \leq N} \|\hat{Q}_{i,z\delta}\|^2 \geq C^2 J^{-2r} \right) \\
& \leq P \left(\theta_{NT}^2 \max_{1 \leq i \leq N} J \sum_{r=1}^p \sum_{j=1}^J \left(\frac{1}{T} \sum_{t=1}^T B_{rit,j}^J \right)^2 \geq C^2 J^{-2r} \right) \\
& \leq P \left(\max_{1 \leq i \leq N} J \sum_{r=1}^p \sum_{j=1}^J \left(\frac{1}{T} \sum_{t=1}^T B_{rit,j}^J \right)^2 \geq C \right) \\
& = o(N^{-1})
\end{aligned}$$

Similarly, we could prove that for large enough $C > 0$

$$P \left(\max_{1 \leq i \leq N} \left\| \frac{1}{T} \sum_{t=1}^T z_{it} \frac{1}{T} \sum_{t=1}^T \delta_{h_i,it} \right\| \geq CJ^{-r} \right) = o(N^{-1})$$

Thus $P \left(\max_{1 \leq i \leq N} \|\hat{Q}_{i,\tilde{z}\tilde{\delta}}\| \geq CJ^{-r} \right) = o(N^{-1})$.

(ii) For the second part, since

$$\begin{aligned}
& \|\hat{Q}_{i,\tilde{z}\tilde{u}}\| \\
& \leq \sum_{r=1}^p \left\| \frac{1}{T} \sum_{t=1}^T \sum_{r=1}^p \sqrt{J} B_{rit}^J u_{it} \right\| + \sum_{r=1}^p \left\| \frac{1}{T} \sum_{t=1}^T \sqrt{J} B_{rit}^J \frac{1}{T} \sum_{t=1}^T u_{it} \right\|
\end{aligned}$$

Consider $\left\| \frac{1}{T} \sum_{t=1}^T \sqrt{J} B_{rit}^J u_{it} \right\|$, By Lemma 2.1, for any $c > 0$ and $v > 0$,

$$P \left(\max_{1 \leq i \leq N} \left\| \frac{1}{T} \sum_{t=1}^T \sqrt{J} B_{rit}^J u_{it} \right\| \geq cJ^{\frac{1}{2}} T^{-\frac{1}{2}} (\ln T)^{3+v} \right) = o(N^{-1})$$

Similarly,

$$P \left(\max_{1 \leq i \leq N} \left\| \frac{1}{T} \sum_{t=1}^T \sqrt{J} B_{rit}^J \frac{1}{T} \sum_{t=1}^T u_{it} \right\| \geq c J^{\frac{1}{2}} T^{-\frac{1}{2}} (\ln T)^{3+v} \right) = o(N^{-1})$$

This completes the proof.

□

Chapter 3

A Network Sample Selection Model

3.1 Motivation

The problem of selection bias is pervasive whenever there is non-random sampling. Since Heckman (1974), there grows a large literature on how to correct for selection bias in various different models (e.g., Kyriazidou (1997), Greene (1994), Terza (1998), Das, Newey, and Vella (2003) and Newey (2009)). The most popular method is the Heckman Selection model (HSM), which includes two equations: the selection equation (or link formation equation in network data models) and the outcome equation. However, it remains a challenge to apply the HSM to network data models when bilateral fixed effects are introduced in the selection equation to control for unobserved heterogeneity from both sides. To fill this gap, I propose a network sample selection model in which 1) bilateral fixed effects enter the pairwise outcome equation additively; 2) link formation depends on latent variables from both sides nonparametrically.

The link formation equation follows Auerbach (2019). I assume that each knot is represented by a latent variable (discrete or continuous), deciding whether it forms a link with another knot. Then I use a statistic introduced by Auerbach (2019) to measure the distance between two different knots. In the outcome equation, I introduce a four-cycle structure to difference out additive bilateral fixed effects. Using the distance statistics from the link formation equation, I use the kernel function to control for selection bias.

My paper is closely related to two strands of literature: sample selection and network formation. In the former literature, Heckman develops the HSM in a series of papers (Heckman (1974), Heckman (1976), Heckman (1979) and Heckman (1990)). Ahn and Powell (1993), Powell (1994), Kyriazidou (1997), Andrews and Schafgans (1998) and Newey (2009) generalize HSM to semiparametric models and Das, Newey, and Vella (2003) considers a nonparametric version. My paper is close to Ahn and Powell (1993) and Kyriazidou (1997) in the sense that we both difference out fixed effects and use kernel function to control for selection bias, but their methods are not applicable directly to network selection models when bilateral fixed effects are present.

In the network formation literature, I mainly discuss random utility models. To construct networks, there are several different ways of modeling: from subgraphs (e.g, Chandrasekhar and Jackson (2014)); to consider strategic interactions between links(e.g., Ridder and Sheng (2015), Sheng (2018), De Paula, Richards-Shubik, and Tamer (2018), Menzel (2015), Jackson and Wolinsky (1996), Bala and Goyal (2000), Jackson (2008) Goldsmith-Pinkham and Imbens (2013)); to assume that different links are conditional independent (e.g., Auerbach (2019), Graham (2017), Candelaria (2016) and Chernozhukov, Fernandez-Val, and Weidner (2018)). My paper falls into the last category.

In addition, there are also several papers that develop sample selection models using network data, like Johnsson and Moon (2017), Hsieh and Lee (2016) and Chernozhukov, Fernández-Val, and Luo (2018).

My contributions are three-fold. First, I introduce a fully nonparametric link formation model and study pairwise outcomes. Whether two knots are connected is purely decided by some unobserved latent variables. Fewer parametric or functional form assumptions are required, thus avoiding as much model specification bias as possible. The model is used in Auerbach (2019) and discussed in Johnsson and Moon (2017), but I first apply it to the studies of pairwise outcomes. Furthermore, I generalize this model to the directed network, which is new in the literature.

Second, I contribute to the partially linear model literature by using a novel four-cycle structure in network data models. I explore this structure to difference out additive bilateral fixed effects. Although Graham (2017) also considered a similar structure, I use it in a very different way.

Third, compared with the traditional approach dealing with two-way fixed effects in the link formation equation, I no longer need to deal with the incidental parameter problem encountered in Fernández-Val and Weidner (2016), Chernozhukov, Fernandez-Val, and Weidner (2018), and Chernozhukov, Fernández-Val, and Luo (2018), making the analysis much more straightforward.

To further illustrate the applications of my method, I discuss the following example.

Example: Determinants of Trade Flows

To study the determinants of trade flows between countries is important. (for instance, Helpman, Melitz, and Rubinstein (2008), Rose (2004) and Haveman and Hummels) It is naturally a network problem where knots are different countries and links are pairwise imports and exports between them. However, only about 50% of all country pairs are with non-zero trade flows (See Figure 3.1 from Helpman, Melitz, and Rubinstein (2008)).

Figure 3.1: Trade Flows

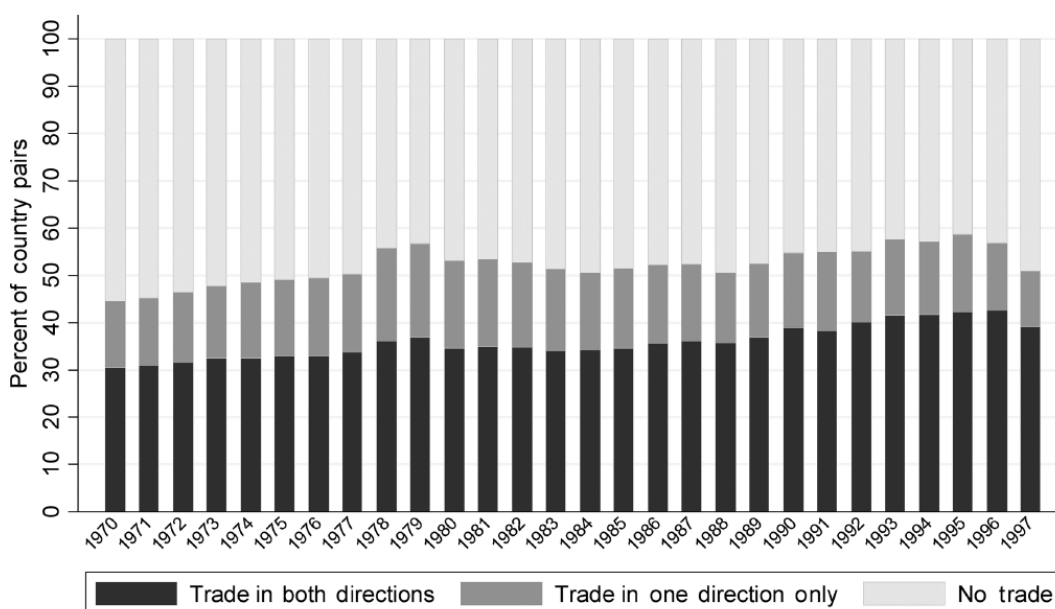


FIGURE I

Distribution of Country Pairs Based on Direction of Trade

Note. Constructed from 158 countries.

Whether two countries trade with each other is a mutually rational choice. Thus it is crucial to correct for the selection bias. Simultaneously, to control for unobserved heterogeneity from both sides, it is natural to add bilateral fixed effects to both the link formation and outcome equations.

Outline: The rest of the paper is organized as follows. Section 3.2 discusses the model. Section 3.3 presents the estimation strategy. Section 3.4 studies the asymptotic properties of the estimator. Section 3.5 extends the model to directed networks. Section 3.6 concludes.

All proofs of the main results are given in the Appendix.

Notation: There are $i = 1, 2, \dots, N$ agents (individuals, households, firms, countries, etc) randomly sampled from the population and $n = \binom{N}{2} = \frac{1}{2}N(N - 1)$ pairs (dyads). $D_{ij} = 1$ if i and j are connected and 0 otherwise. I assume the links are undirected, and there is no self-connection. The $N \times N$ adjacent matrix is denoted as \mathbf{D} . I consider only dense networks where $P(D_{ij} = 1) = \rho_0 > 0$.

3.2 Model Setup

In Helpman, Melitz, and Rubinstein (2008), the model setting is:

$$D_{ij} = \mathbf{1}\{z'_{ij}\gamma + A_i + A_j + \eta_{ij} \geq 0\}$$

$$y_{ij} = \begin{cases} x'_{ij}\beta + B_i + B_j + \varepsilon_{ij} & \text{if } D_{ij} = 1 \\ 0 & \text{if otherwise} \end{cases}$$

where D_{ij} indicates whether two countries i and j trade with each other, y_{ij} is the size of trade flows between country i and j . In the link formation equation, z_{ij} are pairwise characteristics; A_i and A_j are bilateral fixed effects and η_{ij} is the error term. x_{ij} , B_i , B_j and ε_{ij} are similarly defined.

The goal is to estimate β .

The two-step approach is usually used: 1) estimate the link formation equation and get $\hat{\gamma}$, \hat{A}_i and \hat{A}_j . Then let $\hat{\theta} \equiv z'_{ij}\hat{\gamma} + \hat{A}_i + \hat{A}_j$. (See Graham (2017) with logistic error term and Chernozhukov, Fernandez-Val, and Weidner (2018) on distribution regression.) 2) the outcome equation could be expressed as

$$\begin{aligned} \mathbf{E}[y_{ij}|D_{ij} = 1] &= \mathbf{E}[x'_{ij}\beta + B_i + B_j|D_{ij} = 1] + \mathbf{E}[\varepsilon_{ij}|D_{ij}=1] \\ &= \mathbf{E}[x'_{ij}\beta + B_i + B_j|D_{ij} = 1] + f(\hat{\theta}) \end{aligned}$$

However, there are two main issues with this approach. First, the link formation equation follows a parametric form, and certain particular distribution assumptions need to be made for η_{ij} . (Normal or Logistic distributions are two common choices.). Second, the estimation of A_i suffers an incidental parameter problem and is hard to deal with.

To fix the potential issues in the traditional approach, I propose an alternative model where the link formation is purely nonparametric, and the incidental parameter problem is no longer an issue. The model I consider is:

$$D_{ij} = \mathbf{1}\{\eta_{ij} \leq f(\omega_i, \omega_j)\}$$

$$y_{ij} = \begin{cases} x'_{ij}\beta + B_i + B_j + \varepsilon_{ij} & \text{if } D_{ij} = 1 \\ 0 & \text{if otherwise} \end{cases}$$

where ω_i , ω_j and η_{ij} follow standard uniform distribution, which is a harmless normalization. $f : [0, 1]^2 \rightarrow [0, 1]$ is Lebesgue-measurable and symmetric in its arguments. ε_{ij} and η_{ij} are correlated, causing sample selection bias.

3.2.1 Explanation of the Link Formation Process

I will use international trade as an example to better explain the formation process. Please step back and rethink it.

- (i) Suppose US and UK trade with the same countries, from the perspective of the formation, they are the same. See Figure 3.2, where 1, 2, 3 denote different countries.
- (ii) Add some randomness, if US and UK trade with different countries with the same probability, they are the same. I denote that they are of [the same type](#). See Figure 3.3.

Figure 3.2: Trade with Same Countries

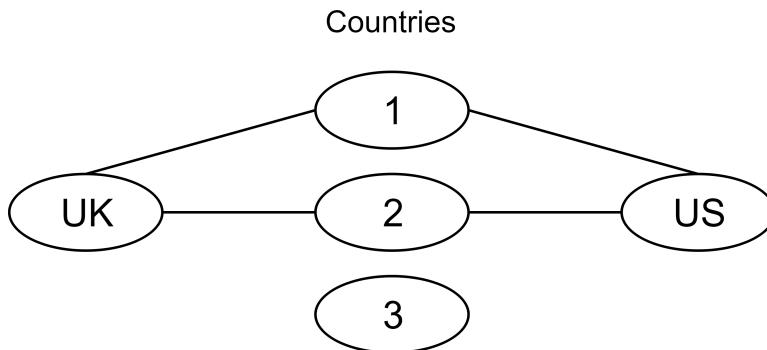
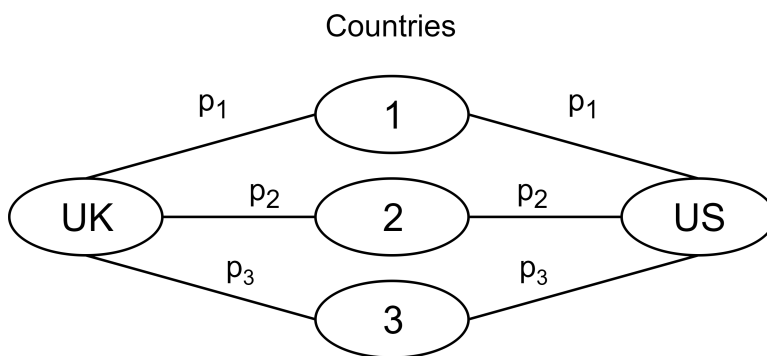
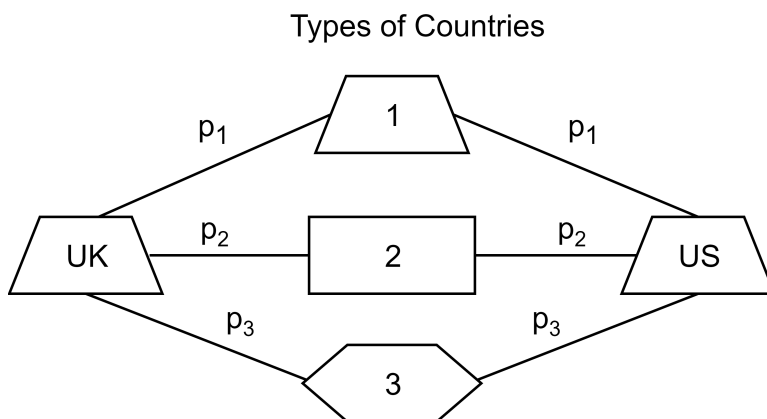


Figure 3.3: Trade with Same Countries with Same Probabilities



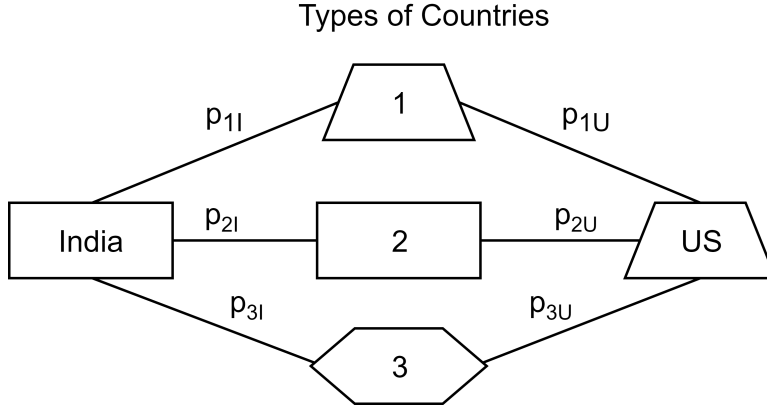
(iii) Thus, I could classify different countries into different types, which are indicated by different shapes. See Figure 3.4.

Figure 3.4: Countries of the Same Type



(iv) On the contrary, countries of different types would trade with the same type of countries with different probabilities. See Figure 3.5, where $p_{1I} \neq p_{1U}$ or $p_{2I} \neq p_{2U}$ or $p_{3I} \neq p_{3U}$.

Figure 3.5: Countries of Different Types



Mathematically, I formalize this intuition as

$$D_{ij} = \mathbf{1}\{\eta_{ij} \leq f(\omega_i, \omega_j)\}$$

3.3 Estimation Strategy

To estimate the model, I first introduce some assumptions on the model setup.

Assumption 3.1. (i) *The random sequence $\{x_{ij}, \varepsilon_{ij}\}$ are independent and identically distributed. $\{B_i\}$ are i.i.d. $\{B_i\}$ and $\{x_{ij}\}$ are independent of $\{\varepsilon_{ij}\}$.*

(ii) *η_{ij} and ω_i follow standard uniform marginals. $\{\eta_{ij}\}$ and $\{\omega_i\}$ are independent.*

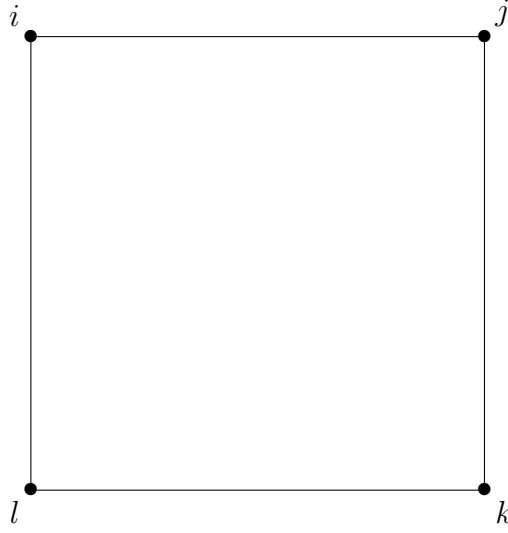
(iii) *ε_{ij} and η_{mn} are independent if $(i, j) \neq (m, n)$.*

(iv) *$f : [0, 1]^2 \rightarrow [0, 1]$ is Lebesgue-measurable and symmetric in its arguments.*

I consider a four-cycle structure demonstrated in Figure 3.6:

Let $T_{ijkl} = D_{ij}D_{jk}D_{kl}D_{il}$, then

Figure 3.6: Four Cycle



$$\mathbf{E}[y_{ij}|T_{ijkl=1}] = \mathbf{E}[x'_{ij}\beta + B_i + B_j + \varepsilon_{ij}|T_{ijkl} = 1]$$

$$\mathbf{E}[y_{jk}|T_{ijkl=1}] = \mathbf{E}[x'_{jk}\beta + B_j + B_k + \varepsilon_{jk}|T_{ijkl} = 1]$$

$$\mathbf{E}[y_{kl}|T_{ijkl=1}] = \mathbf{E}[x'_{kl}\beta + B_k + B_l + \varepsilon_{kl}|T_{ijkl} = 1]$$

$$\mathbf{E}[y_{il}|T_{ijkl=1}] = \mathbf{E}[x'_{il}\beta + B_i + B_l + \varepsilon_{il}|T_{ijkl} = 1]$$

Using Assumption 3.1, I have that

$$\mathbf{E}[\varepsilon_{ij}|T_{ijkl} = 1] = \mathbf{E}[\varepsilon_{ij}|D_{ij} = 1]$$

Define $\Lambda(f(\omega_i, \omega_j)) \equiv \mathbf{E}[\varepsilon_{ij}|D_{ij} = 1] = \mathbf{E}[\varepsilon_{ij}|\eta_{ij} \leq f(\omega_i, \omega_j)]$. To estimate β , there are three steps.

Step 1: I difference out the fixed effects B . Consider

$$\begin{aligned}
& \mathbf{E}[y_{ij} + y_{kl} - y_{jk} - y_{il} | T_{ijkl} = 1] \\
&= \mathbf{E}[(x_{ij} + x_{kl} - x_{jk} - x_{il})' \beta | T_{ijkl} = 1] \\
&\quad + (\Lambda(f(\omega_i, \omega_j)) - \Lambda(f(\omega_j, \omega_k))) \\
&\quad + (\Lambda(f(\omega_k, \omega_l)) - \Lambda(f(\omega_i, \omega_l)))
\end{aligned}$$

Define

$$y_{ij,kl} = y_{ij} + y_{kl} - y_{jk} - y_{il}$$

$$x_{ij,kl} = x_{ij} + x_{kl} - x_{jk} - x_{il}$$

Then I have

$$\begin{aligned}
& \mathbf{E}[y_{ij,kl} | T_{ijkl} = 1] \\
&= \mathbf{E}[x'_{ij,kl} \beta | T_{ijkl} = 1] \\
&\quad + (\Lambda(f(\omega_i, \omega_j)) - \Lambda(f(\omega_j, \omega_k))) \\
&\quad + (\Lambda(f(\omega_k, \omega_l)) - \Lambda(f(\omega_i, \omega_l)))
\end{aligned}$$

Step 2 : To control for the selection bias, I want to find i and k similar enough such that $(\Lambda(f(\omega_i, \omega_j)) - \Lambda(f(\omega_j, \omega_k))) \approx 0$ and $(\Lambda(f(\omega_k, \omega_l)) - \Lambda(f(\omega_i, \omega_l))) \approx 0$. (Note that f is symmetric in its arguments.)

I utilize the result from Auerbach (2019) and define average agent degree and average codegree.

Definition. *Average agent degree:*

$$\frac{1}{N-1} \sum_{i \neq j} D_{ij}$$

Then the agent's population degree is $\int f_{\omega_i}(\tau) d\tau$. Define $f_{\omega_i}(\cdot) := f(\omega_i, \cdot)$.

Definition. *Average codegree of i and j :*

$$\frac{1}{N-2} \sum_{t \neq i, j} D_{it} D_{jt}$$

The population codegree of i and j : $p(\omega_i, \omega_j) = \int f_{\omega_i}(\tau) f_{\omega_j}(\tau) d\tau$. Define $p_{\omega_i}(\cdot) := p(\omega_i, \cdot)$.

Mathematically, I use $\|f_{\omega_i} - f_{\omega_k}\|_2$ to measure the similarity between i and k . If $\|f_{\omega_i} - f_{\omega_k}\|_2$ is small enough,

$$(\Lambda(f(\omega_i, \omega_j)) - \Lambda(f(\omega_j, \omega_k))) + (\Lambda(f(\omega_k, \omega_l)) - \Lambda(f(\omega_i, \omega_l)))$$

is negligible.

Now the target is to find a statistic to bound $\|f_{\omega_i} - f_{\omega_k}\|_2$. It turns out that

$$\|p_{\omega_i} - p_{\omega_k}\|_2 = 0 \implies \|f_{\omega_i} - f_{\omega_k}\|_2 = 0$$

The sample analogue of $\|p_{\omega_i} - p_{\omega_k}\|_2$ is

$$\hat{\delta}_{ik} = \left(\frac{1}{N} \sum_{t=1}^N \left(\frac{1}{N-2} \sum_{s=1}^N D_{ts} (D_{is} - D_{ks}) \right)^2 \right)^{1/2}$$

which is used to measure the distance between i and k .

Step 3: Define

$$\varepsilon_{ij,kl} = \varepsilon_{ij} + \varepsilon_{kl} - \varepsilon_{jk} - \varepsilon_{il}$$

Thus the estimator could be constructed as

$$\begin{aligned} \hat{\beta} &= \left(\sum_{i<j<k<l} \frac{1}{4!} \sum_{\pi \in \Pi_4} T_{ijkl} x_{ij,kl} x'_{ij,kl} K\left(\frac{\hat{\delta}_{ik}}{h_N}\right) \right)^{-1} \\ &\quad \left(\sum_{i<j<k<l} \frac{1}{4!} \sum_{\pi \in \Pi_4} T_{ijkl} x_{ij,kl} y_{ij,kl} K\left(\frac{\hat{\delta}_{ik}}{h_N}\right) \right) \\ &= \beta + \left(\binom{N}{4}^{-1} \sum_{i<j<k<l} \frac{1}{4!} \sum_{\pi \in \Pi_4} T_{ijkl} x_{ij,kl} x'_{ij,kl} K\left(\frac{\hat{\delta}_{ik}}{h_N}\right) \right)^{-1} \\ &\quad \left(\binom{N}{4}^{-1} \sum_{i<j<k<l} \frac{1}{4!} \sum_{\pi \in \Pi_4} T_{ijkl} x_{ij,kl} \varepsilon_{ij,kl} K\left(\frac{\hat{\delta}_{ik}}{h_N}\right) \right) \end{aligned}$$

where Π_4 is a permutation; $K(\cdot)$ is a kernel function and h_N is the bandwidth.

3.4 Asymptotic Properties

This section includes three subsections. Subsection 3.4.1 studies the consistency of the estimator. Subsection 3.4.2 and 3.4.3 discusses the asymptotic distribution of the estimator when ω_i is discrete or continuous, respectively.

3.4.1 Consistency

More assumptions are required.

Assumption 3.2. (i) x_{ij} , ε_{ij} both have finite fourth moments.

(ii) $\mathbf{E}[T_{ijkl} x_{ij,kl} x'_{ij,kl} \mid \|f_{\omega_i} - f_{\omega_k}\|_2 = 0] = \Gamma_0$ is positive definite.

(iii) $\mathbf{E}[T_{ijkl} x_{ij,kl} \varepsilon_{ij,kl} \mid \|f_{\omega_i} - f_{\omega_k}\|_2 = 0] = 0$.

Assumption 3.3. (i) $\lim_{h \rightarrow 0} \mathbf{E}[T_{ijkl} x_{ij,kl} x'_{ij,kl} \mid \|f_{\omega_i} - f_{\omega_k}\|_2 = h] = \Gamma_0$ is positive definite.

(ii) $\lim_{h \rightarrow 0} \mathbf{E}[T_{ijkl} x_{ij,kl} \varepsilon_{ij,kl} \mid \|f_{\omega_i} - f_{\omega_k}\|_2 = h] = 0$.

The following assumption imposes restrictions on the kernel function and bandwidth.

Assumption 3.4. (i) The bandwidth sequences $h_N \rightarrow 0$, $N^{1-\gamma} h_N^2 \rightarrow \infty$ for some $\gamma > 0$ as $N \rightarrow \infty$.

(ii) Let

$$r_N = \mathbf{E} \left[K \left(\frac{\|p_{\omega_i} - p_{\omega_k}\|_2}{h_N} \right) \right]$$

and $Nr_N \rightarrow \infty$ as $N \rightarrow \infty$.

(iii) K is supported, bounded, differentiable on $[0, 1]$, strictly positive on $[0, 1)$ and has bounded first derivative.

Remark. (i) The kernel functions could be Epanechnikov, Quartic, Triweight, etc, and I restrict to the positive part.

(ii) $Nr_n \rightarrow \infty$ implies that the number of pairs "close" enough increases with N .

(iii) Assumptions 3.2, 3.3, 3.4 together imply that if I randomly draw 4 agents i, j, k, l , the probability that 1) $T_{ijkl} = 1$ and 2) i and k are close enough is large than 0.

I need to utilize two lemmas from Auerbach (2019).

Lemma 3.1. Under Assumption 3.1 and 3.4, I have

$$\max_{i \neq k} |\hat{\delta}_{ik} - \|p_{\omega_i} - p_{\omega_j}\|_2| = o_{a.s.}(n^{-\gamma/4} h_N)$$

Lemma 3.2. Under Assumption 3.1, I have that $\forall \varepsilon > 0$, there exists a $\delta > 0$ such that with probability $1 - \varepsilon^2/4$

$$\|p_{\omega_i} - p_{\omega_k}\|_2 \leq \delta \rightarrow \|f_{\omega_i} - f_{\omega_k}\|_2 \leq \varepsilon$$

They formalize the intuition that

$$\hat{\delta}_{ik} \xrightarrow{\text{Lemma 1}} \|p_{\omega_i} - p_{\omega_j}\|_2 \xrightarrow{\text{Lemma 2}} \|f_{\omega_i} - f_{\omega_k}\|_2$$

The following theorem demonstrates the consistency of the estimator.

Theorem 3.1. *Suppose Assumptions 3.1, 3.2, 3.3, 3.4 hold, $\hat{\beta} \rightarrow \beta$ in probability.*

3.4.2 Asymptotic Distribution when ω_i has Finite Support

An additional technical assumption is required to study the asymptotic distribution of the estimator.

Assumption 3.5. $x_{ij} = g(\mathbf{X}_i, \mathbf{X}_j)$, where g is symmetric and Lebesgue-measurable.

The following theorem establishes the asymptotic distribution of β when ω is discrete.

Theorem 3.2. *Suppose Assumptions 3.1-3.5 hold, using Theorem 1.1 of Chatterjee et al. (2006) and Theorem 1 of Graham (2017), when ω_i is finite, I can get*

$$\frac{\sqrt{n\alpha_{2,N}^{-1}}c'(\hat{\beta} - \beta)}{\sqrt{c'\Gamma_0^{-1}\tilde{\Delta}_N\Gamma_0^{-1}c}} \xrightarrow{d} N(0, 36)$$

for any $k \times 1$ vector of real constants c , where

$$\tilde{\Delta}_N = \frac{1}{n} \sum_{i=1}^n \tilde{\Delta}_i$$

$$\tilde{\Delta}_i = \alpha_{2,N}^{-1} \mathbf{E}[\bar{u}_i \bar{u}_i' \mid \mathbf{X}, \boldsymbol{\omega}]$$

$$\bar{u}_i = \mathbf{E}[r_N^{-1} u_{ijkl} \mid X_i, X_j, \omega_i, \omega_j, \varepsilon_{ij}]$$

3.4.3 Asymptotic Distribution when ω_i is Continuous

When ω_i is Continuous, it becomes more challenging, and more technical assumptions are required.

Assumption 3.6. *There exists an integer K and a partition of $[0, 1)$ into K equally spaced, adjacent and non-intersecting intervals $\cup_{t=1}^K [x_t^1, x_t^2)$ with $x_1^1 = 0$ and $x_K^2 = 1$ such that for any $t \in \{1, \dots, K\}$ and almost every $x, y \in [x_t^1, x_t^2)$ and $s \in [0, 1]$, $|f(x, s) - f(y, s)| \leq C_6|x - y|^\alpha$, for some $C_6 \geq 0$ and $\alpha > 0$.*

Assumption 3.6 imposes more structure restrictions on f .

Assumption 3.7. $\mathbf{E}[T_{ijkl} x_{ij,kl} \varepsilon_{ij,kl} \mid \|f_{\omega_i} - f_{\omega_k}\|_2 = h] \leq C_7 h^\zeta$ for some $C_7, \zeta > 0$ and for all h in a small neighborhood to the right of 0.

Assumption 3.7 imposes more restrictions on the conditional expectation.

More regularity conditions on the bandwidth sequence and the kernel function are also needed.

Assumption 3.8. *The bandwidth sequence $h_n = C_8 \times n^{-\rho}$ for $\rho \in (\frac{\alpha}{4+8\alpha}, \frac{\alpha}{2+4\alpha})$ and some $C_8 > 0$. $K(\sqrt{u})$ is supported, bounded, and twice differentiable on $[0, 1]$, and strictly positive on $[0, 1)$.*

The following theorem establishes the asymptotic distribution of β when ω is continuous.

Theorem 3.3. *Suppose Assumptions 3.1-3.3 and 3.5-3.8 hold. Further assume that $a \times \zeta > 1/2$, using the theorem 1.1 of Chatterjee et al. (2006) and Theorem 1 of Graham (2017), when ω_i is continuous, I can get*

$$\frac{\sqrt{n\alpha_{2,N}^{-1}}c'(\hat{\beta} - \beta_{h_n})}{\sqrt{c'\Gamma_0^{-1}\tilde{\Delta}_N\Gamma_0^{-1}c}} \xrightarrow{d} N(0, 36)$$

for any $k \times 1$ vector of real constants c , where

$$\beta_{h_n} = \beta + \frac{1}{\Gamma_0 r_N} \mathbf{E} \left[\underset{128}{T_{ijkl} x_{ij,kl} \varepsilon_{ij,kl} K\left(\frac{\delta_{ik}}{h_N}\right)} \right]$$

Remark. *There are several methods in the literature to deal with the bias brought by the kernel method. First, one could introduce additional smoothness conditions and use jackknife bias correction. Second, I could choose slight smaller h to get a consistent estimator while sacrificing some efficiency. This is not the focus of the paper and thus omitted.*

3.5 Extension to Directed Networks

Our approach could be easily extended to the directed networks, which is of practical importance. The model should be revised as

$$D_{i \rightarrow j} = \mathbf{1}\{\eta_{ij} \leq f(\omega_i, \omega_j)\}$$

$$y_{i \rightarrow j} = \begin{cases} x'_{ij}\beta + B_{i \rightarrow} + B_{j \leftarrow} + \varepsilon_{ij} & \text{if } D_{i \rightarrow j} = 1 \\ 0 & \text{if otherwise} \end{cases}$$

where $D_{i \rightarrow j} = 1$ if i exports to j ; $f : [0, 1]^2 \rightarrow [0, 1]$ is not symmetric anymore; $y_{i \rightarrow j}$ is the trade flow from i to j ; $B_{i \rightarrow}$ is country i 's export fixed effect and $B_{j \leftarrow}$ is country j 's import fixed effect.

The definitions should be changed accordingly.

Definition. *Average out degree:*

$$\frac{1}{N-1} \sum_{j \neq i} D_{i \rightarrow j}$$

Define $f_{\omega_i \rightarrow}(\cdot) := f(\omega_i, \cdot)$, and the agent's population out degree is $\int f_{\omega_i \rightarrow}(\tau) d\tau$.

Definition. *Average in degree:*

$$\frac{1}{N-1} \sum_{j \neq i} D_{j \rightarrow i}$$

Define $f_{\omega_i \leftarrow}(\cdot) := f(\cdot, \omega_i)$, and the agent's population out degree: $\int f_{\omega_i \leftarrow}(\tau) d\tau$.

Definition. Average out codegree of i and j :

$$\frac{1}{N-2} \sum_{t \neq i, j} D_{i \rightarrow t} D_{j \rightarrow t}$$

The population out codegree of i and j : $p_{\rightarrow}(\omega_i, \omega_j) = \int f_{\omega_i \rightarrow}(\tau) f_{\omega_j \rightarrow}(\tau) d\tau$. Define $p_{\omega_i \rightarrow}(\cdot) := p(\omega_i, \cdot)$.

Definition. Average in codegree of i and j :

$$\frac{1}{N-2} \sum_{t \neq i, j} D_{t \rightarrow i} D_{t \rightarrow j}$$

The population in codegree of i and j : $p_{\leftarrow}(\omega_i, \omega_j) = \int f_{\omega_i \leftarrow}(\tau) f_{\omega_j \leftarrow}(\tau) d\tau$. Define $p_{\omega_i \leftarrow}(\cdot) := p(\omega_i, \cdot)$.

To measure the similarities of i and k , I use the following two different strategies:

- Use $\|p_{\omega_i \rightarrow} - p_{\omega_k \rightarrow}\|_2$ to bound $\|f_{\omega_i \rightarrow} - f_{\omega_k \rightarrow}\|_2$.
- Use $\|p_{\omega_i \leftarrow} - p_{\omega_k \leftarrow}\|_2$ to bound $\|f_{\omega_i \leftarrow} - f_{\omega_k \leftarrow}\|_2$.

The structure needs some modifications as well. Instead of a non-directed four-cycle, I use a directed one.

The construction of the estimator would be:

Let $T_{ij,kl} = D_{i \rightarrow j} D_{k \rightarrow j} D_{i \rightarrow l} D_{k \rightarrow l}$.

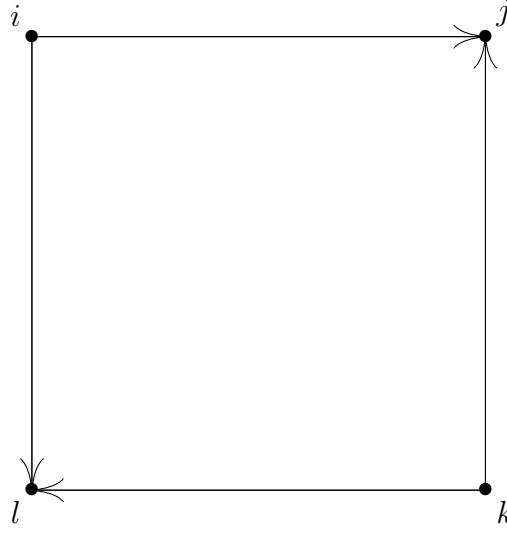
$$\mathbf{E}[y_{i \rightarrow j} | T_{ij,kl} = 1] = x'_{ij} \beta + B_{i \rightarrow} + B_{j \leftarrow} + \Lambda(f(\omega_i, \omega_j))$$

$$\mathbf{E}[y_{k \rightarrow j} | T_{ij,kl} = 1] = x'_{kj} \beta + B_{k \rightarrow} + B_{j \leftarrow} + \Lambda(f(\omega_k, \omega_j))$$

$$\mathbf{E}[y_{k \rightarrow l} | T_{ij,kl} = 1] = x'_{kl} \beta + B_{k \rightarrow} + B_{l \leftarrow} + \Lambda(f(\omega_k, \omega_l))$$

$$\mathbf{E}[y_{i \rightarrow l} | T_{ij,kl} = 1] = x'_{il} \beta + B_{i \rightarrow} + B_{l \leftarrow} + \Lambda(f(\omega_i, \omega_l))$$

Figure 3.7: Directed Four Cycle



Then

$$\begin{aligned}
 & \mathbf{E}[y_{i \rightarrow j} + y_{k \rightarrow l} - y_{k \rightarrow j} - y_{i \rightarrow l} | T_{ij,kl} = 1] \\
 &= (x_{ij} + x_{kl} - x_{kj} - x_{il})' \beta \\
 & \quad + (\Lambda(f(\omega_i, \omega_j)) - \Lambda(f(\omega_k, \omega_j))) \\
 & \quad + (\Lambda(f(\omega_k, \omega_l)) - \Lambda(f(\omega_i, \omega_l)))
 \end{aligned}$$

One could use $\|f_{\omega_i \rightarrow} - f_{\omega_k \rightarrow}\|_2$ to measure the similarity of ω_i and ω_k as exporters.

Alternatively, rearrange the above equation, I get that

$$\begin{aligned}
 & \mathbf{E}[y_{i \rightarrow j} + y_{k \rightarrow l} - y_{k \rightarrow j} - y_{i \rightarrow l} | T_{ij,kl} = 1] \\
 &= (x_{ij} + x_{kl} - x_{kj} - x_{il})' \beta \\
 & \quad + (\Lambda(f(\omega_i, \omega_j)) - \Lambda(f(\omega_i, \omega_l))) \\
 & \quad + (\Lambda(f(\omega_k, \omega_l)) - \Lambda(f(\omega_k, \omega_j)))
 \end{aligned}$$

which means I could also use $\|f_{\omega_j \leftarrow} - f_{\omega_l \leftarrow}\|_2$ to measure the similarity of ω_j and ω_l as

importers.

The following procedures are very similar and thus omitted.

3.6 Conclusion

I propose a network sample selection model. In this model, the link formation depends on two latent variables from both sides nonparametrically. Using the statistics offered by Auerbach (2019), I could measure the distance between two knots. Then in the outcome equation, I propose a four-cycle structure to difference out bilateral fixed effects. At the same time, I use kernel function to control for selection bias. The asymptotic properties of the estimators are studied. Finally, I extend the model to directed networks.

Appendix

Proof of Theorem 3.1

Proof. The estimator is constructed as:

$$\begin{aligned}
 \hat{\beta} &= \left(\sum_{i<j<k<l} \frac{1}{4!} \sum_{\pi \in \Pi_4} T_{ijkl} x_{ij,kl}^2 K\left(\frac{\hat{\delta}_{ik}}{h_N}\right) \right)^{-1} \\
 &\quad \left(\sum_{i<j<k<l} \frac{1}{4!} \sum_{\pi \in \Pi_4} T_{ijkl} x_{ij,kl} y_{ij,kl} K\left(\frac{\hat{\delta}_{ik}}{h_N}\right) \right) \\
 &= \beta + \left(\binom{N}{4}^{-1} \sum_{i<j<k<l} \frac{1}{4!} \sum_{\pi \in \Pi_4} T_{ijkl} x_{ij,kl}^2 K\left(\frac{\hat{\delta}_{ik}}{h_N}\right) \right)^{-1} \\
 &\quad \left(\binom{N}{4}^{-1} \sum_{i<j<k<l} \frac{1}{4!} \sum_{\pi \in \Pi_4} T_{ijkl} x_{ij,kl} \varepsilon_{ij,kl} K\left(\frac{\hat{\delta}_{ik}}{h_N}\right) \right)
 \end{aligned}$$

Then

$$\begin{aligned}
 \hat{\beta} &= \beta + \left(\binom{N}{4}^{-1} \sum_{i<j<k<l} \frac{1}{4!} \sum_{\pi \in \Pi_4} T_{ijkl} x_{ij,kl}^2 K\left(\frac{\hat{\delta}_{ik}}{h_N}\right) \right)^{-1} \\
 &\quad \left(\binom{N}{4}^{-1} \sum_{i<j<k<l} \frac{1}{4!} \sum_{\pi \in \Pi_4} T_{ijkl} x_{ij,kl} \varepsilon_{ij,kl} K\left(\frac{\hat{\delta}_{ik}}{h_N}\right) \right) \\
 &= \beta + \left(\binom{N}{4}^{-1} \sum_{i<j<k<l} \hat{v}_{ijkl} \right)^{-1} \left(\binom{N}{4}^{-1} \sum_{i<j<k<l} \hat{u}_{ijkl} \right) \\
 &= \beta + \hat{V}_N^{-1} \hat{U}_N
 \end{aligned}$$

Remark: For these items with replacing $\hat{\delta}_{ik}$ with $\|p_{\omega_i} - p_{\omega_k}\|_2$, I similarly define v_{ijkl} , V_N , u_{ijkl} and U_N .

The estimator could be simplified as

$$\hat{\beta} = \beta + \hat{V}_N^{-1} \hat{U}_N \approx \beta + V_N^{-1} U_N$$

To prove that $\hat{\beta}$ is consistent, I just need to prove that

•

$$\frac{1}{r_N} V_N - \frac{1}{r_N} \mathbf{E} \left[T_{ijkl} x_{ij,kl}^2 K \left(\frac{\delta_{ik}}{h_N} \right) \right] \xrightarrow{p} 0$$

•

$$\frac{1}{r_N} U_N - \frac{1}{r_N} \mathbf{E} \left[T_{ijkl} x_{ij,kl} \varepsilon_{ij,kl} K \left(\frac{\delta_{ik}}{h_N} \right) \right] \xrightarrow{p} 0$$

where

$$r_N = E \left[K \left(\frac{\|p_{\omega_i} - p_{\omega_k}\|_2}{h_N} \right) \right]$$

By Chebyshev's inequality,

$$P(|X - \mathbb{E}[X]| > \varepsilon) \leq \frac{\text{Var}(X)}{\varepsilon^2}$$

It is sufficient to prove that $\frac{1}{r_N^2} \text{Var}(U_N) \rightarrow 0$ and $\frac{1}{r_N^2} \text{Var}(V_N) \rightarrow 0$ as $N \rightarrow \infty$.

Consider $\text{Var}(U_N)$, define

$$\Delta_{q,N} = \text{Cov}(r_N^{-1} u_{ijkl}, r_N^{-1} u_{mnop})$$

where $\{i, j, k, l\}$ and $\{m, n, o, p\}$ have q agents in common.

Observation: $\Delta_{0,N} = 0$.

So as long as $\Delta_{q,N} < \infty$, $q = 1, 2, 3, 4$, Consistency is trivial.

$\Delta_{q,N} < \infty$, $q = 1, 2, 3, 4$ is guaranteed by that x_{ij}, ε_{ij} has finite fourth moments and the kernel function is bounded. □

Proof of Theorem 3.2

Proof. Following the proof of Theorem 3.1 We focus on $\frac{1}{r_N} U_N$.

One important observation:

- $\Delta_{1,N} = 0$ as well. Actually, for any two dyads sharing no links, their contribution to $\text{Var}(U_N)$ is trivial.

Remark: Consider

$$u_1 = \frac{1}{r_N} T_{ijkl} x_{ij,kl} \varepsilon_{ij,kl} K\left(\frac{\delta_{ik}}{h_N}\right)$$

$$u_2 = \frac{1}{r_N} T_{mnop} x_{mn,op} \varepsilon_{mn,op} K\left(\frac{\delta_{mo}}{h_N}\right)$$

where they share no links (possibly sharing 0, 1 or 2 agents.).

Since

$$\varepsilon_{ij,kl} = \text{selection bias} + \text{link specific error term}$$

The first part is 0 if $\omega_i = \omega_k$. the second item makes 0 contribution to the covariance if they share no links.

Then

$$\begin{aligned} \frac{1}{r_N^2} \text{Var}(U_N) &= \binom{N}{4}^{-1} \binom{4}{2} \binom{N-4}{2} \Delta_{2,N} \\ &+ \binom{N}{4}^{-1} \binom{4}{3} \binom{N-4}{1} \Delta_{3,N} \\ &+ \binom{N}{4}^{-1} \binom{4}{4} \binom{N-4}{4} \Delta_{4,N} \end{aligned}$$

As long as $\Delta_{2,N}$, $\Delta_{3,N}$ and $\Delta_{4,N}$ make the same order nontrivial contributions to $\text{Var}(U_N)$.

The first item dominates.

Using Theorem 1.1 of Chatterjee et al. (2006) and Theorem 1 of Graham (2017), I finish the proof. □

Bibliography

- Abraham, Christophe, Pierre-André Cornillon, ERIC Matzner-Løber, and Nicolas Molinari. 2003. “Unsupervised curve clustering using B-splines.” *Scandinavian journal of statistics* 30 (3):581–595.
- Ahn, Hyungtaik and James L Powell. 1993. “Semiparametric estimation of censored selection models with a nonparametric selection mechanism.” *Journal of Econometrics* 58 (1-2):3–29.
- Ai, Chunrong and Xiaohong Chen. 2003. “Efficient estimation of models with conditional moment restrictions containing unknown functions.” *Econometrica* 71 (6):1795–1843.
- Ai, Chunrong and Qi Li. 2008. “Semi-parametric and Non-parametric methods in panel data models.” In *The Econometrics of Panel Data*. Springer, 451–478.
- Ando, Tomohiro and Jushan Bai. 2014. “Asset pricing with a general multifactor structure.” *Journal of Financial Econometrics* 13 (3):556–604.
- . 2016. “Panel data models with grouped factor structure under unknown group membership.” *Journal of Applied Econometrics* 31 (1):163–191.
- . 2017. “Clustering huge number of financial time series: A panel data approach with high-dimensional predictors and factor structures.” *Journal of the American Statistical Association* 112 (519):1182–1198.

- Andrews, Donald WK and Marcia MA Schafgans. 1998. “Semiparametric estimation of the intercept of a sample selection model.” *The Review of Economic Studies* 65 (3):497–517.
- Arellano, Manuel and Bo Honoré. 2001. “Panel data models: some recent developments.” In *Handbook of econometrics*, vol. 5. Elsevier, 3229–3296.
- Atkin, David and Dave Donaldson. 2015. “Who’s getting globalized? The size and implications of intra-national trade costs.” Tech. rep., National Bureau of Economic Research.
- Atkin, David, Amit K Khandelwal, and Adam Osman. 2017. “Exporting and firm performance: Evidence from a randomized experiment.” *The Quarterly Journal of Economics* 132 (2):551–615.
- Auerbach, Eric. 2019. “Identification and estimation of a partially linear regression model using network data.” *arXiv preprint arXiv:1903.09679* .
- Bala, Venkatesh and Sanjeev Goyal. 2000. “A noncooperative model of network formation.” *Econometrica* 68 (5):1181–1229.
- Baltagi, Badi H, Georges Bresson, and Alain Pirotte. 2008. “To pool or not to pool?” In *The econometrics of panel data*. Springer, 517–546.
- Baltagi, Badi H, James M Griffin, and Weiwen Xiong. 2000. “To pool or not to pool: Homogeneous versus heterogeneous estimators applied to cigarette demand.” *Review of Economics and Statistics* 82 (1):117–126.
- Baltagi, Badi H and Dan Levin. 1992. “Cigarette taxation: Raising revenues and reducing consumption.” *Structural Change and Economic Dynamics* 3 (2):321–335.
- Baltagi, Badi Hani. 2015. *The Oxford handbook of panel data*. Oxford Handbooks.
- Bester, C Alan and Christian B Hansen. 2016. “Grouped effects estimators in fixed effects models.” *Journal of Econometrics* 190 (1):197–208.

- Bonhomme, Stéphane and Elena Manresa. 2015. “Grouped patterns of heterogeneity in panel data.” *Econometrica* 83 (3):1147–1184.
- Browning, Martin and Jesus Carro. 2007. “Heterogeneity and microeconometrics modeling.” *Econometric Society Monographs* 43:47.
- Browning, Martin and Jesus M Carro. 2010. “Heterogeneity in dynamic discrete choice models.” *The Econometrics Journal* 13 (1):1–39.
- . 2014. “Dynamic binary outcome models with maximal heterogeneity.” *Journal of Econometrics* 178 (2):805–823.
- Candelaria, Luis E. 2016. “A semiparametric network formation model with multiple linear fixed effects.” *Duke University* .
- Chandrasekhar, Arun G and Matthew O Jackson. 2014. “Tractable and consistent random graph models.” Tech. rep., National Bureau of Economic Research.
- Chatterjee, Sourav et al. 2006. “A generalization of the Lindeberg principle.” *The Annals of Probability* 34 (6):2061–2076.
- Chen, Xiaohong. 2007. “Large sample sieve estimation of semi-nonparametric models.” *Handbook of econometrics* 6:5549–5632.
- Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. “Double/debiased machine learning for treatment and structural parameters.” *The Econometrics Journal* 21 (1):C1–C68. URL <https://doi.org/10.1111/ectj.12097>.
- Chernozhukov, Victor, Iván Fernández-Val, and Siyi Luo. 2018. “Distribution regression with sample selection, with an application to wage decompositions in the UK.” *arXiv preprint arXiv:1811.11603* .

- Chernozhukov, Victor, Ivan Fernandez-Val, and Martin Weidner. 2018. “Network and panel quantile effects via distribution regression.” *arXiv preprint arXiv:1803.08154* .
- Chiou, Jeng-Min and Pai-Ling Li. 2007. “Functional clustering and identifying substructures of longitudinal data.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69 (4):679–699.
- Clemens, Michael A, Ethan G Lewis, and Hannah M Postel. 2018. “Immigration restrictions as active labor market policy: Evidence from the mexican bracero exclusion.” *American Economic Review* 108 (6):1468–87.
- Das, Mitali, Whitney K Newey, and Francis Vella. 2003. “Nonparametric estimation of sample selection models.” *The Review of Economic Studies* 70 (1):33–58.
- De Paula, Áureo, Seth Richards-Shubik, and Elie Tamer. 2018. “Identifying preferences in networks with bounded degree.” *Econometrica* 86 (1):263–288.
- Durlauf, Steven N, Andros Kourtellos, and Artur Minkin. 2001. “The local Solow growth model.” *European Economic Review* 45 (4-6):928–940.
- Fan, Jianqing, Jinchi Lv, and Lei Qi. 2011. “Sparse high-dimensional models in economics.” *Annu. Rev. Econ.* 3 (1):291–317.
- Fernández-Val, Iván and Martin Weidner. 2016. “Individual and time effects in nonlinear panel models with large N, T.” *Journal of Econometrics* 192 (1):291–312.
- Goldsmith-Pinkham, Paul and Guido W Imbens. 2013. “Social networks and the identification of peer effects.” *Journal of Business & Economic Statistics* 31 (3):253–264.
- Graham, Bryan S. 2017. “An econometric model of network formation with degree heterogeneity.” *Econometrica* 85 (4):1033–1063.
- Greene, William H. 1994. “Accounting for excess zeros and sample selection in Poisson and negative binomial regression models.” .

- Hahn, Jinyong and Hyungsik Roger Moon. 2010. “Panel data models with finite number of multiple equilibria.” *Econometric Theory* 26 (3):863–881.
- Haveman, Jon and David Hummels. 2004. “Alternative hypotheses and the volume of trade: the gravity equation and the extent of specialization.” *Canadian Journal of Economics/Revue canadienne d’économique* .
- Heckman, James. 1974. “Shadow prices, market wages, and labor supply.” *Econometrica: journal of the econometric society* :679–694.
- . 1990. “Varieties of selection bias.” *The American Economic Review* 80 (2):313–318.
- Heckman, James J. 1976. “The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models.” In *Annals of economic and social measurement, volume 5, number 4*. NBER, 475–492.
- . 1979. “Sample selection bias as a specification error.” *Econometrica: Journal of the econometric society* :153–161.
- Helpman, Elhanan, Marc Melitz, and Yona Rubinstein. 2008. “Estimating trade flows: Trading partners and trading volumes.” *The quarterly journal of economics* 123 (2):441–487.
- Hsiao, Cheng and M Hashem Pesaran. 2008. “Random coefficient models.” In *The econometrics of panel data*. Springer, 185–213.
- Hsiao, Cheng and A Kamil Tahmiscioglu. 1997. “A panel analysis of liquidity constraints and firm investment.” *Journal of the American Statistical Association* 92 (438):455–465.
- Hsieh, Chih-Sheng and Lung Fei Lee. 2016. “A social interactions model with endogenous friendship formation and selectivity.” *Journal of Applied Econometrics* 31 (2):301–319.
- Huang, Jian, Joel L Horowitz, and Fengrong Wei. 2010. “Variable selection in nonparametric additive models.” *Annals of statistics* 38 (4):2282.

- Jackson, Matthew O. 2008. “Average distance, diameter, and clustering in social networks with homophily.” In *International Workshop on Internet and Network Economics*. Springer, 4–11.
- Jackson, Matthew O and Asher Wolinsky. 1996. “A strategic model of social and economic networks.” *Journal of economic theory* 71 (1):44–74.
- Johnsson, Ida and Hyungsik Roger Moon. 2017. “Estimation of peer effects in endogenous social networks: Control function approach.” *USC-INET Research Paper* (17-25).
- Ke, Zheng Tracy, Jianqing Fan, and Yichao Wu. 2015. “Homogeneity pursuit.” *Journal of the American Statistical Association* 110 (509):175–194.
- Kyriazidou, Ekaterini. 1997. “Estimation of a panel data sample selection model.” *Econometrica: Journal of the Econometric Society* :1335–1364.
- Lee, Kevin, M Hashem Pesaran, and Ron Smith. 1997. “Growth and convergence in a multi-country empirical stochastic Solow model.” *Journal of applied Econometrics* 12 (4):357–392.
- Lin, Chang-Ching and Serena Ng. 2012. “Estimation of panel data models with parameter heterogeneity when group membership is unknown.” *Journal of Econometric Methods* 1 (1):42–55.
- Luan, Yihui and Hongzhe Li. 2003. “Clustering of time-course gene expression data using a mixed-effects model with B-splines.” *Bioinformatics* 19 (4):474–482.
- Mammen, Enno, Bård Støve, and Dag Tjøstheim. 2009. “Nonparametric additive models for panels of time series.” *Econometric Theory* 25 (2):442–481.
- Mátyás, László and Patrick Sevestre. 2013. *The econometrics of panel data: handbook of theory and applications*, vol. 28. Springer Science & Business Media.

- Menzel, Konrad. 2015. “Strategic network formation with many agents.” Tech. rep., Working papers, NYU.
- Merlevède, Florence, Magda Peligrad, Emmanuel Rio et al. 2009. “Bernstein inequality and moderate deviations under strong mixing conditions.” In *High dimensional probability V: the Luminy volume*. Institute of Mathematical Statistics, 273–292.
- Miao, Ke, Liangjun Su, and Wendun Wang. 2020. “Panel threshold regressions with latent group structures.” *Journal of Econometrics* 214 (2):451–481.
- Newey, Whitney K. 1997. “Convergence rates and asymptotic normality for series estimators.” *Journal of econometrics* 79 (1):147–168.
- . 2009. “Two-step series estimation of sample selection models.” *The Econometrics Journal* 12:S217–S229.
- Ni, Zhong-Xin, Da-Zhong Wang, and Wen-Jun Xue. 2015. “Investor sentiment and its nonlinear effect on stock returns—New evidence from the Chinese stock market based on panel quantile regression model.” *Economic Modelling* 50:266–274.
- Phillips, Peter CB and Donggyu Sul. 2007. “Transition modeling and econometric convergence tests.” *Econometrica* 75 (6):1771–1855.
- Powell, James L. 1994. “Estimation of semiparametric models.” *Handbook of econometrics* 4:2443–2521.
- Profit, Stefan and Stefan Sperlich. 2004. “Non-uniformity of job-matching in a transition economy—A nonparametric analysis for the Czech Republic.” *Applied Economics* 36 (7):695–714.
- Ridder, Geert and Shuyang Sheng. 2015. “Estimation of large network formation games.” Tech. rep., Working papers, UCLA.

- Rose, Andrew K. 2004. “Do we really know that the WTO increases trade?” *American Economic Review* 94 (1):98–114.
- Sarafidis, Vasilis and Neville Weber. 2015. “A partially heterogeneous framework for analyzing panel data.” *Oxford Bulletin of Economics and Statistics* 77 (2):274–296.
- Sheng, Shuyang. 2018. “A structural econometric analysis of network formation games through subnetworks.” *forthcoming in Econometrica, mimeo UCLA* .
- Sperlich, Stefan, Dag Tjøstheim, and Lijian Yang. 2002. “Nonparametric estimation and testing of interaction in additive models.” *Econometric Theory* 18 (2):197–251.
- Su, Liangjun and Qihui Chen. 2013. “Testing homogeneity in panel data models with interactive fixed effects.” *Econometric Theory* 29 (6):1079–1135.
- Su, Liangjun and Gaosheng Ju. 2018. “Identifying latent grouped patterns in panel data models with interactive fixed effects.” *Journal of Econometrics* 206 (2):554–573.
- Su, Liangjun, Zhentao Shi, and Peter CB Phillips. 2016. “Identifying latent structures in panel data.” *Econometrica* 84 (6):2215–2264.
- Su, Liangjun and Aman Ullah. 2011. “Nonparametric and semiparametric panel econometric models: estimation and testing.” *Handbook of empirical economics and finance* :455–497.
- Su, Liangjun, Xia Wang, and Sainan Jin. 2019. “Sieve estimation of time-varying panel data models with latent structures.” *Journal of Business & Economic Statistics* 37 (2):334–349.
- Sun, Yixiao. 2005. “Estimation and inference in panel structure models.” *Available at SSRN 794884* .
- Tarpey, Thaddeus. 2007. “Linear transformations and the k-means clustering algorithm: applications to clustering curves.” *The American Statistician* 61 (1):34–40.

- Terza, Joseph V. 1998. “Estimating count data models with endogenous switching: Sample selection and endogenous treatment effects.” *Journal of econometrics* 84 (1):129–154.
- Ullah, Aman and Nilanjana Roy. 1998. “Nonparametric and semiparametric econometrics of panel data.” *STATISTICS TEXTBOOKS AND MONOGRAPHS* 155:579–604.
- Vogt, Michael and Oliver Linton. 2017. “Classification of non-parametric regression functions in longitudinal data models.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79 (1):5–27.
- . 2020. “Multiscale clustering of nonparametric regression curves.” *Journal of Econometrics* .
- Wang, Wuyi, Peter CB Phillips, and Liangjun Su. 2018. “Homogeneity pursuit in panel data models: Theory and application.” *Journal of Applied Econometrics* 33 (6):797–815.