

**UCLA**

**UCLA Electronic Theses and Dissertations**

**Title**

Essays on Econometrics

**Permalink**

<https://escholarship.org/uc/item/4kr5w52k>

**Author**

Zhou, Wenyu

**Publication Date**

2020

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA  
Los Angeles

Essays on Econometrics

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

by

Wenyu Zhou

2020

© Copyright by  
Wenyu Zhou  
2020

# ABSTRACT OF THE DISSERTATION

Essays on Econometrics

by

Wenyu Zhou

Doctor of Philosophy in Economics

University of California, Los Angeles, 2020

Professor Rosa Liliana Matzkin, Co-Chair

Professor Shuyang Sheng, Co-Chair

This dissertation consists of four main chapters that study network social interaction models and panel models with grouped heterogeneity. Chapter 1 and Chapter 2 are representative work finished during my early exploration of economics. Chapter 3 and Chapter 4 are completed during the last two years of my Ph.D. studies.

Chapter 1 studies a network social interaction model with heterogeneous links. I show that the endogenous and exogenous social interaction effects as well as the strength of network links are identified under some mild conditions. I adopt the nonlinear least squares method to estimate the unknown parameters using data of a single network. I also investigate the finite sample performance of the estimation method through Monte Carlo simulations and apply the model to analyze an online social network.

Chapter 2 studies social interactions model with both in-group and out-group effects. The in-group effect follows the standard setup in the literature, while the out-group effect is introduced by assuming the economic outcome also depends on its out-group average value. I present a network game with limited information of outside groups that rationalizes the econometric model. I show that both effects are identified under a set of mild regularity conditions. I propose to estimate the model using the two-stage least squares (2SLS) method

and establish the asymptotic normality of the estimators. The finite sample performance of the estimators are investigated through Monte Carlo simulations.

Chapter 3 studies a semiparametric panel quantile regression model with grouped heterogeneity. The model can capture both time-variant and time-invariant effects of explanatory variables when group-specific heterogeneity directly affects the coefficients. A series-based estimation method is developed to estimate the parameters of interest and the group memberships. I investigate the asymptotic properties of the estimators and propose an information criterion to estimate the number of groups. The finite sample performance of the estimation method and the information criterion are investigated through Monte Carlo simulations. I apply the model to study the effect of foreign direct investment (FDI) on economic growth. My empirical findings show that FDI has large and significant heterogeneous effects on economic growth, especially for low-income countries, and such effect diminishes as the GDP per capita increases. None of these findings have been documented in previous literature.

In Chapter 4 (joint with Hualei Shang), we study a nonparametric additive panel regression model with grouped heterogeneity. The model is a valuable extension to the heterogeneous panel model studied in Su et al. (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. Besides, we present the decision rule for group classification and establish its consistency. A BIC-type information criterion is developed to determine the group pattern of each nonparametric component. We investigate the finite sample performance of the estimation method and the information criterion through Monte Carlo simulations. Results show that both work very well. Finally, we apply the model to study the demand for cigarettes in the United States using panel data of 46 states from 1963 to 1992.

The dissertation of Wenyu Zhou is approved.

Ying Nian Wu

Zhipeng Liao

Shuyang Sheng, Committee Co-Chair

Rosa Liliana Matzkin, Committee Co-Chair

University of California, Los Angeles

2020

## DEDICATIONS

*To my parents...  
for their love and support.*

# Contents

<b>1</b>	<b>A Network Social Interaction Model with Heterogeneous Links</b>	<b>1</b>
1.1	Introduction . . . . .	2
1.2	Model . . . . .	3
1.2.1	Setup . . . . .	3
1.2.2	An Inherent Classification Criterion . . . . .	4
1.2.3	A Simple Example . . . . .	5
1.3	Identification . . . . .	6
1.4	Simulation and Empirical Application . . . . .	7
1.4.1	Simulation . . . . .	8
1.4.2	Empirical Application . . . . .	9
1.5	Conclusion . . . . .	10
<b>2</b>	<b>Social Interaction Models with Out-group Effects</b>	<b>13</b>
2.1	Introduction . . . . .	13
2.2	Setup . . . . .	15
2.2.1	The Model . . . . .	15
2.2.2	The Microfoundation . . . . .	17
2.3	Identification and Estimation . . . . .	18
2.3.1	Identification . . . . .	18
2.3.2	Estimation . . . . .	19



2.4	Monte Carlo Simulations . . . . .	20
2.5	Conclusion . . . . .	22
<b>3</b>	<b>Semiparametric Quantile Panel Regression with Grouped Heterogeneity</b>	<b>23</b>
3.1	Introduction . . . . .	24
3.2	Model . . . . .	26
3.2.1	Motivating Applications . . . . .	28
3.3	Estimation . . . . .	30
3.3.1	Series Approximation . . . . .	30
3.3.2	Implementation . . . . .	31
3.4	Asymptotic Properties . . . . .	32
3.5	Monte Carlo Simulations . . . . .	40
3.5.1	Data Generating Processes . . . . .	40
3.5.2	The Determination of the Number of Groups . . . . .	43
3.5.3	Implementation Details . . . . .	44
3.5.4	Simulation Results . . . . .	46
3.6	Empirical Application . . . . .	48
3.6.1	Background . . . . .	48
3.6.2	Data and Estimation . . . . .	50
3.6.3	Empirical Results . . . . .	52
3.7	Conclusion . . . . .	54
<b>4</b>	<b>Nonparametric Additive Panel Regression Models with Grouped Heterogeneity</b>	<b>79</b>
4.1	Introduction . . . . .	80
4.2	Model . . . . .	83
4.3	Estimation . . . . .	84
4.3.1	Sieve Approximation . . . . .	84

4.3.2	Penalized Estimation of $h(x)$ and $f(x)$ . . . . .	86
4.4	Asymptotic Properties . . . . .	88
4.4.1	Preliminary Rates of Convergence . . . . .	88
4.4.2	Classification Consistency . . . . .	92
4.4.3	The Oracle Property and Asymptotic Distributions . . . . .	93
4.4.4	Determination of Number of Groups . . . . .	96
4.5	Simulation . . . . .	98
4.5.1	Data Generating Process . . . . .	98
4.5.2	Simulation Results . . . . .	101
4.6	Empirical Illustration . . . . .	103
4.7	Conclusion . . . . .	107
<b>5</b>	<b>Appendix</b>	<b>108</b>

# List of Figures

3.1	Figure 1: The Plots of Nonparametric Coefficients in Different DGPs (top panel: solid line for $\beta_{G_1,\tau}(\cdot)$ and dashed line for $\beta_{G_2,\tau}(\cdot)$ in DGP 1. Middle panel: solid line for $\beta_{G_1,\tau}(\cdot)$ , dashed line for $\beta_{G_2,\tau}(\cdot)$ and dash-dotted line for $\beta_{G_3,\tau}(\cdot)$ in DGP 2, DGP 3, and $\beta_{G_{11},\tau}(\cdot)$ , $\beta_{G_{21},\tau}(\cdot)$ ) and $\beta_{G_{31},\tau}(\cdot)$ in DGP 4. Bottom panel: solid line for $\beta_{G_{12},\tau}(\cdot)$ , dashed line for $\beta_{G_{22},\tau}(\cdot)$ and dash-dotted line for $\beta_{G_{32},\tau}(\cdot)$ in DGP 4. . . . .	56
3.2	Figure 2: Estimated Functional Coefficients ( $\tau = 0.25$ ) . . . . .	62
3.3	Figure 2.1: Bootstrap 95% Confidence Band of the Functional Coefficient (Pooling Case, $\tau = 0.25$ ) . . . . .	63
3.4	Figure 2.2: Bootstrap 95% Confidence Band of the Functional Coefficient (Group 1, $\tau = 0.25$ ) . . . . .	64
3.5	Figure 2.3: Bootstrap 95% Confidence Band of the Functional Coefficient (Group 2, $\tau = 0.25$ ) . . . . .	65
3.6	Figure 2.4: Bootstrap 95% Confidence Band of the Functional Coefficient (Group 3, $\tau = 0.25$ ) . . . . .	66
3.7	Figure 3: Estimated Functional Coefficients ( $\tau = 0.50$ ) . . . . .	67
3.8	Figure 3.1: Bootstrap 95% Confidence Band of the Functional Coefficient (Pooling Case, $\tau = 0.50$ ) . . . . .	68
3.9	Figure 3.2: Bootstrap 95% Confidence Band for the Functional Coefficient (Group 1, $\tau = 0.50$ ) . . . . .	69

3.10	Figure 3.3: Bootstrap 95% Confidence Band of the Functional Coefficient (Group 2, $\tau = 0.50$ ) . . . . .	70
3.11	Figure 3.4: Bootstrap 95% Confidence Band of the Functional Coefficient (Group 3, $\tau = 0.50$ ) . . . . .	71
3.12	Figure 4: Estimated Functional Coefficients ( $\tau = 0.75$ ) . . . . .	72
3.13	Figure 4.1: Bootstrap 95% Confidence Band of the Functional Coefficient (Pooling Case, $\tau = 0.75$ ) . . . . .	73
3.14	Figure 4.2: Bootstrap 95% Confidence Band of the Functional Coefficient (Group 1, $\tau = 0.75$ ) . . . . .	74
3.15	Figure 4.3: Bootstrap 95% Confidence Band of the Functional Coefficient (Group 2, $\tau = 0.75$ ) . . . . .	75
3.16	Figure 4.4: Bootstrap 95% Confidence Band of the Functional Coefficient (Group 3, $\tau = 0.75$ ) . . . . .	76
4.1	Plot of $\hat{h}_1(x)$ . . . . .	105
4.2	Plot of $\hat{h}_2(x)$ . . . . .	106

# List of Tables

1.1	Monte Carlo Simulation Results (1000 draws) . . . . .	11
1.2	Estimation Results of the Weibo Dataset . . . . .	12
2.1	Finite Sample Performance of the 2SLS Estimators (1000 draws) . . . . .	21
2.2	Simulation Results of the Mis-specified Model (1000 draws) . . . . .	22
3.1	Table 4: Some Summary Statistics of the Data Set . . . . .	51
3.2	Table 5: Values of the Information Criterion for Different $K$ . . . . .	51
3.3	Table 1: Finite Sample Performance of the Information Criterion . . . . .	57
3.4	Table 2: Bias and RMSE of the Coefficient Estimates . . . . .	58
3.5	Table 3: Empirical Rate of Correct Classification . . . . .	59
3.6	Continued Table 3: Empirical Rate of Correct Classification . . . . .	60
3.7	Table 6: Estimates of Parametric Coefficients in Model 3.15 . . . . .	61
3.8	Table 7: Group Classification of 90 Countries in the Data Set . . . . .	77
3.9	Table 8: Group-Specific Summary Statistics . . . . .	78
4.1	Simulation Results of the Group-specific Parameters in DGP 1 . . . . .	102
4.2	Simulation Results of the Group-specific Parameters in DGP 2 . . . . .	102
4.3	Simulation Results of the Group-specific Parameters in DGP 3 . . . . .	103
5.1	Descriptive Statistics of the Weibo Data Set . . . . .	117

## ACKNOWLEDGMENT

I wish to first thank my advisors Rosa Matzkin, Shuyang Sheng, Zhipeng Liao and Ying Nian Wu for their invaluable guidance and continuous support. I am also grateful to Jinyong Hahn, Denis Chetverikov, Andres Santos and other econometrics seminar participants at the University of California, Los Angeles. All errors are mine.

Chapter 3 is the original manuscript of Zhou (2019) published in *Economics Letters*, available online:

<https://www.sciencedirect.com/science/article/abs/pii/S0165176519301296>.

## VITA

### Education

#### University of California, Los Angeles

M.S. in Statistics	2019
C.Phil. in Economics	2018
M.A. in Economics	2017

#### Peking University

B.A. in Finance	2015
B.S. in Applied Mathematics	2015

### Publications

Zhou, Wenyu. "A network social interaction model with heterogeneous links." *Economics Letters* 180 (2019): 50-53.

### Honors and Awards

UCLA Dissertation Year Fellowship	2020
UCLA Economics Department Fellowship	2015-2019
Cathy Bank Fellowship	2017
College-level Outstanding Graduate, Peking University	2015
Sun Hung Kai Properties Fellowship, Peking University	2014
Zhang Wenjin Fellowship, Peking University	2013
Guanghua Fellowship, Peking University	2012

## Teaching Experience

Teaching Assistant, Department of Economics, UCLA

(1) Statistics for Economists (Fall 2016, Winter 2017, Fall 2017)

(2) Econometrics (Spring 2017, Winter 2018, Fall 2018, Winter 2019, Spring 2019)

(3) Statistical Reasoning (Summer 2019)

(4) Principles of Economics (Spring 2018, Spring 2020)

(5) Master-level Econometrics (Fall 2019)

(6) Computational Finance and Data Analysis (Winter 2020)



# Chapter 1

## A Network Social Interaction Model with Heterogeneous Links

## 1.1 Introduction

There is a growing literature on social interaction models. Researchers have paid considerable attention to identification and estimation of social interaction models under different network settings (e.g., Bramoullé et al. (2009); Liu and Lee (2010); Blume et al. (2015)). In these studies, the strength of network links, which is represented by a so-called adjacency matrix, is either assumed to be known or the same. But most current network datasets only contain information on the link profile (coded as 1 and 0), making it difficult for researchers to know the strength of links *ex ante*. If the heterogeneity in links is ignored, the estimated social interaction effects may be biased and can generate misleading policy implications. For example, previous studies (Carrell et al. (2009); Lin (2010)) have found that there exist positive social interaction effects on student academic achievement. However, a strong social interaction effect may only exist between students with similar past academic performance (Carrell et al. (2013)), and simply organizing students with polarized academic performance into one classroom may not help to improve the overall academic performance of the whole class. As pointed out in Jackson et al. (2017), it is important to capture the heterogeneity in links if one wants to have a better understanding of social interaction effects.

In this chapter, we study a network social interaction model which enables researchers to use information about the link classification to account for the heterogeneity in links. To the best of our knowledge, there is limited existing literature on estimating the heterogeneity in links using cross-sectional data of a single network. The idea is to classify network links into different groups and impose group-level fix effects to control the heterogeneity. We also propose a simple data-driven classification criterion to determine the types of links, which does not require any extra information except for the adjacency matrix. We show that the endogenous and exogenous social interaction effects as well as the strength of network links can be identified under some mild conditions. We adopt the nonlinear least squares (NLS) method for estimation, which has been used in Wang and Lee (2013) and Liu et al. (2017). Additionally, we investigate the finite sample performance of the model through Monte Carlo

simulations and apply the method to analyze an online social network.

## 1.2 Model

### 1.2.1 Setup

Researchers are interested in studying social interaction effects in some network which consists of  $N$  agents. The network may contain more than one component<sup>1</sup> and can be represented by a  $N \times N$  adjacency matrix  $\mathbf{W}$ , where  $\mathbf{W}_{ij} = 1$  if  $i$  links to  $j$ , and  $\mathbf{W}_{ij} = 0$  otherwise. Following the convention, we assume  $\mathbf{W}_{ii} = 0$  and there are no isolated agents. We focus on undirected networks<sup>2</sup>, i.e.,  $\mathbf{W}_{ij} = \mathbf{W}_{ji}$ . Researchers also have information on the link classification and all links are classified into  $K$  groups. The classification can also be represented by a  $N \times N$  matrix  $\mathbf{M}$ , where  $\mathbf{M}_{ij}$  denotes the group identity of the link  $(i, j)$ . Let  $c_k$  represent the strength of links in group  $k$ , where  $k \in \{1, \dots, K\}$ . Without loss of generality, we normalize the strength of links in the first group to be 1, i.e.,  $c_1 = 1$ . The structural model is given by

$$\mathbf{y} = \lambda \mathbf{G} \mathbf{y} + \mathbf{x} \beta + \mathbf{G} \mathbf{x} \delta + \epsilon, \quad (1.1)$$

where  $\mathbf{y} = (y_1, \dots, y_N)' \in \mathbf{R}^N$  is the vector of agents' outcome variables,  $\mathbf{x} = (x'_1, \dots, x'_N)' \in \mathbf{R}^{N \times p}$  is the matrix of agents' exogenous characteristics. For the sake of simplicity in notation, we assume  $p = 1$  in the rest of the chapter. We define the  $N \times N$  matrix  $\mathbf{G}$  to be the *strength-adjusted adjacency matrix*, which reflects the influence of link strength on the scale of social interaction effects.  $\mathbf{G}$  is constructed using the original

---

<sup>1</sup>Components are parts of the network that are disconnected from each other.

<sup>2</sup>The model can be generalized to cover the directed networks.

adjacency matrix  $\mathbf{W}$  and the classification matrix  $\mathbf{M}$  as follows

$$\mathbf{G}_{ij} = \frac{\mathbf{W}_{ij}c_{\mathbf{M}_{ij}}}{\sum_{i=1}^N \mathbf{W}_{ij}c_{\mathbf{M}_{ij}}}, \quad (1.2)$$

where  $\mathbf{M}_{ij}$  is the group of the link  $(i, j)$  and  $c_{\mathbf{M}_{ij}}$  represents the corresponding link strength. In addition,  $\lambda$  is the endogenous social interaction effect and  $\delta$  is the exogenous social interaction effect. Therefore, the parameters of interest are  $\theta = (\lambda, \beta, \delta, c_2, \dots, c_K)'$ , as we assume that  $c_1$  is normalized to be 1. It is worth noting that we can easily generalize the model by allowing for component-specific unobserved fixed effects and similar techniques of local difference can be applied to avoid the incidental parameters problem as in Bramoullé et al. (2009) and Lin (2010).

### 1.2.2 An Inherent Classification Criterion

In most cases, researchers can classify links according to the characteristics of the agents in the network. For example, when studying the social interaction effects on student academic achievement, links can be classified based on students' previous academic performance. Besides utilizing extra information, the adjacency matrix  $\mathbf{W}$  can also provide useful information for link classification. Sociology and computer science literature suggest that codegree could be used as a measure of link strength (Marsden and Campbell (1984); Gilbert and Karahalios (2009)). Motivated by such evidence, we propose to use codegree

$$k_{ij} = \sum_{l=1}^N \mathbf{W}_{il}\mathbf{W}_{lj} \quad (1.3)$$

as a default link classification criterion, which is easy to implement as the codegree matrix equals  $\mathbf{W}^2$ . Besides, since most social networks are sparse, such a criterion also guarantees the number of link groups  $K$  is bounded, which ensures an accurate estimation. Other statistical properties of network data such as centrality may also be useful for link

classification.

### 1.2.3 A Simple Example

Here we present a simple example in which we use codegree as a classification criterion.

Consider the following network

$$\mathbf{W} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}.$$

**Step 1:** Calculate the classification matrix  $\mathbf{M} = \mathbf{W} \circ \mathbf{W}^2$ , where  $\circ$  is the hadamard product,

$$\mathbf{M} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

**Step 2:** Construct the *strength-adjusted adjacency matrix*  $\mathbf{G}$  using the original adjacency matrix  $\mathbf{W}$  and the classification matrix  $\mathbf{M}$ . Let  $c_1, c_2, c_3$  denote the strength for links  $(i, j)$  such that  $\mathbf{W}_{ij} = 1$  and  $\mathbf{M}_{ij} = 0, 1, 2$ , respectively. We normalize  $c_1$  to be 1. According to

the equation (1.2), the strength-adjusted adjacency matrix is given by

$$\mathbf{G} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{1+2c_2} & 0 & \frac{c_2}{1+2c_2} & 0 & \frac{c_2}{1+2c_2} & 0 & 0 \\ 0 & \frac{c_2}{2c_2+c_3} & 0 & \frac{c_2}{2c_2+c_3} & \frac{c_3}{2c_2+c_3} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{c_2}{2+2c_2+c_3} & \frac{c_3}{2+2c_2+c_3} & \frac{c_2}{2+2c_2+c_3} & 0 & \frac{1}{2+2c_2+c_3} & \frac{1}{2+2c_2+c_3} \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}.$$

### 1.3 Identification

In this section, we discuss the identification issue of the model.

**Assumption 1:**  $\mathbb{E}[\epsilon|\mathbf{x}] = 0$  and  $\mathbb{E}[\mathbf{x}'\mathbf{x}]$  is non-singular.

**Assumption 2:**  $|\lambda| < 1$ .

**Assumption 3:**  $\lambda\beta + \delta \neq 0$ .

**Assumption 4:**  $0 < c_k < M$  for all  $k = 1, \dots, K$ , where  $M$  is a constant.

**Assumption 5:** There exist  $i, j \in V$  such that  $(\mathbf{G}^2)_{ii} \neq (\mathbf{G}^2)_{jj}$ .

**Assumption 6:** The network contains either (1) one component, or (2)  $L$  components such that every non-empty proper subset  $P$  of  $\{1, \dots, L\}$  satisfies  $c(\mathbf{G}_P) \cap c(\mathbf{G}_{P^c}) \neq \emptyset$ , where  $\mathbf{G}_P$  represents the subnetwork that is comprised of all components in  $P$  and  $c(\mathbf{G}_P)$  denotes the set of link identities of all links in this subnetwork.

Assumptions 1-5 are standard in the literature of social interaction models. It is worth noting that Assumption 2 implies that  $(\mathbf{I} - \lambda\mathbf{G})$  is invertible, which enables us to rewrite (1.1) into

the reduced form

$$\mathbf{y} = (\mathbf{I} - \lambda\mathbf{G})^{-1}(\beta\mathbf{I} + \delta\mathbf{G})\mathbf{x} + (\mathbf{I} - \lambda\mathbf{G})^{-1}\epsilon, \quad (1.4)$$

which will be used for the estimation of parameters. Assumption 3 rules out the cases in which endogenous and exogenous social interactions are zero or balance each other out. Assumption 4 basically ensures that link strength is positive and finite. Assumption 5 relates to the asymmetry of network structure as pointed out in De Paula et al. (2018) and implies the network independence condition in Bramoullé et al. (2009). Assumption 6 ensures that all link identities in the network can be compared with the normalized group such that  $c_1 = 1$ . If there are multiple components, we must have some links acting as an information bridge between different components to make each link comparable with links in the normalized group.

**Proposition 1:** Under Assumptions 1-6, parameters  $\theta = (\lambda, \beta, \delta, c_2, \dots, c_K)'$  in the social interaction model defined in Section 2 are identified if (1) the sign of  $(\lambda\beta + \delta)$  is known, or (2)  $\lambda > 0$ .

Proof: See the appendix.

## 1.4 Simulation and Empirical Application

Since the model is nonlinear in parameters, we use the nonlinear least squares (NLS) method for estimation following Wang and Lee (2013) and Liu et al. (2017). The NLS estimator of  $\theta = (\lambda, \beta, \delta, c_2, \dots, c_K)'$  is given by

$$\hat{\theta}_{NLS} = \arg \min_{\theta \in \Theta_0} [\mathbf{y} - \mathbf{h}(\mathbf{x}, \theta)]' [\mathbf{y} - \mathbf{h}(\mathbf{x}, \theta)] \quad (1.5)$$

where  $\mathbf{h}(\mathbf{x}, \theta) = (\mathbf{I} - \lambda\mathbf{G}(\theta))^{-1}(\beta\mathbf{I} + \delta\mathbf{G}(\theta))\mathbf{x}$ . Following a similar argument in Wang and Lee (2013) and Liu et al. (2017), we can show the NLS estimator is consistent and asymptotically normal. A sketch of the asymptotic analysis is included in the online appendix.

### 1.4.1 Simulation

To investigate the finite sample performance of the proposed NLS estimator and demonstrate the insight of the model, we conduct Monte Carlo experiments based on the following specification

$$y_i = \lambda \sum_{j=1}^N g_{ij} y_j + x_i \beta + \delta \sum_{j=1}^N g_{ij} x_j + \epsilon_i, \quad (1.6)$$

where  $x_i$  is drawn from independent  $N(0, 2)$  distributions,  $\epsilon_i$  is drawn from independent  $N(0, 1)$  distributions and  $y_i$  is generated according to equation (1.4). We set  $\lambda = 0.5$ ,  $\beta = 1$ ,  $\delta = 0.8$  and generate 3 networks using the Erdős-Rényi model with  $N = \{100, 200, 400\}$ , requiring the average degree to be 5 and that there be no isolated agents. Such network settings aims mimic sparse social networks in reality. There are three link identities in our experiments and the probability of a link belongs to each group equals  $\frac{1}{3}$ . We normalize the strength of one group to be 1, i.e.,  $c_1 = 1$ . For parameters that control the strength of links, we consider 3 cases which mimic different real-world scenarios: (1)  $(c_1, c_2, c_3) = (1, 1, 1)$ , (2)  $(c_1, c_2, c_3) = (1, 1.1, 1.2)$ , (3)  $(c_1, c_2, c_3) = (1, 2, 4)$ . In this first case, there is no heterogeneity in links, while the heterogeneity becomes more severe in the second and in the third case. We conduct 1000 repetitions for each setting<sup>3</sup> and report the mean and standard deviation of the empirical distribution of the estimates in Table 1.1.

First, the simulation results show that the parameters of interest can be consistently estimated using the nonlinear least squares method in each scenario. It is worth noting that the estimation method works well for sparse social networks, as the density of the network with 400 agents in our third setting is only 1.3%. Second, it can be seen from the simulation results that both the bias and the standard deviation decrease when the size of the network increases. Third, we find that the standard deviations of  $c_2$  and  $c_3$ , which measure the strength of links, increase when the heterogeneity in links becomes more severe. This phenomenon may be attributed to the functional form of the elements in the

---

<sup>3</sup>The initial values for parameters are  $(\lambda, \beta, \delta, c_2, c_3) = (0, 0.5, 0.5, 1, 1)$ .



strength-adjusted adjacency matrix  $\mathbf{G}$ . For example, the corresponding variation in  $\frac{c_2}{1+c_2+c_3}$  is small with respect to the change in  $c_3$  when  $c_3$  is much larger than  $c_2$ , which may increase the difficulty of finding the global minimizer in equation (1.5).

### 1.4.2 Empirical Application

As an empirical illustration, we apply the model to analyze a social network of Internet celebrities using data collected from Sina Weibo, which is a popular Twitter-type social media in China. Our data set includes account information, individual characteristics and the link profile of 426 social media influencers, which mainly consist of singers and actors. Two agents are linked if they follow each other on Sina Weibo. Descriptive statistics are included in the appendix. We are interested in evaluating the social interaction effects on their willingness to interact with their followers, which is measured as the average number of daily posts in 2018. We cluster the links into three groups following two designs. In the first design, links are classified based on the disparity of social influence of two sides, which is measured by the ratio of their numbers of followers. One third of the links with largest ratios are classified into the first group and we normalize the strength of this group to be 1. The rest of links are classified evenly into two groups according to their ratios. Similarly, in the second design, links are classified based on their codegrees and we normalize the link strength of the group with the smallest codegrees. The estimation results are reported in Table 1.2.

There are several interesting findings about the estimation results. We find a positively significant endogenous social interaction effect on these social media influencers' willingness to interact with their followers. The results also show that male social media influencers tend to interact with their followers more frequently and the number of followers has a significant positive effect on their willingness to post new messages on the social media. In addition, we find significant heterogeneity in links and the magnitudes of endogenous social interaction effects are similar to each other in both designs. However, selecting the best link classification

criterion is beyond the scope of this chapter and we leave it as future work.

## 1.5 Conclusion

In this chapter, we have studied a network social interaction model, which enables researchers to detect the heterogeneity in links. We propose to estimate the model using the nonlinear least squares method. We investigate the finite sample performance of the NLS estimator through Monte Carlo simulations and apply the model to analyze the social interaction effects in a real social network.

Table 1.1: Monte Carlo Simulation Results (1000 draws)

Parameters	N=100		N=200		N=400	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
<i>No Heterogeneity</i>						
$\lambda = 0.5$	0.505	0.470	0.496	0.381	0.503	0.285
$\beta = 1$	1.000	0.103	1.001	0.077	0.999	0.054
$\delta = 0.8$	0.794	0.635	0.801	0.489	0.798	0.366
$c_2 = 1$	1.012	0.567	1.005	0.349	1.002	0.242
$c_3 = 1$	1.015	0.525	1.001	0.351	1.004	0.248
<i>Weak Heterogeneity</i>						
$\lambda = 0.5$	0.497	0.501	0.498	0.377	0.502	0.283
$\beta = 1$	1.001	0.115	1.001	0.078	0.999	0.053
$\delta = 0.8$	0.799	0.643	0.799	0.489	0.794	0.364
$c_2 = 1.1$	1.110	0.639	1.105	0.398	1.103	0.276
$c_3 = 1.2$	1.213	0.634	1.209	0.442	1.205	0.298
<i>Strong Heterogeneity</i>						
$\lambda = 0.5$	0.497	0.413	0.498	0.314	0.503	0.233
$\beta = 1$	0.999	0.098	1.001	0.073	0.999	0.049
$\delta = 0.8$	0.802	0.578	0.797	0.419	0.799	0.304
$c_2 = 2$	2.073	1.668	2.020	1.057	2.017	0.747
$c_3 = 4$	4.128	3.173	4.061	2.061	4.036	1.402

Table 1.2: Estimation Results of the Weibo Dataset

	Design 1		Design 2	
	Coef.	<i>t</i> -value	Coef.	<i>t</i> -value
Endogenous effect	0.148	2.086	0.162	2.239
<i>Own characteristics</i>				
Age	-0.021	-1.473	-0.025	-1.564
Male	0.182	1.981	0.167	2.101
Number of followers	0.127	2.416	0.119	2.355
<i>Exogenous effects</i>				
Age	-0.014	-0.339	-0.009	-0.198
Male	0.196	1.207	0.183	1.331
Number of followers	0.098	1.254	0.109	1.586
<i>Link strength</i>				
$c_2$	1.284	2.032	1.197	1.802
$c_3$	2.193	2.741	1.635	2.327

# Chapter 2

## Social Interaction Models with Out-group Effects

### 2.1 Introduction

Ever since the seminal work of Manski (1993), social interaction models have attracted considerable attention from both theoretical and empirical sides; see Jackson et al. (2017) and Kline and Tamer (2020) for a comprehensive review. The key feature of such models is that the economic outcome of interest is not only determined by one's own characteristics but also by his peers. For example, students' academic achievement, measured by GPA, are also affected by their friends' performance (Lin (2010)).

Motivated by the fact that many real-world networks can further decomposed into subgroups, a large amount of literature has focused on social interaction models with group structures; see Lee (2007), Liu and Lee (2010) and Bramoullé et al. (2009), among many others. All these studies assume that individuals can only be affected by their within-group friends. Such setting, however, can be restrictive in reality because potential group-level interaction effects are completely ignored. We illustrate this point using the same example of students' academic achievement. Suppose that all students in some city form a network.

This single network can be further decomposed by treating each school as a group. It is likely that a student's GPA may not only be affected by students in his school but also by the average academic performance of students in other schools if all students need to compete together, such as taking the city-level high school entrance examination.

In this paper, we regard the social interaction effect induced by individuals outside the group as *the out-group social interaction effect*. To introduce such effect into the classic social interaction models, we assume that one's economic outcome depends not only on his friends' economic outcomes but also the average value of other groups. This setting is motivated by the observation that one may not know the situation of other groups as well as of his own group. For example, it is likely that students have more information of the academic achievement of his peers in the same school than in other schools.

We show that both *the in-group social interaction effect* and *the out-group social interaction effect* are identified under a set of assumptions that have been made in previous studies (Bramoullé et al. (2009)). To estimate the parameters of interest, we adopt the two-stage least squares estimation method developed in Kelejian and Prucha (1998) and establish the asymptotic normality of the estimators. We investigate the finite sample performance of the 2SLS estimators through Monte Carlo simulations, which show they performs very well.

Our paper contributes to the literature of social interaction models by first introducing the *the out-group social interaction effect*. It is noteworthy that ignoring *the out-group social interaction effect* may lead to a significant bias of *the in-group social interaction effect* because these two effects are often positively correlated in practice. We illustrate this observation based on numerical experiments in Section 2.4. With the model and the asymptotic results developed in this paper, one can conveniently test whether *the in-group social interaction effect* alone is enough to capture all the interaction effects in real-world network data sets, making our model an appealing choice for empirical studies.

The rest of the paper is organized as follows. Section 2.2 presents the econometric model and a network game as its microfoundation. Section 2.3 studies the identification and the

2SLS estimation of the model. Section 2.4 investigates the finite sample performance of the proposed estimators through Monte Carlo simulations. Section 2.5 concludes. The online appendix offers proofs.

*Notations.* For any real vector or matrix  $A$ , we use  $A^\top$  to denote the transpose of  $A$  and  $A^{-1}$  to denote its inverse. We use  $A_{ij}$  to denote the  $ij$ th element of a matrix  $A$ . For two positive integers  $a$  and  $b$ , we let  $\mathbf{0}_{a \times b}$  denote the  $a \times b$  matrix consists of zeros and  $\mathbf{1}_a$  denote the  $a$ -dimensional unit vector. For a sequence of random variables  $X_n$ , we let  ${}_{n \rightarrow \infty} X_n$  denote its probability limit,  $\xrightarrow{p}$  and  $\xrightarrow{d}$  denote convergence in probability and in distribution, respectively.

## 2.2 Setup

### 2.2.1 The Model

Suppose we have data of a single network which consists of  $n$  individuals and  $K$  groups. We let  $G_k$  denote the  $k$ th group. In the group  $G_k$ ,  $k = 1, \dots, K$ , there are  $n_k$  individuals, so  $n = n_1 + \dots + n_k$ . The corresponding  $n_k \times n_k$  adjacency matrices  $W_k$  are observed. Without loss of generality, we let  $G_1 = \{1, \dots, n_1\}, \dots, G_K = \{\sum_{k=1}^{K-1} n_k + 1, \dots, \sum_{k=1}^K n_k\}$  denote the group structure and we use  $G(i)$  to represent the individual  $i$ 's group for  $i = 1, \dots, n$ . Following the literature (e.g., Lee (2007) and Bramoullé et al. (2009)), we assume that links only exist within groups. The social interaction model with both in-group and out-group effects is given by:

$$y_i = \lambda_1 \sum_{j \in G(i)} W_{G(i),ij} y_j + \lambda_2 \bar{y}_{-G(i)} + x_i^\top \beta + \epsilon_i, \quad (2.1)$$

where  $\lambda_1$  measures *the in-group social interaction effect*,  $\lambda_2$  measures *the out-group social interaction effect*,  $W_{G(i),ij}$  is the  $ij$ th element of the adjacency matrix of the group  $G(i)$ ,  $\bar{y}_{-G(i)}$  is the average value of the economic outcome outside the group  $G(i)$ , i.e.,  $\bar{y}_{-G(i)} =$

$1/(n - n_{G(i)}) \sum_{j \notin G(i)} y_j$ ,  $x_i$  is a  $p \times 1$  vector of nonstochastic<sup>1</sup> individual-specific characteristics and  $\epsilon_i$  is the error term.

To facilitate our discussion, we rewrite the model in its matrix form:

$$\mathbf{Y} = \lambda_1 \mathbf{W}_1 \mathbf{Y} + \lambda_2 \mathbf{W}_2 \mathbf{Y} + \mathbf{X}\beta + \epsilon, \quad (2.2)$$

where  $\mathbf{Y} = (y_1, \dots, y_n)^\top$ ,  $\mathbf{X} = (x_1, \dots, x_n)^\top$  and the two adjacency matrices are given by

$$\mathbf{W}_1 = \begin{bmatrix} W_1 & \mathbf{0}_{n_1 \times n_2} & \cdots & \mathbf{0}_{n_1 \times n_K} \\ \mathbf{0}_{n_2 \times n_1} & W_2 & \cdots & \mathbf{0}_{n_2 \times n_K} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{n_K \times n_1} & \mathbf{0}_{n_K \times n_2} & \cdots & W_K \end{bmatrix} \in \mathbb{R}^{n \times n},$$

and

$$\mathbf{W}_2 = \begin{bmatrix} \mathbf{0}_{n_1 \times n_1} & \frac{1}{n-n_1} \mathbf{1}_{n_1} \mathbf{1}_{n_2}^\top & \cdots & \frac{1}{n-n_1} \mathbf{1}_{n_1} \mathbf{1}_{n_K}^\top \\ \frac{1}{n-n_2} \mathbf{1}_{n_2} \mathbf{1}_{n_1}^\top & \mathbf{0}_{n_2 \times n_2} & \cdots & \frac{1}{n-n_2} \mathbf{1}_{n_2} \mathbf{1}_{n_K}^\top \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n-n_K} \mathbf{1}_{n_K} \mathbf{1}_{n_1}^\top & \frac{1}{n-n_K} \mathbf{1}_{n_K} \mathbf{1}_{n_2}^\top & \cdots & \mathbf{0}_{n_K \times n_K} \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

**Remark 1:** If  $\lambda_2 = 0$ , then model (2.1) becomes a simplified version of the models studied in Bramoullé et al. (2009) and Lee (2007). The main difference is that we do not include group-specific fixed effects here for the sake of simplicity<sup>2</sup>.

<sup>1</sup>It is a convention in the literature of social interaction models to assume that the individual characteristics  $X$  are nonstochastic; see Lee (2004) and Lee (2007), among many others.

<sup>2</sup>The identification results can be derived similarly for models with fixed effects but estimation procedure would be much more complicated; see Lee (2007) for more details. We leave model (2.1) with group-specific fixed effects as a future research direction.



## 2.2.2 The Microfoundation

In this subsection, we present a network game with limited information of outside groups as a microfoundation for model (2.1) following the literature (Bramoullé et al. (2007)). Consider a network game in which each individual maximizes his utility by setting the optimal level of  $y_i$ . We assume that any individual  $i$  has full information of other individuals in his group but only knows the average value of the economic outcome outside his group, i.e.,  $\mathcal{F}_i = \{\pi_i, W_{G(i)}, \mathbf{Y}_{G(i)}, \mathbf{X}_{G(i)}, \bar{y}_{-G(i)}\}$ , where  $\pi_i$  is the individual-specific heterogeneity in marginal return of  $y_i$ ,  $\mathbf{Y}_{G(i)}$  is a  $n_{G(i)}$ -dimensional vector of economic outcomes of the group  $G(i)$  and  $\mathbf{X}_{G(i)}$  is defined in the similar fashion. Each individual  $i$  is supposed to have the following utility function:

$$u_i(y_i; \mathcal{F}_i) = \underbrace{(\pi_i + \lambda_1 \sum_{j \in G(i)} W_{G(i),ij} y_j + \lambda_2 \bar{y}_{-G(i)})}_{\text{benefit}} y_i - \underbrace{\frac{1}{2} y_i^2}_{\text{cost}}, \quad (2.3)$$

where the term  $(\pi_i + \lambda_1 \sum_{j \in G(i)} W_{ij}^{G(i)} y_j + \lambda_2 \bar{y}_{-G(i)})$  measures the marginal return of  $y_i$ . It is noteworthy that individual's marginal return now depends not only on his in-group friends but also the average value of the economic variable outside his group. From the first order condition, the individual  $i$ 's best response function is given by:

$$y_i = \pi_i + \lambda_1 \sum_{j \in G(i)} W_{G(i),ij} y_j + \lambda_2 \bar{y}_{-G(i)}. \quad (2.4)$$

If we let  $\pi_i = x_i^\top \beta + \epsilon_i$ , the best response function (5) becomes the econometric model (2.1). We next characterize the unique interior Nash equilibrium of the network game defined above.

**Assumption 1.** *The adjacency matrix  $W_k$  is row-normalized with  $W_{k,ij} \geq 0$ ,  $W_{k,ii} = 0$  for  $k = 1, \dots, K$  and  $1 \leq i \leq j \leq n_k$ .*

**Assumption 2.**  $|\lambda_1| + |\lambda_2| < 1$ .

Assumption 1 is standard in the literature of social interaction models (e.g., Lee (2004), Bramoullé et al. (2009) and Liu and Lee (2010)). Assumption 1 requires that the group-specific adjacency matrices to be row-normalized and individuals do not link to themselves. Assumption 2 restricts the sum of the absolute values of the in-group and out-group social interaction effects, which ensures the Nash equilibrium of the network game is unique.

**Proposition 1.** *If Assumptions 1 and 2 hold, the matrix  $(\mathbf{I} - \lambda_1 \mathbf{W}_1 - \lambda_2 \mathbf{W}_2)$  is invertible and the network game with payoff function (2.3) has a unique interior Nash equilibrium in pure strategies:*

$$\mathbf{Y} = (\mathbf{I} - \lambda_1 \mathbf{W}_1 - \lambda_2 \mathbf{W}_2)^{-1} \mathbf{\Pi},$$

where  $\mathbf{\Pi} = (\pi_1, \dots, \pi_n)^\top$ .

**Proof:** See the online appendix.

## 2.3 Identification and Estimation

### 2.3.1 Identification

In this subsection, we show that the parameters in model (2.1) are identified under a set of mild assumptions. Let  $\theta = (\lambda_1, \lambda_2, \beta^\top)^\top$  denote the vector of true parameters.

**Assumption 3.**  $\beta_i \neq 0$  for all  $i = 1, \dots, p$ .

**Assumption 4.** For  $i = 1, \dots, n$ ,  $v_i$  is *i.i.d* distributed with  $E[v_i] = 0$  and  $\text{Var}(v_i) = \sigma_\epsilon^2 < \infty$ .

Assumption 3 ensures that all individual characteristics can be used as valid instrumental variables. Assumption 4 requires that the error terms are *i.i.d*. Both assumptions have been made in most previous studies (Kline and Tamer (2020)). The next proposition establishes the identifiability of the parameters.

**Proposition 2.** *If Assumptions 1, 2, 3 and 4 hold, the parameters of interest  $\theta = (\lambda_1, \lambda_2, \beta^\top)^\top$  are identified.*

**Proof:** See the online appendix.

### 2.3.2 Estimation

We next discuss the estimation of model (2.1). Given the fact that the OLS estimators are inconsistent because of the famous reflection problem (Manski (1993)), we propose to estimate the parameters using the 2SLS method developed in Kelejian and Prucha (1998). Let  $\mathbf{Z} = (\mathbf{W}_1\mathbf{Y}, \mathbf{W}_2\mathbf{Y}, \mathbf{X})$  denote the design matrix of model (2.2) and  $\mathbf{H}$  denote the matrix of instrumental variables, for example,  $\mathbf{H} = (\mathbf{W}_1\mathbf{X}, \mathbf{W}_2\mathbf{X}, \mathbf{X})$ . The 2SLS estimators are then given by:

$$\hat{\theta}_{2SLS} = (\mathbf{Z}^\top \mathbf{P}_H \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{P}_H \mathbf{Y}, \quad (2.5)$$

where  $\mathbf{P}_H = \mathbf{H}(\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top$ . Next we establish the asymptotic properties of the proposed 2SLS estimators.

**Assumption 5.** *There exists a generic positive constant  $c$  and  $s_k$  such that  $\lim_{n \rightarrow \infty} \frac{n_k}{n} = s_k > c$  for all  $k = 1, \dots, K$ .*

**Assumption 6.** *The column sums of the group-specific adjacency matrices  $W_k$ ,  $k = 1, \dots, K$  are bounded uniformly.*

**Assumption 7.** *The nonstochastic matrix  $\mathbf{X}$  have full column rank and its elements are bounded in absolute values uniformly.*

**Assumption 8.** *The matrix of instrumental variables  $\mathbf{H}$  has full column rank  $k \geq p + 2$  for all  $n$  large enough. In addition,  $\mathbf{H}$  consists of a subset of the linearly independent columns of  $(\mathbf{X}, \mathbf{W}_1\mathbf{X}, \mathbf{W}_2\mathbf{X}, \mathbf{W}_1\mathbf{W}_1\mathbf{X}, \mathbf{W}_2\mathbf{W}_2\mathbf{X} \dots)$ , where the subset contains at least the linearly independent columns of  $(\mathbf{X}, \mathbf{W}_1\mathbf{X}, \mathbf{W}_2\mathbf{X})$ .*

**Assumption 9.**  *$Q_{HH} = \lim_{n \rightarrow \infty} n^{-1} \mathbf{H}'\mathbf{H}$  exists and is finite and nonsingular. Furthermore,  $Q_{HZ} = \lim_{n \rightarrow \infty} n^{-1} \mathbf{H}'\mathbf{Z}$  exists and is finite and has full column rank.*

Assumption 5 requires that each group contains a substantial number of individuals, which is reasonable for most empirical applications. Furthermore, this condition together with Assumption 6 ensure that the matrices  $\mathbf{W}_1$  and  $\mathbf{W}_2$  have uniformly bounded row and column sums. Assumptions 5-9 are standard in the literature of social interaction models, e.g., Kelejian and Prucha (1998) and Liu and Saraiva (2015). The asymptotic distribution of the 2SLS estimators are given in Proposition 3.

**Proposition 3.** *If the data is generated by model (2.1) and Assumptions 1-9 hold, then*

$$\sqrt{n}(\hat{\theta}_{2SLS} - \theta) \xrightarrow{d} N(0, [Q_{HZ}^\top Q_{HH}^{-1} Q_{HZ}]^{-1}).$$

Notice that  $Q_{HZ}$  and  $Q_{HH}$  can be calculated directly using observed data, so it is straightforward to conduct statistical inference on  $\lambda_1$  and  $\lambda_2$  with the help of general  $t$  tests.

## 2.4 Monte Carlo Simulations

To investigate the finite sample performance of the proposed estimators, we conduct Monte Carlo simulations based on the following specification:

$$y_i = \lambda_1 \sum_{j \in G(i)} W_{G(i),ij} y_j + \lambda_2 \bar{y}_{-G(i)} + x_{i1} \beta_1 + x_{i2} \beta_2 + \epsilon_i. \quad (2.6)$$

We consider two sets of parameters, which represent cases of weak out-group effect and strong out-group effect, respectively: (1)  $\lambda_1 = 0.60$ ,  $\lambda_2 = 0.20$  and  $\beta_1 = \beta_2 = 1$ ; (2)  $\lambda_1 = 0.20$ ,  $\lambda_2 = 0.60$  and  $\beta_1 = \beta_2 = 1$ . The individual characteristics  $x_{i1}$  and  $x_{i2}$  are drawn from independent  $N(0, 2)$  distributions and the error term  $\epsilon_i$  is drawn from standard normal distributions. When implementing the 2SLS method, we let  $\mathbf{H} = (\mathbf{X}, \mathbf{W}_1 \mathbf{X}, \mathbf{W}_2 \mathbf{X})$ . We fix the group size to be 50 and consider three different settings:  $n = 100, 200, 400$ , which consist of 2, 4, 8 groups, respectively. The group-specific adjacency matrices  $W_k$  are constructed

following the specification in Liu and Lee (2010): for the  $i$ th row of  $W_k$  ( $i = 1, \dots, 50$ ), we draw a integer  $m_{ki}$  randomly from the set of integers  $[0, 1, 2, 3, 4]$ . If  $i + m_{ki} < 50$  we set the  $(i + 1)$ th,  $\dots$ ,  $(i + m_{ki})$ th elements of the  $i$ th row of  $W_k$  to be ones and the rest elements in that row to be zeros. Otherwise, the entries of ones will be wrapped around such that the first  $(m_{ki} - 50)$  entries of the  $i$ th row will be ones. In the case of  $m_{ki} = 0$ , the  $i$ th row of  $W_k$  will have all zeros. We then normalize the matrix  $W_k$  by its row sums. The number of repetitions in each experiment is 1000. The simulation results are reported in Table 2.1.

Table 2.1: Finite Sample Performance of the 2SLS Estimators (1000 draws)

Parameters	$n = 100$		$n = 200$		$n = 400$	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
<i>Case 1</i>						
$\lambda_1 = 0.6$	0.5959	0.0362	0.6002	0.0331	0.5991	0.0098
$\lambda_2 = 0.2$	0.1876	0.0880	0.1956	0.1640	0.1892	0.1098
$\beta_1 = 1$	0.9927	0.0625	0.9993	0.0535	0.9997	0.0289
$\beta_2 = 1$	0.9958	0.0634	0.9991	0.0519	0.9994	0.0287
<i>Case 2</i>						
$\lambda_1 = 0.2$	0.2003	0.0236	0.2002	0.0162	0.2011	0.0117
$\lambda_2 = 0.6$	0.6066	0.1309	0.6018	0.0898	0.6018	0.0865
$\beta_1 = 1$	1.0002	0.0544	1.0007	0.0364	0.9988	0.0242
$\beta_2 = 1$	0.9991	0.0538	0.9993	0.0368	1.0003	0.0250

The simulation results in Table 2.1 show that the 2SLS estimation method works well for our model as both the bias and the standard error of the estimates are relatively small compared with their true values. We next investigate the estimation bias of the in-group social interaction effect if the out-group effect is ignored. In this case, we adopt the standard 2SLS estimation method in Kelejian and Prucha (1998) for estimation and take  $\mathbf{H} = (\mathbf{X}, \mathbf{W}_1\mathbf{X}, \mathbf{W}_1^2\mathbf{X})$  as instrumental variables. The estimation results are shown in Table 2.2.

Table 2.2: Simulation Results of the Mis-specified Model (1000 draws)

Parameters	$n = 100$		$n = 200$		$n = 400$	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
<i>Case 1: <math>\lambda_2 = 0.2</math></i>						
$\lambda_1 = 0.6$	0.5793	0.0688	0.5799	0.0401	0.5904	0.0207
$\beta_1 = 1$	0.9646	0.0941	0.9712	0.0596	0.9783	0.0375
$\beta_2 = 1$	0.9671	0.0920	0.9671	0.0609	0.9781	0.0392
<i>Case 2: <math>\lambda_2 = 0.6</math></i>						
$\lambda_1 = 0.2$	0.3645	0.0979	0.3206	0.0908	0.2845	0.0735
$\beta_1 = 1$	1.1067	0.1801	1.0765	0.1208	1.0574	0.0888
$\beta_2 = 1$	1.1082	0.1731	1.0714	0.1194	1.0565	0.0901

The results in Table 2.2 indicate that ignoring the our-group social interaction effect will lead to substantiate estimation bias of the in-group social interaction effect. This problem is especially severe when the out-group effect is large (Case 2). In this sense, the model proposed in this paper can become an appealing choice for empiricists to deal with potential out-group social interaction effect in the data.

## 2.5 Conclusion

In this paper, we study a new class of social interaction models with both in-group and out-group effects. We provide a network game with limited information of outside groups which rationalizes the econometric model. We show that the parameters of interest are identified under a set of mild conditions. We propose to estimate the model using the 2SLS method developed in Kelejian and Prucha (1998) and establish the asymptotic properties of the estimators. We investigate the finite sample properties of the 2SLS estimators through Monte Carlo simulations which show the estimation method performs very well.

## Chapter 3

Semiparametric Quantile Panel

Regression with Grouped

Heterogeneity

## 3.1 Introduction

In recent years, quantile panel regression models have attracted considerable interest in both theoretical and applied econometrics as they can take advantage of panel data structure but still reserve the merits of quantile regression. Even though substantial progress has been made by previous studies, the current literature still cannot simultaneously handle several important features of real-world data sets. First, almost all existing literature uses individual fixed effects to control heterogeneity, which has been suggested by many empirical studies as problematic if the heterogeneity can directly affect the data-generating process (Lee et al. (1997), Durlauf et al. (2001), Cavalcanti et al. (2011) and Zhu et al. (2016)). Second, most prior studies focused on linear models, which prevents them from capturing potentially time-varying or nonlinear quantile effects.

Motivated by such facts, we propose a semiparametric quantile panel regression model with grouped heterogeneity that can account for the above two issues. Our model is fairly general as it includes several prior models as special cases, such as Wang et al. (2009), Cai et al. (2018) and Zhang et al. (2019), etc. The key feature of our model is that the conditional quantile function consists of both a parametric component and a nonparametric component: the parametric component takes a linear structure, which captures effects of explanatory variables that do not vary across different time periods; the nonparametric component takes a varying coefficient structure, which captures effects that evolve with certain time-varying variables. At the same time, we allow the unobserved grouped heterogeneity to affect both parametric and functional coefficients by assuming individuals belonging to different latent groups have heterogeneous coefficients.

The flexible setting of our model brings new challenges for estimation because not only the parameters of interest but also individuals' group memberships have to be estimated. To meet these requirements, we develop a series-based estimation method that mimics the  $k$ -means clustering method in the statistics literature. We establish the asymptotic properties of the estimators of both parametric and functional coefficients as well as the



group memberships. More specifically, we show the asymptotic normality for the estimators of the parametric coefficients and the convergence rate of the functional coefficients, which can achieve the optimal convergence rate in Stone (1982) under some regularity conditions. We also propose an information criterion to estimate the true number of groups using observed panel data. The finite sample performance of the estimation method and the information criterion are investigated through Monte Carlo simulations. The results show that both work very well.

We illustrate the usefulness of our model by applying it to study an important topic in the empirical growth literature: the effect of foreign direct investment (FDI) on economic growth. Many previous studies have looked into this question using various econometric methods and data sources; see Carkovic and Levine (2005), Kottaridi and Stengos (2010) and Cai et al. (2018), among many others. However, they all ignored the potential grouped heterogeneity in such effect<sup>1</sup>, which is likely the main reason that their estimates are small and insignificant. Based on the most recent data, we show that the effect of FDI on economic growth is in fact heterogeneous, large and significant, especially for low-income countries. In addition, we find that the scale of the effect decreases as GDP per capita increases, which is also ignored by prior studies.

Our paper contributes to the literature in three aspects. First, our paper complements the literature on quantile panel regression by considering a semiparametric model with grouped heterogeneity, which is of much potential interest for empiricists. Second, our paper contributes to the thriving literature on panel models with grouped heterogeneity by generalizing the clustering estimation method to the semiparametric setting. Third, our paper also contributes to the literature on economic growth by building a new data set and disclosing some important new results which have been largely ignored in prior studies. The literature that is most relevant to our paper includes Wei et al. (2006), Wang et al. (2009) and Zhang et al. (2019) on quantile panel regression, Bonhomme and Manresa

---

<sup>1</sup>In this paper, we call the estimates in the literature "the pooling estimates."

(2015), Su et al. (2016), Su et al. (2019) and Zhang et al. (2019) on panel models with grouped heterogeneity, and Kottaridi and Stengos (2010), Durlauf et al. (2001) and Cai et al. (2018) on economic growth.

The rest of the paper is organized as follows. In Section 4.2, we introduce the model and compare it with related models in the literature. In Section 4.3, we describe the estimation method in detail. Section 4.4 studies the asymptotic properties of the proposed estimators. Section 4.5 reports the Monte Carlo simulation results. An empirical application is presented in Section 4.6. Finally, Section 4.7 concludes. Proofs are included in the Appendix.

**Notation:** For any matrix  $A$ , we denote  $A^{-1}$  as its Moore-Penrose generalized inverse and  $\|A\|$  as its Frobenius norm. If  $A$  is also a squared matrix, we denote  $\lambda_{\max}(A)$  and  $\lambda_{\min}(A)$  as its largest and smallest eigenvalues. 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  $h(\cdot)$ , we let  $\|h\|_2$  be its  $L_2$ -norm. For a set  $G$ , its cardinality is denoted by  $|G|$ . We let  $(N, T) \rightarrow \infty$  denote  $N$  and  $T$  diverging to infinity jointly,  $\xrightarrow{p}$  convergence in probability,  $\xrightarrow{d}$  convergence in probability. As a general rule for this paper, we write  $c$  as positive generic constants.

## 3.2 Model

Suppose we observe panel data of  $N$  individuals. For any individual  $i$ , we observe their data for  $T$  periods, i.e.,  $\{y_{it}, x_{it}, z_{it}, u_{it}\}_{t=1}^T$ . At a given quantile level  $\tau \in (0, 1)$ , we assume the response variable  $y_{it}$  is generated according to the following model:

$$y_{it} = x'_{it}\alpha_{i,\tau} + z'_{it}\beta_{i,\tau}(u_{it}) + e_{it,\tau}, \quad i = 1, \dots, N \text{ and } t = 1, \dots, T, \quad (3.1)$$

where  $y_{it}$  denotes the outcome of individual  $i$  in period  $t$ ,  $x_{it} = (x_{it,1}, \dots, x_{it,p})' \in \mathbb{R}^p$  and  $z_{it} = (z_{it,1}, \dots, z_{it,q})' \in \mathbb{R}^q$  are two vectors of observable explanatory variables,  $u_{it} \in \mathbb{R}$

is an observable smoothing variable<sup>2</sup> and we further assume  $u_{it} \in [0, 1]$  without loss of generality,  $e_{it,\tau}$  is the error term whose  $\tau$ -th quantile conditional on  $(x_{it}, z_{it}, u_{it})$  equals zero,  $\alpha_{i,\tau} = (\alpha_{i\tau,1}, \dots, \alpha_{i\tau,p})'$  is a  $p \times 1$  vector of individual-specific coefficients which does not vary with time and  $\beta_{i,\tau}(u) = (\beta_{i\tau,1}(u), \dots, \beta_{i\tau,q}(u))'$  is a vector of functional coefficients of  $u$ . To ensure identifiability, we assume that only  $x_{it}$  contains the constant term. For each individual  $1 \leq i \leq N$ , the parameters to be estimated are  $\theta_{i,\tau} = (\alpha'_{i,\tau}, \beta_{i,\tau}(u))' \in \mathbb{R}^{p+q}$ , and we let  $\theta_{i,\tau}^0$ ,  $\alpha_{i,\tau}^0$  and  $\beta_{i,\tau}^0$  denote their true values, respectively.

We assume that the model (3.1) is associated with a latent group structure with  $K^0$  groups, which means that there exists a disjoint partition of  $\{1, 2, \dots, N\}$ , denoted as  $\mathcal{G} = \{G_1, \dots, G_{K^0}\}$ , that uniquely classifies  $N$  individuals into  $K^0$  groups. Throughout the paper, we regard the true number of groups  $K^0$  as an unknown parameter to be estimated. By assumption, we have  $G_i \cap G_j = \emptyset$  for any  $i \neq j$  and  $\cup_{k=1}^{K^0} G_k = \{1, \dots, N\}$ . Following the literature, we assume each individual's group membership is time-invariant and independent of the explanatory variables. However, we allow both the number of groups  $K^0$  and the group structure  $\mathcal{G}$  to depend on the given quantile level  $\tau$ , i.e.,  $K^0(\tau)$  and  $\mathcal{G}(\tau)$ , but we suppress such dependence in notation on purpose for simplicity. We let  $N_k$  denote the number of individuals in group  $k$ , i.e.,  $N_k = |G_k|$ , for  $k = 1, \dots, K^0$ . We also assume that for any individual  $1 \leq i \leq N$ , the parameter  $\theta_{i,\tau}^0$  is determined by the individual  $i$ 's group membership:

$$\theta_{i,\tau}^0 = \sum_{k=1}^{K^0} \theta_{G_k,\tau}^0 \cdot 1\{i \in G_k\}, \quad (3.2)$$

where  $\theta_{G_k,\tau}^0$  is the parameter shared by all individuals in the group  $G_k$ , for all  $1 \leq k \leq K^0$ . At the same time, we assume  $\theta_{G_k,\tau}^0 \neq \theta_{G_l,\tau}^0$  for any  $k \neq l$ , which means that individuals belonging to different groups have heterogeneous parameters.

Here we briefly discuss the relationship between our model and some related models that have been studied in the literature.

---

<sup>2</sup>Here  $u_{it}$  could also be a random vector and the estimation method is similar to that used in the scalar case. However, this may come at the cost of the curse of dimensionality, which requires more observations to ensure estimation accuracy.

(1) If  $\beta_{i,\tau}(u) = 0$ , the model (3.1) becomes the linear quantile panel regression model with group heterogeneity without individual fixed effects, which is a simplified version of the models studied in Zhang et al. (2019):

$$y_{it} = x'_{it}\alpha_{i,\tau} + e_{it,\tau}$$

where  $\alpha_{i,\tau} = \sum_{k=1}^{K^0} \alpha_{G_k,\tau} \cdot 1\{i \in G_k\}$ .

(2) If the grouped heterogeneity is only associated with the intercept term, the model (3.1) becomes the one studied in Gu and Volgushev (2019):

$$y_{it} = x'_{it}\alpha_\tau + \lambda_{i,\tau} + e_{it,\tau},$$

where  $\lambda_{i,\tau}$  is the individual fixed effect and  $\lambda_{i,\tau} = \sum_{k=1}^{K^0} \lambda_{G_k,\tau} \cdot 1\{i \in G_k\}$ .

(3) If there does not exist any grouped heterogeneity and  $z_{it} = 1$ , the model (3.1) becomes the partially linear quantile panel regression model studied in He et al. (2002):

$$y_{it} = x'_{it}\alpha_\tau + \beta_\tau(u_{it}) + e_{it,\tau}.$$

(4) If there does not exist any grouped heterogeneity, the model (3.1) becomes the semiparametric quantile panel regression model studied in Wang et al. (2009):

$$y_{it} = x'_{it}\alpha_\tau + z'_{it}\beta_\tau(u_{it}) + e_{it,\tau}.$$

### 3.2.1 Motivating Applications

In this subsection, we present two potential empirical applications for the model (3.1).

#### Application 1 (the Effect of FDI on Economic Growth)

Evaluating the effect of FDI on economic growth is an important research topic in the macroeconomics literature; see Kottaridi and Stengos (2010) for a comprehensive review.

However, all previous literature has ignored the potential unobserved heterogeneity which can cause countries to have different economic growth patterns (Durlauf et al. (2001)). Such heterogeneity can be attributed to culture, institution and other unobserved or immeasurable factors. Following the literature, we propose to study the quantile effect of FDI on economic growth using the following model:

$$y_{it} = \alpha_{i\tau,1} + \alpha_{i\tau,2} \log\left(\left(\frac{I^d}{Y}\right)_{it}\right) + \alpha_{i\tau,3}n_{it} + \alpha_{i\tau,4}h_{it} + \alpha_{i\tau,5}h_{it} \cdot \left(\frac{I^f}{Y}\right)_{it} + \beta_{i,\tau}(u_{it}) \cdot \left(\frac{I^f}{Y}\right)_{it} + e_{it,\tau},$$

where  $y_{it}$  is the growth rate of GDP per capita of country  $i$  in period  $t$ ,  $\left(\frac{I^d}{Y}\right)_{it}$  is the ratio of the domestic investment to the GDP of country  $i$  in period  $t$ ,  $\left(\frac{I^f}{Y}\right)_{it}$  is the ratio of foreign direct investment to the GDP of country  $i$  in period  $t$ ,  $n_{it}$  is the population growth rate,  $h_{it}$  is the human capital measured by the mean years of schooling, and  $u_{it}$  is the GDP per capita of country  $i$  in period  $(t - 1)$ .

In this model, the main purpose of introducing the functional coefficient  $\beta_{i,\tau}(u_{it})$  is to address the potentially nonlinear and time-varying effect of FDI on economic growth. This is because the effect is likely to depend on the absolute level of economic development, which is measured by GDP per capita in the last period. We introduce the grouped heterogeneity because prior studies have shown its existence (Durlauf et al. (2001) and Kottaridi and Stengos (2010)) and ignoring such heterogeneity may result in inconsistent estimates (Su et al. (2016)). A detailed analysis of this empirical application is conducted in Section 4.6.

### **Application 2 (the Effect of Investor Sentiment on Stock Returns)**

Behavioral finance has attracted considerable attention in the last few decades. One important finding in this field is that investor sentiment can significantly affect stock returns; see Baker and Wurgler (2006) and Schmeling (2009), among many others. Based on this observation, various econometric models have been adopted in prior studies to quantify such effect. Ni et al. (2015) considered an interesting linear quantile panel

regression model as follows:

$$r_{it+1} = \beta_{\tau} \text{SENT}_{it} + x'_{it} \alpha_{\tau} + e_{it,\tau},$$

where  $r_{it+1}$  denotes the stock return of firm  $i$  in period  $t + 1$ ,  $\text{SENT}_{it}$  measures investor sentiment, and  $x_{it}$  is a vector of firm-specific explanatory variables. Even though Ni et al. (2015) did not address the grouped heterogeneity in their econometric model, they found strong evidence indicating that the effect of investor sentiment on stock returns is heterogeneous across different pre-specified subgroups. For example, they found that investor sentiment has a larger impact on firms with small market values. However, regression analysis based on a pre-specified group structure is somewhat arbitrary since the true group structure can be caused by unobserved factors, such as the public's preference of CEOs and the quality of corporate governance. Therefore, we can instead use the following model:

$$r_{it+1} = \beta_{i,\tau}(u_{it}) \cdot \text{SENT}_{it} + x'_{it} \alpha_{i,\tau} + e_{it,\tau},$$

where  $u_{it}$  is some variable which affects the effect of investor sentiment on stock returns, e.g., trading volume, and the stocks are classified by a data-driven estimation method rather than arbitrary designation.

## 3.3 Estimation

### 3.3.1 Series Approximation

We propose estimating the model (3.1) using a series-based clustering method. Let  $P(u) = (P_1(u), \dots, P_J(u))'$  denote the vector of basis functions on  $[0, 1]$ . Some basis functions that have been widely used in the literature include B-splines, Legendre polynomials and power series. We choose B-spline polynomials for simulation and empirical studies because they have stable numerical performance and are computationally easy; see

Chen (2007) and Schumaker (2007) for more details on the series estimation method.

Given the vector of basis functions  $P(u)$ , we approximate  $\beta_{il,\tau}^0(u)$  for all  $i = 1, \dots, N$  and  $l = 1, \dots, q$  by  $\beta_{il,\tau}^0(u) \approx \sum_{s=1}^J P_s(u) \gamma_{il,s,\tau}^0 = P(u)' \gamma_{il,\tau}^0$ , where  $\beta_{il,\tau}^0(u)$  is the  $l$ -th element of the  $q \times 1$  vector  $\beta_{i,\tau}^0(u)$ ,  $\gamma_{il,\tau}^0 = (\gamma_{il,1,\tau}^0, \dots, \gamma_{il,J,\tau}^0)'$  is a  $J \times 1$  vector such that  $P(u)' \gamma_{il,\tau}^0$  can approximate  $\beta_{il,\tau}^0(u)$  sufficiently well (which will be defined formally in Section 4.4) for all  $i = 1, \dots, N$  and  $l = 1, \dots, q$ . Therefore, the model (3.1) can be written as:

$$y_{it} = \sum_{s=1}^p x_{it,s} \alpha_{is,\tau}^0 + \sum_{l=1}^q \sum_{s=1}^J z_{it,l} P_s(u_{it}) \gamma_{il,s,\tau}^0 + \epsilon_{it,\tau} = x'_{it} \alpha_{i,\tau}^0 + w'_{it} \gamma_{i,\tau}^0 + \epsilon_{it,\tau}, \quad (3.3)$$

where  $w_{it} = (z_{it,1} P(u_{it})', \dots, z_{it,q} P(u_{it})')' \in \mathbb{R}^{qJ \times 1}$ ,  $\gamma_{i,\tau}^0 = (\gamma_{i1,\tau}^0, \dots, \gamma_{iq,\tau}^0)' \in \mathbb{R}^{qJ \times 1}$  and  $\epsilon_{it,\tau} = e_{it,\tau} + z'_{it} \beta_{i,\tau}^0 - w'_{it} \gamma_{i,\tau}^0$ .

### 3.3.2 Implementation

Since we assume that the parameters of interest exhibit grouped heterogeneity, we need to estimate the group-specific parameters  $\theta_{G_k,\tau}^0 = (\alpha_{G_k,\tau}^0, \beta_{G_k,\tau}^0)'$  as well as the group memberships  $g_\tau^0 = \{g_{1,\tau}^0, \dots, g_{N,\tau}^0\} \in \Gamma_{K^0}^N$ , where  $\Gamma_{K^0}^N$  denotes the set of all partitions of  $N$  individuals into  $K^0$  groups  $\{G_1, \dots, G_{K^0}\}$ . Without loss of generality, we denote  $g_i = G_k = k$  if individual  $i$  is in the group  $G_k$ . We then have  $\theta_{i,\tau}^0 = \theta_{G_k,\tau}^0$  if  $g_i = k$  for all  $i = 1, \dots, N$ . Given the approximated linear model (3.3), we propose estimating the individual-specific parameter  $\theta_{i,\tau}^0$  and  $g_{i,\tau}^0$  for  $i = 1, \dots, N$  using the following steps.

**Step 1:** Estimate the group-specific parameters  $\delta_\tau$  and the group memberships  $g_\tau$  by minimizing the following objective function:

$$(\hat{\delta}_\tau, \hat{g}_\tau) = \arg \min_{\delta_\tau \in \mathbb{R}^{(p+qJ) \cdot K^0}, g_\tau \in \Gamma_{K^0}^N} \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \rho_\tau(y_{it} - x'_{it} \alpha_{g_i,\tau} - w'_{it} \gamma_{g_i,\tau}), \quad (3.4)$$

where  $\rho_\tau(u) = (\tau - \mathbf{1}(u < 0))u$  is the check function,  $g_\tau = \{g_{1,\tau}, \dots, g_{N,\tau}\}$  is a partition of  $N$  individuals into  $K^0$  groups,  $\delta_\tau = (\delta'_{G_1,\tau}, \dots, \delta'_{G_{K^0},\tau})'$  consists of the group-specific parameters

of  $K^0$  groups, where  $\delta_{G_{k,\tau}} = (\alpha'_{G_{k,\tau}}, \gamma'_{G_{k,\tau}})'$  for all  $k = 1, \dots, K^0$ .

**Step 2:** Recover  $\hat{\alpha}_{G_{k,\tau}}$  and  $\hat{\gamma}_{G_{k,\tau}}$  for all  $k = 1, \dots, K^0$  from  $\hat{\delta}_\tau$ , directly. The individual-specific parameters are determined by:

$$\begin{aligned} \hat{\alpha}_{i,\tau} &= \hat{\alpha}_{G_{k,\tau}}, & \text{if } \hat{g}_i = k \text{ for and } l = 1, \dots, q \\ \hat{\beta}_{il,\tau}(u) &= P(u)' \hat{\gamma}_{G_{kl,\tau}}. \end{aligned} \tag{3.5}$$

The vector of the functional coefficients is then given by  $\hat{\beta}_{i,\tau}(u) = (\hat{\beta}_{i1,\tau}(u), \dots, \hat{\beta}_{iq,\tau}(u))'$  for all  $i = 1, \dots, N$ .

The main intuition behind the optimization problem (3.4) is to find  $K^0$  vectors of the group-specific parameters and classify  $N$  individual-specific parameters by minimizing the quantile loss function. However, solving the optimization problem (3.4) directly is infeasible when  $N$  is large since the number of partitions in the set  $\Gamma_{K^0}^N$  is too large to search exhaustively. To ease the computational burden, we adopt a two-step iterative algorithm which shares a similar spirit with the  $k$ -means clustering method and has been previously used in Bonhomme and Manresa (2015) and Zhang et al. (2019). The details of the algorithm are provided in the Appendix.

### 3.4 Asymptotic Properties

In this section, we discuss the asymptotic properties of the proposed estimators. We first introduce some notations. For all  $i = 1, \dots, N$  and  $t = 1, \dots, T$ , let  $f(e_{it,\tau}|x_{it}, z_{it})$  be the density function of  $e_{it,\tau}$  conditional on  $(x_{it}, z_{it})$  and  $F(e_{it,\tau}|x_{it}, z_{it})$  be the cumulative distribution function of  $e_{it,\tau}$  conditional on  $(x_{it}, z_{it})$ . Recall that  $w_{it} = (z_{it,1}P(u_{it})', \dots, z_{it,q}P(u_{it})')' \in \mathbb{R}^{qJ \times 1}$  for  $J \geq 1$ . We next introduce some assumptions that are sufficient to show the consistency of the estimators.

**Assumption 1.** (*Data Generating Process*)



- (i)  $(y_{it}, x_{it}, z_{it}, u_{it})$  is independent across different individuals  $i$  and time periods  $t$ , and identically distributed for all  $i \in G_k$ .
- (ii) For all  $1 \leq i \leq N$  and  $1 \leq t \leq T$ , we have: (1)  $x_{it}$  and  $z_{it}$  are bounded almost surely; (2)  $u_{it}$  has a bounded support on  $[0, 1]$ .
- (iii) The eigenvalues of the matrix  $E[z_{it}z'_{it}|u_{it} = u]$  are bounded and bounded away from zero and infinity uniformly for all  $u \in [0, 1]$ .
- (iv) The conditional density of  $u_{it}$  given  $x_{it}$  is uniformly bounded and bounded away from zero on the support of  $x_{it}$ .
- (v) The conditional density function  $f(e|x, z)$  is bounded and bounded away from zero and have a bounded first derivative in the neighborhood of zero.

Assumption 1 specifies the restrictions on the data generating process. Assumptions 1(i) specify the dependence structure of the data. In this paper, we consider the *i.i.d.* case for the sake of technical simplicity. It is noteworthy that the estimation method and the corresponding asymptotic results developed in this paper can be generalized to time series data, say, the  $\beta$ -mixing processes, with the cost of notational heaviness. Some Monte Carlo simulation results for time series are available upon request. The first part of Assumption 1(ii) imposes uniform moment restrictions on the explanatory variables  $x_{it}$  and  $z_{it}$ , which is standard in the literature on quantile regression models; see Koenker et al. (2017). The second part of Assumption 1(ii) is a standard assumption and a common practice for varying coefficient models and series estimation. This assumption can be easily satisfied by transforming  $u_{it}$  to some bounded random variables using a one-to-one mapping. Assumption 1(iii), (iv) and (v) are standard technical assumptions for quantile regression models and series estimation, which ensure the estimators are fully identified and thus well-defined.

**Assumption 2.** ( *Series Approximation* )

(i) The eigenvalues of  $E[P(u_{it})P(u_{it})']$  and  $\int_0^1 P(u)P(u)'du$  are bounded and bounded away from below and above by some generic constants. In addition, there exists a sequence of positive constants  $\zeta_0(J)$  such that  $\sup_{u \in [0,1]} \|P(u)\| \leq \zeta_0(J)$  and the  $L_2$  norm of  $P(u)$  is bounded and bounded away from below and above.

(ii) For  $1 \leq i \leq N$  and  $1 \leq l \leq q$ , there exists some  $\gamma_{il}^0 \in \mathbb{R}^J$  and  $\kappa > 0$  such that  $\sup_{u \in [0,1]} |\beta_{il,\tau}^0(u) - P(u)' \gamma_{il}^0| = O_p(J^{-\kappa})$ .

(iii)  $(N, T) \rightarrow \infty$ ,  $J \rightarrow \infty$ ,  $J^3 \zeta_0^2(J) (\log(T))^2 / T \rightarrow 0$ .

Assumption 2 relates to the series approximation theory. Assumption 2(i) is standard in the literature on series approximation method; see Chen (2007). The first part of this assumption ensures that the population design matrix of basis functions is well-defined, while the second part specifies the upper bound of the Euclidean norm of the vector of basis functions. Assumption 2(ii) assumes that  $\beta_{il,\tau}^0(u)$  can be approximated sufficiently well by basis functions, which further implies that the group-specific parameters  $\beta_{G_k,l,\tau}^0(u)$  can also be well-approximated. This assumption is standard in the series-based semiparametric estimation literature, for example, Newey (1997). Furthermore, it is worth noting that this assumption holds if the function being approximated belongs to the so-called Hölder ball, which is a subspace of smooth functions; see Chen (2007) for more details. Assumption 2(iii) imposes restrictions on the growth rate of  $N$  and  $T$ , which are useful for showing the consistency of the estimators.

**Assumption 3.** ( *Grouped Heterogeneity* )

(i) For any group  $G_k \in \mathcal{G} = \{G_1, \dots, G_{K^0}\}$ , there exists a generic constant  $c > 0$  such that

$$\lim_{N \rightarrow \infty} \sum_{i=1}^N 1\{g_{i,\tau}^0 = G_k\} = \frac{N_k}{N} = \pi_k > c.$$

(ii) There exists a generic constant  $c > 0$  such that either

$$\min_{G_k, G_l \in \mathcal{G}} \|\alpha_{G_k,\tau}^0 - \alpha_{G_l,\tau}^0\| > c$$

or

$$\min_{G_k, G_l \in \mathcal{G}} \left\| \beta_{G_k, \tau}^0 - \beta_{G_l, \tau}^0 \right\|_2 > c$$

or both hold, for any  $G_k, G_l \in \mathcal{G} = \{G_1, \dots, G_{K^0}\}$  and  $k \neq l$ .

Assumption 3 is about the unobserved grouped heterogeneity in the data. Both Assumption 3(i) and Assumption 3(ii) are standard in the literature of panel regression models with grouped heterogeneity (Su et al. (2016), Su et al. (2019), etc). Assumption 2(i) requires that each latent group consists of a substantive amount of individuals, which guarantees that the grouped heterogeneity in the data does not vanish asymptotically. Assumption 2(ii) assumes that the group-specific parameters are well-separated across different groups. It is worth mentioning that Assumption 2(ii) only requires that either parametric or nonparametric coefficients are separated across different groups.

A direct outcome of Assumption 3(ii) and 2(ii) is that the oracle group-specific parameters,  $\delta_{G_k, \tau}^0$ , are well-separated for any  $G_k, G_l \in \mathcal{G}$ , which is formally stated in the following lemma.

**Lemma 1.** *Suppose Assumptions 1, 2 and 3 hold, for any  $G_k, G_l \in \mathcal{G} = \{G_1, \dots, G_{K^0}\}$  and  $k \neq l$ , we have*

$$\left\| \delta_{G_k, \tau}^0 - \delta_{G_l, \tau}^0 \right\| > c > 0,$$

for some constant  $c$ , where  $\delta_{G_k, \tau}^0 = (\alpha_{G_k, \tau}^{\prime}, \gamma_{G_k, \tau}^{\prime})' \in \mathbb{R}^{(p+qJ) \times 1}$  and  $\gamma_{G_k, \tau}^0 = (\gamma_{G_k, 1, \tau}^{\prime}, \dots, \gamma_{G_k, q, \tau}^{\prime})' \in \mathbb{R}^{qJ \times 1}$ .

The above lemma implies that different groups have different values of  $\delta_{G_k, \tau}^0$ . Intuitively, when  $(N, T)$  is large, the estimates  $\hat{\delta}_{i, \tau}$  will converge to  $\delta_{i, \tau}^0$ , so the consistency follows under some regularity conditions. We are now ready to present the asymptotic properties of the proposed estimators. The following theorem formally establishes the consistency of the estimators of coefficients.

**Theorem 1.** *Suppose Assumptions 1, 2 and 3 hold and the correct number of groups  $K^0$  is known, let  $\hat{\mathcal{G}} = \{\hat{G}_1, \dots, \hat{G}_{K^0}\}$  be the estimated group memberships, then for all  $k = 1, \dots, K^0$ ,*

we have

$$\left\| \hat{\alpha}_{\hat{G}_{\sigma(k),\tau}} - \alpha_{G_k,\tau}^0 \right\| = o_p(1),$$

and

$$\left\| \hat{\beta}_{\hat{G}_{\sigma(k),l,\tau}} - \beta_{G_{kl},\tau}^0 \right\|_2 = o_p(1),$$

for all  $l = 1, \dots, q$ , where  $(\sigma(1), \sigma(2), \dots, \sigma(K))$  is a suitable permutation of  $(1, \dots, K^0)$ .

Theorem 2 gives the preliminary convergence rate of the estimators of both parametric and functional coefficients. The estimators of the parametric coefficients  $\hat{\alpha}_{\hat{G}_{\sigma(k),\tau}}$  and the estimators of the functional coefficients  $\hat{\beta}_{\hat{G}_{\sigma(k),l,\tau}}$  are consistent when  $T \rightarrow \infty$  and  $J \rightarrow \infty$ . The estimation error is determined by both  $T$  and  $J$ , while  $J \rightarrow \infty$  alone ensures that the approximation error is asymptotically negligible. It is noteworthy that we also find that the rates are only determined by  $T$  rather than  $N$ , which is consistent with the findings in Su et al. (2019). The reason we need to introduce the permutation  $\sigma$  here is that the estimation procedure 3.3.2 only generates a partition  $\hat{\mathcal{G}} = \{\hat{G}_1, \dots, \hat{G}_{K^0}\}$  of  $N$  individuals into  $K^0$  groups. The induced subgroups  $\{\hat{G}_1, \dots, \hat{G}_{K^0}\}$  are not labeled or ordered, so the permutation  $\sigma$  here can be understood as a one-to-one mapping from the estimated groups  $\hat{\mathcal{G}}$  to the latent group  $\mathcal{G}$ . The second part of Theorem 2 implies that the estimated individual group memberships  $\{\hat{g}_1, \dots, \hat{g}_N\}$  are consistent. Finally, it is worth pointing out that we need to know the correct number of groups *ex ante* in order to achieve the consistency of the estimators of both coefficients and group memberships. Otherwise the estimates will be inconsistent and misleading. The estimation of  $K^0$  will be discussed in Section 3.5.2. The next corollary establish the consistency of the estimators of group memberships.

**Corollary 1.** *Suppose Assumptions 1, 2 and 3 hold and the correct number of groups  $K^0$  is known, let  $\hat{g}_\tau = \{\hat{g}_{1,\tau}, \dots, \hat{g}_{N,\tau}\}$  be the estimated group memberships, for all  $i = 1, \dots, N$ , we have*

$$\lim_{T \rightarrow \infty} P(\hat{g}_{i,\tau} = g_{i,\tau}^0) = 1.$$

We next study the asymptotic distribution of the estimators of the parametric coefficients and the refined convergence rate of the functional coefficients. We first introduce some other notations. For each group  $G_k \in \mathcal{G}$ , we let  $f_{G_k, \tau}(e|x_{it}, z_{it}, u_{it})$  be the probability density function of  $e_{it, \tau}$  conditional on  $(x_{it}, z_{it}, u_{it})$  and  $F_{G_k, \tau}(e|x_{it}, z_{it}, u_{it})$  be the cumulative probability function of  $e_{it, \tau}$  conditional on  $(x_{it}, z_{it}, u_{it})$ . For the sake of notational simplicity, we let  $f_{G_k, \tau}$  and  $F_{G_k, \tau}$  denote  $f_{G_k, \tau}(e|x_{it}, z_{it}, u_{it})$  and  $F_{G_k, \tau}(e|x_{it}, z_{it}, u_{it})$  throughout the rest of this paper, respectively. We define  $\phi_{G_k}(z, u) = E[x_{it}|z_{it} = z, u_{it} = u]$  to be a function from  $\mathbb{R}^q \otimes \mathbb{R}$  to  $\mathbb{R}^p$  for all  $i \in G_k$ . Next, we let  $\mathcal{H}_{G_k}$  be the subspace of varying-coefficient-form functions on  $\mathbb{R}^q \times [0, 1]$  to  $\mathbb{R}$ , i.e.,

$$\mathcal{H}_{G_k} = \{h : h(z, u) = z_1 h_1(u) + \dots + z_q h_q(u), E[z_{it, j}^2 h_j(u_{it})^2] < \infty, \text{ for } i \in G_k \text{ and } j = 1, \dots, q\}.$$

We let  $m_{G_k, l}(z, u)$  be the projection of the  $l$ -th element of  $\phi_{G_k}(z, u) = E[x_{it}|z_{it} = z, u_{it} = u]$  onto the space  $\mathcal{H}_{G_k}$  for all  $i \in G_k$  and  $l = 1, \dots, p$ , which is defined by

$$m_{G_k, l}(z, u) = \arg \min_{h \in \mathcal{H}_{G_k}} E \left[ f_{G_k, \tau}(0) (\phi_{G_k, l}(z_{it}, u_{it}) - h(z_{it}, u_{it}))^2 \right].$$

In addition, we let  $M_{G_k}(z, u) = (m_{G_k, 1}(z, u), \dots, m_{G_k, p}(z, u))'$ . It is worth noting that for any  $m_{G_k, l}(z, u)$ , there exist corresponding  $\{h_{G_k, ls}(u), s = 1, \dots, q\}$  such that  $m_{G_k, l}(z, u) = \sum_{s=1}^q z_s h_{G_k, ls}(u)$ , which is ensured by definition. Finally, for all  $i \in G_k$ , we let

$$\Gamma_{G_k, \tau} = E \left[ f_{G_k, \tau}(0) (x_{it} - M(z_{it}, u_{it})) (x_{it} - M(z_{it}, u_{it}))' \right], \quad (3.6)$$

and

$$\Omega_{G_k, \tau} = E \left[ \tau(1 - \tau) (x_{it} - M(z_{it}, u_{it})) (x_{it} - M(z_{it}, u_{it}))' \right]. \quad (3.7)$$

We next introduce assumptions that are sufficient to establish the asymptotic distribution of  $\hat{\alpha}_{G_k, \tau}$  and the refined convergence rate of the functional coefficients  $\hat{\beta}_{G_k, \tau}$ .

**Assumption 4.** ( *Asymptotic Normality* )

(i) For all  $k = 1, \dots, K^0$ ,  $s = 1, \dots, q$  and  $l = 1, \dots, p$ , there exists some  $\eta_g^0 \in \mathbb{R}^J$  and  $\kappa > 0$  such that  $\sup_{u \in [0,1]} |h_{G_k,ls}(u) - P(u)' \eta_g^0| = O_p(J^{-\kappa})$ .

(ii) Let  $\hat{\delta}_\tau^*$  be the estimator to the optimization problem 3.4 such that  $g_\tau^0$  is known. The following condition holds  $\hat{\delta}_\tau = \hat{\delta}_\tau^* + o_p(c_{NT})$ , where  $c_{NT}$  is a sequence of real numbers.

(iii)  $J^3 \zeta_0^2(J) (\log(NT))^2 / (NT) \rightarrow 0$ ,  $J \rightarrow \infty$  and  $\sqrt{NT} \cdot J^{-\kappa} \rightarrow 0$ .

(iv) Both  $\Gamma_{G_{k,\tau}}$  and  $\Omega_{G_{k,\tau}}$  are positive definite for all  $k = 1, \dots, K^0$ .

Assumption 4(i) is similar to Assumption 2(ii), which ensures that  $h_{G_k,ls}(u)$  can also be approximated sufficiently well by basis functions. Assumption 4(ii) is a high-level assumption which ensures the asymptotic equivalence between the general estimator to the optimization problem 3.4 and the oracle estimator. This assumption can be implied by other low-level assumptions with more technical complexity. Assumption 4(iii) imposes restrictions on the joint convergence rates of both  $N$  and  $T$ . There are two implications of this assumption. First, it implies the asymptotic equivalence between the proposed estimators and the oracle estimators. Second, it ensures the approximation error is asymptotic negligible. Assumption 4(iv) ensures the covariance matrix is well-defined. The next theorem establishes the asymptotic distribution of the parametric coefficients and the refined convergence rate of the functional coefficients.

**Theorem 2.** Suppose Assumptions 1, 2, 3 and 4 hold, let  $\hat{\mathcal{G}} = \{\hat{G}_1, \dots, \hat{G}_{K^0}\}$  be the estimated group structure, for all  $k = 1, \dots, K^0$ , there exists a suitable permutation of  $(1, \dots, K^0)$ , denoted by  $\sigma = (\sigma(1), \sigma(2), \dots, \sigma(K^0))$ , such that

(i)

$$\sqrt{N_{\sigma(k)} T} (\hat{\alpha}_{\hat{G}_{\sigma(k),\tau}} - \alpha_{G_{k,\tau}}^0) \xrightarrow{d} N(0, \Gamma_{G_{k,\tau}}^{-1} \Omega_{G_{k,\tau}} \Gamma_{G_{k,\tau}}^{-1}),$$

(ii)

$$\left\| \hat{\beta}_{\hat{G}_{\sigma(k),\tau}^l} - \beta_{G_{k,\tau}^l}^0 \right\|_2 = O_p((N_k T)^{-1/2} J^{1/2} + J^{-\kappa}),$$

for all  $l = 1, \dots, q$ , where  $\Gamma_{G_k, \tau}$  and  $\Omega_{G_k, \tau}$  are defined in equations 3.6 and 3.7.

Theorem 2 establishes the asymptotic distribution of the parametric coefficients  $\hat{\alpha}_{\hat{G}_{\sigma(k), \tau}}$  as well as the refined convergence rate of the nonparametric coefficients  $\hat{\beta}_{\hat{G}_{\sigma(k), \tau}}$ . It is worth noting that when  $J$  has the order of  $(N_k T)^{1/(1+2\kappa)}$ , we have  $\left\| \hat{\beta}_{\hat{G}_{\sigma(k)l, \tau}} - \beta_{G_k l, \tau}^0 \right\|_2 = O_p((N_k T)^{-\kappa/(1+2\kappa)})$ , which achieves the optimal convergence rate for nonparametric estimators in Stone (1982). For the inference of  $\hat{\alpha}_{\hat{G}_{\sigma(k), \tau}}$ , there are two main approaches in the literature. The first approach is to estimate the asymptotic covariance matrix directly, which requires one to construct the empirical counterparts of  $\Gamma_{G_k, \tau}$  and  $\Omega_{G_k, \tau}$ . However, it is well known that estimating the conditional density function  $f_{G_k, \tau}(0)$  is a non-trivial task and the existing estimation methods can have poor finite sample performance because the estimates can be sensitive to the value of the bandwidth (Koenker and Machado (1999)). A more popular approach is to use the bootstrap method to construct the confidence interval based on resampled data, and we follow this approach.

## 3.5 Monte Carlo Simulations

### 3.5.1 Data Generating Processes

We consider four data generating processes (DGPs). For each DGP, we consider six settings of  $(N, T)$ , including (1)  $(N, T) = (50, 40)$ ; (2)  $(N, T) = (50, 80)$ ; (3)  $(N, T) = (50, 120)$ ; (4)  $(N, T) = (100, 40)$ ; (5)  $(N, T) = (100, 80)$  and (6)  $(N, T) = (100, 120)$ . We utilize these six settings to study the influence of sample size and length of panels on estimation accuracy. When estimating these models, we follow the estimation procedure proposed in subsection 3.3.2, which is based on a single quantile for estimation, and we consider  $\tau = 0.25, 0.5$  and  $0.75$ , separately.

**DGP 1:** In this benchmark data generating process, we assume  $y_{it}$  is generated according to the following specification:

$$y_{it} = x_{it}\alpha_i + z_{it}\beta_i(u_{it}) + \epsilon_{it} \quad (3.8)$$

for all  $i = 1, \dots, N$  and  $t = 1, \dots, T$ , where  $x_{it} \stackrel{iid}{\sim} \text{Unif}(-1, 1)$ ,  $z_{it} \stackrel{iid}{\sim} \text{Unif}(-1, 1)$ ,  $u_{it} \stackrel{iid}{\sim} \text{Unif}(0, 1)$  and  $\epsilon_{it} \stackrel{iid}{\sim} N(0, 1)$ . Therefore, the conditional quantile function of  $y_{it}$  is given by

$$Q_\tau(y_{it}|x_{it}, z_{it}, u_{it}) = x_{it}\alpha_i + z_{it}\beta_i(u_{it}) + \Phi^{-1}(\tau),$$

where  $\Phi^{-1}(\tau)$  is the inverse cumulative density function of the standard normal distribution. Notice that when  $\tau = 0.5$ , we have  $\Phi^{-1}(\tau) = 0$ . When we estimate the model, we always include the constant term along with  $x$  to accommodate the potentially nonzero intercept term. We next introduce the grouped heterogeneity on the model. In this benchmark DGP, we assume there exist two groups, i.e.,  $\mathcal{G} = \{G_1, G_2\}$  and each group contains exactly half of the individuals, i.e.,  $N_1 = N_2 = \frac{N}{2}$ . Furthermore, we make the following assumptions on



the group-specific parameters:

$$\alpha_i = \begin{cases} 1 & \text{if } i \in G_1, \\ 2 & \text{if } i \in G_2, \end{cases}$$

and

$$\beta_i(u) = \begin{cases} -2u^2 + 5u & \text{if } i \in G_1, \\ u^2 + 2u & \text{if } i \in G_2. \end{cases}$$

**DGP 2:** We let the second DGP share the same specification as DGP 1, i.e.,

$$y_{it} = x_{it}\alpha_i + z_{it}\beta_i(u_{it}) + \epsilon_{it} \quad (3.9)$$

for all  $i = 1, \dots, N$  and  $t = 1, \dots, T$ , where  $x_{it} \stackrel{iid}{\sim} \text{Unif}(-1, 1)$ ,  $z_{it} \stackrel{iid}{\sim} \text{Unif}(-1, 1)$ ,  $u_{it} \stackrel{iid}{\sim} \text{Unif}(0, 1)$  and  $\epsilon_{it} \stackrel{iid}{\sim} N(0, 1)$ . However, in this DGP, we increase the number of groups to three, i.e.,  $\mathcal{G} = \{G_1, G_2, G_3\}$  and the ratio of individuals within these three groups is fixed to  $N_1 : N_2 : N_3 = 3 : 3 : 4$ . In addition, we assume the group-specific parameters are given by

$$\alpha_i = \begin{cases} 1 & \text{if } i \in G_1, \\ 2 & \text{if } i \in G_2, \\ 3 & \text{if } i \in G_3, \end{cases}$$

and

$$\beta_i(u) = \begin{cases} -2u^2 + 5u & \text{if } i \in G_1, \\ u^2 + 2u & \text{if } i \in G_2, \\ 4u^2 - u & \text{if } i \in G_3. \end{cases}$$

Notice that we keep DGP 2 similar to DGP 1 except for the number of groups. Therefore, these two DGPs together can allow us to investigate the influence of the number of groups  $K^0$  on the finite sample performance of the estimation method as well as the information

criterion.

**DGP 3:** In this data generating process, we consider the following specification with heteroscedastic error

$$y_{it} = x_{it}\alpha_i + z_{it}\beta_i(u_{it}) + x_{it}\epsilon_{it}, \quad (3.10)$$

for all  $i = 1, \dots, N$  and  $t = 1, \dots, T$ , where  $x_{it} \stackrel{iid}{\sim} \text{Unif}(0, 1)$ ,  $z_{it} \stackrel{iid}{\sim} \text{Unif}(-1, 1)$ ,  $u_{it} \stackrel{iid}{\sim} \text{Unif}(0, 1)$  and  $\epsilon_{it} \stackrel{iid}{\sim} N(0, 1)$ . Since the model contains a heteroscedastic error term, the parametric coefficient  $\alpha_i$  is quantile-dependent. To be more clear, the conditional quantile function of  $y_{it}$  given  $(x_{it}, z_{it}, u_{it})$  is given by

$$Q_\tau(y_{it}|x_{it}, z_{it}, u_{it}) = x_{it}(\alpha_i + \Phi^{-1}(\tau)) + z_{it}\beta_i(u_{it}).$$

The group-specific parameters and the ratio of individuals in each group are the same as in DGP 2. By comparing the simulation results of DGP 2 and DGP 3, we can see how heteroscedastic errors influence the finite sample performance.

**DGP 4:** In this data generating process, we consider the following specification:

$$y_{it} = x_{it}\alpha_i + z_{it,1}\beta_{i,1}(u_{it}) + z_{it,2}\beta_{i,2}(u_{it}) + \epsilon_{it}, \quad (3.11)$$

for all  $i = 1, \dots, N$  and  $t = 1, \dots, T$ , where  $x_{it} \stackrel{iid}{\sim} \text{Unif}(-1, 1)$ ,  $z_{it,1} \stackrel{iid}{\sim} \text{Unif}(-1, 1)$ ,  $z_{it,2} \stackrel{iid}{\sim} \text{Unif}(-1, 1)$ ,  $u_{it} \stackrel{iid}{\sim} \text{Unif}(0, 1)$  and  $\epsilon_{it} \stackrel{iid}{\sim} N(0, 1)$ . The conditional quantile function of  $y_{it}$  given  $(x_{it}, z_{it}, u_{it})$  is

$$Q_\tau(y_{it}|x_{it}, z_{it}, u_{it}) = x_{it}\alpha_i + z_{it,1}\beta_{i,1}(u_{it}) + z_{it,2}\beta_{i,2}(u_{it}) + \Phi^{-1}(\tau).$$

We still assume there exist three groups of individuals, i.e.,  $\mathcal{G} = \{G_1, G_2, G_3\}$  and the ratio of individuals within these three groups is fixed to  $N_1 : N_2 : N_3 = 3 : 3 : 4$ . The group-specific

parameters are given by

$$\alpha_i = \begin{cases} 1 & \text{if } i \in G_1, \\ 2 & \text{if } i \in G_2, \\ 3 & \text{if } i \in G_3, \end{cases}$$

and

$$\beta_{i,1}(u) = \begin{cases} -u^2 + 4u & \text{if } i \in G_1, \\ u^2 + 2u & \text{if } i \in G_2, \\ 3u^2 & \text{if } i \in G_3, \end{cases}$$

and

$$\beta_{i,2}(u) = \begin{cases} \sin(\pi u) & \text{if } i \in G_1, \\ -\sin(\pi u) & \text{if } i \in G_2, \\ \cos(\pi u) & \text{if } i \in G_3. \end{cases}$$

The plots of the nonparametric coefficients in these four DGPs are shown in Figure 1.

### 3.5.2 The Determination of the Number of Groups

There are two main methods of determining the number of groups  $K^0$  in the literature. The first method is to use a BIC-type information criterion (Bonhomme and Manresa (2015) Su et al. (2016), Su et al. (2019), Wang et al. (2019) and Gu and Volgushev (2019), etc.). The main idea of this method is that the information criterion, which trades off between the model fitness and the model complexity, is able to select the correct number of groups consistently under some regularity conditions. The second one is the cross validation with averaging method, which has been used in Zhang et al. (2019). We follow the literature of the first method by proposing a BIC-type information criterion to select  $K^0$ .

Let  $K_{\min}$  and  $K_{\max}$  be the possible minimum and maximum number of groups, respectively.

In most cases, the minimum number of groups is set to be one, meaning there exists no

grouped heterogeneity. However, currently there is no consensus on how to choose the maximum number of groups  $K_{\max}$  besides that it must be larger than the correct number of groups  $K^0$ .<sup>3</sup> We let  $\mathcal{K} = \{K_{\min}, K_{\min} + 1, \dots, K_{\max}\}$  denote the set of possible values for the number of groups.

**Step 1:** For a given quantile  $\tau$  and any  $K \in \mathcal{K}$ , estimate the group structure  $\hat{\mathcal{G}} = \{\hat{G}_1, \dots, \hat{G}_K\}$  and the group-specific parameters  $\hat{\delta}_\tau^{(K)}$  using the two-step estimation method in Section 4.3. Then calculate the following information criterion

$$IC(K, \tau) = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \rho_\tau(y_{it} - x'_{it} \hat{\alpha}_{\hat{g}_i, \tau}^{(K)} - w'_{it} \hat{\gamma}_{\hat{g}_i, \tau}^{(K)}) + c_{NT} \cdot (p + qJ)K, \quad (3.12)$$

where  $c_{NT}$  is a tuning parameter which depends on  $(N, T)$ .

**Step 2:** Choose  $\hat{K}_\tau$  such that

$$\hat{K}_\tau = \arg \min_{K \in \mathcal{K}} IC(K, \tau).$$

We use this information criterion to determine the number of groups both in the Monte Carlo simulations and in the empirical application.

### 3.5.3 Implementation Details

In this subsection, we describe the details of the implementation and evaluation procedures used in the Monte Carlo simulations. For each data generating process, we conduct 1000 repetitions of the simulation. In each repetition, we pretend the true number of groups is unknown and use the information criterion 3.12 to determine the number of groups  $K$ . When estimating the model, we assume  $K^0$  is known and follow the procedure in Section 4.3 for estimation. It is worth mentioning that such arrangement is standard in the literature;

---

<sup>3</sup>This is because if  $K_{\max}$  is strictly smaller than  $K^0$ , the coefficient estimators and the estimated group memberships will be inconsistent. On the other hand, if the number of groups used for estimation is larger than the correct number of groups, the coefficient estimators will still be consistent but individuals of a common group might be classified into two or more groups (Liu et al. (2019)).

see Su et al. (2016), Zhang et al. (2019) and Su et al. (2019), etc. When computing the information criterion, we let the tuning parameter  $c_{NT} = c \cdot \log(NT)/(NT)$ , where  $c$  is some positive constants. Although there are no established results for determining  $c$ , we find that the information criterion with  $c = 0.80$  performs satisfactorily among multiple alternatives<sup>4</sup>. In the estimation part, we use cubic B-splines to form basis functions and let  $J = \lfloor (NT)^{1/5} \rfloor$ , where  $\lfloor c \rfloor$  denotes the largest integer that is smaller than the constant  $c$ .

To evaluate the finite sample performance of the information criterion, we report the empirical probability of selecting different numbers of groups:  $\hat{P}(\hat{K} = K) = N_K/N_{\text{sim}}$ , where  $N_K$  is the number of repetitions in which  $K$  is selected by the information criterion and  $N_{\text{sim}}$  is the total number of repetitions which equals 1000 in our case. To evaluate the finite sample performance of the estimation method, we report the average rate of correct classification (CC Rate) and the average root mean square error (RMSE) for all individuals. As pointed out in Section 4.3, we use three strategies to generate thirteen initial values and pick the one that gives the lowest loss in order to ease the problem of local optimum. The average rate of correct classification (CC Rate) is given by

$$\text{CC rate} = \frac{1}{N_{\text{sim}}} \sum_{j=1}^{N_{\text{sim}}} \left\{ \frac{1}{N} \sum_{i=1}^N I(\hat{g}_{i,\tau}^{(j)} = g_{i,\tau}^{0,(j)}) \right\},$$

where  $j$  denotes the  $j$ -th repetition. It is worth noting that we need to find the proper permutation  $\sigma$  of  $(1, \dots, K^0)$  which is defined in Section 4.3 when calculating the CC rate. This is because the group indexes given by the estimation procedure may not coincide with the true group indexes in the data-generating process, making it necessary to define a suitable mapping from the estimated group indexes to the true group indexes. When the number of true groups is  $K^0$ , there will be a total number of  $K^0!$  possible permutations and we pick the one which gives the highest rate of correct classification. The RMSE for all individuals

---

<sup>4</sup>This value of  $c$  is chosen from the grid  $[0.1, 0.2, \dots, 2.0]$  based on multiple Monte Carlo simulations.

is given by

$$\text{RMSE}(\hat{\alpha}_\tau) = \frac{1}{N_{\text{sim}}} \sum_{j=1}^{N_{\text{sim}}} \sqrt{\frac{1}{N} \sum_{i=1}^N \left\| \hat{\alpha}_{i,\tau}^{(j)} - \alpha_{i,\tau}^{0,(j)} \right\|^2},$$

and

$$\text{RMSE}(\hat{\beta}_\tau) = \frac{1}{N_{\text{sim}}} \sum_{j=1}^{N_{\text{sim}}} \sqrt{\frac{1}{N} \sum_{i=1}^N \left\{ \frac{1}{R} \sum_{r=1}^R \left\| \hat{\beta}_{i,\tau}^{(j)}\left(\frac{r}{R}\right) - \beta_{i,\tau}^{0,(j)}\left(\frac{r}{R}\right) \right\|^2 \right\}},$$

where  $R$  is the number of evaluation points and we let  $R = 10000$  in the simulation.

### 3.5.4 Simulation Results

Table 1, Table 2 and Table 3 report the simulation results for four DGPs based on 1000 repetitions. Table 1 reports the empirical probability that a specific number of groups is selected at  $\tau = 0.50$ . The results show that the information criterion can effectively select the true number of groups with high probability when  $(N, T)$  is large. More specifically, we can see that  $T$  has a large impact on the performance of the information criterion while the impact of  $N$  is relatively moderate. The simulation results also show that the model complexity affects the performance of the information criterion as the empirical probability of selecting the true number of groups in DGP 4 is on average lower than that in DGP 1, 2 and 3. When we calculate the information criterion, we let  $c = 0.80$  and  $\tau = 0.50$ . It is worth noting that we can also use other quantiles to determine the number of groups as long as the group heterogeneity exists at this quantile. Furthermore, we also find that the value of the tuning parameter  $c$  plays an important role in determining the finite sample performance of the information criterion, especially when the length of panels  $T$  is relatively small. If  $c$  is too small, the number of groups  $\hat{K}$  selected by the information criterion will be larger than the true  $K^0$ , and the reverse conclusion holds if  $c$  is too large. We have also evaluated the finite sample performance of the information criterion using different values of  $c$ , and these results are available upon request. In practice, we recommend researchers try different values of  $c$  when determining the number of groups. As mentioned above, using  $\hat{K}$  that is larger than  $K^0$  usually does not harm consistency, so it can be a good idea to choose

the number of groups such that the values of the estimated group parameters are no longer sensitive to a larger  $\hat{K}$ .

Table 2 reports the bias and the RMSEs for the estimates of the coefficients at  $\tau = 0.25, 0.50$  and  $0.75$ . The simulation results show that the estimators perform well in finite samples as both the bias and root mean square errors are small when  $(N, T)$  is sufficiently large in all four DGPs. In addition, it also shows that the RMSEs are on average smaller at  $\tau = 0.50$  than those at  $\tau = 0.25$  and  $0.75$ . Table 3 shows the empirical rate of correct classification which is defined in equation 3.5.3. It can be learned from the results that the empirical rate of correct classification increases when either  $N$  or  $T$  increases. In DGP 1, which is the simplest setting, the CC rate is above 99% when  $T$  equals 120 for both  $N = 50$  and  $N = 100$ , which means that on average less than one individual will be mis-classified. Similar patterns can also be found in other data generating processes.

In conclusion, these simulation results together have demonstrated that the information criterion and the proposed estimation method perform well in finite samples, so we follow the same procedures to determine  $K^0$  and estimate the model in our empirical application.

## 3.6 Empirical Application

### 3.6.1 Background

The evaluation of FDI flows on economic growth is an important research topic in macroeconomic literature. There are two main approaches studying such effect, using either country-level data or firm-level data. However, studies using firm-level data often suffer the sample selection problem, resulting in their failing to capture the positive spillover effects of FDI on the host country (Kottaridi and Stengos (2010)). In the seminal paper De Gregorio (1992), the authors first utilized country-level data to study the effect of FDI on economic growth. For the next two decades, this approach was adopted by multiple studies which used various data sources and econometric models, including Balasubramanyam et al. (1996), Zhang (2001), Durham (2004), and Carkovic and Levine (2005), among many others. We refer readers to Kottaridi and Stengos (2010) for a comprehensive review of the literature.

The econometric setting in our empirical application builds on the ones in Kottaridi and Stengos (2010) and Cai et al. (2018). Kottaridi and Stengos (2010) adopted the following specification as their benchmark model:

$$y_{it} = \alpha_0 + \alpha_1 D_j + \alpha_2 \log\left(\left(\frac{I^d}{Y}\right)_{it}\right) + \alpha_3 n_{it} + \alpha_4 \log(x_{it}) + \alpha_5 \left(\frac{I^f}{Y}\right)_{it} + \alpha_6 h_{it} + \epsilon_{it}, \quad (3.13)$$

where  $y_{it}$  is the growth rate of GDP per capita of country  $i$  in period  $t$ ,  $D_j$  is a dummy variable for different regions, which include Africa, America, Asia, EU, etc,  $\left(\frac{I^d}{Y}\right)_{it}$  is the ratio of domestic investment to GDP,  $n_{it}$  is the population growth rate,  $x_{it}$  is GDP per capita at the beginning of each period  $t$ ,  $\left(\frac{I^f}{Y}\right)_{it}$  is the ratio of domestic investment to GDP and  $h_{it}$  is the country-specific human capital.

Even though model 3.13 and its variants have been frequently used in the literature, there still exist two big problems. First, this specification fails to capture the grouped heterogeneity,



which is well-known to be important in the literature of economic growth. Ignoring the grouped heterogeneity in the data will result in inconsistent estimates and thus misleading policy implications. Second, this model assumes that the marginal effect of FDI is the same for countries of different levels of development and across different time periods, which can be restrictive in reality. It is likely that the effect of FDI on economic growth is more critical for developing countries than developed countries. Motivated by the second point, Cai et al. (2018) considered the following quantile panel regression model:

$$Q_\tau(y_{it}|u_i, X_{it}) = \alpha_i + \alpha_1 \log\left(\left(\frac{I^d}{Y}\right)_{it}\right) + \alpha_2 n_{it} + \alpha_3 h_{it} + \alpha_4 \left(\frac{I^f}{Y}\right)_{it} \cdot h_{it} + \beta(u_i) \cdot \left(\frac{I^f}{Y}\right)_{it}, \quad (3.14)$$

where  $u_i$  is the logarithm of GDP per capita of country  $i$  in the initial period. Compared with model 3.13, model 3.14 allows the marginal effect of FDI to depend on the initial value of GDP per capita. However, such setting can still be restrictive since the marginal effect of FDI should vary with the absolute value of GDP per capita ( $u_{it}$ ) at the current period. Furthermore, the problem of potential grouped heterogeneity is still left unsolved in model 3.14, which may harm the plausibility of the policy implications generated by the model.

The drawbacks in model 3.13 and 3.14 have motivated us to use the following model to evaluate the effect of FDI on economic growth:

$$y_{it} = \alpha_{i\tau,1} + \alpha_{i\tau,2} \log\left(\left(\frac{I^d}{Y}\right)_{it}\right) + \alpha_{i\tau,3} n_{it} + \alpha_{i\tau,4} h_{it} + \alpha_{i\tau,5} h_{it} \cdot \left(\frac{I^f}{Y}\right)_{it} + \beta_{i,\tau}(u_{it}) \cdot \left(\frac{I^f}{Y}\right)_{it} + e_{it,\tau}. \quad (3.15)$$

where  $u_{it}$  is GDP per capita of country  $i$  in period  $(t-1)$ ,  $h_{it} \cdot \left(\frac{I^f}{Y}\right)_{it}$  is the joint effect of human capital and FDI, and  $e_{it,\tau}$  is an error term whose  $\tau$ -th quantile conditional on  $(x_{it}, z_{it}, u_{it})$  is zero. It is easy to see that our model 3.15 includes Kottaridi and Stengos (2010) and Cai et al. (2018) as special cases. The major difference is that we allow the marginal quantile effect of FDI on economic growth to depend on the level of economic development, measured by GDP per capita in the last period, while in Kottaridi and Stengos (2010) and Cai et al. (2018), they either ignored such dependence or restrict it only to the initial period; see Cai

et al. (2018) for more details.

### 3.6.2 Data and Estimation

In this subsection, we describe the data sources and the estimation procedure. Our data set includes balanced panel data of 90 countries from 1970 - 2010. The variables contained in this data set include GDP per capita growth rate ( $y_{it}$ ), the ratio of domestic gross fixed capital formation to GDP ( $(\frac{I^d}{Y})_{it}$ ), human capital measured as mean years of schooling for total population ( $h_{it}$ ), population growth rate ( $n_{it}$ ), GDP per capita measured by US dollars in 2010 constant values ( $u_{it}$ ) and the ratio of FDI flow to GDP ( $(\frac{I^f}{Y})_{it}$ ). The data of GDP per capita growth rate, GDP per capita, population growth rate and domestic investment are downloaded from the World Development Indicators<sup>5</sup>, the data of mean years of schooling is collected from the Barro-Lee Dataset<sup>6</sup> and the World Development Indicators, and the data of FDI flow is taken from the United Nations Conference on Trade and Development (UNCTAD) website<sup>7</sup>, which is measured in 2010 constant dollars. Since time series data is vulnerable to macroeconomic shocks such as wars, financial crisis and so on, we follow the convention in the literature to take five-year moving averages<sup>8</sup>. In the end, we build a balanced panel data set with  $N = 90$  and  $T = 41$ . Some summary statistics are presented in Table 4.

It is worthwhile to mention that we scale the variable of GDP per capita to have values between zero and one since it is required by our estimation method in Section 4.3. We use the following transformation:

$$\tilde{u}_{it} = \Phi\left(\frac{u_{it} - \bar{u}}{\hat{\sigma}_u}\right), \quad (3.16)$$

where  $\Phi(\cdot)$  is the c.d.f. of the standard normal distribution,  $\bar{u}$  is the sample mean and  $\hat{\sigma}_u$  and the sample standard error of the variable  $u_{it}$ .

---

<sup>5</sup>See: <http://datatopics.worldbank.org/world-development-indicators/>

<sup>6</sup>See: <http://www.barrolee.com/>

<sup>7</sup>See: <https://unctad.org/en/Pages/DIAE/FDI%20Statistics/FDI-Statistics.aspx>

<sup>8</sup>This is a common practice in the literature of economic growth; see Durham (2004), Kottaridi and Stengos (2010) and Cai et al. (2018), etc.

Table 4: Some Summary Statistics of the Data Set

	Mean	S.D.	Min	Median	Max	$q_{0.25}$	$q_{0.75}$
GDP per Capita Growth Rate (%)	1.68	2.94	-15.28	1.79	22.73	0.20	3.17
GDP per Capita (in 2010 Value)	11129	14203	224	3810	89772	1248	18415
FDI Flows (% of GDP)	1.88	2.75	-15.04	1.07	27.43	0.37	2.38
Domestic Investment (% of GDP)	22.84	7.71	3.03	22.32	74.79	18.30	26.35
Mean Years of Schooling	6.00	2.93	0.35	5.92	13.05	3.56	8.22
Population Growth Rate (%)	1.84	1.07	-4.44	1.92	6.27	1.01	2.65

We first apply the information criterion 3.12 to determine the number of groups in the data. We take  $c = 0.80$  in the calculation since it performs satisfactorily in the Monte Carlo simulations. However, since the length of panels in our data set is relatively short, which may lead to considerable estimation errors in finite samples, we also calculate the information criterion using  $c = 1.2$  and  $c = 1.6$  and compare the results with the case of  $c = 0.80$ . The values of the information criterion based on the 0.50 quantile are reported in Table 5.

Table 5: Values of the Information Criterion for Different  $K$ 

$\tau$	$c$	Number of Groups				
		$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
0.25	0.8	0.8102	0.7408	<b>0.7148</b>	0.7247	0.7273
	1.2	0.8182	0.7454	<b>0.7391</b>	0.7438	0.7569
	1.6	0.8262	0.7729	<b>0.7690</b>	0.7900	0.7989
0.50	0.8	0.9681	0.9011	<b>0.8781</b>	0.8807	0.8873
	1.2	0.9761	0.9171	<b>0.9076</b>	0.9125	0.9276
	1.6	0.9841	0.9376	<b>0.9326</b>	0.9404	0.9604
0.75	0.8	0.7830	0.7399	<b>0.7147</b>	0.7202	0.7273
	1.2	0.7910	0.7561	<b>0.7487</b>	0.7611	0.7705
	1.6	0.7990	0.7715	<b>0.7686</b>	0.7879	0.8001

Based on the results in Table 5, it is clear that the information criterion achieves the smallest

value when  $K = 3$ , so we set the number of groups to be three in the estimation step. We next implement the estimation method in Section 4.3 to estimate model 3.15. To ease the issue of potential local optimum, we also generate thirteen initial values<sup>9</sup> and estimate the model one by one. The final coefficient estimates are set to be the ones generated by the initial value which gives the smallest quantile loss. The corresponding covariance matrix and the point-wise 95% confidence band are estimated using the bootstrap method. Besides the 0.50 quantile, we also estimate the model at the 0.25 and 0.75 quantiles.

### 3.6.3 Empirical Results

In this subsection, we report the estimation results of model 3.15 using panel data of 90 countries from 1970 – 2010. The estimates of the parametric coefficients in model 3.15 are reported in Table 6. The third column in Table 6 corresponds to the pooling estimates, i.e., the case in which the grouped heterogeneity is ignored. The value and scale of the pooling estimates are comparable to their counterparts in Kottaridi and Stengos (2010). Here we find that the coefficient for domestic investment is significant at the 1% significance level and its value increases with  $\tau$ . The sign of the coefficient for population growth is correct. However, it is only significant at the 10% significance level when  $\tau = 0.25$  and  $\tau = 0.50$ . The estimate of the coefficient for human capital is positive but insignificant for the whole sample. Similar to Kottaridi and Stengos (2010), we also find that the joint effect of human capital and FDI is small and negative, but is insignificant at all three quantiles.

The columns 4-6 in Table 6 correspond to the three sets of group-specific coefficient estimates. Compared with the pooling estimates, there are several interesting findings. First, we find that the coefficient estimate of domestic investment differs significantly across three groups. Under the current labeling, the effect of domestic investment is strongest for countries in Group 3 and weakest for countries in Group 1, and such effect is significant at the 1% significance level for all groups and quantiles except for Group 1 at

---

<sup>9</sup>These eleven initial values are generated following the same rule in the Monte Carlo simulations in Section 4.5: five from Strategy 1, three from Strategy 2 and five from Strategy 3.

the 0.25 quantile. Second, we find that the coefficient estimate of population growth of Group 2 is positive and significant at the 10% significance level when  $\tau = 0.25$  and  $\tau = 0.50$ . Third, though the effect of human capital on economic growth is insignificant in the pooling case, we find it is significant for some groups and quantile levels. For example, the coefficient estimate of human capital is significant at the 5% level for countries in Group 2 at the 0.50 quantile. Similarly, it shows that the joint effect of human capital and FDI is weakly significant for some groups and quantile levels, such as Group 1 and Group 3 at the 0.25 quantile.

The functional coefficient estimates are shown in Figure 2, 3, 4 with the plots of confidence band shown in the subsequent figures. The group labeling here is the same as that for parametric coefficients in Table 6. For the pooling estimate  $\hat{\beta}_\tau(u)$ , its value is relatively small (less than 0.50 at the 0.25 and 0.50 quantile) and positive for most values in support of  $[0, 1]$ , meaning that the FDI only has a small positive effect on economic growth for all countries. The scale of such effect is comparable to the estimate in Kottaridi and Stengos (2010) using only initial values of GDP per capita. However, it is insignificant at the 95% level as the point-wise 95% confidence band of the pool estimates includes the zero function for a large proportion of  $[0, 1]$ .

The pooling estimates  $\hat{\beta}_\tau(u)$  may not provide a concrete picture of the fact, as there are many unobserved factors that can affect the effect of FDI on economic growth, such as culture and level of corruption. Such hypothesis is supported by the group-specific estimates  $\beta_{G_k, \tau}(u)$  for  $k = 1, 2, 3$ . There are two interesting findings. First, we find that the effect of FDI on economic growth can be classified into three groups. For Group 1 and Group 3, such effect is large and positive when GDP per capita is low and the effect for Group 1 is larger than that for Group 3. For Group 2, the effect of FDI on economic growth is negative when GDP per capita is low. To have a better sense of how the effect of FDI varies with the value of GDP per capita, we can conduct some simple calculations based on the summary statistics in Table 4 and the transformation 3.16. From these plots of  $\hat{\beta}_{G_k, \tau}(u)$ , we can see that the effect

is relatively large when  $u_{it} \leq 0.3$ , which maps to GDP per capita of \$3680. Notice that the median value of GDP per capita is \$3810, meaning that the effect of FDI is in fact quite large for nearly half of the countries in our data set. From the point-wise 95% confidence band, we can see that the effect is significant when GDP per capita is small. This finding further substantiates our hypothesis in the beginning: FDI plays a more important role in low-income countries in terms of economic growth than in high-income countries. In addition, these plots also show that the effect of FDI is heterogeneous for low-income countries. It is worth mentioning that the pooling estimates  $\hat{\beta}_\tau(u)$  cannot reveal this important pattern because the grouped heterogeneity is fully ignored and the underlying group-specific effects average each other out. Second, we can see that the scale of the effect of FDI on economic growth decreases with GDP per capita for all three groups, which can be easily learned from the trends of the plots in the second panel.

Finally, we report the group classification and the group-specific summary statistics for these 90 countries based on the 0.50 quantile in Table 7 and Table 8. Turning to the group-specific summary statistics in Table 8, we can see that the average GDP per capita is higher in Group 1 than in Group 2 and 3. On the other hand, Group 3 has the highest GDP per capita growth rate and Group 1 has the lowest GDP per capita growth rate. However, identifying the latent factors that determine the grouped heterogeneity is beyond the scope of this paper and we leave it as a subject for future research.

### 3.7 Conclusion

In this paper, we have studied a new semiparametric quantile panel regression model with grouped heterogeneity. This model can simultaneously handle both the time-variant and nonlinear effects of explanatory variables and the unobserved grouped heterogeneity in coefficients. To estimate the model, we develop a series-based estimation method and a two-step iterative algorithm for computation and establish the asymptotic properties of the

proposed estimators. An information criterion is developed to determine the number of groups. The finite sample performance of the estimation method and the information criterion are investigated through Monte Carlo simulations, which show both perform very well. The model has been applied to study the effect of FDI on economic growth. Some new results have emerged. First, we find that there exist three patterns in the effect of FDI on economic growth. Second, the effect of FDI is in fact large and significant for low-income countries. Third, the scale of the effect diminishes as GDP per capita increases.

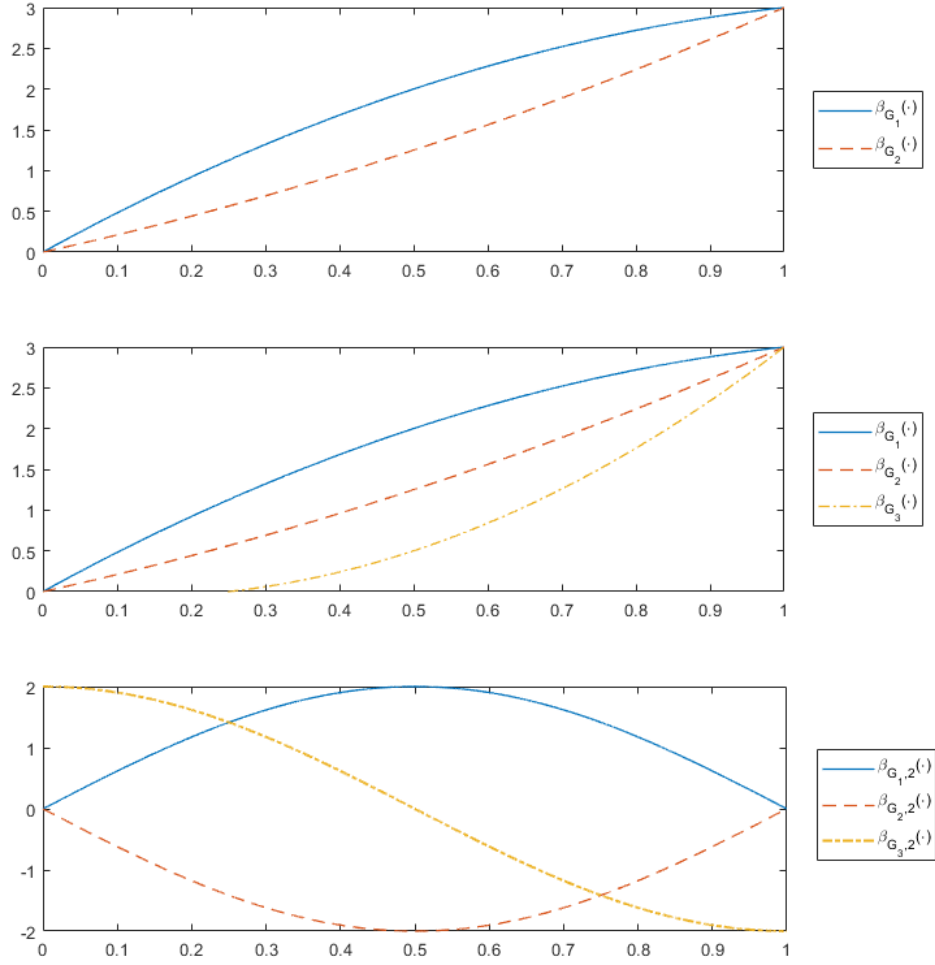


Figure 1: The Plots of Nonparametric Coefficients in Different DGPs (top panel: solid line for  $\beta_{G_1,\tau}(\cdot)$  and dashed line for  $\beta_{G_2,\tau}(\cdot)$  in DGP 1. Middle panel: solid line for  $\beta_{G_1,\tau}(\cdot)$ , dashed line for  $\beta_{G_2,\tau}(\cdot)$  and dash-dotted line for  $\beta_{G_3,\tau}(\cdot)$  in DGP 2, DGP 3, and  $\beta_{G_1,1,\tau}(\cdot)$ ,  $\beta_{G_2,1,\tau}(\cdot)$  ) and  $\beta_{G_3,1,\tau}(\cdot)$  in DGP 4. Bottom panel: solid line for  $\beta_{G_1,2,\tau}(\cdot)$ , dashed line for  $\beta_{G_2,2,\tau}(\cdot)$  and dash-dotted line for  $\beta_{G_3,2,\tau}(\cdot)$  in DGP 4.



Table 1: Finite Sample Performance of the Information Criterion

	$N$	$T$	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
DGP 1	50	40	0.120	0.880	0.000	0.000	0.000
	50	80	0.001	0.996	0.003	0.000	0.000
	50	120	0.000	0.996	0.004	0.000	0.000
	100	40	0.002	0.977	0.021	0.000	0.000
	100	80	0.000	0.989	0.011	0.000	0.000
	100	120	0.000	0.990	0.010	0.000	0.000
DGP 2	50	40	0.000	0.178	0.822	0.000	0.000
	50	80	0.000	0.067	0.933	0.000	0.000
	50	120	0.000	0.025	0.974	0.001	0.000
	100	40	0.000	0.160	0.836	0.004	0.000
	100	80	0.000	0.019	0.971	0.010	0.000
	100	120	0.000	0.000	0.989	0.011	0.000
DGP 3	50	40	0.000	0.099	0.892	0.009	0.000
	50	80	0.000	0.002	0.961	0.037	0.000
	50	120	0.000	0.001	0.984	0.015	0.000
	100	40	0.000	0.002	0.966	0.032	0.000
	100	80	0.000	0.000	0.977	0.023	0.000
	100	120	0.000	0.000	0.989	0.010	0.001
DGP 4	50	40	0.000	0.069	0.875	0.054	0.002
	50	80	0.000	0.000	0.960	0.040	0.000
	50	120	0.000	0.000	0.982	0.028	0.000
	100	40	0.000	0.000	0.871	0.123	0.006
	100	80	0.000	0.000	0.964	0.035	0.001
	100	120	0.000	0.000	0.988	0.012	0.000

<sup>1</sup> We set  $\tau = 0.5$  when calculating the information criterion.

<sup>2</sup> The empirical probability in this table is calculated based on 1000 repetitions.

Table 2: Bias and RMSE of the Coefficient Estimates

		Bias( $\hat{\alpha}_\tau$ )			RMSE( $\hat{\alpha}_\tau$ )			RMSE( $\hat{\beta}_\tau$ )				
$N$	$T$	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$		
DGP 1	50	40	-0.0607	-0.0477	-0.0414	0.0806	0.0681	0.0830	0.1860	0.1603	0.1837	
	50	80	0.0333	-0.0932	-0.0447	0.0512	0.0438	0.0526	0.1195	0.1072	0.1211	
	50	120	-0.0242	0.0161	-0.0151	0.0390	0.0360	0.0407	0.1071	0.0950	0.1083	
	100	40	0.0093	0.0100	-0.0424	0.0587	0.0487	0.0598	0.1301	0.1113	0.1299	
	100	80	-0.0115	0.0219	-0.0192	0.0373	0.0315	0.0353	0.0941	0.0829	0.0937	
	100	120	0.0229	0.0125	0.0204	0.0292	0.0238	0.0281	0.0819	0.0722	0.0816	
	DGP 2	50	40	-0.1064	0.0327	0.2108	0.1246	0.0996	0.1250	0.2424	0.2030	0.2469
		50	80	-0.0366	-0.0027	-0.0464	0.0673	0.0583	0.0678	0.1521	0.1319	0.1501
		50	120	0.0473	-0.0115	0.0596	0.0511	0.0456	0.0508	0.1331	0.1175	0.1320
100		40	-0.1175	0.0167	0.0478	0.0886	0.0687	0.0858	0.1710	0.1418	0.1703	
100		80	-0.0171	-0.0261	0.0226	0.0478	0.0411	0.0476	0.1171	0.1038	0.1171	
100		120	-0.0517	0.0017	0.0178	0.0361	0.0323	0.0365	0.1014	0.0899	0.1008	
DGP 3		50	40	0.0343	0.0391	0.0401	0.0602	0.0511	0.0601	0.0548	0.0468	0.0545
		50	80	-0.0401	-0.0157	0.0127	0.0408	0.0369	0.0407	0.0349	0.0297	0.0345
		50	120	0.0166	0.0371	-0.0041	0.0339	0.0299	0.0336	0.0298	0.0253	0.0298
	100	40	-0.0256	0.0186	-0.0262	0.0407	0.0368	0.0409	0.0346	0.0301	0.0343	
	100	80	0.0157	-0.0056	-0.0165	0.0282	0.0257	0.0287	0.0246	0.0213	0.0251	
	100	120	0.0003	-0.0061	0.0078	0.0242	0.0207	0.0239	0.0215	0.0183	0.0213	
	DGP 4	50	40	0.0831	0.0382	-0.0162	0.0882	0.0783	0.0872	0.2631	0.2358	0.2635
		50	80	-0.0199	0.0082	-0.0018	0.0608	0.0551	0.0618	0.1854	0.1648	0.1861
		50	120	0.0441	-0.0619	0.0091	0.0506	0.0450	0.0491	0.1615	0.1453	0.1622
100		40	0.0271	-0.0208	0.0502	0.0620	0.0549	0.0610	0.1846	0.1655	0.1850	
100		80	0.0065	-0.0343	0.0244	0.0442	0.0394	0.0428	0.1404	0.1256	0.1400	
100		120	0.0154	0.0207	-0.0043	0.0355	0.0311	0.0360	0.1216	0.1095	0.1232	

Table 3: Empirical Rate of Correct Classification

	$N$	$T$	CC Rate		
			$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$
DGP 1	50	40	0.9342 (0.0378)	0.9592 (0.0302)	0.9295 (0.0368)
	50	80	0.9847 (0.0171)	0.9934 (0.0114)	0.9853 (0.0168)
	50	120	0.9960 (0.0086)	0.9987 (0.0050)	0.9953 (0.0094)
	100	40	0.9365 (0.0251)	0.9615 (0.0191)	0.9358 (0.0252)
	100	80	0.9851 (0.0124)	0.9939 (0.0077)	0.9861 (0.0121)
	100	120	0.9962 (0.0063)	0.9990 (0.0030)	0.9963 (0.0060)
DGP 2	50	40	0.8957 (0.0553)	0.9373 (0.0403)	0.8919 (0.0574)
	50	80	0.9784 (0.0214)	0.9914 (0.0137)	0.9782 (0.0213)
	50	120	0.9941 (0.0110)	0.9985 (0.0054)	0.9941 (0.0108)
	100	40	0.9078 (0.0333)	0.9447 (0.0249)	0.9094 (0.0334)
	100	80	0.9805 (0.0151)	0.9919 (0.0091)	0.9801 (0.0144)
	100	120	0.9947 (0.0075)	0.9985 (0.0037)	0.9949 (0.0071)

<sup>1</sup> The values in the parentheses are the corresponding standard errors.

Continued Table 3: Empirical Rate of Correct Classification

	$N$	$T$	CC Rate		
			$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$
DGP 3	50	40	0.9950 (0.0097)	0.9985 (0.0052)	0.9946 (0.0104)
	50	80	0.9998 (0.0016)	0.9999 (0.0006)	0.9999 (0.0012)
	50	120	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
	100	40	0.9952 (0.0067)	0.9988 (0.0034)	0.9955 (0.0067)
	100	80	0.9999 (0.0008)	1.0000 (0.0000)	0.9999 (0.0008)
	100	120	1.0000 (0.0000)	1.0000 (0.0000)	0.9999 (0.0003)
DGP 4	50	40	0.9979 (0.0062)	0.9997 (0.0021)	0.9982 (0.0059)
	50	80	1.0000 (0.0000)	1.0000 (0.0000)	0.9999 (0.0011)
	50	120	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
	100	40	0.9986 (0.0036)	0.9998 (0.0015)	0.9985 (0.0039)
	100	80	0.9999 (0.0003)	1.0000 (0.0000)	0.9999 (0.0003)
	100	120	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)

<sup>1</sup> The values in the parentheses are the corresponding standard errors.

Table 6: Estimates of Parametric Coefficients in Model 3.15

$\tau$	Coefficient	$K = 3$			
		$K = 1$	Group 1	Group 2	Group 3
0.25	$\log((\frac{I^d}{Y})_{it})$	1.869*** (0.651)	0.749** (0.346)	2.952*** (0.935)	3.946*** (0.828)
	$n_{it}$	-0.453* (0.270)	-0.545* (0.315)	0.622** (0.298)	-0.673* (0.369)
	$h_{it}$	0.106 (0.079)	0.299 (0.194)	0.315* (0.183)	-0.137 (0.104)
	$h_{it} \cdot (\frac{I^f}{Y})_{it}$	-0.044 (0.050)	-0.138* (0.081)	0.043 (0.053)	0.189* (0.098)
0.50	$\log((\frac{I^d}{Y})_{it})$	2.281*** (0.626)	0.802*** (0.331)	3.024*** (0.861)	4.196*** (0.898)
	$n_{it}$	-0.304* (0.181)	-0.359* (0.194)	0.583* (0.315)	-0.265 (0.178)
	$h_{it}$	0.027 (0.042)	0.303* (0.176)	0.265** (0.127)	-0.139 (0.142)
	$h_{it} \cdot (\frac{I^f}{Y})_{it}$	-0.028 (0.039)	-0.089 (0.124)	0.124 (0.082)	0.138* (0.079)
0.75	$\log((\frac{I^d}{Y})_{it})$	2.443*** (0.869)	1.466*** (0.519)	3.009*** (0.896)	4.987*** (0.971)
	$n_{it}$	-0.058 (0.037)	-0.161* (0.096)	0.090 (0.158)	-0.189* (0.101)
	$h_{it}$	-0.043 (0.036)	-0.170 (0.106)	0.064 (0.124)	-0.171* (0.098)
	$h_{it} \cdot (\frac{I^f}{Y})_{it}$	-0.006 (0.022)	-0.248* (0.134)	0.056 (0.048)	0.011 (0.037)

<sup>1</sup> \*, \*\*, \*\*\* denote the estimated coefficient is significant at the 10%, 5% and 1% level, respectively. The standard errors are calculated via bootstrapping.

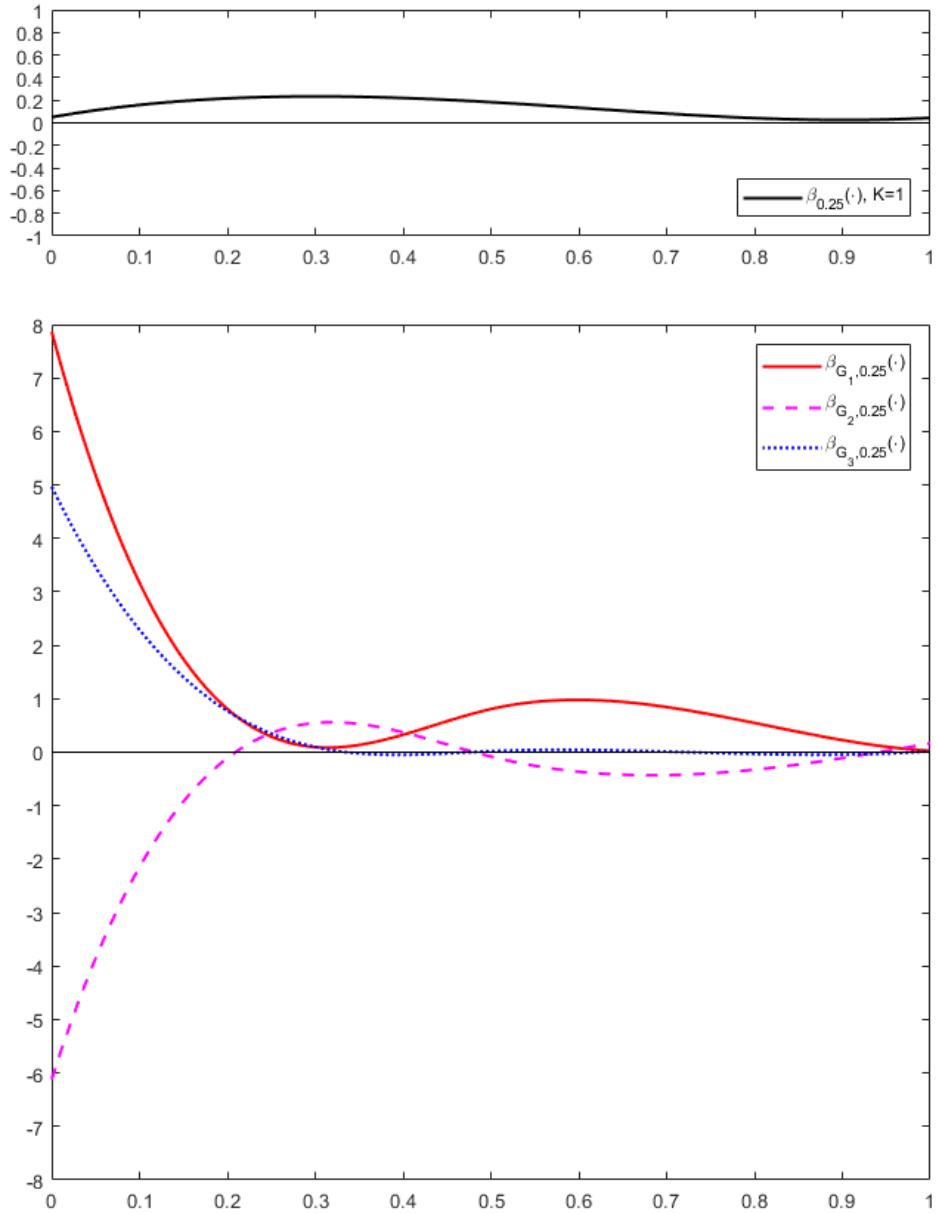


Figure 2: Estimated Functional Coefficients ( $\tau = 0.25$ )

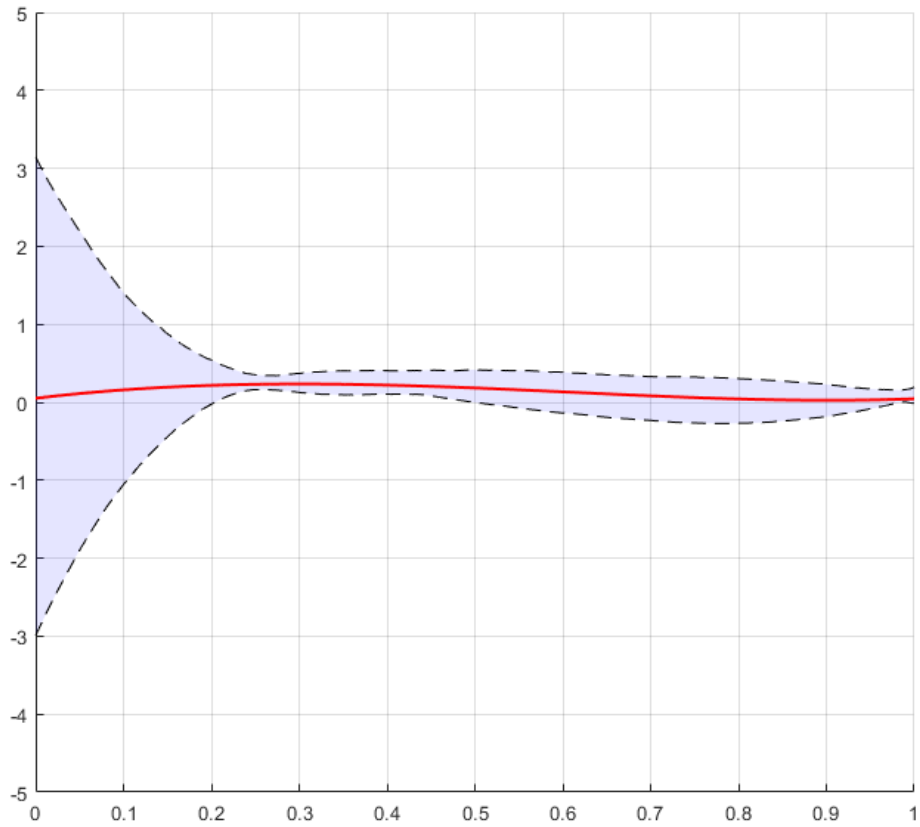


Figure 2.1: Bootstrap 95% Confidence Band of the Functional Coefficient (Pooling Case,  $\tau = 0.25$ )

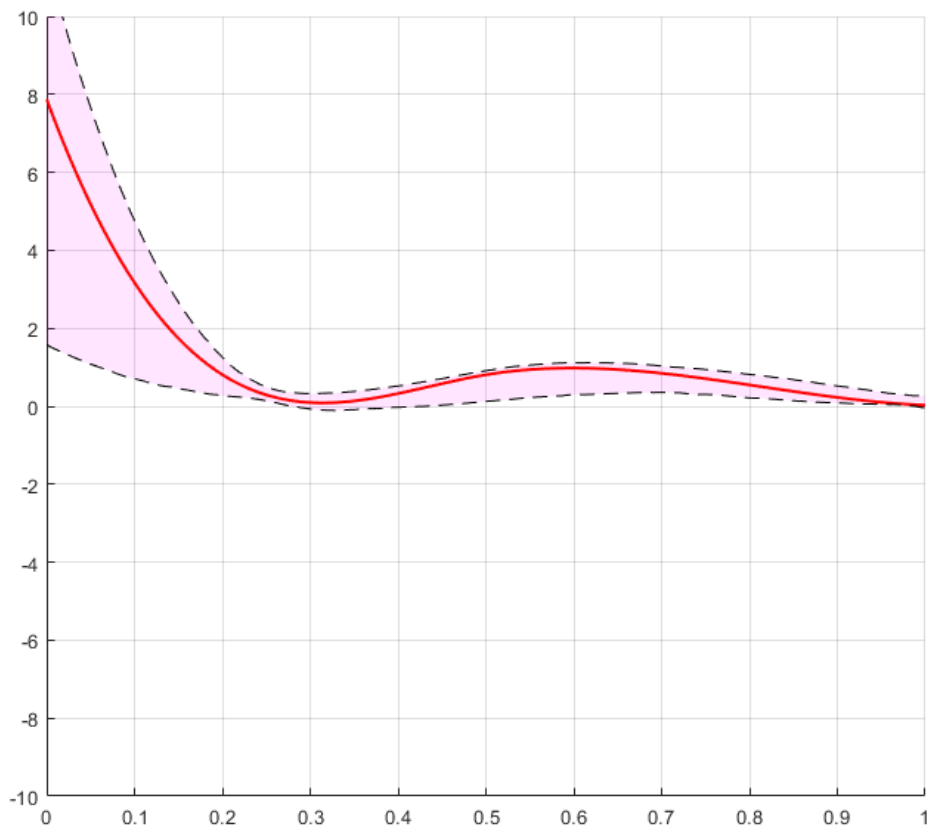


Figure 2.2: Bootstrap 95% Confidence Band of the Functional Coefficient (Group 1,  $\tau = 0.25$ )



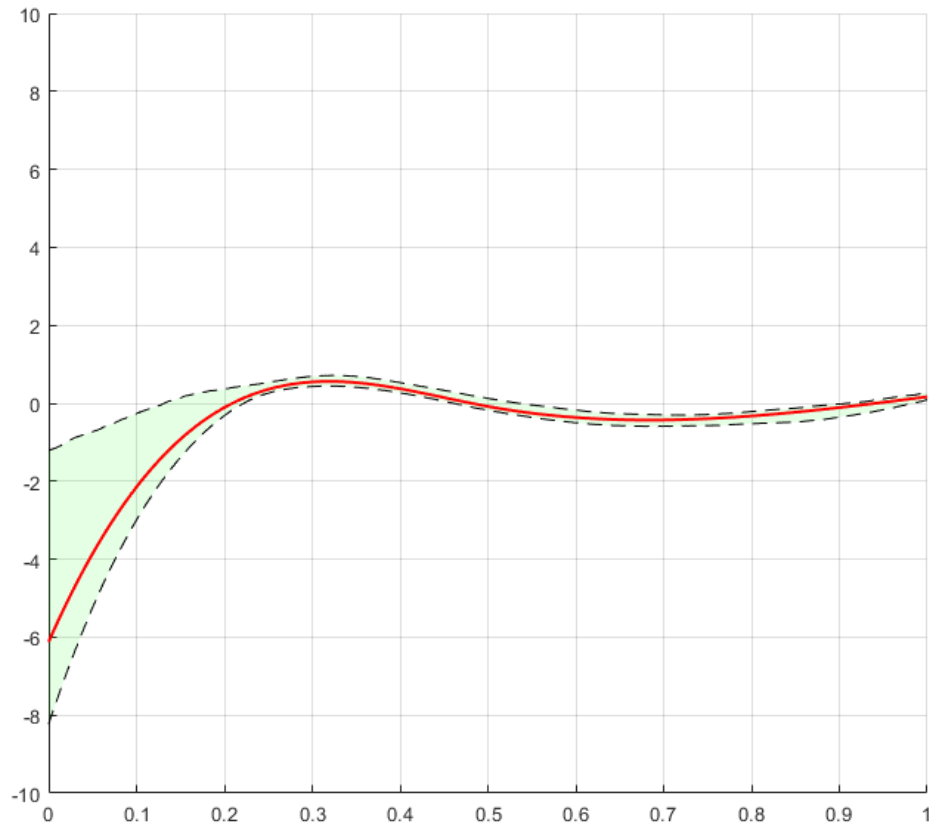


Figure 2.3: Bootstrap 95% Confidence Band of the Functional Coefficient (Group 2,  $\tau = 0.25$ )

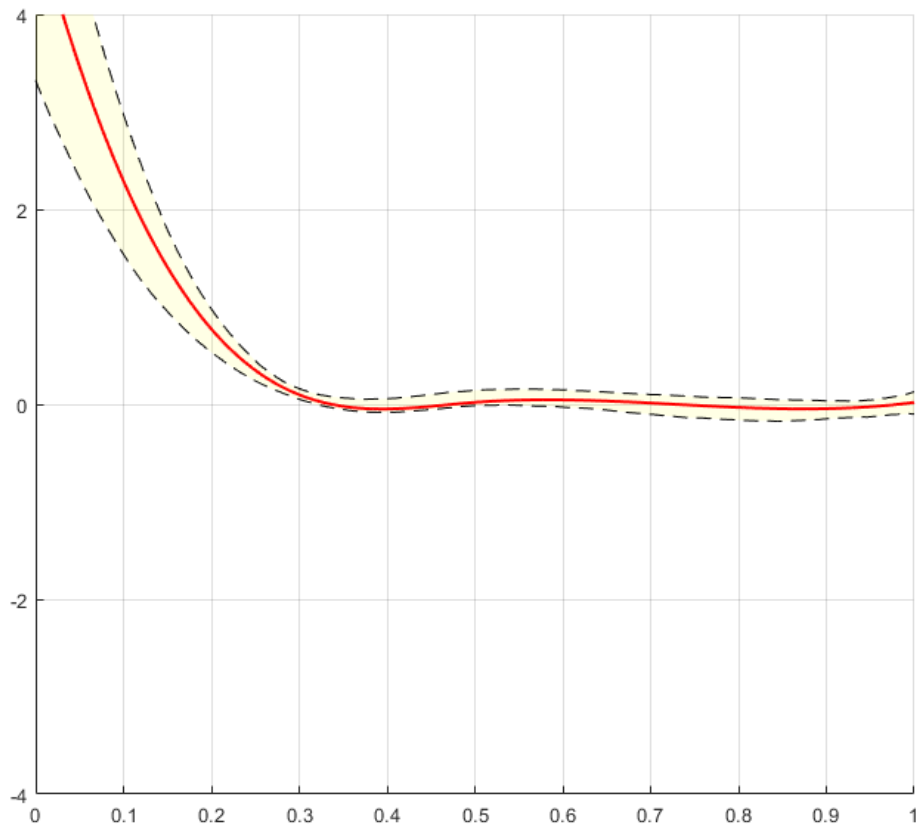


Figure 2.4: Bootstrap 95% Confidence Band of the Functional Coefficient (Group 3,  $\tau = 0.25$ )

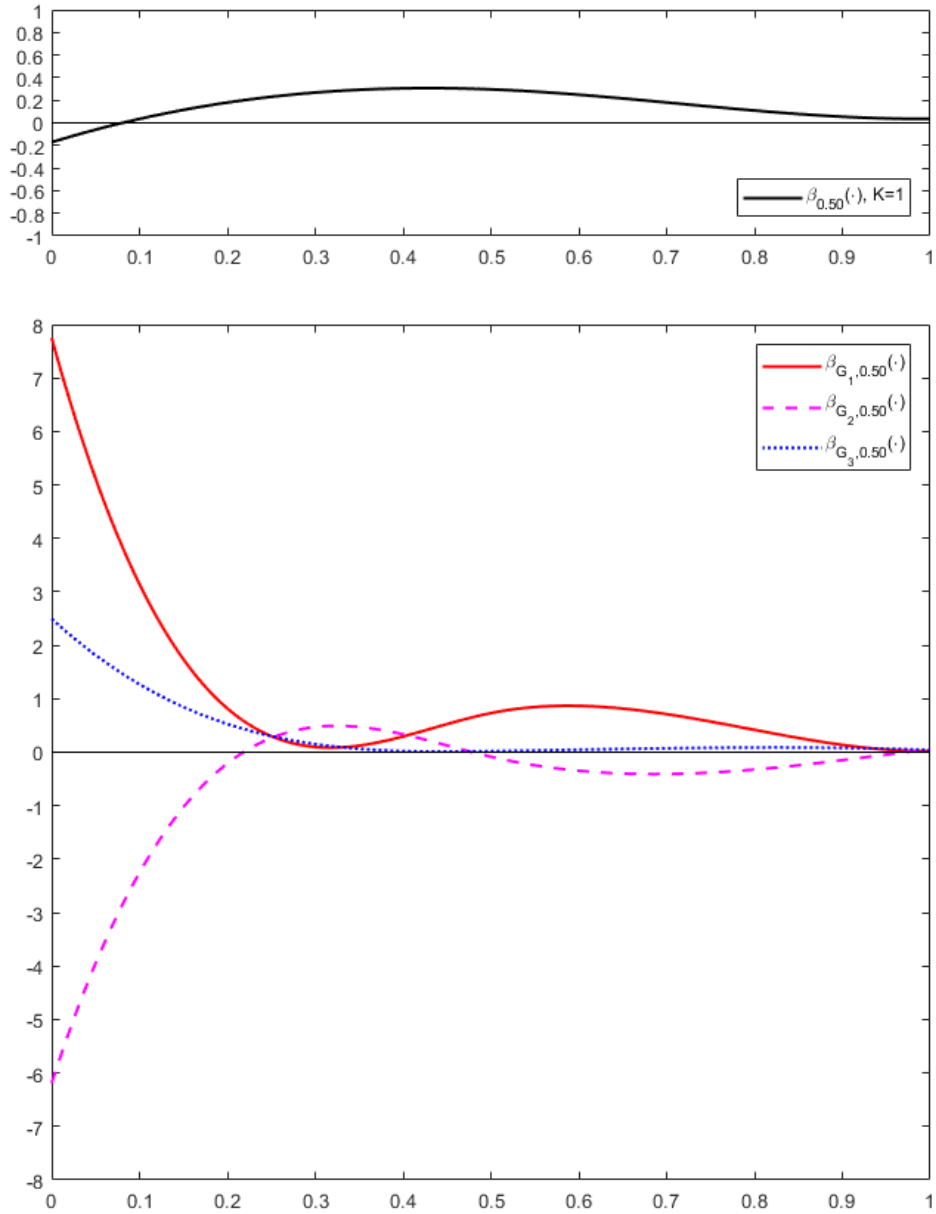


Figure 3: Estimated Functional Coefficients ( $\tau = 0.50$ )

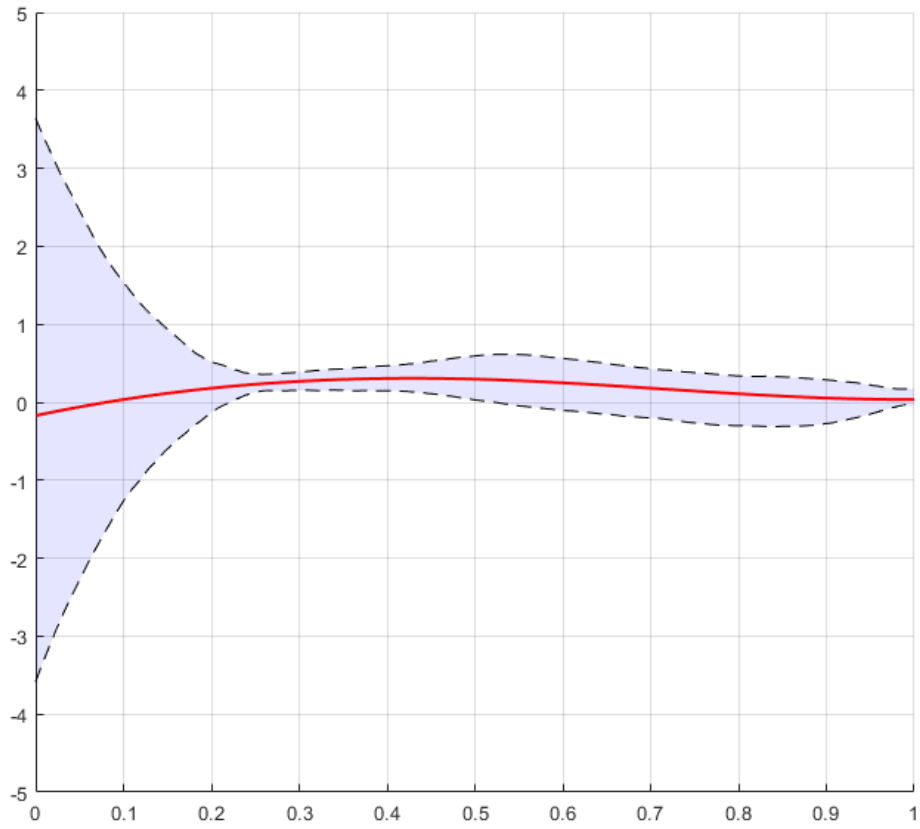


Figure 3.1: Bootstrap 95% Confidence Band of the Functional Coefficient (Pooling Case,  $\tau = 0.50$ )

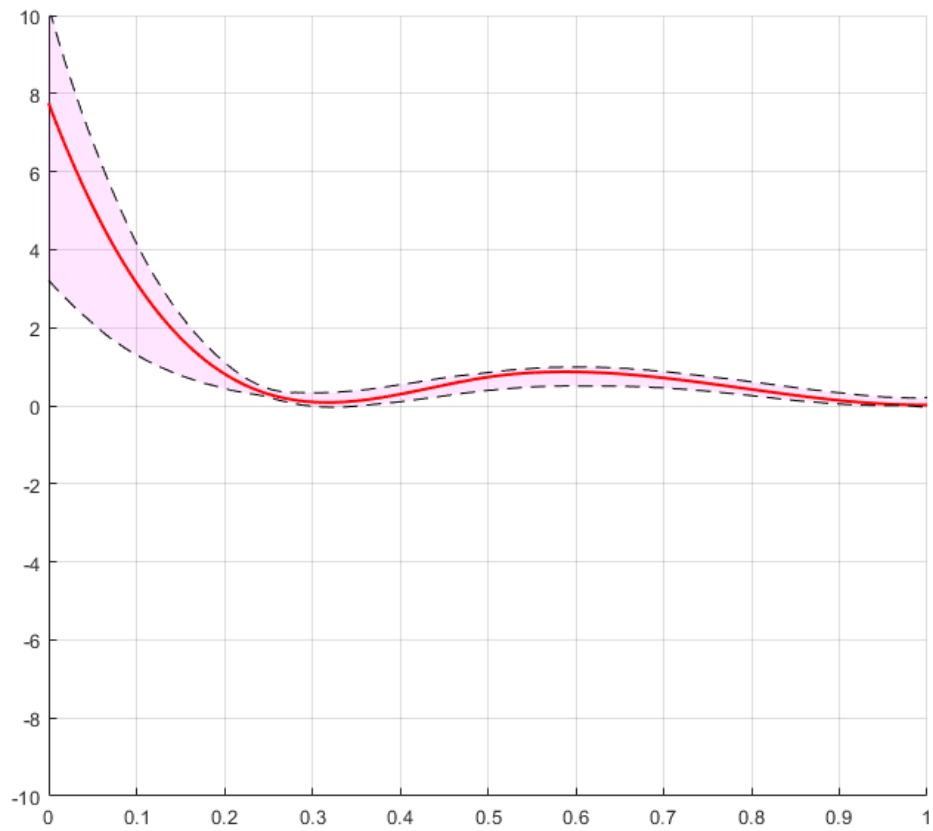


Figure 3.2: Bootstrap 95% Confidence Band for the Functional Coefficient (Group 1,  $\tau = 0.50$ )

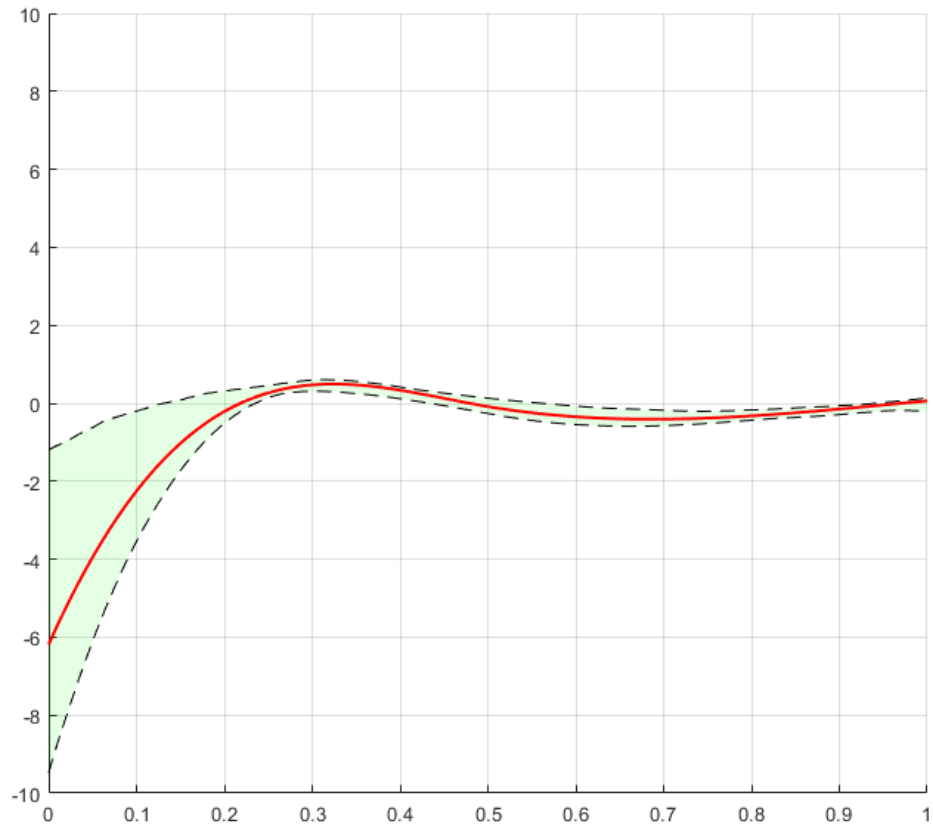


Figure 3.3: Bootstrap 95% Confidence Band of the Functional Coefficient (Group 2,  $\tau = 0.50$ )

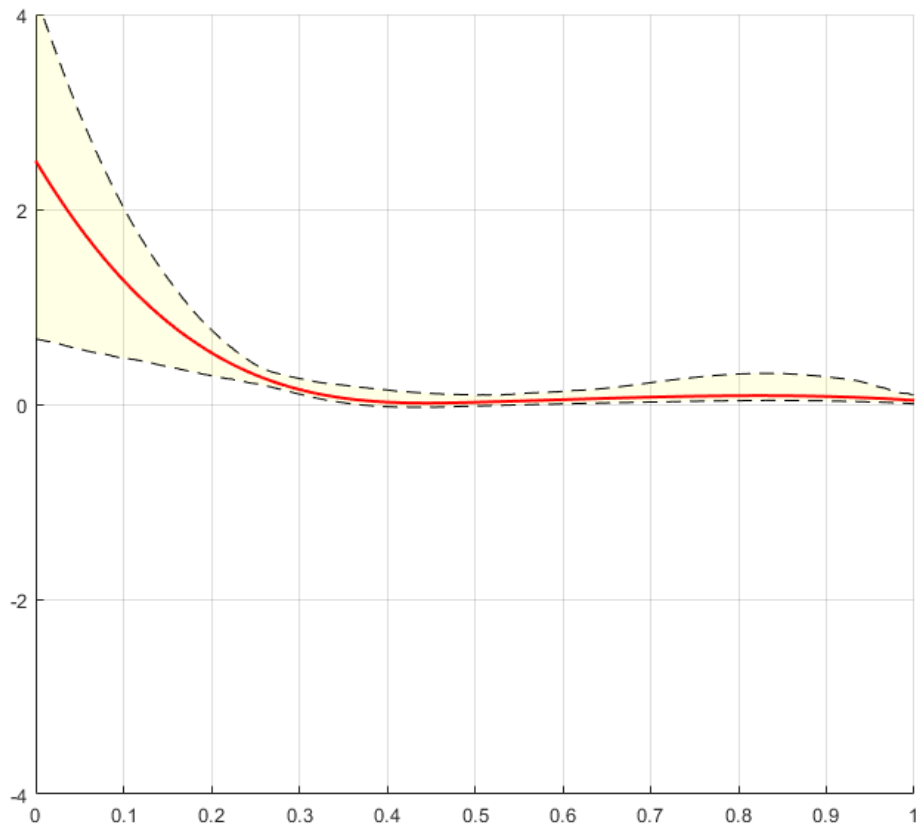


Figure 3.4: Bootstrap 95% Confidence Band of the Functional Coefficient (Group 3,  $\tau = 0.50$ )

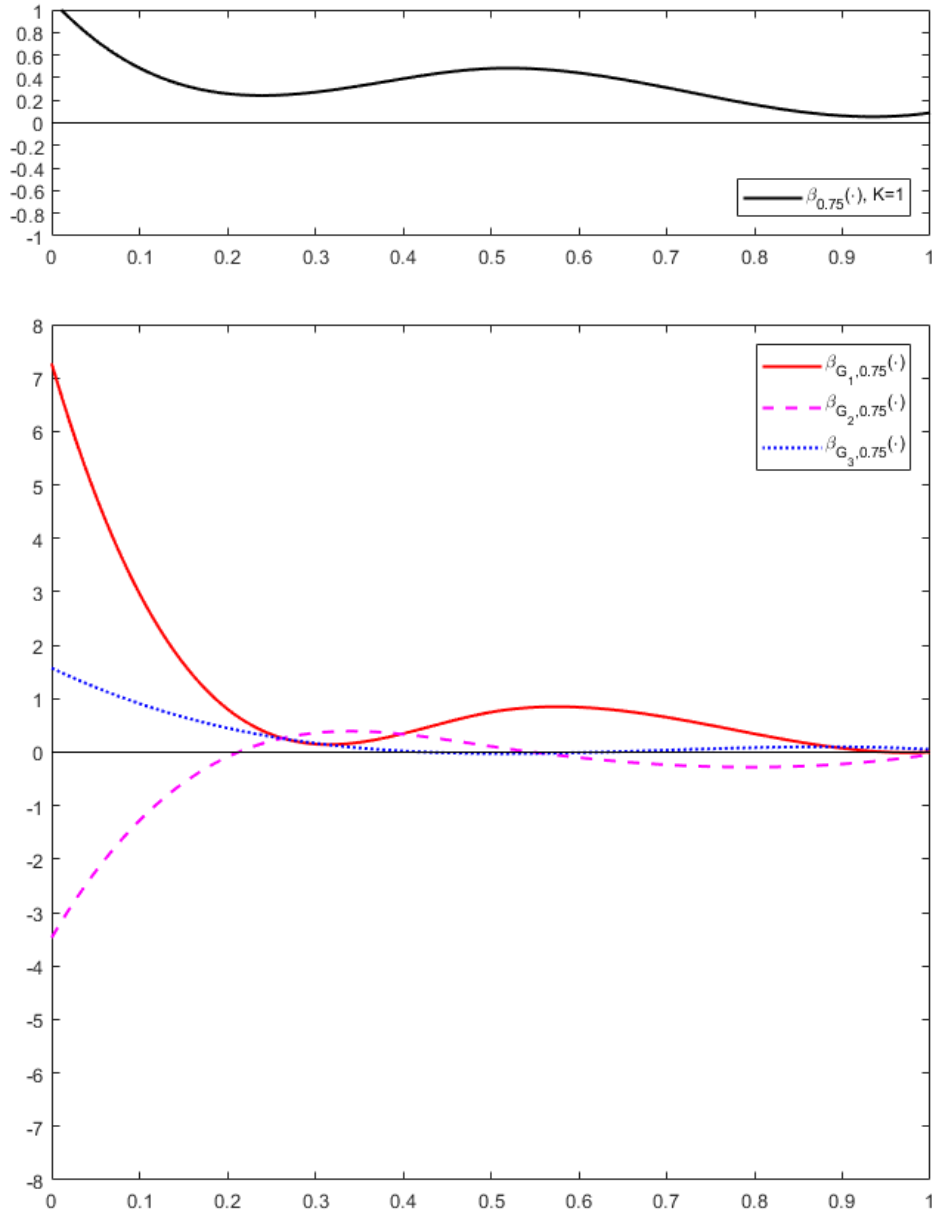


Figure 4: Estimated Functional Coefficients ( $\tau = 0.75$ )



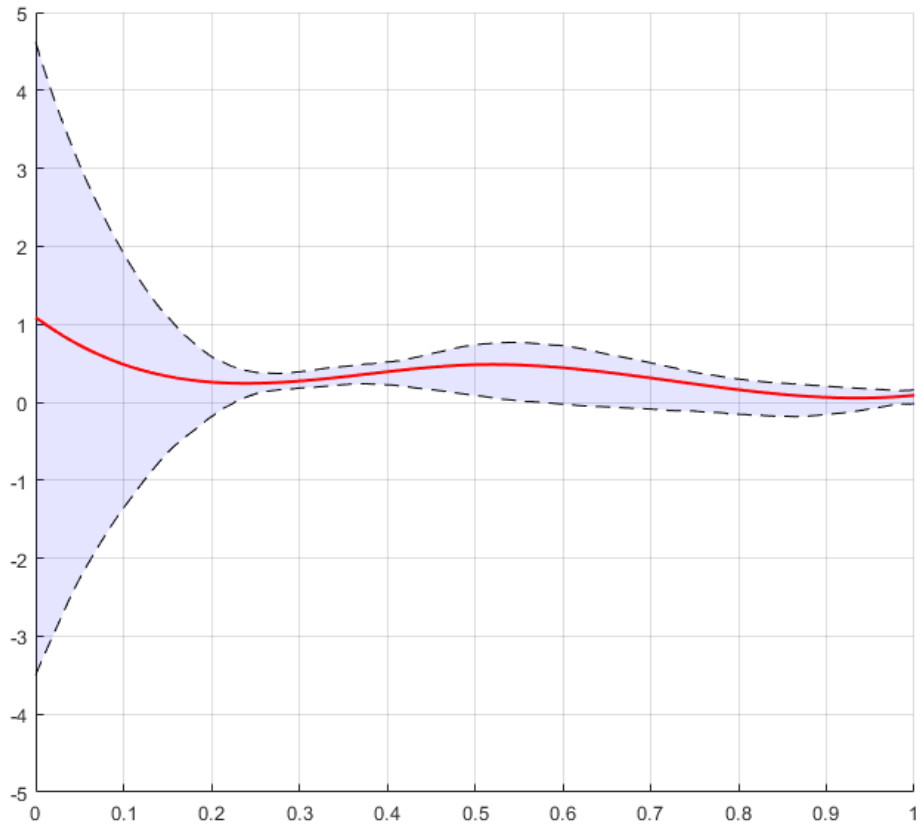


Figure 4.1: Bootstrap 95% Confidence Band of the Functional Coefficient (Pooling Case,  $\tau = 0.75$ )

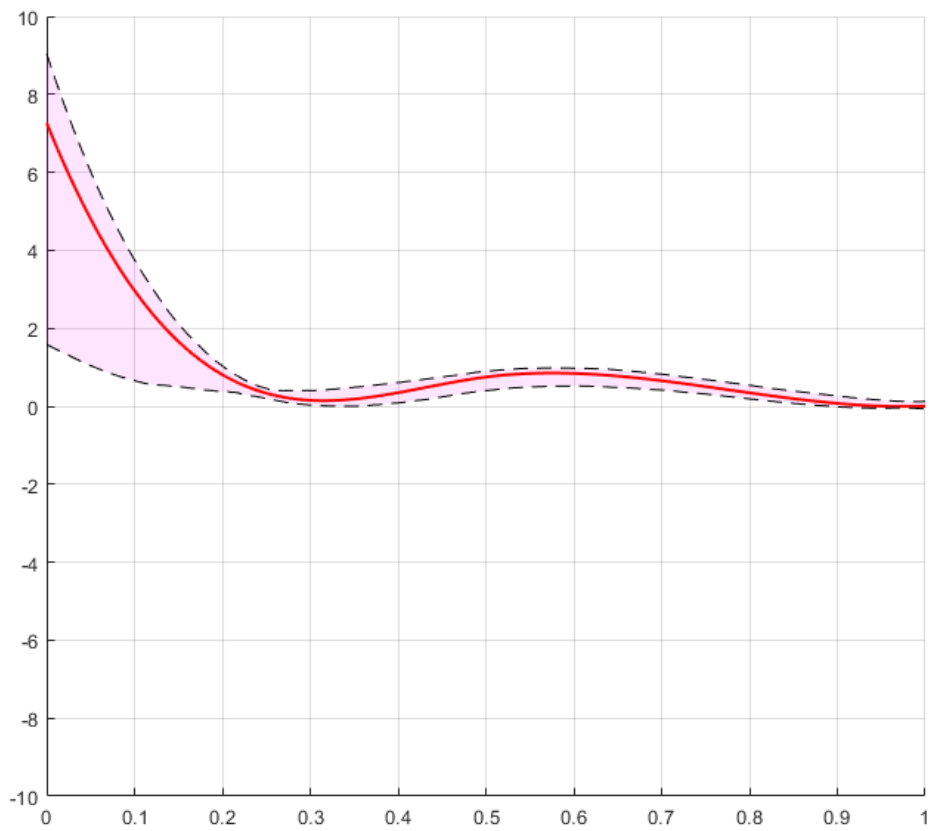


Figure 4.2: Bootstrap 95% Confidence Band of the Functional Coefficient (Group 1,  $\tau = 0.75$ )

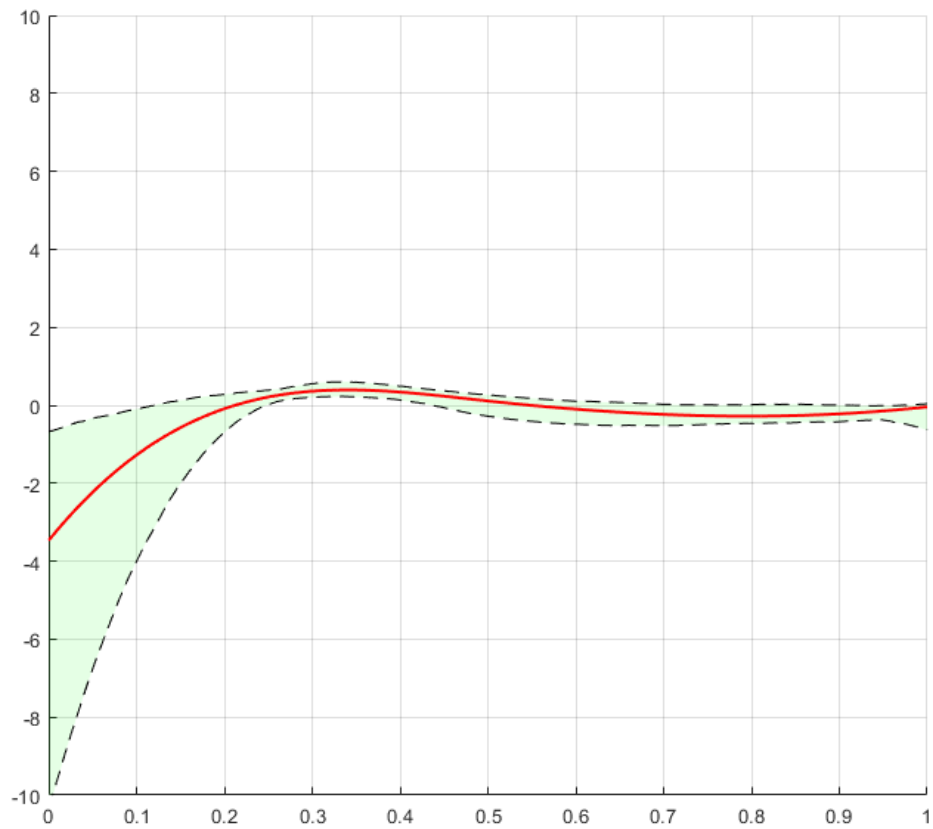


Figure 4.3: Bootstrap 95% Confidence Band of the Functional Coefficient (Group 2,  $\tau = 0.75$ )

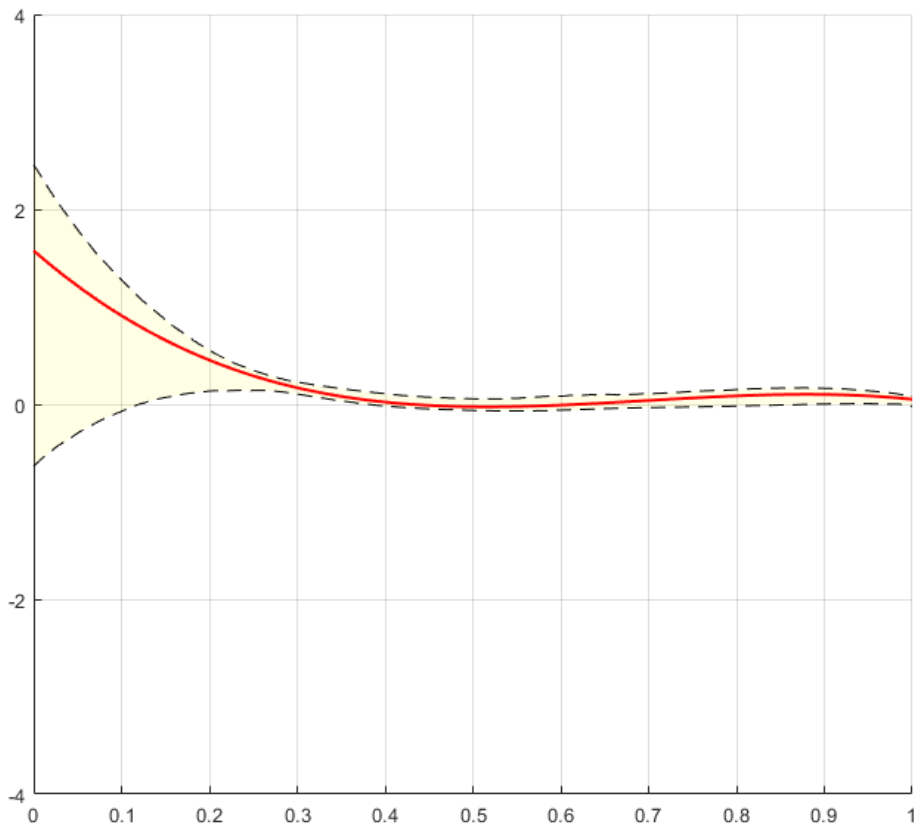


Figure 4.4: Bootstrap 95% Confidence Band of the Functional Coefficient (Group 3,  $\tau = 0.75$ )

Table 7: Group Classification of 90 Countries in the Data Set

Group Membership	Country
Group 1 ( $N = 27$ )	Argentina, Australia, Belgium, Bolivia, Canada, Guyana, Central African Republic, Chad, Dem. Rep. of the Congo, Haiti, Israel, Jamaica, Madagascar, Mauritania, Netherlands, New Zealand, Nicaragua, Niger, Nigeria, Papua New Guinea, Spain, Suriname, Sweden, Trinidad and Tobago, Uruguay, United States of America, United Kingdom
Group 2 ( $N = 33$ )	Algeria, Austria, Bahamas, Burundi, Cameroon, Congo, Denmark, Ecuador, Costa Rica, France, Gabon, Gambia, Germany, Greece, Guatemala, Honduras, Iran, Italy, Japan, Kenya, Malawi, Mali, Mexico, Morocco, Peru, Philippines, Portugal, Saudi Arabia, Senegal, Eswatini, Togo, Tunisia, Venezuela (Bolivarian Rep. of)
Group 3 ( $N = 30$ )	Brazil, Sri Lanka, Chile, Colombia, Benin, Ghana, India, Dominican Republic, El Salvador, Fiji, Finland, Iceland, Iraq, Ireland, Republic of Korea, Malaysia, Oman, Norway, Pakistan, Paraguay, Rwanda, Seychelles, Sierra Leone, Singapore, Zimbabwe, Sudan, Thailand, Turkey, Egypt, Burkina Faso

Table 8: Group-Specific Summary Statistics

<b>Group 1</b>	Mean	S.D.	Min	Median	Max	$q_{0.25}$	$q_{0.75}$
GDP Per Capita Growth Rate (%)	0.94	2.78	-11.88	1.38	12.81	-0.52	2.56
GDP Per Capita (in 2010 Value)	13358	14735	283	6131	52235	1195	26496
FDI Flows (% of DGP)	2.11	3.31	-15.04	1.29	27.43	0.49	2.75
Domestic Investment (% of GDP)	23.34	9.23	3.98	22.40	74.79	18.56	26.36
Mean Years of Schooling	6.78	3.41	0.47	7.05	13.05	3.52	9.84
Population Growth Rate (%)	1.57	1.01	-0.78	1.48	3.79	0.70	2.49
<b>Group 2</b>							
GDP Per Capita Growth Rate (%)	1.45	2.91	-15.28	1.51	22.73	0.07	2.87
GDP Per Capita (in 2010 Value)	11135	13315	224	3783	59794	1484	18904
FDI Flows (% of DGP)	1.49	1.91	-4.59	0.97	12.78	0.32	1.96
Domestic Investment (% of GDP)	23.06	6.93	4.77	22.58	56.71	19.16	26.53
Mean Years of Schooling	5.52	2.53	0.35	5.50	12.08	3.53	7.43
Population Growth Rate (%)	1.98	1.12	-0.21	2.25	6.13	1.12	2.76
<b>Group 3</b>							
GDP Per Capita Growth Rate (%)	2.61	2.87	-8.53	2.57	16.75	1.15	4.13
GDP Per Capita (in 2010 Value)	9118	14374	275	3403	89772	1046	9116
FDI Flows (% of DGP)	2.08	2.94	-3.93	1.05	19.87	0.38	2.60
Domestic Investment (% of GDP)	22.14	6.94	3.02	21.71	45.89	17.57	25.96
Mean Years of Schooling	5.83	2.71	0.82	5.83	11.83	3.62	7.72
Population Growth Rate (%)	1.92	1.02	-4.44	1.94	6.28	1.25	2.54

## Chapter 4

### Nonparametric Additive Panel

### Regression Models with Grouped

### Heterogeneity

## 4.1 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 et al. (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 et al. (2016), Vogt and Linton (2017), Miao et al. (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 chapter, 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 et al. (2002), Profit and Sperlich (2004), Mammen et al. (2009) and Huang et al. (2010), etc. Therefore, this chapter 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 chapters 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 et al. (2016). Su et al. (2019) applied a similar sieve-approximation-based estimation method to estimate time-varying coefficients panel models. However, the time-varying coefficients considered in Su et al. (2019) are nonrandom; thus, the asymptotic properties derived in their chapter 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 that there is grouped heterogeneity in the effect of real per capita disposable income on cigarette sales. The rest of the chapter is organized as follows. We introduce the nonparametric additive panel regression model with grouped heterogeneity in Section 4.2. In Section 4.3, we describe the proposed sieve-approximation-based *Classifier*-Lasso estimation method. Section 4.4 establishes the asymptotic properties of the proposed estimator. Section 4.5 reports the Monte Carlo simulation results. An empirical application is presented in Section 4.6. Finally, Section 4.7 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 chapter, we write  $c$  as positive generic constants that are independent of  $n$  in different places.

## 4.2 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:

$$y_{it} = \mu_i + \sum_{j=1}^p h_{i,j}(x_{it,j}) + u_{it}, \quad u_{it} = \sigma_i(x_{it})\varepsilon_{it}, \quad (4.1)$$

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,  $\mu_i$  denotes the unobserved individual fixed effect which can be correlated with  $x_{it}$ ,  $\varepsilon_{it}$  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 chapter, 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  $\{h_{i,1}^0, \dots, h_{i,p}^0\}_{i=1}^N$ :

$$h_{i,j}^0(x) = \sum_{k=1}^{K_j^0} f_{k,j}^0(x) \mathbf{1}\{i \in G_{k,j}^0\} \quad \text{for any } x \in [0, 1] \text{ and } j = 1, \dots, p, \quad (4.2)$$

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  $f_j = \left( f_{1,j}, \dots, f_{K_j^0,j} \right)'$  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  $\{f_{1,j}(x), \dots, f_{K_j,j}(x)\}$  for  $j = 1, \dots, p$ ; (3) identifying the group memberships  $\{G_{1,j}^0, \dots, G_{K_j,j}^0\}$  for  $j = 1, \dots, p$ . It is worth noting that the nonparametric additive panel regression model given by equations 4.1 and 4.2 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. We can avoid the curse of dimensionality and capture the nonlinearity in the marginal effects of explanatory variables because of the additive structure. Therefore, our model can become an appealing choice for empirical studies in economics, sociology, and many other fields.

### 4.3 Estimation

In this section, we propose the sieve-approximation-based *Classifier*-Lasso estimation method. This section includes two subsections. In Subsection 4.3.1, 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 4.3.2, we introduce the optimization problem and the related estimators.

#### 4.3.1 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 chapter, we use the B-spline polynomials of order  $\kappa$  (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 4.1 could be expressed

as

$$y_{it} = \mu_i + \sum_{j=1}^p z'_{it,j} \theta_{i,j} + e_{it} \quad (4.3)$$

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 4.2 implies

$$\theta_{i,j}^0 = \sum_{k=1}^{K_j^0} \eta_{k,j}^0 \mathbf{1}\{i \in G_{k,j}^0\}. \quad (4.4)$$

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

### 4.3.2 Penalized Estimation of $h(x)$ and $f(x)$

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  $\mu_i$  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

$$y_{it} - \bar{y}_i = \sum_{j=1}^p (z_{it,j} - \bar{z}_{i,j})' \theta_{i,j} + e_{it} - \bar{e}_i, \quad (4.5)$$

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 4.5 could be written as

$$\tilde{y}_{it} = \sum_{j=1}^p \tilde{z}'_{it,j} \theta_{i,j} + \tilde{e}_{it}. \quad (4.6)$$

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:

$$Q_{NT,\lambda}(\theta, \eta) = Q_{NT}(\theta) + \frac{\lambda}{N} \sum_{i=1}^N \sum_{j=1}^p \prod_{k=1}^{K_j^0} \|\theta_{i,j} - \eta_{k,j}\|_F, \quad (4.7)$$

where

$$Q_{NT}(\theta) = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \left( \tilde{y}_{it} - \sum_{j=1}^p \tilde{z}'_{it,j} \theta_{i,j} \right)^2. \quad (4.8)$$

In equations 4.7 and 4.8, 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  $\eta_j = (\eta'_{1,j}, \dots, \eta'_{K_j,j})'$ .  $\lambda$  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 4.7. Then  $\{\hat{h}_{i,1}(x), \dots, \hat{h}_{i,p}(x)\}$  for  $i = 1, \dots, N$ , and  $\{\hat{f}_{1,j}(x), \dots, \hat{f}_{K_j,j}(x)\}$  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 et al. (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  $l = 1, \dots, K_j$ .

## 4.4 Asymptotic Properties

In this section, we establish the asymptotic properties for the estimators proposed in Section 4.3. This section include five subsections. They are organized as follows: in Subsection 4.4.1, 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 4.4.2 presents the results of classification consistency. After that, Subsection 4.4.3 reports the asymptotic distribution of group-specific parameters  $f_{k,j}$ , for  $j = 1, \dots, p$  and  $k = 1, \dots, K_j^0$ . Subsection 4.4.4 discusses how to determine the number of groups.

### 4.4.1 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 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 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 et al. (2016), Su et al. (2019), Vogt and Linton (2017), etc. For more discussions on this, we refer readers to Su et al. (2019).

Assumption 1(ii) imposes moment restrictions for  $u_{it}$ .

Assumption 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 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  $\pi$  and  $\eta$  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

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

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

Similarly, we have

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

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

we can get

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

if  $i \notin G_{k,j}^0$ . This result guarantees that the penalty item in the criterion function 4.7 could shrink the individual coefficients to some group-specific parameters.

Assumption 1(v) is a standard identification condition for sieve estimation. As demonstrated in Section 4.3.2, 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 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.** 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 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 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 1.

**Theorem 1.** Suppose Assumption 1, 2 hold, then

$$(i) \quad \|\hat{\theta}_i - \theta_i^0\|_F = O_p(J^{-r} + J^{\frac{1}{2}}T^{-\frac{1}{2}} + \lambda) \text{ and } \|\hat{\theta}_{i,j} - \theta_{i,j}^0\|_F = O_p(J^{-r} + J^{\frac{1}{2}}T^{-\frac{1}{2}} + \lambda) \text{ for } i = 1, 2, \dots, N, j = 1, \dots, p.$$

$$(ii) \quad \frac{1}{N} \sum_{i=1}^N \|\hat{\theta}_i - \theta_i^0\|_F^2 = O_p(J^{-2r} + JT^{-1}) \text{ and } \frac{1}{N} \sum_{i=1}^N \|\hat{\theta}_{i,j} - \theta_{i,j}^0\|_F^2 = O_p(J^{-2r} + JT^{-1}) \text{ for } j = 1, \dots, p.$$

(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 1(i) and (ii) give the pointwise and mean square convergence rates of  $\hat{\theta}_{i,j}$  for  $j = 1, \dots, p$ . In Theorem 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  $\lambda$ . However, in Theorem 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  $\lambda$ .

By Assumption 2 and Theorem 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

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

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

#### 4.4.2 Classification Consistency

To ensure the group classification's consistency, we need to impose more assumptions, which are given in Assumption 3.

**Assumption 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 3 imposes restrictions on  $\lambda$  and some further ones on  $J$ . Intuitively, we require that  $\lambda$  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 et al. (2016) and Su et al. (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.** *Suppose Assumption 1, 2 and 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 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  $k = 1, \dots, K_j^0$ .

### 4.4.3 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  $\text{Var}(u_i|x_i) = \Sigma_i^{1/2} V_i \Sigma_i^{1/2}$ , where

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

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

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  $\eta$  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  $\eta$  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 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  $\delta_{NT}$  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

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

Assumption 4 is analogous to Assumption A.3 in Su et al. (2019). Assumption 4(i) imposes restrictions on the covariance matrix of  $\varepsilon_i$ . Assumption 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 5.** As  $(N, T) \rightarrow \infty$ ,  $NTJ^{-2r} \rightarrow 0$ .

Assumption 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 3 establishes the asymptotic distribution of the estimated functional of  $f_{k,j}$ .

**Theorem 3.** Suppose Assumption 1, 2, 3, 4 and 5 hold. Then for any  $j \in \{1, \dots, p\}$ ,  $k \in \{1, \dots, K_j^0\}$ ,

(i)

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

(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  $V_{G^0}$ .

Theorem 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.

#### 4.4.4 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  $\lambda$  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  $\rho_{NT}$  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 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 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 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 7. Similarly, Assumption 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 4.** Suppose Assumptions 1, 2, 3, 4, 5, 6 and 7 hold. Then  $P(\hat{K}(\lambda) = K^0) \rightarrow 1$  as  $(N, T) \rightarrow \infty$ .

Theorem 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.

## 4.5 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.

### 4.5.1 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  $s = 1, \dots, p$ ,  $\mu_i$  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 et al. (2019) and simulate with 100 repetitions. We let the tuning parameter  $\lambda = (NT)^{-1/8}$ , which satisfies all the related assumptions on  $\lambda$  given in Section 4.4 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 chapter, 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.

### 4.5.2 Simulation Results

Table 4.1, Table 4.2, and Table 4.3 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 a satisfying performance. Second, the correct classification rate is higher in DGP 1 than in DGP 2 and DGP 3 when  $(N, T)$  is fixed. This 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 4.4 and the simulation findings in Su et al. (2016) and Su et al. (2019). Based on these findings, we recommend using the post-Lasso estimators in empirical studies.

Table 4.1: Simulation Results of the 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 4.2: Simulation Results of the 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 4.3: Simulation Results of the 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

## 4.6 Empirical Illustration

In this section, we apply the model and the estimation method developed in this chapter 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 et al. (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 et al. (2000), they modeled the cigarettes sales using a dynamic linear panel regression model which is specified as

$$\ln y_{it} = \alpha + \beta_1 \ln y_{i,t-1} + \beta_2 \ln x_{it,1} + \beta_2 \ln x_{it,2} + \beta_3 \ln x_{it,3} + u_{it}, \quad (4.9)$$

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 et al. (2000) estimated the model 4.9 using various estimation techniques such as OLS, 2SLS, and Shrinkage OLS. However, the estimation results in 4.9 can give very different policy implications since the signs of  $\beta$ '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. There fore, 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 chapter. 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 et al. (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:

$$\ln y_{it} = \beta_1 \ln y_{i,t-1} + h_{i,1}(x_{it,1}) + h_{i,2}(x_{it,2}) + \alpha_i + u_{it}, \quad (4.10)$$

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 4.10 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.

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 4.1.

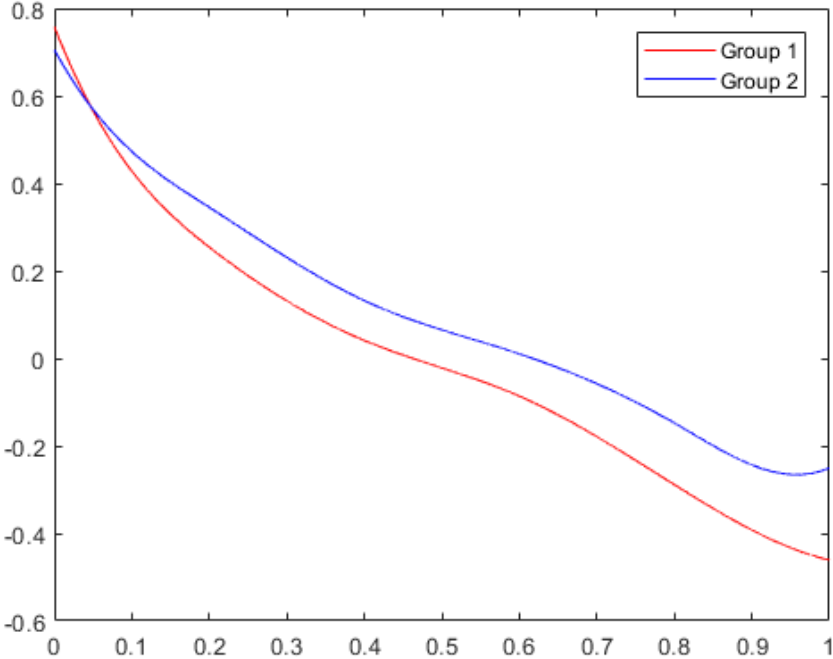


Figure 4.1: Plot of  $\hat{h}_1(x)$

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 4.1, 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 4.2.

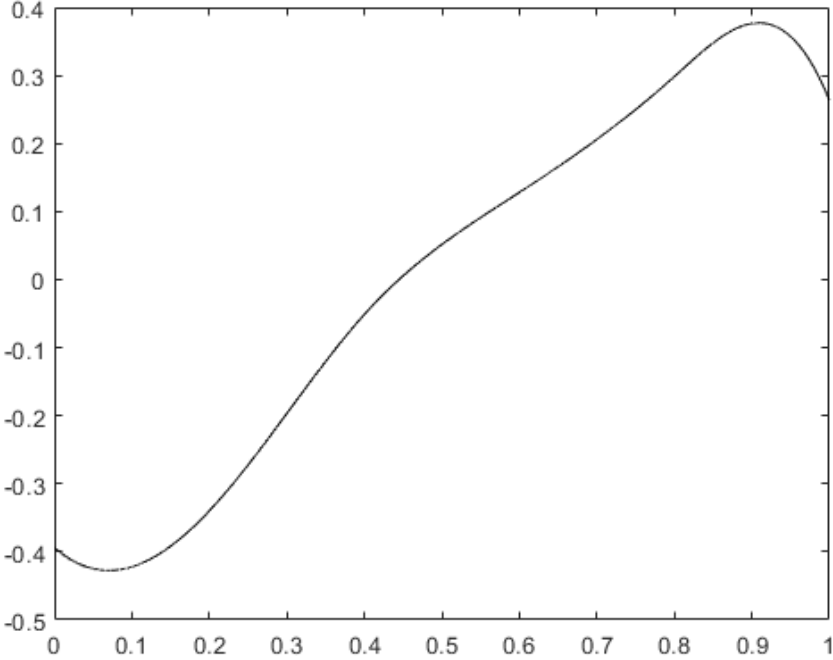


Figure 4.2: Plot of  $\hat{h}_2(x)$

Figure 4.2 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 et al. (2009). It is worth noting that the estimated function of  $h_2(x)$  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.

## 4.7 Conclusion

In this chapter, 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. Besides, we show that the proposed estimation method enjoys the so-call oracle property, which means that parameters are estimated as if the latent group structure is known in advance. Such finding is consistent with Su et al. (2016) and Su et al. (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 chapter to estimate the demand for cigarettes in the United States using panel data of 46 American states from 1963 to 1992.

# Chapter 5

## Appendix

### Proofs of the Main Results in Chapter 1

**Proof of Proposition 1:** We prove the proposition following a similar identification strategy in De Paula et al. (2018). We first show the parameters of interest  $\theta = (\lambda, \beta, \delta, c_2, \dots, c_K)'$  are locally identified by checking conditions of the Theorem 6 in Rothenberg (1971). Without loss of generality, we assume  $c_1$  is normalized to be 1. So,  $\theta \in \mathbb{R}^{K+2}$ . Let

$$H(\theta) = (I - \lambda G(\theta))^{-1}(\beta I + \delta G(\theta)). \quad (5.1)$$

Notice that  $H(\theta)$  is continuously differential in  $\theta$ . Following the definition of identification in Hurwicz (1950) and Bramoullé et al. (2009), the reduced form parameter  $H = (I - \lambda G)^{-1}(\beta I + \delta G)$  is globally identified by Assumption 1. Also notice that the parameters space  $\Theta$  is open by assumption. So, to show the local identification of  $\theta$ , we just need to prove the derivative matrix of  $H(\theta)$  has full rank. We next calculate the derivative matrix of  $H(\theta)$ , denoted as  $\nabla_H$ . First, notice

$$\frac{\partial H}{\partial \lambda} = (I - \lambda G)^{-1} G (I - \lambda G)^{-1} (\beta I + \delta G),$$

$$\frac{\partial H}{\partial \delta} = (I - \lambda G)^{-1} G,$$

$$\frac{\partial H}{\partial \beta} = (I - \lambda G)^{-1},$$

$$\frac{\partial H}{\partial c_k} = \lambda (I - \lambda G)^{-1} \Delta_{c_k} (I - \lambda G)^{-1} (\beta I + \delta G) + \delta (I - \lambda G)^{-1} \Delta_{c_k},$$

where  $\Delta_{c_k}$  is a  $N \times N$  matrix in which the  $(i, j)$ th element equals  $\frac{\partial G_{ij}(\theta)}{\partial c_k}$ . We then can write out the derivative matrix  $\nabla_H$  by vectorizing those derivatives and combining them into a  $N^2 \times (K + 2)$  matrix.

$$\nabla_H = [\text{vec}(\frac{\partial H}{\partial \lambda}), \text{vec}(\frac{\partial H}{\partial \delta}), \text{vec}(\frac{\partial H}{\partial \beta}), \text{vec}(\frac{\partial H}{\partial c_2}), \dots, \text{vec}(\frac{\partial H}{\partial c_K})]$$

By the Theorem 6 in Rothenberg (1971),  $\theta$  is locally identified if  $\text{rank}(\nabla_H) = (K + 2)$ . Suppose not, then there exist some  $(K + 2)$ -dimensional vector  $a = (a_\lambda, a_\delta, a_\beta, a_2, \dots, a_K)$  such that  $a \neq 0$  and

$$a_\lambda \cdot \text{vec}(\frac{\partial H}{\partial \lambda}) + a_\delta \cdot \text{vec}(\frac{\partial H}{\partial \delta}) + a_\beta \cdot \text{vec}(\frac{\partial H}{\partial \beta}) + a_2 \cdot \text{vec}(\frac{\partial H}{\partial a_2}) + \dots + a_K \cdot \text{vec}(\frac{\partial H}{\partial c_K}) = 0,$$

which is equivalent to

$$a_\lambda \cdot \frac{\partial H}{\partial \lambda} + a_\delta \cdot \frac{\partial H}{\partial \delta} + a_\beta \cdot \frac{\partial H}{\partial \beta} + a_2 \cdot \frac{\partial H}{\partial c_2} + \dots + a_K \cdot \frac{\partial H}{\partial c_K} = 0. \quad (5.2)$$

Plugging in the expressions of derivatives, we have

$$\left\{ \sum_{i=2}^K a_i [\lambda(I-\lambda G)^{-1} \Delta_{c_i} (I-\lambda G)^{-1} (\beta I + \delta G) + \delta (I-\lambda G)^{-1} \Delta_{c_i}] \right\} + a_\lambda (I-\lambda G)^{-1} G (I-\lambda G)^{-1} (\beta I + \delta G) + a_\delta (I-\lambda G)^{-1} G + a_\beta (I-\lambda G)^{-1} = 0. \quad (5.3)$$

Multiplying both sides of equation (5.3) by  $(I - \lambda G)$  on the left, we get

$$\left\{ \sum_{i=2}^K a_i [\lambda \Delta_{c_i} (I - \lambda G)^{-1} (\beta I + \delta G) + \delta \Delta_{c_i}] \right\} + a_\lambda G (I - \lambda G)^{-1} (\beta I + \delta G) + a_\delta G + a_\beta I = 0.$$

Let  $C = \sum_{i=2}^K a_i \Delta_{c_i}$ . Because  $|\lambda| < 1$  by Assumption 2 and  $G$  is row-normalized, we can commute  $(\beta I + \delta G)$  and  $(I - \lambda G)^{-1}$ . Thus we have

$$\lambda C (\beta I + \delta G) (I - \lambda G)^{-1} + \delta C + a_\lambda G (\beta I + \delta G) (I - \lambda G)^{-1} + a_\delta G + a_\beta I = 0. \quad (5.4)$$

Multiplying both sides of equation (5.4) by  $(I - \lambda G)$  on the right, we get

$$\lambda C (\beta I + \delta G) + \delta C (I - \lambda G) + a_\lambda G (\beta I + \delta G) + a_\delta G (I - \lambda G) + a_\beta (I - \lambda G) = 0.$$

After some algebra, we have

$$(\delta + \lambda \beta) C + a_\beta I + (\beta a_\lambda - \lambda a_\beta + a_\delta) G + (a_\lambda \delta - \lambda a_\delta) G^2 = 0. \quad (5.5)$$

Since the model assumes  $W_{ii} = 0$ , we have  $G_{ii} = 0$  and thus  $C_{ii} = 0$  for  $i = 1, \dots, N$ . So, equation (5.5) implies

$$a_\beta + (a_\lambda \delta - \lambda a_\delta) (G_{ii}^2) = 0,$$

for all  $i = 1, \dots, N$ . By Assumption 5, there exist  $i, j \in V$  such that  $G_{ii}^2 \neq G_{jj}^2$  and notice  $G_{ii}^2 > 0$  since there is no isolated agents, we have  $a_\beta = 0$  and  $a_\lambda \delta - \lambda a_\delta = 0$ . So, we can

further simplify equation (5.5) and get

$$(\delta + \lambda\beta)C + (\beta a_\lambda + a_\delta)G = \left( \sum_{i=2}^K a_i (\delta + \lambda\beta) \Delta_{c_i} \right) + (\beta a_\lambda + a_\delta)G = 0. \quad (5.6)$$

We only need to show the matrices  $\{\Delta_{c_2}, \dots, \Delta_{c_K}, G\}$  are linear independent. We abuse notations a little bit here by using  $c_k$  to represent both the group identity  $k$  and the strength of links in group  $k$ . We call a link group  $k$  is directly comparable to the normalized group  $c_1$  if there exists some row  $i$  in  $G$  such that the denominator of  $G_{ij} \neq 0$  ( $j \in V$ ) takes the form  $(b \cdot c_1 + \sum_{j \in P(i)} e_j c_j)$  and  $k \in P(i)$ , where  $P(i)$  is set of identities of all links connect to  $i$  except for the normalized group,  $b$  equals the number of links connect to  $i$  that are in the normalized group  $c_1$  and  $e_j$  equals the number of links connect to  $i$  that are in the group  $j$  except for the normalized group. Suppose  $\{\Delta_{c_2}, \dots, \Delta_{c_K}, G\}$  are linearly dependent, then there exist  $d = (d_2, \dots, d_K, d_G) \neq 0$  such that

$$d_2 \Delta_{c_2} + \dots + d_K \Delta_{c_K} + d_G G = 0. \quad (5.7)$$

By the definition of the strength-adjusted adjacency matrix  $G$ , there must exist some agent  $i$  such that  $i$ 's links can be classified into the normalized group and other groups (otherwise  $G$  will be a matrix which only contains constants). Without loss of generality, we assume that  $i$ 's links can be classified into the first  $p$  groups, where  $p \leq K$ . Then there exists a finite

sequence  $\{j_1, \dots, j_p\}$ ,  $j_l \in \{1, \dots, N\}$ ,  $l = 1, \dots, p$  such that

$$\begin{aligned} G_{ij_1} &= \frac{1}{b + \sum_{j=1}^p g_j c_j}, \\ G_{ij_2} &= \frac{c_2}{b + \sum_{j=1}^p g_j c_j}, \\ &\vdots \\ G_{ij_p} &= \frac{c_p}{b + \sum_{j=1}^p g_j c_j}, \end{aligned}$$

where  $b$  is the number of  $i$ 's links that are classified into the normalized group,  $g_j$  is the number of  $i$ 's links that are classified into the group  $j$ . Notice that

$$\begin{aligned} \frac{\partial G_{ij_1}}{\partial c_j} &= \frac{-g_j}{(b + \sum_{j=1}^p g_j c_j)^2}, \\ \frac{\partial G_{ij_l}}{\partial c_l} &= \frac{b + \sum_{j=1, j \neq l}^p g_j c_j}{(b + \sum_{j=1}^p g_j c_j)^2}, \\ \frac{\partial G_{ij_l}}{\partial c_k} &= \frac{-g_k c_l}{(b + \sum_{j=1}^p g_j c_j)^2}. \end{aligned}$$

Then equation (5.7) implies the following system of linear equations:

$$\left\{ \begin{array}{l} d_1(-g_1) + d_2(-g_2) + \dots + d_p(-g_p) + d_G(b + \sum_{j=1}^p g_j c_j) = 0, \end{array} \right. \quad (5.8)$$

$$\left\{ \begin{array}{l} d_1(b + \sum_{j=1, j \neq 1}^p g_j c_j) + d_2(-g_2 c_1) + \dots + d_p(-g_p c_1) + d_G(b + \sum_{j=1}^p g_j c_j) c_1 = 0, \end{array} \right. \quad (5.9)$$

$$\left\{ \begin{array}{l} d_1(-g_1 c_2) + d_2(b + \sum_{j=1, j \neq 2}^p g_j c_j) + \dots + d_p(-g_p c_2) + d_G(b + \sum_{j=1}^p g_j c_j) c_2 = 0, \end{array} \right. \quad (5.10)$$

$$\left\{ \begin{array}{l} \dots \\ d_1(-g_1 c_p) + d_2(-g_2 c_p) + \dots + d_p(b + \sum_{j=1, j \neq p}^p g_j c_j) + d_G(b + \sum_{j=1}^p g_j c_j) c_p = 0. \end{array} \right. \quad (5.11)$$



Calculate  $(5.14) - (5.8) \times c_1$ , we get

$$d_1(b + \sum_{j=1}^p g_j c_j) = 0.$$

We have  $g_j > 0$  and  $b > 0$  by construction and  $c_j > 0$  by Assumption 4, so  $d_1 = 0$ . Similarly, we have  $d_2 = d_3 = \dots = d_p = 0$  and  $d_G = 0$ . Therefore, we have shown that if group  $c_j$  is directly comparable to the normalized group  $c_1$ , then  $d_j = 0$ . Let  $D_1$  denote the set of link groups that are directly comparable to the normalized group  $c_1$ . The next step is to consider the link groups that are directly comparable to the link groups in the set  $D_1$  but not directly comparable to the normalized link. Let  $D_2$  denote the set of such link groups. By definition, there exists an agent  $i'$  and his links can be classified into the link groups in  $D_1$  and link groups in  $D_2$ . Without loss generality, we assume all links of the agents  $i'$  can be classified into  $p_1 + p_2$  groups, i.e.,  $P(i') = \{1, \dots, p_1, p + 1, p + 2, \dots, p + p_2\}$ , where  $p$  is defined above and  $p_1 \neq p$ . So, link groups  $p + 1, \dots, p + p_2$  are not directly comparable to the normalized group  $c_1$ , while link groups  $1, \dots, p_1$  are directly comparable to the normalized group  $c_1$ . Using similar idea, there exist elements in  $G$  such that

$$G_{i'm_k} = \frac{c_k}{\sum_{j=1}^{p_1} g_j c_j + \sum_{j=p+1}^{p+p_2} g_j c_j}$$

where  $m_k \in V$ ,  $k = 1, \dots, p_1, p + 1, p + 2, \dots, p + p_2$ . Then equation (5.7) implies the following system of linear equations:

$$\left\{ \begin{array}{l} d_1 \left( \sum_{j=2}^{p_1} g_j c_j + \sum_{j=p+1}^{p+p_2} g_j c_j \right) + d_2 (-g_2 c_1) + \dots + d_{p_2} (-g_{p_2} c_1) + d_G \left( \sum_{j=1}^{p_1} g_j c_j + \sum_{j=p+1}^{p+p_2} g_j c_j \right) c_1 = 0 \\ d_1 (-g_1 c_2) + d_2 \left( \sum_{j=1, j \neq 2}^{p_1} g_j c_j + \sum_{j=p+1}^{p+p_2} g_j c_j \right) + \dots + d_{p_2} (-g_{p_2} c_1) + d_G \left( \sum_{j=1}^{p_1} g_j c_j + \sum_{j=p+1}^{p+p_2} g_j c_j \right) c_2 = 0 \\ \vdots \\ d_1 (-g_1 c_{p_2}) + d_2 (-g_2 c_{p_2}) + \dots + d_{p_2} \left( \sum_{j=1}^{p_1} g_j c_j + \sum_{j=p+1}^{p+p_2-1} g_j c_j \right) + d_G \left( \sum_{j=1}^{p_1} g_j c_j + \sum_{j=p+1}^{p+p_2} g_j c_j \right) c_{p_2} = 0 \end{array} \right.$$

Notice that we have already shown  $d_1 = d_2 = \dots = d_{p_1} = d_G = 0$ . Using a similar strategy as above, it can be shown that  $d_{p+1} = d_{p+2} = \dots = d_{p+p_2} = 0$ . Therefore, we have shown that if any link group  $k$  is directly comparable to some link group which is either directly or indirectly comparable to the normalized group  $c_1$ , then  $d_k = 0$ . Under Assumption 6, if there is only one component in the network and there is no isolated agent (the network is connected), then every link group is either directly comparable or indirectly comparable to  $c_1$ , so  $d_1 = \dots = d_K = d_G = 0$ , which implies  $\{\Delta_{c_2}, \dots, \Delta_{c_K}, G\}$  is linearly independent. If there are multiple components, by Assumption 6, all links are also either directly or indirectly comparable to the normalized group. If this is not true, we can always find a non-empty proper subset  $P$  of  $\{1, \dots, L\}$  such that  $c(G_p) \cap c(G_{p^c}) = \emptyset$ , which contradicts the Assumption 6. In summary, under Assumptions 1-6,  $\{\Delta_{c_2}, \dots, \Delta_{c_K}, G\}$  are linearly independent.

Therefore, combining the above results with equation (5.6), we have

$$a_i(\delta + \lambda\beta) = 0, \quad \text{for } i = 2, \dots, K,$$

and

$$\beta a_\lambda + a_\delta = 0 \tag{5.12}$$

By Assumption 3,  $\lambda\beta + \delta \neq 0$ , so  $a_2 = \dots = a_K = 0$ . Also notice that we have shown

$$a_\lambda\delta - \lambda a_\delta = 0 \tag{5.13}$$

Combining equation (5.12) and equation (5.13) together, we have  $(\lambda\beta + \delta)a_\lambda = 0$ . By Assumption 3, which states  $\lambda\beta + \delta \neq 0$ , we have  $a_\lambda = 0$ , which implies  $a_\delta = 0$ . Finally, we have shown

$$a_\lambda = a_\delta = a_\beta = a_2 = \dots = a_K = 0,$$

which implies the derivative matrix  $\nabla_H$  has full rank. By the Theorem 6 in Rothenberg (1971),  $\theta$  is locally identified. The global identification result then follows from the Corollary 2 and Corollary 3 in De Paula et al. (2018). Therefore, under Assumptions 1-6,  $\theta$  is identified if (1) the sign of  $(\lambda\beta + \delta)$  is known, or (2)  $\lambda > 0$ .

**Asymptotic Analysis of the NLS estimator:** Here we provide a brief asymptotic analysis of the NLS estimator following a similar strategy in Wang and Lee (2013). To simplify the analysis, in this part, we also assume that  $\mathbf{x}$  is non-stochastic and elements of  $\epsilon$  are i.i.d. with zero mean and variance  $\sigma_0$  following Wang and Lee (2013). Let  $h(\mathbf{x}_n, \theta) = (I - \lambda\mathbf{G}_n(\theta))^{-1}(\beta\mathbf{I} + \delta\mathbf{G}_n(\theta))\mathbf{x}_n$ ,  $Q_n(\theta) = I - \lambda\mathbf{G}_n(\theta)$ ,  $\mathbf{G}_{n0} = \mathbf{G}_n(\theta_0)$ ,  $Q_{n0} = Q_n(\theta_0)$  and  $L_n(\theta) = [\mathbf{y} - \mathbf{h}(\mathbf{x}_n, \theta)]'[\mathbf{y} - \mathbf{h}(\mathbf{x}_n, \theta)]$ , where  $\theta_0$  is the true value of  $\theta$ .

Notice that

$$\begin{aligned} \frac{1}{n}\mathbb{E}[L_n(\theta)] &= \frac{1}{n}[(\beta_0 - \beta)I + (\mathbf{G}_{n0}\delta_0 - \mathbf{G}\delta)\mathbf{x} + (\lambda_0\mathbf{G}_{n0} - \lambda\mathbf{G})Q_{n0}'^{-1}(\beta_0I + \mathbf{G}_{n0}\delta_0)]'Q_n(\theta)'^{-1} \\ &\quad \cdot Q_n(\theta)^{-1}[(\beta_0 - \beta)I + (\mathbf{G}_{n0}\delta_0 - \mathbf{G}\delta)\mathbf{x} + (\lambda_0\mathbf{G}_{n0} - \lambda\mathbf{G})Q_{n0}'^{-1}(\beta_0I + \mathbf{G}_{n0}\delta_0)] \\ &\quad + \sigma_0^2\frac{1}{n}tr(Q_{n0}'^{-1}Q_{n0}^{-1}). \end{aligned}$$

Since the model is uniquely identified at  $\theta_0$ ,  $\frac{1}{n}\mathbb{E}[L_n(\theta)]$  is uniquely minimized at  $\theta_0$  (Instead of assuming  $\mathbb{E}[\mathbf{x}'\mathbf{x}]$  is non-singular, we need to assume  $\lim_{n \rightarrow \infty} \mathbf{x}'_n\mathbf{x}$  has full rank. ). To show the consistency of  $\hat{\theta}_{NLS}$ , we need to check  $\frac{1}{n}L_n(\theta)$  converges in probability to  $\frac{1}{n}\mathbb{E}[L_n(\theta)]$

uniformly in  $\theta \in \Theta$ , where  $\Theta$  is a compact convex set which contains  $\theta_0$  in its interior and  $\frac{1}{n}\mathbb{E}[L_n(\theta)]$  is uniformly equicontinuous on  $\Theta$ . Under a set of similar conditions in Wang and Lee (2013), these two conditions are satisfied, which implies that  $\hat{\theta}_{NLS}$  is consistent. The proof is similar in spirit to the proof of Proposition 2.1 in Wang and Lee (2013), which is omitted here. Since  $\hat{\theta}_{NLS}$  minimizes  $L_n(\theta)$ , we have  $\frac{\partial L_n(\hat{\theta}_{NLS})}{\partial \theta} = 0$  and by Taylor expansion at  $\theta_0$ , we have  $\sqrt{n}(\hat{\theta}_{NLS} - \theta_0) = -[\frac{1}{n}\frac{\partial^2 L_n(\tilde{\theta})}{\partial \theta \partial \theta'}]^{-1} \frac{1}{\sqrt{n}} \frac{\partial L_n(\theta_0)}{\partial \theta}$ . Notice that

$$\begin{aligned} \frac{1}{\sqrt{n}} \frac{\partial L_n(\theta_0)}{\partial \theta} &= \frac{2}{\sqrt{n}} \frac{\partial h'_n(\mathbf{x}_n, \theta_0)}{\partial \theta} [Y_n - h_n(\mathbf{x}_n, \theta_0)] \\ &= -\frac{2}{\sqrt{n}} \frac{\partial h'_n(\mathbf{x}_n, \theta_0)}{\partial \theta} Q_{n0}^{-1} \epsilon_n \end{aligned}$$

Denote

$$\begin{aligned} A_{n,\lambda_0} &= \frac{\partial h_n(\mathbf{x}_n, \theta)}{\partial \lambda} = (I_n - \lambda_0 \mathbf{G}_{n0})^{-1} \mathbf{G}_{n0} (I - \lambda_0 \mathbf{G}_{n0})^{-1} (\beta_0 I_n + \delta \mathbf{G}_{n0}) \mathbf{x}_n, \\ A_{n,\beta_0} &= \frac{\partial h_n(\mathbf{x}_n, \theta)}{\partial \beta} = (I_n - \lambda \mathbf{G}_{n0})^{-1} \mathbf{x}_n, \\ A_{n,\delta_0} &= \frac{\partial h_n(\mathbf{x}_n, \theta)}{\partial \delta} = (I_n - \lambda \mathbf{G}_{n0})^{-1} \mathbf{G}_{n0} \mathbf{x}_n, \\ A_{n,c_k} &= \frac{\partial h_n(\mathbf{x}_n, \theta)}{\partial c_k} = [\lambda_0 (I_n - \lambda_0 \mathbf{G}_{n0})^{-1} \Delta_{c_k} (I_n - \lambda_0 \mathbf{G}_{n0})^{-1} (\beta_0 I_n + \delta_0 \mathbf{G}_{n0}) + \delta_0 (I_n - \lambda_0 \mathbf{G}_{n0})^{-1} \Delta_{c_k}] \mathbf{x}_n, \end{aligned}$$

for  $k = 2, \dots, K$ . Let  $A_n = [A_{n,\lambda_0}, A_{n,\beta_0}, A_{n,\delta_0}, A_{n,c_2}, \dots, A_{n,c_K}]'$ , we then have

$$\frac{1}{\sqrt{n}} \frac{\partial L_n(\theta_0)}{\partial \theta} \xrightarrow{d} N(0, \lim_{n \rightarrow \infty} \frac{4}{n} \sigma_0^2 A_n Q_{n0}^{-1} Q_{n0}'^{-1} A_n'). \quad (5.14)$$

On the other hand, since  $\hat{\theta}_{NLS}$  converges in probability to  $\theta_0$ , we have  $\tilde{\theta} \xrightarrow{p} \theta$ . By a similar uniform law of large numbers in Wang and Lee (2013), we have  $\frac{1}{n} \frac{\partial^2 L_n(\tilde{\theta})}{\partial \theta \partial \theta'} \xrightarrow{p} \frac{1}{n} \frac{\partial^2 L_n(\theta_0)}{\partial \theta \partial \theta'}$ . Notice that

$$\frac{1}{n} \frac{\partial^2 L_n(\tilde{\theta})}{\partial \theta \partial \theta'} = \frac{2}{n} \frac{\partial h'_n(\mathbf{x}_n, \theta)}{\partial \theta} \frac{\partial h_n(\mathbf{x}_n, \theta)}{\partial \theta} + o_p(1) = \frac{2}{n} A_n A_n' + o_p(1), \quad (5.15)$$

Combining the above results, we finally have

$$\sqrt{n}(\hat{\theta}_{NLS} - \theta_0) \xrightarrow{d} N(0, \lim_{n \rightarrow \infty} n(A_n A_n')^{-1} A_n Q_{n0}^{-1} Q_{n0}'^{-1} A_n' (A_n A_n')^{-1})$$

The asymptotic properties of the case in which the dimension of  $\beta$  is large than 1 can be derived in a similar fashion, so we omit the derivation here.

### Descriptive Statistics of the Weibo Data Set:

Here we provide the descriptive statistics for the Sina Weibo data set which is used in empirical application of the paper.

Table 5.1: Descriptive Statistics of the Weibo Data Set

	Mean	S.D.	Min.	Max.
Average number of daily posts	1.258	0.405	0.022	2.419
Age	35.815	9.043	18	64
Male	0.561	0.497	0	1
Number of followers (millions)	8.342	12.196	0.863	123.722

## Proofs of the Main Results in Chapter 2

*Notations.* For any real vector or matrix  $A$ , we use  $A^\top$  to denote the transpose of  $A$ ,  $\lambda_{\max}(A)$  to denote its largest eigenvalue and  $\|A\|_1$  to denote its maximum absolute column sum norm. We use  $A_{ij}$  to denote the  $ij$ th element of a matrix  $A$ . For two positive integers  $a$  and  $b$ , we let  $\mathbf{0}_{a \times b}$  denote the  $a \times b$  matrix consists of zeros and  $\mathbf{1}_a$  denote the  $a$ -dimensional unit vector. For a sequence of random variables  $X_n$ , we let  $\text{plim}_{n \rightarrow \infty} X_n$  denote its probability limit,  $\xrightarrow{p}$  and  $\xrightarrow{d}$  denote convergence in probability and in distribution, respectively.

### Proof of Proposition 1

We first show that the matrix  $(\mathbf{I} - \lambda_1 \mathbf{W}_1 - \lambda_2 \mathbf{W}_2)$  is invertible. Define  $\mathbf{W} = \lambda_1 \mathbf{W}_1 + \lambda_2 \mathbf{W}_2$ , which is a  $n \times n$  matrix. By definition, we just need to show  $(\mathbf{I} - \mathbf{W})$  is invertible. Notice that a sufficient condition to ensure the invertibility of  $(\mathbf{I} - \mathbf{W})$  is  $|\lambda_{\max}(\mathbf{W})| < 1$ ; see, for example, Seber (2008). Let's first consider the structure of  $\mathbf{W}$ , which is shown as follows:

$$\mathbf{W} = \begin{bmatrix} \lambda_1 W_1 & \frac{\lambda_2}{n-n_1} \mathbf{1}_{n_1} \mathbf{1}_{n_2}^\top & \dots & \frac{\lambda_2}{n-n_1} \mathbf{1}_{n_1} \mathbf{1}_{n_K}^\top \\ \frac{\lambda_2}{n-n_2} \mathbf{1}_{n_2} \mathbf{1}_{n_1}^\top & \lambda_1 W_2 & \dots & \frac{\lambda_2}{n-n_2} \mathbf{1}_{n_2} \mathbf{1}_{n_K}^\top \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\lambda_2}{n-n_K} \mathbf{1}_{n_K} \mathbf{1}_{n_1}^\top & \frac{\lambda_2}{n-n_K} \mathbf{1}_{n_K} \mathbf{1}_{n_2}^\top & \dots & \lambda_1 W_K \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

By Assumption 1, it is easy to see that the row sums of  $\mathbf{W}$  all equal  $(\lambda_1 + \lambda_2)$ , which implies the matrix  $\frac{1}{\lambda_1 + \lambda_2} \mathbf{W}$  is row-normalized. Next, notice that for any row-normalized matrix  $A$ , we have  $|\lambda_{\max}(\mathbf{A})| = 1$  (Banerjee et al. (2014)), which implies  $|\lambda_{\max}(\mathbf{W})| = |\lambda_1 + \lambda_2|$ . Then by Assumption 2, we have  $|\lambda_{\max}(\mathbf{W})| = |\lambda_1 + \lambda_2| < |\lambda_1| + |\lambda_2| < 1$ . So, we have  $(\mathbf{I} - \mathbf{W})$  is invertible.

We next show there exists a unique interior Nash equilibrium in the network game defined in the main text. For individual  $i$ ,  $i = 1, \dots, n$ , the first order condition of his utility

maximization problem implies:

$$y_i = \pi_i + \lambda_1 \sum_{j \in G(i)} W_{G(i),ij} y_j + \lambda_2 \bar{y}_{-G(i)}.$$

By the definition of Nash equilibrium in pure strategies, we have the following system of equations:

$$\mathbf{Y} = \lambda_1 \mathbf{W}_1 \mathbf{Y} + \lambda_2 \mathbf{W}_2 \mathbf{Y} + \mathbf{\Pi}.$$

Since  $(\mathbf{I} - \lambda_1 \mathbf{W}_1 - \lambda_2 \mathbf{W}_2)$  is invertible under Assumptions 1 and 2, we have

$$\mathbf{Y} = (\mathbf{I} - \lambda_1 \mathbf{W}_1 - \lambda_2 \mathbf{W}_2)^{-1} \mathbf{\Pi}.$$

### Proof of Proposition 2

We show the identification of the parameters of interest following the method in Bramoullé et al. (2009). We first illustrate the idea by considering a special case  $p = 1$ , i.e.,  $x_i$  is a scalar. By the reduced form of the model and Assumption 4, we have:

$$E[\mathbf{Y}|\mathbf{X}] = (\mathbf{I} - \lambda_1 \mathbf{W}_1 - \lambda_2 \mathbf{W}_2)^{-1} \beta \mathbf{X}.$$

By the definition of identification in Bramoullé et al. (2009), we need to show the structural parameters  $(\lambda_1, \lambda_2, \beta)^\top$  is unique. Suppose there exists a set of different parameters  $(\tilde{\lambda}_1, \tilde{\lambda}_2, \tilde{\beta})^\top$  such that

$$(\mathbf{I} - \lambda_1 \mathbf{W}_1 - \lambda_2 \mathbf{W}_2)^{-1} \beta = (\mathbf{I} - \tilde{\lambda}_1 \mathbf{W}_1 - \tilde{\lambda}_2 \mathbf{W}_2)^{-1} \tilde{\beta}. \quad (5.16)$$

By the above equation, we have  $(\beta - \tilde{\beta})\mathbf{I} + (\lambda_1 \tilde{\beta} - \tilde{\lambda}_1 \beta)\mathbf{W}_1 + (\lambda_2 \tilde{\beta} - \tilde{\lambda}_2 \beta)\mathbf{W}_2 = \mathbf{0}_{n \times n}$ . Then notice that  $\mathbf{I}$ ,  $\mathbf{W}_1$  and  $\mathbf{W}_2$  are linearly independent because of Assumption 1 and the model setup, so we have  $\beta = \tilde{\beta}$ ,  $\lambda_1 \tilde{\beta} = \tilde{\lambda}_1 \beta$  and  $\lambda_2 \tilde{\beta} = \tilde{\lambda}_2 \beta$ . Then by Assumption 3,  $\beta \neq 0$ , we have  $\beta = \tilde{\beta}$ ,  $\lambda_1 = \tilde{\lambda}_1$  and  $\lambda_2 = \tilde{\lambda}_2$ , which implies all the parameters are identified.

We next consider the general case  $p > 1$ . The reduced form of the model is given by:

$$E[\mathbf{Y}|\mathbf{X}] = (\mathbf{I} - \lambda_1 \mathbf{W}_1 - \lambda_2 \mathbf{W}_2)^{-1} \mathbf{X} \beta.$$

If the parameters are not identified, then there must exist another set of parameters  $(\tilde{\lambda}_1, \tilde{\lambda}_2, \tilde{\beta}^\top)^\top$  such that

$$(\mathbf{I} - \lambda_1 \mathbf{W}_1 - \lambda_2 \mathbf{W}_2)^{-1} \mathbf{X} \beta = (\mathbf{I} - \tilde{\lambda}_1 \mathbf{W}_1 - \tilde{\lambda}_2 \mathbf{W}_2)^{-1} \mathbf{X} \tilde{\beta}, \quad (5.17)$$

for any value of  $\mathbf{X}$ . By Assumption 3 and linear algebra, the equation (2) implies  $\beta = \tilde{\beta}$  and  $(\mathbf{I} - \lambda_1 \mathbf{W}_1 - \lambda_2 \mathbf{W}_2)^{-1} = (\mathbf{I} - \tilde{\lambda}_1 \mathbf{W}_1 - \tilde{\lambda}_2 \mathbf{W}_2)^{-1}$ . Then we have  $(\lambda_1 - \tilde{\lambda}_1) \mathbf{W}_1 + (\lambda_2 - \tilde{\lambda}_2) \mathbf{W}_2 = \mathbf{0}_{n \times n}$ . Then by the structure of  $\mathbf{W}_1$  and  $\mathbf{W}_2$ , we have  $\lambda_1 = \tilde{\lambda}_1$  and  $\lambda_2 = \tilde{\lambda}_2$ . So, all the parameters are identified.

### Proof of Proposition 3

Recall  $\mathbf{Z} = (\mathbf{W}_1 \mathbf{Y}, \mathbf{W}_2 \mathbf{Y}, \mathbf{X})$ ,  $\mathbf{P}_H = \mathbf{H}(\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top$  and define  $\hat{\mathbf{Z}} = \mathbf{P}_H \mathbf{Z}$ . Then we have

$$\begin{aligned} \hat{\theta}_{2\text{SLS}} &= (\mathbf{Z}^\top \mathbf{P}_H \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{P}_H \mathbf{Y} \\ &= (\hat{\mathbf{Z}}^\top \hat{\mathbf{Z}})^{-1} \hat{\mathbf{Z}}^\top \mathbf{Y} \\ &= (\hat{\mathbf{Z}}^\top \hat{\mathbf{Z}})^{-1} \hat{\mathbf{Z}}^\top (\mathbf{Z} \theta + \epsilon) \\ &= \theta + [\mathbf{Z}^\top \mathbf{H}(\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \mathbf{Z}]^{-1} [\mathbf{Z}^\top \mathbf{H}(\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \epsilon], \end{aligned}$$

where the last equality holds because  $\mathbf{H}$  contains  $\mathbf{X}$ . The above equation directly implies

$$\sqrt{n}(\hat{\theta}_{2\text{SLS}} - \theta) = [(\frac{1}{n} \mathbf{Z}^\top \mathbf{H})(\frac{1}{n} \mathbf{H}^\top \mathbf{H})^{-1}(\frac{1}{n} \mathbf{H}^\top \mathbf{Z})]^{-1} [(\frac{1}{n} \mathbf{Z}^\top \mathbf{H})(\frac{1}{n} \mathbf{H}^\top \mathbf{H})^{-1} \frac{1}{\sqrt{n}} \mathbf{H}^\top \epsilon].$$

By Assumption 9, we have  $\lim_{n \rightarrow \infty} \frac{1}{n} \mathbf{H}^\top \mathbf{Z} = Q_{\text{HZ}}$  and  $\lim_{n \rightarrow \infty} \frac{1}{n} \mathbf{H}^\top \mathbf{H} = Q_{\text{HH}}$ . We next show that the elements of  $\mathbf{H}$  are bounded in absolute value. By Assumption 7 and Assumption 8, the elements of  $\mathbf{H}$  will be bounded in absolute value if  $\mathbf{W}_1$  and  $\mathbf{W}_2$  have uniformly



bounded row and columns sums. We consider  $\mathbf{W}_1$  first. The row sums of  $\mathbf{W}_1$  is uniformly bounded by Assumption 1. The column sums of  $\mathbf{W}_1$  is uniformly bounded by Assumption 6 and the structure of  $\mathbf{W}_1$ . We next consider  $\mathbf{W}_2$ . The row sums of  $\mathbf{W}_2$  are uniformly bounded because of the structure of  $\mathbf{W}_2$ . The column sums of  $\mathbf{W}_2$  is uniformly bounded because of Assumption 5 and its structure. To see this, simply notice that  $\lim_{n \rightarrow \infty} \|\mathbf{W}_2\|_1 \leq \frac{n}{n - (\max_k s_k)n} \leq \frac{1}{1 - \max_k s_k} \leq \frac{1}{(K-1)c}$ , where  $c$  is the positive constant defined in Assumption 5. So, we have shown the row and column sums of  $\mathbf{W}_1$  and  $\mathbf{W}_2$  are uniformly bounded. Combining this with Assumption 7 and Assumption 8, the elements of  $\mathbf{H}$  is bounded in absolute values. By Assumption 9,  $\lim_{n \rightarrow \infty} n^{-1} \mathbf{H}' \mathbf{H}$  exists and is finite and nonsingular. We then apply the Theorem A.1 in Kelejian and Prucha (1998) and the Slutsky's Theorem to conclude  $\sqrt{n}(\hat{\theta}_{2SLS} - \theta) \xrightarrow{d} N(0, [Q_{HZ}^\top Q_{HH}^{-1} Q_{HZ}]^{-1})$ .

### An Iterative Algorithm for the Optimization Problem (3.4) in Chapter 3:

- (i) Choose some initial value of  $\hat{\delta}_\tau$ , which is denoted by  $\hat{\delta}_\tau^{(0)}$  and set  $s = 0$ .
- (ii) Given the group-specific parameters  $\hat{\delta}_\tau^{(s)}$ , estimate the group memberships  $\hat{g}_\tau^{(s+1)}$  by

$$\hat{g}_{i,\tau}^{(s+1)} = \arg \min_{g_{i,\tau} \in \{1, \dots, K^0\}} \frac{1}{T} \sum_{t=1}^T \rho_\tau(y_{it} - x'_{it} \alpha_{g_{i,\tau}}^{(s)} - w'_{it} \gamma_{g_{i,\tau}}^{(s)}),$$

for all  $i = 1, \dots, N$ .

- (iii) Given the group memberships  $\hat{g}_\tau^{(s+1)}$ , estimate the group-specific parameters  $\hat{\delta}_\tau^{(s+1)}$  by solving the following optimization problem

$$\hat{\delta}_{G_k,\tau}^{(s+1)} = \arg \min_{\delta_\tau \in \mathbb{R}^{p+qJ}} \frac{1}{NT} \sum_{\hat{g}_i^{(s+1)}=k} \sum_{t=1}^T \rho_\tau(y_{it} - x'_{it} \alpha_{\hat{g}_i^{(s+1)},\tau} - w'_{it} \gamma_{\hat{g}_i^{(s+1)},\tau}),$$

for all  $k = 1, \dots, K^0$ .

- (iv) Repeat Step 3 and Step 4 until the group memberships  $\hat{g}_\tau$  or the group-specific parameters  $\hat{\delta}_\tau$  converge.

The iterative algorithm here consists of two main parts: (1) estimating group memberships  $g_\tau$  according to given values of group-specific parameters (Step 2); (2) updating the group-specific parameters  $\delta_\tau$  based on group memberships  $g_\tau$  (Step 3). As pointed out in Bonhomme and Manresa (2015) and Liu et al. (2019), the solution produced by the iterative algorithm can be sensitive to the choice of initial values of  $\delta_\tau^{(0)}$  because it is not guaranteed to be a global optimum. To address this issue, Bonhomme and Manresa (2015) recommended trying different initial values and picking the one with the minimum loss, and we follow their practice in our paper. Here, we propose three strategies for generating initial values.

**Strategy 1** Randomly assign  $N$  individuals into  $K^0$  groups (each group consists of  $N/K^0$  individuals). Then estimate  $\delta_{G_k,\tau}$  using individuals assigned to the group  $G_k$ , for all  $k = 1, \dots, K^0$ .

**Strategy 2** First, estimate  $\delta_{i,\tau}$  using  $\{y_{it}, x_{it}, z_{it}\}_{t=1}^T$  for all  $i = 1, \dots, N$ . Second, implement the standard multivariate  $k$ -means algorithm with  $K^0$  groups using the estimates  $\{\delta_{1,\tau}, \dots, \delta_{N,\tau}\}$  as input. We consider three different cases (1) only use the parametric coefficients  $\alpha_i$ ; (2) only use the functional coefficients  $\gamma_i$ ; (3) both  $\alpha_i$  and  $\gamma_i$  are used. This practice gives us three sets of initial values.

**Strategy 3** First, treat all individuals as a single group and estimate  $\delta_\tau^*$  using all the data. Second, generate  $\delta_{G_k,\tau}^{(0)} = \delta_\tau^* + a \cdot U_k$ <sup>1</sup>, where  $U_k$  is a  $(p + qJ)$ -dimensional normal random vector with mean zero and the diagonal variance matrix whose elements are  $\{|\delta_{1,\tau}^*|^2, \dots, |\delta_{p+qJ,\tau}^*|^2\}$  for all  $k = 1, \dots, K^0$ . Here  $\delta_s^*$  denotes the  $s$ -th element of the  $(p + qJ)$ -dimensional  $\delta_\tau^*$ .

**Remark 1** In Monte Carlo simulations, we estimate the model using thirteen initial values: five from Strategy 1, five from Strategy 3, and three generated using Strategy 2. For each initial value, we estimate the parameters of interest using the proposed iterative algorithm and record the corresponding loss in 3.4, then we pick the one that gives the minimum loss. We find that the iterative algorithm with this procedure performs very well in terms of estimation bias and standard deviation and runs quite fast<sup>2</sup>. So we keep the same procedure in the empirical application section.

---

<sup>1</sup>In Monte Carlo simulations, we find that the Strategy 3 with  $a = 1.0$  performs satisfactorily, so we also let  $a = 1.0$  in the empirical application.

<sup>2</sup>Our estimation method can be carried out easily with popular statistical software, including R, matlab or State with minor modification on the built-in function, such as the ‘quanreg’ command in R or the ‘qreg’ command in Stata. It takes a computer equipped with Intel i7-7700K CPU and 16G 2400MHZ RAM about 20 minutes to run 1000 repetitions of each DGP in Section 4.5.

## Proofs of the Main Results in Chapter 3

### Proof of Lemma 1:

By definition we have  $\delta_{G_k, \tau}^0 = (\alpha_{G_k, \tau}^{0'}, \gamma_{G_k, \tau}^{0'})' \in \mathbb{R}^{p+qJ}$  and  $\theta_{G_k, \tau}^0 = (\alpha_{G_k, \tau}^{0'}, \beta_{G_k, \tau}^{0'})' \in \mathbb{R}^{p+q}$  for all  $k = 1, \dots, K^0$ . For two different groups  $G_k, G_l \in \mathcal{G} = \{G_1, \dots, G_{K^0}\}$ , since

$$\left\| \delta_{G_k, \tau}^0 - \delta_{G_l, \tau}^0 \right\|^2 = \left\| \alpha_{G_k, \tau}^0 - \alpha_{G_l, \tau}^0 \right\|^2 + \left\| \gamma_{G_k, \tau}^0 - \gamma_{G_l, \tau}^0 \right\|^2,$$

To show  $\left\| \delta_{G_k, \tau}^0 - \delta_{G_l, \tau}^0 \right\|^2 > 0$ , it is sufficient to consider two different cases:

(1)

$$\left\| \alpha_{G_k, \tau}^0 - \alpha_{G_l, \tau}^0 \right\| > c,$$

(2)

$$\left\| \alpha_{G_k, \tau}^0 - \alpha_{G_l, \tau}^0 \right\| = 0,$$

but

$$\left\| \beta_{G_k, \tau}^0 - \beta_{G_l, \tau}^0 \right\|_2 > c.$$

In the first case, notice that by Assumption 2(ii), there exists a positive constant  $c$  such that

$$\begin{aligned} c &< \left\| \alpha_{G_k, \tau}^0 - \alpha_{G_l, \tau}^0 \right\| \\ &\leq \left\| \delta_{G_k, \tau}^0 - \delta_{G_l, \tau}^0 \right\|. \end{aligned} \tag{5.18}$$

In the second case, by Assumption 2(ii), there exists a positive constant  $c$  and some  $1 \leq$

$m \leq q$  such that

$$\begin{aligned}
c &< \left\| \beta_{G_k m, \tau}^0 - \beta_{G_l m, \tau}^0 \right\|_2 \\
&\leq \left\| \beta_{G_k m, \tau}^0 - P(u)' \gamma_{G_k m, \tau}^0 \right\|_2 + \left\| P(u)' \gamma_{G_k m, \tau}^0 - P(u)' \gamma_{G_l m, \tau}^0 \right\|_2 + \left\| \beta_{G_l m, \tau}^0 - P(u)' \gamma_{G_l m, \tau}^0 \right\|_2 \\
&= o(1) + \left\{ (\gamma_{G_k m, \tau}^0 - \gamma_{G_l m, \tau}^0)' \left( \int P(u) P(u)' du \right) (\gamma_{G_k m, \tau}^0 - \gamma_{G_l m, \tau}^0) \right\}^{\frac{1}{2}} \\
&\leq o(1) + \sqrt{\lambda_{\max} \left( \int P(u) P(u)' du \right)} \left\| \gamma_{G_k m, \tau}^0 - \gamma_{G_l m, \tau}^0 \right\|, \tag{5.19}
\end{aligned}$$

where the second inequality is by triangular inequality, the third inequality is by Assumption 2(ii) and  $J \rightarrow \infty$  and the definition of  $L_2$  norm, the fourth inequality is by Assumption 2(i) and the inequality  $c'Ac \leq \lambda_{\max}(A)\|c\|^2$ , where  $A$  is some  $p \times p$  matrix and  $c$  is a  $p \times 1$  constant vector. Then equation 5.19 implies that  $\left\| \gamma_{G_k m, \tau}^0 - \gamma_{G_l m, \tau}^0 \right\| > c$  for some constant  $c > 0$  when  $J \rightarrow \infty$ . Finally, we have  $\left\| \delta_{G_k, \tau}^0 - \delta_{G_l, \tau}^0 \right\| \geq \left\| \gamma_{G_k m, \tau}^0 - \gamma_{G_l m, \tau}^0 \right\| > c > 0$  when  $J \rightarrow \infty$ . Combining the results of the above two cases together, Lemma 1 is proved.

**Proof of Theorem 1:** For all  $1 \leq i \leq N$ , we let  $\mathbf{W}_i = (w_{i1}, \dots, w_{iT})' \in \mathbb{R}^{T \times qJ}$ ,  $\mathbf{Y}_i = (y_{i1}, \dots, y_{iT})' \in \mathbb{R}^{T \times 1}$ ,  $\mathbf{X}_i = (x_{i1}, \dots, x_{iT})' \in \mathbb{R}^{T \times p}$ ,  $\mathbf{R}_i = (R_{i1}, \dots, R_{iT})' \in \mathbb{R}^{T \times 1}$  and  $\mathbf{e}_i = (e_{i1}, \dots, e_{iT})' \in \mathbb{R}^{T \times 1}$ , where  $R_{it} = w'_{it} \gamma_i^0 - z'_{it} \beta(u_{it})$  denotes the approximation error, where  $\gamma_i^0 = (\gamma_{i1}^0, \dots, \gamma_{i,q}^0)$ , which is defined in Assumption 3(ii). Furthermore, we let

$$\mathbf{f}_i = \begin{bmatrix} f_{i1}(0) & 0 & \dots & 0 \\ 0 & f_{i2}(0) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & f_{iT}(0) \end{bmatrix} \in \mathbb{R}^{T \times T},$$

where  $f_{it}(0)$  is the conditional density function of  $e_{it}$  evaluated at zero, and we let  $\psi(u) =$

$\tau - 1\{u \leq 0\}$ . For simplicity, we suppress the subscript  $\tau$ . For individual  $i$ , we have

$$\mathbf{Y}_i = \mathbf{W}_i \gamma_i^0 + \mathbf{X}_i \alpha_i^0 - \mathbf{R}_i + \mathbf{e}_i. \quad (5.20)$$

Define  $\tilde{\mathbf{W}}_i = \mathbf{W}_i (\mathbf{W}'_i \mathbf{f}_i \mathbf{W}_i)^{-1} \mathbf{W}'_i \mathbf{f}_i$ . Then equation 5.20 can be rewritten as

$$\begin{aligned} \mathbf{Y}_i = & \mathbf{W}_i (\mathbf{W}'_i \mathbf{f}_i \mathbf{W}_i)^{-1/2} \{ (\mathbf{W}'_i \mathbf{f}_i \mathbf{W}_i)^{1/2} \gamma_i^0 + (\mathbf{W}'_i \mathbf{f}_i \mathbf{W}_i)^{-1/2} \mathbf{W}'_i \mathbf{f}_i \mathbf{X}_i \alpha_i^0 \} \\ & + T^{-1/2} (\mathbf{X}_i - \tilde{\mathbf{W}}_i \mathbf{X}_i) (T^{1/2} \alpha_i^0) - \mathbf{R}_i + \mathbf{e}_i. \end{aligned} \quad (5.21)$$

Furthermore, for individual  $i$ , we let  $v_{1,it}$  be the  $t$ -th row of  $\mathbf{W}_i (\mathbf{W}'_i \mathbf{f}_i \mathbf{W}_i)^{-1/2}$  and  $v_{2,it}$  be the  $t$ -th row of  $T^{-1/2} (\mathbf{X}_i - \tilde{\mathbf{W}}_i \mathbf{X}_i)$ , and  $v_{it} = (v_{1,it}, v_{2,it})$ . And we let

$$\begin{aligned} \theta_{i1}^0 &= (\mathbf{W}'_i \mathbf{f}_i \mathbf{W}_i)^{1/2} \gamma_i^0 + (\mathbf{W}'_i \mathbf{f}_i \mathbf{W}_i)^{-1/2} \mathbf{W}'_i \mathbf{f}_i \mathbf{X}_i \alpha_i^0 \in \mathbb{R}^{qJ}. \\ \theta_{i2}^0 &= T^{1/2} \alpha_i^0 \in \mathbb{R}^p. \end{aligned}$$

Let  $\theta_i^0 = (\theta_{i1}^0, \theta_{i2}^0)'$ , we have

$$y_{it} = v_{it} \theta_i^0 - R_{it} + e_{it} = v_{1,it} \theta_{i1}^0 + v_{2,it} \theta_{i2}^0 - R_{it} + e_{it}.$$

Let  $\Delta$  denote a set which contains  $K^0$  different elements in  $\mathbb{R}^{p+qJ}$ . Consider the following objective function:

$$L_{NT}(\Delta) = \frac{1}{N} \sum_{i=1}^N \min_{\delta_i \in \Delta} \left[ \frac{1}{T} \sum_{t=1}^T \rho_\tau(y_{it} - x'_{it} \alpha_i - w'_{it} \gamma_i) \right], \quad (5.22)$$

where  $\delta_i = (\alpha'_i, \gamma'_i)' \in \mathbb{R}^{p+qJ}$  and  $w_{it}$  is defined in the main text. Comparing the above objective function with the one in the optimization problem 3.4, the only difference is that the group memberships are concentrated out here. It is straightforward to see that the solution minimizes the objective function 5.22 also solves the optimization problem in the

main text (3.4), i.e.,

$$\hat{\Delta} = \inf_{\Delta \in \mathbb{R}^{(p+qJ) \cdot K^0}} \frac{1}{N} \sum_{i=1}^N \min_{\delta_i \in \Delta} \left[ \frac{1}{T} \sum_{t=1}^T \rho_\tau(y_{it} - x'_{it}\alpha_i - w'_{it}\gamma_i) \right]. \quad (5.23)$$

In fact, since the optimization problem 5.23 is equivalent to 3.4 with concentrated  $g$ , we have  $\hat{\delta} = \hat{\Delta}$ , where  $\hat{\delta}$  is defined the the main text, which implies that we only need to focus on the solution that solves 5.23.

Notice that we have  $y_{it} = v_{1,it}\theta_{i,1}^0 + v_{2,it}\theta_{i,2}^0 - R_{it} + e_{it}$  by definition, which further gives us  $y_{it} - x'_{it}\alpha_i - w'_{it}\gamma_i = e_{it} - v_{1,it}(\theta_{i,1}^0 - \theta_{i,1}) - v_{2,it}(\theta_{i,2}^0 - \theta_{i,2}) - R_{it} = e_{it} - v_{it}\theta_i - R_{it}$ , where  $\theta_i = (\theta_{i,1}^0 - \theta'_{i,1}, \theta_{i,2}^0 - \theta'_{i,2})'$  and  $v_{it} = (v_{1,it}, v_{2,it})$ . By Lemma 7, we have the following results.

(1) For any individual  $1 \leq i \leq N$ , for any constant  $\varepsilon > 0$ , there exists a constant  $c$  such that

$$\lim_{T \rightarrow \infty} \Pr \left( \inf_{\|\theta_i\| > c \cdot J^{\frac{1}{2}}} \sum_{t=1}^T \rho_\tau(e_{it} - v_{it}\theta_i - R_{it}) > \sum_{t=1}^T \rho_\tau(e_{it} - R_{it}) \right) > 1 - \varepsilon.$$

(2) Let  $\Theta = \{\theta : \|\theta_i\| \leq c \cdot J^{\frac{1}{2}}\}$  for some finite constant  $c$ . For  $\theta_i^*$  and  $\tilde{\theta}_i$  such that  $\theta_i^*, \tilde{\theta}_i \in \Theta$ , we have

$$\lim_{T \rightarrow \infty} \Pr \left( \sum_{t=1}^T \rho_\tau(e_{it} - v_{it}\theta_i^* - R_{it}) = \sum_{t=1}^T \rho_\tau(e_{it} - v_{it}\tilde{\theta}_i - R_{it}) \right) > 1 - \varepsilon.$$

for any constant  $\varepsilon > 0$ .

We next show that claims (1) and (2) imply the consistency of the estimators. First, by Lemma 1, we have  $\|\theta_{G_k}^0 - \theta_{G_j}^0\| > c > 0$ , for some  $c > 0$  as  $J \rightarrow \infty$  for all  $1 \leq k \neq j \leq K^0$ , which implies that different groups will have different "true" parameters, i.e., the elements contained in  $\Delta$  are different from each other. This further implies the corresponding  $\theta_i^0$  are different for individuals in different groups.

Second, we show that  $\|\hat{\Delta} - \Delta^0\| = o_p(1)$  by showing  $\|(\hat{\alpha}'_i, \hat{\gamma}'_i)' - (\alpha_i^0, \gamma_i^0)'\| = o_p(1)$ . We claim that for any group  $1 \leq k \leq K^0$ , there exists  $\hat{\delta}_{\sigma(k)} \in \hat{\Delta}$  such that the corresponding  $\hat{\theta}_{\sigma(k)} \in \Theta$  with probability approaching 1. Suppose not, then there are a non-negligible proportion of

individuals such that their  $\hat{\theta}_i > cJ^{1/2}$ . Without loss of generality, we assume there exists a such group indexed by  $k$ . Then we have

$$\sum_{i \in G_k} \sum_{t=1}^T \rho_\tau(e_{it} - v_{it}\hat{\theta}_i - R_{it}) > \sum_{i \in G_k} \sum_{t=1}^T \rho_\tau(e_{it} - R_{it}) \quad (5.24)$$

with probability approaching 1 by (1). Combining 5.24 and (2), we then have

$$\sum_{i=1}^N \sum_{t=1}^T \rho_\tau(e_{it} - v_{it}\hat{\theta}_i - R_{it}) > \sum_{i=1}^N \sum_{t=1}^T \rho_\tau(e_{it} - R_{it}) + O_p(1). \quad (5.25)$$

Since  $\hat{\Delta}$  solves the optimization problem 3.4, we then have  $\hat{\theta}_{\sigma(k)} \in \Theta$  for all  $1 \leq k \leq K^0$  with probability approaching 1, which gives us  $\|\hat{\theta}_{i,1} - \theta_{i,1}^0\| = O_p(J^{1/2})$  and  $\|\hat{\theta}_{i,2} - \theta_{i,2}^0\| = O_p(J^{1/2})$  for all  $1 \leq i \leq N$ . By the definition of  $\hat{\theta}_{i,2}$ , we have

$$\|\hat{\alpha}_i - \alpha_i^0\| = O(T^{-1/2}J^{1/2}).$$

For  $\hat{\theta}_{i,1}$ , notice that

$$\hat{\theta}_{i,1} - \theta_{i,1}^0 = (\mathbf{W}'_i \mathbf{f}_i \mathbf{W}_i)^{\frac{1}{2}}(\hat{\gamma}_i - \gamma_i^0) + (\mathbf{W}'_i \mathbf{f}_i \mathbf{W}_i)^{-\frac{1}{2}} \mathbf{W}'_i \mathbf{f}_i \mathbf{X}_i(\hat{\alpha}_i - \alpha_i^0),$$

which implies that

$$\left\| (\mathbf{W}'_i \mathbf{f}_i \mathbf{W}_i)^{\frac{1}{2}}(\hat{\gamma}_i - \gamma_i^0) \right\| \leq \left\| \hat{\theta}_{i,1} - \theta_{i,1}^0 \right\| + \left\| (\mathbf{W}'_i \mathbf{f}_i \mathbf{W}_i)^{-\frac{1}{2}} \mathbf{W}'_i \mathbf{f}_i \mathbf{X}_i(\hat{\alpha}_i - \alpha_i^0) \right\|.$$

By Lemma 3, the eigenvalues of  $\mathbf{W}'_i \mathbf{f}_i \mathbf{W}_i$  and  $\mathbf{W}_i(\mathbf{W}'_i \mathbf{f}_i \mathbf{W}_i)^{-1} \mathbf{W}'_i$  are bounded and bounded away from zero and above, which implies that  $\|\hat{\gamma}_i - \gamma_i^0\| = O_p(T^{-1/2}J^{1/2})$ . So for all  $1 \leq i \leq$



$N$  and  $1 \leq l \leq q$ , we have

$$\begin{aligned} \left\| \hat{\beta}_{il} - \beta_{il}^0 \right\|_2 &= \left\| \hat{\beta}_{il} - P(u)' \gamma_{il}^0 + P(u)' \gamma_{il}^0 - \beta_{il}^0 \right\|_2 \\ &\leq \left\| \hat{\beta}_{il} - P(u)' \gamma_{il}^0 \right\|_2 + \left\| P(u)' \gamma_{il}^0 - \beta_{il}^0 \right\|_2 \\ &= O_p(T^{-1/2} J^{1/2} + J^{-\kappa}). \end{aligned}$$

Finally, notice that  $\hat{\delta}_i$ ,  $i = 1, \dots, N$ , takes  $K^0$  different values because of the setup of the optimization problem, we have  $\left\| \hat{\alpha}_{\hat{G}_{\sigma(k)}} - \alpha_{G_k}^0 \right\| = O(T^{-1/2} J^{1/2})$ . and

$$\left\| \hat{\beta}_{\hat{G}_{\sigma(k)l}} - \beta_{G_{kl}}^0 \right\|_2 = O_p(T^{-1/2} J^{1/2} + J^{-\kappa}), \text{ for all } k = 1, \dots, K^0.$$

Then when  $T \rightarrow \infty$  and  $J \rightarrow \infty$ , which is ensured by Assumption 2 (iii), we have

$$\left\| \hat{\alpha}_{\hat{G}_{\sigma(k)}} - \alpha_{G_k}^0 \right\| = o_p(1), \text{ and } \left\| \hat{\beta}_{\hat{G}_{\sigma(k)l}} - \beta_{G_{kl}}^0 \right\|_2 = o_p(1), \text{ for all } k = 1, \dots, K^0 \text{ and } 1 \leq l \leq q.$$

**Proof for Corollary 1:** By Lemma 1, we have

$$\left\| \delta_{G_k, \tau}^0 - \delta_{G_l, \tau}^0 \right\| > c > 0,$$

for some constant  $c > 0$ , and any  $G_k, G_l \in \mathcal{G} = \{G_1, \dots, G_{K^0}\}$  and  $k \neq l$ . This implies

$$\begin{aligned} \left\| \theta_{i1}^0 - \theta_{-i1}^0 \right\| &= \left\| (\mathbf{W}'_i \mathbf{f}_i \mathbf{W}_i)^{1/2} (\gamma_i^0 - \gamma_{-i1}^0) + (\mathbf{W}'_i \mathbf{f}_i \mathbf{W}_i)^{-1/2} \mathbf{W}'_i \mathbf{f}_i \mathbf{X}_i (\alpha_i^0 - \alpha_{-i1}^0) \right\| \\ &\geq c \cdot T^{1/2}, \end{aligned} \tag{5.26}$$

for some constant  $c > 0$ , where  $\gamma_{-i1}^0$  and  $\alpha_{-i1}^0$  is the pseudo true group-specific parameters of the group other than  $i$ 's. The above result further implies that  $\left\| \theta_i^0 - \theta_{-i}^0 \right\| \geq c \cdot T^{1/2}$ , for some constant  $c > 0$  and  $i = 1, \dots, N$ . By Theorem 1, we have

$$\left\| \hat{\theta}_i - \theta_i^0 \right\| = o_p(1). \tag{5.27}$$

Then, by 5.26 and 5.27 together, we have  $\left\| \hat{\theta}_i - \hat{\theta}_{-i} \right\| > c \cdot T^{1/2}$  for some constant  $c > 0$  under

Assumptions 1, 2 and 3. By the Claim (1) and Claim (2) in the proof of Theorem 1, we then have

$$\lim_{T \rightarrow \infty} \Pr \left( \sum_{t=1}^T \rho_{\tau}(e_{it} - v_{it} \hat{\theta}_{-i} - R_{it}) > \sum_{t=1}^T \rho_{\tau}(e_{it} - v_{it} \hat{\theta}_i - R_{it}) \right) > 1 - \varepsilon.$$

Since  $\hat{\theta}_i$  minimizes the loss for individual  $i$  and by the nature of the optimization problem, the above result then gives us  $\lim_{T \rightarrow \infty} P(\hat{g}_i = g_i^0) = 1$ .

**Proof for Theorem 2:**

To prove Theorem 2, we introduce more notations for the sake of simplicity. Recall that  $w_{it} = (z_{it,1}P(u_{it})', \dots, z_{it,q}P(u_{it})')' \in \mathbb{R}^{qJ}$ . Suppose all individuals are perfectly classified into  $K^0$  groups. Without loss of generality, we can only consider the asymptotic distribution of the  $k$ -th group  $G_k$  which consists of  $N_k$  individuals. For all  $i \in G_k$ , we let

$$\begin{aligned} \mathbf{W}_{G_k} &= (w_{i1}, \dots, w_{iT}, \dots, w_{N_k1}, \dots, w_{N_kT})' \in \mathbb{R}^{N_k T \times qJ}, \\ \mathbf{Y}_{G_k} &= (y_{11}, \dots, y_{1T}, \dots, y_{N_k1}, \dots, y_{N_kT})' \in \mathbb{R}^{N_k T \times 1}, \\ \mathbf{X}_{G_k} &= (x_{11}, \dots, x_{1T}, \dots, x_{N_k1}, \dots, x_{N_kT})' \in \mathbb{R}^{N_k T \times p}, \\ \mathbf{R}_{G_k} &= (R_{11}, \dots, R_{1T}, \dots, R_{N_k1}, \dots, R_{N_kT})' \in \mathbb{R}^{N_k T \times 1}, \\ \mathbf{e}_{G_k} &= (e_{11}, \dots, e_{1T}, \dots, e_{N_k1}, \dots, e_{N_kT})' \in \mathbb{R}^{N_k T \times 1}, \end{aligned}$$

where  $R_{it} = w'_{it} \gamma_{G_k}^0 - z'_{it} \beta(u_{it})$  denotes the approximation error,  $\gamma_{G_k}^0 = (\gamma_{G_k,1}^0, \dots, \gamma_{G_k,q}^0)$ , which

is defined in Assumption 2(ii). Furthermore, we let

$$\mathbf{f}_{G_k} = \begin{bmatrix} f_{11}(0) & 0 & \dots & \dots & 0 \\ 0 & f_{12}(0) & \dots & \dots & 0 \\ \vdots & \vdots & \ddots & \dots & \vdots \\ \vdots & \vdots & \dots & \ddots & \vdots \\ 0 & 0 & \dots & \dots & f_{N_k T}(0) \end{bmatrix} \in \mathbb{R}^{N_k T \times N_k T},$$

where  $f_{it}(0)$  is the conditional density function of  $e_{it}$  evaluated at zero for all  $i \in G_k$ , and we let  $\psi(u) = \tau - 1\{u \leq 0\}$ . For simplicity, we suppress the subscripts  $G_k$  and  $\tau$ . At this moment, the semiparametric quantile panel regression model can be rewritten as

$$\mathbf{Y} = \mathbf{W}\gamma^0 + \mathbf{X}\alpha^0 - \mathbf{R} + \mathbf{e}. \quad (5.28)$$

We further define  $\tilde{\mathbf{W}} = \mathbf{W}(\mathbf{W}'\mathbf{f}\mathbf{W})^{-1}\mathbf{W}'\mathbf{f}$ . Then equation 5.28 can be rewritten as

$$\begin{aligned} \mathbf{Y} = & \mathbf{W}(\mathbf{W}'\mathbf{f}\mathbf{W})^{-1/2} \{ (\mathbf{W}'\mathbf{f}\mathbf{W})^{1/2} \gamma_0 + (\mathbf{W}'\mathbf{f}\mathbf{W})^{-1/2} \mathbf{W}'\mathbf{f}\mathbf{X}\alpha^0 \} \\ & + (N_k T)^{-1/2} (\mathbf{X} - \tilde{\mathbf{W}}\mathbf{X}) ((N_k T)^{1/2} \alpha^0) - \mathbf{R} + \mathbf{e}. \end{aligned} \quad (5.29)$$

Furthermore, we let  $v_{1,it}$  be the  $((i-1) \times T + t)$ -th row of  $\mathbf{W}(\mathbf{W}'\mathbf{f}\mathbf{W})^{-1/2}$  and  $v_{2,it}$  be the  $((i-1) \times T + t)$ -th row of  $(N_k T)^{-1/2}(\mathbf{X} - \tilde{\mathbf{W}}\mathbf{X})$ , and  $v_{it} = (v_{1,it}, v_{2,it})$ , for  $i = 1, \dots, N_k T$ .

And we let

$$\theta_1^0 = (\mathbf{W}'\mathbf{f}\mathbf{W})^{1/2} \gamma_0 + (\mathbf{W}'\mathbf{f}\mathbf{W})^{-1/2} \mathbf{W}'\mathbf{f}\mathbf{X}\alpha^0 \in \mathbb{R}^{qJ}.$$

$$\theta_2^0 = (N_k T)^{1/2} \alpha^0 \in \mathbb{R}^p.$$

Let  $\theta^0 = (\theta_1^0, \theta_2^0)'$ , we then have

$$y_{it} = v_{it}\theta^0 - R_{it} + e_{it} = v_{1,it}\theta_1^0 + v_{2,it}\theta_2^0 - R_{it} + e_{it}.$$

Define the so-called oracle estimator and establish its asymptotic properties. The oracle estimator is given by

$$\hat{\delta}_{G_k, \tau}^* = \arg \min_{\delta \in \mathbb{R}^{p+qJ}} \frac{1}{NT} \sum_{i \in G_k} \sum_{t=1}^T \rho_\tau(y_{it} - x'_{it}\alpha - w'_{it}\gamma),$$

where  $\delta = (\alpha', \gamma)'$ . The oracle estimator assumes that the correct group memberships are known *ex ante*, so there is no estimation error from the estimation of group memberships. We first establish the asymptotic properties of the oracle estimators following the similar idea in Wei et al. (2006). The main difference is that Wei et al. (2006) uses spline polynomials as basis functions while we do not assign specific basis. First recall the model can be rewritten as  $y_{it} = v_{1,it}\theta_1^0 + v_{2,it}\theta_2^0 - R_{it} + e_{it}$ , where  $v_{1,it}, v_{2,it}, \theta_1^0, \theta_2^0, R_{it}$  are defined in the discussion before Theorem 2. We can rewrite the optimization problem 3.4 as an optimization problem on  $\theta_1$  and  $\theta_2$ , i.e.,

$$(\hat{\theta}_1, \hat{\theta}_2) = \arg \min_{\theta_1, \theta_2} \sum_{i \in G_k} \sum_{t=1}^T \rho_\tau(e_{it} - v_{1,it}(\theta_1 - \theta_1^0) - v_{2,it}(\theta_2 - \theta_2^0) - R_{it}). \quad (5.30)$$

We can further transform the optimization problem as follows. let  $\tilde{\theta} = (\tilde{\theta}'_1, \tilde{\theta}'_2)' = (\hat{\theta}'_1 - \theta_1^0, \hat{\theta}'_2 - \theta_2^0)'$ , which solves

$$\tilde{\theta} = \arg \min_{\theta_1, \theta_2} \sum_{i \in G_k} \sum_{t=1}^T \rho_\tau(e_{it} - v_{1,it}\theta_1 - v_{2,it}\theta_2 - R_{it}). \quad (5.31)$$

Then by Lemma 6(i) and (iii), for any constant  $\varepsilon > 0$ , there exists a constant  $c$  such that

$$\Pr\left(\inf_{\|\theta\| > c.J^{\frac{1}{2}}} \sum_{i \in G_k} \sum_{t=1}^T \rho_\tau(e_{it} - v_{it}\theta - R_{it}) > \sum_{i \in G_k} \sum_{t=1}^T \rho_\tau(e_{it} - R_{it})\right) > 1 - \varepsilon.$$

Then by the fact that  $\tilde{\theta} = (\tilde{\theta}'_1, \tilde{\theta}'_2)$  minimizes the optimization problem 5.30, the above result implies that  $\|\tilde{\theta}\| = \|\hat{\theta} - \theta^0\| = O_p(J^{\frac{1}{2}})$ , which further implies  $\|\tilde{\theta}_1\| = \|\hat{\theta}_1 - \theta_1^0\| = O_p(J^{\frac{1}{2}})$  and  $\|\tilde{\theta}_2\| = \|\hat{\theta}_2 - \theta_2^0\| = O_p(J^{\frac{1}{2}})$ . By the definition of  $\theta_1$ , we have

$$\hat{\theta}_1 - \theta_1^0 = (\mathbf{W}'\mathbf{f}\mathbf{W})^{\frac{1}{2}}(\hat{\gamma} - \gamma^0) + (\mathbf{W}'\mathbf{f}\mathbf{W})^{-\frac{1}{2}}\mathbf{W}'\mathbf{f}\mathbf{X}(\hat{\alpha} - \alpha^0),$$

which implies that

$$\left\|(\mathbf{W}'\mathbf{f}\mathbf{W})^{\frac{1}{2}}(\hat{\gamma} - \gamma^0)\right\| \leq \|\hat{\theta}_1 - \theta_1^0\| + \left\|(\mathbf{W}'\mathbf{f}\mathbf{W})^{-\frac{1}{2}}\mathbf{W}'\mathbf{f}\mathbf{X}(\hat{\alpha} - \alpha^0)\right\|.$$

Notice that by Assumption 1(v) and Lemma 1, the eigenvalues of  $\mathbf{W}'\mathbf{f}\mathbf{W}$  and  $\mathbf{W}(\mathbf{W}'\mathbf{f}\mathbf{W})^{-1}\mathbf{W}'$  are also bounded and bounded away from zero and above, which implies that

$$\left\|\sqrt{N_k T}(\hat{\gamma} - \gamma^0)\right\| = O_p(J^{\frac{1}{2}}),$$

so we have  $\|\hat{\gamma} - \gamma^0\| = O_p((N_k T)^{-1/2} J^{1/2})$ . We next consider the asymptotic distribution of  $\sqrt{N_k T}(\hat{\alpha} - \alpha^0)$ . Let  $\theta_2^* = (\frac{1}{N_k T} \Gamma_{N_k T})^{-1} \sum_{i \in G_k} \sum_{t=1}^T v_{2,it} \psi(e_{it})$ . By Lemma 4 and Lemma 5, we have

$$\theta_2^* \xrightarrow{d} N(0, \Gamma_{G_k, \tau}^{-1} \Omega_{G_k, \tau} \Gamma_{G_k, \tau}^{-1}).$$

To derive the asymptotic distribution of  $\hat{\alpha}$ , notice that by definition, we have  $\tilde{\theta}_2 = \hat{\theta}_2 - \theta_2^0 = \sqrt{N_k T}(\hat{\alpha} - \alpha^0)$ . So we only need to show  $\|\tilde{\theta}_2 - \theta_2^*\| = o_p(1)$ . By definition, we have  $\Pr(\|\theta_2^*\| < M) \rightarrow 1$  for any  $M > 0$  as  $(N, T) \rightarrow \infty$ . In addition, by the previous results, we have  $\|\tilde{\theta}_1\| = O_p(J^{1/2})$ . Then we define

$$\tilde{D}_{it}(\theta_2, \theta_2^*) = \rho_{\tau}(e_{it} - v_{1,it}\tilde{\theta}_1 - v_{2,it}\theta_2 - R_{it}) - \rho_{\tau}(e_{it} - v_{1,it}\tilde{\theta}_1 - v_{2,it}\theta_2^* - R_{it}).$$

Then by Lemma 6(ii), for any  $\varepsilon > 0$ , we have

$$\sup_{\|\theta_2 - \theta_2^*\| \leq \varepsilon} \left| \sum_{i \in G_k} \sum_{t=1}^T \{ \tilde{D}_{it}(\theta_2, \theta_2^*) - E[\tilde{D}_{it}(\theta_2, \theta_2^*)] - v_{2,it}(\theta_2 - \theta_2^*)\psi(e_{it}) \} \right| = o_p(1).$$

On the other hand, we have

$$\begin{aligned} & \sup_{\theta_2 - \theta_2^* \leq \varepsilon} \left| \sum_{i \in G_k} \sum_{t=1}^T \tilde{D}_{it}(\theta_2, \theta_2^*) + \left( \sum_{i \in G_k} \sum_{t=1}^T \psi(e_{it}) v_{2,it} \right) (\theta_2 - \theta_2^*) \right. \\ & \quad \left. - \frac{1}{2N_k T} \theta_2' \Gamma_{N_k T} \theta_2 + \frac{1}{2N_k T} \theta_2^* \Gamma_{N_k T} \theta_2^* \right| \\ &= \sup_{\theta_2 - \theta_2^* \leq \varepsilon} \left| \sum_{i \in G_k} \sum_{t=1}^T \tilde{D}_{it}(\theta_2, \theta_2^*) + \frac{1}{N_k T} (\theta_2 - \theta_2^*)' \Gamma_{N_k T} \theta_2^* \right. \\ & \quad \left. - \frac{1}{2N_k T} \theta_2' \Gamma_{N_k T} \theta_2 + \frac{1}{2N_k T} \theta_2^* \Gamma_{N_k T} \theta_2^* \right| \\ &= \sup_{\theta_2 - \theta_2^* \leq \varepsilon} \left| \sum_{i \in G_k} \sum_{t=1}^T \tilde{D}_{it}(\theta_2, \theta_2^*) + \frac{1}{2N_k T} (\theta_2 - \theta_2^*)' \Gamma_{N_k T} (\theta_2 - \theta_2^*) \right| \\ &= o_p(1), \end{aligned}$$

where the first equality is by the definition of  $\theta_2^*$ , the second equality is by simple algebra and the last equality is by Lemma 6(iv) and the triangular inequality. Notice that as  $(N, T) \rightarrow \infty$ ,  $\frac{1}{2N_k T} (\theta_2 - \theta_2^*)' \Gamma_{N_k T} (\theta_2 - \theta_2^*) > 0$  when  $\|\theta_2 - \theta_2^*\| > \varepsilon$ , by Lemma 4 and Assumption 4, we have

$$\begin{aligned} & \Pr \left( \inf_{\|\theta_2 - \theta_2^*\| > \varepsilon} \sum_{i \in G_k} \sum_{t=1}^T \rho_\tau(e_{it} - v_{1,it} \tilde{\theta}_1 - v_{2,it} \theta_2 - R_{it}) \right. \\ & \quad \left. > \sum_{i \in G_k} \sum_{t=1}^T \rho_\tau(e_{it} - v_{1,it} \tilde{\theta}_1 - v_{2,it} \theta_2^* - R_{it}) \right) \rightarrow 1, \end{aligned}$$

when  $(N, T) \rightarrow \infty$ . Then by the fact that  $\tilde{\theta} = (\tilde{\theta}_1, \tilde{\theta}_2)$  solves the optimization problem 5.31, the above result implies  $\tilde{\theta}_2 = \theta_2^* + o_p(1)$ , and we have

$$\tilde{\theta}_2 = \hat{\theta}_2 - \theta_2^0 = \sqrt{N_k T} (\hat{\alpha} - \alpha^0) \xrightarrow{d} N(0, \Gamma_{G_k, \tau}^{-1} \Omega_{G_k, \tau} \Gamma_{G_k, \tau}^{-1}).$$

By Assumption 4, we have  $\hat{\delta}_{\sigma(k),\tau} = \hat{\delta}_{G_k,\tau}^* + o_p(c_{NT})$ , for all  $k = 1, \dots, K^0$ . This further implies that  $\|\hat{\gamma}_{\sigma(k),\tau} - \hat{\gamma}_{G_k,\tau}^*\| = o_p(c_{NT})$  and  $\|\hat{\alpha}_{\sigma(k),\tau} - \hat{\alpha}_{G_k,\tau}^*\| = o_p(c_{NT})$ . Then for all  $l = 1, \dots, q$  and  $k = 1, \dots, K^0$ , we have

$$\begin{aligned} \left\| \hat{\beta}_{\hat{G}_{\sigma(k)l,\tau}} - \beta_{G_kl,\tau}^0 \right\|_2 &= \left\| \hat{\beta}_{\hat{G}_{\sigma(k)l,\tau}} - P(u)' \hat{\gamma}_{G_kl,\tau}^* + P(u)' \hat{\gamma}_{G_kl,\tau}^* - \beta_{G_kl,\tau}^0 \right\|_2 \\ &\leq \left\| \hat{\beta}_{\hat{G}_{\sigma(k)l,\tau}} - P(u)' \hat{\gamma}_{G_kl,\tau}^* \right\|_2 + \left\| P(u)' \hat{\gamma}_{G_kl,\tau}^* - \beta_{G_kl,\tau}^0 \right\|_2 \\ &\leq o_p(c_{NT}) + \left\| P(u)' \hat{\gamma}_{G_kl,\tau}^* - P(u)' \gamma_{G_kl,\tau}^0 \right\|_2 + \left\| \beta_{G_kl,\tau}^0 - P(u)' \gamma_{G_kl,\tau}^0 \right\|_2 \\ &= O_p((N_k T)^{-1/2} J^{1/2} + J^{-\kappa}), \end{aligned}$$

where the second and third inequalities are by triangular inequality, the last equality is by Assumptions 2(ii) and 4. On the other hand, for  $k = 1, \dots, K^0$ ,

$$\begin{aligned} \sqrt{N_{\sigma(k)} T} (\hat{\alpha}_{\hat{G}_{\sigma(k),\tau}} - \alpha_{G_k,\tau}^*) &= o_p(\sqrt{N_{\sigma(k)} T} c_{NT}) \\ &= o_p(\sqrt{NT} c_{NT}) \\ &= o_p(1), \end{aligned}$$

where the first equality is by the previous result, the second equality is by Assumption 3(i) and the last equality is by Assumption 4. Therefore, we have

$$\begin{aligned} \sqrt{N_{\sigma(k)} T} (\hat{\alpha}_{\hat{G}_{\sigma(k),\tau}} - \alpha_{G_k,\tau}^0) &= \sqrt{N_{\sigma(k)} T} (\hat{\alpha}_{\hat{G}_{\sigma(k),\tau}} - \alpha_{G_k,\tau}^*) + \sqrt{N_{\sigma(k)} T} (\hat{\alpha}_{\hat{G}_{\sigma(k),\tau}} - \alpha_{G_k,\tau}^0) \\ &= o_p(1) + \sqrt{N_{\sigma(k)} T} (\hat{\alpha}_{\hat{G}_{\sigma(k),\tau}} - \alpha_{G_k,\tau}^0) \\ &\xrightarrow{d} N(0, \Gamma_{G_k,\tau}^{-1} \Omega_{G_k,\tau} \Gamma_{G_k,\tau}^{-1}). \end{aligned}$$

The proof of Theorem 2 is thus finished.

**Some Useful Lemmas:**

**Lemma 2.** Under Assumptions 1, 2, 3 and 4, the eigenvalues of

$$\frac{1}{N_k T} \sum_{i \in G_k} \sum_{t=1}^T w_{it} w'_{it} = \frac{1}{N_k T} \mathbf{W}'_{G_k} \mathbf{W}_{G_k}$$

and

$$\frac{1}{T} \sum_{t=1}^T w_{it} w'_{it} = \frac{1}{T} \mathbf{W}'_i \mathbf{W}_i$$

are bounded and bounded away from zero by some generic constants in probability.

**Proof of Lemma 2:** We first show that the eigenvalues of  $E[1\{i \in G_k\}w_{it}w'_{it}]$ , i.e., the expectation of  $w_{it}w'_{it}$  for individuals in the group  $k$ , are bounded and bounded away from zero by some generic constants. Let  $\mathbf{b} = (b'_1, \dots, b'_q)' \in \mathbb{R}^{qJ}$  be an arbitrary vector of nonzero constants, where  $b_i = (b_{i1}, \dots, b_{iJ})' \in \mathbb{R}^J$ . Notice that for all  $i \in G_k$ ,

$$\begin{aligned} \mathbf{b}' E[w_{it}w'_{it}] \mathbf{b} &= E[(b'_1 P(u_{it}), \dots, b'_q P(u_{it}))' (z_{it} z'_{it}) (b'_1 P(u_{it}), \dots, b'_q P(u_{it}))] \\ &= E[(b'_1 P(u_{it}), \dots, b'_q P(u_{it}))' E[z_{it} z'_{it} | u_{it}] (b'_1 P(u_{it}), \dots, b'_q P(u_{it}))] \\ &\leq E[\lambda_{\max}(E[z_{it} z'_{it} | u_{it}]) \|(b'_1 P(u_{it}), \dots, b'_q P(u_{it}))\|^2] \\ &\leq c_1 \cdot \sum_{j=1}^q b'_j E[P(u_{it}) P(u_{it})'] b_j \\ &\leq c_2 \cdot \lambda_{\max}(E[P(u_{it}) P(u_{it})']) \|\mathbf{b}\|^2 \\ &\leq c_3 \|\mathbf{b}\|^2, \end{aligned}$$

where the first equality is by definition, the second equality is by law of iterative expectation, the third inequality is by the basic inequality  $c' A c \leq \lambda_{\max}(A) \|c\|^2$ , where  $A$  is some  $K \times K$  matrix and  $c$  is a  $K \times 1$  constant vector and  $K \geq 1$ , the fourth inequality is by Assumption 1(iii) and the sixth inequality is by Assumption 2(i). The above calculation implies that the eigenvalues of  $E[w_{it}w'_{it}]$  are bounded from above by some generic constant and we can similarly show that its eigenvalues are also bounded away from zero.



Next, we show for  $i \in G_k$ , we have

$$\frac{1}{N_k T} \sum_{i \in G_k} \sum_{t=1}^T w_{it} w'_{it} \xrightarrow{p} E[1\{i \in G_k\} w_{it} w'_{it}]. \quad (5.32)$$

Define  $S_{it}^{J_1, J_2, l_1, l_2} = P_{J_1}(u_{it}) P_{J_2}(u_{it}) z_{it, l_1} z_{it, l_2}$ , for  $1 \leq J_1, J_2 \leq J$  and  $1 \leq l_1, l_2 \leq q$ . Notice that  $S_{it}^{J_1, J_2, l_1, l_2}$  are just entries of the  $qJ \times qJ$  matrix  $w_{it} w'_{it}$ . Therefore, to show equation 5.32, we only need to show  $\frac{1}{N_k T} \sum_{i \in G_k} \sum_{t=1}^T S_{it}^{J_1, J_2, l_1, l_2} \xrightarrow{p} E[S_{it}^{J_1, J_2, l_1, l_2}]$ . For the sake of notational simplicity, we now suppress the subscripts  $J_1, J_2, l_1, l_2$  when writing  $S_{it}$ . First notice that

$$\begin{aligned} \text{Var}(S_{it}) &= E[S_{it}^2] - (E[S_{it}])^2 \\ &\leq E[S_{it}^2] \\ &= E[P_{J_1}^2(u_{it}) P_{J_2}^2(u_{it}) z_{it, l_1}^2 z_{it, l_2}^2] \\ &\leq c \cdot (\sup_u \|P(u_{it})\|)^2 E[P_{J_1}(u_{it}) P_{J_2}(u_{it})] \\ &\leq c \cdot \zeta_0^2(J). \end{aligned}$$

The fourth inequality is by Assumption 2(i) and the last inequality is by the fact that  $E[P_{J_1}(u_{it}) P_{J_2}(u_{it})] = e'_{J_1} E[P(u_{it}) P(u_{it})'] e_{J_2} \leq \lambda_{\max}(E[P(u_{it}) P(u_{it})'])$ , where  $e_{J_i}$  is a  $J \times 1$  vector in which the  $J_i$ -th entry is 1 and the rest entries are zero. Then by Chebyshev's inequality, for any  $\epsilon > 0$ , we have

$$\mathbb{P}\left(\left|\frac{1}{N_k T} \sum_{i=1}^{N_k} \sum_{t=1}^T S_{it} - E[S_{it}]\right| > \epsilon\right) \leq O\left(\frac{\zeta_0^2(J)}{N_k T \epsilon^2}\right) = o(1),$$

by Assumption 4. Combining the above results together, we thus have proved the first part of Lemma 2. For  $\frac{1}{T} \sum_{i \in G_k} \sum_{t=1}^T w_{it} w'_{it}$ , the proof is exactly the same as before, so we omit the details to save space.

**Lemma 3.** Let

$$\mathbf{M} = (M(z_{11}, u_{11}), \dots, M(z_{1T}, u_{1T}), \dots, M(z_{N_k 1}, u_{N_k 1}), \dots, M(z_{N_k T}, u_{N_k T}))' \in \mathbb{R}^{N_k T \times p},$$

where  $M(z_{it}, u_{it})$  is defined in the main text, and recall  $\tilde{\mathbf{W}} = \mathbf{W}(\mathbf{W}'\mathbf{f}\mathbf{W})^{-1}\mathbf{W}'\mathbf{f}$ . Under Assumptions 1, 2, 3 and 4, we have

$$(i) \quad \|\mathbf{W}\| = O_p(\sqrt{N_k T J})$$

$$(ii) \quad \|\tilde{\mathbf{W}}\mathbf{X} - \tilde{\mathbf{W}}\mathbf{M}\| = O_p(J)$$

**Proof of Lemma 3.** We first show  $\|\mathbf{W}\| = O_p(\sqrt{N_k T J})$ . Let's consider the norm of each row in  $\mathbf{W}$ . Notice that

$$\begin{aligned} \text{Var}(\|w_{it}\|) &= E\|w_{it}\|^2 - (E\|w_{it}\|)^2 \\ &\leq E\|w_{it}\|^2 \\ &= E[w'_{it}w_{it}] \\ &= E[\text{tr}(w_{it}w'_{it})] \\ &= \text{tr}(E[w_{it}w'_{it}]) \\ &= O_p(J), \end{aligned}$$

where the first to the fifth inequalities are by definition and simple algebra, and the last equality is by Lemma 2. Therefore, there exists a constant  $c$  such that  $E\|w_{it}\| \leq c \cdot J$ . Then by Markov's inequality, for any  $\epsilon > 0$ , there exists  $c_0 = \sqrt{\frac{c}{\epsilon}J}$  such that

$$P(\|w_{it}\| > c_0) \leq \frac{E\|w_{it}\|^2}{c_0^2} = \epsilon,$$

which implies that  $\|w_{it}\| = O_p(\sqrt{J})$ . Finally, by Assumption 1(i), we have  $\|\mathbf{W}\| = O_p(\sqrt{N_k T J})$ .

Next, we show  $\|\tilde{\mathbf{W}}\mathbf{X} - \tilde{\mathbf{W}}\mathbf{M}\| = O_p(J)$ . First, notice that

$$\|\tilde{\mathbf{W}}\mathbf{X} - \tilde{\mathbf{W}}\mathbf{M}\| = \|\mathbf{W}(\mathbf{W}'\mathbf{f}\mathbf{W})^{-1}\mathbf{W}'\mathbf{f}(\mathbf{X} - \mathbf{M})\|.$$

We thus consider the convergence rates of  $\|\mathbf{W}\|$ ,  $\|(\mathbf{W}'\mathbf{f}\mathbf{W})^{-1}\|$ , and  $\|\mathbf{W}'\mathbf{f}(\mathbf{X} - \mathbf{M})\|$ . The convergence rate of  $\|\mathbf{W}\|$  has been derived above, which is  $O_p(J)$ . In addition, by Lemma 1, we have  $\|(\mathbf{W}'\mathbf{f}\mathbf{W})^{-1}\| = O_p(\frac{1}{N_k T})$ . Then for  $\|\mathbf{W}'\mathbf{f}(\mathbf{X} - \mathbf{M})\|$ , consider one arbitrary row in  $\mathbf{W}'\mathbf{f}(\mathbf{X} - \mathbf{M})$ , which is  $f_{it}(0)w_{it}(x_{it} - m(z_{it}, u_{it}))'$ . Notice that

$$\begin{aligned} \text{Var}(\|f_{it}(0)w_{it}(x_{it} - m_{z_{it}, u_{it}})'\|) &\leq E[\|f_{it}(0)w_{it}(x_{it} - m(z_{it}, u_{it}))'\|^2] \\ &\leq c \cdot E\|w_{it}\|^2 \\ &= O_p(J), \end{aligned}$$

where the second inequality is by Assumption 1(ii) and Assumption 1(iv), and the last inequality is by the rate of convergence of  $E\|w_{it}\|^2$  which is derived above. Similarly as before, by Markov inequality and Assumption 1(i), we have  $\|\mathbf{W}'\mathbf{f}(\mathbf{X} - \mathbf{M})\| = O_p(\sqrt{N_k T J})$ .

Finally, we have

$$\begin{aligned} \|\tilde{\mathbf{W}}\mathbf{X} - \tilde{\mathbf{W}}\mathbf{M}\| &\leq \|\mathbf{W}\| \|(\mathbf{W}'\mathbf{f}\mathbf{W})^{-1}\| \|\mathbf{W}'\mathbf{f}(\mathbf{X} - \mathbf{M})\| \\ &= O_p\left(\frac{1}{N_k T} \sqrt{N_k T J} \sqrt{N_k T J}\right) \\ &= O_p(J). \end{aligned}$$

**Lemma 4.** Define

$$\Gamma_{N_k T} = (\mathbf{X} - \tilde{\mathbf{W}}\mathbf{X})'\mathbf{f}(\mathbf{X} - \tilde{\mathbf{W}}\mathbf{X}),$$

and assume Assumptions 1,2, 3 and 4 hold, we have

$$\frac{1}{N_k T} \Gamma_{N_k T} \xrightarrow{p} \Gamma,$$

where  $\Gamma = E \left[ f_{it}(0)(x_{it} - M(z_{it}, u_{it}))(x_{it} - M(z_{it}, u_{it}))' \right]$  for all  $i \in G_k$ .

**Proof of Lemma 4:** Define  $\tilde{\Gamma} = (\mathbf{X} - \mathbf{M})' \mathbf{f}(\mathbf{X} - \mathbf{M})$ . To show  $\frac{1}{N_k T} \Gamma_{N_k T}$  converges in probability to  $\Gamma$ , we first show that  $\left\| \frac{1}{N_k T} \Gamma_{N_k T} - \frac{1}{N_k T} \tilde{\Gamma} \right\| = o_p(1)$  and later show  $\frac{1}{N_k T} \tilde{\Gamma} \xrightarrow{p} \Gamma$ .

By some simple algebra, we have

$$\begin{aligned}
\Gamma_{N_k T} - \tilde{\Gamma} &= (\mathbf{X} - \tilde{\mathbf{W}}\mathbf{X})' \mathbf{f}(\mathbf{X} - \tilde{\mathbf{W}}\mathbf{X}) - (\mathbf{X} - \mathbf{M})' \mathbf{f}(\mathbf{X} - \mathbf{M}) \\
&= (\mathbf{X} - \mathbf{M} + \mathbf{M} - \tilde{\mathbf{W}}\mathbf{X})' \mathbf{f}(\mathbf{X} - \mathbf{M} + \mathbf{M} - \tilde{\mathbf{W}}\mathbf{X}) - (\mathbf{X} - \mathbf{M})' \mathbf{f}(\mathbf{X} - \mathbf{M}) \\
&= (\mathbf{X} - \mathbf{M})' \mathbf{f}(\mathbf{M} - \tilde{\mathbf{W}}\mathbf{X}) + (\mathbf{M} - \tilde{\mathbf{W}}\mathbf{X})' \mathbf{f}(\mathbf{X} - \mathbf{M}) + (\mathbf{M} - \tilde{\mathbf{W}}\mathbf{X})' \mathbf{f}(\mathbf{M} - \tilde{\mathbf{W}}\mathbf{X}) \\
&= (\mathbf{X} - \mathbf{M})' \mathbf{f}(\mathbf{M} - \tilde{\mathbf{W}}\mathbf{M}) + (\mathbf{X} - \mathbf{M})' \mathbf{f}(\tilde{\mathbf{W}}\mathbf{M} - \tilde{\mathbf{W}}\mathbf{X}) + (\mathbf{M} - \tilde{\mathbf{W}}\mathbf{M})' \mathbf{f}(\mathbf{X} - \mathbf{M}) \\
&\quad + (\tilde{\mathbf{W}}\mathbf{M} - \tilde{\mathbf{W}}\mathbf{X})' \mathbf{f}(\mathbf{X} - \mathbf{M}) + (\mathbf{M} - \tilde{\mathbf{W}}\mathbf{M})' \mathbf{f}(\mathbf{M} - \tilde{\mathbf{W}}\mathbf{M}) \\
&\quad + (\mathbf{M} - \tilde{\mathbf{W}}\mathbf{M})' \mathbf{f}(\tilde{\mathbf{W}}\mathbf{M} - \tilde{\mathbf{W}}\mathbf{X}) + (\tilde{\mathbf{W}}\mathbf{M} - \tilde{\mathbf{W}}\mathbf{X})' \mathbf{f}(\mathbf{M} - \tilde{\mathbf{W}}\mathbf{M}) \\
&\quad + (\tilde{\mathbf{W}}\mathbf{M} - \tilde{\mathbf{W}}\mathbf{X})' \mathbf{f}(\tilde{\mathbf{W}}\mathbf{M} - \tilde{\mathbf{W}}\mathbf{X}).
\end{aligned}$$

First,  $\|\mathbf{X} - \mathbf{M}\| = O_p((N_k T)^{\frac{1}{2}})$  by Assumption 1. Second, we have  $\|\mathbf{M} - \tilde{\mathbf{W}}\mathbf{M}\| = O_p((N_k T)^{\frac{1}{2}} J^{-\kappa})$  because of Assumption 4. Third, by Lemma 3, we have  $\|\tilde{\mathbf{W}}\mathbf{M} - \tilde{\mathbf{W}}\mathbf{X}\| = O_p(J)$ . Then, by Assumption 4(iii), we have

$$\begin{aligned}
\frac{1}{N_k T} \Gamma_{N_k T} - \frac{1}{N_k T} \tilde{\Gamma} &= O_p(J^{-\kappa}) + O_p((N_k T)^{-\frac{1}{2}} J) + O_p(J^{-\kappa}) \\
&\quad + O_p(J^{-\kappa}) + O_p(J^{-2\kappa}) + O_p((N_k T)^{-\frac{1}{2}} J^{1-\kappa}) \\
&\quad + O_p((N_k T)^{-\frac{1}{2}} J^{1-\kappa}) + O_p(J^2) \\
&= o_p(1)
\end{aligned}$$

So, this implies that  $\frac{1}{N_k T} \Gamma_{N_k T} = \frac{1}{N_k T} \tilde{\Gamma} + o_p(1)$ . Furthermore, notice that  $\frac{1}{N_k T} \tilde{\Gamma} \xrightarrow{p} \Gamma$  by the common law of large numbers. So, we have  $\frac{1}{N_k T} \Gamma_{N_k T} \xrightarrow{p} \Gamma$ .

**Lemma 5.** Recall that  $\psi(u) = \tau - 1\{u \leq 0\}$ . Under Assumptions 1, 2, 3 and 4, let

$\tilde{\theta}_2 = \sum_{i \in G_k} \sum_{t=1}^T v_{2,it} \psi(e_{it})$ , we have

$$\tilde{\theta}_2 \xrightarrow{d} N(0, \Omega_{G_k, \tau}).$$

**Proof of Lemma 5:** Define  $\psi(\mathbf{e}) = (\psi(e_{11}), \dots, \psi(e_{1T}), \dots, \psi(e_{N_k 1}), \dots, \psi(e_{N_k T}))$ . By definition, we have

$$\begin{aligned} \tilde{\theta}_2 &= \sum_{i \in G_k} \sum_{t=1}^T v_{2,it} \psi(e_{it}) \\ &= \frac{1}{\sqrt{N_k T}} (\mathbf{X} - \tilde{\mathbf{W}}\mathbf{X})' \psi(\mathbf{e}) \\ &= \frac{1}{\sqrt{N_k T}} (\mathbf{X} - \mathbf{M} + \mathbf{M} - \tilde{\mathbf{W}}\mathbf{M} + \tilde{\mathbf{W}}\mathbf{M} - \tilde{\mathbf{W}}\mathbf{X})' \psi(\mathbf{e}) \\ &= \frac{1}{\sqrt{N_k T}} (\mathbf{X} - \mathbf{M})' \psi(\mathbf{e}) + \frac{1}{\sqrt{N_k T}} (\mathbf{M} - \tilde{\mathbf{W}}\mathbf{M})' \psi(\mathbf{e}) + \frac{1}{\sqrt{N_k T}} (\tilde{\mathbf{W}}\mathbf{M} - \tilde{\mathbf{W}}\mathbf{X})' \psi(\mathbf{e}) \end{aligned}$$

First, notice that the first term  $\frac{1}{\sqrt{N_k T}} (\mathbf{X} - \mathbf{M})' \psi(\mathbf{e}) \xrightarrow{d} N(0, \Omega_{G_k, \tau})$  by the usual central limit theorem under Assumptions 1, 2, 3 and 4. The second term is  $o_p(1)$  because  $\frac{1}{\sqrt{N_k T}} \|\mathbf{M} - \tilde{\mathbf{W}}\mathbf{M}\| = O_p((N_k T)^{\frac{1}{2}} J^{-\kappa})$  and Assumption 4. Third, the third term is also  $o_p(1)$  because  $\|\tilde{\mathbf{W}}\mathbf{M} - \tilde{\mathbf{W}}\mathbf{X}\| = O_p(J)$  by Lemma 3. Therefore,

$$\tilde{\theta}_2 = \frac{1}{\sqrt{N_k T}} (\mathbf{X} - \mathbf{M})' \psi(\mathbf{e}) + o_p(1),$$

which implies  $\tilde{\theta}_2 \xrightarrow{d} N(0, \Omega_{G_k, \tau})$  by the Slutsky's theorem.

**Lemma 6.** Let  $C_{it} = \rho_\tau(e_{it} - v_{it}\theta - R_{it}) - \rho_\tau(e_{it} - R_{it}) + v_{it}\theta\psi(e_{it})$  and  $D_{it}(\theta_1, \theta_2) = \rho_\tau(e_{it} - v_{1,it}\theta_1 - v_{2,it}\theta_2 - R_{it}) - \rho_\tau(e_{it} - v_{1,it}\theta_1 - R_{it}) + v_{2,it}\theta_2\psi(e_{it})$ . Under Assumptions 1, 2, 3, 4 and  $(N, T) \rightarrow \infty$ , we have

(i)

$$\sup_{\|\theta\| \leq c \cdot J^{\frac{1}{2}}} J^{-1} \left| \sum_{i \in G_k} \sum_{t=1}^T \{C_{it} - E[C_{it}]\} \right| = o_p(1).$$

(ii)

$$\sup_{\|\theta\|_1 \leq c_1 \cdot J^{\frac{1}{2}}, \|\theta\|_2 \leq c_2} \left| \sum_{i \in G_k} \sum_{t=1}^T \{D_{it}(\theta_1, \theta_2) - E[D_{it}(\theta_1, \theta_2)]\} \right| = o_p(1),$$

(iii)

$$\Pr\left( \inf_{\|\theta\|=c \cdot J^{\frac{1}{2}}} \left| J^{-1} \sum_{i \in G_k} \sum_{t=1}^T \left\{ E[C_{it}(\theta)] - v_{it}\theta\psi(e_{it}) \right\} \right| > 1 \right) \rightarrow 1,$$

(iv)

$$\sup_{\|\theta\|_1 \leq c_1 \cdot J^{\frac{1}{2}}, \|\theta\|_2 \leq c_2} \left| \sum_{i \in G_k} \sum_{t=1}^T E[D_{it}(\theta_1, \theta_2)] - \frac{1}{2N_k T} \theta_2' \Gamma_{N_k T} \theta_2 \right| = o_p(1).$$

where  $c, c_1, c_2$  are any positive constants.

**Lemma 7.** Let  $C_{it} = \rho_\tau(e_{it} - v_{it}\theta - R_{it}) - \rho_\tau(e_{it} - R_{it}) + v_{it}\theta\psi(e_{it})$  and  $D_{it}(\theta_1, \theta_2) = \rho_\tau(e_{it} - v_{1,it}\theta_1 - v_{2,it}\theta_2 - R_{it}) - \rho_\tau(e_{it} - v_{1,it}\theta_1 - R_{it}) + v_{2,it}\theta_2\psi(e_{it})$ . Under Assumptions 1, 2 and 3 and  $T \rightarrow \infty$ , for all  $1 \leq i \leq N$ ,

(i)

$$\sup_{\|\theta\| \leq c \cdot J^{\frac{1}{2}}} J^{-1} \left| \sum_{t=1}^T \{C_{it} - E[C_{it}]\} \right| = o_p(1).$$

(ii)

$$\sup_{\|\theta\|_1 \leq c_1 \cdot J^{\frac{1}{2}}, \|\theta\|_2 \leq c_2} \left| \sum_{t=1}^T \{D_{it}(\theta_1, \theta_2) - E[D_{it}(\theta_1, \theta_2)]\} \right| = o_p(1),$$

(iii)

$$\Pr\left( \inf_{\|\theta\|=c \cdot J^{\frac{1}{2}}} \left| J^{-1} \sum_{t=1}^T \left\{ E[C_{it}(\theta)] - v_{it}\theta\psi(e_{it}) \right\} \right| > 1 \right) \rightarrow 1,$$

where  $c, c_1, c_2$  are any positive constants.

**Proof of Lemma 6 and 7:** Lemma 6 and 7 can be proved following the same argument in Wei et al. (2006), so we omit the details here.

## Proofs of the Main Results in Chapter 4

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

**Theorem 1.** *Suppose Assumption 1, 2 hold, then*

$$(i) \quad \|\hat{\theta}_i - \theta_i^0\|_F = O_p(J^{-r} + J^{\frac{1}{2}}T^{-\frac{1}{2}} + \lambda) \text{ and } \|\hat{\theta}_{i,j} - \theta_{i,j}^0\|_F = O_p(J^{-r} + J^{\frac{1}{2}}T^{-\frac{1}{2}} + \lambda) \text{ for } i = 1, 2, \dots, N, j = 1, \dots, p.$$

$$(ii) \quad \frac{1}{N} \sum_{i=1}^N \|\hat{\theta}_i - \theta_i^0\|_F^2 = O_p(J^{-2r} + JT^{-1}) \text{ and } \frac{1}{N} \sum_{i=1}^N \|\hat{\theta}_{i,j} - \theta_{i,j}^0\|_F^2 = O_p(J^{-2r} + JT^{-1}) \text{ for } j = 1, \dots, p.$$

$$(iii) \quad \|\hat{\eta}_{(k),j} - \eta_{k,j}^0\|_F = O_p(J^{-r} + J^{\frac{1}{2}}T^{-\frac{1}{2}}), \text{ for } k = 1, \dots, K_j^0, j = 1, \dots, p, \text{ where } (\hat{\eta}_{(1),j}, \dots, \hat{\eta}_{(K_j^0),j}) \text{ is a suitable permutation of } (\hat{\eta}_{1,j}, \dots, \hat{\eta}_{K_j^0,j}) \text{ for } j = 1, \dots, p.$$

*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 4,  $\mu_{\min}(\hat{Q}_{i,\bar{z}\bar{z}}) > \underline{c} > 0$  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 4,  $\|\hat{Q}_{i,\bar{z}\bar{e}}\| = O_p(J^{-r} + J^{\frac{1}{2}} T^{-\frac{1}{2}})$ , 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  $c_{1i,NT}(\hat{\theta}, \theta^0, \hat{\eta}) = O_p(1)$ .

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 1(iii), we could get that  $\gamma^0$  and  $\pi^0$  lie in the interior of  $\mathcal{R}_b$  and  $\Pi_b$  respectively, thus  $\theta^0 \in \Theta_b$  and  $\eta^0 \in \mathcal{H}_b$ .

Then we search over  $\Theta_b$  and  $\mathcal{H}_b$  to minimize the objective function 4.7, 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  $\varepsilon$ , 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.

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

where the last inequality holds w.p.a 1 by Lemma 4.

By Lemma 4,  $\frac{1}{N} \sum_{i=1}^N \left\| \hat{Q}_{i, \tilde{z}\tilde{e}} \right\|^2 = O_p(J^{-2r} + JT^{-1})$ , then  $\frac{m_{JT}^{-2}}{N} \sum_{i=1}^N \left\| \hat{Q}_{i, \tilde{z}\tilde{e}} \right\|^2 = O_p(1)$ , 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.

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

(iii) Further consider  $c_{1ji,NT}(\hat{\theta}, \theta^0, \eta)$ , where  $\hat{\theta}$  and  $\eta$  lie in the interior of  $\Theta_b$  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 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  $\|\hat{\eta}_{k,j} - \eta_{k,j}^0\| = O_p(J^{-r} + J^{\frac{1}{2}} T^{-\frac{1}{2}})$ .

□

**Theorem 2.** *Suppose Assumption 1, 2 and 3 hold, then*

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

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

*Proof.* (i) For any  $i \in G_{k,j}^0$  and  $l \neq k$ , by Theorem 1,  $\|\hat{\theta}_{i,j} - \hat{\eta}_{l,j}\| \xrightarrow{p} \|\eta_{k,j}^0 - \eta_{l,j}^0\| \neq 0$ . Suppose that  $\|\hat{\theta}_{i,j} - \hat{\eta}_{k,j}\| \neq 0$  for some  $i \in G_{k,j}^0$ , which means that  $i \notin \hat{G}_{k,j}$ , then the first order condition with respect to  $\theta_{i,j}$  is

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

where  $\hat{e}_{im,j} = \frac{\hat{\theta}_{i,j} - \hat{\eta}_{m,j}}{\|\hat{\theta}_{i,j} - \hat{\eta}_{m,j}\|}$  if  $\|\hat{\theta}_{i,j} - \hat{\eta}_{m,j}\| \neq 0$  and  $\hat{e}_{im,j} \leq 1$  otherwise.

From the proof of Theorem 1, I have that

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

Let  $\mu_{1,JT} = (J^{-r} + J^{\frac{1}{2}} T^{-\frac{1}{2}} (\ln T)^3 + \lambda) (\ln T)^v$  and  $\mu_{2,JT} = (J^{-r} + J^{\frac{1}{2}} T^{-\frac{1}{2}} (\ln T)^3) (\ln T)^v$  for some  $v > 0$ . By Lemma 4, 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  $\hat{c}_{ik,j} = \prod_{l=1, l \neq k}^{K_j^0} \|\hat{\theta}_{i,j} - \hat{\eta}_{l,j}\|$ , then

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

Similarly let  $c_{ik,j}^0 = \prod_{l=1, l \neq k}^{K_j^0} \|\theta_{i,j}^0 - \eta_{l,j}^0\|$ . Define  $\bar{c}_{k,j}^0 = \max_{i \in G_{k,j}^0} c_{ik,j}^0$  and  $\underline{c}_{k,j}^0 = \min_{i \in G_{k,j}^0} c_{ik,j}^0$ .

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

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

Define

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

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

Let  $\phi_{ik,j} = \frac{\hat{\theta}_{i,j} - \hat{\eta}_{k,j}}{\|\hat{\theta}_{i,j} - \hat{\eta}_{k,j}\|}$ . Conditional on  $\Xi_{kNT,j}$ , we have that uniformly in  $i \in G_{k,j}^0$ , with

probability  $1 - o(N^{-1})$ ,

$$\begin{aligned}
|\phi'_{ik,j} \hat{A}_{i2,j}| &\geq 2\bar{c} \|\hat{\theta}_{i,j} - \hat{\eta}_{k,j}\| + \lambda \hat{c}_{ik,j} \geq \lambda \frac{\mathcal{C}_{k,j}^0}{2} \\
|\phi'_{ik,j} \hat{A}_{i3,j}| &\leq 2 \|\hat{Q}_{i,\bar{z}\bar{\delta},j}\| \leq 2C\theta_{NT} \\
|\phi'_{ik,j} \hat{A}_{i4,j}| &\leq 2\bar{c} \|\hat{\eta}_{k,j} - \eta_{k,j}^0\| \leq 2\bar{c}c \left( J^{-r} + J^{\frac{1}{2}} T^{-\frac{1}{2}} (\ln T)^3 \right) (\ln T)^v \\
|\phi'_{ik,j} \hat{A}_{i5,j}| &\leq \max_{i \in G_{k,j}^0} \|\hat{A}_{i5,j}\| \leq C\lambda\mu_{1,JT} \\
|\phi'_{ik,j} \hat{A}_{i6,j}| &\leq C\mu_{1,JT}
\end{aligned}$$

Then

$$\begin{aligned}
&\left| \phi'_{ik,j} (\hat{A}_{i2,j} + \hat{A}_{i3,j} + \hat{A}_{i4,j} + \hat{A}_{i5,j} + \hat{A}_{i6,j}) \right| \\
&\geq \phi'_{ik,j} \hat{A}_{i2,j} - \left| \phi'_{ik,j} (\hat{A}_{i3,j} + \hat{A}_{i4,j} + \hat{A}_{i5,j} + \hat{A}_{i6,j}) \right| \\
&\geq \lambda \frac{\mathcal{C}_{k,j}^0}{2} - \left[ 2C\theta_{NT} + 2\bar{c}c \left( J^{-r} + J^{\frac{1}{2}} T^{-\frac{1}{2}} (\ln T)^3 \right) (\ln T)^v + C\lambda\mu_{1,JT} + C\mu_{1,JT} \right] \\
&\geq \lambda \frac{\mathcal{C}_{k,j}^0}{4}
\end{aligned}$$

where we use Assumption 2 and 3.

Thus

$$\begin{aligned}
&P(\hat{\mathbf{E}}_{ik,j}) \\
&= P(i \notin \hat{G}_{k,j} | i \in G_{k,j}^0) \\
&= P(-\hat{A}_{i1,j} = \hat{A}_{i2,j} + \hat{A}_{i3,j} + \hat{A}_{i4,j} + \hat{A}_{i5,j} + \hat{A}_{i6,j}) \\
&\leq P\left( \left| \phi'_{ik,j} \hat{A}_{i1,j} \right| \geq \left| \phi'_{ik,j} (\hat{A}_{i2,j} + \hat{A}_{i3,j} + \hat{A}_{i4,j} + \hat{A}_{i5,j} + \hat{A}_{i6,j}) \right| \right) \\
&\leq P\left( \left| \hat{A}_{i1,j} \right| \geq \lambda \frac{\mathcal{C}_{k,j}^0}{4}, \Xi_{kNT,j} \right) + P(\Xi_{kNT,j}^c) \\
&= o(N^{-1})
\end{aligned}$$



Thus, with probability  $1 - o(N^{-1})$  such that  $\|\theta_{i,j} - \eta_{k,j}\|$  is not differentiable with respect to  $\theta_{i,j}$  for some  $i \in G_{k,j}^0$ , which means that  $P(\|\theta_{i,j} - \eta_{k,j}\| = 0 | i \in G_{k,j}^0) = 1 - o(N^{-1})$ .

Then

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

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

(ii) The proof is similar to Su et al. (2016) Theorem 2.2 (ii) and thus omitted. □

**Theorem 3.** *Suppose Assumption 1, 2, 3, 4 and 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

$$V_{k,j,B} = B^J(x)' \left( \hat{Q}_{G_{k,j}^0} \right)^{-1} \frac{1}{N_{k,j}} \sum_{i \in G_{k,j}^0} \frac{1}{T} W_i' \Sigma_i^{\frac{1}{2}} V_i \Sigma_i^{\frac{1}{2}} W_i \left( \hat{Q}_{G_{k,j}^0} \right)^{-1} B^J(x)$$

*Proof.* The proof of Theorem 3 is similar to the one in Su et al. (2019) and thus is omitted.  $\square$

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

**Lemma 2.** *Let  $\xi_{it}$  be a  $\mathcal{R}^{d_\xi}$  random variable and  $\mathbf{E}[\xi_{it}] = 0$  for all  $i, t$ . For each  $i = 1, \dots, N$ ,  $\xi_{it}$  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  $0 < c_\alpha < \infty$ ,  $0 < \rho < \infty$ .  $\xi_{it}$  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 et al. (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  $\lambda$  satisfies that  $T^{-\frac{1}{2}} (\ln T)^3 = o(\lambda)$ . For convenience, we could choose  $\lambda = T^{-\frac{1}{2}} (\ln T)^{3+v}$  for some  $v > 0$ .  $\square$

**Lemma 3.** *Let  $\xi_{it}$  be a  $\mathcal{R}^{d_\xi}$  random variable and  $\mathbf{E}[\xi_{it}] = 0$  for all  $i, t$ . For each  $i = 1, \dots, N$ ,  $\xi_{it}$  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  $0 < c_\alpha < \infty$ ,  $0 < \rho < \infty$ .  $\xi_{it}$  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  $\tau_\xi$  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  $\mathbf{E}[\xi_{it}] = 0$ . we prove the lemma by showing that for any  $c > 0$

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

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

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

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 4.** *Suppose that Assumption 1 and 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 3, 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 3,

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

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  $\|\hat{Q}_{i,\tilde{z}\tilde{\delta}}\| \leq \|\hat{Q}_{i,\tilde{z}\tilde{\delta}}\| + \|\hat{Q}_{i,\tilde{z}\tilde{u}}\|$ .

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,

$$\mathbf{E} \left[ \|\hat{Q}_{i,z\delta}\|^2 \right] = \sum_{r=1}^p \mathbf{E} \left[ \left\| \frac{1}{T} \sum_{t=1}^T \sqrt{J} B_{rit}^J \delta_{h_i,it} \right\|^2 \right]$$

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

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

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

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

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



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  $1 \leq r \leq p$ ,

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

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  $1 \leq j \leq J$ ,

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

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, 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.

□

# Bibliography

- Abraham, Christophe, Pierre-André Cornillon, ERIC Matzner-Løber, and Nicolas Molinari (2003) “Unsupervised curve clustering using B-splines,” *Scandinavian journal of statistics*, Vol. 30, No. 3, pp. 581–595.
- Ai, Chunrong and Xiaohong Chen (2003) “Efficient estimation of models with conditional moment restrictions containing unknown functions,” *Econometrica*, Vol. 71, No. 6, pp. 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, pp. 451–478.
- Ando, Tomohiro and Jushan Bai (2014) “Asset pricing with a general multifactor structure,” *Journal of Financial Econometrics*, Vol. 13, No. 3, pp. 556–604.
- (2016) “Panel data models with grouped factor structure under unknown group membership,” *Journal of Applied Econometrics*, Vol. 31, No. 1, pp. 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*, Vol. 112, No. 519, pp. 1182–1198.
- Arellano, Manuel and Stéphane Bonhomme (2012) “Identifying distributional characteristics in random coefficients panel data models,” *The Review of Economic Studies*, Vol. 79, No. 3, pp. 987–1020.

- Arellano, Manuel and Bo Honoré (2001) “Panel data models: some recent developments,” in *Handbook of econometrics*, Vol. 5: Elsevier, pp. 3229–3296.
- Atkin, David and Dave Donaldson (2015) “Who’s getting globalized? The size and implications of intra-national trade costs,” Technical report, 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*, Vol. 132, No. 2, pp. 551–615.
- Bai, Jushan (2009) “Panel data models with interactive fixed effects,” *Econometrica*, Vol. 77, No. 4, pp. 1229–1279.
- Baker, Malcolm and Jeffrey Wurgler (2006) “Investor sentiment and the cross-section of stock returns,” *The journal of Finance*, Vol. 61, No. 4, pp. 1645–1680.
- Balasubramanyam, Venkataraman N, Mohammed Salisu, and David Sapsford (1996) “Foreign direct investment and growth in EP and IS countries,” *The economic journal*, Vol. 106, No. 434, pp. 92–105.
- Baltagi, Badi H, Georges Bresson, and Alain Pirotte (2008) “To pool or not to pool?” in *The econometrics of panel data*: Springer, pp. 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*, Vol. 82, No. 1, pp. 117–126.
- Baltagi, Badi H and Dan Levin (1992) “Cigarette taxation: Raising revenues and reducing consumption,” *Structural Change and Economic Dynamics*, Vol. 3, No. 2, pp. 321–335.
- Baltagi, Badi H, Dong Li et al. (2002) “Series estimation of partially linear panel data models with fixed effects,” *Annals of economics and finance*, Vol. 3, No. 1, pp. 103–116.



- Baltagi, Badi Hani (2015) *The Oxford handbook of panel data*: Oxford Handbooks.
- Banerjee, Sudipto, Bradley P Carlin, and Alan E Gelfand (2014) *Hierarchical modeling and analysis for spatial data*: CRC press.
- Belloni, Alexandre, Victor Chernozhukov et al. (2013) “Least squares after model selection in high-dimensional sparse models,” *Bernoulli*, Vol. 19, No. 2, pp. 521–547.
- Belloni, Alexandre, Victor Chernozhukov, Denis Chetverikov, Christian Hansen, and Kengo Kato (2018) “High-dimensional econometrics and regularized GMM,” *arXiv preprint arXiv:1806.01888*.
- Bester, C Alan and Christian B Hansen (2016) “Grouped effects estimators in fixed effects models,” *Journal of Econometrics*, Vol. 190, No. 1, pp. 197–208.
- Blume, Lawrence E, William A Brock, Steven N Durlauf, and Rajshri Jayaraman (2015) “Linear social interactions models,” *Journal of Political Economy*, Vol. 123, No. 2, pp. 444–496.
- Bonhomme, Stéphane, Thibaut Lamadon, and Elena Manresa (2017) “Discretizing unobserved heterogeneity,” *University of Chicago, Becker Friedman Institute for Economics Working Paper*, No. 2019-16.
- Bonhomme, Stéphane and Elena Manresa (2015) “Grouped patterns of heterogeneity in panel data,” *Econometrica*, Vol. 83, No. 3, pp. 1147–1184.
- Bramoullé, Yann, Habiba Djebbari, and Bernard Fortin (2009) “Identification of peer effects through social networks,” *Journal of econometrics*, Vol. 150, No. 1, pp. 41–55.
- (2019) “Peer effects in networks: A survey.”
- Bramoullé, Yann, Rachel Kranton et al. (2007) “Public goods in networks,” *Journal of Economic Theory*, Vol. 135, No. 1, pp. 478–494.

- Browning, Martin and Jesus Carro (2007) “Heterogeneity and microeconometrics modeling,” *Econometric Society Monographs*, Vol. 43, p. 47.
- Browning, Martin and Jesus M Carro (2010) “Heterogeneity in dynamic discrete choice models,” *The Econometrics Journal*, Vol. 13, No. 1, pp. 1–39.
- (2014) “Dynamic binary outcome models with maximal heterogeneity,” *Journal of Econometrics*, Vol. 178, No. 2, pp. 805–823.
- Bühlmann, Peter and Sara Van De Geer (2011) *Statistics for high-dimensional data: methods, theory and applications*: Springer Science & Business Media.
- Cai, Zongwu, Linna Chen, and Ying Fang (2018) “A semiparametric quantile panel data model with an application to estimating the growth effect of FDI,” *Journal of Econometrics*, Vol. 206, No. 2, pp. 531–553.
- Cai, Zongwu and Qi Li (2008) “Nonparametric estimation of varying coefficient dynamic panel data models,” *Econometric Theory*, Vol. 24, No. 5, pp. 1321–1342.
- Cai, Zongwu and Zhijie Xiao (2012) “Semiparametric quantile regression estimation in dynamic models with partially varying coefficients,” *Journal of Econometrics*, Vol. 167, No. 2, pp. 413–425.
- Canay, Ivan A (2011) “A simple approach to quantile regression for panel data,” *The Econometrics Journal*, Vol. 14, No. 3, pp. 368–386.
- Carkovic, Maria and Ross Levine (2005) “Does foreign direct investment accelerate economic growth?” *Does foreign direct investment promote development*, Vol. 195.
- Carrell, Scott E, Richard L Fullerton, and James E West (2009) “Does your cohort matter? Measuring peer effects in college achievement,” *Journal of Labor Economics*, Vol. 27, No. 3, pp. 439–464.

- Carrell, Scott E, Bruce I Sacerdote, and James E West (2013) “From natural variation to optimal policy? The importance of endogenous peer group formation,” *Econometrica*, Vol. 81, No. 3, pp. 855–882.
- Carro, Jesus M (2007) “Estimating dynamic panel data discrete choice models with fixed effects,” *Journal of Econometrics*, Vol. 140, No. 2, pp. 503–528.
- Cavalcanti, Tiago V de V, Kamiar Mohaddes, and Mehdi Raissi (2011) “Growth, development and natural resources: New evidence using a heterogeneous panel analysis,” *The Quarterly Review of Economics and Finance*, Vol. 51, No. 4, pp. 305–318.
- Chen, Heng, Xuan Leng, and Wendun Wang (2019) “Latent Group Structures with Heterogeneous Distributions: Identification and Estimation.”
- Chen, Jia, Jiti Gao, and Degui Li (2012) “Semiparametric trending panel data models with cross-sectional dependence,” *Journal of Econometrics*, Vol. 171, No. 1, pp. 71–85.
- Chen, Xiaohong (2007) “Large sample sieve estimation of semi-nonparametric models,” *Handbook of econometrics*, Vol. 6, pp. 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*, Vol. 21, No. 1, pp. C1–C68, URL: <https://doi.org/10.1111/ectj.12097>, DOI: 10.1111/ectj.12097.
- 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)*, Vol. 69, No. 4, pp. 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*, Vol. 108, No. 6, pp. 1468–87.

- Cytrynbaum, Max (2020) “Blocked Clusterwise Regression,” *arXiv preprint arXiv:2001.11130*.
- De Boor, Carl (2001) “A practical guide to splines. 2001,” *Appl. Math. Sci.*
- De Boor, Carl, Carl De Boor, Etats-Unis Mathématicien, Carl De Boor, and Carl De Boor (1978) *A practical guide to splines*, Vol. 27: springer-verlag New York.
- De Gregorio, José (1992) “Economic growth in latin america,” *Journal of development economics*, Vol. 39, No. 1, pp. 59–84.
- De Paula, Áureo, Imran Rasul, and Pedro Souza (2018) “Recovering social networks from panel data: identification, simulations and an application,” *Working Paper*.
- Durham, J Benson (2004) “Absorptive capacity and the effects of foreign direct investment and equity foreign portfolio investment on economic growth,” *European economic review*, Vol. 48, No. 2, pp. 285–306.
- Durlauf, Steven N, Andros Kourtellos, and Artur Minkin (2001) “The local Solow growth model,” *European Economic Review*, Vol. 45, No. 4-6, pp. 928–940.
- Fan, Jianqing, Wolfgang Härdle, Enno Mammen et al. (1998) “Direct estimation of low-dimensional components in additive models,” *The Annals of Statistics*, Vol. 26, No. 3, pp. 943–971.
- Fan, Jianqing, Jinchi Lv, and Lei Qi (2011) “Sparse high-dimensional models in economics,” *Annu. Rev. Econ.*, Vol. 3, No. 1, pp. 291–317.
- Galvao, Antonio F and Kengo Kato (2016) “Smoothed quantile regression for panel data,” *Journal of econometrics*, Vol. 193, No. 1, pp. 92–112.
- Galvao Jr, Antonio F (2011) “Quantile regression for dynamic panel data with fixed effects,” *Journal of Econometrics*, Vol. 164, No. 1, pp. 142–157.

- Gilbert, Eric and Karrie Karahalios (2009) “Predicting tie strength with social media,” in *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 211–220, ACM.
- Graham, Bryan S, Jinyong Hahn, Alexandre Poirier, and James L Powell (2018) “A quantile correlated random coefficients panel data model,” *Journal of Econometrics*, Vol. 206, No. 2, pp. 305–335.
- Gu, Jiaying and Stanislav Volgushev (2019) “Panel data quantile regression with grouped fixed effects,” *Journal of Econometrics*, Vol. 213, No. 1, pp. 68–91.
- Hahn, Jinyong and Guido Kuersteiner (2002) “Asymptotically unbiased inference for a dynamic panel model with fixed effects when both  $n$  and  $T$  are large,” *Econometrica*, Vol. 70, No. 4, pp. 1639–1657.
- (2011) “Bias reduction for dynamic nonlinear panel models with fixed effects,” *Econometric Theory*, Vol. 27, No. 6, pp. 1152–1191.
- Hahn, Jinyong and Hyungsik Roger Moon (2010) “Panel data models with finite number of multiple equilibria,” *Econometric Theory*, Vol. 26, No. 3, pp. 863–881.
- Hansen, Bruce E (2008) “Uniform convergence rates for kernel estimation with dependent data,” *Econometric Theory*, Vol. 24, No. 3, pp. 726–748.
- He, Xuming, Zhong-Yi Zhu, and Wing-Kam Fung (2002) “Estimation in a semiparametric model for longitudinal data with unspecified dependence structure,” *Biometrika*, Vol. 89, No. 3, pp. 579–590.
- Henderson, Daniel J, Raymond J Carroll, and Qi Li (2008) “Nonparametric estimation and testing of fixed effects panel data models,” *Journal of Econometrics*, Vol. 144, No. 1, pp. 257–275.
- Hsiao, Cheng (2014) *Analysis of panel data*, No. 54: Cambridge university press.

- Hsiao, Cheng and M Hashem Pesaran (2004) “Random coefficient panel data models.”
- (2008) “Random coefficient models,” in *The Econometrics of Panel Data*: Springer, pp. 185–213.
- Hsiao, Cheng and A Kamil Tahmiscioglu (1997) “A panel analysis of liquidity constraints and firm investment,” *Journal of the American Statistical Association*, Vol. 92, No. 438, pp. 455–465.
- Hsieh, Chih-Sheng and Lung Fei Lee (2016) “A social interactions model with endogenous friendship formation and selectivity,” *Journal of Applied Econometrics*, Vol. 31, No. 2, pp. 301–319.
- Huang, Jian, Joel L Horowitz, and Fengrong Wei (2010) “Variable selection in nonparametric additive models,” *Annals of statistics*, Vol. 38, No. 4, p. 2282.
- Huang, Wenxin, Sainan Jin, and Liangjun Su (2018) “Identifying Latent Grouped Patterns in Cointegrated Panels,” *Econometric Theory*, pp. 1–47.
- Hurwicz, Leonid (1950) “Generalization of the concept of identification,” *Statistical inference in dynamic economic models*, Vol. 10, pp. 245–57.
- Jackson, Matthew O, Brian W Rogers, and Yves Zenou (2017) “The economic consequences of social-network structure,” *Journal of Economic Literature*, Vol. 55, No. 1, pp. 49–95.
- Kawahara, Hiroyuki and Katsumi Shimotsu (2009) “Nonparametric identification of finite mixture models of dynamic discrete choices,” *Econometrica*, Vol. 77, No. 1, pp. 135–175.
- Kato, Kengo, Antonio F Galvao Jr, and Gabriel V Montes-Rojas (2012) “Asymptotics for panel quantile regression models with individual effects,” *Journal of Econometrics*, Vol. 170, No. 1, pp. 76–91.
- Ke, Zheng Tracy, Jianqing Fan, and Yichao Wu (2015) “Homogeneity pursuit,” *Journal of the American Statistical Association*, Vol. 110, No. 509, pp. 175–194.

- Kelejian, Harry H and Ingmar R Prucha (1998) “A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances,” *The Journal of Real Estate Finance and Economics*, Vol. 17, No. 1, pp. 99–121.
- Kline, Brendan and Elie Tamer (2020) “Econometric analysis of models with social interactions,” in *The Econometric Analysis of Network Data*: Elsevier, pp. 149–181.
- Koenker, Roger (2004) “Quantile regression for longitudinal data,” *Journal of Multivariate Analysis*, Vol. 91, No. 1, pp. 74–89.
- Koenker, Roger and Gilbert Bassett (1978) “Regression quantiles,” *Econometrica: journal of the Econometric Society*, pp. 33–50.
- Koenker, Roger, Victor Chernozhukov, Xuming He, and Limin Peng (2017) *Handbook of quantile regression*: CRC press.
- Koenker, Roger and Jose AF Machado (1999) “Goodness of fit and related inference processes for quantile regression,” *Journal of the American Statistical Association*, Vol. 94, No. 448, pp. 1296–1310.
- Kottaridi, Constantina and Thanasis Stengos (2010) “Foreign direct investment, human capital and non-linearities in economic growth,” *Journal of Macroeconomics*, Vol. 32, No. 3, pp. 858–871.
- Laitner, John (2000) “Structural change and economic growth,” *The Review of Economic Studies*, Vol. 67, No. 3, pp. 545–561.
- Lamarche, Carlos (2010) “Robust penalized quantile regression estimation for panel data,” *Journal of Econometrics*, Vol. 157, No. 2, pp. 396–408.
- Lee, Kevin, M Hashem Pesaran, and Ron Smith (1997) “Growth and convergence in a multi-country empirical stochastic Solow model,” *Journal of Applied Econometrics*, Vol. 12, No. 4, pp. 357–392.

- Lee, Lung-Fei (2004) “Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models,” *Econometrica*, Vol. 72, No. 6, pp. 1899–1925.
- (2007) “Identification and estimation of econometric models with group interactions, contextual factors and fixed effects,” *Journal of Econometrics*, Vol. 140, No. 2, pp. 333–374.
- Lee, Lung-fei, Xiaodong Liu, and Xu Lin (2010) “Specification and estimation of social interaction models with network structures,” *The Econometrics Journal*, Vol. 13, No. 2, pp. 145–176.
- Lee, Lung-fei and Jihai Yu (2010) “Estimation of spatial autoregressive panel data models with fixed effects,” *Journal of Econometrics*, Vol. 154, No. 2, pp. 165–185.
- Li, Qi (2000) “Efficient estimation of additive partially linear models,” *International Economic Review*, Vol. 41, No. 4, pp. 1073–1092.
- Lian, Heng, Xinghao Qiao, and Wenyang Zhang (2019) “Homogeneity pursuit in single index models based panel data analysis,” *Journal of Business & Economic Statistics*, pp. 1–44.
- Lin, Chang-Ching and Serena Ng (2012) “Estimation of panel data models with parameter heterogeneity when group membership is unknown,” *Journal of Econometric Methods*, Vol. 1, No. 1, pp. 42–55.
- Lin, Xu (2010) “Identifying peer effects in student academic achievement by spatial autoregressive models with group unobservables,” *Journal of Labor Economics*, Vol. 28, No. 4, pp. 825–860.
- Linton, Oliver and Jens Perch Nielsen (1995) “A kernel method of estimating structured nonparametric regression based on marginal integration,” *Biometrika*, pp. 93–100.
- Liu, Ruiqi, Zuofeng Shang, Yonghui Zhang, and Qiankun Zhou (2019) “Identification and estimation in panel models with overspecified number of groups,” *Journal of Econometrics*.



- Liu, Xiaodong and Lung-fei Lee (2010) “GMM estimation of social interaction models with centrality,” *Journal of Econometrics*, Vol. 159, No. 1, pp. 99–115.
- Liu, Xiaodong, Eleonora Patacchini, and Edoardo Rainone (2017) “Peer effects in bedtime decisions among adolescents: a social network model with sampled data,” *The Econometrics Journal*, Vol. 20, No. 3, pp. S103–S125.
- Liu, Xiaodong, Eleonora Patacchini, and Yves Zenou (2014) “Endogenous peer effects: local aggregate or local average?” *Journal of Economic Behavior & Organization*, Vol. 103, pp. 39–59.
- Liu, Xiaodong and Paulo Saraiva (2015) “GMM estimation of SAR models with endogenous regressors,” *Regional Science and Urban Economics*, Vol. 55, pp. 68–79.
- Lu, Xun and Liangjun Su (2017) “Determining the number of groups in latent panel structures with an application to income and democracy,” *Quantitative Economics*, Vol. 8, No. 3, pp. 729–760.
- Luan, Yihui and Hongzhe Li (2003) “Clustering of time-course gene expression data using a mixed-effects model with B-splines,” *Bioinformatics*, Vol. 19, No. 4, pp. 474–482.
- Mammen, Enno, Bård Støve, and Dag Tjøstheim (2009) “Nonparametric additive models for panels of time series,” *Econometric Theory*, Vol. 25, No. 2, pp. 442–481.
- Manski, Charles F (1993) “Identification of endogenous social effects: The reflection problem,” *The review of economic studies*, Vol. 60, No. 3, pp. 531–542.
- (2000) “Economic analysis of social interactions,” *Journal of economic perspectives*, Vol. 14, No. 3, pp. 115–136.
- Marsden, Peter V and Karen E Campbell (1984) “Measuring tie strength,” *Social forces*, Vol. 63, No. 2, pp. 482–501.

- Mátyás, László and Patrick Sevestre (2013) *The econometrics of panel data: handbook of theory and applications*, Vol. 28: Springer Science & Business Media.
- Mele, Angelo (2017) “A structural model of dense network formation,” *Econometrica*, Vol. 85, No. 3, pp. 825–850.
- 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, pp. 273–292.
- Miao, Ke, Liangjun Su, and Wendun Wang (2020) “Panel threshold regressions with latent group structures,” *Journal of Econometrics*, Vol. 214, No. 2, pp. 451–481.
- Newey, KW and Daniel McFadden (1994) “Large sample estimation and hypothesis,” *Handbook of Econometrics, IV, Edited by RF Engle and DL McFadden*, pp. 2112–2245.
- Newey, Whitney K (1991) “Uniform convergence in probability and stochastic equicontinuity,” *Econometrica: Journal of the Econometric Society*, pp. 1161–1167.
- (1994) “Kernel estimation of partial means and a general variance estimator,” *Econometric Theory*, Vol. 10, No. 2, pp. 1–21.
- (1997) “Convergence rates and asymptotic normality for series estimators,” *Journal of econometrics*, Vol. 79, No. 1, pp. 147–168.
- 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*, Vol. 50, pp. 266–274.
- Oka, Tatsushi and Zhongjun Qu (2011) “Estimating structural changes in regression quantiles,” *Journal of Econometrics*, Vol. 162, No. 2, pp. 248–267.
- Okui, Ryo and Wendun Wang (2018) “Heterogeneous structural breaks in panel data models,” *Available at SSRN 3031689*.

- Phillips, Peter CB and Donggyu Sul (2007) “Transition modeling and econometric convergence tests,” *Econometrica*, Vol. 75, No. 6, pp. 1771–1855.
- Profit, Stefan and Stefan Sperlich (2004) “Non-uniformity of job-matching in a transition economy—A nonparametric analysis for the Czech Republic,” *Applied Economics*, Vol. 36, No. 7, pp. 695–714.
- Robinson, Peter M (1991) “Time-varying nonlinear regression,” in *Economic Structural Change*: Springer, pp. 179–190.
- (2012) “Nonparametric trending regression with cross-sectional dependence,” *Journal of Econometrics*, Vol. 169, No. 1, pp. 4–14.
- Rothenberg, Thomas J (1971) “Identification in parametric models,” *Econometrica: Journal of the Econometric Society*, pp. 577–591.
- Sarafidis, Vasilis and Neville Weber (2015) “A partially heterogeneous framework for analyzing panel data,” *Oxford Bulletin of Economics and Statistics*, Vol. 77, No. 2, pp. 274–296.
- Schmeling, Maik (2009) “Investor sentiment and stock returns: Some international evidence,” *Journal of empirical finance*, Vol. 16, No. 3, pp. 394–408.
- Schumaker, Larry (2007) *Spline functions: basic theory*: Cambridge University Press.
- Seber, George AF (2008) *A matrix handbook for statisticians*, Vol. 15: John Wiley & Sons.
- Seo, Myung Hwan and Yongcheol Shin (2016) “Dynamic panels with threshold effect and endogeneity,” *Journal of Econometrics*, Vol. 195, No. 2, pp. 169–186.
- Sheng, Shuyang (2014) “A structural econometric analysis of network formation games,” *Working Paper*.

- Sim, Nicholas and Hongtao Zhou (2015) “Oil prices, US stock return, and the dependence between their quantiles,” *Journal of Banking & Finance*, Vol. 55, pp. 1–8.
- Song, Song, Ya’acov Ritov, and Wolfgang K Härdle (2012) “Bootstrap confidence bands and partial linear quantile regression,” *Journal of Multivariate Analysis*, Vol. 107, pp. 244–262.
- Sperlich, Stefan, Dag Tjøstheim, and Lijian Yang (2002) “Nonparametric estimation and testing of interaction in additive models,” *Econometric Theory*, Vol. 18, No. 2, pp. 197–251.
- Steinley, Douglas (2006) “K-means clustering: a half-century synthesis,” *British Journal of Mathematical and Statistical Psychology*, Vol. 59, No. 1, pp. 1–34.
- Stone, Charles J (1982) “Optimal global rates of convergence for nonparametric regression,” *The annals of statistics*, pp. 1040–1053.
- Su, Liangjun and Qihui Chen (2013) “Testing homogeneity in panel data models with interactive fixed effects,” *Econometric Theory*, Vol. 29, No. 6, pp. 1079–1135.
- Su, Liangjun and Gaosheng Ju (2018) “Identifying latent grouped patterns in panel data models with interactive fixed effects,” *Journal of Econometrics*, Vol. 206, No. 2, pp. 554–573.
- Su, Liangjun, Zhentao Shi, and Peter CB Phillips (2016) “Identifying latent structures in panel data,” *Econometrica*, Vol. 84, No. 6, pp. 2215–2264.
- Su, Liangjun and Aman Ullah (2011) “Nonparametric and semiparametric panel econometric models: estimation and testing,” *Handbook of empirical economics and finance*, pp. 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*, Vol. 37, No. 2, pp. 334–349.

- Sun, Yiguo, Raymond J Carroll, and Dingding Li (2009) “Semiparametric estimation of fixed effects panel data varying coefficient models,” *Advances in Econometrics*, Vol. 25, pp. 101–129.
- Sun, Yixiao (2005) “Estimation and inference in panel structure models,” *Available at SSRN 794884*.
- Tang, Xiwei, Fei Xue, and Annie Qu (2019) “Individualized Multi-directional Variable Selection,” *Journal of the American Statistical Association*, No. just-accepted, pp. 1–29.
- Tarpey, Thaddeus (2007) “Linear transformations and the k-means clustering algorithm: applications to clustering curves,” *The American Statistician*, Vol. 61, No. 1, pp. 34–40.
- Tibshirani, Robert (1996) “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society: Series B (Methodological)*, Vol. 58, No. 1, pp. 267–288.
- Tibshirani, Robert, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight (2005) “Sparsity and smoothness via the fused lasso,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 67, No. 1, pp. 91–108.
- Ullah, Aman and Nilanjana Roy (1998) “Nonparametric and Semiparametric Econometrics of Panel Data: Econometrics of Victoria, Victoria, British Columbia,” in *Handbook of Applied Economic Statistics*: CRC Press, pp. 599–601.
- 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)*, Vol. 79, No. 1, pp. 5–27.
- (2019) “Multiscale clustering of nonparametric regression curves,” *arXiv preprint arXiv:1903.01459*.
- (2020) “Multiscale clustering of nonparametric regression curves,” *Journal of Econometrics*.

- Wang, Huixia Judy, Zhongyi Zhu, Jianhui Zhou et al. (2009) “Quantile regression in partially linear varying coefficient models,” *The Annals of Statistics*, Vol. 37, No. 6B, pp. 3841–3866.
- Wang, Wei and Lung-Fei Lee (2013) “Estimation of spatial autoregressive models with randomly missing data in the dependent variable,” *The Econometrics Journal*, Vol. 16, No. 1, pp. 73–102.
- Wang, Wuyi, Peter CB Phillips, and Liangjun Su (2018) “Homogeneity pursuit in panel data models: Theory and application,” *Journal of Applied Econometrics*, Vol. 33, No. 6, pp. 797–815.
- (2019) “The heterogeneous effects of the minimum wage on employment across states,” *Economics Letters*, Vol. 174, pp. 179–185.
- Wei, Ying, Xuming He et al. (2006) “Conditional growth charts,” *The Annals of Statistics*, Vol. 34, No. 5, pp. 2069–2097.
- Wooldridge, Jeffrey M (2005) “Fixed-effects and related estimators for correlated random-coefficient and treatment-effect panel data models,” *Review of Economics and Statistics*, Vol. 87, No. 2, pp. 385–390.
- Yang, Chao and Lung-fei Lee (2017) “Social interactions under incomplete information with heterogeneous expectations,” *Journal of Econometrics*, Vol. 198, No. 1, pp. 65–83.
- Zhang, Kevin Honglin (2001) “Does foreign direct investment promote economic growth? Evidence from East Asia and Latin America,” *Contemporary economic policy*, Vol. 19, No. 2, pp. 175–185.
- Zhang, Yingying, Huixia Judy Wang, and Zhongyi Zhu (2019) “Quantile-regression-based clustering for panel data,” *Journal of Econometrics*, Vol. 213, No. 1, pp. 54–67.
- Zhou, Wenyu (2019) “A network social interaction model with heterogeneous links,” *Economics Letters*, Vol. 180, pp. 50–53.

Zhu, Huiming, Lijun Duan, Yawei Guo, and Keming Yu (2016) “The effects of FDI, economic growth and energy consumption on carbon emissions in ASEAN-5: evidence from panel quantile regression,” *Economic Modelling*, Vol. 58, pp. 237–248.